6-25-2022

Data-Driven Analytical Predictive Modeling for Pancreatic Cancer, Financial & Social Systems

Aditya Chakraborty
University of South Florida

Data-Driven Analytical Predictive Modeling for Pancreatic Cancer, Financial & Social Systems

by

Aditya Chakraborty

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics & Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Chris P. Tsokos, Ph.D.
Kandethody M. Ramachandran, Ph.D.
Lu Lu, Ph.D.
Yicheng Tu, Ph.D.

Date of Approval:
June 21, 2021

Keywords: Pancreatic Adenocarcinoma, Parametric Survival Models, Survival Monitoring Indicator (SMI), Subjective Well Being (SWB), Healthcare Business Segment (HBS), Stochastic Modeling in Finance

Dedication

This doctoral dissertation is dedicated to my major professor Dr. Chris P Tsokos, my parents, and my beloved wife for their unconditional love and support.

Acknowledgments

Many people deserve special thanks and acknowledgment for their guidance, support and friendship during my time here at University of South Florida, Department of Mathematics & Statistics. First and foremost is my supervisor Professor Dr. Chris P Tsokos. His willingness and enthusiasm to always explore one step more has made the work of developing this thesis an absolute joy. For his deep insights into research, education and life, I am eternally grateful. I am also grateful to the committee members, Dr. Ram, Dr. Tu, Dr. Lu Lu. My heartiest appreciation also goes to all the faculties and staff of the department of Mathematics & Statistics who contributed in diverse ways to make my experience a token of memory at University of South Florida. I am fully grateful to all my friends here at USF for immense emotional support and encouragement I have experienced from them throughout this whole journey. Most importantly, my heartfelt gratitude goes to my mother Mrs. Anuja Chakraborty, my father Mr. Rabindranath Chakraborty, my sister Mrs. Aditi Chakraborty, and my wife Mrs. Ananya Dey. I wouldn't express my debt of gratitude to them in words. Their incessant love, support, encouragement were the fuel to keep going forward during this endeavor. Last but not the least, I would like to thank Dr. Subhash C Bagui , and Dr. Bikas K Sinha for their constant support and motivation. I would also like to thank my friend Dr. Ruhul Ali Khan for being a constant source of inspiration to me.

Table of Contents

List of Tables

List of Figures

Abstract

Pancreatic cancer is one of the most deathly disease and becoming an increasingly common cause of cancer mortality. It continues giving rise to massive challenges to clinicians and cancer researchers. The combined five-year survival rate for pancreatic cancer is extremely low, about 5 to 10 percent, owing to the fact that a large number of the patients are diagnosed at stage IV when the disease has metastasized. Our study investigates if there exists any statistical significant difference between the median survival times and also the survival probabilities of male and female pancreatic cancer patients at different cancer stages, and irrespective of stages. Also, we investigated if there exists any parametric probability distribution function that best fits the male and female patient survival times in different stages of cancer , irrespective of stages , and performed the parametric survival analysis by using SEER cancer database.

We also have developed a data-driven survival model to predict the survival times of individual pancreatic patients using extreme gradient boosting, which was done based the NIH PLCO (Prostate, Lung, Colorectal and Ovarian) cancer data. Most importantly, we have identified ten risk factors that contribute significantly to the survival of the patient diagnosed with pancreatic adenocarcinoma. Once we identify these risk factors, we rank them with respect to the percentage of contribution to pancreatic cancer. For example, the top three most contributing risk factors of pancreatic adenocarcinoma are the age of the patient (35.5 %), current body mass index (BMI) (24.3 %), and the number of years smoking cigarette (14.93 %). The proposed predictive analytical model is 96.42% accurate. This model has been statistically tested to give excellent predictions.

We have developed a stochastic model that is a function of Stochastic growth intensity factor (SGIF) and a Survival Index $\mathcal{SI}$, that we have introduced. The $\mathcal{SI}$ identifies the survival

rate of pancreatic cancer patients as a function of time, and SGIF monitors the behavior of pancreatic cancer patients at a specific time. The $\mathcal{SI}$ is an important decision-making indicator that conveys three important conditions of the pancreatic cancer patients at a specific time.

- The patients' survival time is increasing.
- The patients' survival remains the same.
- The patients' survival time is decreasing.

The $\mathcal{SI}$ offers a number of important uses on the subject matter. For example, in the case of pancreatic cancer patients, they have three different treatments.

- Chemotherapy only (C)
- Radiation only (R)
- Chemotherapy and Radiation both (C+R)

The proposed $\mathcal{SI}$ can be used to evaluate the effectiveness of the administered treatment to a given patient. That is, if the treatment worsens the patient's cancer, the treatment has no effect on cancer, or the treatment is effective on the cancer. To our knowledge, there is no such analytical model that offers this important evaluation of different treatments. The flexibility of our model lies in the fact that it can incorporate any number of additional treatments. Furthermore, our study categories pancreatic cancer patients from three race groups, Caucasian, African American, and other in utilizing the proposed analytical model. In addition, our analysis is performed at four different stages of pancreatic cancer and three different age groups, 40 to 59, 60 to 79, and 80 and older.

Our statistical analysis includes some other important findings. For example, are there any significant differences in the survival rate between male and female pancreatic cancer patients? We have also found that the Generalized Pareto probability distribution function best characterizes the survival times of pancreatic cancer patients. This finding is important in obtaining a more powerful measurement/estimation of the survival analysis of the subject patients. That is, it gives more accurate results than the classical methods that are commonly used.

We also built predictive models for healthcare business segment (HBS) by utilizing the S&P 500 stock data. We identified the most significant financial and economic indicators, along with the significant interactions, that affect the stock return of the segment by ranking those. We identified the optimum levels of the financial indicators for which the stock price is maximized via analytical modeling. Finally, we developed an analytical procedure that can monitor and predict the Average Weekly Percentage Return (AWPR) of the HBS.

We also developed a data-driven analytical predictive model to predict the subjective well being (SWB)/happiness score by utilizing the world happiness data. The developed analytical model predicts the happiness of an individual based on certain socio-economic factors. After building the model, we ranked the attributable factors, and significant interacting effects according to the percentage of contribution of the happiness score. Finally, we have implemented clustering algorithm to categorize individual countries of the world in three different clusters based on their predicted happiness score. We have compared the happiness scores for different clusters and have done some exploratory data analysis to understand which indicators contribute the most to each cluster. Finally, we validated our clustering mechanism based on three popular machine learning classification algorithms and obtained excellent accuracy.

Chapter 1: A Modern Approach of Survival Analysis of Patients with Pancreatic Cancer

Journal article: “A modern approach of survival analysis of patients with pancreatic cancer,” by Chakraborty A, Tsokos CP. Am J Cancer Res. 2021 Oct 15;11(10):4725-4745. PMID: 34765290; PMCID: PMC8569348.Copyright 2021 by Copyright Holder. Used with permission.

1.1 Introduction

Pancreatic Adenocarcinoma is one of the most fatal human cancers and continues to be a major unsolved health problem at the start of the 21st century. It has been estimated that this disease causes 30,000 deaths per year in the USA [89]. It is the fourth leading cause of cancer death in the USA and leads to an estimated 227,000 deaths per year worldwide. The incidence and number of deaths caused by pancreatic tumours have been gradually increasing, even as incidence and mortality of other common cancers have been declining. Despite developments in detection and management of pancreatic cancer, only about 4% of patients will live 5 years after diagnosis, [134]. The normal pancreas consists of digestive enzyme-secreting acinar cells, bicarbonate-secreting ductal cells, centro-acinar cells that are the geographical transition between acinar and ductal cells, hormone-secreting endocrine islets and relatively inactive stellate cells. The majority of malignant neoplasms of the pancreas are adenocarcinomas. Rare pancreatic neoplasms include neuroendocrine tumours (which can secrete hormones such as insulin or glucagon) and acinar carcinomas (which can release digestive enzymes into the circulation).Specifically, ductal adenocarcinoma is the most common malignancy of the pancreas; this tumour (commonly referred to as pancreatic cancer) presents a substantial health problem, with an estimated 367,000 new cases diagnosed worldwide in 2015 and an associated 359,000 deaths in the same year[80][47]. After the detection of pancreatic cancer, doctors usually perform some additional tests to understand better if cancer has been spread or the locations of spreading areas of the cancer. Imaging tests, such as a PET scan, help doctors identify the presence of cancerous growths. With these tests, doctors try to establish cancer's stage of a given patient with pancreatic cancer. Staging helps explicate the advancement of cancer. It also assists doctors in deciding treatment options. Once a diagnosis has been made, the doctor allocates the patient a stage based on the following test results:

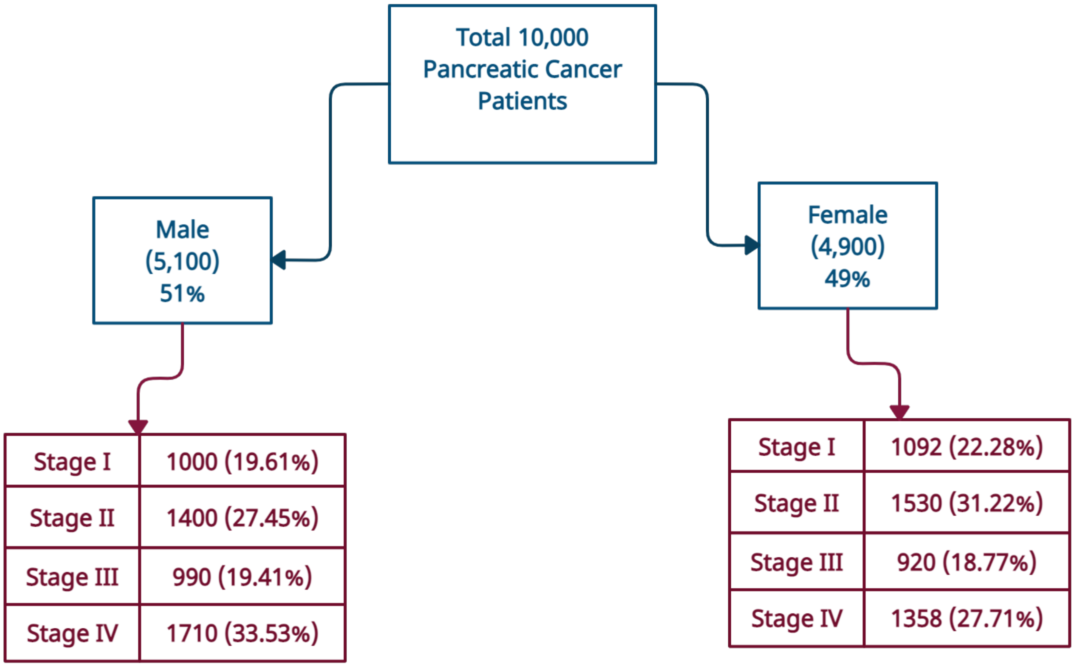
1. Stage I: Tumors exist solely in the pancreas.

2. Stage II: Tumors have spread to adjacent abdominal tissues or lymph nodes.

3. Stage III: The cancer has spread to major blood vessels and lymph nodes.

4. Stage IV: Tumors have spread to other organs, such as the liver, lung, bone, etc.

Although in most of the cases, pancreatic cancer disease remains irremediable, most researches studying this type of cancer, have focused on how to improve the survival times of patients diagnosed with pancreatic cancer in different stages. The Kaplan-Meier (KM) method has been widely used for analyzing cancer survivorship data in recent times due to the simplicity of its usage. It is often used to compare the survival difference of several groups of patients based on the log-rank test of the null hypothesis that there is no significant difference among the groups. Our study presents a parametric and non-parametric survival analysis of the survival times of patients diagnosed with Pancreatic Cancer. We believe that finding the unique probability distribution that characterizes the probabilistic behavior of the survival times is important so that we can proceed to obtain the survival function that is driven by the given data. Such an analysis is more powerful than the non-parametric approach. Feigl and Zelen,[46] have shown that assumption of exponential distribution works well for studying some of the survival of cancer-related studies, [141][140][70]. However, assuming such a probability distribution without justification might lead to misleading results. Thus, it is important to identify the correct probability distribution of the survival times of patients among any number of groups (for male/female, different age groups,etc.). In the present study, we identify the probability distribution that fits the survival times the best and proceed to obtain the survival function of male and female patients in four different stages. We also compare our results with the commonly used Kaplan-Meier (KM) method. The structure of the paper will be as follows: In Section 2.1, we provide the data discussion and perform the non-parametric Wilcoxon test to investigate if there exist any significant difference between the male and female patients at any individual stages. In section 2.2, we discuss the stage based descriptive analysis with graphical representation. In section 3, we discuss in detail the parametric survival analysis of pancreatic cancer patients at different stages. In section 4, we investigate the significant difference of overall survival times of male and female patients by log-rank test [97][81] , and discuss in detail about the overall parametric survival analysis of patients irrespective of stages. We also describe elaborately the parameter estimation procedure of GP probability distribution in Section 4.3. In Section 5 we present the KM estimate and compare the median survival times of patients using the descriptive, parametric, and non-parametric methods. In Section 6, we compare the survival probability estimates of patients using the Generalized Pareto (GP) probability distribution and non-parametric KM estimates. Sections 7 and 8 provide results & discussion, and conclusion, respectively.



Stage I	1000 (19.61%)
Stage II	1400 (27.45%)
Stage III	990 (19.41%)
Stage IV	1710 (33.53%)

Stage I	1092 (22.28%)
Stage II	1530 (31.22%)
Stage III	920 (18.77%)
Stage IV	1358 (27.71%)

Figure 1.1: Pancreatic Cancer Data Sorted by Gender and Stages

1.2 Methodology

1.2.1 Data Description

The data for our study has been extracted from the Surveillance, Epidemiology and End Results (SEER) database. The data contains information on patients diagnosed with pancreatic adenocarcinoma . We are concerned with the survival time (in months) and cause-specific death (deaths due to pancreatic cancer) for each patient. The survival time of patients is one of the most crucial factors used in all cancer research. It is necessary to evaluate the severity of cancer, which helps to decide the prognosis and help identify the correct treatment methods. We considered a random sample of 10,000 patients diagnosed with pancreatic cancer including male and female. A schematic diagram of the data used in this study with additional details is shown in Figure 1.2, below. As the following schematic diagram illustrates, in our dataset, we have information on survival times regarding 5,100 male and 4,900 female patients diagnosed with pancreatic cancer.

Before we proceed with performing the parametric analysis of the survival times of patients, we need to investigate whether there is a difference in the true median survival times of genders, i.e., male and female patients in different stages of cancer. For this purpose, We use the two-samples Wilcoxon Rank Sum test using the following hypothesis.

H_0: There is no significant difference between the true median survival times of male (μ_M) and true median survival times of female (μ_F) patients at stage $i.i = 1, 2, 3, 4$. That is, $\mu_M = \mu_F$

Vs.

H_1: Differences exist between male and female median survival times at stage i. That is, $\mu_M \neq \mu_F$.

After we analyze the data for male and female patients in each stages, we proceed to perform the combined analysis for all stages, classified by gender. The following Table 1.1 illustrates the test results along with the p-values in different stages for male and female pancreatic cancer patients.

Table 1.1: Wilcoxon Test Results for Different Stages, Classified by Gender

Stages	**P-Values**	**Result**
I	0.75	Difference does not Exist
II	0.25	Difference does not Exist
III	0.84	Difference does not Exist
IV	0.001	Difference Exists

As, results of the above Table 1.1 suggests, there does not exist significant difference between the male and female pancreatic cancer patient survival times in stage I, stage II , and stage III. However, in stage IV, the difference is significant. In the next section, we proceed to identify the parametric probability distributions and survival functions of the survival times of patients along with some important descriptive statistics.

1.2.2 Descriptive Analysis of Pancreatic Cancer Patients in Different Stages-A Gender Based Classification

We plotted the histogram and probability density function (pdf) to investigate the distribution of the survival times of patients in different stages, as shown in the following Figures. We see that the probability distribution of the survival times are right-skewed. The following Table 1.2, illustrates the different descriptive statistics for male and female patients in four different stages.

Table 1.2: Descriptive Statistics of Survival Time (in month) of Pancreatic Cancer Patients Classified by Gender in Different Stages.

Gender	Mean	Median	Std. Dev.	Skewness	Kurtosis	Std. Error
Combined (Stage I)	30.6	20	31.5	1.33	1.14	0.76
Combined (Stage II)	21.44	14	23.50	2.14	5.10	.33
Combined (Stage III)	16.92	8	14.71	3.73	20.01	.37
Male (Stage IV)	6.7	3	12.73	4.78	30.44	.18
Female (Stage IV)	7.50	3.11	13.67	4.63	27.80	.20

We now proceed to identify the most appropriate probability distributions that drives the survival times of patients in different stages (I , II, III , and IV), classified by gender. We came to know from last section that there does not exist any significant difference between male and female survival times in stages I, II, and III. However, we found significant difference in survival times of male and female patients in stage IV. We have obtained the best fits for each stages and estimated their individual parameter estimates. Identification of the most suitable probability distribution is crucial, since it gives the better survival probability estimates for both male and female patients in each of the stages that is driven by the specific probability distribution. Once, we obtain the parameter estimates from the probability distributions at each of the stages, we can obtain the probability density functions (pdfs), cumulative distribution functions (cdfs), and parametric survival function ($S(t)$) driven by the specific probability distribution for male and female patients individually.

1.2.3 Parametric Analysis of Pancreatic Cancer Survival Time for Different Stages

Johnson (1949) [74] proposed systems of different frequency curves based on transformations of the following form

$$z = \gamma + \delta f\Big(\frac{x-\zeta}{\lambda}\Big) \ ,$$

where z is a unit Normal variable, f is a function taking different forms S_L, S_B, and S_U. Our data in Stage I follows Johnson S_B probability distribution with parameters γ (shape parameter), δ (shape parameter), ζ (location parameter), and λ (scale parameter). In Stage II, and Stage III, the data follows a generalized extreme value (GEV) probability distribution. Chakraborty & Tsokos [28] describes in detail about the parameter estimation procedure of acute myeloid leukemia cancer data modeled by GEV probability distribution using probability weighted moment (pwm). In Stage IV, the data follows a generalised pareto (GP) probability distribution. In section 4.3, we discuss in detail about the parameter estimation procedure of generalized pareto (GP) probability distribution for overall survival times of the patients. We now proceed to discuss the parameter estimation process of the Johnson S_B distribution in Stage I. SFIEKIERS [119] has given a brief summary about the parameter estimation procedure of Johnson S_B probability distribution using moments of transformed values of a random Variable. Let T be a random variable denoting the survival times of patients in Stage I. Then, the p.d.f of T is given by,

$$f(t) = \frac{\delta}{\sqrt{2\pi}} \frac{\lambda}{(t-\zeta)(\lambda+\zeta-t)} exp\Big[-\frac{1}{2}\Big(\gamma + \delta ln\Big(\frac{t-\zeta}{\lambda+\zeta-t}\Big)\Big)^2\Big],$$

where

$$\zeta < t < \zeta + \lambda \ , \ -\infty < \zeta < \infty \ , \ \lambda > 0 \ , \ -\infty < \gamma < \infty \ , \ \delta > 0.$$

From the data, the extreme order statistics t_{min} and t_{max} are determined. In our case, in Stage I, $t_{min} = 0$, and $t_{max} = 155$. Since, ζ , and λ are the location and spread parameters respectively, $\hat{\zeta} = t_{min} = 0$ (since, the minimum value of the survival time t is 0) , and $\hat{\lambda} = (t_{max} - t_{min}) = (155 - 0) = 155 = t_{max}$.

Given the estimated values, $\hat{\zeta}$, and $\hat{\lambda}$, we proceed with the following transformation, that is, the values of t_i are transformed to:

$$f_i = ln\Big(\frac{t_i - \hat{\zeta}}{\hat{\lambda} + \hat{\zeta} - t_i}\Big).$$

The estimates of the other parameters $\hat{\gamma}$, and $\hat{\delta}$ take the following form:

$$\hat{\gamma} = -\frac{\bar{f}}{S_f}$$

and

$$\hat{\delta} = \frac{1}{S_f},$$

where $\bar{f} = \frac{\sum_i f_i}{n}$, and $S_f = \sqrt{\frac{\sum_i (f_i - \bar{f})^2}{n}}$.

The validity of the model assumptions are justified using the goodness of fit tests. Soukissian [122] fitted a Johnson S_B probability distribution to the wind speed data and used Kolmogorov–Smirnov (K–S), and Anderson–Darling (A–D) tests to to justify goodness of fit assumptions. We followed the same approach using Kolmogorov–Smirnov (K–S), Anderson–Darling (A–D) , and Cramér–von Mises (CVM) goodness of fit tests.

The following Table 1.3, provides the goodness of fit tests results along with the p-values for all probability distributions in the four different stages.

Table 1.3: Goodness of Test for Four Stages.

Stages	Gender	Prob. Distribution	GOF Tests	p-Values
I	Combined	Johnson S_B	A-D	.11
			K-S	.13
II	Combined	GEV	A-D	.27
			K-S	.21
III	Combined	GEV	A-D	.09
			K-S	.1
IV	Male	GPD	CVM	.22
			K-S	.18
IV	Fenale	GPD	CVM	.19
			K-S	.17

Table 1.4: Probability Distributions and Parameter Estimates of Survival Time (in month) of Pancreatic Cancer Patients for Different Stages.

Stages	Gender	Probability Distributions	Parameter Estimates
I	Combined	4-P Johnson S_B	$\hat{\gamma} = 1.2, \hat{\delta} = 0.62, \hat{\lambda} = 155, \hat{\zeta} = 0$
II	Combined	Gen. Extreme Value (GEV)	$\hat{\mu} = 10.18, \hat{\sigma} = 10.83, \hat{k} = 0.32$
III	Combined	Gen. Extreme Value (GEV)	$\hat{\mu} = 5.54, \hat{\sigma} = 6.07, \hat{k} = 0.37$
IV	Male	3-P Gen. Pareto (GP)	$\hat{\mu} = 0, \hat{\sigma} = 4.12, \hat{k} = 0.25$
IV	Female	3-P Gen. Pareto (GP)	$\hat{\mu} = 0, \hat{\sigma} = 4.63, \hat{k} = 0.41$

As the p-values shown in Table 1.3 of the given data, we fail to reject the fact, that the observations (survival times) follow the specified probability distributions in each of the four stages. The following Table 1.4, provides the specific probability distributions in each stages and their individual parameter estimates (approximate), classified by gender.

The following Table 1.5, illustrates the analytical forms of the probability density functions of male and female patients for the different stages, with the parametric estimates.

Table 1.5: Probability Distributions with their Parameter Estimates of the Survival Times (in months) of Pancreatic Cancer Patients Classified by Gender for Different Stages.

Gender	Analytical Forms
Combined (I)	$f(t) = \frac{.62}{\sqrt{2\pi}} \frac{155}{t(155-t)} exp\Big[-\frac{1}{2}\big(1.2 + .62ln\big(\frac{t}{155-t}\big)\big)^2\Big]$
Combined (II)	$f(t) = \frac{1}{10.83} exp\Big[-\big(1-.32\big(\frac{t-10.18}{10.83}\big)\big)^{3.125}\Big]\big(1-.32\big(\frac{t-10.18}{10.83}\big)\big)^{2.125}$
Combined (III)	$f(t) = \frac{1}{6.07} exp\Big[-\big(1-.32\big(\frac{t-5.54}{6.07}\big)\big)^{2.7}\Big]\big(1-.32\big(\frac{t-5.54}{6.07}\big)\big)^{1.7}$
Male (IV)	$f(t) = \frac{1}{4.12}\Big[1-.25\big(\frac{t}{4.12}\big)\Big]^3$
Female (IV)	$f(t) = \frac{1}{4.63}\Big[1-.34\big(\frac{t}{4.63}\big)\Big]^{1.44}$

The following Figure illustrates the probability density function (pdf) and cumulative distribution function (cdf) of the patients in stage I.

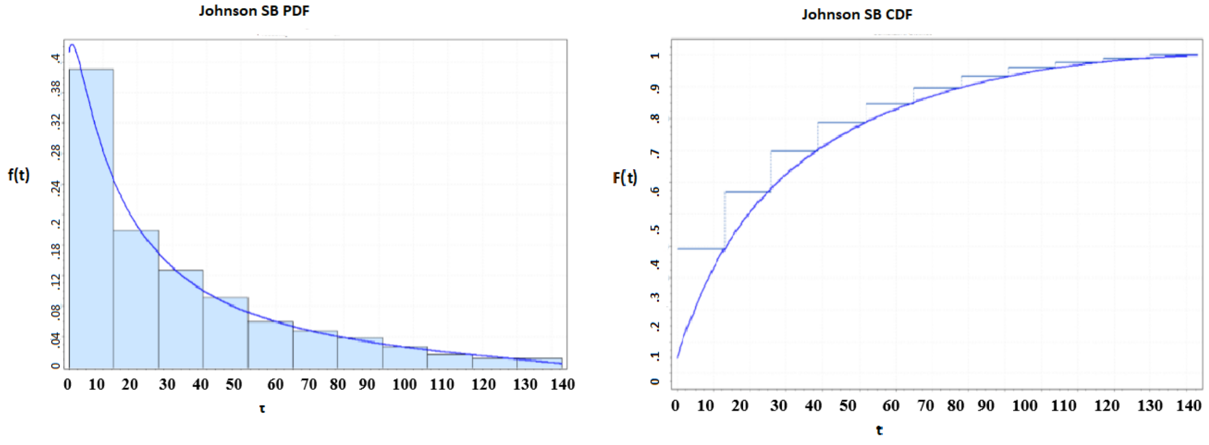


Figure 1.2: Histogram, cdf and Probability Density of Survival Times of Pancreatic Cancer Patients in Stage I

The following Figure 1.3 shows the histogram, pdf and cdf plots of stage II pancreatic cancer patients.

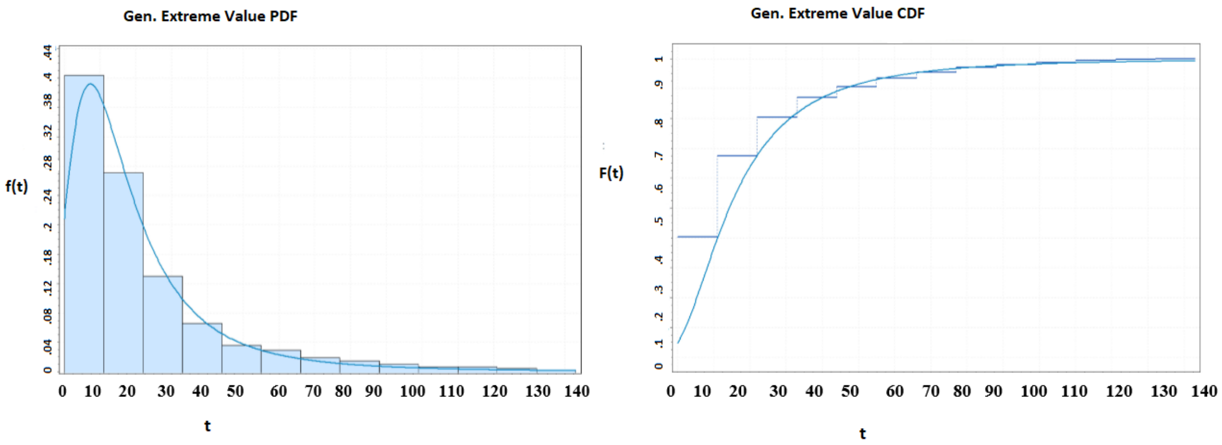


Figure 1.3: Histogram, cdf and Probability Density of Survival Times of Pancreatic Cancer Patients in Stage II

The following Figure 1.4 shows the histogram, pdf and cdf plots of stage III pancreatic cancer patients.

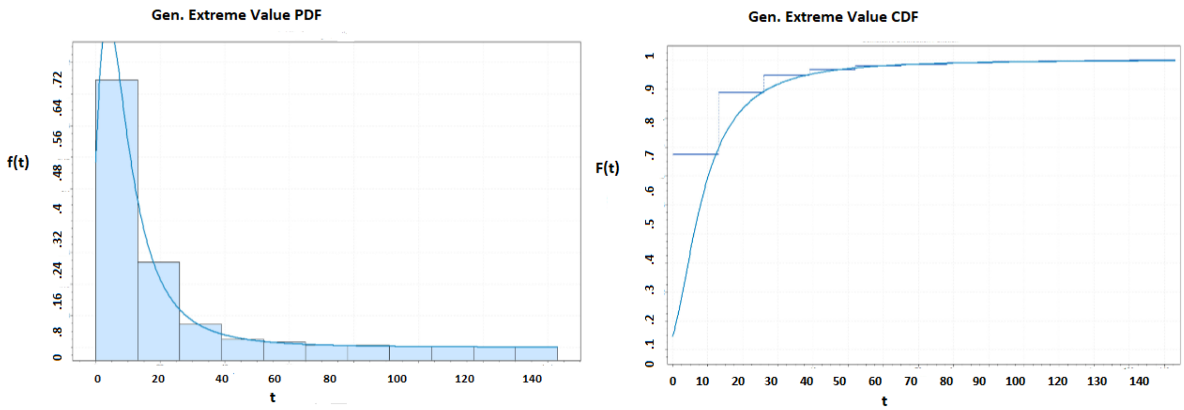


Figure 1.4: Histogram, cdf and Probability Density of Survival Times of Pancreatic Cancer Patients in Stage III

The following two figures (Figure 1.5 and Figure 1.6) describe the histogram, pdf, and cdf of male and female survival time respectively in stage IV.

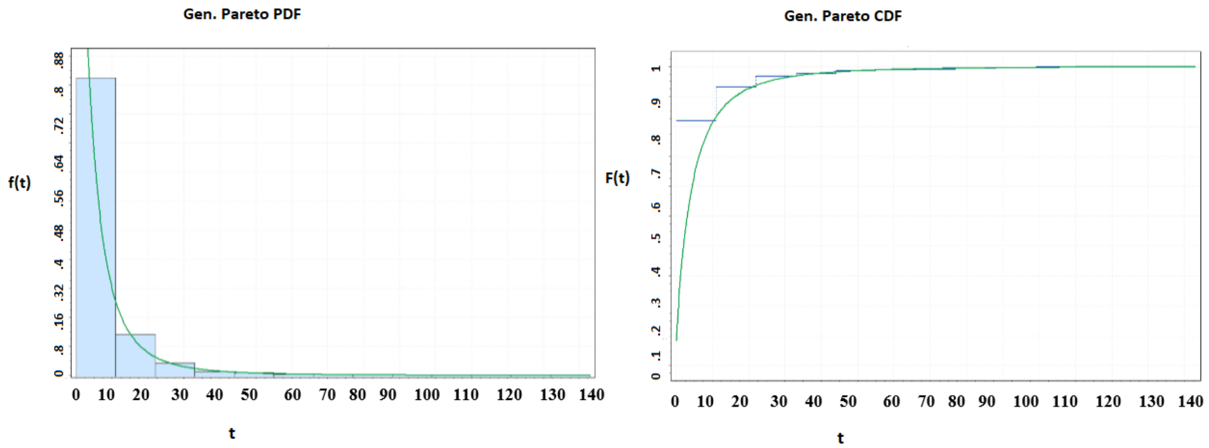


Figure 1.5: Histogram, cdf and Probability Density of Survival Times of Male Pancreatic Cancer Patients in Stage IV

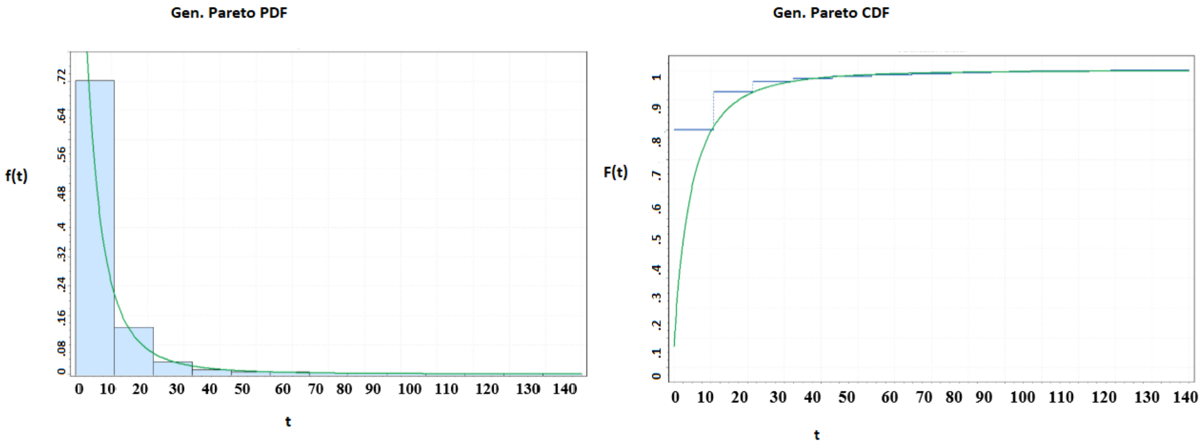


Figure 1.6: Histogram, cdf and Probability Density of Survival Times of Female Pancreatic Cancer Patients in Stage IV

1.3 Parametric Survival Analysis for Different Stages

Once we have the analytical structures of the survival times of patients in different stages, driven by different parametric probability distributions, we can express the survival function $S(t)$ analytically as a function of the cumulative distribution function (cdf). Now we proceed to express the analytical forms of the survival functions for the four different stages. The estimate of the parametric survival function of patients diagnosed with pancreatic cancer in Stage I is given by

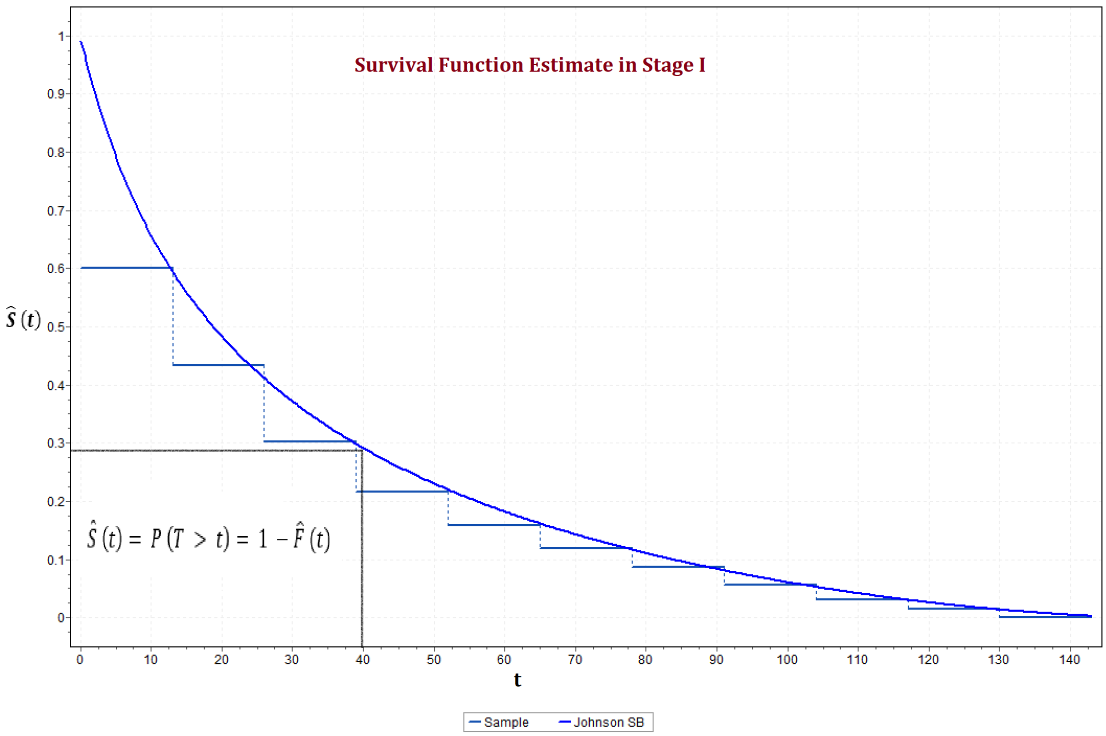


Figure 1.7: Parametric Survival Plot of Pancreatic Cancer Patients in Stage I

$$
\begin{aligned}
\hat{S}_I(t;\hat{\zeta},\hat{\lambda},\hat{\gamma},\hat{\delta}) &= 1-\hat{F}_I(t;\hat{\zeta},\hat{\lambda},\hat{\gamma},\hat{\delta}) \\
&= 1-\Phi\Big[\hat{\gamma}+\hat{\delta}ln\Big(\frac{t-\hat{\zeta}}{\hat{\lambda}-t+\hat{\zeta}}\Big)\Big] \qquad (1.1)\\
&= 1-\Phi\Big[1.2+.62ln\Big(\frac{t}{155-t}\Big)\Big] \;\;,\;\; t\geq 0.
\end{aligned}
$$

where $\Phi(\cdot)$ is the cdf of a standard normal probability distribution and $\hat{F}_I(t;\hat{\zeta},\hat{\lambda},\hat{\gamma},\hat{\delta})$ is the estimated cdf of Johnson S_B probability distribution. The survival function $\hat{S}(\cdot;\cdot)$ can be used to estimate the probability that a patient diagnosed with pancreatic cancer would survive beyond time t, which is denoted by $P(T \geq t)$. For example, we can compute the probability that a male patient diagnosed with pancreatic cancer would survive beyond 30 months. For example, for $t = 40$ in equation (1.1), we estimate the probability is 0.29 approximately. Thus, we can infer that a randomly chosen patient classified in Stage I with pancreatic cancer has a 29% chance of survival beyond 40 months, as shown by Figure 1.7.

Similarly, the estimate of parametric survival function of patients, driven by GEV probability distribution function diagnosed with pancreatic cancer in Stage II is given by

$$\begin{aligned}\hat{S}_{II}(t;\hat{\mu},\hat{\sigma},\hat{k}) &= 1 - \hat{F}_{II}(t;\hat{\mu},\hat{\sigma},\hat{k}) \\ &= 1 - exp\Big[-\Big(1 - k\Big(\frac{t-\hat{\mu}}{\hat{\sigma}}\Big)\Big)^{\frac{1}{k}}\Big] \\ &= 1 - exp\Big[-\Big(1 - .32\Big(\frac{t-10.18}{10.83}\Big)\Big)^{\frac{1}{.32}}\Big] \ , \ \ t \geq 10.18.\end{aligned} \tag{1.2}$$

As the following survival plot for Stage II patients illustrates, patients in stage II have comparatively lower survival probability than stage I patients, which is quite natural. With reference to the last example, we can predict the survival probability as 13% for a Stage II patient, surviving beyond 40 months.

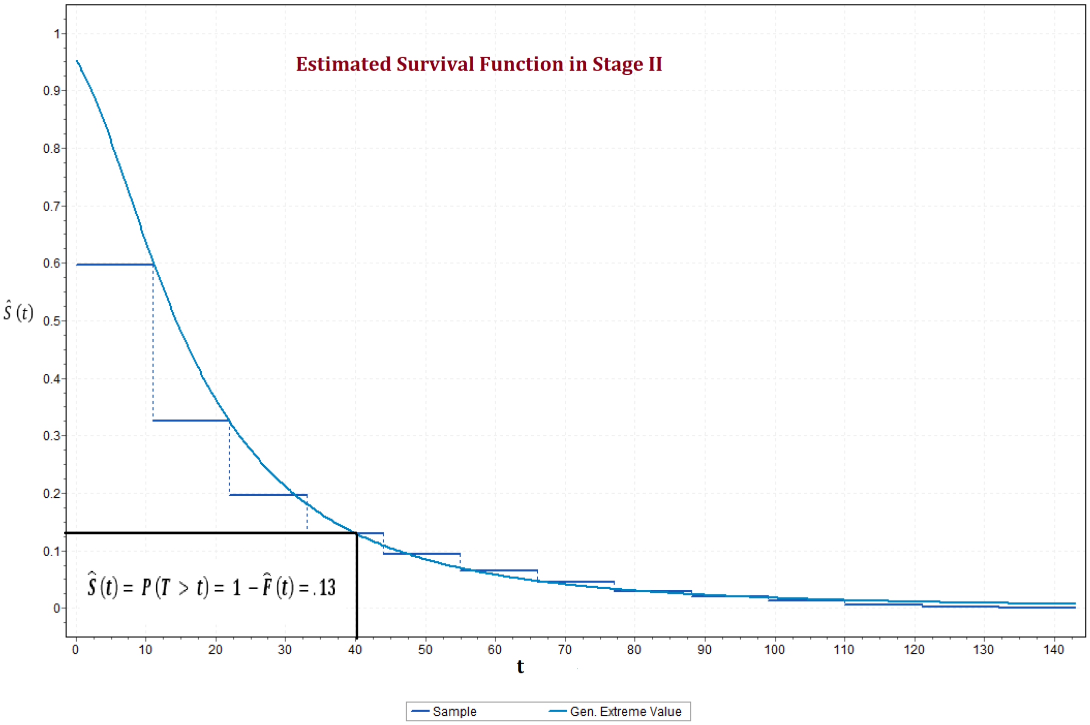


Figure 1.8: Parametric Survival Plot of Pancreatic Cancer Patients in Stage II

Now we proceed to express the GEV in analytical form for the stage III patients in a similar manner. The survival function at stage III can be given by,

$$\begin{aligned}\hat{S}_{III}(t;\hat{\mu},\hat{\sigma},\hat{k}) &= 1-\hat{F}_{III}(t;\hat{\mu},\hat{\sigma},\hat{k})\\ &= 1-exp\Big[-\Big(1-k\Big(\frac{t-\hat{\mu}}{\hat{\sigma}}\Big)\Big)^{\frac{1}{k}}\Big] \qquad (1.3)\\ &= 1-exp\Big[-\Big(1-.37\Big(\frac{t-5.54}{6.07}\Big)\Big)^{\frac{1}{.37}}\Big]\ ,\quad t\geq 5.54.\end{aligned}$$

From the following Figure 1.9, we see that the survival probability is decreasing and it is approximately 5% for a randomly chosen patient who will survive beyond t = 40 months after the patient is diagnosed with pancreatic cancer, Stage III.

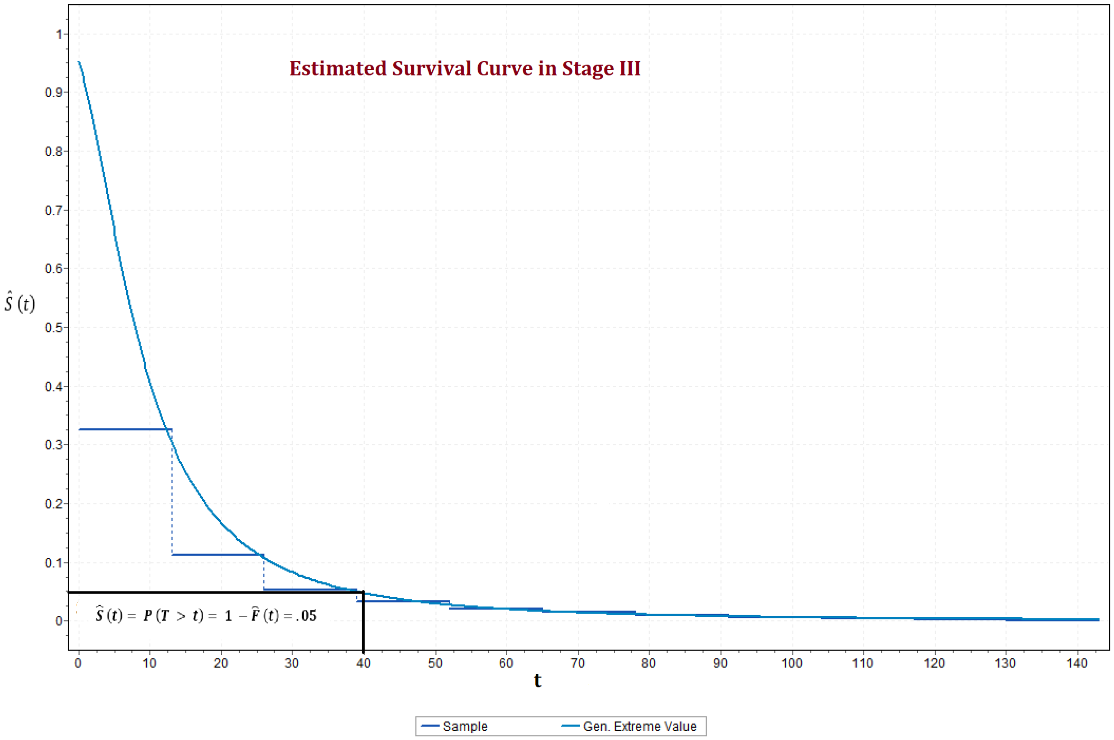


Figure 1.9: Parametric Survival Plot of Pancreatic Cancer Patients in Stage III

Results from Table 1.1, suggested that there is a significant difference between the true mean survival times of stage IV patients, classified by gender. Thus, we now proceed to express the analytical forms of the survival times for male and female patients separately at Stage IV. The parametric survival function, driven by GPD, at stage IV male patients is expressed as,

$$
\begin{aligned}
\hat{S}_{IV}(t;\hat{\mu},\hat{\sigma},\hat{k}) &= 1-\hat{F}_{IV}(t;\hat{\mu},\hat{\sigma},\hat{k}) \\
&= 1-\Big[1-\Big(\Big[1+\hat{k}\Big(\frac{t-\hat{\mu}}{\hat{\sigma}}\Big)\Big]^{-\frac{1}{\hat{k}}}\Big)\Big] \qquad (1.4)\\
&= \Big(\Big[1+.25\Big(\frac{t}{4.12}\Big)\Big]^{-\frac{1}{.25}}\Big)\ ,\ \ t\geq 0.
\end{aligned}
$$

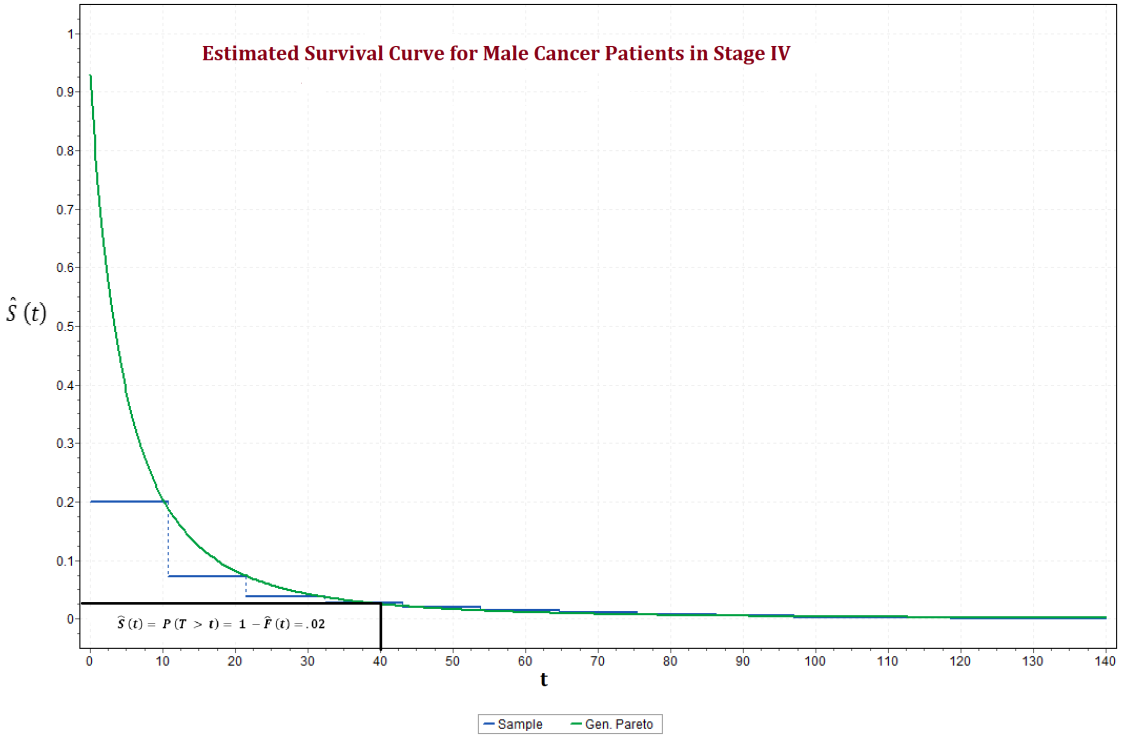


Figure 1.10: Parametric Survival Plot of Male Pancreatic Cancer Patients in Stage IV

Similarly, The parametric survival function, driven by GPD, at stage IV female patients is given by,

$$
\begin{aligned}
\hat{S}_{IV}(t;\hat{\mu},\hat{\sigma},\hat{k}) &= 1-\hat{F}_{IV}(t;\hat{\mu},\hat{\sigma},\hat{k}) \\
&= 1-\Big[1-\Big(\Big[1+\hat{k}\Big(\frac{t-\hat{\mu}}{\hat{\sigma}}\Big)\Big]^{-\frac{1}{\hat{k}}}\Big)\Big] \qquad (1.5)\\
&= \Big(\Big[1+.41\Big(\frac{t}{4.63}\Big)\Big]^{-\frac{1}{.41}}\Big)\ ,\ \ t\geq 0.
\end{aligned}
$$

As, the following two figures indicates, the survival probabilities are extremely low (2% for male patients and 3% for female patients) for surviving beyond $t = 40$ months after the diagnosis at Stage IV.

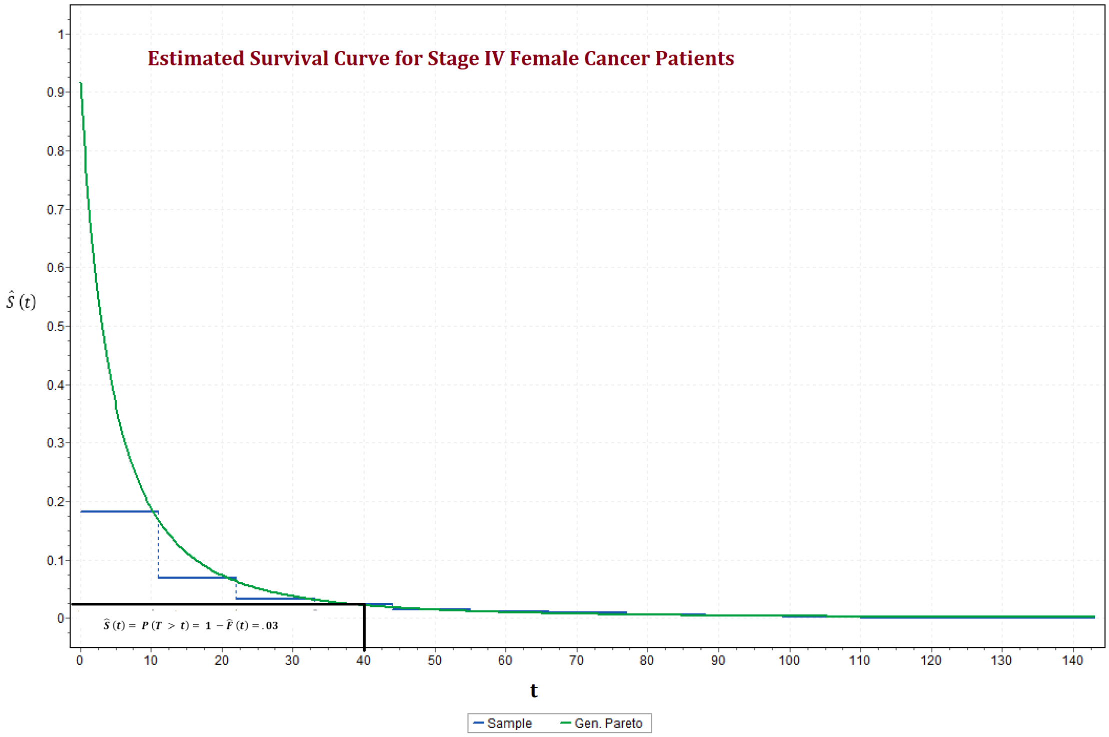


Figure 1.11: Parametric Survival Plot of Female Pancreatic Cancer Patients in Stage IV

In the next section, we will discuss in detail the combined analysis of male and female patients irrespective of stages.

1.4 Parametric Analysis of The Survival Times of Patients with Pancreatic Cancer-A Combined Analysis

So far, we discussed about the parametric analytical forms of the survival times of patients in different stages. We also computed the survival functions of patients in different stages. We found that there is no significant different in the survival times of male and female patients except stage IV. We now proceed to do the same for the combined data, irrespective of stage. At first, we will check if there exists significant difference between the true mean survival times of male and female pancreatic cancer patients. For this purpose, we use the log rank test and found that there is insufficient sample evidence to reject the hypothesis that the distribution of mean survival times between the Male and Female patients diagnosed with pancreatic cancer is the same. The following Figure 1.12 illustrates the behavior of overall survival curves of male and female patients. The male and female survival curves are highlighted in blue and yellow, respectively.

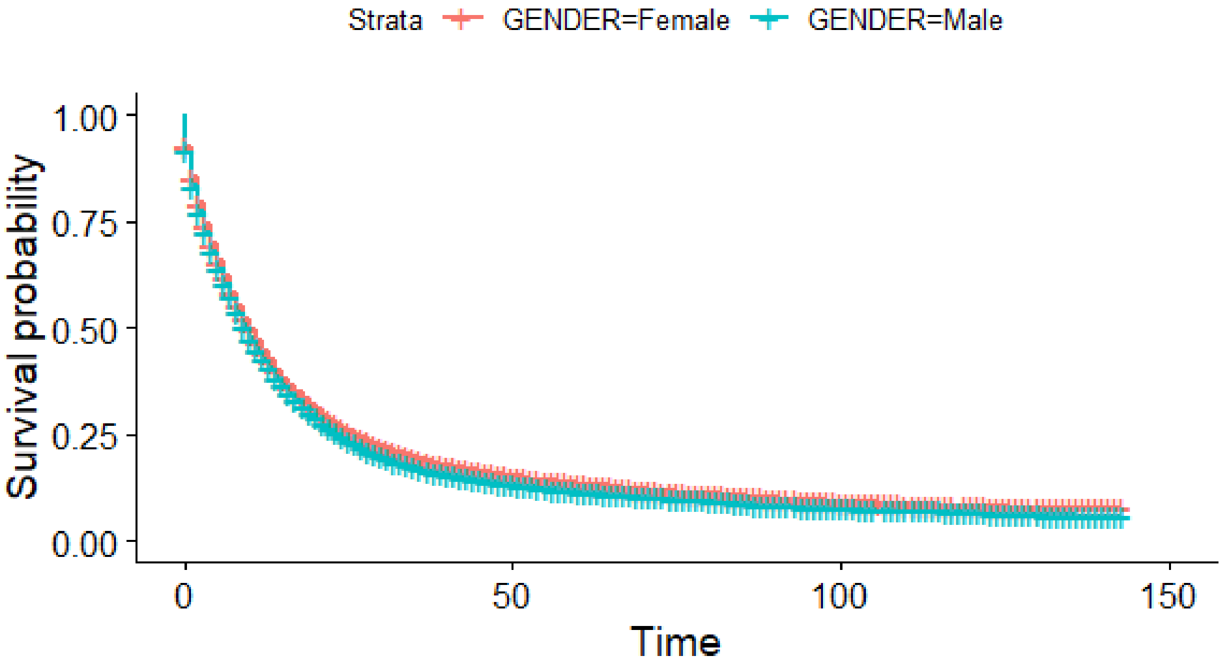


Figure 1.12: Log-rank Test for Difference in Survival Times of Gender.

As the above Figure 1.12 illustrates, the survival curve of males (skyblue) and the survival curve of females (red), are almost identical which implies that they exhibit similar characteristics.

1.4.1 Descriptive Statistics of the Survival Times of Pancreatic Cancer Patients

In this section, we proceed to analyze the combined survival data descriptively. We plotted the histogram and probability density function (pdf) to investigate the probability distribution of the survival times of pancreatic cancer patients. We can see that the probability distribution of the overall survival time is right-skewed. Table 1.6 displays the descriptive statistics of the overall survival times for pancreatic cancer patients. We see that the mean (average) survival times patients diagnosed with pancreatic cancer is 18 months. It implies that a randomly chosen patient diagnosed with pancreatic cancer is expected to survive for 18 months on an average. Also, the median survival time is nine months, which implies that the probability/chance of survival of a male or female patient beyond nine months, is approximately 50%. A negative (less than zero) skewed value implies that data distribution is left or negatively skewed, and a positive skewed value suggests that data is right or positive skewed. Thus, the positive skewed value of 3.07, as shown in Table 1.6, for patients diagnosed with pancreatic cancer, is further evidence to support the right-skewed behavior of the data, as shown in Figure 14 above. Kurtosis supports the assessment of the extreme values of the data, and its positive value illustrates a leptokurtic behavior of the distribution. In contrast, a negative value shows a platykurtic behavior of the data distribution. Thus, the kurtosis value of 12.67 in Table 1.6 attests to the leptokurtic behavior of the survival data.

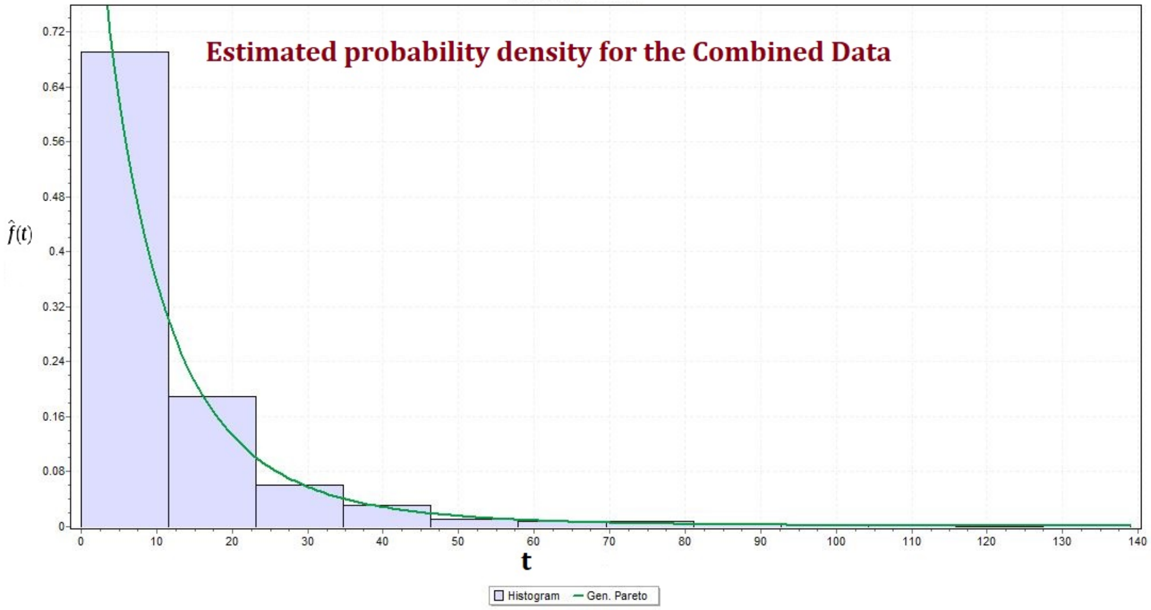


Figure 1.13: Histogram and Probability Density of Survival Times of Combined Pancreatic Cancer Patients

Table 1.6: Descriptive Statistics of Survival Time (in month) of Overall Pancreatic Cancer Patients Classified.

Descriptive Statistics	Measures
Mean	10.87
Median	6
Std. Dev.	14.63
Skewness	3.07
Kurtosis	12.67
Std. Error	.24

Table 1.6, above illustrates the different descriptive statistics for survival times of all patients combined, diagnosed with pancreatic cancer.

1.4.2 Three Parameter Generalized Pareto (GP) Probability Estimation of The Survival Times of Patients with Pancreatic Cancer

We perform a parametric analysis of the survival times of patients diagnosed with pancreatic cancer to identify the underlying probability distribution, which characterizes the probabilistic behavior of the survival times of patients (both genders). In the attempt to obtain the best-fitted probability distribution, a number of classical distributions were tested to fit the data. We used the famous Anderson-Darling test [3] and Cramér–von Mises test [33] identify the best probability distribution function that characterizes the probabilistic behavior of the survival times patients. Also, we estimate the expected survival times and median survival times that is driven by the best fitted probability distribution. The best-fitted probability distribution that characterizes the probabilistic behavior of the survival times of the male and female patients accurately is the three parameter (3-P) Generalized Pareto (GP) probability distribution. Table 1.7 below shows the goodness of fit (GOF) results of the 3-P GPD distribution.

Table 1.7: Goodness-of-fit Test of the GPD of the Survival Times of Male and Female.

Statistical Tests	**P-Values Male**	**P-Values Female**
Kolmogorov-Smirnov	0.27	.38
Cramér–von Mises	0.22	.18

The above results show that we fail to reject the null hypothesis that the subject data (survival times for males and females) follow a GP probability distribution. In this section, we define the probability density function (pdf) of the Generalized Pareto distribution (GPD) and the statistical approach to obtain approximate estimates of its parameters. In the domain of probability theory and statistics, the GPD is a family of continuous probability distributions developed based on the extreme value theory,[38]. The GPD is a generalization of the Pareto distribution (PD). The PD was studied extensively by Arnold (1983), and the problem of estimation in the PD was considered by Arnold and Press (1989) [4]. It has been used broadly by several researchers to model data arising from several fields. Hosking and Wallis [66] used the GPD to model the annual maximum flood of the River Nidd at Hunsingore, England. Grimshaw [58] used it to model tensile strength data from a random sample of nylon carpet fibers. Other estimation procedures and uses of the GPD in extreme value analysis using numerical optimization have been illustrated by Castillo and Daoudi [39]. Let T be a random variable following GPD with location parameter μ , scale parameter $\sigma > 0$ and shape parameter k. That is, $T \sim GPD(\mu, \sigma, k)$ with the domain $\mu \leq x \leq \mu - \frac{\sigma}{k}$, when $k < 0$ and $\mu \leq t < \infty$, when $k \geq 0$. Then, the probability density function (pdf) of T is given as follows:

$$f_{GPD}(t; \mu, \sigma, k) = \begin{cases} \frac{1}{\sigma}\big(\big[1 + k\big(\frac{t-\mu}{\sigma}\big)\big]^{-\frac{1}{k}-1}\big) & , k \neq 0 \\ \frac{1}{\sigma}exp\big(-\frac{(t-\mu)}{\sigma}\big) & , k = 0 \end{cases} \tag{1.6}$$

The corresponding cumulative distribution function (cdf) is given as follows:

$$F_{GPD}(t;\mu,\sigma,k) = \begin{cases} 1-\big(\big[1+k\big(\frac{t-\mu}{\sigma}\big)\big]^{-\frac{1}{k}}\big) & ,\ k \neq 0 \\ 1-exp\big(-\frac{(t-\mu)}{\sigma}\big) & ,\ k = 0 \end{cases} \tag{1.7}$$

There are several methods for estimating the parameters μ, σ, and k of the GP distribution. Some of these methods include elemental percentile method (EPM) proposed by Castillo and Hadi [23]. Grimshaw [58] proposed an algorithm for computing the maximum likelihood estimation (MLE) of the parameters of the GPD. Hosking & Wallis [66] derived a parameter and quantile estimation mechanism based on Probability-weighted moments (PWM). Zhang [143] proposed an improved maximum likelihood estimation using the empirical Bayesian method to overcome the non-existence problem of the PWM estimator. Castillo and Hadi [24] proposed a more efficient optimization algorithm for estimators of the GPD parameters where the proposed estimators are defined for all possible values of the parameters. The performance of the estimators were found to be better than the method of moments (MOM) and Probability-Weighted Moments (PWM) estimates. Pham, Tsokos, & Choi [105] proposed a GP parameter estimation method for censored data and validated their results using sensitivity and specificity test. Singh & Gao [120] developed a parameter estimation method using the principle of maximum entropy (POME) for 3-P GPD. Since, we have enough data to analyze, we can choose any well known method for our parameter estimation purpose. In the next subsection, we discuss briefly about the parameter estimation procedure of 3-P GPD by pwm method.

1.4.3 Parameter Estimation of 3-P GPD Using the Method of Probability-Weighted Moments (PWM)

The probability-weighted moments (PWM) of a random variable T with cumulative distribution function $F(t) = P(T \leq t)$ is given by,

$$M_{p,r,s} = E[T^p\{F(t)\}^r\{1-F(t)\}^s] \quad , \tag{1.8}$$

where p, r, and s are real numbers. Probability-weighted moments can be expressed as a function of the the inverse distribution function $F^{-1}(t) = t(F)$ in closed form by,

$$M_{p,r,s} = \int_0^1 \{t(F)\}^p F^r \{1-F\}^s] \quad . \tag{1.9}$$

The two special cases of $M_{p,r,s}$ which are commonly used are

$$\begin{aligned} &\alpha_s = M_{1,0,s} = E[T\{1-F(t)\}^s] \quad , (s = 0, 1, 2, ...) \\ &\quad \text{and} \\ &\beta_r = M_{1,r,0} = E[T\{F(t)\}^r] \quad , (r = 0, 1, 2, ...) \quad , \end{aligned} \tag{1.10}$$

where T inside the $E[\cdot]$ is the inverse distribution of T, denoted by $t(F)$. To estimate the parameters of GPD , we use $\alpha_s = M_{1,0,s} = E[T\{1-F(t)\}^s]$ according to the approach used by Singh & Gao [120].

From (1.7) we can solve for T to obtain the inverse cdf, $t(F)$. The inverse distribution function is given by,

$$t(F) = \begin{cases} \mu + \frac{\sigma}{k}\{1-(1-F^k)\} & \text{if } k \neq 0 \\ \mu - \sigma log(1-F) & \text{if } k = 0 \quad . \end{cases} \tag{1.11}$$

The analytical form of α_s for the 3-P GPD is given as follows. Using expressions (1.10) and (1.11). From (1.10), we have

$$\begin{aligned} \alpha_s &= M_{1,0,s} \\ &= \int_0^1 \big[\mu + \frac{\sigma}{k}\{1-(1-F^k)\}\big][1-F^s]dF \\ &= \frac{1}{s+1}\Big(\mu + \frac{\sigma}{k}\Big) - \frac{\sigma}{k}\Big(\frac{1}{k+s+1}\Big) \quad , (s = 0, 1, 2, ...). \end{aligned} \tag{1.12}$$

Thus, for $k \neq 0$, the probability-weighted moments (PWM) of the 3-P GP distribution is given by (12). In equation (1.12) , substituting $s = 0, r = 1$, and $r = 2$ we can obtain explicit expressions of α_0, α_1, and α_2 in terms of μ, σ and k. That is,

$$\alpha_0 = \Big(\mu + \frac{\sigma}{k}\Big) - \frac{\sigma}{k}\Big(\frac{1}{k+1}\Big), \tag{1.13}$$

$$\alpha_1 = \frac{1}{2}\Big(\mu + \frac{\sigma}{k}\Big) - \frac{\sigma}{k}\Big(\frac{1}{k+2}\Big), \tag{1.14}$$

and

$$\alpha_2 = \frac{1}{3}\Big(\mu + \frac{\sigma}{k}\Big) - \frac{\sigma}{k}\Big(\frac{1}{k+3}\Big). \tag{1.15}$$

The **PWM** estimates of the parameters $(\hat{\mu}, \hat{\sigma}, \hat{k})$ can be obtained by solving the equations (1.13), (1.14) and (1.15) for μ, σ, and k. After solving the above three equations, we obtain the explicit expressions of the PWM estimates [120] as follow:

$$\hat{k} = \frac{\alpha_0 - 8\alpha_1 - 9\alpha_2}{-\alpha_0 + 4\alpha_1 - 3\alpha_2} \quad . \tag{1.16}$$

$$\hat{\sigma} = \frac{(\alpha_0 - 2\alpha_1)(\alpha_0 - 3\alpha_2)(-4\alpha_1 + 6\alpha_2)}{(-\alpha_0 + 4\alpha_1 - 3\alpha_2)^2} \quad . \tag{1.17}$$

and

$$\hat{\mu} = \frac{2\alpha_0\alpha_1 - 6\alpha_0\alpha_2 + 6\alpha_1\alpha_2}{-\alpha_0 + 4\alpha_1 - 3\alpha_2} \quad . \tag{1.18}$$

Table 1.8 below shows the approximate parameter estimates of survival times driven by 3-P GP probability distribution.

Table 1.8: Parameter Estimates of 3-P GP Probability Distribution of the Survival Times of Pancreatic Cancer Patients.

Estimates	**Measures**
Location ($\hat{\mu}$)	.65
Scale ($\hat{\sigma}$)	8.9
Shape ($\hat{k}$)	0.22

Now substituting the parameter estimates of μ, σ, k in (1.6) to obtain the analytical form of the probability density function (pdf) of patients' survival times. The analytical form of the GP probability density function (pdf) for combined pancreatic cancer survival time is given by:

$$f_{Combined}(t) = \frac{1}{8.9}\left[1 + .22\left(\frac{t - .22}{8.9}\right)\right]^{-5.54} \quad , \quad t \geq \ .22. \tag{1.19}$$

The above probability density function characterize the probabilistic behavior of the overall survival times of male and female patients with pancreatic cancer.

We now proceed to calculate the expected survival times $E(T)$ of patients driven by GP probability distribution. Using estimates given in Table 1.8, we can find the expectations and median survival times for the patients that follow $GPD(.65, 8.9, 0.22)$ distribution.

The expected value of a random variable T following $GPD(\mu, \sigma, k)$ is given by

$$E(T) = \hat{\mu} + \frac{\hat{\sigma}}{1 - \hat{k}} \quad , \quad \hat{k} < 1. \tag{1.20}$$

Using equation (1.20), the expected survival time for pancreatic cancer patients following $GPD(.65, 8.9, 0.22)$ is

$$E(T) = .65 + \frac{8.9}{1 - .22} = 12.06 \quad \text{months.}$$

The median of the survival time T of $GPD(\mu, \sigma, k)$ is given by,

$$Med_{GPD}(t; \mu, \sigma, k) = \hat{\mu} + \frac{\hat{\sigma}(2^{\hat{k}} - 1)}{\hat{k}} \tag{1.21}$$

From equation (1.21), the overall median survival times of male and female pancreatic patients together is given by,

$$Med(T) = .65 + \frac{8.9(2^{.22} - 1)}{.22} = 7.31 \quad \text{months.}$$

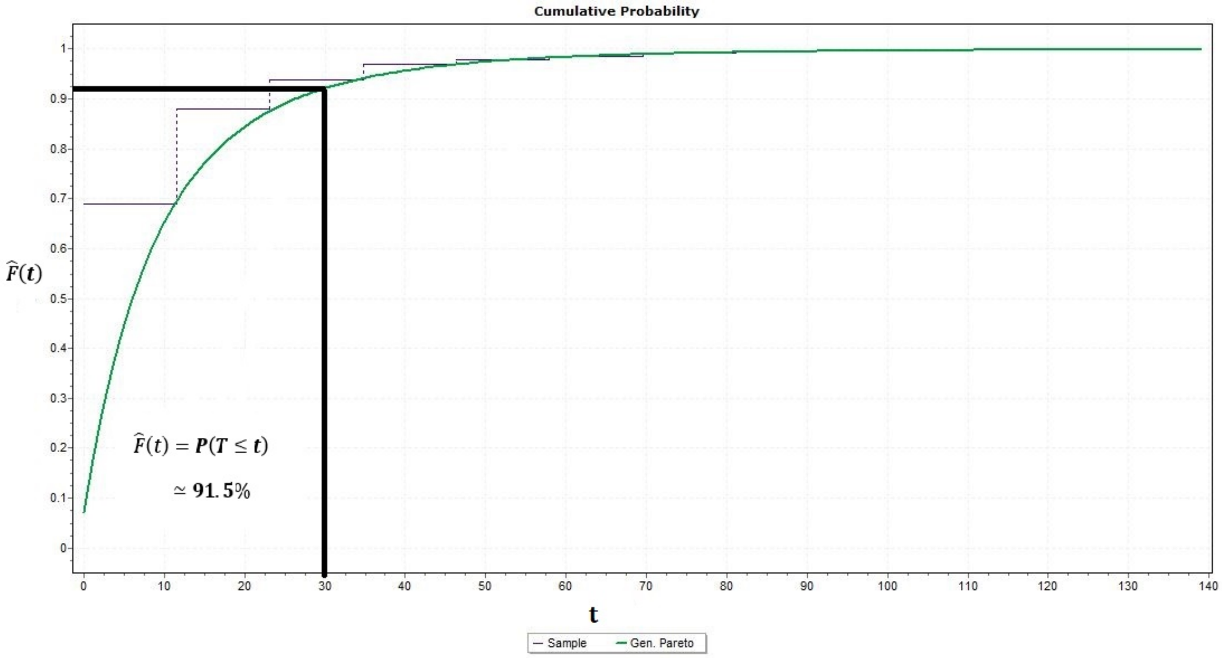


Figure 1.14: cdf Plot for the Survival Times of overall Pancreatic Cancer Patients

Once we have the analytical forms of the pdf , we can obtain the cumulative distribution functions (cdf) of the the random variable T. The analytical form of the GPD cdf is given by:

$$F_{Combined}(t) = 1 - \left[1 + .22\left(\frac{t - .65}{8.9}\right)^{-4.54}\right], \ \ t \geq .65. \tag{1.22}$$

The Figure 1.14 illustrates the cdf plot of the overall survival times.

As the above figure illustrates, the cdf plot is very helpful to estimate the probabilities that a certain male or female patient diagnosed with pancreatic cancer will survive up to a specific point of time. For example, from Figure 1.14 above, the probability that a randomly diagnosed patient will survive up to time $t = 30$ months is approximately 91.5%. In the next section, we will present the parametric survival analysis of the overall survival times of pancreatic cancer patients, which is one of the most important aspects of this study.

1.4.4 Parametric Survival Analysis

Estimation of a parametric survival function is a process to evaluate the survival probabilities of male or female pancreatic cancer patients as a function of the survival time. We have determined the cdf of the survival times for patients diagnosed with pancreatic cancer patients in Equation (1.22). Now, we can proceed to estimate the survival function $S(t)$. Thus, the parametric survival function of patients, irrespective of stages, diagnosed with pancreatic cancer is given by,

$$\begin{aligned}\hat{S}(t; \hat{\mu}, \hat{\sigma}, \hat{k}) &= 1 - \hat{F}(t; \hat{\mu}, \hat{\sigma}, \hat{k}) \\ &= \left[1 + .22\left(\frac{t - .65}{8.9}\right)^{-4.54}\right], \ \ t \geq .65.\end{aligned} \tag{1.23}$$

The survival function $\hat{S}(\cdot;\cdot)$ can be used to estimate the probability that a randomly selected patient diagnosed with pancreatic cancer would survive beyond time t, which is denoted by $P(T \geq t)$. For example, we can compute the probability that a patient diagnosed with pancreatic cancer would survive beyond 30 months. That is, for $t = 30$ in equation (1.23), we estimate the probability as 0.09. Thus, we can infer that a randomly chosen pancreatic cancer patient has a 9% chance of survival beyond 30 months. Figure 1.15 below, describes the parametric survival plot for pancreatic cancer patients, generated using GP probability distribution.

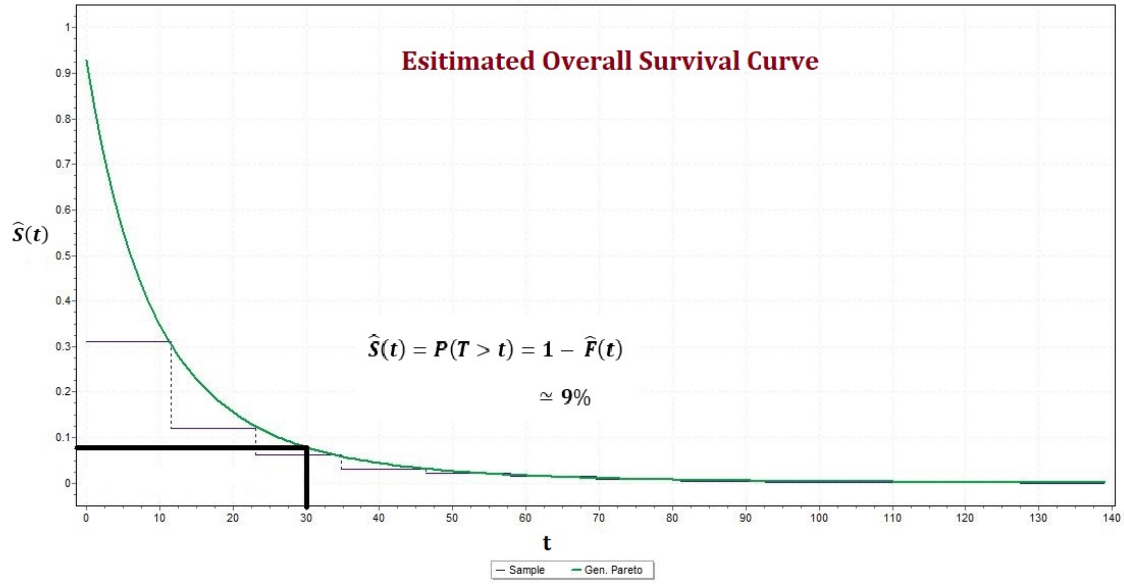


Figure 1.15: Parametric Survival Plot of Overall Pancreatic Cancer Patients

In the next section, we discuss briefly the non-parametric Kaplan-Meier Survival function for pancreatic cancer .

1.5 Kaplan-Meier Estimation of Survival Probability of the Survival Times of Patients with Pancreatic Cancer

The most frequently used parametric estimation methods for distributions of lifetimes are probably the fitting of a normal probability distribution to the observations or their logarithms by calculating the mean and variance and fitting an exponential distribution by estimating the mean alone. Such assumptions about the form of the distribution are naturally advantageous insofar as they are correct; the estimates are simple and relatively efficient, and a complete distribution is obtained even though the observations may be restricted in range. However, non-parametric estimates have the important functions of suggesting or confirming such assumptions and of supplying the estimate itself in case suitable parametric assumptions are not known. The Kaplan–Meier (KM) estimator [15], also known as the product-limit estimator, is a non-parametric statistic used to estimate the survival function from data related to survival time. In health science, it is generally used to measure the fraction of patients living for a certain amount of time after treatment. It was developed by Edward L. Kaplan and Paul Meier (1958). It is defined as the product over the failure times of the conditional probabilities of surviving to the next failure time. Formally, it is given by,

$$\hat{S}(t) = \prod_{t_i \le t} (1 - \hat{q}_i) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \quad , \tag{1.24}$$

where n_i is the number of patients at risk at time t_i, and d_i is the number of individual patients who fail(die) at that time.

The following Figure 1.16, demonstrates the overall non-parametric survival curve for patients diagnosed with pancreatic cancer.

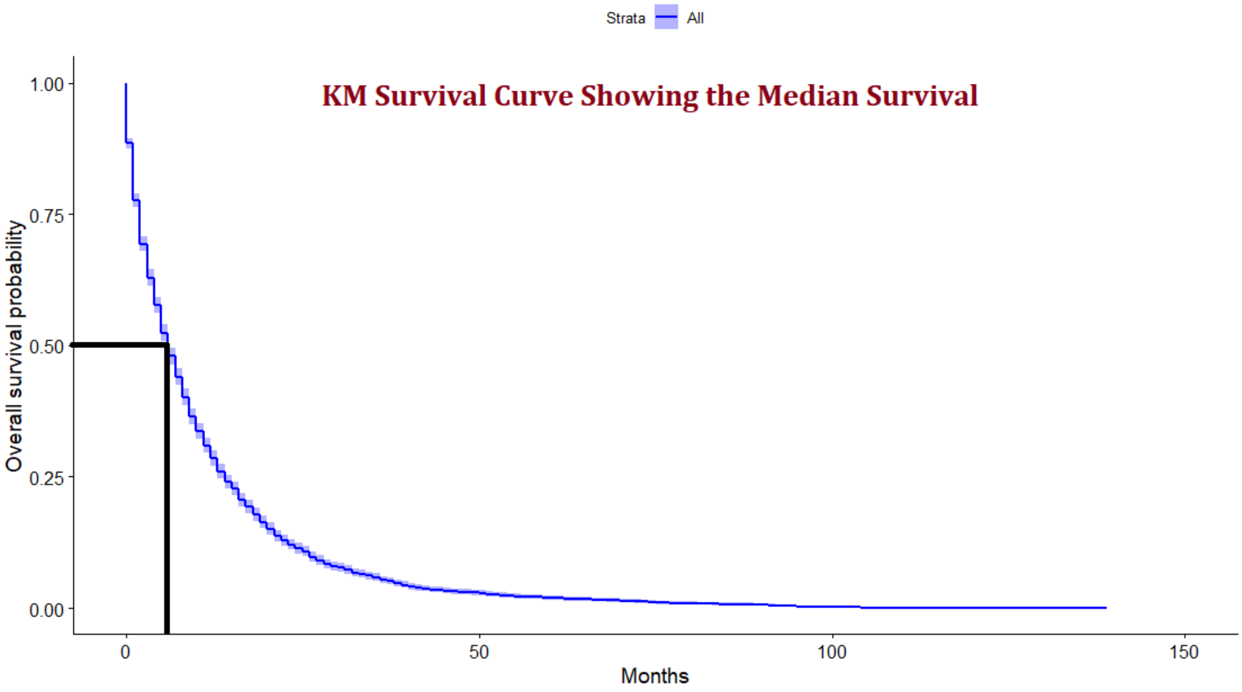


Figure 1.16: Overall KM Survival Plot for Pancreatic Cancer Patients

1.5.1 Median Survival Using KM Estimate

Median survival time is a statistic that indicates how long a group of patients will survive with an illness in general or after a specific treatment has been implemented. It is usually expressed in months or years. Median survival time is when half the patients who are susceptible to a certain disease are anticipated to be alive. It signifies that the probability of surviving beyond that specific time is 50%. It gives an approximate indication of survival and the prognosis of a group of patients with cancer. Median survival is frequently reported in almost every cancer treatment studies. Generally, the median survival time is defined as, $\hat{t}_{med} = inf\{t : \hat{S}(t) \leq 0.5\}$ (see [123] for details). It means that it is the smallest t such that the estimated survival function $\hat{S}(t)$ is less than or equal to 0.5. The median survival times, computed using non-parametric KM estimator, for the pancreatic cancer patients are given as *six* which is evident from above Figure 17. It is very interesting to note that the median survival time we obtained by the descriptive method (Table 5) is exactly same as what we obtained from non-parametric method. However, the median survival times we obtained using the parametric method (implementing the GPD) is significantly *higher* than the descriptive and non-parametric methods. The following Table 1.9 compares the median survival times for all patients diagnosed with pancreatic cancer, computed using the *three* methods.

Table 1.9: Table of Comparison of the Median Survival Times for All Pancreatic Cancer Patients.

Methods	**Median Survival Time**
Descriptive	6
Parametric	7.31
Non-Parametric	6

1.6 Comparison of GP Probability Distribution with the Kaplan-Meier (KM) Estimation of the Survival Function

In the parametric analysis (section 4.2), we found that patients' survival times (both male and female) with pancreatic cancer follows a Generalized Pareto (GP) distribution. In section 5, we performed a non-parametric analysis using the Kaplan-Meier to estimate a randomly selected patients' survival probability.

Table 1.10: Table of Comparison of Estimated Survival Probabilities of Pancreatic Cancer Patients Computed Using Parametric and Non-Parametric Procedures.

t	$\hat{S}_P(t)$	$\hat{S}_{KM}(t)$
0	.96	.88
1	.87	.77
2	.81	.69
3	.77	.62
4	.7	.57
5	.63	.52
6	.57	.47
7	.51	.44
8	.47	.4
9	.43	.36
10	.39	.33

We now compare the survival probability estimates obtained from GP probability distribution with the non-parametric Kaplan-Meier survival estimates of the survival times of the pancreatic cancer patients. The importance of the survival function of the two methods is to estimate the survival probability of a patient diagnosed with pancreatic cancer beyond a given time. The survival probabilities corresponding to a specific time (in months) are shown in Table 1.10 for comparison purposes. We observe that the probability estimates computed by the GP survival function are *significantly higher* than that of Kaplan-Meier probability estimates. Since parametric methods are more powerful, robust, and efficient than non-parametric methods, we must use the parametric estimates of the probabilities as the most accurate estimates.

In the above Table 1.10, $\hat{S}_P(t)$ is the parametric survival probability estimates for pancreatic cancer patients using GP probability distribution. $\hat{S}_{KM}(t)$ is the non-parametric survival probability estimates for pancreatic cancer patients using the non-parametric KM estimate.

1.7 Results and Discussions

Given the risk posed by pancreatic cancer in the past several years, it is imperative to investigate the prognosis and enhance the therapeutic/treatment strategy of pancreatic cancer. The primary treatment for most types of pancreatic cancer is chemotherapy, sometimes, along with a targeted therapy drug. A stem cell transplant might follow this. Surgery and radiation therapy do not fall under crucial treatments for pancreatic cancer, but they might be used in exceptional circumstances. Also, the treatment approach for children with pancreatic cancer can be slightly different from that used for adults. Different research approaches and methodologies have been developed to treat pancreatic cancer patients to boost their survival times. Chakraborty & Tsokos [28] performed a data-driven research on Acute Myeloid Leukemia (AML) by doing some parametric and non-parametric analysis to improve the survival probabilities of patients of different gender groups. In our present study,

- We analyzed a total of 10,000 patient information and have shown that there was *no* significant difference between the overall survival times of male and female pancreatic cancer patients.

- We identified a well-defined probability distribution that characterizes the survival times of a total of 10,000 patient (5,100 male and 4,900 female) diagnosed with pancreatic cancer and used the information to estimate the parametric survival function driven by generalized Pareto (GP) probability distribution.

- We have tested if there is any significant difference between the mean survival times of male and female patients in each of the four stages.

- We have identified the probability distributions of male and female survival times in four different cancer stages, and derived their analytical forms. Also we derived the parametric survival functions in each stages, driven by different parametric probability distributions.

- We compared the median survival times of patients using descriptive, parametric, and non-parametric methods and obtained very consistent results.

- We calculated the overall survival probabilities utilizing the frequently used non-parametric Kaplan-Meier (KM) cancer survivorship analysis method and compared those estimates with the parametric probability estimates obtained from GP probability distribution.

In the first part of our analysis, we tried to investigate if there exits any statistically significant difference between the survival times of male and female pancreatic cancer patients in each stages using Wilcoxon test. We found that there exists significant difference only in stage IV. Then, we proceed to find the most appropriate probability distributions in each stages that best characterize the survival time data and estimated the parameters of the distributions. We then compute the analytical structures of the survival functions in each stages driven by several probability distribution. This is one of the most important aspects of our study, as, the survival function predicts the probability of surviving a randomly selected patient beyond a particular time at a specific stage after diagnosed by pancreatic cancer, which is crucial. We believe that finding the most accurate probability distribution that represents the probabilistic behavior of the survival times for a given cancer patient can lead to estimating the survival probability with much more accuracy and efficiency. After we analyzed the data from individual stages we proceed to analyze the combined survival data irrespective of stages. After we analyzed the survival data for individual stages, we wanted to verify if there is any statistical significant difference between the *overall* male and female survival time. For that purpose we used the the Log-Rank test. We found that there does not exist any statistical significant difference between the survival times of both males and females diagnosed with pancreatic cancer. So, we proceed to perform the analysis of the overall patient survival times, irrespective of cancer stages. We found that a GP probability distribution best characterizes the overall survival time's probabilistic behavior. We found that the GP distribution most often estimates higher survival probabilities compared to the KM survival function, given by Table 1.10. We know that KM estimates are very frequently and commonly used tool to analyze the cancer survivorship data, but they are not the best estimates. Statistically, the parametric techniques are considered to be more robust and efficient than the non-parametric counterpart. Therefore, our finding of the parametric GP probability distribution gives better results in estimating the survival probability of the patients diagnosed with pancreatic cancer than the Kaplan-Meier. By obtaining the best parametric probability distribution that characterizes the survival times, we can find the survival function and estimate the survival rate and compare the results of two or more entities with a high degree of accuracy.

1.8 Conclusion

We have determined the survival probability of patients diagnosed with pancreatic cancer using different statistical methods; the parametric Generalized Pareto (GP) distribution , and the non-parametric Kaplan-Meier (KM) estimation. We found the parametric method to give often higher estimates of the survival probabilities than the non-parametric KM method. The parametric survival analysis's difficulty is the fundamental inherent assumption that the survival times under study follow a specific probability distribution. But if we can overcome such restriction, we can obtain a more robust and efficient result from the parametric analysis, which has greater statistical power. We can also evaluate the hazard function, which determines the rate at which patients die with pancreatic cancer, after finding the most appropriate parametric distribution. Depending on the two different methods utilized for estimating the probability of survival of patients diagnosed with pancreatic cancer, we impart the following important recommendations.

- Given the information regarding male and female cancer patients' survival times, it is customary to investigate first if there exists any statistically significant difference between the male and female patients' true median survival times. If the difference is significant, we must perform a separate analysis for each of the two groups. In the present study, we found that there is **no** significant difference between overall survival times of male and female patients diagnosed with pancreatic cancer.

- After identifying the appropriate probability distributions of male and female cancer patients, if we have further data available regarding the different stages, it is essential to identify the analytical forms of the probability distributions that drive the survival data in each of the four individual stages.

- If we have information available, then the *stage by stage* analysis most appropriate reflects the survival probability of patients in individual stages.

- If the only information provided about the patient is the survival time, then estimating the survival probability using the parametric technique will yield more accurate, robust, and efficient results than the commonly used non-parametric Kaplan-Meier survival estimate.

- However, if there is no unique or well-defined parametric probability distribution are found, we propose using the kernel density estimate or Kaplan-Meier (KM) estimate of the survival probabilities.

Although the use of non-parametric Kaplan-Meier survival analysis may, in certain circumstances, result in a similar or higher probability estimate of the survival rate (if we include the censored observations in our study), the parametric analysis remains more powerful, robust, and efficient when there is no information about the censored individuals. Hence, the parametric analysis must be considered the first stage of data analysis of any given cancer survivorship data. This study provides a more effective and plausible method for estimating the survival probability and analysis of cancer survivorship data to further enhance the therapeutic/treatment process of pancreatic cancer.

Chapter 2: Survival Analysis for Pancreatic Cancer Patients using Cox-Proportional Hazard (CPH) Model

Journal article: "Survival Analysis for Pancreatic Cancer Patients using Cox-Proportional Hazard (CPH) Model," by Chakraborty, A., & Tsokos, C, 2021, Global Journal Of Medical Research. doi:10.34257/GJMRFVOL21IS3PG29 CC-BY-NC 2021 by Copyright Holder. Used with Permission.

2.1 Introduction

The incidence and number of deaths caused by pancreatic tumors have been gradually increasing, even as incidence and mortality of other common cancers have been declining. Despite developments in detection and management of pancreatic cancer, only about 4% of patients will live five years after diagnosis, [134]. The normal pancreas consists of digestive enzyme-secreting acinar cells, bicarbonate-secreting ductal cells, centroacinar cells that are the geographical transition between acinar and ductal cells, hormone-secreting endocrine islets and relatively inactive stellate cells. The majority of malignant neoplasms of the pancreas are adenocarcinomas. Rare pancreatic neoplasms include neuroendocrine tumors (which can secrete hormones such as insulin or glucagon) and acinar carcinomas (which can release digestive enzymes into the circulation). Particularly, ductal adenocarcinoma is the most frequent kind of malignancy of the pancreas; this tumor (commonly referred to as pancreatic cancer) presents a substantial health problem, with an estimated 367,000 new cases diagnosed worldwide in 2015 and an associated 359,000 deaths in the same year[80]. After the detection of pancreatic cancer, doctors usually perform some additional tests to understand better if cancer has been spread or the spreading area of cancer. Different imaging tests, such as a PET scan, can help doctors identify the presence of cancerous growths. With these tests, doctors try to establish cancer's stage. Staging helps explicate how advanced the cancer is. It also assists doctors in deciding the treatment options. The following are the description of the stages used in our dataset according to the definition of the Surveillance, Epidemiology, and End Results (SEER) database.

1. **Localized**: There is no sign that the cancer has spread outside of the pancreas.
2. **Regional**: The cancer has spread from the pancreas to nearby structures or lymph nodes.
3. **Distant**: The cancer has spread to distant parts of the body such as the lungs, liver or bones.

Although, in most cases, pancreatic cancer remains incurable, researchers have focused on how to improve the survival times of patients diagnosed with pancreatic cancer. Cox proportional hazard model/ Cox model [37] has been used extensively in the literature of cancer research to address the hazard of an individual patient with respect to specific risk factors. It is also useful to assess the association between different treatments and the survival time of patients. Perera and Tsokos [102] developed a statistical model with Non-Linear Effects and Non-Proportional Hazards for Breast Cancer Survival Analysis. In their study, the authors have identified the effects of age and breast cancer tumor size at diagnosis on the hazard function, which have a non-linear effect. Also, they have addressed the different assumptions of the proportional hazard model. Asano, Hirakawa, and Hamada [5] used an imputation-based receiver operating characteristic curve (AUC) to evaluate the predictive accuracy of the cure rate from the PH cure model. They also illustrated the estimation of the imputation-based AUCs using breast cancer data. Yong & Tsokos [140] have evaluated the effectiveness of widely used Kaplan-Meier (KM) model, non-parametric Kernel density (KD) models with the Cox PH model, using both Monte Carlo simulations on the breast cancer data. Du, Li et al. (2018) [45] compared a flexible parametric survival model (FPSM) and Cox model using Markov transition probabilities from a cohort study data investigating ischemic stroke outcomes in Western China. The FPSM produced hazard ratio and baseline cumulative hazard estimates similar to those obtained using the Cox proportional hazards model. Mamudu & Tsokos developed a semi-parametric Cox model for Multiple Myeloma Cancer (MMC) patients and addressed the validity of the assumptions of the model.

In our study, we used the semi-parametric Cox-PH survival analysis of the survival times to estimate the survival rate of patients diagnosed with pancreatic cancer. We utilized the Cox-PH model to analyze the proportion of survival time, taking into account the fifteen risk factors that are identified in section 2.1. We assessed the relationship between the proportion of survival time as a function of the attributable risk factors and two-way interactions based on the Cox proportional hazard (PH) model. The significant attributable risk factors identified were meticulously investigated and selected based on the step-wise model selection method, with the final model representing the model with the least AIC. The final Cox-PH model was validated to satisfy all the main assumptions of the Cox-PH model.

2.2 Methodology

2.2.1 Data Description

The data for our study has been obtained from The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial system of the National Cancer Institute (NIH) database. The data contains information on patients diagnosed with pancreatic adenocarcinoma. We are concerned with the survival time (in days) and cause-specific death (deaths due to pancreatic cancer) for each patient. The survival time of patients is one of the most important factors used in all cancer research. It is important to evaluate the severity of cancer, which helps to decide the prognosis and help identify the correct treatment methods. There were a total of 677 patient information in our study after eliminating the missing observations for which several risk factors were missing. In our study, the response variable is the survival time of patients (in days). There are a total of *fifteen* risk factors used in our survival model.

Twelve of them are categorical, and three of them are numeric variables. The description of the risk factors is as follows.

1. Age (Numeric) (X_1): Age of diagnosis of the patient.

2. Stage (Categorical) (X_2): Pancreatic Cancer Stages, categorized as a) localized, b) regional, and c) distant

3. Aspirin (Categorical) (X_3): Does the person use Aspirin Regularly?

4. Ibuprofen (Categorical) (X_4): Does the person use Ibuprofen Regularly?

5. Relatives (Categorical) (X_5): The number of first-degree relatives with pancreatic cancer.

6. Diabetes (Categorical) (X_6): Did the patient ever have diabetes?

7. Heart attack (Categorical) (X_7): Did the participant ever have coronary heart disease or a heart attack?

8. Emphysema (Categorical) (X_8): Did the patient ever have emphysema?

9. Sex (Categorical) (X_9): Sex of the individual.

10. BMI (numeric) (X_{10}): Current Body Mass Index (BMI) at Baseline (In lb/in2)

11. Cigarette Years (numeric) (X_{11}) : The total number of years the patient smoked.

12. Diverticulosis (Categorical) (X_{12}): Did the participant ever have diverticulitis or diverticulosis?

13. Smoke (Categorical) (X_{13}): Has the patient ever smoked cigarettes regularly for six months or longer?

14. Gallbladder (Categorical) (X_{14}): Did the individual ever have gall bladder stones or inflammation?

15. Hypertension (Categorical) (X_{15}): Did the individual ever have high blood pressure?

A schematic diagram of the data used in our study with the description of risk factors is shown in Figure 2.1, below.

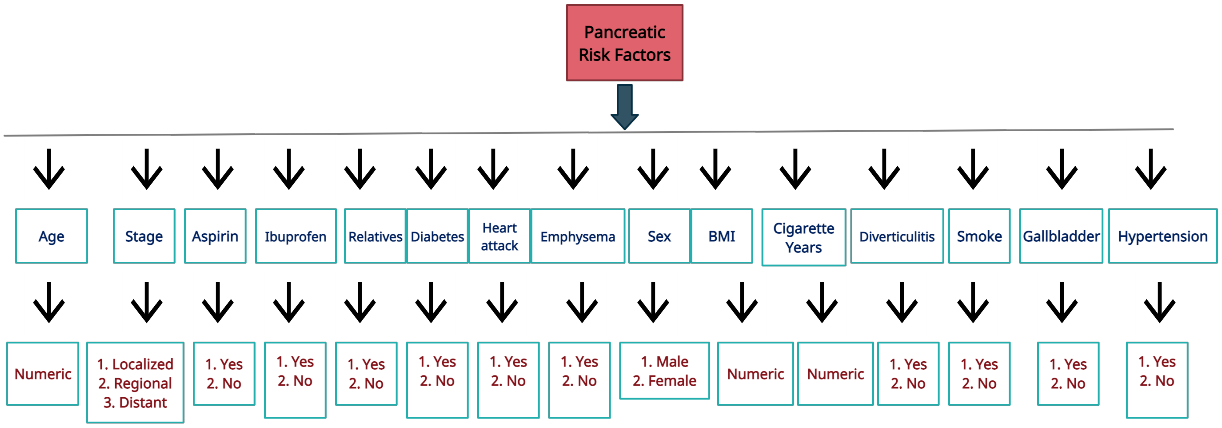


Figure 2.1: Pancreatic Cancer Data with Relevant Risk Factors

As the above Figure illustrates, we see that twelve out of fifteen risk factors are categorical, having two or more categories. Before we proceed with our main analysis, it is very important to investigate if there is any statistically significant difference between the survival times of male and female patients diagnosed with pancreatic cancer. If any significant differences are found, separate analyses for each gender should be performed. To answer this question, we used the non-parametric Wilcoxon rank-sum test with continuity correction and obtained a p-value of .47, indicating that there is not enough sample evidence to reject the following null hypothesis (H_0) at a 5% level of significance.

H_0: There is no statistically significant difference between the survival times of male and female patients.

Thus we proceeded with our analysis and modeling by combining the male and female data together to constitute our sample size.

2.3 Brief Description of Cox Proportional Hazard (CPH) Model

The CPH model, proposed by Sir David Cox, is a statistical method that can be used for survival-time (time-to-event) outcomes on one or more risk factors and their interactions. In survival analysis, the Cox model has been widely recommended for semi-parametric modeling of the survival time relationship as a function of the risk factors. Kleinbaum & Klein [81] gives a good introductory review of the background and methodology, and more detailed descriptions have been provided by Kalbeisch , and Prentice [76]. In this section, we give a brief review of the Cox proportional hazards model. An important aspect of the Cox PH model is the hazard function $h(t)$. It measures the rate of the event of occurrence (death) as a function of time t. We define the hazard function as follows; Let random variable T denotes the survival time with cumulative density function $F_T(t)$, given by

$$F_T(t) = P(T \leq t) = \int_0^t f(t)dt \ ,$$

where $f(t) = \frac{dF_T(t)}{dt}$ is the probability density function (pdf) of the random variable T. The survival function at time t is defined as:

$$S(t) = P(T \geq t) = 1 - F_T(t) = \int_t^\infty f(t)dt \ . \tag{2.1}$$

$S(t)$ gives the probability that a specific individual would survive beyond time t. Since $S(t)$ is a probability, $0 \leq S(t) \leq 1$ and $S(0) = 1$, for $T \geq 0$ from (1) we have,

$$f(t) = \frac{dF_T(t)}{dt} = -\frac{dS(t)}{dt} \ . \tag{2.2}$$

For continuous survival data, the hazard function plays a very important role. It aims to quantify the *instantaneous risks* that an event will occur at time t. It is defined as the follows:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P\{t \le T < t + \Delta t \mid T \ge t\}}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P\{t \le T < t + \Delta t\}}{\Delta t} \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)} \ . \end{aligned} \tag{2.3}$$

Combining (2) and (3), we obtain,

$$h(t) = -\frac{d}{dt} log\{S(t)\} \ . \tag{2.4}$$

Integrating both sides of equation (4) gives an expression for the survival function $S(t)$ in terms of the hazard function $h(t)$. That is,

$$S(t) = exp\Big[-\int_0^t h(u)du\Big] \ . \tag{2.5}$$

Now, from (3) and (5) we can express the pdf $f(t)$ as a function of $S(t)$ and $h(t)$ given by,

$$f(t) = h(t)exp\Big[-\int_0^t h(u)du\Big] \ . \tag{2.6}$$

From (3) the cumulative hazard function $H(t)$ can be expressed as:

$$H(t) = \int_0^t h(u)du = -lnS(t) \ . \tag{2.7}$$

Now, suppose $X_i = (X_{i1}, X_{i1}, \ldots, X_{ip})$ are the realized values of the risk factor for the i^{th} subject. Then, the Cox PH model (not including time-dependent risk factors or non-proportional hazards) can be expressed in term of the hazard as:

$$h_i(t) = \lambda_0(t)exp\Big[\sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j \neq k} \eta_{jk} X_{ij} X_{ik}\Big] \ , \ j,k = 1,2,\ldots,p. \tag{2.8}$$

In the above expression, λ_0 is called the *baseline hazard* which can be thought of as the hazard function for an individual for which all value of the risk factors are 0. β_j measures the impact of X_{ij} on $h_i(t)$. η_{jk} is the interaction coefficient between j^{th} and k^{th} risk factor of the i^{th} individual and measures the impact of $X_{ij}X_{ik}$ on $h_i(t)$. From (8), it is clear that the individual hazard is a function of the risk factors and their interactions and is connected through baseline hazard. From (8), we can write,

$$ln\Big\{\frac{h_i(t)}{h_k(t)}\Big\} = \Big[\sum_{j=1}^{p}\beta_j X_{ij} + \sum_{j\neq k}\eta_{jk}X_{ij}X_{ik}\Big]\ j \neq k \tag{2.9}$$

From the above expression we see that the ratio of log hazard of the i^{th} and k^{th} individual is constant over time. Thus, the name *proportional* in the Cox PH model. We interpret the hazard ratio (HR) in the following ways:

1. HR $= 1$; implies that there is no hazard effect. Thus, the risk factors have no relationship with the event probability, thus, no influence on the length of survival.

2. HR > 1 (i.e. equivalently $\beta_i > 0$), implies an increase in hazard. That is, the risk factors have a positive association with the event probability, thus, a negative association with the length of survival (bad prognostic factor).

3. HR < 1(i.e. equivalently $\beta_i < 0$), implies a decrease in hazard. That is, the risk factors are negatively associated with the probability of the event, thus, positively associated with the length of survival (good prognostic factor).

A detailed description of the hazard ratio have been provided in [112] [22].

2.4 Statistical Data Analysis and Survival Modeling

We now proceed to develop our most parsimonious statistical model using Cox PH. We initially started by fitting the Cox-PH model to the survival times t as a function of all fifteen risk factors given in Figure 2.2 together with their two-way interactions. So, there were fifteen risk factors and $\binom{15}{2} = 105$ two-way interaction terms. We used a stepwise model selection procedure to select the best model with the minimum Akaike information criterion ($AIC = 2ln(L) + 2k$, where L is the value of the maximum likelihood function of the model and k represents the number of estimated model parameters)[2]. AIC gives an estimation of the relative amount of information missing in the model; hence, the smaller the AIC value, the better the quality of the model. It also deals with the risk associated with overfitting or under-fitting the model. One of the most important assumptions of the Cox PH is proportionality. Initially, all of the risk factors and two-way interactions except *age* satisfied the assumption.

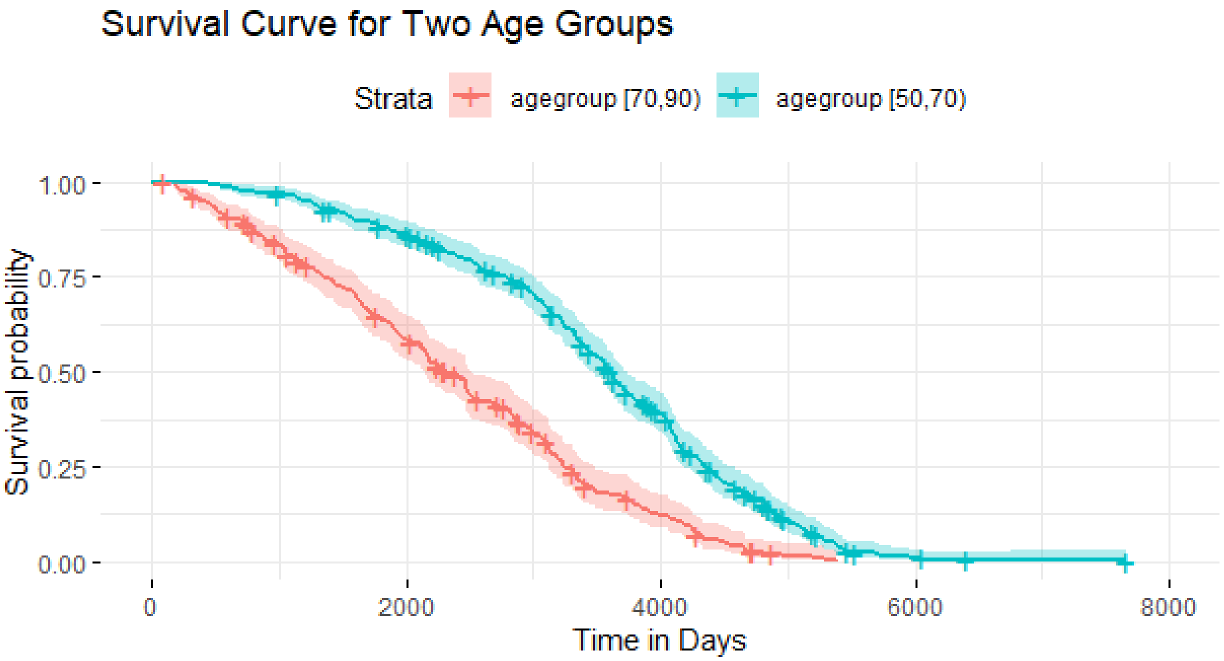


Figure 2.2: The Estimated Survival Curve for the Two different Age Groups

The range of the variable age was [50-90). So, we divided the range into two categories, say [50,70), and [70,90). Now, we use *stratification* on the variable age. Stratification is one of the tools used by researchers when one of the risk factors does not satisfy the proportionality assumption. The stratification will produce hazard ratios for all other risk factors in the presence of two hazards intrinsic to the level of age. Since age violated the proportional hazards assumption, stratifying it will help meet the PH assumption and provide more valid estimates for all other risk factors. The stratified model allows the baseline hazard $\lambda_0(t)$ to vary between strata but controls the effect of the risk factors to be the same for each stratum. For each subject in strata $s, s = 1, 2$, we have from (8),

$$h_i(t) = \lambda_{0s}(t) exp\Big[\sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j \neq k} \eta_{jk} X_{ij} X_{ik}\Big] \ , \ \ j, k = 1, 2, \ldots, p. \ (s = 1, 2) \qquad (2.10)$$

However, it is not possible to get an estimate of the risk factor (age) separately after stratification. The following Figure 2.2 illustrates the survival curve for the two age groups.

We observe from Figure 2.2 that the age group [70,90) (highlighted in pink) is much more vulnerable than the age group [50,70) (highlighted in blue) in terms of survival probabilities. That is, a randomly selected patient in the age group [50,70) has a higher survival probability than a patient in the group [70,90), which is quite plausible.

The cumulative hazard function, $H(t)$, of the two age groups is given below by Figure 2.4.

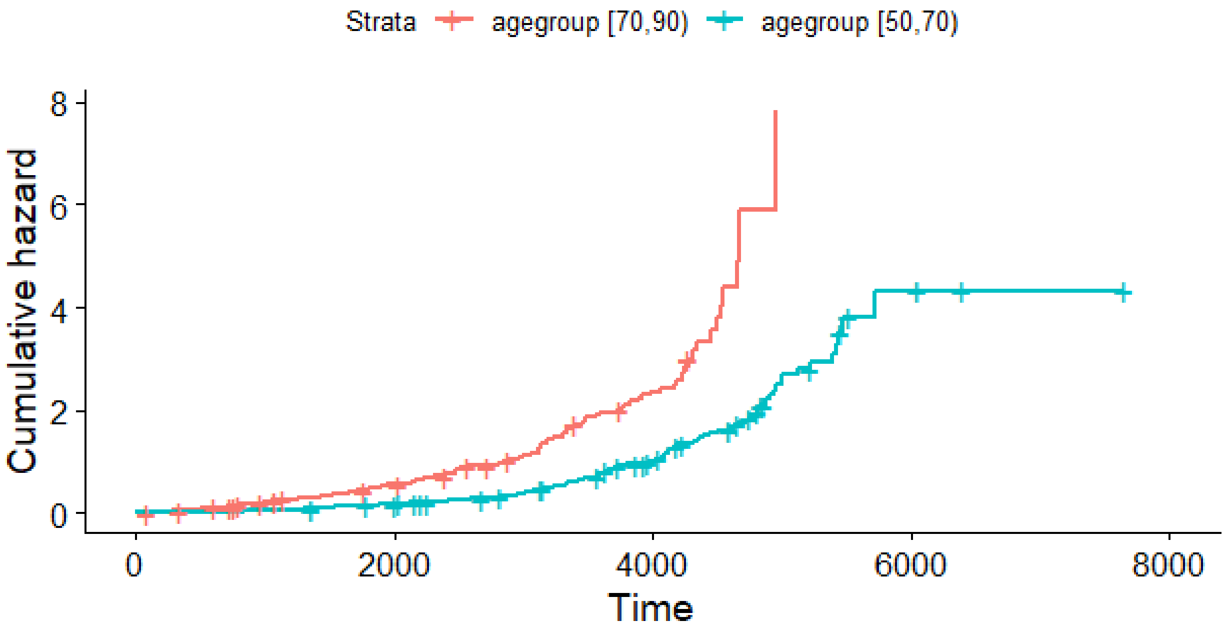


Figure 2.3: Cumulative Hazard Functions of the Two Age Groups

Table 2.1: Table Showing the Count of Different Categories of Risk Factors

Risk Factors		**Count**
Stage	Localized	135
	Regional	178
	Distant	364
Aspirin	Yes	333
	No	344
Ibuprofen	Yes	168
	No	509
Relatives	Yes	650
	No	27
Diabetes	Yes	83
	No	594
Heart attack	Yes	84
	No	593
Emphysema	Yes	19
	No	658
Sex	Male	388
	Female	289
BMI		677
Cigarette Years		677
Diverticulosis	Yes	41
	No	636
Smoke	Yes	404
	No	273
Gallbladder	Yes	98
	No	579
Hypertension	Yes	256
	No	421

Figure 2.3 suggests that the cumulative hazard for patients in the age group [70,90) is more than patients belonging to [50,70). We see that the cumulative hazard is the same for two age groups, almost up to $t = 1000$ days. After that, the cumulative hazard is exponentially increasing for the age group [70,90). However, for the age group [50,70), the cumulative hazard has an increasing pattern up to $t = 5800$ days approximately. After that, the graph has a steady pattern. The step-wise procedure produced *seven* out of fourteen significant risk factors and *ten* two-way interaction terms. There were some risk factors that did not contribute to the hazard individually, but, interacting with other risk factors, their effect was significant. Thus, we added those risk factors in our proposed model. That is why there are thirteen individual risk factors and ten interactions in the model (11). In the following model (11), we denote **"Y"** to indicate yes of a specific answer of a risk factor. That is, the specific category possesses the characteristic. For example, to answer the question "does the patient ever have diabetes?" the individual answers "yes." To describe any particular category of the risk factor *stage*, we use **L**, **R**, and **D** which are the first letters of Localized, Regional, and Distant. To describe male and female category of the variable *Sex*, we use the letters **M** and **F**, respectively. The most parsimonious model that we found after removing the insignificant ($p - value \geq 0.05$) term from the model is given as follows:

$$ln\widehat{\left[\frac{h_i(t)}{\lambda_0(t)}\right]} = \begin{cases} 0.3X_{2R} + .5X_{2D} - .53X_{3Y} \\ +.61X_{4Y} - .37X_{15Y} + .87X_{6Y} \\ -.6X_{5Y} - .7X_{8Y} \\ -.35X_{9F} + .0037X_{11} - .51X_{12Y} + .15X_{13Y} \\ +.28X_{14Y} - .56X_{4Y}X_{13Y} + .41X_{3Y}X_{9F} \\ +.6X_{3Y}X_{15Y} + .01X_{2R}X_{11} + .68X_{12Y}X_{9F} \\ +.32X_{15Y}X_{9F} - .47X_{15Y}X_{14Y} \\ -.52X_{2R}X_{4Y} + 2.18X_{2R}X_{8Y} + .8X_{15Y}X_{12Y} \end{cases} \tag{2.11}$$

Thus, the proposed statistical model consists of thirteen individual risk factors and ten interactions that contributes to the hazard.

2.4.1 Estimating The Survival Function

The above equation (10) can be written as:

$$h_i(t; X_{ij}, X_{ij}X_{ik}) = h_{0s}(t)exp\Big[\sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j\neq k} \hat{\eta}_{jk} X_{ij} X_{ik}\Big], \;\; j \neq k \tag{2.12}$$

We can express the Cox-PH model (11) in the form of the survival function, $S(t)$, by employing equation (5) from Section 3. Thus, the survival function of the Cox-PH model can be expressed as;

$$\begin{aligned}
\hat{S}_i(t; X_{ij}, X_{ij}X_{ik}) &= exp\Big[-\int_0^t h_i(t; X_{ij}, X_{ij}X_{ik})dt\Big] \\
&= exp\Big[-\int_0^t h_{0s}(t)exp\Big[\sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j\neq k} \hat{\eta_{jk}} X_{ij}X_{ik}\Big]dt\Big] \\
&= exp\Big[exp\Big[\sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j\neq k} \hat{\eta_{jk}} X_{ij}X_{ik}\Big]\Big(-\int_0^t h_{0s}(t)dt\Big)\Big] \qquad (2.13) \\
&= exp\Big(-\int_0^t h_{0s}(t)dt\Big)^{\Big[\sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j\neq k} \hat{\eta_{jk}} X_{ij}X_{ik}\Big]} \\
&= \big[S_{0s}(t)\big]^{\Big[\sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j\neq k} \hat{\eta_{jk}} X_{ij}X_{ik}\Big]}
\end{aligned}$$

where $\hat{S_{is}}(t; X_{ij}, X_{ij}X_{ik})$ is the survival function at time t for i^{th} individual and s^{th}, $(s = 1, 2)$ stratum. $S_{0s}(t)$ is the baseline survivor function for each stratum $s = 1, 2$. After the estimation of $\hat{\beta}$ and $\hat{\eta}_{jk}$ by partial likelihood [56], $S_{0s}(t)$ can be estimated by a non-parametric maximum likelihood method [48]. The co-efficient estimates of parameters $\hat{\beta}$ and $\hat{\eta}_{jk}$ are given in the third column of Table 2.2.

Table 2 below displays the estimates of the model coefficients/parameters, their hazard ratios (HR) $(exp(\hat{\beta}))$, standard error of coefficients, statistical significance, and 95% confidence interval. We proceed to rank the significant contributing risk factors and their significant interactions based on the prognostic effect on the survival times of patients diagnosed with pancreatic cancer using the hazard ratio (HR). Thus, we rank from the most contributing risk factor to the least contributing risk factor to pancreatic cancer patient's death or survival times.

Table 2.2: Ranking of the Significant Contributing Risk Factors and Interactions Based on Prognostic Effect to the Survival Time Using the Hazard Ratios

Rank	**Risk Factors**	**coeff($\hat{\beta}$)**	**HR** [$exp(\hat{\beta})$]	[$S.E(\hat{\beta})$]	**Lower 95%**	**Upper 95%**
1	$X_{2R}X_{8Y}$	2.18	8.84	.96	1.32	59.1
2	X_{6Y}	.87	2.39	.33	1.2	4.6
3	$X_{15Y}X_{12Y}$	.8	2.28	.38	1.07	4.87
4	$X_{12Y}X_{9F}$	.68	1.98	.39	.92	4.25
5	X_{4Y}	.61	1.834	.25	1.27	2.62
6	$X_{3Y}X_{15Y}$	.6	1.831	.18	1.11	3.02
7	X_{2D}	.5	1.63	.17	1.16	2.3
8	$X_{3Y}X_{9F}$	.41	1.5	.18	1.06	2.13
9	$X_{15Y}X_{9F}$	.32	1.37	.18	.96	1.96
10	$X_{2R}X_{11}$	0.01	1.01	.007	.99	1.05
11	X_{9F}	-.35	.7	.13	.54	.91
12	X_{15Y}	-.37	.69	.16	.5	.95
13	$X_{15Y}X_{14Y}$	-.47	.63	.26	.42	.94
14	$X_{13Y}X_{4Y}$	-.46	.63	.2	.42	.94
15	X_{3Y}	-.53	.6	.13	.45	.77
16	$X_{2R}X_{4Y}$	-.52	.59	.3	.33	1.05
17	X_{5Y}	-.6	.55	.2	.35	.84

The above Table describes different information, including the hazard ratio of all *seven* significant risk factors and all *ten* significant interactions used in the model. A positive estimated coefficient/weight ($\hat{\beta} > 0$) implies higher hazard rate, and thus a bad prognostic factor. on the contrary , a negative estimated coefficient/weight ($\hat{\beta} < 0$) implies a lower hazard rate, and thus a good prognostic factor. For example, $\hat{\beta}_{9F} = -0.35$ from Table 2, implies females are good prognostic of the survival time of pancreatic cancer; thus, females have a lower risk of death (higher survival rates) of cancer than males. The $exp(\hat{\beta})$ is the hazard ratio (HR). Thus, $exp(-0.35) = .7 < 1$ for gender female means being a female has a reduced risk of dying with pancreatic cancer than being a male. The ranking of the significant risk factors from Table 2, based on the HR, shows that the interaction between **cancer stage (Regional)** and **patient having Emphysema** ($X_{2R}X_{8Y}$) is the highest prognostic factor to the survival of pancreatic cancer, followed by patients having diabetes (X_{6Y}), and Relatives who have pancreatic cancer (X_{5Y}) is the least prognostic factor. We also provide the 95% confidence interval of the hazard ratios (HR) corresponding to the risk factors; that is,

$$P[UCL \leq HR \leq LCL] \geq 95\%$$

where UCL and LCL are the upper and lower confidence limits and we are at least 95% confident that the hazard ratios will fall into the limits. The following Table provides the three popular global tests of significance which our model is based on. As, the following table shows, our proposed model (2.11) is *highly significant* based on all the three statistical tests.

Table 2.3: Global Statistical Significance of the Model

Test	**Test Statistics Value**	**df**	**p-value**
Likelihood Ratio Test	96.6	34	$7 * 10^{-8}$
Wald Test	100.8	34	$2 * 10^{-8}$
Score (log-rank) Test	109.9	34	$6 * 10^{-10}$

2.5 Assumptions of Cox PH Model and Validation of the Proposed Model

In order to apply the CPH model, we must verify that the following three key assumptions are satisfied, prior to its implementation. Failure to satisfy these assumptions will bring about inaccurate decisions about the subject matter.

1. **Proportional hazard (PH) assumption**: The *proportional hazard* assumption of the Cox model can be validated depending on formal statistical tests. A non-statistical

significance of all risk factors along with the interactions in the model with the global test is an evidence that the PH assumption is well-grounded. Another way to verify the PH assumption is by investigating the plot of scaled Schoenfeld residuals [138] against the transformed time. The Schoenfeld residuals are independent of time; a non-random pattern against time is evidence of a violation of the PH assumption. We calculate the Schoenfeld residuals for each of the risk factors and all interactions.

The data consists of times $T_1, T_2, \ldots, T_n$ which are either observed survival times or censored times with censoring indicators $\delta_1, \delta_2, \ldots, \delta_n$. $\delta_i = 1$ implies T_i is observed, and $\delta_i = 0$ implies T_i is censored. Suppose there are p fixed covariates/risk factors $Z_1, Z_2, \ldots, Z_n$ and $\mathscr{R}_i$ be the risk set at time T_i denoted as $\mathscr{R}_i = \{j : T_j \geq T_i\}$. Given the setup, the *partial likelihood*, proposed by Cox (1975) is defined by:

$$L(\beta) = \sum_{i=1}^{n} \delta_i \Big[\beta^T Z_i - log \Big[\sum_{j \in \mathscr{R}_i} exp(\beta^T Z_j) \Big] \Big] \quad . \tag{2.14}$$

Let $\hat{\beta}$ be the usual estimator of β that minimizes $L(\beta)$ in (13). Also, let $t_{(i)}$ be the i^{th} ordered observed survival time and $Z_{(i)}$ and $\mathscr{R}_i$ the corresponding covariate vector and risk set. Then SCHOENFELD'S RESIDUALS are defined as follows:

$$\hat{r}_i = Z_{(i)} - \frac{\sum_{j \in \mathscr{R}_i} Z_j exp(\hat{\beta}^T Z_j)}{\sum_{j \in \mathscr{R}_i} exp(\hat{\beta}^T Z_j)} \quad . \tag{2.15}$$

The following Figures 2.4 and 2.5 illustrate the plot of the scaled Shoenfeld residual against time for all risk factors and interaction terms used in the model (11), respectively. It shows that there is no pattern as a function of time. Thus, the residuals are randomly scattered with no systematic departures from the horizontal fitted smoothing spline deep line (that is, the residuals are independent of times).

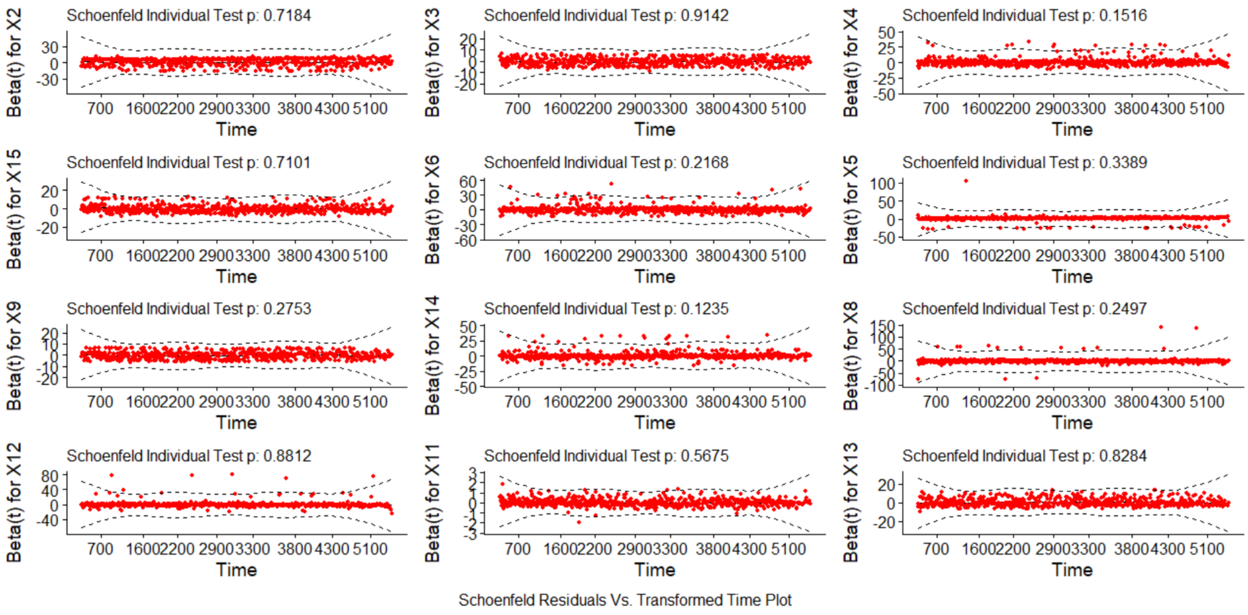


Figure 2.4: Testing Proportional Hazard Assumption For Individual Risk Factors

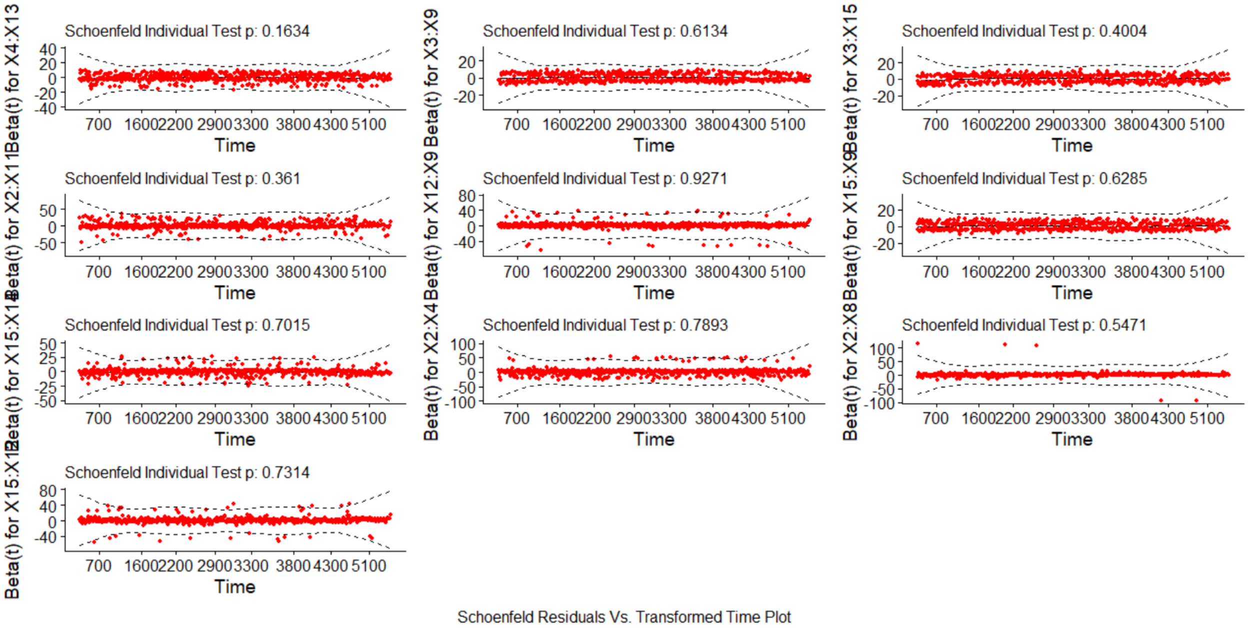


Figure 2.5: Testing Proportional Hazard Assumption For All Interactions

A formal test for the PH assumption is given in the following Table.The covariates and the global test are non-statistically significant given by the large p-values. This is a further justification of the validity of the PH assumption for our proposed model.

Table 2.4: Testing Proportional Hazard Assumption

Risk Factors	χ^2	**p-value**
X_2	.66	.72
X_3	.01	.91
X_4	2.05	.15
X_{15}	.14	.71
X_7	3.39	.1
X_6	1.5	.21
X_8	1.3	.25
X_{12}	.02	.88
X_{14}	2.37	.12
X_{13}	.05	.82
X_{11}	.32	.56
X_{10}	2.56	.11
X_5	2.16	.34
X_9	1.19	.27
$X_4 \bigcap X_{13}$	1.94	.16
$X_3 \bigcap X_9$	.25	.61
$X_3 \bigcap X_{15}$	.71	.4
$X_2 \bigcap X_{11}$	.04	.36
$X_{12} \bigcap X_9$	.008	.93
$X_{15} \bigcap X_9$	.23	.63
$X_{15} \bigcap X_{14}$	.14	.7
$X_2 \bigcap X_4$	.47	.79
$X_2 \bigcap X_8$	1.2	.55
$X_{15} \bigcap X_{12}$	.12	.73
GLOBAL	44.17	.1

We have included all fourteen risk factors and ten interaction terms in the table. The number of terms in Table 2.4 is greater than Table 2.2 since we have included all of the fourteen individual risk factors used in our analysis in Table 2.5.1.

2. **Linear Functional Form of continuous Risk Factors**: Often, many researchers assume that the continuous risk factors in the Cox PH model have a linear form. However, one should verify this assumption before implementation of the model. Representing the Martingale residuals against continuous covariates is a graphical form, is a common approach to identify the nonlinearity or, in other words, to assess the functional form of a covariate. For a given continuous covariate, the plot patterns may suggest that the variable is not properly fit. Nonlinearity is not a problem for categorical risk factors. So we only investigate plots of martingale residuals against the only continuous covariate X_{11}. Sometimes, these plots can help select the appropriate functional forms of the risk factors in the Cox model. The *martingle residual*, proposed by Therneau and Grambsch [129] is given by,

$$\hat{M}_i = \delta_i - \hat{\Gamma}_0(t_i) exp\Big[\sum_{j=1}^{p} \hat{\beta}_j X_{ij} + \sum_{j \neq k} \hat{\eta_{jk}} X_{ij} X_{ik}\Big], \ \ j \neq k. \ ,$$

where δ_i denotes the event indicator for i^{th} observation, $\hat{\Gamma}_0(t_i)$ is the estimated cumulative hazard at the final follow-up time for the i^{th} observation. Martingale residuals, M_i, have a skewed distribution.

We have, $\hat{M}_i = 1$ for for maximum possible values and $\hat{M}_i = -\infty$ for minimum possible values. Positive values of $\hat{M}_i$ indicate those patients expired too early compared to expected survival times.

On the contrary, negative values of $\hat{M}_i$ correspond to patients who were alive for a long time.

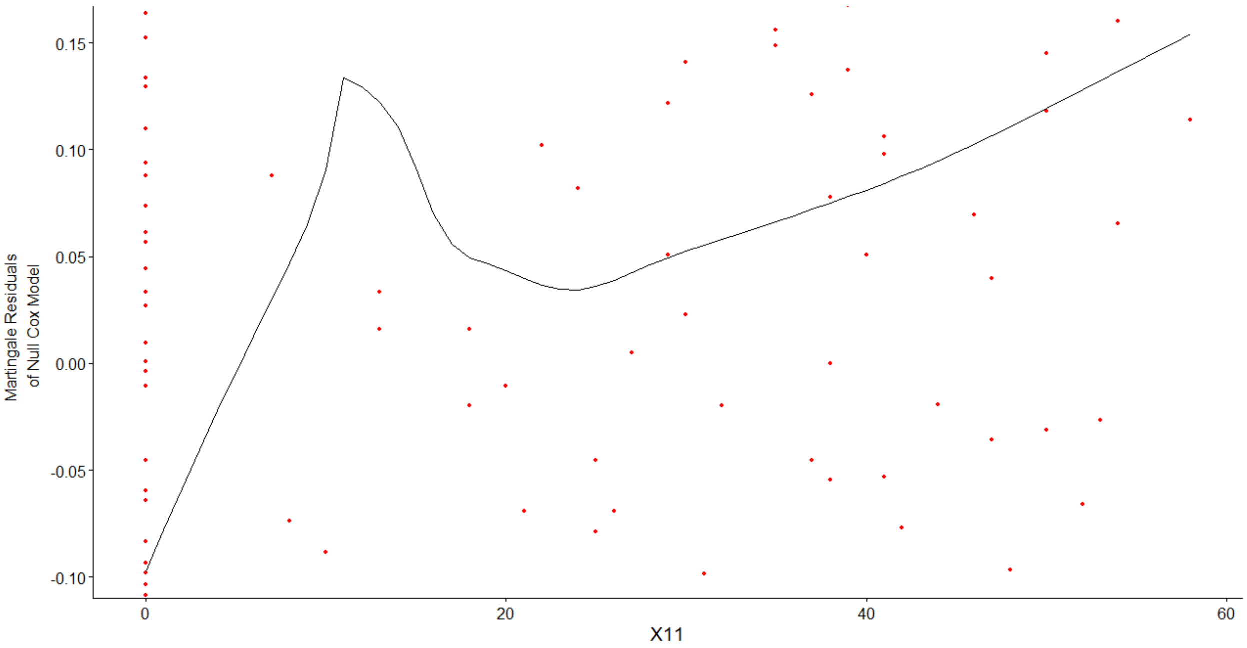


Figure 2.6: Validating the Linearity Assumption of the Continuous Covariate

As shown in the Figure 2.6, the data points are fairly linear for almost all points except around $X_{11} = 10$. The continuous covariate X_{11} is the *number of cigarette smoking years* of an individual patient. There are several patients who did not smoke at all (indicated by the points around zero). If we omit these observations, the pattern of the graph is fairly linear and increasing.

3. **Testing influential observations and Outliers**: Often influential observations can cause problems with modeling results. In order to check the influential observations, we visualized the dfbeta values. The dfbeta values estimates the influence of the i^{th} - patient observation on the regression coefficients β_j. A high value of dfbeta must be investigated carefully.

   Another method for checking influential observations is by assessing the *deviance residuals* (symmetric/normalized transformation of the Martingale residuals) plot. The deviance residual is defined by

$$d_i = sin(\hat{M}_i)\sqrt{2}\sqrt{-\hat{M}_i - \delta_i log(\delta_i - \hat{M}_i)}.$$

In the above equation, $\hat{M}_i$ implies $d_i = 0$. The square root shrinks the large negative martingale residuals, while the logarithm transformation expands those residuals that are close to zero. The distribution of the residuals must approximately be symmetrical around mean zero and standard deviation of one. A very large/small/distant deviance residual values indicate influential observations or outliers. Figure 2.7 below implies that none of the observations is exceedingly influential individually, on average.

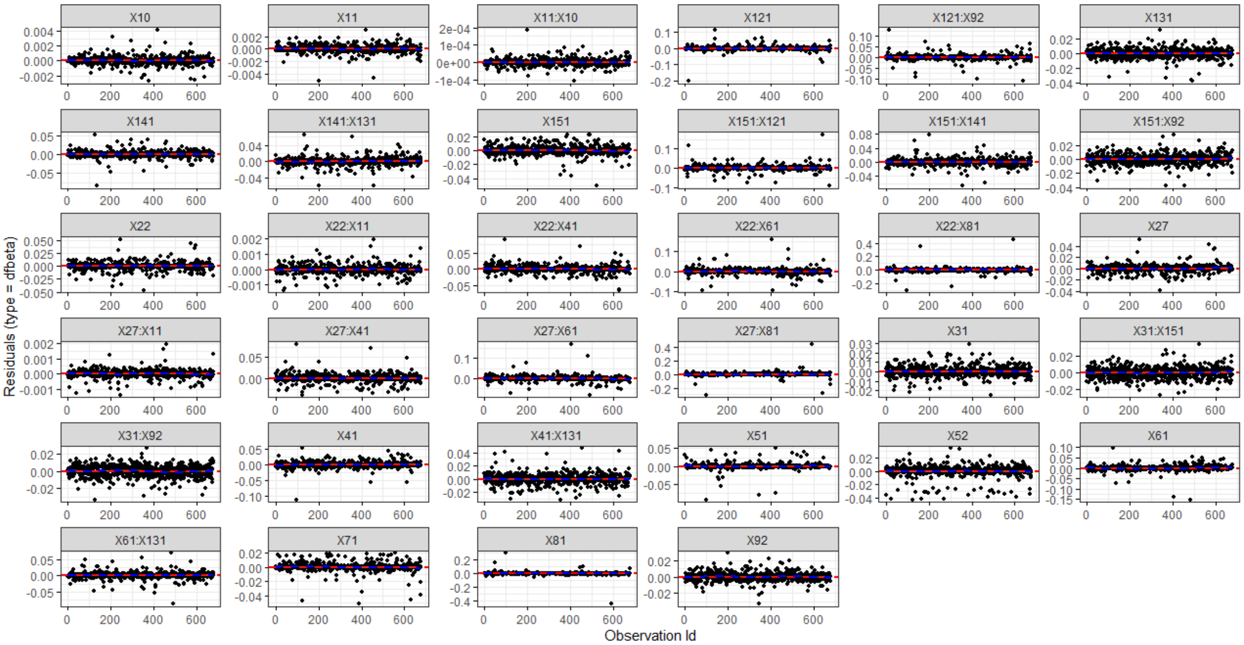


Figure 2.7: Assessing Influential Observations in the Model by dfbeta

The following Figure 2.8 plots the deviance residual and the residual pattern looks fairly symmetrical around zero. The mean deviance residual for our model is .2 which is very small.

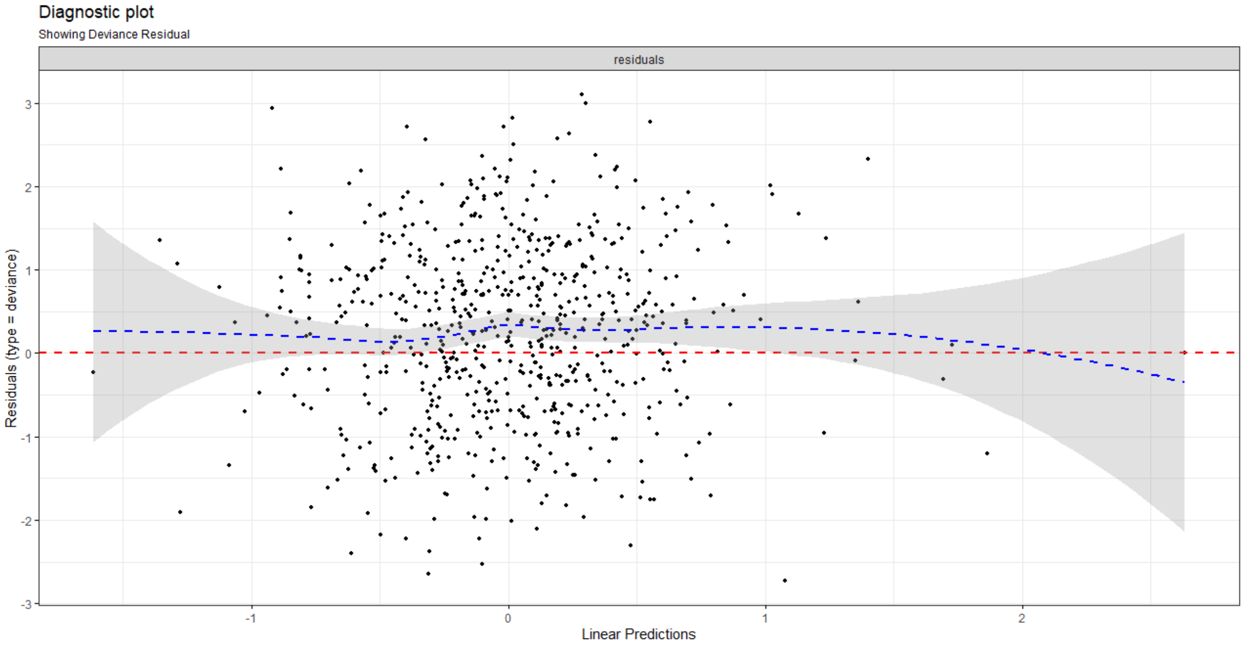


Figure 2.8: Assessing Influential Observations in the Model by Deviance Residual

2.6 Results and Discussions

Given the risk posed by pancreatic cancer in the past few years, it is imperative to investigate the clinical diagnosis and enhance the therapeutic/treatment strategy of pancreatic cancer. The primary treatment for most types of pancreatic cancer is chemotherapy. Sometimes, with chemotherapy, specific therapy drugs are used. Usually, surgery and radiation therapy do not fall under crucial treatments for pancreatic cancer, but they might be used in exceptional circumstances. Also, the treatment approach for children with pancreatic cancer can be slightly different from that used for adults. Several research approaches and statistical methodologies [27] [28] have been developed to cure pancreatic cancer patients and boost their survival times. Chakraborty & Tsokos [28] performed data-driven research on pancreatic cancer patients by performing parametric analysis to improve the survival probabilities of patients of different cancer stages. In the present study, we initially investigated if there exists any statistically significant difference between the *true* mean survival times of the male and female pancreatic cancer patients using the Wilcoxon two-sample rank-sum test.

The p-value ($.47 > .05$) of the test result suggests that there is no evidence of a significant difference between the true mean survival times of the males and the females. Hence, we proceed to perform to develop the Cox-PH (CPH) model with the combined information of male and female patients. While developing the CPH model, it is very important to justify the model assumptions. In the preliminary analysis, we found that all of the risk factors except age (X_1) did not satisfy the proportional hazard assumption. Thus, we introduced stratification in our model by dividing the covariate age into two groups. By doing stratification, we obtained more valid estimates of the other covariates, and the proportional hazard assumption was satisfied for all risk factors, including age. Performing stratification, we restrict the effect of the covariates to be the same for each stratum. Our final developed Cox-PH model given by equation (2.11) identified all the significant risk factors along with all the significant interaction terms as contributing to the hazard. After building our model, we proceed to rank all significant individual risk factors and all possible significant interactions according to the hazard ratio, as shown in Table 2.2. From Table 2.2, we observe that X_{6Y} (patients having diabetes), X_{4Y} (patients taking ibuprofen regularly), X_{2D} (patients who are in stage **distant** (Cancer has spread to distant parts of the body)), X_{9F} (sex), and X_{15Y} (hypertension) are the most contributing risk factors individually to the survival of patients with a hazard ratio (HR) of 2.39, 1.83, 1.63, .7, and .7, respectively. For the risk factor X_{6Y}, HR $= 2.39$ indicates a strong association between the patients having diabetes and increased risk of death due to pancreatic cancer. Keeping the other covariates constant, being a diabetic patient has a 2.39-fold increase in the hazard of death; that is, 2.39-fold increased risk (or decreased survival). It is important to note that according to the American Cancer Society, one of the main risk factors of pancreatic cancer is diabetes which is supported by our study. Also, we have found that those who take ibuprofen regularly have an increased risk of 1.83-fold than those who do not take the medication on a regular basis. Also, being a female has approximately 30% less hazard than a male patient. Among the most significant interactions we have $X_{2R}X_{8Y}$, $X_{15Y}X_{12Y}$, $X_{12Y}X_{9F}$, $X_{3Y}X_{15Y}$, $X_{3Y}X_{9F}$,

$X_{15Y}X_{9F}$, and $X_{2R}X_{11}$ with hazard ratio 8.84, 2.28, 1.98, 1.83, 1.5, 1.37, and 1.01 respectively. The most contributing risk factor is an interaction term ($X_{2R}X_{8Y}$) (patients with emphysema and cancer stage regional with HR = 8.84). However, they do not contribute significantly to survival. We see that X_{15Y} (hypertension) has a lower risk of survival (HR = .79). However, interacting with X_{12Y} (diverticulosis), it has a hazard ratio of 2.28. Also, interacting with X_{3Y} (person who uses Aspirin Regularly), it has a hazard ratio of 2.28. It is also important to note that X_{3Y} individually has lower risk (better survival) with HR = .6. Although X_{12Y} (diverticulosis) and X_{9F} (female) has a hazard ratio less than one, their combined effect remains significant with HR = 1.98.

2.7 Conclusion

In this study, we have estimated the survival probabilities of patients diagnosed with pancreatic cancer using the semi-parametric Cox proportional hazard (CPH) model. We believe the proposed Cox-PH model given by equation (2.11) gives an accurate estimate of the survival probability of patients diagnosed with pancreatic cancer. The stratification of the age produced more reliable estimates of the risk factor included in the CPH model. We identified seven significant risk factors and ten significant interaction terms as contributing to the survival probability of patients diagnosed with pancreatic cancer, as described in Table 2.2. We also ranked those risk factors and their interactions based on the hazard ratio. There have not enough studies been done in the literature that incorporates the **significant interaction effect** of two risk factors. Interaction effects play a major role as a prognostic factor in addition to the individual risk factors in the CPH model. We found some of the risk factors used in our study individually have hazard less than one, but by combining with some other risk factor, the hazard was more than 1.5, and the combined effect was significant. Our final proposed Cox-PH model is of very high quality, robust, and efficient, given by the fact that it satisfies all the major assumptions described in Section 5. The stepwise model selection procedure was utilized to carefully assess and select the risk factors and the interaction term based on their statistical significance to the survival probability. Depending on the survival analysis of the survival times based on the CPH model of the pancreatic cancer patients, we recommend the following.

1. Besides the survival time of patients, if any additional details regarding some of the potential risk factors are known, then use of the Cox proportional hazard (CPH) model can reflect a better picture of covariate effect on survival via hazard ratio.

2. Before implementing the developed CPH model, one should be careful about the fact that the CPH model assumptions are satisfied. In our present analysis, we justified the key assumptions of the CPH model.

3. The significant two-way interaction effects of the risk factors in the CPH model should not be excluded because they can significantly influence the prediction accuracy of the model and survival rate of pancreatic cancer patients, which might lead to serious clinical and therapeutic/treatment issues.

4. The ranking of the individual and interacting risk factors can be wisely used in pancreatic cancer research to improve the treatment options.

2.8 Acknowledgement

The authors are thankful to the National Cancer Institute (NIH) for making the data available.

Chapter 3: A data-driven Predictive Model for Pancreatic Cancer Patients Using Extreme Gradient Boosting

3.1 Introduction

This study focuses on building a efficient survival model based on the risk factors and identify the most contributing factors influencing the survival times of patients diagnosed with pancreatic cancer. In this study, we developed a real data-driven machine learning predictive model with 800 pancreatic cancer patients' information and *ten* risk factors to predict their survival times. To check the validity of the model, we compared the model's performance with ten *deep neural network* models, grown sequentially with different activation functions and optimizers. We also constructed an ensemble model using Gradient Boosting Machine (GBM). Our proposed XGBoost model outperformed all competing models we considered with regards to root mean square error (RMSE). After developing the model, we ranked all the individual risk factors according to their individual contribution to the response predictions, which is extremely important for pancreatic research organizations to spend their resources on the risk factors causing/influencing the particular type of cancer. The three most influencing risk factors affecting the survival of pancreatic cancer patients are found to be the age of the patient, current BMI, and cigarette smoking years with contributing percentages 35.5%, 24.3%, and 14.93%, respectively. Our proposed predictive model is approximately 96.42% accurate in predicting the survival times of the patients diagnosed with pancreatic cancer and performs excellently on test data. The analytical model can be implemented for prediction purposes for the survival times of pancreatic cancer patients, given a set of risk factors.

The response variable of our study is the survival time (in years). Although in most cases, pancreatic cancer remains incurable, researchers have concentrated on how to enhance the survival rates of individuals with pancreatic cancer.

In our study, we developed a non-linear predictive model using Extreme Gradient Boosting (XGBoost) to estimate the survival time of patients diagnosed with pancreatic cancer. Given a set of risk factors (described in Section 2.2), our model predicts the survival of patients with a high degree of accuracy. We also compared our proposed model's accuracy (in terms of RMSE) with Gradient Boosting Machines (GBM) and different deep learning models. In recent years, researchers are prone to using sophisticated machine learning and deep learning algorithms in cancer research because of their high predictive power and learning abilities from data [26][35][27] [28]. There is an increased tendency in the studies published in recent years that applied semi-supervised ML techniques for modeling cancer survival which address both labeled and unlabeled data.[101]. Kourou, Exarchos, et al., 2015 [84] presented a detailed review about the most recent ML research methods applicable to cancer prediction/prognosis with case studies. Ahmad, Eshlaghy, et al., [1] used different ML and DL algorithms like Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN) and compared their performance to predict the recurrence of breast cancer using 10-fold cross-validation. Hayward, Alvarez, et al., [63] developed different predictive models for the clinical performance of pancreatic cancer patients based on machine learning methods. The predictive performance of machine learning (ML) is compared with linear and logistic regression techniques. According to their study, ML offers techniques for improved prediction of clinical performance, and thus, these techniques can be considered as valuable alternatives to the conventional multivariate regression methods in clinical research. Wang & Yoon [90] suggested an online gradient boosting (GAOGB) model based on a genetic algorithm for incremental breast cancer (BC) prognosis. Their proposed GAOGB model was evaluated on the SEER database in terms of accuracy, the area under the curve (AUC), sensitivity, specificity, retraining time, and variation at each iteration. Ma, Meng, et al.,[91] suggested a classification model that uses the power of extreme gradient boosting (XGBoost) in complicated multi-omics data to focus on early-stage and late-stage malignancies separately. Their XGBoost model was applied to four types of cancer data downloaded from The Cancer Genome Atlas (CGA), and the model's performance was compared with other popular machine learning methods (ML) methods. The authors investigated the efficacy of XGBoost on the diagnostic categorization of malignancies in their study and found XGBoost as a robust predictive algorithm. Chen, Jia, et al.,[32] proposed a non-parametric model for survival analysis that utilizes an ensemble of regression trees to determine the variation of hazard functions with respect to the associated risk factors. The scientists used GBMCI (gradient boosting machine for concordance index) software to develop their model and tested its effectiveness against other conventional survival models using a large-scale breast cancer prognostic dataset. In their study, they found the GBMCI to be consistently outperforming other methods based on a number of covariate settings.

3.2 Materials and methods

3.2.1 Data Description

The study data has been obtained from National Cancer Institute (NIH). The data contains information on patients diagnosed with pancreatic adenocarcinoma. We treated the survival time (in days) as the response in developing our model and considered cause-specific death (deaths due to pancreatic cancer) for each patient. Patient survival time is one of the most crucial factors in all cancer studies. It is critical to assess the severity of cancer since it helps to determine the prognosis and find the best treatment options. There were a total of 800 patient information in our study after eliminating the missing observations for which several risk factors were missing. In our study, the response variable is the survival time of patients (in days). There are a total of *ten* risk factors used in our predictive analysis. Seven of those are categorical in nature, and three of them are numeric variables. The descriptions of the risk factors are as follows.

1. panc_exitage (Numeric) (X_1): Age of diagnosis of the patient.
2. Stage (Categorical) (X_2): Pancreatic Cancer Stages, categorized as a) localized, b) regional, and c) distant
3. asp (Categorical) (X_3): Does the person use Aspirin Regularly?
4. ibup (Categorical) (X_4): Does the person use Ibuprofen Regularly?
5. fh_Cancer (Categorical) (X_5): The number of first-degree relatives with any type of cancer.
6. Sex (Categorical) (X_6): Sex of the individual.
7. BMI (numeric) (X_7): Current Body Mass Index (BMI) at Baseline (In lb/in2)
8. Cigarette Years (numeric) (X_8): The total number of years the patient smoked.
9. gallblad_f (Categorical) (X_9): Did the individual ever have gall bladder stones or inflammation?
10. hyperten_f (Categorical) (X_{10}): Did the individual ever have high blood pressure?

A schematic diagram of the data used in our study with the description of risk factors is shown in Figure 3.1 below. As Figure 3.1 illustrates, seven out of ten risk factors are categorical, having two or more categories.

Before starting our analysis of the data, one important question is if there is any statistically significant difference between the survival times of male and female patients diagnosed with pancreatic cancer. To answer this question, we used the non-parametric Wilcoxon rank-sum test with continuity correction and obtained a p-value of .47, which suggests that there is no statistically significant difference between the true mean survival times of patients from both genders at 5% level of significance. Therefore, we performed our analysis by combining the information of males and females.

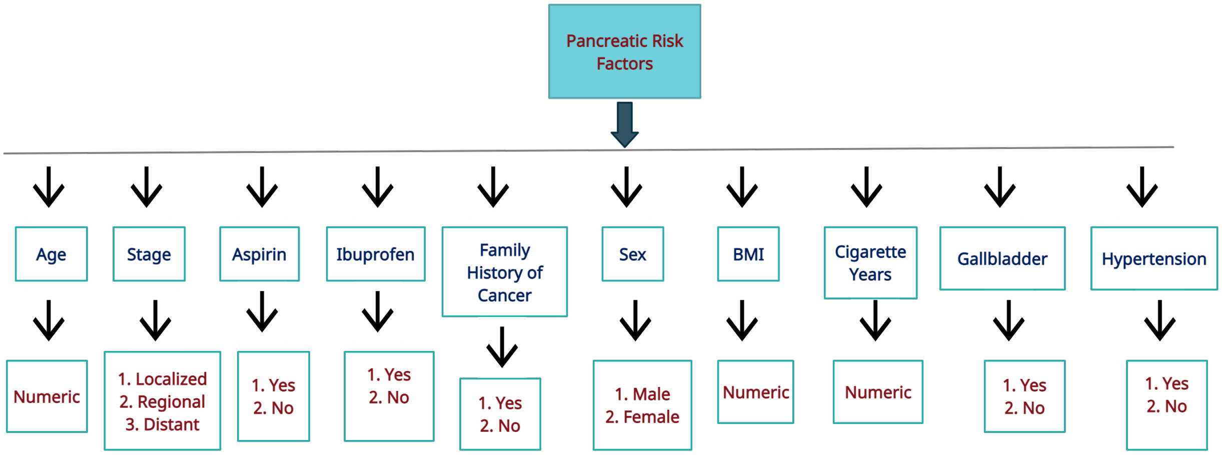


Figure 3.1: Pancreatic Cancer Data with Relevant Risk Factors

3.3 A Brief Overview of Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XGBoost)

In the literature of machine learning, 'Boosting' is a collection of algorithms that transforms the ensemble of a weak learner to strong learners iteratively. Boosting is an ensemble method for improving the model predictions of any given learning algorithm.Gradient boosting machines (GBM), as introduced by Friedman (2001) [50], are a prominent family of machine-learning (ML) algorithms that have demonstrated significant success in a wide range of applied and experimental fields. They are highly customizable to the specific requirement of the application and can be implemented with respect to different loss functions. In this section, we will go through the theoretical notions of gradient boosting briefly.

Let us assume the problem of classical supervised learning problem where we have n risk factors $X = (x_1, x_2, \ldots, x_n)$ and y as a continuous response variable. Given the data, training of the model is performed by obtaining the optimal model parameters θ that best fit the training data x_i and response y_i. To train the model, we define the following objective function to quantify how well the model fits the training data.

$$O(\theta) = L(\theta) + \varrho(\theta) \tag{3.1}$$

where $L(\theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the training loss (mean square error) function that measures the predictive power of our model is with respect to the training data. $\varrho(\theta)$ is the regularization term that helps to prevent model overfitting and controls the complexity of the model.

3.3.1 Decision Tree Ensembles

In our study, we use boosted decision tree ensemble method to train our model. Boosting combines a learning algorithm in an additive manner to achieve a strong learner from many sequentially connected weak learners. A decision tree's major goal is to partition the input space variables into similar rectangular sections using a tree-based rule system. Each tree split corresponds to an if-then rule applied to a single input variable. A decision tree's structure naturally stores and represents the interactions between predictor variables (risk factors). The number of splits, or equivalently, the *interaction depth*, is typically used to parameterize these trees. It is also possible to have one of the variables split numerous times in a row. A tree stump is a special example of a decision tree with just one split (i.e., a tree with two terminal nodes). As a result, if one wishes to fit an additive model using tree base-learners, the tree stumps can be used. Small trees and tree stumps produce remarkably accurate results in many real-world applications.

3.3.2 Model Structure

Mathematically, we can write our analytical model in the form:

$$\hat{y} = \hat{f}(x) = \sum_{i=1}^{K} \hat{f}_i(x), \quad \hat{f}_i \in \mathcal{F} \tag{3.2}$$

where $\mathcal{F}$ is the collection of all possible regression trees, K is the number of regression trees, and $\hat{f}_i$ are the additive functions (additive trees) in $\mathcal{F}$.

$f(x) = w_{q(x)}(q : \mathbb{R}^m \longrightarrow \{1, 2, \ldots, T\}, w \in \mathbb{R}^T)$. Here, q indicates the tree structure that maps an input to the relevant leaf index at which it finishes up. The number of leaves in the tree is denoted by T. Individual regression trees accommodate a continuous score on each of its leaves. w_i represents the score on i^{th} leaf. The tree structures of $\hat{f}_i$ are intractable to learn at once. Hence, we use the following additive strategy. Let $\hat{y_i}^{(t)}$ be the predicted value of the i^{th} observation at step t. Then,

$$\begin{aligned}
\hat{y_i}^{(0)} &= 0 \\
\hat{y_i}^{(1)} &= \hat{f}_1(x_i) = \hat{y_i}^{(0)} + \hat{f}_1(x_i) \\
\hat{y_i}^{(2)} &= \hat{f}_1(x_i) + \hat{f}_2(x_i) = \hat{y_i}^{(1)} + \hat{f}_2(x_i) \\
&\vdots \\
\hat{y_i}^{(t)} &= \sum_{j=1}^{t} \hat{f}_j(x_i) = \hat{y_i}^{(t-1)} + \hat{f}_t(x_i).
\end{aligned} \tag{3.3}$$

Now we have introduced the model; our goal is to define an objective function mathematically and proceed to minimize it. From Equation (1) in Section (3), we have

$$O(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{K} \varrho(\hat{f}_j), \tag{3.4}$$

where $l(\cdot,\cdot)$ is a convex differentiable function that measures the difference between actual y_i and predicted $\hat{y}_i$. $\varrho(\hat{f}_j) = \gamma T + \frac{1}{2}\lambda(\| w \|)^2$. T is the number of leaves in the tree. γ and λ are the model hyper-parameters. From Equation (3) and Equation (4), at the t^{th} iteration, the objective function can be written as

$$\begin{aligned}
O^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{i=1}^{t} \varrho(\hat{f}_i) \\
&= \sum_{i=1}^{n} l(y_i, (\hat{y_i}^{(t-1)} + \hat{f}_t(x_i))) + \sum_{i=1}^{t} \varrho(\hat{f}_i)
\end{aligned} \tag{3.5}$$

Since, we use mean-square error loss function, the above equation takes the following form:

$$
\begin{aligned}
O^{(t)} &= \sum_{i=1}^{n}(y_i - ({\hat{y}_i}^{(t-1)}) + \hat{f}_t(x_i))^2 + \sum_{i=1}^{t}\varrho(\hat{f}_i) \\
&= \sum_{i=1}^{n}(y_i - {\hat{y}_i}^{(t-1)})^2 + \sum_{i=1}^{n}(\hat{f}_t(x_i))^2 \\
&- 2\sum_{i=1}^{n}(y_i - {\hat{y}_i}^{(t-1)})\hat{f}_t(x_i) + \sum_{i=1}^{t}\varrho(\hat{f}_i) \\
&= \sum_{i=1}^{n}(y_i - {\hat{y}_i}^{(t-1)})^2 + \sum_{i=1}^{n}(\hat{f}_t(x_i))^2 \\
&- 2\sum_{i=1}^{n}(y_i - {\hat{y}_i}^{(t-1)})\hat{f}_t(x_i) + \sum_{i=1}^{t-1}\varrho(\hat{f}_i) + \varrho(\hat{f}_t) \\
&= \underbrace{-2\sum_{i=1}^{n}(y_i - {\hat{y}_i}^{(t-1)})\hat{f}_t(x_i) + \sum_{i=1}^{n}(\hat{f}_t(x_i))^2 + \varrho(\hat{f}_t)}_{\text{function of t}} + c
\end{aligned}
\tag{3.6}
$$

where $c = \sum_{i=1}^{n}(y_i - {\hat{y}_i}^{(t-1)})^2 + \sum_{i=1}^{t-1}\varrho(\hat{f}_i)$ is a constant term (not a function of t). From the above expression, the optimal weights of the leaf can be computed that minimizes the objective function. For details, see [32], [121]. In the next section, we discuss briefly the hyper-parameters for Gradient Boosted Machines (GBMs).

3.3.3 Model Tuning Gradient Boosted Machine (GBM)

Although GBMs are highly flexible, they can take significant time to tune and find the optimal combination of hyperparameters. If the learning algorithm is not applied properly with the optimal combination of the hyperparameters, the model is prone to overfitting the data; this suggests that it will predict the training data rather than the functional relationship between the risk factors and response variables. The following are the most typical hyperparameters seen in most GBM implementations:

3.3.3.1 Number of trees

It represents the total number of trees required to match the model. GBMs frequently necessitates a large number of trees. However, GBMs, unlike random forests, can overfit. Hence, the goal is to use cross-validation to estimate the appropriate number of trees that minimize the loss function of interest.

3.3.3.2 Depth of Trees

The complexity of the boosted ensemble is determined by the number of splits in each tree. It is in charge of the depth of the individual trees. Naturally, numbers range from 3 to 8; however, it is not uncommon to have a tree depth of 1. [61].

3.3.3.3 Shrinkage

The introduction of regularization by shrinkage is the traditional strategy to controlling model complexity. Shrinkage is employed in the context of GBMs to reduce or decrease the influence of each additionally fitted base-learner. It decreases the number of incremental steps, penalizing the significance of each successive iteration. The idea behind this strategy is to take many modest steps to improve a model rather than taking a few enormous steps. If one of the boosting iterations is found to be incorrect, the adverse impact can be simply addressed in the following steps. The shrinking effect is usually denoted as the parameter $\lambda \in (0, 1]$ and is applied to the final step in the gradient boosting algorithm. [67].

3.3.3.4 Subsampling

The subsampling approach has been demonstrated to increase the model's generalization features while minimizing the required computation resources. The objective of this approach is to incorporate some unpredictability into the fitting procedure. Only a random subset of the training data is used to fit a consecutive base-learner at each learning iteration. Frequently, training data is sampled without replacement (SWOR). Using less than 100% of the training observations implies the implementation of stochastic gradient descent (SGD). This helps to reduce overfitting and keep the loss function gradient from being trapped in a local minimum or plateau.

Extreme Gradient Boosting (XGBoost) performs in a similar mechanism as GBM using ensemble additive training. Both XGBoost and GBM follow the principle of gradient boosting. However, XGBoost uses some more regularized model parameters to reduce overfitting and obtain the bias-variance trade-off, which improves the performance of the model. For more theoretic and practical applications, see [18] [55]. In the next section, we discuss the statistical data analysis and results.

3.4 Statistical Analysis and Results

One of the most important goals of our study is to predict the survival times of pancreatic cancer patients with the highest degree of accuracy. For that purpose, a number of machine learning (ML) and deep learning (DL) models have been tested and validated on our data. We used Feed forward Deep Learning Models [98] [125] with different layers, optimizer, and activation functions [88]. The best deep learning model that we have obtained is a dense feed-forward network with RMSE **.38** on the test data. However, our proposed XGBoost model does the prediction task with significantly lower RMSE **.04** on test data.

As described in Section 2.1, in our data, we have seven categorical and three numeric risk factors. Usually, most of the ML and DL algorithms do not accept categorical/factor inputs. This implies that the categorical risk factors must be converted to a numerical form. However, in our case, 70% of the risk factors are non-numeric in nature. To overcome this problem, we used a sophisticated technique, termed as ”one-hot-encoding” [114]. It is a tool to convert the categorical predictors to numeric in ML algorithms to do a better job in prediction. After we convert the risk factors to numeric scale, we perform *Min-Max normalization* on the set of risk factors. Min-Max normalization is a tool used in ML tasks to adjust the predictors and response when they are in different scale. Usually, it make all the predictors to fall into [0,1]. It is defined as follows:

$$y^* = \frac{y - min(y)}{max(y) - min(y)} \tag{3.7}$$

where y and y^* are the original response value, and the normalized value of response respectively. After training the XGBoost model, we can back transform to get the original prediction of the response. In our data set, the minimum and maximum responses are .21 years and 21 years respectively. Hence, min(y) = .21 years, max(y) = 21 years, and max(y) - min(y) = (21 - .21) = 20.79 years. Now, we can back transform (7) in the following manner:

$$\begin{aligned} y &= min(y) + y^*[max(y) - min(y)] \\ &= .21 + 20.79y^* \end{aligned} \tag{3.8}$$

We also performed the z-score standardization with the data but, the min-max normalization provided better performance with XGBoost. After normalizing the data, we divided the data into 70% training and 30% test data.

At first, we perform the GBM algorithm on the data. In order to find the best combination of hyperparameters, we performed *grid search* mechanism [12] that iterates through every possible combination of hyperparameter values and enables us to select the most suitable combination. To perform a grid search, we create our grid of hyper-parameter combinations. We searched across 54 models with varying learning rates (shrinkage), tree depth (interaction.depth), and the minimum number of observations allowed in the trees' terminal nodes (n.minobsinnode). We also introduced stochastic gradient descent (SGD) in the grid search (bag.fraction < 1).

The following Table 3.1 shows the combinations of the hyperparameters (abbreviated by S, I.D, N.M, and B.F, respectively) we used for the grid search to obtain 54 models.

Table 3.1: Hyper-parameters and Their Combinations in the Grid Search

Hyper-parameters	**Value Combination**
Shrinkage (S)	(.01, .1, .3)
interaction.depth (I.D)	(2, 3, 5)
n.minobsinnode (N.M)	(5, 10)
bag.fraction (B.F)	(.65, .8, 1)

We loop through each hyperparameter combination and apply the grid search on 1,000 trees. After around 30 minutes, our grid search completes, and we the estimated hyper-parameters for all **54 models**. The following Table 3.2 shows **top ten** models (ascending order of RMSE) with the particular choices of the hyper-parameters.

Table 3.2: Top 10 Models with Hyper-parameters for GBM

S	I.D	N.M	B.F	O.T	min_RMSE
.3	**5**	**5**	**.8**	**47**	**0.03217434**
.3	5	10	1	87	0.03354224
.1	5	5	.8	140	0.03358716
.1	3	5	.8	232	0.03376142
.1	3	5	1	413	0.03376934
.3	5	10	.65	126	0.03377321
.1	5	10	.8	206	0.03380464
.1	2	5	.65	603	0.03382063
.01	5	5	.65	1000	0.03382830
.3	3	10	1	76	0.03386993

From the above table, we see that, while training the model, we obtain the minimum RMSE (**0.03217434**) for the following optimal values of the hyper-parameters in the model:

- shrinkage (S): 0.3
- interaction.depth (I.D): 5
- n.minobsinnode (N.M): 5
- bag.fraction (B.F): 0.8
- optimal_trees (O.T): 47

Now we have the optimal values of the hyper-parameters, we utilize 5-fold cross-validation to train our model with the hyper-parameters. The RMSE we obtained in the test data set using GBM is **0.04222367**.

Now we proceed to perform the data analysis with XGBoost, which is more sophisticated than GBM and has more options to set the hyper-parameters to reduce overfitting. It has several hyperparameters options to train the model. We shall describe briefly the hyperparameters we used for training the model according to the definition given in the **R software** module [117].

- **nrounds**: Controls the maximum number of iterations.
- **eta**: Controls the learning rate, or how quickly the model learns data patterns.
- **max_depth (MW)**: The depth of the tree is controlled by this variable. Typically, the greater the depth, the more complex the model grows, increasing the likelihood of overfitting.
- **min_child_weight (MCW)**: It denotes the smallest number of instances required in a child node in the context of a regression problem. It aids in preventing overfitting by avoiding potential feature interactions.
- **subsample (SS)**: It regulates the number of samples (observations) provided to a tree.

- **colsample_bytree (CSBT)**: It controls the number of predictors given to a tree.

Similar to GBM, we perform a grid search with different combinations of hyperparameters. We trained 243 different hyper-parameter combinations to model. The following Table shows **top ten** models (ascending order of RMSE) with the particular choices of the hyperparameters.

Table 3.3: Top 10 Models with Hyper-parameters for XGBoost

eta	M.D	MCW	SS	CSBT	OT	min_RMSE
.05	**7**	**1**	**.8**	**.8**	**158**	**.0304000**
.05	7	3	1	.8	182	0.0305060
.01	7	1	.8	.65	713	0.0305134
.05	7	3	.8	.8	141	0.0306156
.05	7	3	1	.8	134	0.0306568
.01	7	1	.65	.65	762	0.0307100
.01	7	1	.8	.65	725	0.0307280
.05	7	1	.65	.8	174	0.0307378
.01	7	1	.65	.8	725	0.0307526
.01	7	1	1	.8	816	0.0307682

From the above table we see that the mimimum RMSE (**.0304**) was achieved while training the data when

- eta = 0.05
- max_depth (MD) = 7
- min_child_weigh (MCW) = 1
- subsample (SS) = 0.8
- colsample_bytree (CSBT) = 0.8

- optimal_trees (OT) = 158

Therefore, our final XGBoost ensemble model can be expressed as follows

$$\hat{y^*} = \hat{f}(x) = \sum_{i=1}^{158} \hat{f}_i(x), \ \ \hat{f}_i \in \mathcal{F} \tag{3.9}$$

where $\mathcal{F}$ is the collection of all possible regression trees and $\hat{f}_i$ are the additive functions (additive trees) in $\mathcal{F}$. Our analytical model provides the best results with the optimal values of the six hyper-parameters mentioned above. With the optimal values of the hyper-parameters, we train our model with 5-fold cross-validation and obtained an RMSE of **0.04127676** in test data, which is better than what we obtained using GBM.

We can provide the algorithm to obtain the best analytical model with the optimal hyper-parameters in the following manner:

Algorithm for Obtaining Optimal Analytical Model

Input

- Input Vector: $X = (x_1, x_2, \ldots, x_n)$.
- response y as output.
- Number of iteration T decided by the researcher.
- Mean Square Error Loss Function $L(\theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)$.
- Decision tree as base (weak) learner to be combined in the ensemble.

Algorithm

- for $t = 1$ to T do
    1. Initially, a decision tree is fitted to the data: $\hat{f}_1(x) = y$.
    2. Next, the subsequent decision tree is fitted to the prior tree's residuals: $d_1(x) = y - \hat{f}_1(x)$

3. The latest tree is then added to the algorithm : $\hat{f}_2(x) = \hat{f}_1(x) + d_1(x)$.

4. The succeeding decision tree is fitted to the residuals of $\hat{f}_2 : d_2(x) = y - \hat{f}_2(x)$.

5. The new tree is then added to our algorithm: $\hat{f}_3(x) = \hat{f}_2 + d_2(x)$

6. Use cross-validation while training the model to decide the stopping criteria of the training process.

7. Create a hyper-parameter grid with some user provided values and perform grid search mechanism to find optimal combination of the hyper-parameters.

8. The final analytical model is the sum of all the decision tree base learners with optimal values of the hyper-parameter along with optimal number of trees T^*: $\hat{f} = \sum_{i=1}^{T*} \hat{f}_i$.

- end.

3.4.1 Validation of the Proposed Model

After developing our proposed analytical model, it is most important to validate the model so that we can implement it to obtain the best results. In developing the model, we used 70% of the training data and obtained an RMSE of **.034**. It is a usual tendency of a good model to have a predictive performance in the test data set close to the training data set. When we implement our model on the test data set, we obtained an RMSE of .0422, which is very close to what we have obtained in the training set, implying that our model performs well on the unseen/future data set. We can predict the survival times (in years) by back-transforming the scaled response using equation (8) from Section 4 and compare how good the prediction is. The following Table 3.4 shows the actual and estimated predictions of the pancreatic survival times (in years).

Table 3.4: Predicted and Actual Response

Predicted Response	Actual Response
1.5849055	1.7806254
2.1938655	2.0418507
2.3095083	2.0542900
2.5678812	2.1577326
2.1382802	2.3273000
3.5089106	3.7427615
3.2106355	3.3957704
2.4213239	2.5643014
1.2646362	1.6215333
1.5551881	1.8559159
2.1867148	2.4340161
2.9590347	3.2622116

From the above table, we see that the predictions are very close to the actual response. To validate our prediction accuracy, we also performed Wilcoxon's rank-sum test with continuity correction to check if the actual and predicted responses are significantly different. The test produced a p-value of **.5** ($> .05$), implying that there is insufficient sample evidence to reject the null hypothesis that both actual and predicted responses are the same. Thus, the test suggests there is no significant difference between the actual and predicted responses at a 5% level of significance.

3.4.2 Comparison with Different Models

The XGBoost method performed really well and was about 96% accurate. We compared the proposed boosted regression tree (using XGBoost) model with different deep learning models to validate its performance. Deep learning models are efficient with a large amount of data to train to address the complex structure of features. We used activation functions like rectified linear unit (ReLU), Exponential Linear Unit (ELU), scaled exponential linear units (SELU), and Hyperbolic Tangent (tanh) in different layers of the deep network and used optimizer like stochastic gradient descent (SGD), Root Mean Square Propagation (RMSprop), and Adam (derived from adaptive moment estimation). In some models, we introduced dropouts and batch normalization, and in some models, we did not. Adding dropouts [51] and batch-normalization usually prevents overfitting in the networks and boosts the performance. The theoretical details and applications of the optimizer and activation functions can be found in [32]. Each of the models is trained using 300 epochs and batch size = 32. Table 3.4.6 compares different deep learning models in terms of root mean square error (RMSE) and mean absolute error (MAE) in the test data. In the following table, the activation function, optimizer, dropout, and batch normalization are abbreviated as AF, OPT., DROP., and BN , respectively. We considered ten deep learning sequential models with three dense layers containing units 100, 90, and 50, respectively. As Table 6 illustrates, the best deep learning model (DL6) with minimum RMSE (.378) is the model where we use *tanh* activation function in each of the three hidden layers, use optimizer *Adam*, use *dropout* with *batch-normalization*. The following Figure 3.2 illustrates the graph of RMSE and MAE of DL6 while training.

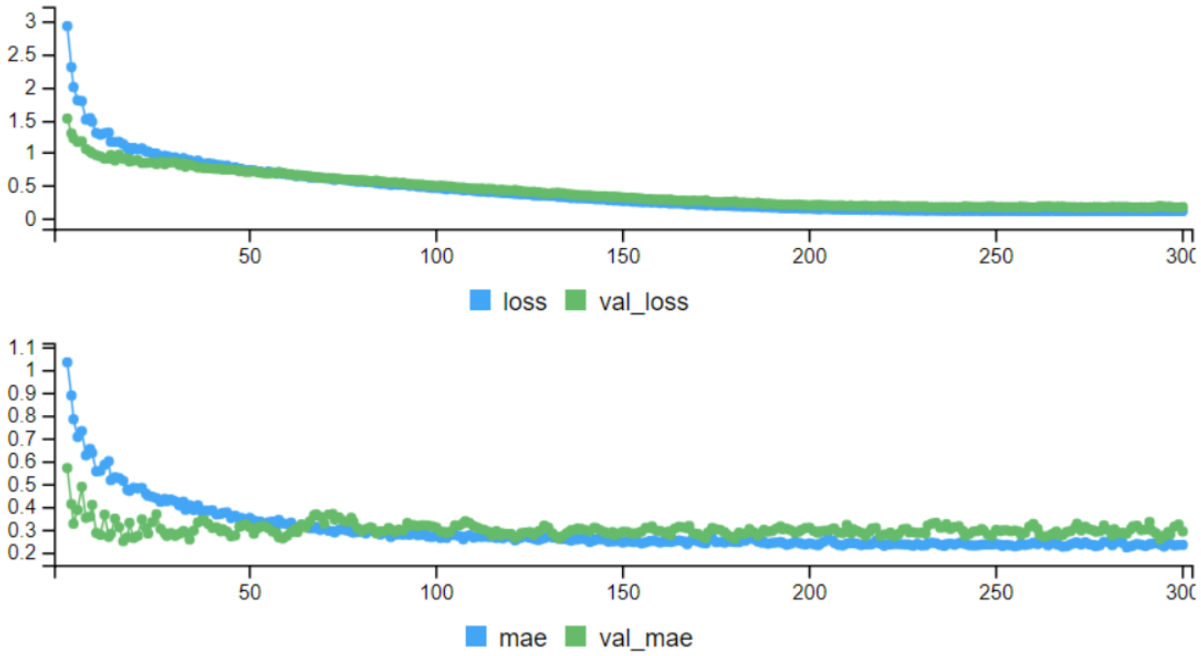


Figure 3.2: RMSE and MAE of DL6 for Training and Validation Data

The following Table 3.5 compares the boosted regression tree model using GBM and XGBoost in terms of RMSE and MAE in test data.

Table 3.5: Comparison of Different GBM & XGBoost Models in Terms of RMSE and MAE in Test Data

MODEL	RMSE	MAE
XGBoost	.0412	.034
GBM	.0422	.039

As the above Table 3.4.5 illustrates, the XGBoost performs the best with the minimum RMSE.

3.4.3 Ranking of Risk Factors and Prediction of the Survival Time

Once we have found the best-performing model, it is important to rank the pancreatic risk factors according to their relative importance. We rank the contributing risk factor in

Table 3.6: Comparison of Different Deep Learning Models in Terms of RMSE and MAE in the Test Data

Model	Unit	AF	OPT.	DROP.	BN	RMSE	MAE
DL1	(100,90,50)	(tanh,tanh,relu)	RMSprop	yes	yes	.381	.26
DL2	(100,90,50)	(ReLU,ReLU,ReLU)	Adam	yes	yes	.391	.24
DL3	(100,90,50)	(ReLU,ReLU,ReLU)	SGD	yes	yes	.9	.255
DL4	(100,90,50)	(ReLU,ReLU,ReLU)	RMSprop	Yes	Yes	.391	.25
DL5	(100,90,50)	(ReLU,ReLU,ReLU)	Adam	No	No	.39	.26
DL6	**(100,90,50)**	**(tanh,tanh,tanh)**	**Adam**	**Yes**	**Yes**	**.378**	**.249**
DL7	(100,90,50)	(ELU,ELU,ReLU)	Adam	Yes	Yes	.388	.234
DL8	(100,90,50)	(ReLU,SELU,ELU)	Adam	Yes	Yes	.385	.232
DL9	(100,90,50)	(ReLU,ReLU,ReLU)	Adam	No	Yes	.49	.4
DL10	(100,90,50)	(ReLU,ReLU,ReLU)	Adam	No	No	.51	.3

survival time using the measure *Gain* [118]. The gain denotes the relative impact of a certain risk factor to the model, which is computed by considering each predictor's contributions to each tree in the model. A higher value of this metric for a specific risk factor, compared to another risk factor, implies that the risk factor with a higher gain is more important for generating a prediction. From Figure 3.3, we see that the top five most contributing risk factors in the model are age, current bmi, the number of years a patient smoked cigarette, people who have family history of cancer, and people who took aspirin on a regular basis. Table 3.7 illustrates the percentage contributions of the risk factors to the response survival times. From Table 3.4.7, see that the risk factors explains **96.42%** of the total variation of the response.

3.5 Conclusion

In cancer research, one of the most important aspects is to estimate the survival times of the patients. It results in improved management, more efficient use of resources, and the provision of specialized treatment alternatives. It is imperative to investigate the clinical diagnosis and enhance the therapeutic/treatment strategy of pancreatic cancer. Pancreatic cancer is one of the deadliest cancer, and most of the cases, detected in later stages (stage III /IV). Once a patient is diagnosed with pancreatic cancer, he/she or his/her family members

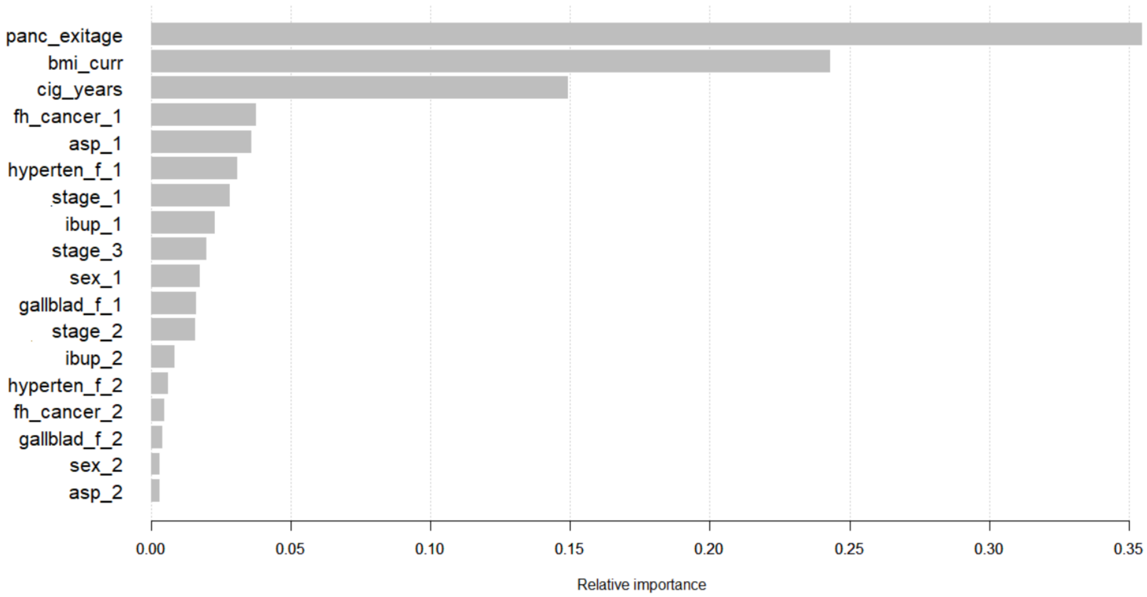


Figure 3.3: The Relative Importance of Risk Factors Used in the XGBoost Model

would be interested in knowing how long is the expected/predicted survival. This question is usually asked by patients with a terminal illness to their doctors. However, it is impossible to provide the exact answer to these questions; doctors provide an answer which is mainly subjective. If we have a model based on real data that answer the questions given a particular choice of risk factors, it would be very helpful to the doctors and medical professionals. Also, if we have some more relevant risk factors, we can incorporate those in this model. This would be very helpful for healthcare professionals and patients with terminal illnesses. Given a collection of risk factors, we can build a questionnaire (attached in Appendix I) that can address the patient information who are diagnosed with pancreatic cancer. Based on their response, the estimate of the survival times can be obtained very accurately. To our knowledge, there is no such model that is as accurate as our predictive analytical model. In this study,

1. We have developed a boosted ensemble regression tree model using XGBoost that is very accurate and performs well on test data set, given a collection of risk factors (numeric and categorical).

Table 3.7: Risk Factors and Their Percentage of Contribution to The Response

Risk Factors	% Contribution
panc_exitage	35.5
bmi_curr	24.3
cig_years	14.93
fh_cancer_1	3.76
asp_1	3.6
hyperten_f_1	3.1
stage_1	2.82
ibup_1	2.29
stage_3	1.96
sex_1	1.73
gallblad_f_1	1.6
stage_2	1.57
ibup_2	.83
hyperten_f_2	.61
fh_cancer_2	.45
gallblad_f_2	.4
sex_2	.29
asp_2	.28

2. We ranked all the risk factors according to their relative importance in the boosted model. This ranking provides the percentage of contribution of the individual risk factors to the response, survival time.

3. We have compared the performance of the XGBoost model with the GBM model and other ten deep learning sequential models with different activation functions and optimizers. The XGBoost model produced an RMSE and MAE of **.0412** and **.034** which is the smallest on the test data compared to all of the other models.

4. Our proposed analytical model can be implemented to any future data set containing information on different risk factors relating to the subject study to obtain very good predictive performance.

Chapter 4: A Stochastic Model for Monitoring the Behavior of Pancreatic Cancer Patients at Different Stages as a function of time

In this study, we have introduced a modern analytical approach using *Survival Index* ($\mathcal{SI}$) to monitor and evaluate the behavior of survival times pancreatic cancer patients. We have considered survival times of patients from three race groups (Caucasian, African-American, and Others) at four different cancer stages, categorized in three different age groups ([40-59), [60-79), and [80-above)). There are a total of 108 patient groups who received three different treatments; only chemotherapy (C), only radiation (R), and a combination of chemotherapy and radiation (C + R). Our analytical method is helpful to predict the pattern of survival intensities based on the *Survival Index* ($\mathcal{SI}$) as a function of time t; which necessarily provides information if the specific treatment has been useful for the particular patient group. We also introduced the concept of *Relative Change in Intensities (RCI)* for patients diagnosed with the subject cancer, which gives the approximate change in the stochastic growth intensity function (SGIF) $\zeta(t)$, for each unit time change. Finally, we have developed an analytical algorithm to compare the survival intensities for any two specific groups out of a total of 108 patient groups without actually computing the stochastic growth intensities. Our analytical methodology based on *Survival Index* ($\mathcal{SI}$) and **stochastic growth intensity function** $\zeta(t)$ is useful and effective for any subject cancer and can be implemented as a modern approach to monitor and evaluate cancer mortality rate as a function of time. The adaptability of our technique stems from the fact that our algorithm may be used to any number of patient groups of any age, of any race, at any specific cancer stage, and receiving any unique treatment or combination.

4.1 Introduction

Given the destructive nature of pancreatic cancer, it remains one of the major threats devastating human existence. However, there are various treatment options (chemotherapy, radiation, surgery, immunotherapy, targeted therapy) to cure the lethal carcinogenic disease; very few studies have been conducted to understand at which stage a particular treatment option is the most effective. Also, it is crucial to understand how the treatment options are affecting the mortality of patients from a specific race belonging to a particular age group, at different cancer stages; which essentially means by applying a particular treatment of interest or combination of both if the mortality of a patient from a specific race at a specific stage is increasing, decreasing, or staying the same. There are several data-driven research in the literature to understand the nature of pancreatic cancer at different stages and what risk factors are the major cause of this type of cancer, [26] [80]. In our study, we have introduced a new analytical approach by defining the Survival Indicator ($\mathcal{SI}$) to monitor the behavior of the cancer survivorship for patients from different age groups, different cancer stages, and from different races, as a stochastic realization of time. The present study uses data from the Surveillance, Epidemiology, and End Results (SEER) database, which contains information on patients diagnosed with pancreatic adenocarcinoma. The analytical model we propose is based on the survival times (in months) and cause-specific death (deaths due to pancreatic cancer) for each patient. The survival times of patients are one of the most pivotal factors used in all cancer research. It is necessary to evaluate the severity of cancer, which helps to determine the prognosis and help identify the correct treatment options. We have extracted a sufficiently large random sample of patients diagnosed with pancreatic adenocarcinoma from different races (white, black, others), and four cancer stages which contain the information of different treatment options (chemotherapy (C), radiation (R), combination of both (C + R)). We have categorized the information for *three* different age groups; 40 to 59, 60 to 79, and 80 and above. The schematic diagrams of the data used in this study for different races, cancer stages, and age groups are shown in Table 4.1, Table 4.2, and Table 4.3 below.

Table 4.1: Showing the Number of Patients for White Population in Different Cancer Stages, Categorized by Age Groups

WHITE			
Age: [40 - 59)			
Stages	C	R	C+R
I	148	12	123
II	1206	52	1351
III	514	33	663
IV	5070	123	556
Age: [60 - 79)			
Stages	C	R	C+R
I	568	75	490
II	3406	249	3268
III	1286	117	1358
IV	11263	305	869
Age: [80 - Above)			
Stages	C	R	C+R
I	237	118	162
II	704	144	478
III	281	451	210
IV	1783	132	120

Table 4.2: Showing the Number of Patients for Black Population in Different Cancer Stages, Categorized by Age Groups

BLACK			
Age: [40 - 59)			
Stages	C	R	C+R
I	34	3	29
II	211	14	249
III	104	13	126
IV	1000	45	101
Age: [60 - 79)			
Stages	C	R	C+R
I	88	15	82
II	394	45	391
III	212	15	203
IV	1476	68	113
Age: [80 - Above)			
Stages	C	R	C+R
I	18	18	5
II	61	15	33
III	17	4	11
IV	157	15	10

Table 4.3: Showing the Number of Patients for Other (American Indian/AK Native, Asian/Pacific Islander) Race Groups in Different Cancer Stages, Categorized by Age Groups

OTHERS			
Age: [40 - 59)			
Stages	C	R	C+R
I	16	1	12
II	118	10	104
III	44	6	65
IV	461	16	62
Age: [60 - 79)			
Stages	C	R	C+R
I	50	9	37
II	263	22	244
III	149	18	147
IV	918	34	101
Age: [80 - Above)			
Stages	C	R	C+R
I	25	14	6
II	63	16	43
III	27	6	21
IV	134	16	20

4.2 Methodology

4.2.1 Analytical Method for Developing the Survival Indicator ($\mathcal{SI}$)

In the context of pancreatic cancer research, research scientists would like to investigate the survival rate pattern as a function of time for patients belonging to a specific race, age group, cancer stages, and specific treatments they received. For example, researchers would be interested in monitoring and evaluating if the failure rate of survival time of a patient belonging to the Caucasian race receiving chemotherapy from age group [60-79) at Stage IV shows an increasing or decreasing trend. As a result, it is critical to track how the survival rate changes over time as a result of the application of a certain treatment. In this regard, We define *stochastic growth intensity factor (SGIF)* that measures the rate of change of a survival time as a stochastic realization of time. The analytical structure of the SGIF function is:

$$\zeta(t; \mathcal{SI}; \phi) = \frac{\mathcal{SI}}{\vartheta}\left(\frac{t}{\phi}\right)^{\mathcal{SI}-1} \quad , \quad \mathcal{SI} > 0, \vartheta > 0, t > 0 \quad , \tag{4.1}$$

Where $\mathcal{SI}$ and ϕ are the shape and scale parameters, respectively, and t denotes the time behavior of the incident under investigation.

For n survival times, $t_1 < t_2 < \ldots < t_n$, (where $t_1 < t_2 < \ldots < t_n$ are the observed and successive), the joint probability density function, $f(t_1, \ldots, t_n)$ can be expressed in terms of $\zeta(t; \mathcal{SI}; \phi)$ as follows,

$$\begin{aligned} f(t_1, \ldots, t_n) &= \prod_{i=1}^{n} \left(\zeta(t_i\right) exp\left[-\int_0^{t_n} \zeta(y)dy\right] \\ &= \prod_{i=1}^{n} \frac{\mathcal{SI}}{\phi}\left(\frac{t_i}{\phi}\right)^{\mathcal{SI}-1} exp\left[-\int_0^{t_n} \frac{\mathcal{AI}}{\phi}\left(\frac{y}{\phi}\right)^{\mathcal{SI}-1} dy\right] \\ &= \frac{\mathcal{SI}^n}{\phi^{n\mathcal{SI}}}\left(\prod_{i=1}^{n}\right)^{\mathcal{SI}-1} exp\left[-\left(\frac{t_n}{\phi}\right)^{\mathcal{SI}}\right], \\ & where \ \ t_1 < t_2 < \ldots < t_n. \end{aligned} \tag{4.2}$$

Implementing the method of Maximum Likelihood Method (MLE) of parameter estimation, we can estimate the parameters $\mathcal{SI}$ and ϕ from (4).The likelihood function for (4) when $T_1 = t_1; T_2 = t_2, \ldots, T_n = t_n$ can be expressed as

$$\begin{aligned} \mathscr{L} = L(t; \mathcal{SI}; \phi) &= \prod_{i=1}^{n} f_i(t \mid t_1, \ldots, t_{i-1}) \\ &= \left(\frac{\mathcal{SI}}{\phi}\right)^n \prod_{i=1}^{n} \left(\frac{t_i}{\phi}\right)^{\mathcal{SI}-1} exp\Bigg[-\left(\frac{t_n}{\phi}\right)^{\mathcal{SI}}\Bigg] \quad . \end{aligned} \tag{4.3}$$

The parameter, $\mathcal{SI}$ is a function of t_n, the maximum failure time or the largest value of the phenomenon of interest. We compute the estimate of $\mathcal{SI}$ by equating the partial derivative of $\mathscr{L}$ with respect to $\mathcal{SI}$ and setting it equal to zero, then solving for $\mathcal{SI}$, given by,

$$\frac{\partial \mathscr{L}}{\partial \mathcal{I}} = 0; \hat{\mathcal{SI}} = \frac{n}{\sum_{i=1}^{n} log\left(\frac{t_n}{t_i}\right)} \quad . \tag{4.4}$$

The parameter ϕ is a function of $\mathcal{SI}$. In a similar way, as above, the estimate of ϕ is computed by equating the partial derivative of $\mathscr{L}$ with respect to ϕ to zero and then, substituting the estimate of $\mathcal{SI}$, given by,

$$\frac{\partial \mathscr{L}}{\partial \phi} = 0; \hat{\phi} = \frac{t_n}{n^{\frac{1}{\mathcal{I}}}} \quad . \tag{4.5}$$

In the context of cancer survivorship, we formally define the *Survival Indicator* ($\mathcal{SI}$) as follows.

Definition 4.2.1. *The Survival Indicator ($\mathcal{SI}$) for a patient group belonging to a particular race, from a specific age group is an* ***index*** *based on the survival time, that determines the improvement or deterioration of survival of that particular group at a specified cancer stage when any definite treatment or a combination of more than one treatment is administered. Mathematically, it can be expressed as follows.*

$$\mathcal{SI}_{jk}^{lm} = \frac{n_{jk}^{lm}}{\sum_{i=1}^{n} log\left(\frac{(t_n)_{jk}^{lm}}{(t_i)_{jk}^{lm}}\right)} \tag{4.6}$$

where $\mathcal{SI}_{jk}^{lm}$ is the $\mathcal{SI}$ of the $j^{th}, (j = 1 = C, 2 = R, 3 = C + R)$ treatment group, at Stage $k, k = 1, 2, 3, 4$, for age group $l, (l = 1 = [40 - 59), 2 = [60 - 79), 3 = [80 - above))$ belonging to race $m, (m = 1 \equiv$ white , $2 \equiv$ black, and $3 \equiv$ others).

The term $(t_n)_{jk}^{lm}$ is the largest time to death, and n_{jk}^{lm} is the number of patients.

For example, $\mathcal{SI}_{32}^{12}$ represents the index indicator value for the **black** patient group, under age group **[80 - above)** at **Stage II** who received only **chemotherapy**. Now, we can express the stochastic growth intensity function (SGIF) for any specific group in the following way:

$$\zeta(t; \mathcal{SI}_{jk}^{lm}; \phi) = \frac{\mathcal{SI}_{jk}^{lm}}{\phi}\left(\frac{t}{\phi}\right)^{\mathcal{SI}_{jk}^{lm}-1} \quad , \quad \mathcal{SI}_{jk}^{lm} > 0, \phi > 0, t > 0 \quad . \tag{4.7}$$

We will show that how $\mathcal{SI}_{jk}^{lm}$ depends on the interpretation of the *SGIF* ζ.

- **Case 1: $\zeta(t)$ is decreasing with time,that is, the patient survival rate is improving as a function of time t**

For $\zeta(t)$ being a decreasing function of t, we have,

$$\begin{aligned}
&\zeta(t) < \zeta(t-1) \quad , for \quad t-1 < t \\
&\Rightarrow \frac{\mathcal{SI}_{jk}^{lm}}{\phi}\left(\frac{t}{\phi}\right)^{\mathcal{SI}_{jk}^{lm}-1} < \frac{\mathcal{SI}_{jk}^{lm}}{\phi}\left(\frac{t-1}{\phi}\right)^{\mathcal{SI}_{jk}^{lm}-1} \\
&\Rightarrow \left(\frac{t}{\phi}\right)^{\mathcal{SI}_{jk}^{lm}-1} < \left(\frac{t-1}{\phi}\right)^{\mathcal{SI}_{jk}^{lm}-1} \\
&\Rightarrow \left(\frac{t-1}{t}\right)^{\mathcal{SI}_{jk}^{lm}-1} > 0
\end{aligned}$$

Replacing t with $(t-1)$, in the above inequality, we have $(\frac{t-2}{t-1})^{\mathcal{SI}_{jk}^{lm}-1} > 0$. Again, replacing $(t-1)$ with $(t-2)$, in above inequality gives us $(\frac{t-2}{t-2})^{\mathcal{SI}_{jk}^{lm}-1} > 0$. Proceeding in a similar manner, we end up with $(\frac{t_1}{t_0})^{\mathcal{SI}_{jk}^{lm}-1} > 0$, where t_0 is the initial time of death.

Arranging all the above inequalities and expressing them in product form gives us,

$$
\begin{aligned}
&\left[\left(\frac{t-1}{t}\right)^{\mathcal{SI}_{jk}^{lm}-1}\left(\frac{t-2}{t-1}\right)^{\mathcal{SI}_{jk}^{lm}-1}\cdots\left(\frac{t_2}{t_1}\right)^{\mathcal{SI}_{jk}^{lm}-1}\left(\frac{t_1}{t_0}\right)^{\mathcal{SI}_{jk}^{lm}-1}\right] > 0\\
&\Rightarrow \left(\frac{1}{tt_0}\right)^{\mathcal{SI}_{jk}^{lm}-1} \quad > 0
\end{aligned}
$$

Since, $t, t_0 > 0$, in order to satisfy the above inequality, $\mathcal{SI}_{jk}^{lm}$ must satisfy, $\mathcal{SI}_{jk}^{lm} - 1 < 0 \Rightarrow \mathcal{SI}_{jk}^{lm} < 1$.

- **Case 2: $\zeta(t)$ is increasing with time, That is, the patient survival rate is deteriorating as a function of time t.**

For $V(t)$ being a increasing function of t, proceeding with the similar logic, we end up with $\mathcal{SI}_{jk}^{lm} > 1$.

- **Case 3: $\zeta(t)$ is constant; that is, the patient survival rate is constant.**

For $\zeta(t)$ being an independent function of t, proceeding with a similar argument, we end up having $\mathcal{SI}_{jk}^{lm} = 1$.

Now, provided the estimates of $\mathcal{SI}_{jk}^{lm}$ and θ, we can calculate the value of the *SGIF*, $zeta(\cdot)$ (given in (7)), which is utilized in modeling the survival growth of a *specific patient group*, receiving any treatment or combination at any given time t. $\zeta(t)$ is a measure of the rate of change in survival growth as a function of time when a patient deteriorates/improves with the use of any given treatment (radiation/chemotherapy/combination of both). A decrease in $\zeta(t)$ implies that *SGIF* is decreasing or an improvement in the survival rate of a patient diagnosed with pancreatic cancer as a function of time. This means that $\mathcal{SI} < 1$. A rise in $\zeta(t)$ suggests that *SGIF* is increasing, implying that $\mathcal{SI} > 1$. This means that the survival rate is decreasing with respect to time. When there is no change in $\zeta(t)$, it implies that $\mathcal{SI} = 1$; thus death rate is constant, and the NHPP becomes a homogeneous Poisson process (HPP) (Rigdon & Basu, 2010). Therefore, the behavior of the change in the cancer survival growth model is dependent on $\mathcal{SI}$ of the intensity function. That is, we can use $\mathcal{SI}$ to monitor the survival rate of patients as a function of time.

4.2.2 Analytical Method for Developing the *Relative Change in Intensity (RCI)*

The *SGIF* ($\zeta(t)$) plays a major role in deciding the pattern of the mortality rate of a group of patients as a function of time under the application of a specific treatment group. Depending upon the values of the Survival Indicator ($\mathcal{SI}$) ($\lesseqgtr 1$), it can predict that if the survival rate increases, decreases, or staying constant. However, what can be said about the *SGIF*s of two groups of patients receiving two different treatments where both the Survival Indicator ($\mathcal{SI}$) is less than 1, or greater than 1? In this section, we investigate how the the *SGIF* changes for two different ($\mathcal{SI}$), where both ($\mathcal{SI}$) is ≤ 1 or > 1.

For any two different groups (can be of a different race, age-group, cancer stage, or treatment group) let $\mathcal{SI}_1$, and $\mathcal{SI}_2$ (for calculation simplicity, we use only one suffix, instead of four) be the survival indicator for two different groups, and ϕ_1 and ϕ_2 be the corresponding scale parameters. The corresponding *SGIF*s be $\zeta_1(t)$, and $\zeta_2(t)$, respectively.

- **Case 1:** $\mathcal{SI}_1 < \mathcal{SI}_2$

From Equation (4.9), we have,

$$\begin{aligned}\frac{\zeta_1(t)}{\zeta_2(t)} &= \underbrace{\left(\frac{\mathcal{SI}_1}{\mathcal{SI}_2}\right)\left(\frac{\theta_2^{\mathcal{SI}_2}}{\theta_1^{\mathcal{SI}_1}}\right)}_{\text{Constant}} t^{(\mathcal{SI}_1-\mathcal{SI}_2)} \\ &= Ct^{(\mathcal{SI}_1-\mathcal{SI}_2)}.\end{aligned} \tag{4.8}$$

From (10), we see that $\frac{\zeta_1(t)}{\zeta_2(t)}$ is a decreasing function of time, since $\mathcal{SI}_1$ ¡ $\mathcal{SI}_2$. Let $h(t) = \frac{\zeta_1(t)}{\zeta_2(t)}$. Then we have,

$$
\begin{aligned}
h'(t) = \frac{\zeta_2(t)\zeta_1(t)' - \zeta_1(t)\zeta_2(t)'}{(\zeta_2(t))^2} < 0 \\
&\implies \zeta_2(t)\zeta_1(t)' < \zeta_1(t)\zeta_2(t)' \\
&\implies \frac{\zeta_1(t)'}{\zeta_1(t)} < \frac{\zeta_2(t)'}{\zeta_2(t)} \\
&\implies t\frac{\zeta_1(t)'}{\zeta_1(t)} < t\frac{\zeta_2(t)'}{\zeta_2(t)}.
\end{aligned}
\tag{4.9}
$$

The term $t\frac{\zeta(t)'}{\zeta(t)}$ in the above expression can be expressed as:

$$
\begin{aligned}
t\frac{\zeta(t)'}{\zeta(t)} &= \lim_{x\to t}\left[\frac{\zeta(x)-\zeta(t)}{x-t}\frac{t}{\zeta(t)}\right] \\
&= \lim_{x\to t}\left[\frac{\zeta(x)-\zeta(t)}{\zeta(t)}\frac{t}{x-t}\right] \\
&= \lim_{x\to t}\frac{1-\frac{\zeta(x)}{V(t)}}{1-\frac{x}{t}} \\
&\cong \frac{\%\Delta\zeta(t)}{\%\Delta t}.
\end{aligned}
$$

In the above expression $\frac{\%\Delta\zeta(t)}{\%\Delta t}$ is the ratio of relative percent change in $\zeta(t)$ with respect to relative percent change in t. It can also be thought of as the approximate change in the *SGIF* $\zeta(t)$ for each unit time change. We define the term as *Relative Change in Intensity (RCI)*. Thus,

$$
RCI = t\frac{\zeta(t)'}{\zeta(t)} \tag{4.10}
$$

From (11), it can also be noted that,

$$
\begin{aligned}
RCI &= t\frac{\zeta(t)'}{\zeta(t)} \\
&= \left[\frac{\frac{\zeta(t)'}{\zeta(t)}}{\frac{1}{t}}\right] \\
&= \frac{\frac{d}{dt}log\zeta(t)}{\frac{d}{dt}logt}
\end{aligned}
\tag{4.11}
$$

Hence, we can also see RCI as the rate of change of the intensity function $\zeta(t)$ in the logarithmic scale with respect to the rate of the chance of time in the logarithmic scale. Now we proceed to define the *Relative Change in Intensity (RCI)* formally.

Definition 4.2.2. *RCI is the the ratio of the relative percent change in the death rate $\zeta(t)$ with respect to relative percent change in the survival time t.*

Definition 4.2.3. *RCI can also be defined as the ratio of the SGIF $\zeta(t)$ and the rate of the survival time in logarithmic scale.*

In this context, it is important to note that,

$$\frac{\zeta(t)'}{\zeta(t)} = \frac{d}{dt} ln\zeta(t) \tag{4.12}$$

Equation (4.12) is the exact rate of change of the log of the intensity function $\zeta(t)$ with respect to time t. From Equation (4.11), we have $RCI_1 < RCI_2$, when $0 < \mathcal{SI}_1 < \mathcal{SI}_2$. That is, if we have prior knowledge about the survival indicator $\mathcal{SI}$ for any two different patient groups where one is less than the other, we can conclude that the *relative change in intensity (RCI)* for the patient group, for which $\mathcal{SI}$ is less, is smaller than the competitive group.

- **Case 2:** $\mathcal{SI}_1 > \mathcal{SI}_2$:

Following a similar approach as in Case 1, we have $RCI_1 > RCI_2$. That is, the *relative change in intensity (RCI)* for the patient group which has greater $\mathcal{SI}$ index, is greater than the competitive group.

- **Case 3:** $\mathcal{SI}_1 = \mathcal{SI}_2$:

Following the similar approach as in Case 1, we have $RCI_1 = RCI_2$. That is, the *relative change in intensity (RCI)* for two patient groups are the same if they have the same survival index $\mathcal{SI}$.

4.2.3 Deriving the Criterion for the stochastic growth intensity $\zeta(t)$ and Time t Based on the Survival Indicator ($\mathcal{SI}$)

Now, we have derived the criterion on *relative change in intensity (RCI)*, we can also determine the range of time t under Case 1; that is, when $\mathcal{SI}_1 < \mathcal{SI}_2$ assuming $\zeta_1(t) \leq \zeta_2(t)$ and $\zeta_1(t) \geq \zeta_2(t)$. We have,

$$\frac{\phi_2^{\mathcal{SI}_2}}{\phi_1^{\mathcal{SI}_1}} = \frac{t_{n_2}}{(n_2)^{1/\phi}} \frac{(n_1)^{1/\phi_1}}{t_{n_1}}. \tag{4.13}$$

Combining Equation (4.10) and (4.15), we have

$$\begin{aligned}\frac{\zeta_1(t)}{\zeta_2(t)} &= \left(\frac{\mathcal{SI}_1}{\mathcal{SI}_2}\right) \frac{t_{n_2}}{(n_2)^{1/\beta_2}} \frac{(n_1)^{1/\beta_1}}{t_{n_1}} t^{(\mathcal{SI}_1 - \mathcal{SI}_2)} \\ &= \delta t^{(\mathcal{SI}_1 - \mathcal{SI}_2)},\end{aligned} \tag{4.14}$$

where

$$\delta = \frac{\mathcal{SI}_1}{\mathcal{SI}_2} \frac{t_{n_2}}{(n_2)^{1/\mathcal{SI}_2}} \frac{(n_1)^{1/\mathcal{SI}_1}}{t_{n_1}}$$

is a constant quantity. We will now proceed to find the range if t under the following assumption

$$\zeta_1(t) \leq \zeta_2(t), \text{and } \mathcal{SI}_1 < \mathcal{SI}_2.$$

By writing the above expression, we are assuming that, while comparing any two patient groups, the group that has the lesser $\mathcal{SI}$ index value, has also the lesser decease/death rate $\zeta(t)$. Let us consider the following.

$$\begin{aligned} &\zeta_1(t) \leq \zeta_2(t) \\ \Longleftrightarrow\ &\frac{\zeta_1(t)}{\zeta_2(t)} \leq 1 \\ \Longleftrightarrow\ &\delta t^{(\mathcal{SI}_1 - SI_2)} \leq 1, (\text{from } (16)) \\ \Longleftrightarrow\ &t^{(\mathcal{SI}_1 - \mathcal{SI}_2)} \leq \frac{1}{\delta} (\text{as } \ \delta > 0) \\ \Longleftrightarrow\ &(\mathcal{SI}_1 - \mathcal{SI}_2) log t \leq log\left(\frac{1}{\delta}\right) \end{aligned}$$

$$\begin{aligned}\iff logt &\geq \frac{log(\frac{1}{\delta})}{(\mathcal{SI}_1 - \mathcal{SI}_2)} \\ &= \frac{-log\delta}{(\mathcal{SI}_1 - \mathcal{SI}_2)} \\ &= \frac{log\delta}{(\mathcal{SI}_2 - \mathcal{SI}_1)}, \text{ as } (\mathcal{SI}_1 - \mathcal{SI}_2) < 0 \\ \iff t &\geq e^{\frac{log\delta}{(\mathcal{SI}_2 - \mathcal{SI}_1)}} \\ &= \left(e^{log\delta}\right)^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}} \\ &= \delta^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}}\end{aligned} \tag{4.15}$$

Hence, from (4.15),

$$\zeta_1(t) \leq \zeta_2(t) \iff t \in [\delta^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}}, \infty).$$

Now, let us consider the case when, $\zeta_1(t) \geq \zeta_2(t)$. Then, proceeding in a similar manner as the previous case, we obtain,

$$\zeta_1(t) \geq \zeta_2(t) \iff t \in (0, \delta^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}}].$$

The above conditions are the *necessary and sufficient* conditions for comparing any two intensities $\zeta_1(t)$ and $\zeta_2(t)$ and obtaining the range of the survival time t. That is, if we have the prior information about the range of the survival time t of any two specific patient groups, we can compare their death intensities $\zeta_1(t)$ and $\zeta_2(t)$ at time t. Conversely, if we have the knowledge that the *SGIF* $\zeta(t)$ of any specific patient group is less/more than the other, we can find the range of the survival time (time to death) for both of the patient groups. This approach can be extended to more than one patient groups for comparison purpose.

4.3 Results

The following tables (Table 4.A, Table 4.B, and Table 4.C) shows the $\mathcal{SI}$ values for Caucasian race group at the four cancer Stages, categorized by three age groups ([40-59), [60-79), and [80-above)), who receive three treatment options (only chemotherapy (C), only radiation (R), and the combination of both (C+R)).

Table 4.A: Showing the $\mathcal{SI}$ and ϕ Values for the Caucasian Race Groups in Different Cancer Stages, for Age Group [40-59)

WHITE						
Age: [40 - 59)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.54	.007	1.09	1.94	.47	.004
Stage II	.45	1.5×10^{-5}	.41	.008	.36	2.4×10^{-7}
Stage III	.39	1.4×10^{-5}	.48	.03	.42	.001
Stage IV	.33	7.2×10^{-10}	.29	5.28×10^{-6}	.37	4.19×10^{-6}

Table 4.B: Showing the $\mathcal{SI}$ and ϕ Values for the Caucasian Race Groups in Different Cancer Stages,for Age Group [50-79)

WHITE						
Age: [60 - 79)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.4	1.74×10^{-5}	.52	.01	.41	4.2×10^{-5}
Stage II	.4	2.15×10^{-7}	.37	4.2×10^{-5}	.46	2.8×10^{-6}
Stage III	.37	4.8×10^{-7}	.33	7.8×10^{-5}	.42	2.8×10^{-6}
Stage IV	.3	4.3×10^{-12}	.32	9.6×10^{-7}	.43	9.65×10^{-6}

Table 4.C: Showing the $\mathcal{SI}$ and ϕ Values for the Caucasian Race Groups in Different Cancer Stages, for Age Group [80-above)

WHITE						
Age: [80 - Above)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.37	4.39×10^{-5}	.41	.0008	.4	.0004
Stage II	.4	8.2×10^{-6}	.43	.0007	.46	.0001
Stage III	.4	6.4×10^{-4}	.5	.02	.53	.002
Stage IV	.33	2.2×10^{-8}	.32	1.5×10^{-5}	.4	.0004

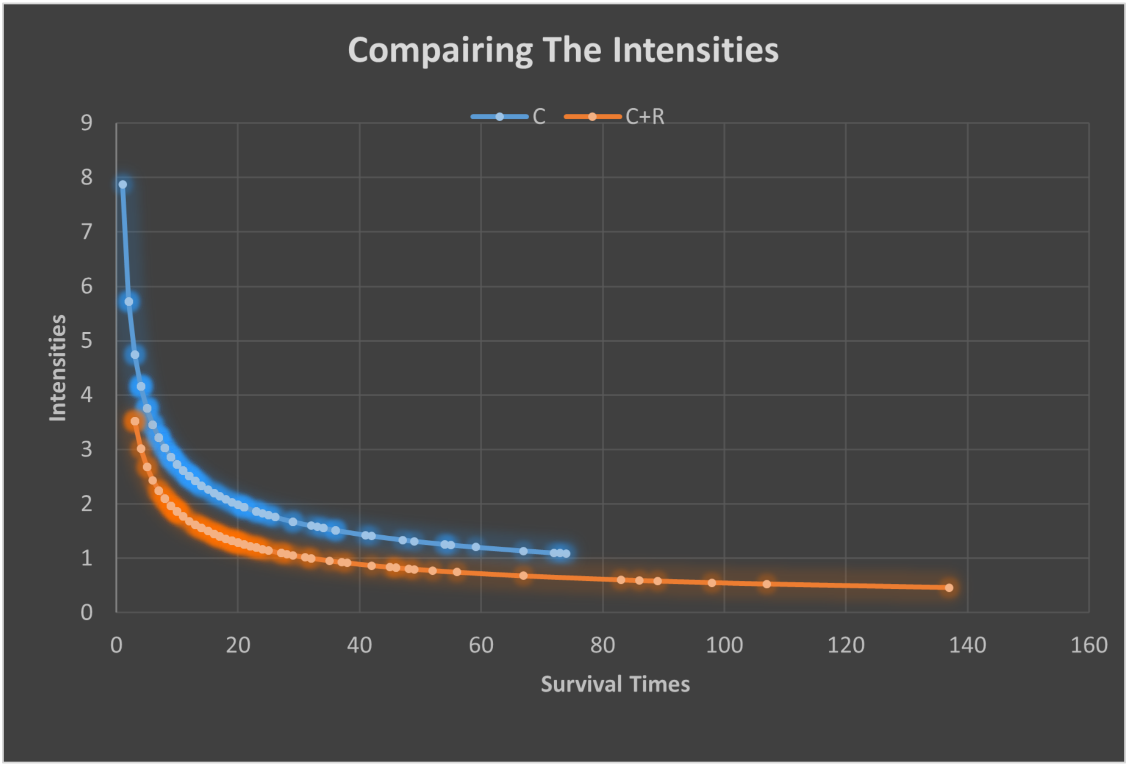


Figure 4.1: Showing the Failure Intensities for Caucasian Race at Stage I, Under Age Group [40-59), Who Received Only Chemotherapy, and the group who received Chemotherapy & Radiation

From Table 4.A, we see that, at Stage I, for age group [40-59), the $\mathcal{SI}$ is .47 for the patient who received chemotherapy and radiation (C+R) together, which is less than the $\mathcal{SI}$ (.54) of the patient group who received only chemotherapy (C). As a consequence, we can infer that, the Relative Change in Intensity (RCI) for C+R group is less than the only C group, which follows from Equation (4.11). In other words, the approximate change in the failure intensity $V(t)$, for each unit time change for group C+R at Stage I, for age group [40-59) is less than the group who receive only C for Caucasian race, which implies that chemotherapy together with radiation has been more effective at Stage I for the particular age group which is also evident from Figure 4.1 above. It is also important to note that, the $\mathcal{SI} = 1.09 (> 1)$ for only radiation (R) group at Stage I implying that the survival intensity is decreasing with time for the particular age group receiving radiation therapy only which is not effective with respect to the survival. The importance of our analytical method is, it can be implemented for any chosen group at any given cancer stage, from any particular age group receiving any specific treatment. The following tables (Table 5.A, Table 5.B, and Table 5.C) shows the $\mathcal{SI}$ values for Black race group at the four cancer Stages, categorized by three age groups ([40-59), [60-79), and [80-above)), who receive three treatment options (only chemotherapy (C), only radiation (R), and the combination of both (C+R)).

Table 5.A: Showing the $\mathcal{SI}$ and ϕ Values for the Black Race Groups in Different Cancer Stages, for Age Group [40 - 59)

BLACK						
Age: [40 - 59)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.12	.6	4.52	12.01	.56	.19
Stage II	.54	.004	.91	1.15	.48	.001
Stage III	.4	.001	.83	1.18	.5	.005
Stage IV	.36	4.45×10^{-7}	.41	.005	.38	.0004

Table 5.B: Showing the $\mathcal{SI}$ and ϕ Values for the Black Race Groups in Different Cancer Stages, for Age Groups [60-79)

BLACK						
Age: [60 - 79)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.4	.0016	1.08	1.19	.43	.005
Stage II	.42	.00005	.38	.0046	.53	.001
Stage III	.42	.00024	1.19	1.03	.65	.014
Stage IV	.35	7.67×10^{-8}	.32	.0001	.64	.02

Table 5.C: Showing the $\mathcal{SI}$ and ϕ Values for the Black Race Groups in Different Cancer Stages, for Age Groups [80-above)

BLACK						
Age: [80 - Above)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.78	.69	.65	.36	1.21	8.19
Stage II	.65	.06	.73	.61	.48	.05
Stage III	1.31	1.95	1.39	6.31	.74	1.49
Stage IV	.43	.0005	.81	.35	1.2	2.05

From the above Table 5, we notice that the survival index ($\mathcal{SI}$) is greater than 1 for African-American patients from age group [40-59), at Stage I receiving only radiation therapy ($\mathcal{SI} = 4.52$), patients from age group [60-79), at Stage I receiving only radiation therapy ($\mathcal{SI} = 1.08$), patients from age group [60-79), at Stage III receiving only radiation therapy ($\mathcal{SI} = 1.19$), patients from age group [80-above), at Stage I receiving both chemotherapy and radiation ($\mathcal{SI} = 1.21$), patients from age group [80-above), at Stage III receiving only chemotherapy ($\mathcal{SI} = 1.31$) and, only radiation therapy ($\mathcal{SI} = 1.39$). We also note that at Stage IV, under the age group [80-above), patients who received both chemotherapy and radiation (C+R), the $\mathcal{SI}$ is 1.2. These results raise red flags regarding implementing the specific treatments to the specific patient groups of African-American race for which $\mathcal{SI}$ is more than 1, implying that the survival rate is deteriorating for these patients.

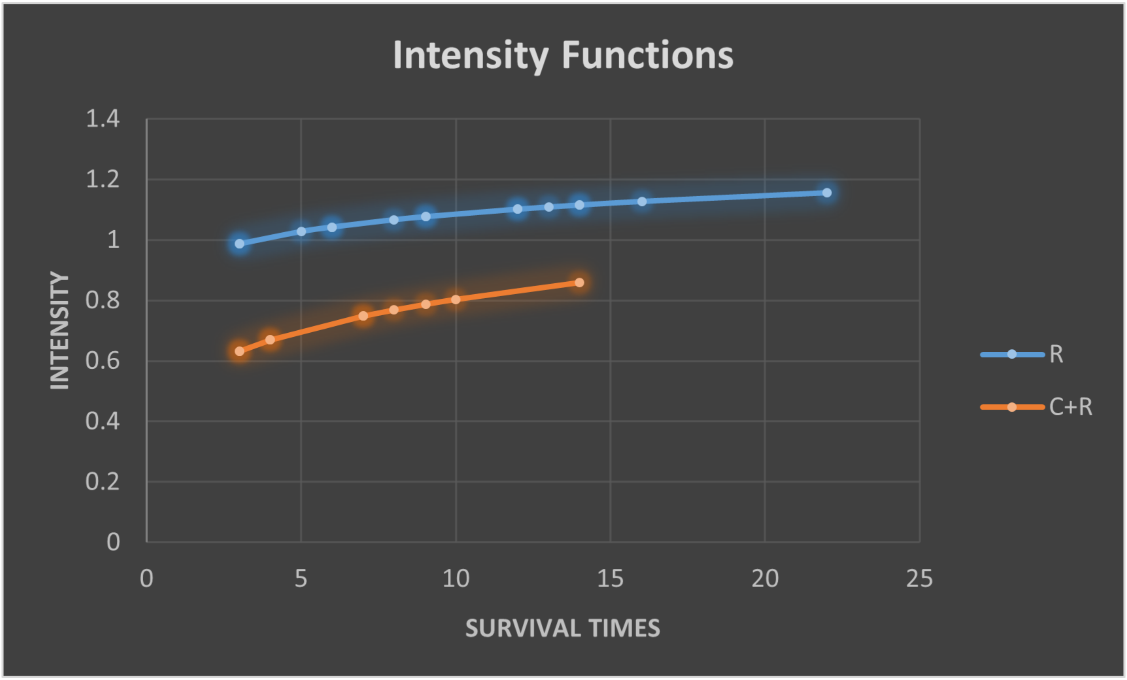


Figure 4.2: Showing the Comparison between the Failure Intensities for African-American Race at Stage I, Under Age Group [60-79), Who Received Only Radiation, and the group who received Chemotherapy & Radiation at Stage IV, under age group [80-above)

We compared the intensities of two specific groups of African-American patients via Figure 4.2 above. From the figure, we see that the failure intensity curve of the patients who received chemotherapy and radiation (C+R) together at Stage IV, under age group [80-above) lies below than the patients who received only radiation (R) at Stage I, under age group [60-79). However, the $\mathcal{SI}$ (1.2) for C+R group is greater than that of $\mathcal{SI}$ (1.08) for R group, the intensity graph for C+R group lies below the graph of R; which necessarily means $\mathcal{SI}_1 < \mathcal{SI}_2$ does not imply $\zeta_1(t) < \zeta_2(t)$ as $\zeta(t; \mathcal{SI}, \phi)$ depends on the time t and another parameter ϕ. However, as Equation (4.11) suggests, $RCI_1 < RCI_2$ if $\mathcal{SI}_1 < \mathcal{SI}_2$.

In our example, $\mathcal{SI}_1 = 1.08 < \mathcal{SI}_2 = 1.2$. Suppose, we are interested RCI_1 and RCI_2 for time $t = 3$. Then we have

$$RCI = \frac{\zeta'(t)}{\zeta(t)} = t(\mathcal{SI} - 1)t^{-1}$$

From the above equation, $RCI_1(t) = RCI_1(3) = 3 \times \frac{(1.08-1)}{3} = .078$ and $RCI_2(3) = 3 \times \frac{(1.2-1)}{3} = .201 > RCI_1(3)$. Hence the relative percentage change in the failure intensity for the patients who received only R at Stage I with respect to the relative percent change in $t = 3$ months is approximately 7.8%. On the other hand, the relative percentage change in the failure intensity for the patients who received both C+R at Stage IV with respect to the relative percent change in $t = 3$ months approximately 20%.

The following tables (Table 6.A, Table 6.B, and Table 6.C) shows the $\mathcal{SI}$ values for Other (American Indian/AK Native, Asian/Pacific Islander) race group at the four cancer Stages, categorized by three age groups ([40-59), [60-79), and [80-above)), who receive three treatment options (only chemotherapy (C), only radiation (R), and the combination of both (C+R)).

Table 6.A: Showing the $\mathcal{SI}$ and ϕ Values for Other (American Indian/AK Native, Asian/Pacific Islander) Race Groups in Different Cancer Stages, for Age Group [40-59)

OTHERS						
Age: [40 - 59)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.55	.74	-	-	.5	.98
Stage II	.59	.02	.96	2.6	.66	.07
Stage III	.63	.14	1.44	5.76	.7	.12
Stage IV	.35	2.62×10^{-6}	.33	.02	.54	.02

Table 6.B: Showing the $\mathcal{SI}$ and ϕ Values for Other (American Indian/AK Native, Asian/Pacific Islander) Race Groups in Different Cancer Stages for Age Group [60-79)

OTHERS						
Age: [60 - 79)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.65	.13	.57	.93	.58	.15
Stage II	.52	.002	.5	.17	.5	.002
Stage III	.51	.003	.5	.13	.61	.01
Stage IV	.31	3.4×10^{-8}	.67	.06	.41	.004

Table 6.C: Showing the $\mathcal{SI}$ and ϕ Values for Other (American Indian/AK Native, Asian/Pacific Islander) Race Groups for Age Group [80-above)

OTHERS						
Age: [80 - Above)						
Treatment	C		R		C+R	
Parameter	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ	$\mathcal{SI}$	ϕ
Stage I	.7	.42	.69	.67	.84	3.19
Stage II	.5	.02	.72	.49	.72	.18
Stage III	.6	.18	1.75	1.1	1.26	1.96
Stage IV	.42	.0003	.77	.3	1.41	1.31

In Table 6.A, the ”-” in Stage I, for age group [40-59) implies that there are insufficient data points to calculate the $\mathcal{SI}$ and ϕ values. From Table 4.3 in Section 4.1, we see that there is only a *single* observation that falls under the category. Since any inference based on a single observation is misleading, we did not calculate the $\mathcal{SI}$ and ϕ values for the specific group of patients. We see that under the age group [40-59) at Stage III, the patients who received only radiation therapy the $\mathcal{SI}$ is 1.44, an indication that the failure intensity is increasing. Also, we see the same scenario for the patients belonging to the age group [80-above) at Stage III who received only radiation ($\mathcal{SI} = 1.75$), for the patients who received chemotherapy and radiation together ($\mathcal{SI} = 1.26$), and for the patients at Stage IV who received chemotherapy and radiation together ($\mathcal{SI} = 1.41$).

4.4 Conclusion

In this study, we focused two main aspects.

- Analytical Development in the subject area.
- Data Analysis and Monitoring the survival Time of a *specific* group of patients.

In Section 2.1, we have defined the Survival Indicator ($\mathcal{SI}$) and explained how it could be implemented in the survival data of pancreatic cancer patients. We have computed the Survival Indicators ($\mathcal{SI}$) for all the all cancer stages, categorized by races and age three age groups. These $\mathcal{SI}$ values play a vital role in deciding the mortality rate of patients as a function of time. We also derived a criterion for the relative change in intensity (RCI) based on $\mathcal{SI}$. The analytical process determines the behavior of RCI when any two $\mathcal{SI} \leq 1$ or ≥ 1 for any two specific groups of patients. Finally, in Section 2.3, we have determined the range of the study time t based on the *SGIF* $\zeta(t)$ of two groups as a function of $\mathcal{SI}$. Our analytical method is useful for determination of the order of any two *SGIF*s $\zeta_1(t)$ and $\zeta_2(t)$ (which one is greater than other) based on the time range $(0, \delta^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}}]$ or $[\delta^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}}, \infty)$. In our study, there are thirty-six groups of patients for each race, totaling ($36 \times 3 = 108$) patient groups. A comparison of the *SGIF*s can be made between any two groups knowing the time range without actually computing the *SGIF*s. Conversely, if we don't have the information regarding the specific survival time t but we know two *SGIF*s $\zeta_i(t)$ and $\zeta_j(t)$ for any two specific groups i and j, ($i \neq j = 1, 2, \ldots, 108$) out of 108 groups, we can estimate the interval for the specific time to death. The whole process can be summarized in an algorithmic form using the following steps:

1. Determine the specific two groups (i and j , $i \neq j = 1, 2, \ldots, 108$) as per requirement.

2. Determine the number of individuals n_i and n_j for the i^{th} and j^{th} groups, respectively.

3. Arrange the observations from the lowest to highest in each groups.

4. Determine the highest observations in each groups t_{n_i} and t_{n_j}.

5. Compute $\mathcal{SI}_i$ and $\mathcal{SI}_j$ using Equation (4.6).

6. Compute $\delta = \frac{\mathcal{SI}_i}{\mathcal{SI}_j} \frac{t_{n_j}}{(n_j)^{1/\mathcal{SI}_j}} \frac{(n_i)^{1/\mathcal{SI}_i}}{t_{n_i}}$.

7. $\zeta_i(t) \leq \zeta_j(t) \iff t \in [\delta^{\frac{1}{(\mathcal{SI}_j - \mathcal{SI}_i)}}, \infty)$ and $\zeta_i(t) \geq \zeta_j(t) \iff t \in (0, \delta^{\frac{1}{(\mathcal{SI}_j - \mathcal{SI}_i)}}]$, where $\zeta_i(t)$ and $\zeta_j(t)$ are the SGIs for i^{th} and j^{th} groups at time t.

8. $SGF_1(t) \leq SGF_2(t) \iff t \in [\delta^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}}, \infty)$
   $SGF_1(t) \geq SGF_2(t) \iff t \in (0, \delta^{\frac{1}{(\mathcal{SI}_2 - \mathcal{SI}_1)}}]$

Chapter 5: S&P 500: Real Data-Driven Analytical Predictive Model For Health Business Segment (HBS).

The S&P consists of 500 large-cap companies that are selected based on size, liquidity, and industry. It is very important to all investors in the stock market. S&P consists of eleven business segments, which are classified according to the type of industry. Investors look at S&P 500 to assess the overall behavior of the stock market. We are concerned with the HBS of the S&P 500, which consists of 59 large-cap health companies. The HBS is the second largest segment that constitutes the S&P 500, about 14.6% of all large-cap health companies. The HBS incorporates businesses that supply medical services, manufacturing medical equipment, development of drugs, provide health insurance, etc. It is one of the important sectors and contributes significantly to the U.S economy, approximately a fifth of the overall gross domestic product (GDP). Thus, healthcare stock behavior has an enormous impact on the global economy; our objective is to develop a non-linear predictive model to predict the weekly average stock price of the 59 stocks. We utilize the average Weekly Closing Price (WCP) of all the HBS from August 2017 to December 2019. In building our analytical model, we have identified six financial and four economic indicators along with thirty-one interactions of the indicators that contribute significantly to the WCP of the HBS stocks. We rank all forty-one indicators as to the percentage of contribution to the WCP. Furthermore, we utilize an analytical optimization process to determine the actual values of these indicators that will maximize the WCP. The proposed analytical model was evaluated by several statistical methods, and it is of very high quality, approximately 96.47% efficient.

5.1 Introduction

Stock price maximization is one of the most significant attributes for value maximization objectives. Stock prices are the most distinguishable of all financial measures that can be used to evaluate the performance of a number of companies. The firms persistently update their information regarding the stock price to reflect any new financial details. Thus, managers are repeatedly judged about their actions, with the benchmark being the stock price performance. Stock prices reflect the long term movement of business decisions of a firm. When firms' stock prices are maximized, investors can realize capital gains instantly by selling their shares of the company. An increase in the stock price is frequently attributed to management's value creation performance. The stock price oscillates over time by showing some dramatic ups and downs. To stay on top of their assets, some investors like to regularly watch these movements. However, if one doesn't keep a track of the stocks on daily basis, keeping track of the net change percentage over time is vital to sustain a prosperous portfolio. Healthcare is a major requirement for everyone, or at least almost everyone needs it at some point in their lives, and when there is something that everybody requires, there's a massive opportunity for the investors. More than 7.8 trillion is spent on healthcare globally. Approximately half of that total, 3.5 trillion, is spent in the U.S.Because the healthcare sector is developing at a higher rate than the global economy, these figures will presumably be considerable by the end of the decade. The indicators we have included in our model have significant relevance in the literature of finance. We have considered six financial indicators and four economic indicators in our proposed model, which will be described in detail in the next section. Many researchers and business analysts strongly believe that dividend yield plays a crucial role in stock returns. Stocks with high dividend yields usually enjoy attractive return advantages over their lower-yielding counterparts. One of the most crucial indicators to influence the return is the price to earnings ratio. Studies[85] have found a direct relationship of price to earnings ratio(P/E ratio) with the stock return, and the returns were changed more by P/E ratio than Price/Earnings-to-Growth(PEG ratio), and thus, stock returns of firms are more affected by the P/E ratio than PEG ratio. Lemmon and Portniaguina[87] have shown that consumer's confidence exhibits forecasting power for the stock return. Tang and Shum[126] have found a significant relationship between the ups and downs of individual stock and the beta risk. We have also considered the US GDP and US personal saving rate as our indicators in the development process of our model as personal saving rate and GDP of a country are vital for a nation as a whole[52]. This is because the current saving rate influences the future consumption and investment in financial assets. In our developed analytical model, the dependent variable is the average weekly closing price for the 59 healthcare stocks, thus, the developed analytical model contains significant contributable variables (indicators) and significant interactions of the indicators. Also, we ranked the indicators according to their percentage of contribution to the response. The validation and quality of our proposed analytical model have been statistically evaluated using R square (R^2), R square adjusted ($R^2_{adjusted}$), and root mean square error ($RMSE$). We also performed residual analysis (section 2.3) to validate our proposed model. To the best of our knowledge, no such statistical model has been constructed to predict the weekly percentage change in healthcare stocks using the proposed logical framework. Therefore, having an appropriate statistical model for the prediction of the weekly average stock price of HBS is important.

5.2 Methodology

5.2.1 The Data and Description of The Indicators

The data of HBS of the S&P 500 that was used to build our analytical model was obtained from the following sources:

1. yahoo finance (https://finance.yahoo.com/)

2. U.S Bureau of Economic Analysis (https://www.bea.gov/)

3. US Bureau of Labor Statistics (https://www.bls.gov/).

Our database that can be summarized by the following schematic diagram.

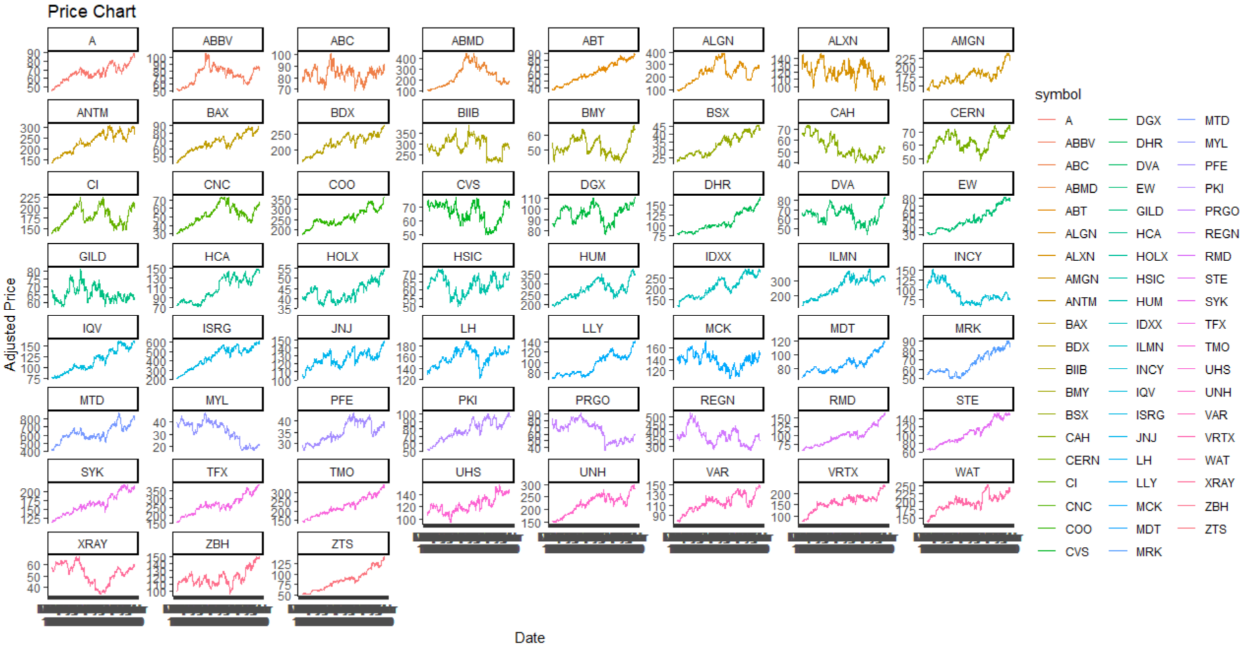


Figure 5.1: Price Chart of The 59 Health Care Stocks

There were 315 pieces of information related to the top 59 healthcare stocks. In structuring the data matrix in a meaningful way, we took the average of all 59 stocks for all the indicators and response, WCP. Our data contains average weekly (five days) information. Our data includes the information from October 2^{nd}, 2017 to December 31^{st}, 2018. We have collected data based on *six* financial indicators and *four* economical indicators. A five day period moving average (MA) method was used for each of the indicators to structure our data. One of the main goals of our study is to understand what indicators and their interactions significantly affect the variation of the stock price of the healthcare management system as a whole.

We have *ten* indicators and thirty-one interactions that drive the average of **WCP** (Weekly Closing Price) as a measure of the response.

The description of the attributable variables (indicators) that the data was collected on is given below.

5.2.1.1 Financial Indicators

The six financial indicators that we have found significantly contribute to WCP are:

1. **Div_Yield**(X_1): *The dividend yield* is a financial measure that demonstrates how much a company disburses in dividends each year with respect to its stock price. It is the annual dividend rate divided by the current share price. It is expressed in a percent form. For instance, if the current stock price is \$50, and the annual dividend is \$1, the dividend yield is 2 percent.

2. **Beta**(X_2): Beta is a risk measure of a stock's volatility of return with respect to the overall market. In general, a stock with a higher beta value tends to have a higher risk and also higher expected returns. It is defined as follows:

$$Beta = \frac{Cov(R_I, R_M)}{Var(R_M)} \quad ,$$

   where R_I is the return on an individual stock, and R_M is the return on the overall market. $Cov(\cdot, \cdot)$ is the covariance between R_I and R_M, i.e., how the changes in stock return are related to the changes in the market return, $Var(\cdot)$ is the variance measure implying how far apart the market data is scattered from their average market return.

3. **PE**(X_3): *The Price-to-Earnings Ratio (P/E ratio)* is the ratio that measures the current share price of a stock with respect to its earning per share (EPS). It is defined as follows:

$$P/ERatio = \frac{\text{Market value per share}}{\text{Earning per share}} \quad .$$

4. **FSCORE**(X_4): *The Piotroski F Score* or *Piotroski Score* was developed by Chicago Accounting Professor Joseph Piotroski, who devised a scale, according to some specific aspects of the company's financial statements. The Piotroski score is a discrete numerical score between 0 to 9 that reflects nine criteria used to decide the strength

of a firm's financial stability. The score is utilized to determine the best value stocks, with nine being the best and zero being the worst. The nine criteria are as follows:

- Positive return on assets in the present year (1 point).
- Positive operating cash flow in the present year (1 point).
- Higher return on assets (ROA) in the current period compared to the ROA in the previous year (1 point).
- Cash flow from operations is higher than Net Income (1 point)
- Lower ratio of long term debt to the current period compared value in the previous year (1 point).
- The higher current ratio in the current year relative to last year (1 point).
- No new shares were issued in the previous year (1 point).
- A higher gross margin in contrast to the previous year (1 point).
- A higher asset turnover ratio compared to last year (1 point).

The company's profit potential is evaluated with the first four criteria. The fifth and sixth criteria focus is on the solvency (including debt) of the company. The last two rules look at the operational efficiency of the company. Companies with an F score of 2 or lower are deemed to be very weak in performance, and companies that score 8 or 9 are considered stable and influential to the financial market.

5. **EBITDA**(X_5): *EBITDA, or Earnings Before Interest, Taxes, Depreciation, and Amortization*, is a measure of a company's overall financial performance and is used as an alternative to simple earnings or net income in some circumstances. EBITDA is frequently used to measure corporate profitability. It is defined as follows:

$$EBITDA = NI + I + T + DE + AE = OP + DE + AE,$$

where, NI = Net Income, I = Interest, T = Taxes, DE = Depreciation expense, AO = Amortization expense and $OP = NI + I + T$.

6. **FCF**(X_6): *Free Cash Flow(FCF)(in Billions)* characterizes the generated cash of a company after deducting away the purchase of assets such as property, equipment, and other major investments from its operating cash flow. Free cash flow is a key measurement since it describes how productive a company is at generating cash. Investors utilize free cash flow to measure whether a company might have sufficient cash, after all the capital expenditures, to pay investors through dividends and share buybacks.

5.2.1.2 Economic Indicators

The four economic indicators that we have found to significantly contribute to WCP are:

1. **US_GDP**(X_7): Gross domestic product of the United States (in trillons).

2. **US_ICS**(X_8): *The Index of Consumer Sentiment*, ICS, or economic well-being was developed at the University of Michigan Survey Research Center to measure the confidence or optimism (pessimism) of consumers in their future well-being and upcoming economic conditions. The index measures short and long-term expectations of business conditions and the individual's perceived economic well-being. Evidence[69] indicates that the ICS is a leading indicator of economic activity as consumer confidence seems to pave the way for major spending decisions.

3. **US_PSR**(X_9): *The U.S. Personal Saving Rate* PSR is personal savings as a percentage of disposable personal income. In other words, it's the percentage of people's incomes left after they pay the essential expenses. The U.S. Bureau of Economic Analysis (BEA) publishes this rate.

4. **US_INFL**(X_{10}): *The Inflation Rate*, INFL, is defined as the percentage increase or decrease in prices(value of a currency) in the course of a given time period, generally, a

month or a year. The percentage indicates how quickly prices rose during the specified period. For instance, if the inflation rate for a gallon of gasoline is 2.5% per year, then gasoline prices will be 2.5% higher next year. Inflation [131] is one of the major metrics used by the US Federal Reserve to estimate the health of the economy and globalization.

In developing the proposed analytical model for the average stock price of the health segment as a function of the different indicators, one of the main assumptions is that the response variable WCP should follow the Gaussian probability distribution. From the following Q-Q plot in Figure 5.2, we see that the values of the response WCP is positively skewed and does not entirely follow a Gaussian probability distribution.

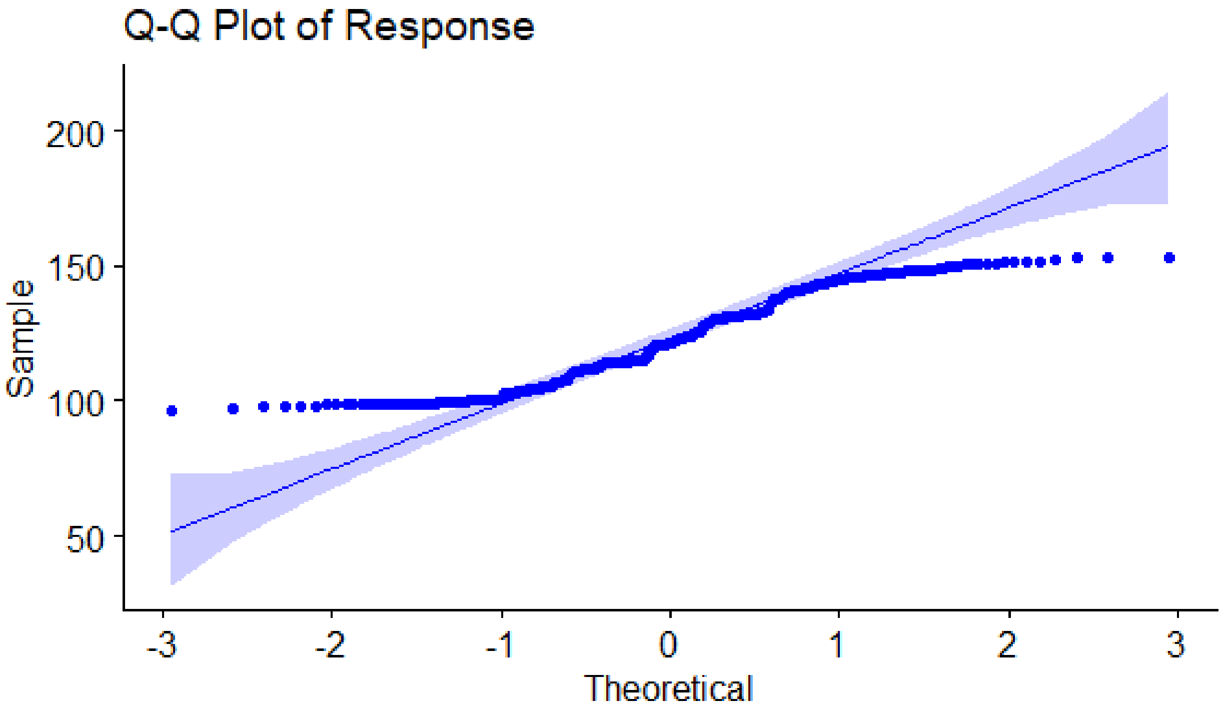


Figure 5.2: Q-Q Plot Of The Response WCP

We have also shown through goodness-of-fit testing (Shapiro-Wilk normality test, a p-value $= 6.4 \times 10^{-11}$) that the subject data does not follow the normal probability distribution as well. Thus, we must address this issue.

5.2.2 Development of the Analytical Model

In developing the analytical model, our main goal is to express our response variable WCP in terms of a non-linear mathematical function of all indicators with a high degree of accuracy. Thus, we proceed to develop the statistical model which is given by the average weekly stock price as a function of the ten indicators which we will show that make a significant contribution to the WCP and all possible interactions as previously discussed. The general analytical form of such model that includes all possible indicators and interactions can be expressed by:

$$WCP = \beta_0 + \sum_i \alpha_i x_i + \sum_j \gamma_j k_j + \epsilon \quad ,$$

where β_0 is the intercept of the model, α_i are the coefficients (weights) of the i^{th} individual attributable variable x_i, γ_j is the coefficient of j^{th} interaction term k_j and ϵ denotes the random disturbance or residual error of the model. One of the main assumptions in constructing our model is that the response variable WCP should follow the Gaussian probability distribution. As we illustrated above, the dependent variable WCP does not support the Gaussian probability distribution. Therefore, we apply a non-linear transformation to the response variable to determine if the transformation can adjust the data of the response to follow a normal probability distribution. We used the Johnson S_B transformation[106] to address the problem which results in equation 5.1, below:

$$z = \gamma + \delta ln\Big(\tfrac{x-\epsilon}{\lambda+\epsilon-x}\Big) \quad , \qquad \epsilon < x < \epsilon + \lambda$$

and

$$TWCP = .18 + 0.5ln\Big(\frac{x - 96.55}{56.63 + 96.55 - x}\Big) \quad . \tag{5.1}$$

Here, $TWCP$ represents the new response variable(transformed) after Johnson's Transformation has been applied. The transformed data were tested and indeed follow the Gaussian probability distribution. Thus, we proceed to estimate the coefficients (weights) of the actual indicators for the transformed data as shown in equation 1. To develop our statistical model, we initially begin with the full statistical model, which included all ten indicators as previously defined and all possible interactions between each pair of indicators. Thus, at first, we start structuring our model with $\binom{n}{k} = 45(n = 10, k = 2)$ potential interaction terms and ten indicators. While we began with the full statistical model, as we mentioned above, we have applied the process to determine the most significant contributions of both the individual indicators and interactions by eliminating the less important indicators and interactions gradually. We used the backward elimination method for this purpose, which is deemed one of the best traditional methods for a small set of feature vectors to tackle the problem of overfitting and perform feature selection. To get better accuracy, we use the log transformation of the indicator PE (X_3) to reduce its high variability. Our statistical analysis has shown that all *ten* indicators significantly contribute to the response, WCP. We now proceed to identify the significant interactions of all ten indicators. Testing the 45 possible interactions of the indicators we found that 31 among them to significantly contribute to the response. Thus, the best proposed statistical model with every significant indicator and interaction that accurately estimates the response WCP are ten indicators individually that significantly contribute and thirty-one interaction terms. Hence, the best preferred analytical model with all significant indicators and interactions that accurately estimates the weekly average stock price of HBS along with estimates of the corresponding weights is given by the following analytical model.

$$\widehat{TWCP} = \begin{cases} .0003 - 0.002X_1 + .038X_2 - .04log(X_3) \\ -0.03X_4 - 0.0001X_5 + .0034X_6 - 0.04X_7 \\ -0.26X_8 - 0.03X_9 - 0.02X_{10} + .51X_1X_3 \\ +.01X_1X_5 - .54X_1X_6 + .29X_1X_7 + 8.2X_1X_8 \\ -.36X_1X_9 - .35X_1X_{10} - .33X_2X_3 + .25X_2X_4 \\ -.88X_2X_6 - .014X_2X_{10} + .9X_3X_5 - .2X_3X_7 \\ +4.96X_3X_8 + .38X_4X_5 - .2X_4X_6 + 5.91X_4X_7 \\ +1.1X_4X_8 - 8.38X_4X_9 - .12X_4X_{10} + .02X_5X_7 \\ +.33X_5X_8 + .015X_5X_9 - .72X_6X_7 - .13X_6X_8 \\ -.54X_6X_9 + .32X_6X_{10} - 5.74X_7X_8 \\ +9.8X_7X_{10} - 2.2X_8X_9 + .36X_9X_{10} \end{cases} \tag{5.2}$$

The *TWCP* estimate is obtained from equation (5.2) above and is based on the Johnson transformation of the data; thus, we will utilize the anti-transformation on equation (3) to estimate the desired, predicted value of the average weekly stock price (WCP) as follows:

$$
\begin{aligned}
\widehat{WCP} &= \hat{\epsilon} + \frac{\hat{\lambda}}{1 + exp\left(\frac{\widehat{TWCP} - \hat{\gamma}}{\hat{\delta}}\right)} \quad .\\
\widehat{WCP} &= 96.55 + \frac{56.63}{1 + exp\left(\frac{\widehat{TWCP} - 0.18}{0.5}\right)} \quad .
\end{aligned} \tag{5.3}
$$

The proposed analytical model will help social researchers, economists, and financial analysts to understand how the weekly stock price varies when any one of the ten indicators is varied, keeping the other indicators fixed. Similarly, with the significant interactions. Most commonly, it will estimate the *predicted estimates* of the response of WCP given the indicators fixed at a specified level. For example, given, $X_1 = 1.36, X_2 = 1.1, X_3 = 48.33, X_4 = 6.27, X_5 = 3.88, X_6 = .94, X_7 = 21.046, X_8 = 94.04, X_9 = 7.86, X_{10} = 1.694$, we obtain the predicted response value as 97.03 (from equation (3)). So, given all the values of the indicators, fixed at a particular level, the weekly average stock price for all healthcare stocks is \$97.03. Furthermore, we illustrate the percentage that the indicators and the interactions contribute to the response , WCP, as we rank them according to their importance (weights), which is shown below in Table 5.1.

The ranking of the indicators that drive the WCP of the HBS is important to the investor. That is, monitoring the behavior of the indicators with respect to current existing data can predict the direction of WCP. Also, the individual health companies that constitute the HBS can utilize the information to increase their company's stock value by concentrating on improving the indicators that contribute most to the WCP. It would also be interesting to compute the percentage of contribution of each indicator individually, combining it with the other indicators, and rank them with respect to the contribution of the response, WCP with reference to Table 5.1. Based on the number of occurrences of each of the ten indicators and their interactions in model 2, we add the contribution percentage to the response, WCP. The total sum of the fourth column of Table 5.2 is more than 100 since we have considered the repeated terms (for example, while determining the percentage of contribution of $GDP(X_7)$,

Table 5.1: Ranking of the Indicators and the Interactions with Respect to the Percentage of Contribution to the Response WCP

Rank	Indicators	Contr.(%)
1	$FCF \cap US\_ICS$	4.53
2	$FSCORE \cap US\_INFL$	4.15
3	$EBITDA \cap US\_ICS$	3.89
4	$GDP \cap US\_ICS$	3.63
5	$PE \cap US\_ICS$	3.41
6	$US\_PSR \cap US\_INFL$	3.34
7	$FSCORE$	3.34
8	$EBITDA$	3.27
9	$BETA \cap US\_INFL$	3.07
10	FCF	3.05
11	$FCF \cap US\_INFL$	3.03
12	DIV_YIELD	2.86
13	US_ICS	2.71
14	$EBITDA \cap US\_PSR$	2.69
15	$BETA$	2.66
16	$US\_ICS \cap US\_PSR$	2.62
17	$DIV\_YIELD \cap US\_ICS$	2.51
18	$BETA \cap FSCORE$	2.49
19	$PE \cap EBITDA$	2.48
20	$DIV\_YIELD \cap EBITDA$	2.47
21	$DIV\_YIELD \cap PE$	2.41
22	PE	2.35
23	$DIV\_YIELD \cap FCF$	2.27
24	$BETA \cap PE$	2.24
25	$BETA \cap FCF$	2.20
26	GDP	2.18
27	$FSCORE \cap US\_PSR$	2.14
28	US_PSR	2.10
29	US_INFL	2.07
30	$US\_ICS \cap FSCORE$	2.02
31	$DIV\_YIELD \cap GDP$	1.97
32	$PE \cap GDP$	1.95
33	$DIV\_YIELD \cap US\_PSR$	1.92
34	$FSCORE \cap EBITDA$	1.87
35	$FSCORE \cap FCF$	1.84
36	$DIV\_YIELD \cap US\_INFL$	1.82
37	$GDP \cap US\_INFL$	1.75
38	$FSCORE \cap GDP$	1.72
30	$FCF \cap US\_PSR$	1.63
40	$EBITDA \cap GDP$	1.62
41	$FCF \cap GDP$	1.59

we considered the interaction term X_7X_1 and other interaction terms with X_7 present in model 2. Also, while determining the percentage of contribution of $DIV\_YIELD(X_1)$, we considered the interaction term X_1X_7 and other interacting terms with X_1 present in model 2. We did the same for all other indicators.

Table 5.2: Ranking of the Indicators With Respect to The Percentage of Contribution to The Response Considering the Number of Occurrence in Model (2), Individually, and Interacting with Other Indicators

Rank	Indicators	No. of Occurrence	Contr.(%)
1	$US\_ICS(X_8)$	8	25.32
2	$FCF(X_6)$	8	20.14
3	$FSCORE(X_4)$	8	19.57
4	$US\_INFL(X_{10})$	7	19.23
5	$EBITDA(X_5)$	7	18.29
6	$DIV\_YIELD(X_1)$	8	18.23
7	$US\_PSR(X_9)$	7	16.44
8	$GDP(X_7)$	8	16.41
9	$PE(X_3)$	6	14.84
10	$BETA(X_2)$	5	12.66

To assess the quality of the proposed analytical model, we use both the coefficient of determination, R^2, and adjusted R^2, which are the critical criteria to evaluate the model accuracy. The regression sum of squares (SSR) measures the variation that is explained by our proposed model. The sum of squared errors (SSE), also termed as the residual sum of squares, is the variation that remains unexplained. We always try to minimize this error in a model The total sum of squares is defined as (SST) = SSE + SSR. R^2, the coefficient of determination is defined as the proportion of the total response variation that is explained by the proposed model, and it measures how well the regression process approximates the real data points. Thus, R^2 is given by

$$R^2 = 1 - \frac{SSE}{SST} \quad .$$

However, R^2 itself does not consider the number of variables in the model and also, there is the problem of the ever-increasing R^2. To address these issues, we use the adjusted R^2, which considers the number of parameters and is given by

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \quad ,$$

where n is the number of points in our data sample, k is the number of independent regressors, i.e., the number of indicators in the model, excluding the constant term. For our final statistical model, the R squared is **96.74%**, and R squared adjusted **96.03%**. Both R squared and adjusted R squared is very high and very close to each other in our model. That is, the developed statistical model explains approximately **97%** of the variation in the response variable, a very high-quality statistical model. Similarly,the indicators that we have included in the model, along with the relevant interactions, estimates approximately **97%** of the total variation in the response variable WCP. The Residual Standard Error (RSE) represents the approximate difference between the observed and predicted outcomes in the proposed model. We obtained an RSE of .21, which implies that the observed response value differs from the predicted response value by .21 on the average. In Table 1, we rank the individual risk factors and interactions(a total of forty-one terms excluding the intersection from equation (3)) with respect to their percentage of contribution to the estimated response WCP by our proposed non-linear analytical model.

5.2.3 Residual Analysis

Once the statistical model has been developed, it is important to check the model assumptions by performing residual analysis. In our case, we have proposed a multiple non-linear statistical model, which is very useful and accurately conveys some important information on the subject matter.

5.2.3.1 Mean Residual

The residual error of the proposed model, that is,

$$\hat{\epsilon} = \text{residual} = \text{observed value-predicted value} = y - \hat{y} \quad ,$$

where y and $\hat{y}$ are the observed and predicted response. $\hat{e}$ is the estimated residual error from the linear fit. The sum of the residuals equals zero, assuming that the regression function is actually the "best fit." In our case, the mean residual is $3.8 * 10^{-18}$, implying that it is almost zero as required and attests to the quality of the developed model.

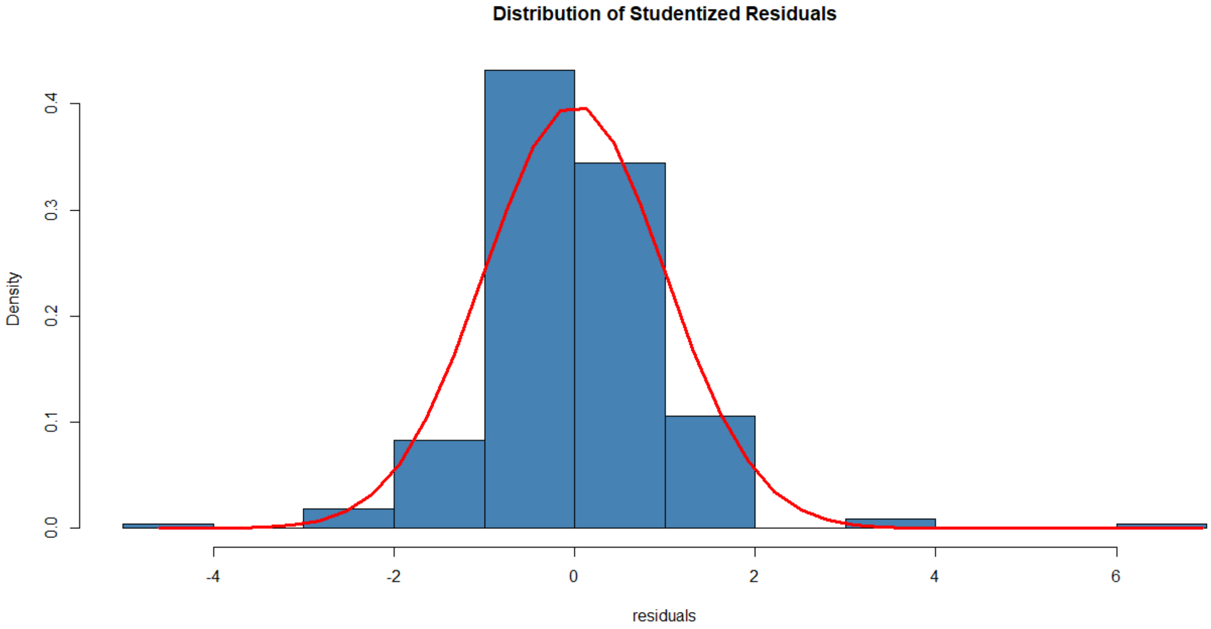


Figure 5.3: Normality of Studentized Residual Plot

5.2.3.2 Normality of the Residuals

One important assumption of our developed model is that the residuals follow the Gaussian probability distribution. From Figure 5.3, we see that the studentized residuals follows a normal pattern.

5.3 Validation of The Proposed Model

We developed our analytical model on 80% training data and validated the model based on 20% testing data. In the testing data(validation data), the test error is the average error that occurs from using the analytical method to predict the response on a new observation. That is a measurement that was not used in training the method. The test error addresses the consistency and accuracy of the analytical model.

Moreover, we performed repeated ten-fold repeated cross-validation(10 times) for our validation testing. The primary objective is that we will use 10-fold cross-validation, then we repeated cross-validation ten times, where each of the repetition folds are split differently. In 10-fold cross-validation, the training set is divided into ten equal subsets. One of the subsets is taken as the testing set in turn, and (10-1) = 9 subsets are taken as a training set in the proposed model. The mean square error E_1 is computed for the held out set.

This procedure is repeated 10 times; each time, a different group of observations is treated as a validation set. This process results in 10 estimates of the test error, $E_i, \quad i = 1, \ldots 10$. The average error of each set throughout the cross-validation process is termed as a cross-validated error. The Figure 5.4 below, illustrates briefly the idea of 10 fold repeated cross-validation, where $E_i, \quad i = 1, \ldots 10$ is the mean square error (MSE) in each iteration and ACVE is the average cross-validated error.

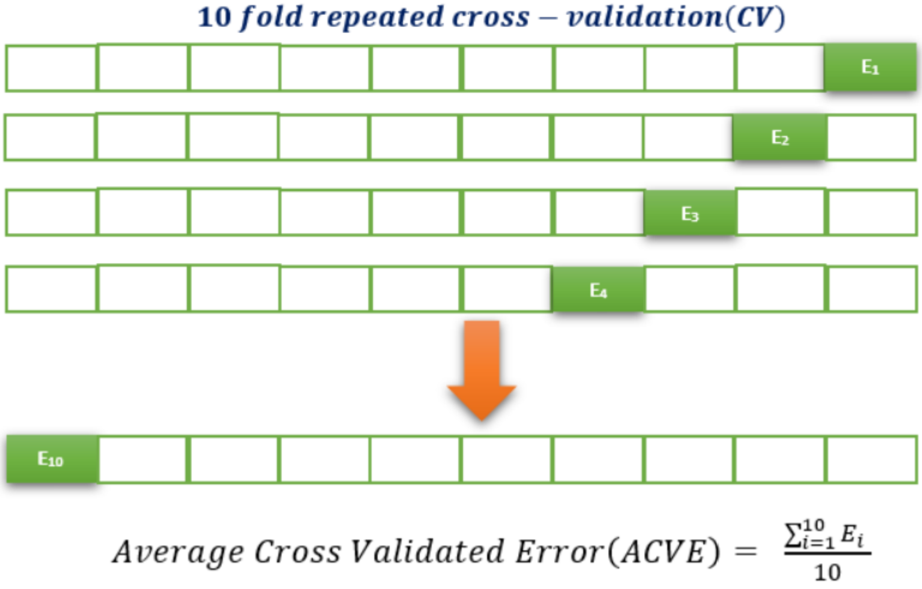

10 fold repeated cross – validation(CV)
E₁
E₂
E₃
E₄
E₁₀
Average Cross Validated Error(ACVE) = $\frac{\sum_{i=1}^{10} E_i}{10}$

Figure 5.4: Brief Illustration Of Repeated Ten Fold Cross Validation

In the validation stage, a high R^2 and low $RMSE$ attests to the good quality of a model. Also, it is expected that the cross-validated error ($RMSE$) and the accuracy (R^2) remains consistent throughout different repeated folds. The following Figure 5, illustrates how the R^2 and $RMSE$ varies in the different folds of the test data.

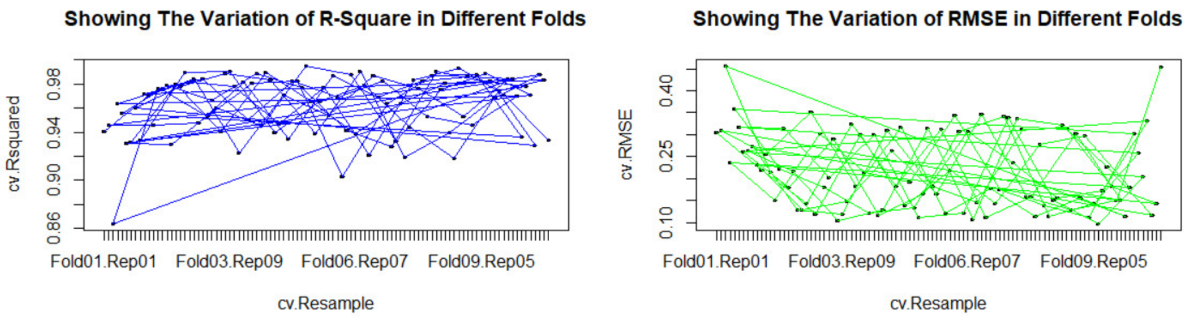


Figure 5.5: Variation of R^2 and $RMSE$ in Different Folds

The above Figure 5.5, illustrates that our R^2 is high and $RMSE$ remains low for different repeated cross-validated folds as expected. Hence, we can conclude that the accuracy of the proposed model is very consistent.

Thus, we have developed a non-linear analytical predictive model for all the fifty-nine healthcare stocks with a high degree of accuracy, which is a function of the combination of the financial and economic indicators along with the significant interactions of indicators that drive the ups and downs of the health segment of S&P 500. In the next section, we will present some methods which will optimize(maximize) our model response(objective function). We will also find the optimum values of all contributing indicators that lead to the optimization of WCP of all the healthcare segment.

5.4 Analytical Method to Optimize the WCP of the Health Business Segment

Once we have developed a high-quality model that identifies the financial and economic indicators and their interactions that predicts the WCP, we proceed to determine the values of the indicators that will maximize the response, WCP. The analytical process is discussed below.

5.4.1 Analytical Approach Using the Desirability Function

We shall use the process of the desirability function for the optimization of the response, WCP, of our proposed model. The desirability functions approach, initially was proposed by Harrington (1980)[60], and have been introduced in the literature with respect to Response Surface Methodology (RSM). The desirability function transforms each of the estimated response $Y_i(x)$ to a desirability value $d_i(Y_i)$, where $0 \leq d_i \leq 1$. For an individual response $Y_i(x)$, a desirability function $d_i(Y_i)$ takes on values within [0,1]. $d_i(Y_i) = 0$, represents entirely an undesirable response Y_i and $d_i(Y_i) = 1$, represents a completely desirable or ideal response. The value of $d_i(Y_i)$ increases as the "desirability" of the corresponding response increases. The individual desirabilities are then merged together using the geometric mean, which gives the overall desirability function, that is,

$$D = [\prod_{i=1}^{k} d_i(Y_i)]^{\frac{1}{k}} \quad ,$$

where k denotes the number of responses. In our model, $k = 1$, the WCP.

Depending on whether a particular response Y_i is to be maximized, minimized, or assigned a target value, different desirability functions $d_i(Y_i)$ can be used. A useful class of desirability functions was proposed by Derringer and Suich,[42]. Let L_i, U_i and T_i be the lower, upper, and target values, respectively, that are desired for response Y_i, with $L_i \leq T_i \leq U_i$.

If there is a specific target set up for the response, then its desirability function is given by,

$$d_i(\hat{Y_i}) = \begin{cases} 0 & , \text{ if } \hat{Y_i}(x) < L_i \\ \left(\frac{\hat{Y_i}(x)-L_i}{T_i-L_i}\right)^s & , \text{ if } L_i \leq \hat{Y_i}(x) \leq T_i \\ \left(\frac{\hat{Y_i}(x)-U_i}{T_i-U_i}\right)^t & , \text{ if } T_i \leq \hat{Y_i}(x) \leq U_i \\ 1 & , \text{ if } \hat{Y_i}(x) > U_i \quad , \end{cases} \tag{5.4}$$

where s and t in the above equation determines how important it is to hit the target. For $t = s = 1$,the desirability function increases linearly towards the direction of T_i. For $s < 1, t < 1$, the desirability function is convex, and for $s > 1, t > 1$, the desirability function is concave[42]. Our objective is to maximize the response, WCP; Thus, the individual desirability function will be,

$$d_i(\hat{Y_i}) = \begin{cases} 0 & , \text{ if } \hat{Y_i}(x) < L_i \\ \left(\frac{\hat{Y_i}(x)-L_i}{T_i-L_i}\right)^s & , \text{ if } L_i \leq \hat{Y_i}(x) \leq T_i \\ 1 & , \text{ if } \hat{Y_i}(x) > T_i \quad , \end{cases} \tag{5.5}$$

where T_i and L_i are chosen by the investor. We propose the following five step algorithm to optimize the response, WCP based on the desirability function method:

1. Develop the statistical model that very accurately predicts the response, WCP, driven by a set of significant indicators.
2. Obtain the constraints on input indicators, for $a < Y_i < b$ and $c < X_i < d$; Y being the response and x being the indicators.
3. Define the desirability function(s) $d_i(Y_i)$ for the response(s) based on the optimization objective.
4. Obtain the optimal values of the response by maximizing the desirability function with respect to the controllable input indicators.
5. Validate the optimization process based on the coefficient of variation R^2 and the ${R^2}_{Adjusted}$.

5.4.2 Numerical Results

Our data-driven non-linear analytical model is a function of *six* financial indicators and *four* economic indicators along with 31 interactions of the indicators. After developing the predictive model with high accuracy, our goal is to maximize the response WCP and find the optimum values of the indicators at which the response is being maximized. We proceed to maximize WCP (within its domain) in the model (5.3), Section 5.2.2. The analytical method of optimization requires the constraints of optimization, as defined in Table 5.3 for the ten indicators. These constraints are the lower and upper boundaries of each of the ten individuals that are used in the modeling. In our optimization technique, we want to make sure that the optimized value of the response WCP falls within its domain, given the specific values of the indicators. All the numerical computations of the analytical optimization have been done using Minitab 19.2.0 software.

Table 5.3: Constraints On The Indicators Showing the Lower and Upper Limits

Indicators	Constraints
$Div\_Yield(X_1)$	$1.02 < X_1 < 1.45$
$Beta(X_2)$	$.84 < X_2 < 1.28$
$PE(X_3)$	$36.57 < X_3 < 90.5$
$FSCORE(X_4)$	$3.55 < X_4 < 6.67$
$EBITDA(X_5)$	$3.88 < X_5 < 4.24$
$FCF(X_6)$	$.55 < X_6 < 1.11$
$US\_GDP(X_7)$	$19.8 < X_7 < 21.15$
$US\_ICS(X_8)$	$91.2 < X_8 < 101.4$
$US\_PSR(X_9)$	$6.7 < X_9 < 8.3$
$US\_INFL(X_{10})$	$1.55 < X_{10} < 2.95$

Table 5.4: Estimated Maximized Response with Optimum Values of the Indicators

Response & Indicators	**Optimum Values**
$WCP(Estimated)$	$155
Div_Yield	1.24
$Beta$	1.06
PE	63.53
$FSCORE$	5.11
$EBITDA$	4.06
FCF	.832
US_GDP	20.5
US_ICS	96.3
US_PSR	7.5
US_INFL	2.61

Table 5.3, above, provides the lower and upper limits of all ten indicators used in our study. Next, using equation (5) of Section 4.1, we will maximize the estimated response from the model (2) and obtain the optimum values of all ten indicators. The following Table 5.4 provides the estimated maximum response WCP along with the optimum values of the indicators.

Thus, with these values of economic and financial indicators, we are at least 95% certain that the response WCP will be maximum. Furthermore, we can tract the numerical behavior of the indicator to determine the direction of WCP. The following Table 5.5 provides the values of R^2, $R^2{}_{Adjusted}$ and desirability value along with the 95% confidence and 95% prediction interval of the estimated response, WCP.

Table 5.5: Some Useful Results Related to The Optimized Response

Estimated Maximized Value	\$155
Desirability	1
R^2	98.84%
$R^2{}_{Adjusted}$	97.85%
95% CI	(139.57, 170.43)
95% PI	(139.06, 170.94)

Thus, with 98% accuracy we have estimated the optimum values of the individual indicators for which the estimated response in \$155. Also, we are at least 95% certain that the maximum WCP is in between \$140 and \$170. Thus, this analytical process in estimating the WCP with very high degree of accuracy is very important to the investors to develop desired strategies by monitoring the behaviors of the financial and economic indicators of Health Segment of the S&P 500.

5.4.3 Graphical Visualization of the Optimization Results

One important aspect of the optimization results is to assist investors to obtain three-dimensional views of the directional behaviors of the identified indicators as they affect the response, WCP. We obtained response surface plots (contour and surface plots) that are very useful in order to obtain desired values of the response and optimum conditions for any two indicators keeping the others fixed at the desired level. In a contour plot, the response surface is observed as a two-dimensional plane where all the points that have a similar response are connected to create contour lines of constant responses. A surface plot generally exhibits a three-dimensional view that may provide a clearer picture of the response's behavior (WCP). Since, the *four* economic indicators are not **controllable**, we will not include those plots and only focus on the combination of *six* financial indicators included in model (5.3) of Section 5.2.2. In this section, we will illustrate different contour and surface plots that will help investors to understand the nature of the relationship between any of the two indicators and the response (WCP). The following six figures (Figure 5.6-Figure 5.11) describe the variation of the estimated response WCP as any single or two indicators varies, keeping the others fixed at a particular level. The usefulness of these visual representations can be interpreted as follows:

Maximizing stock prices and maximizing corporate profit are the essential goals for any company. Both are needed for a company to flourish, and both reflect the overall health and future prosperity of the company. The objective of health companies is to maximize the price of their stocks to comply with their share holders' wishes. The stock price is the discounted sum of all future cash flows. Thus, it reflects all consequences of any decision a company takes at present. Even if it is a current measure, it also reflects the future. So, stock price maximization is vital for shareholders' wealth. Any financial investor willing to invest in the health care segment of S&P 500 may use the following visual representations to select the stocks based on the interacting behavior of any two financial indicators keeping the other fixed at the desired level. Since maximizing stock price account for the maximization of shareholders' wealth, financial managers of a particular health care company of S&P 500 may be interested in looking at the specific ranges of the indicators at which the response, WCP, is maximized. The plots also provide the numerical ranges of the indicators within which the response WCP has increasing/decreasing behavior. These pieces of information are vital to the managers and financial analysts of the companies to make strategic decisions regarding the overall financial health and long term viability of the company.

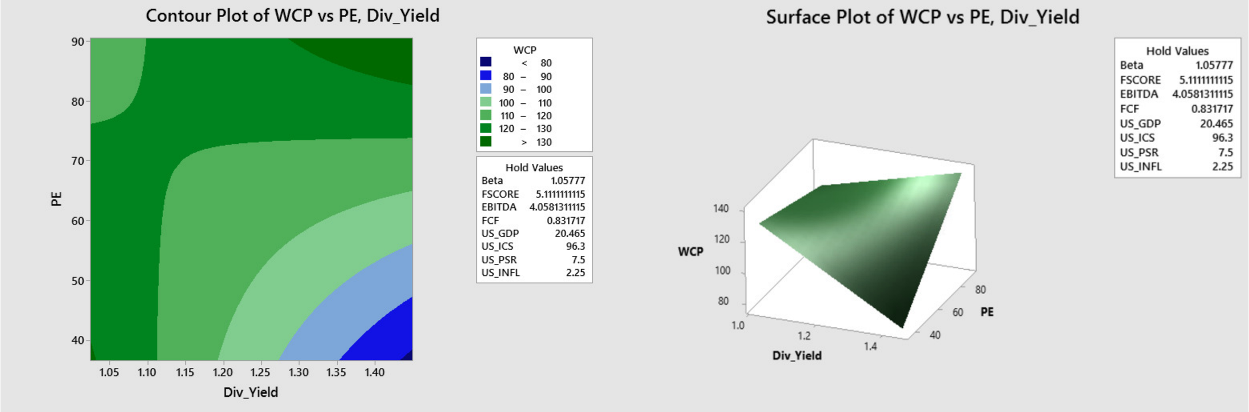


Figure 5.6: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as Div_Yield and PE Varies Keeping Other Indicators Fixed at a Specific Level

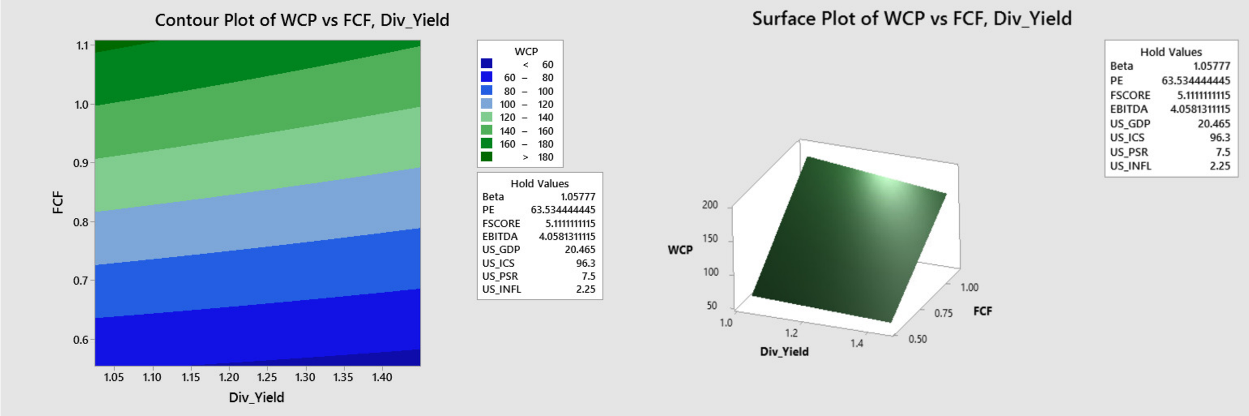


Figure 5.7: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as Div_Yield and FCF Varies Keeping Other Indicators Fixed at a Specific Level

From the Figure 5.6, we see that the estimated response, WCP, is maximized when Div_Yield is more than approximately 1.3, and PE is more than approximately 82, keeping all other indicators fixed at a desired level. Also, as we keep on decreasing the Div_Yield up to 1.1 and keep increasing the PE up to 70, WCP keeps on increasing. This finding may be explained by the fact that with the increase in price-to-earnings (PE) ratio, the price-to-dividends ratio rises as well, thus lowering the dividend yield.

Figure 5.7 above describes how the response WCP changes with the variation of Div_Yield and FCF. The response WCP is maximized (light green region, $120-$160) where FCF remains approximately within the interval [.9,1] throughout the range of Div_Yield. The WCP has an increasing pattern with the increase of FCF. Any financial investor willing to invest in the health care segment of S&P 500 may use the above visual representation to select the stocks whose FCF falls within the specified range.

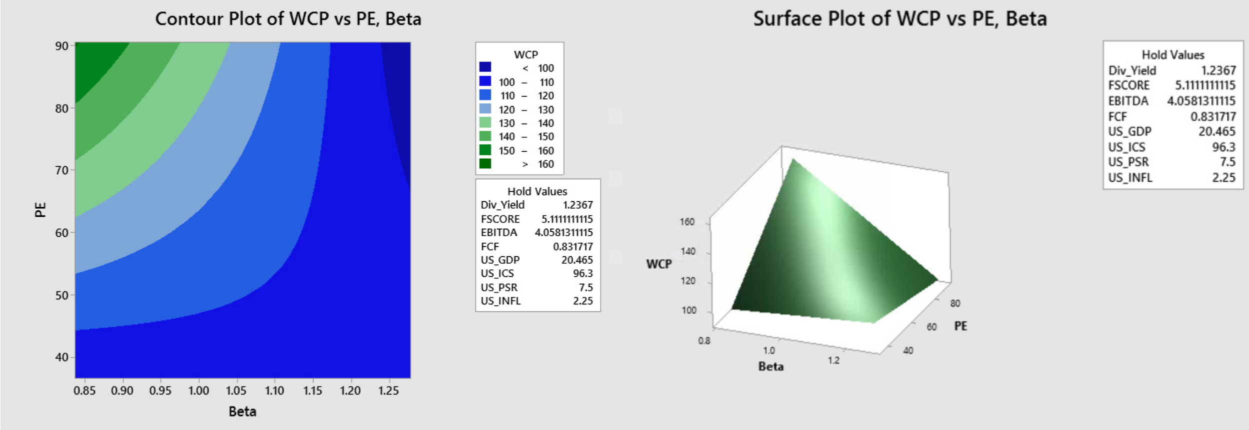


Figure 5.8: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as Beta and PE Varies Keeping Other Indicators Fixed at a Specific Level

Figure 5.8 above describes how the response WCP changes with the variation of Beta and PE. The response WCP is maximized ($140-$160) where Beta remains approximately less than .95 and PE remains approximately more than 71. There is an increasing pattern in WCP with increasing PE ratio and decreasing Beta risk. Hence, we can infer that the response is maximized when the Beta risk is low and the PE ratio is high.

From the following Figure 5.9, we see that the estimated response WCP is maximized in the region where Beta lies approximately below .91 and FSCORE lies approximately 5.5 and below. Also, WCP has an increasing pattern as we keep on decreasing Beta gradually.

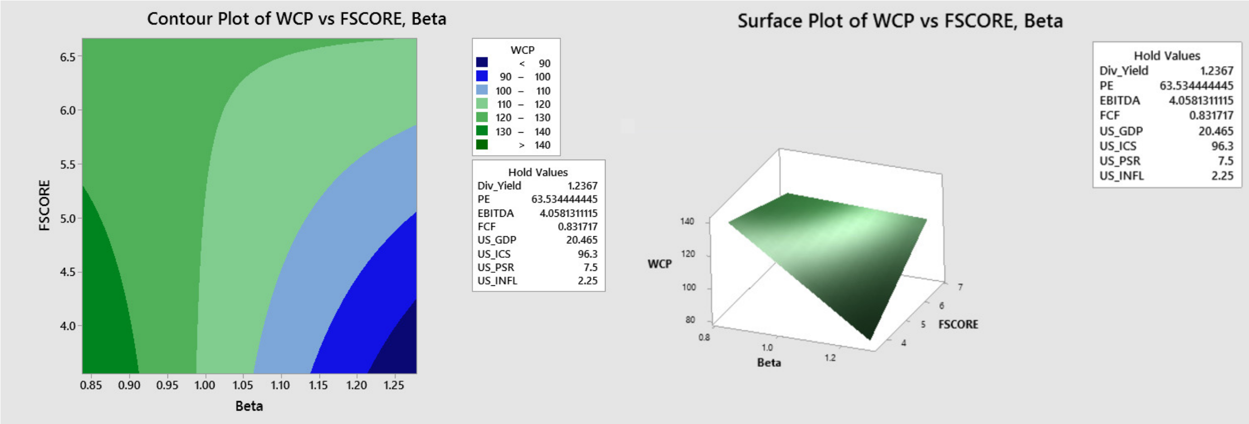


Figure 5.9: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as Beta and FSCORE Varies Keeping Other Indicators Fixed at a Specific Level

From the following Figure 5.10, we see that the estimated response WCP is maximized in the region where Beta lies approximately below 1.17 and FCF lies approximately within the interval [.8,1]. WCP keeps on increasing with the increase of FCF, and it gets maximized Beta risk decreases.

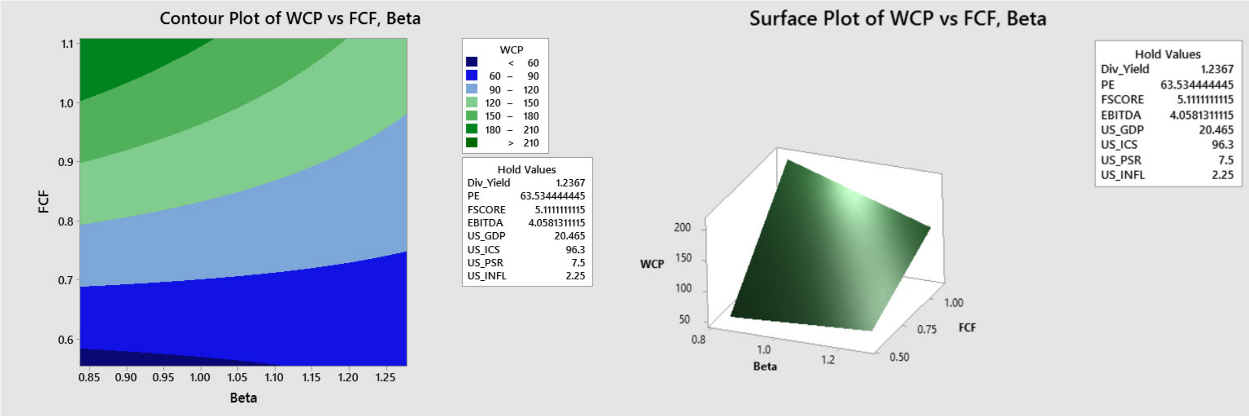


Figure 5.10: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as Beta and FCF Varies Keeping Other Indicators Fixed at a Specific Level

Figure 5.11 below describes how the response WCP changes with the variation of FSCORE and FCF, keeping other indicators at the desired level. The response WCP is maximized, where FCF remains approximately within the interval [.95,1.05] throughout the range of FSCORE. WCP attains its minimum value in the region where both FSCORE and FCF are low (deep blue). Gradually it increases with the increase in both the indicators as desired.

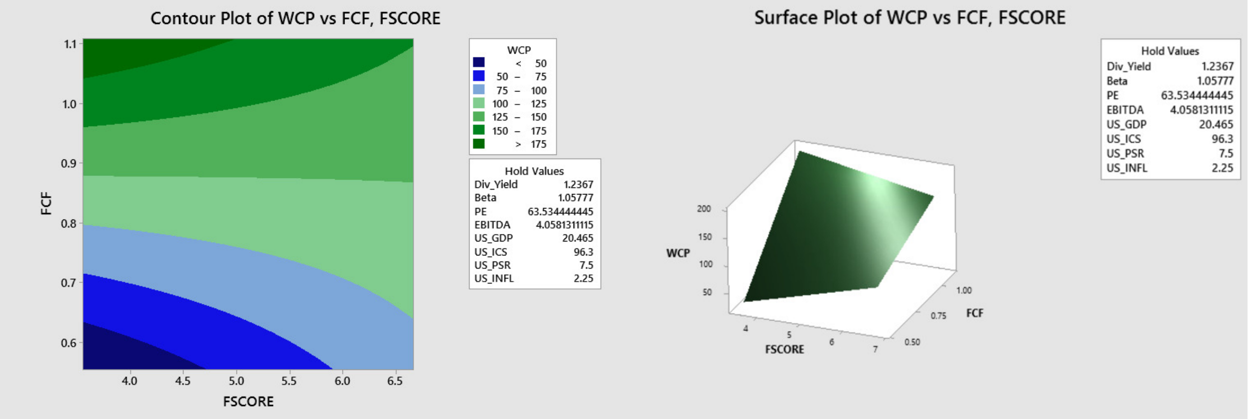


Figure 5.11: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as FSCORE and FCF Varies Keeping Other Indicators Fixed at a Specific Level

5.5 Discussion

In the present study, we have developed a non-linear analytic model that identifies the most significant indicators and the associated interactions responsible for the ups and downs of the 59 healthcare stocks very accurately. After obtaining the significant indicators, along with their significant interactions, we rank them with respect to the percent of contribution to the WCP, as shown in Table 5.1. The highest contributing indicator is the combination of the indicators FCF(X_6) and US_ICS(X_8), contributing 4.53% of the total variation to the response, WCP. The next most significant contribution is also an interaction term that is the combined effect of FSCORE(X_4) and the US_INFL(X_{10}) with a contribution of 4.15% to the response, WCP. Numbers 3, 4, and 5 are respectively the combined interaction effect of EBITDA (X_5) and US_ICS(X_8), interaction between GDP(X_7) and US_ICS(X_8), and interaction between PE(X_3), US_ICS(X_8) with the contribution of 3.89%, 3.63%, and 3.41%, respectively. Hence, summing these indicators up, we identify that they explain more than 96% of the total variability to the response, WCP. Furthermore, utilizing an optimized analytical process to determine the actual values of six financial and four economic indicators in our proposed model will maximize the response, WCP. The R^2 and $R^2{}_{Adjusted}$ from Table 5.5 attests the fact that the optimized model results are very accurate. The desirability value from Table 5.5 indicates that the estimated fit is most desirable/ideal. The following Table 5.6 demonstrates the list of the observed and predicted responses from our data-driven non-linear analytical model. It can be seen clearly that the predictions are very close to the actual observed values and thus, attests to our model's high accuracy and predictive power.

Table 5.6: The List of Observed and Predicted Values of The Response WCP

Observations	**Observed**	**Predicted**
1	155	156
2	152	150
5	151	149
6	148	148
13	141	145
19	144	145
20	143	146
28	139	144
29	138	144
36	149	145
37	150	144
38	149	145
39	152	148
40	153	148
41	153	155
63	136	141
64	135	137
72	153	148
73	154	148
81	143	147
82	142	147
83	140	141
127	131	131
148	123	126
212	111	111
213	111	112
252	104	104
255	104	107
256	105	104
272	98	100

Finally, we have developed an analytical predictive model consisting of ten individual and thirty-one interacting financial and economic indicators that drive the behavior of WCP of the health segment that consists of fifty-nine companies. The proposed model can be used by financial portfolio managers, investors, etc. to select stocks from the health segment of S&P 500 in accordance with the wishes of their clients. In what follows, we will be more specific on the overall findings and usefulness of our innovation.

5.6 Conclusion

We have developed a real data-driven analytical model that very accurately predicts the WCP and identifies some very useful findings of what drives the weekly stock prices of the healthcare segment of the S&P 500. We summarize the important in formations and usefulness that the proposed model offers with a high degree of accuracy.

1. We have identified the individual financial and economic indicators that significantly contribute to the price behavior of the health segment of S&P 500, which consists of fifty-nine stocks.

2. We have identified the significant interactions of the financial, economic, and other indicators that contribute to the WCP of the fifty-nine health stocks.

3. We have ranked the ten individual and thirty-one interactions of the indicators with respect to their percentage of contribution to the WCP of the health stocks of S&P 500.

4. We have developed a non-linear analytical model consisting of ten individual and thirty-one interacting indicators that predict the WCP of the health segment with a 97% accuracy.

5. We compared the original and the estimated response WCP using our analytical model and found that they are very close to each other, indicating the high accuracy of our model.

6. We utilized an analytical optimization process to determine the optimal values of the indicators that will maximize the WCP of the health segment. These values were determined with at least 95% accuracy.

7. We have developed two and three-dimensional contours and surface plots, based on the behavior of the values of the indicators that maximize the WCP of the health stocks. These plots can be used strategically to monitor the behavior of WCP as the significant values of the indicators change.

The above information is important to individual investors, portfolio managers, financial institutions that invest in the health stock of S&P 500. Individual health companies can utilize the usefulness of the proposed model for their strategic planning, their competitive standing among the health segment companies, monitoring and predicting their financial status, among other uses. The proposed analytical model is very accurate, 96.74% in estimating the various findings as stated above that drive the WCP of the 59 health companies. Finally, the derived usefulness of the proposed model is essential for constructive and accurate decision making concerning the financial and economic aspects of the health industry.

Chapter 6: A stock optimization problem in finance: Understanding the financial and economic indicators through analytical predictive modeling

6.1 Introduction

The healthcare sector incorporates businesses that supply medical services, manufacture medical equipment or drugs, provide medical insurance, and facilitate healthcare provision to patients. It is one of the immense sectors and contributes significantly to the U.S. economy, accounting for approximately a fifth of overall gross domestic product(GDP). Since healthcare stock change has an enormous impact on the global economy affecting its overall GDP and other financial factors, we tried to build an analytical prediction model to predict the yearly percentage change of the stocks. The non-linear analytical model we developed includes **five** essential findings. The proposed model's response is the Average Weekly Closing Price (AWCP) of healthcare stock AbbVie Inc.(ABBV) starting from August 1st, 2017, to December 31st, 2019. In addition to predicting the Weekly Stock Price, our model identifies the individual indicators and their corresponding interactions that significantly contribute to the response. We rank these indicators in accordance with their percentage of contributions to the response. We then performed a response surface analysis to find the appropriate values of the indicators that optimize(maximize) the response. Also, we have monitored the optimal ranges of any two indicators that affects the response AWCP with visual illustration. Finally, we compared the original and the predicted responses of AWCP using our analytical model, and found the two set of observations very close to each other testifying the high accuracy of our model. The proposed model offers other useful information on the subject area. Our analytical model has been validated and tested to be of high quality, and our prediction of the weekly stock price is with a high degree of accuracy. The stock price oscillates over time by showing some dramatic ups and downs. Some investors prefer to monitor these changes closely to stay on top of their investments. But even if one doesn't watch the stocks daily, monitoring the net change percentage over time is essential to maintaining a successful portfolio.

While building the statistical model, the response variable is the percentage weekly closing price of ABBV stock; hence, we develop an analytical model containing significant contributable variables and other significant interactions. The proposed statistical model depends on several assumptions, such as linearity, multicollinearity, homoscedasticity, and different assumptions concerning statistical methodology. Our dataset shows that none of the risk factors are highly correlated except for US GDP and dividend yield, as shown in Figure 6.4. This is good for building our model. Our proposed statistical model is useful in predicting individual stock change, given significant risk factors. Also, we ranked the risk factors according to their percentage of contribution to the response. The validation and quality of our proposed analytical model have been statistically evaluated using R square (R^2), R square adjusted (${R^2}_{adj}$), Mean Absolute Error (MAE) and root mean square error (RMSE). Eventually, its usefulness has been illustrated by utilizing different combinations of various risk factors. From 59 healthcare stocks, we selected the stock ABBV based on specific criteria, which will be described in the next section. Best of our knowledge, no such statistical model has been developed under the proposed logical structure to predict the yearly percentage change in healthcare stocks. Therefore, searching for an appropriate statistical model in the prediction of the weekly stock price is imperative.

6.2 Methodology

6.2.1 Selection Of Appropriate Stock

The data has been obtained from Yahoo finance and some other financial and reliable websites. Then, the whole data set has been combined with our analysis. We collected information for the healthcare sector(XLV) of the S&P 500 stock. There were 59 pieces of information related to the top 59 healthcare stocks. One of our study's main goals is to select one stock from the list of 59 healthcare stocks based on a certain meaningful criterion. Initially, our data contained yearly information. Our model's risk factors were based on trailing 12 months (TTM) average starting from December 31^{st}, 2018, to December 31^{st}, 2019. To select the appropriate stock, we performed the K-means clustering algorithm. We performed the clustering based on the following three steps.

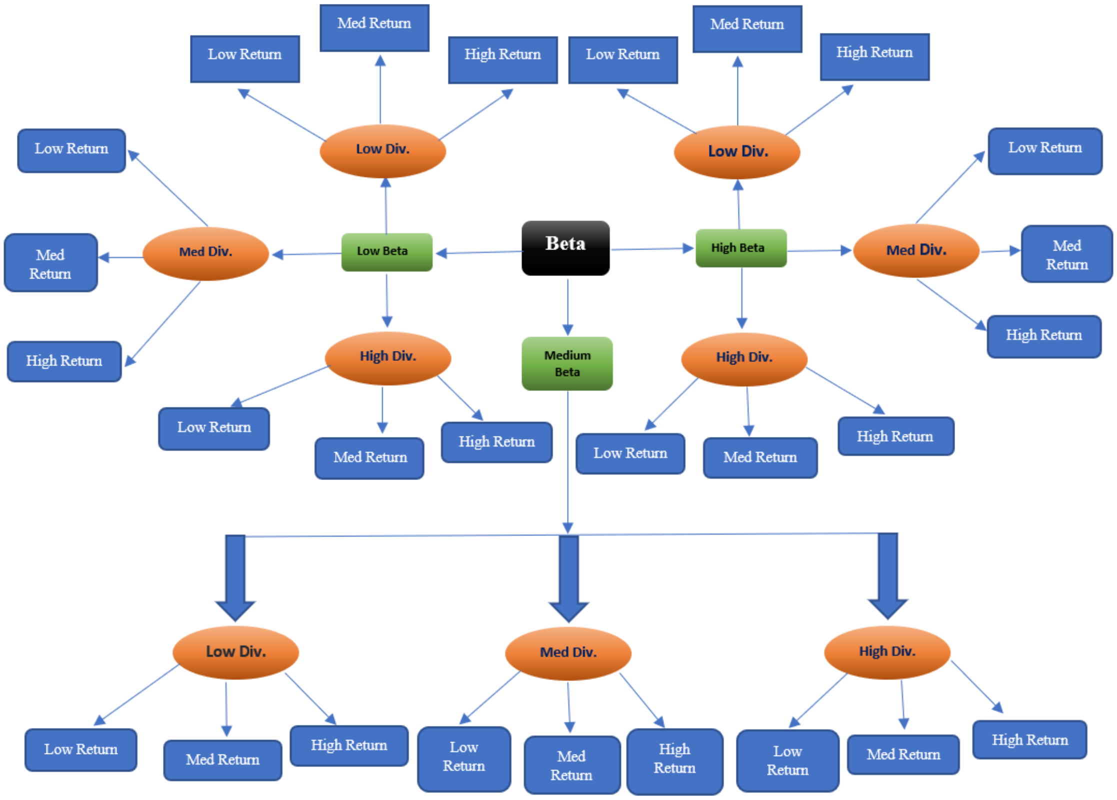


Figure 6.1: Schematic Diagram of Stock Clustering Mechanism

1. Cluster the stocks in three groups(low, medium and high) based on the risk factor beta.

2. After we got the three clusters (high beta, medium beta and low beta), each cluster is further grouped into three categories (low, medium and high) based on the risk factor dividend yield.

3. In the final stage of clustering, we again clustered each group of dividend yield clusters based on the yearly percentage return of stocks.

The following Figure 6.1, demonstrates the schematic diagram of the clustering mechanism.

The above clustering mechanism produced a total of twenty-seven possible clusters to choose from. To select the appropriate stock meaningfully, we focused on the specific cluster comprised of the stocks having low beta risk, high dividend yield, and high yearly percentage return. The following Figure 6.2 shows the stocks having such characteristics.

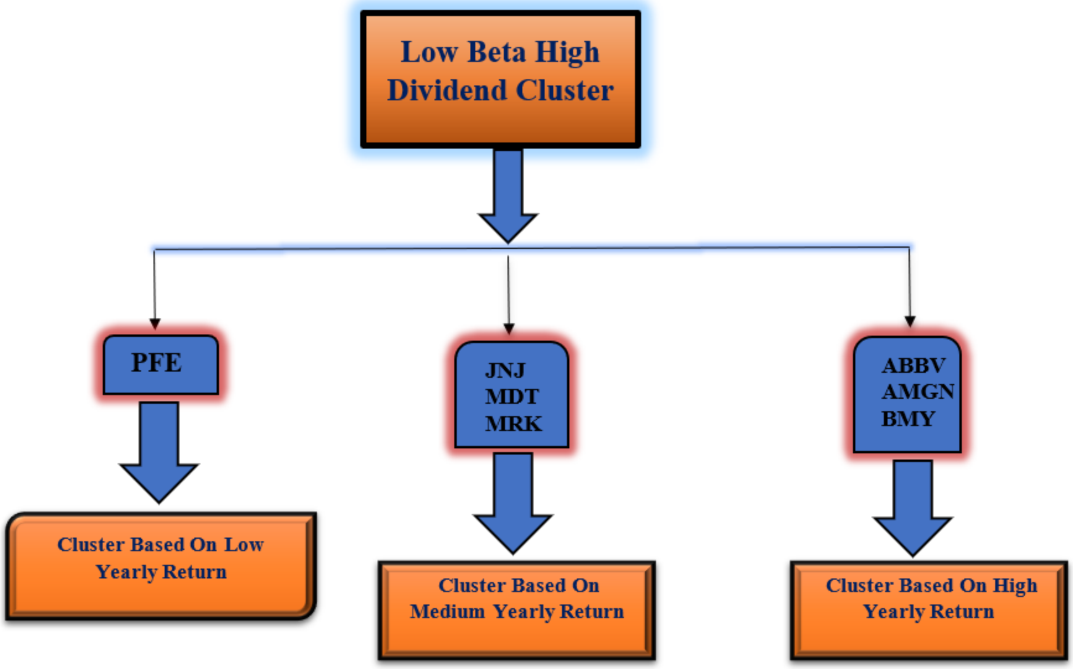

Low Beta High Dividend Cluster
PFE
JNJ
MDT
MRK
ABBV
AMGN
BMY
Cluster Based On Low Yearly Return
Cluster Based On Medium Yearly Return
Cluster Based On High Yearly Return

Figure 6.2: Stocks With Low Beta risk, High Dividend Yield and High Yearly Percentage Return

The above Figure tells us, there are three stocks namely ABBV, AMGN and BMY that matches our selection criterion. Since, our goal is to build an analytical model for one stock we have chosen ABBV, which has the highest dividend yield (5.4%) among the three. Also, AbbVie Inc.(ABBV) is one of the very popular American publicly traded bio-pharmaceutical company. While the company's total revenue for 2019 grew by only 1.6%, U.S. sales of its blockbuster drug, *Humira* (best selling drug in the world for several years), were up by 8.6%, and worldwide sales of another popular drug, *Imbruvica*, were up more than 30%.

6.2.2 The Data and Description of The Indicators

After we select ABBV as the appropriate stock, we collect information on different indicators for the stock. Our data includes the information from August 1st, 2017 to December 31st, 2019. We have collected data based on three financial attributes and three economical attributes. A five day period moving average (MA) method was used for each indicators to structure our data. One of the main goals of our study is to understand what indicators significantly affect the weekly stock price of ABBV. We have six indicators and the **AWCP**(Average Weekly closing price) as a measure of response.
The description of the attributable variables (indicators) that the data data was collected on are given below.

6.2.2.1 Financial indicators

1. **Div_Yield**(X_1):The dividend yield is a financial measure that demonstrates how much a company disburses in dividends each year with respect to its stock price. It is the annual dividend rate divided by the current share price. It is expressed as a percent form. For instance, if the current stock price is $50 and the annual dividend is $1, the dividend yield is 2 percent.

2. **Beta**(X_2): Beta is a risk measure of a stock's volatility of return with regards to the overall market. In general, a stock with higher beta value tends to have a higher risk and also higher expected returns. It is defined as follows:

$$Beta = \tfrac{Cov(R_I,R_M)}{Var(R_M)} \quad ,$$

where R_I is return on an individual stock and R_M is the return on the overall market. Cov(.,.) is the covariance between R_I and R_M, i.e, how the changes in stock return are related to the changes in the market return, Var(.) is the Variance measure implying how how far the market data is scattered from their average market return.

3. **PE**(X_3):The price-to-earnings ratio (P/E ratio) is the ratio that measures the current share price of a stock with respect to its earning per share(EPS). It is defined as follows:

$$P/ERatio = \frac{\text{Market value per share}}{\text{Earning per share}} \quad .$$

6.2.2.2 Economical indicators

4. **US_GDP**(X_4): Gross domestic product of the United States (in trillon).

5. **US_ICS**(X_5): The Index of Consumer Sentiment (ICS) or economic well-being—was developed at the University of Michigan Survey Research Center to measure the confidence or optimism (pessimism) of consumers in their future well-being and coming economic conditions. The index measures short and long-term expectations of business conditions and the individual's perceived economic well-being. Evidence indicates that the ICS is a leading indicator of economic activity as consumer confidence seems to pave the way for major spending decisions.

6. **US_PSR**(X_6): The U.S. personal saving rate is personal saving as a percentage of disposable personal income. In other words, it's the percentage of people's incomes left after they pay taxes and spend money. The U.S. Bureau of Economic Analysis (BEA) publish this rate.

In developing the proposed statistical model for stock price as a function of the attributable variables, one of the main assumptions is that the response variable Change should follow the Gaussian probability distribution. From the following Figure 6.3, we see that the values of the response AWCP is positively skewed and does not entirely follow a Gaussian Probability distribution.

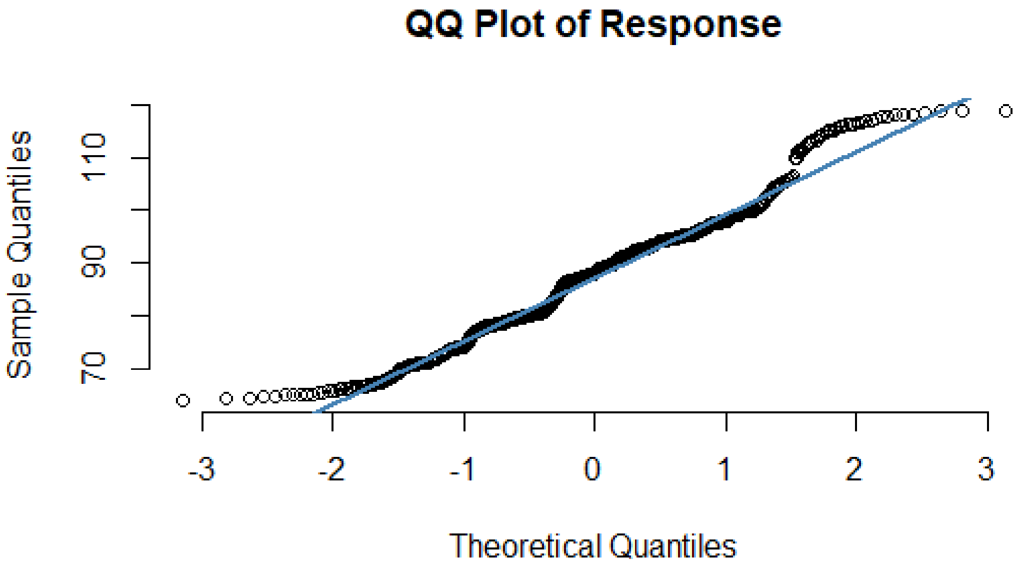
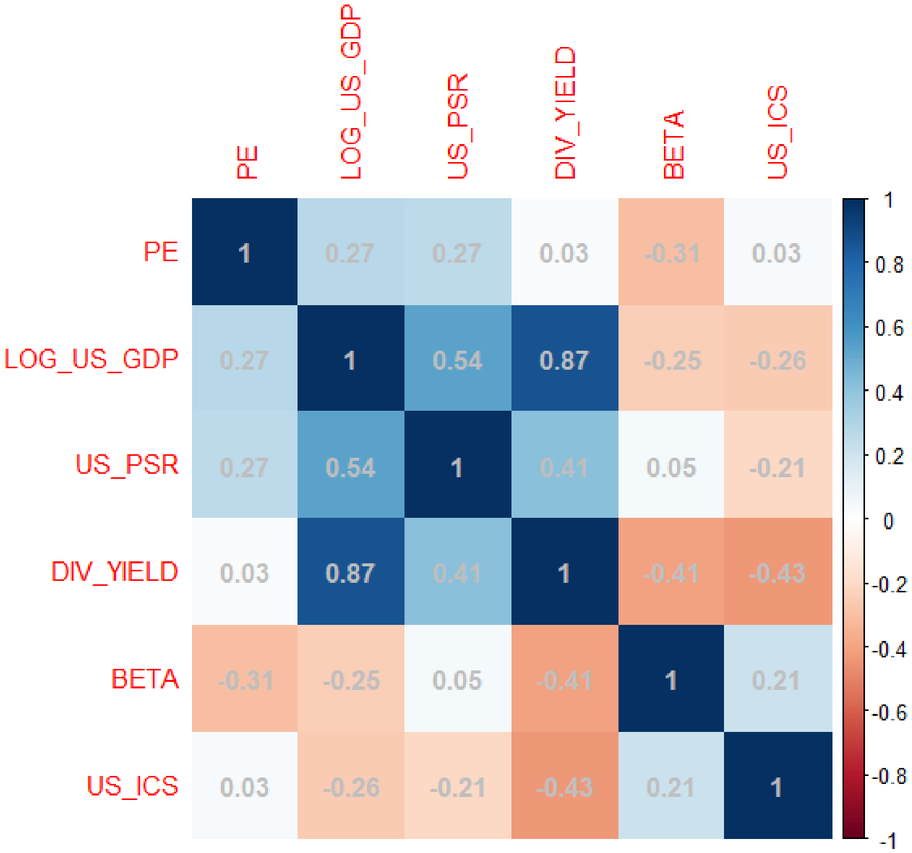


Figure 6.3: Q-Q Plot Of The Response WCP

We have also shown through goodness-of-fit testing (Shapiro-Wilk normality test, A p-value = 2.265×10^{-8}) that the subject data does not follow the normal probability distribution as well. Therefore, the Q-Q plot supports the fact that natural phenomena, such as the weekly average stock price not following the Gaussian probability distribution. The correlation matrix plot comprising the attributable variables is shown in Figure 6.4, where we see that no two variables are highly correlated, and the degree of linear association between any two variables is not high except for $Div\_Yield(X_1)$ and $log(US\_GDP)$. We also considered US consumer Price Index(CPI) initially in our model, but we had to drop it as it was highly correlated (almost perfect correlation with a correlation coefficient of .99) with GDP.



	PE	LOG_US_GDP	US_PSR	DIV_YIELD	BETA	US_ICS
PE	1	0.27	0.27	0.03	-0.31	0.03
LOG_US_GDP	0.27	1	0.54	0.87	-0.25	-0.26
US_PSR	0.27	0.54	1	0.41	0.05	-0.21
DIV_YIELD	0.03	0.87	0.41	1	-0.41	-0.43
BETA	-0.31	-0.25	0.05	-0.41	1	0.21
US_ICS	0.03	-0.26	-0.21	-0.43	0.21	1

Figure 6.4: Correlation Matrix of The Risk Factors

6.2.3 Development of Statistical Model

In developing a statistical model, our main goal is to express our response AWCP in terms of a non-linear mathematical function of all indicators with a high degree of accuracy. Thus, we proceed to develop the statistical model which is given by the average weekly stock price as a function of the six attributable variables (which we believe has a significant contribution to the response) and all possible interactions as previously discussed. One of the pure forms of a model with all possible interactions and additive error structure, in the present situation, could be expressed as follows:

$$AWCP = \beta_0 + \sum_i \alpha_i x_i + \sum_j \gamma_j k_j + \epsilon \quad , \tag{6.1}$$

where β_0 is the intercept of the model, α_i is the coefficient of i^{th} individual attributable variable x_i, γ_j is the coefficient of j^{th} interaction term k_j and ϵ denotes the random disturbance or residual error of the model following a normal distribution with zero mean and constant variance.

One of the main assumptions to construct the above model is that the response variable CLOSE should follow the Gaussian probability distribution. As we illustrated above, the dependent variable Change does not support the Gaussian probability distribution. Therefore, we must apply a non-linear transformation to the response to see if the transformation can adjust the scale of the response to follow a normal probability distribution. We used Johnson transformation[106] for our response which results in equation 6.2, below:

$$z = \gamma + \delta ln(x - \epsilon) \quad , \quad \delta > 0, -\infty < \gamma < \infty, -\infty < \epsilon < \infty, x > \epsilon$$

and

$$ATWCP = -120.5 + 21.87ln(x + 159.83) \quad . \tag{6.2}$$

Here, $TWCP$ represents the new response variable(transformed) after Johnson's Transformation has been applied. Thus, we proceed to estimate the coefficients (weights) of the actual contributable variables for the transformed data in equation 6.2. To develop our statistical model, we initially begin with the full statistical model, which included all six attributable variables as previously defined and five possible interactions between each pair. Thus, at first, we start structuring our model with $\binom{n}{k} = 15(n = 6, k = 2)$ potential interaction terms and six indicator terms. While we began with the full statistical model (twenty-one), as we mentioned above, we have applied the process to determine the most significant contributions of both the individual attributable variables and interactions by eliminating the less important indicators gradually. Moreover, backward elimination is deemed one of the best traditional methods for a small set of feature vectors to tackle the problem of overfitting and perform feature selection. To get better accuracy, we took the log transformation of the indicators $GDP(X_4)$ to reduce its high variability. However, our statistical analysis has shown that six out of six indicators significantly contribute to and twelve interaction terms. The method eliminated three unimportant interaction terms. Thus, the best proposed statistical model with every significant indicators and interactions that estimates accurately the response AWCP are six indicators individually that significantly contribute and twelve interaction term. Hence, the best preferred statistical model with all significantly attributable variables and interactions that estimates the weekly average stock price given by equation (6.3) below.

$$T\widehat{AWC}P = \begin{cases} -.083 - 0.12X_1 - .37X_3 + 5.3X_2 + .0029log(X_4) \\ +1.85X_5 + 6.2X_6 + .186X_1X_2 + .22X_1X_4 + .035X_1X_5 + \\ .39X_1X_6 + .0073X_2X_3 - .023X_3X_4 + .006X_3X_5 + \\ .036X_3X_6 - .22X_2X_4 - .2X_2X_5 - .1X_5X_4 - .41X_4X_6 \end{cases} \tag{6.3}$$

The TAWCP estimate is obtained from equation (6.3) and is based on the Johnson transformation of the data.

Thus, we will utilize the anti-transformation on equation (6.3) to estimate the desired, actual average weekly stock price as follows:

$$\widehat{AWCP} = -159.83 + exp\big(\frac{\widehat{TAWCP} + 120.5}{21.87}\big) \quad . \tag{6.4}$$

The proposed model will help social researchers, economists and financial analysts to understand how the weekly stock price varies when any one of the six attributable variables is varied, keeping the other attributable variables fixed.

Similarly, with the significant interactions. Most commonly, it will estimate the conditional expectation of the response of AWCP given the indicators fixed at a particular level. As for example, given, $X_1 = 4.8, X_2 = .9, X_3 = 31.5, X_4 = 22, X_5 = 99.4$ and $X\_6 = 7.7$, we got our predicted response value as 89.31(from equation (6.4)).

So, given all the values of indicators, fixed at a particular level, the weekly average stock price for ABBV is 89.31$. Furthermore, we illustrate the percentage that the indicators and the interactions contribute to the yearly percentage change in stock as we rank them.

To assess the quality of the proposed analytical model, we use both the coefficient of determination, R^2, and adjusted R^2, which are the critical criteria to evaluate the model accuracy.

The regression sum of squares (SSR), measures the variation that is explained by our proposed model. The sum of squared errors (SSE), also termed as the residual sum of squares, is the variation that remains unexplained. We always try to minimize this error in a model The total sum of squares (SST) = SSE + SSR. R^2, the coefficient of determination is defined as the proportion of the total response variation that is explained by the proposed model, and it measures how well the regression process approximates the real data points. Thus, R^2 is given by

$$R^2 = 1 - \frac{SSE}{SST} \quad .$$

However, R^2 itself does not consider the number of variables in the model. Also, there is the problem of the ever-increasing R^2.

To address these issues, we have the adjusted R^2 which considers the number of parameters and is given by

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \quad ,$$

where n is the number of points in our data sample, k is the number of independent regressors, i.e., the number of indicators in the model, excluding the constant. For our final statistical model, the R squared is **98.89%**, and R squared adjusted **98.85%**. Both R squared and adjusted R squared is very high and very close to each other in our model.

That is, the developed statistical model explains more than **93%** of the variation in the response variable, a very high-quality model. Similarly, the indicators that we included in the the model, along with the relevant interactions, estimates more than **98%** of the total variation in the response variable AWPR.

The Residual Standard Error (RSE), represents the approximate difference between the observed and predicted outcomes in the model. We got a The Residual Standard Error (RSE) of 0.1 from our model, which implies that the observed response value differs from the predicted response value by .1 unit on an average.

In Table 6.1, below, we rank the individual significant attributable variables and interactions with respect to their percentage of contribution to the response AWCP.

Table 6.1: Rank of The Most Significant Indicators and Their Interactions According to The Percentage of Contribution to The Response AWCP

Rank	**Indicators**	**Contribution(%)**
1	Div_Yield	8.95
2	$Div\_Yield \cap US\_PSR$	7.85
3	$PE \cap US\_ICS$	7.81
4	$log(US\_GDP)$	7.79
5	$US\_ICS \cap US\_GDP$	6.8
6	US_ICS	6.54
7	$Div\_Yield \cap US\_ICS$	6.13
8	$PE \cap US\_PSR$	5.11
9	$BETA \cap US\_PSR$	4.76
10	$Div\_Yield \cap BETA$	4.42
11	$BETA$	4.35
12	$Div\_Yield \cap US\_GDP$	4.31
13	$BETA \cap US\_GDP$	3.75
14	$PE \cap US\_GDP$	3.67
15	$US\_PSR \cap US\_GDP$	3.37
16	US_PSR	2.65
17	PE	2.45
18	$PE \cap BETA$	1.63

The ranking is important, given the fact that in a survey or experiment if the group of experimenters or surveyors know beforehand the most important variables which account for the response, they might collect information on those important variables only which will be economical and less time-consuming as they might not be interested in the insignificant variables which contribute very minimum to the response or do not contribute at all. The following Figure 6.5 gives a pictorial representation of all risk factors and interactions that significantly contribute to the response.

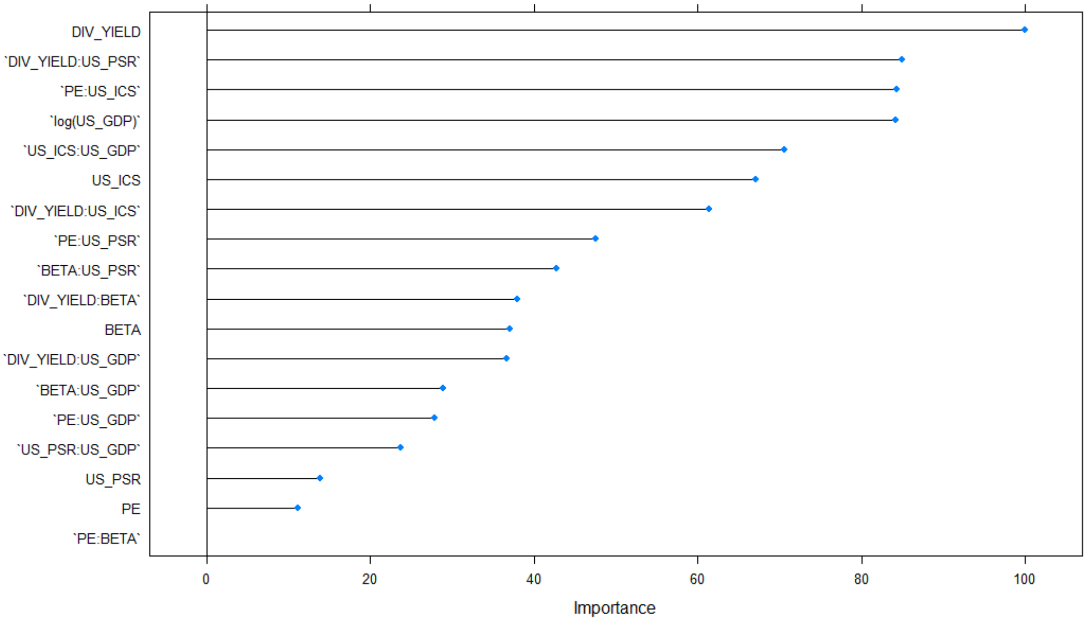


Figure 6.5: Importance plot of The Risk Factors and Interactions According to Their Contributions to The Response

6.2.4 Residual Analysis

Once the statistical model has been developed, it is necessary to check the model assumptions by performing residual analysis. In our case, we have proposed a multiple non-linear regression model, which is very useful and accurately conveys some important information on the subject matter.

6.2.4.1 Mean residual should be zero

When one performs multiple linear regression (or any other type of regression analysis), one gets a line that best fits the data. The entire data points usually don't fall exactly on this regression equation line; they are scattered around. A residual is a vertical distance between a data point and the regression line. Each data point has one residual. The residuals are positive if they are above the regression line and negative if they are below the regression line. If the regression line perfectly passes through the point, the residual at that point is zero. The residual(error) is defined as:

$$\text{residual} = \text{observed value-predicted value} = y - \hat{y} \quad ,$$

where y and $\hat{y}$ are the observed and predicted response. $\hat{e}$ is the estimated residual error from the linear fit.The sum of the residuals equals zero assuming that the regression line is actually the line of "best fit." In our case, the mean residual is $1.97 * 10^{-18}$ implying that it is almost zero.

6.2.4.2 Normality of residual

One important assumption of our model is normality of residual. From Figure 6.6, we see that the studentized residual follows a normal pattern.

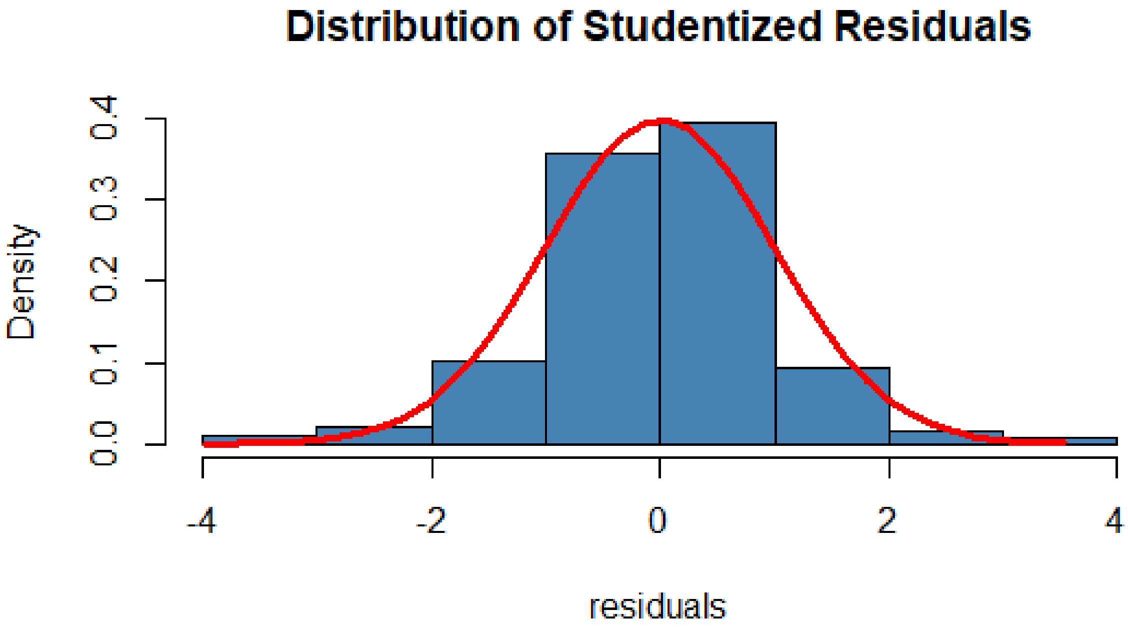


Figure 6.6: Normality of Studentized Residual Plot

6.3 Validation and Prediction Accuracy of The Proposed Model

We developed our analytical model on 80% training data and validated the model based on 20% testing data. In the testing data(validation data) the test error is the average error that occurs from using the analytical method to predict the response on a new observation.That is, a measurement that was not used in training the method. The test error gives an idea about the consistency of the analytical model. We obtained an ac-curacy of 98.7% in terms of R^2 in our validation(testing) set Moreover, we performed repeated ten-fold repeated cross-validation(10 times) for our validation testing.The primary objective is that we will use 10-fold cross-validation, then we repeated cross-validation ten times, where each of the repetition folds are split differently. In 10-fold cross-validation, the training set is divided into ten equal subsets. One of the subsets is taken as the testing set in turn, and (10-1) = 9 subsets are taken as a training set in the proposed model. The error mean square error E_1 is computed for the held out set. This procedure is repeated 10 times; each time, a different group of observations is treated as a validation set. This process results in 10 estimates of the test error, $E_i, \quad i = 1, \ldots 10$. The average error of each set, throughout the cross-validation process is termed as a cross-validated error. The following Figure 8, illustrates briefly the idea of 10 fold repeated cross-validation, where $E_i, \quad i = 1, \ldots 10$ is the mean square error(MSE) in each iteration and ACVE is the average cross-validated error.

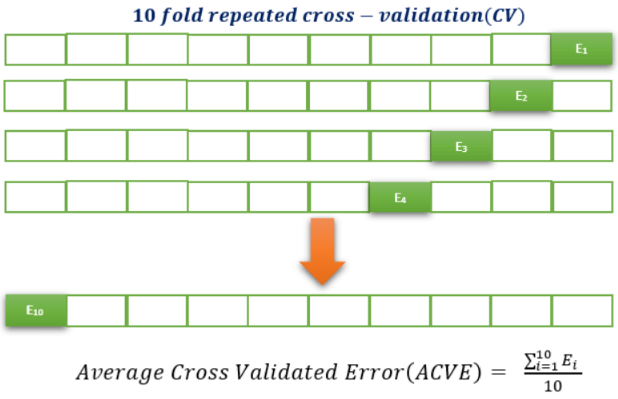

$10\ fold\ repeated\ cross - validation(CV)$
E₁
E₂
E₃
E₄
E₁₀
$Average\ Cross\ Validated\ Error(ACVE) = \frac{\sum_{i=1}^{10} E_i}{10}$

Figure 6.7: Brief Illustration Of Repeated Ten Fold Cross Validation

The R^2, Root Mean Square Error(RMSE) and Mean Absolute Error(MAE) for our model for the test data set are 98.7 , 0.1 and 0.07 respectively. The following Figure 8 illustrates how the R^2 and $RMSE$ varies in the different folds of test data.

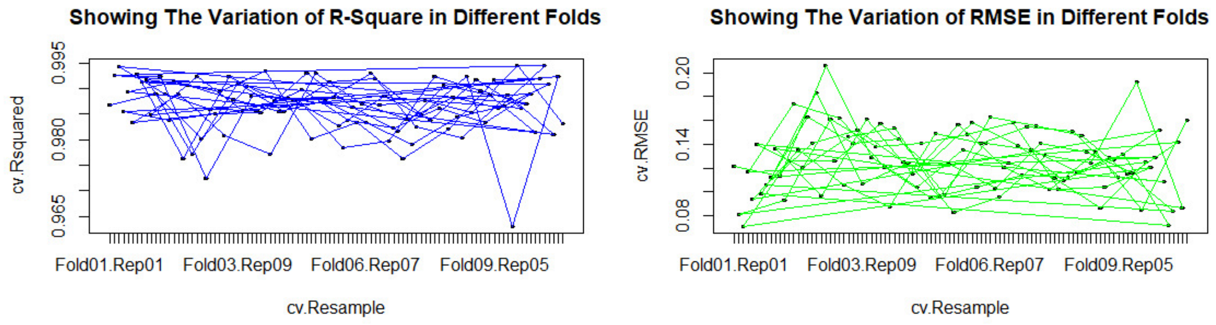


Figure 6.8: Variation of R^2 and $RMSE$ in different folds

The above Figure 6.8 illustrates that the R^2 remains very high(around 98%) and $RMSE$ remains low(around 0.1) for different repeated cross-validated folds the test(validation) data. Hence, we can conclude that, the accuracy of the model is pretty consistent. We have obtained a non-linear analytical model for ABBV weekly stock price with high accuracy, which is a function of the combination of some financial and economic indicators that drive the ups and downs of this particular stock. In the next section, we will be discussing some techniques which will optimize (maximize) our model response (objective function). We will also find the optimum values of the all six factors that lead to the optimization of the weekly response of ABBV stock.

6.4 Response Surface Analysis-A method to optimize The Average Weekly Closing Price For AbbVie Inc.

In the literature of mathematical statistics, response surface methodology (RSM)[42] explores the association between several indicators and one or more response variables. The main idea of RSM is to find the optimum(maximum or minimum) response using an appropriate statistical model. It is also advantageous in finding the optimum values of the indicators used in the model to optimize the response.

6.4.1 A Formal Analytical Approach For The Response Surface Model Using Desirability Function

The concept of desirability function is one of the most frequently used methods in the industry for the optimization of one or more responses. The desirability functions approach, initially proposed by Harrington (1980)[60], are popular in the literature of Response Surface Methodology (RSM). The desirability function transforms each of the estimated response $Y_i(x)$ to a desirability value $d_i(Y_i)$, where $0 \leq d_i \leq 1$.

For individual response $Y_i(x)$, a desirability function $d_i(Y_i)$ takes on values within [0,1]. $d_i(Y_i) = 0$ represents a entirely undesirable value of response Y_i and $d_i(Y_i) = 1$ represents a completely desirable or ideal response value. The value of $d_i(Y_i)$ increases as the "desirability" of the corresponding response increases. The individual desirabilities are then merged together using the geometric mean, which gives the overall desirability D:

$$D = [\textstyle\prod_{i=1}^{k} d_i(Y_i)]^{\frac{1}{k}} \quad ,$$

where k denotes the number of responses. In our study, $k = 1$.
Depending on whether a particular response Y_i is to be maximized, minimized, or assigned a target value, different desirability functions $d_i(Y_i)$ can be used.

A useful class of desirability functions was proposed by Derringer and Suich (1980)[42]. Let L_i, U_i and T_i be the lower, upper, and target values, respectively, that are desired for response Y_i, with $L_i \leq T_i \leq U_i$.

If there is a specific target set up for the response, then its desirability function is given by,

$$d_i(\hat{Y}_i) = \begin{cases} 0 & \text{if } \hat{Y}_i(x) < L_i \\ \left(\frac{\hat{Y}_i(x)-L_i}{T_i-L_i}\right)^s & \text{if } L_i \leq \hat{Y}_i(x) \leq T_i \\ \left(\frac{\hat{Y}_i(x)-U_i}{T_i-U_i}\right)^t & \text{if } T_i \leq \hat{Y}_i(x) \leq U_i \\ 1 & \text{if } \hat{Y}_i(x) > U_i \quad , \end{cases} \tag{6.5}$$

where s and t in the above equation determines how important it is to hit the target. For $t = s = 1$,the desirability function increases linearly towards the direction of T_i. For $s < 1, t < 1$, the desirability function is convex, and for $s > 1, t > 1$, the desirability function is concave. If want the response to be maximized, the individual desirability is given as,

$$d_i(\hat{Y}_i) = \begin{cases} 0 & \text{if } \hat{Y}_i(x) < L_i \\ \left(\frac{\hat{Y}_i(x)-L_i}{T_i-L_i}\right)^s & \text{if } L_i \leq \hat{Y}_i(x) \leq T_i \\ 1 & \text{if } \hat{Y}_i(x) > T_i \quad , \end{cases} \tag{6.6}$$

where T_i is interpreted as a large enough value for the response. Finally, if we want to minimize a response, the desirability function is given as,

$$d_i(\hat{Y}_i) = \begin{cases} 1 & \text{if } \hat{Y}_i(x) < T_i \\ \left(\frac{\hat{Y}_i(x)-L_i}{T_i-U_i}\right)^s & \text{if } T_i \leq \hat{Y}_i(x) \leq U_i \\ 0 & \text{if } \hat{Y}_i(x) > U_i \quad , \end{cases} \tag{6.7}$$

where T_i is interpreted as a small value for the response.

The desirability function approach consists of the following steps:

1. Given the data, fit response models for all k responses (in our study, we have single response, so, k = 1) ;

2. Define individual desirability functions for each responses;

3. Optimize the overall desirability D with respect to the **controllable** indicators.

6.4.2 Numerical Results

In our study, we started developing a data-driven non-liner analytical model with *three* financial indicators and *three* economic indicators. After developing the predictive model with high accuracy, our goal is to maximize the response AWCP and find the optimum values of the indicators at which the response is being maximized. We now proceed to maximize AWCP (within its domain) in model (6.4) in section 6.2.3. The analytical method of optimization, requires the constraints of optimization. The following Table 2 presents the constraints on the ten indicators.

Table 6.2: Constraints On The Indicators Showing the Lower and Upper Limits

Indicators	Lower	Upper
Div_Yield	2.212	6.506
Beta	-.9	2.3
PE	16.55	41.42
US_GDP	19.6	21.9
US_ICS	89.8	101.4
US_PSR	6.7	8.8

Table 6.2 provides the lower and upper limits of all ten indicators used in our study. Next, using the equation (6.3) of section 6.2.3, we will maximize the estimated response from model (6.4) and obtain the optimum values of all ten indicators. The following Table 6.3, provides the estimated maximum response AWCP along with the optimum values of the indicators.

Table 6.3: Estimated maximized Response With Optimum Values of Indicators

Response & Indicators	**Optimum Values**
AWCP (Response)	120$
Div_Yield	4.36
Beta	.7
PE	39.56
US_GDP	19.6
US_ICS	101.4
US_PSR	8.8

The following Table 6.4, provides the values of R^2 , ${R^2}_{Adjusted}$ and desirability value along with the 95% confidence and 95% prediction interval of the estimated response WCP.

Table 6.4: Some Useful Results Related to The Optimized Response

Estimated Maximized Value	120$
Desirability	1
R^2	98.61%
${R^2}_{Adjusted}$	98.57%
95% CI	(110.71, 129.29)
95% PI	(110.28, 129.72)

6.4.3 Graphical Visualization of The Estimated Response surface

One important aspect of the response surface methodology is that, it helps the researchers understand the variation of the response in three dimensional plot in the presence of any two risk factors, keeping the rest of the risk factors fixed at a particular level. Response surface plots (contour and surface plots) are very useful in order to obtain desired values of the response, and optimum conditions for any two indicators keeping the others fixed at a particular level. In a contour plot, the response surface is observed as a two-dimensional plane where all the points that have the similar response are connected to create contour lines of constant responses. A surface plot generally exhibits a three dimensional view that may provide a clearer picture of the response's behavior. Since, the *three* economic indicators are not **controllable**, we won't be including those plots and only focus on the combination of *three* financial indicators included in model (3) of section 2.3. In this section, We will illustrate different contour and surface plots that will help researchers to understand the nature of the relationship between any the two indicators and the response (AWCP). The following three figures (Figure 6.9-Figure 6.11) describe the variation of the estimated response AWCP as any single or two indicators varies keeping the others fixed at particular level.

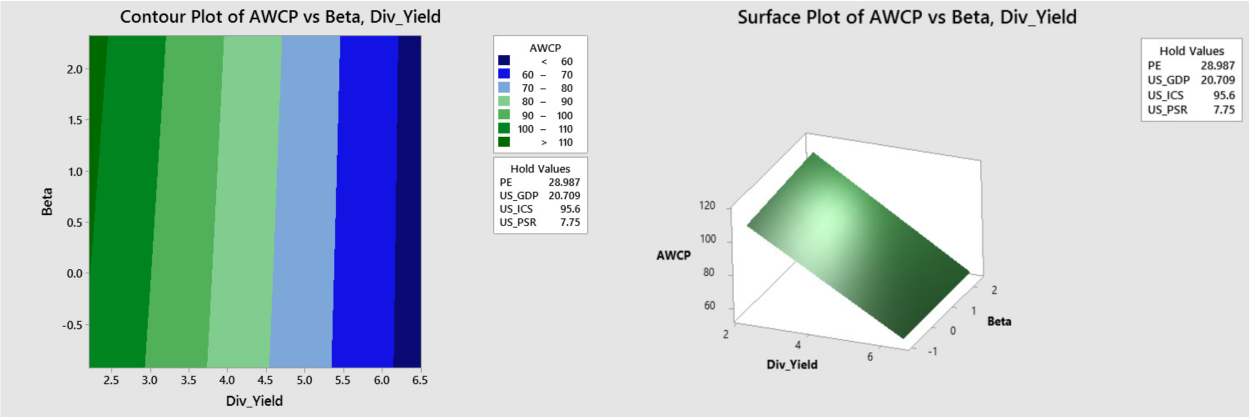


Figure 6.9: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as Div_Yield and Beta Varies Keeping Other Risk Factors Fixed at a Specified Level

From the above Figure 6.9, we see that the estimated response is maximized (greater than 110) for any positive beta when dividend yield is less than approximately 2.5 keeping all other indicators at a constant level.

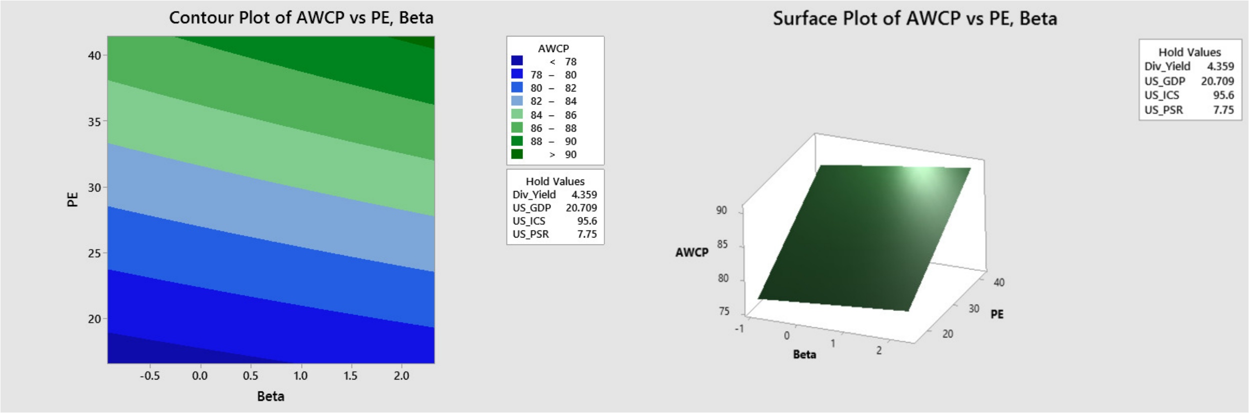


Figure 6.10: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as PE and Beta Varies Keeping Other Risk Factors Fixed at a Specified Level

From the above Figure 6.10, we see that the estimated response is maximized (greater than 90) when PE is greater than approximately 40 and beta is approximately greater than 1.7, keeping all other indicators at a constant level.

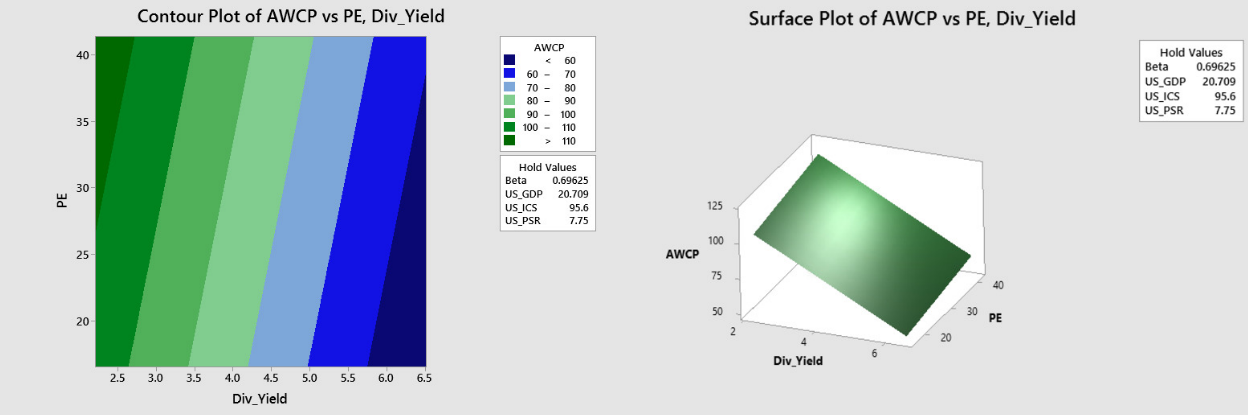


Figure 6.11: Showing the Contour Plot(left) and Surface Plot(right) of The Estimated Response Surface as PE and Div_Yield Varies Keeping Other Risk Factors Fixed at a Specified Level

The above Figure 6.11, demonstrates that the estimated response is maximized (greater than 110) when PE is greater than approximately 17 and Div_Yield is approximately less than 2.5, keeping all other indicators at a constant level.

Table 6.5: The List of Observed and Predicted Values of The Response AWCP

Observations	Observed	Predicted
1	88.1	88.5
246	89.1	90.2
247	89.2	90.4
308	93.5	96.4
373	96.4	100.1
374	96.2	99
375	95	99.5
376	94	98.5
377	93	97.5
378	92.7	96
379	92.3	95
435	92	95.5
436	92	96.5
437	91.8	96.4
438	91.2	96.1
439	91.4	96.1
440	92.4	94.1
449	109.8	106
450	113.2	107.8
451	114.5	102
452	115.9	112
452	115.9	111.8
453	117	113.5
454	118.2	115.2
459	115.4	118.3
460	114.8	118.4
473	112.6	115.8
474	111.3	115.2
475	111.4	114.9
476	111.3	114.4
477	111.8	113.8
482	116.4	111
483	117.9	111
484	116.4	109.4
485	115	108.9
486	113	108.5
496	99.8	102.9
497	100	102.9
504	97.8	97.7
505	97.7	96.6
506	98	97
507	98	97

6.5 Discussion

In our study, we have developed an analytic model, which describes the significant indicators and the and associated interactions responsible for the ups and downs of the stock

ABBV very accurately. After obtaining the significant indicators, along with their significant interactions, we rank them with respect to the percent of contribution to the stock price, as shown in figure 5. The highest contributing risk factor is $DividendYield(X_1)$, contributing 8.95% of the total variation to the response CLOSE. The next most significant contribution is an interaction term that is the combined effect of $DividendYield(X_1)$ and the US Personal Saving Rate(X_6) with a contribution of 7.85% to the response. Numbers 3, 4, and 5 are respectively the combined interaction effect of $PE \cap US\_ICS(X_3 \cap X_5)$, $US\_GDP(X_4)$, and interaction between $US\_ICS \cap US\_GDP(X_5 \cap X_4)$ with the contribution of 7.81%, 7.79%, and 6.8%, respectively. Hence, summing these contribution of indicators up, we identify that they explain more than 98% of the total variability in the ABBV stock price. The following Table 6.5 demonstrates the list of the observed and predicted responses from our data driven non-linear analytical model. From the Table 5, we can see clearly that the predictions are very close to the actual observed values and thus, testify for our model's high accuracy and predictive power.

Moreover, we have performed response surface analysis to maximize the estimated response from our developed analytical model and also obtained the optimum values for the three financial and three economic indicators. The R^2 and ${R^2}_{Adjusted}$ from Table 4, attests the fact that the optimized model is good in terms of accuracy. Also, we have obtained almost similar accuracy in terms of R^2 and ${R^2}_{Adjusted}$ that we obtained from our analytical model (4). The desirability value 1 from table 4 indicates that the estimated fit is most desirable/ideal. We can address the usefulness and importance of the proposed model in the subject area in **six** important categories.

These categories are enumerated below:

1. We have identified and tested the individual attributable variables(indicators) responsible for increase or decrease of ABBV stock price.

2. We have identified the significant interactions that influence the response $AWCP$, in our model.

3. we have ranked the individual attributable variables and interactions as a percentage of contribution to the response.

4. We can obtain excellent predictions of the weekly closing price for the healthcare stock ABBV from our analytical model with a high degree of accuracy.

5. We compared the original and the estimated response AWCP using our analytical model and found that they are very close to each other, indicating the high accuracy of our model.

6. We have performed *response surface analysis*, for this stock ABBV to maximize the predicted response, Average Weekly Closing Price (AWCP) and identified the optimum levels of the risk factors that maximize the predicted response with high degree of accuracy.

7. We have calculated the 95% confidence interval and 95% prediction interval for the estimated response.

6.6 Conclusion

We have developed a real data-driven analytical model that very accurately identifies the following very useful findings concerning the weekly stock price of Abbvie Inc.(ABBV):

- Identifies the significant attributable variables (indicators) that drive the degree to which the particular stock changes(%).

- Identifies the significant interactions of the indicators that contribute to the weekly stock price.

- We rank the individual and interactions of the indicators with respect to their percentage of contribution to the response AWCP.

- The developed analytical model predicts the response change accurately for the specified values of a set of indicators.

- The developed analytical model can be used strategically to increase profit margin by working with the identified indicators.

- Furthermore, we have performed surface response analysis to identify the indicators' **target values** that maximizes the Weekly Stock Price(AWCP) based on the identified values of the indicators.

The developed analytical model has been evaluated by several statistical methods that include the R^2 and $R^2_{adjusted}$ that attest to its high quality. Investment bankers and financial analysts usually keep track of a company's stock price and percentage increase in stock price to measure a company's financial solvency, market performance, and general viability. A steadily rising stock price indicates that a company is moving to the direction of money-making. Besides, if the stockholders are happy, and the company is on its way towards prosperity, the executives are likely to retain their positions with the company. Conversely, if a company is bending over backward, as reflected by a deteriorating stock price, a company's board of directors may decide to fire its top operatives. Thus, decreasing stock price isn't good for a company's higher-ups and financial health as a whole. Our study's findings suggest that financial analysts and quantitative researchers might need to pay more attention to different significant financial and economic attributes, as our research suggests, that will maximize the stock price. One can build a similar model based on other stocks depending on the research interest. Also, since, the individual ups and downs of the stock price of an organization has a positive correlation with current and future increases in the *productivity growth rate* at business cycle, our proposed statistical model can be used for firms' promotion policies, and they may be useful for managers and human resources professionals. Financial analysts can use our model to predict the individual company's percentage change in stock price by using the significant indicators. It will help the different financial firms to identify their financial health, as the higher a stock price is, the more likely a company's prospects become. Identifying those financial institutions are essential as the increased stock price is correlated with an increase in productivity. Finally, our proposed statistical model is highly useful for *decision making* and *strategic planning* on controlling the factors responsible for the company's long-term viability.

Chapter 7: A Stochastic Analytical Model that Monitors the Profit Structure of Healthcare Business Segment (HBS) of S&P 500 as a Function of Real Time.

7.1 Introduction

The healthcare business segment (HBS) incorporates businesses that supply medical services, manufacture medical equipment or drugs, provide medical insurance, and facilitate healthcare provision to patients. It is one of the immense sectors and contributes significantly to the U.S. economy, accounting for approximately a fifth of overall gross domestic product (GDP). Healthcare stock change has an enormous impact on the global economy affecting its overall GDP and other financial factors. In our present study, we created a data-driven stochastic analytical method based on average weekly percentage return (AWPR) of healthcare stocks. Our data contains the information of *fifty-nine* heath care and pharmaceutical stocks from S&P 500. We introduce and define a function called *stochastic growth function (SGF)*, depending on an indicator called **index-indicator** ($\mathcal{I}$) for the stocks that monitors and predicts the profit returns as a function of real time of the whole HBS or any single business. The **index-indicator**, ($\mathcal{I}$), is calculated using the profit returns and will result in one of the three estimates, $\mathcal{I} > 1, \mathcal{I} \approx 1$, and $\mathcal{I} < 1$. Our analytical method found the estimated $\mathcal{I}$ for HBS as a whole to be $\mathcal{I} = 2.12 > 1$ implying that the average profitable returns are increasing as a function of time. Also, we selected a particular stock (**AbbVie Inc.**) from S&P 500 based on the *high dividend, high return, and low beta risk* and found the index indicator ($\mathcal{I}$) to be $\mathcal{I} = .97 \approx 1$, which says that the particular stock is performing approximately the same. Our proposed analytical methodology can be implemented to any single business firm or a whole business sector, that provides improved strategy for monitoring, assessing, and evaluating the profit structure pattern of any given industry as a stochastic realization of time. In the following sections, we proceed to discuss the analytical structure and stochastic methodology for developing such a model to monitor the profitability returns of a firm in real time.

7.2 Methodology

7.2.1 Structure of the Data

We have collected the weekly stock return for all the 59 stocks from the healthcare and pharmaceutical industry. The duration period is from 08-01-2016 to 30-12-2019. All total, we have information for 209 weeks. At first, we computed the weekly return for each of 59 stocks for the time period and then we took the average of all 59 stocks for the 209 weeks. After computing the weekly average return, we multiplied these observation by 100 to obtain the Average Weekly Percentage Return (AWPR). Thus, AWPR gives the percentage gain or loss of an investor on weekly basis. The method that we use to compute the Weekly Percentage Return (WPR) for individual stocks is as follows:

$$WPR_t = \frac{P_t - P_{t-1}}{P_{t-1}} * 100 \quad , t = 1, 2, \ldots, 209.$$

where WPR_t is the Weekly Percentage Return at week t. P_t and P_{t-1} are respectively the closing price of an individual stock at week t and week $t-1$, respectively.
Once we have the Weekly Percentage Return(WPR), we can compute the Average Weekly Percentage Return (AWPR) for all the 209 weeks by taking the average across the 59 stocks. The method for AWPR is as follows:

$$AWPR_t = \frac{1}{59} \sum_{j=1}^{59} WPR_{(t)j} \quad , j = 1, 2, \ldots, 59.$$

where $AWPR_t$ is the Average Weekly Percentage Return at week t and $WPR_{(t)j}$ is the Weekly Percentage Return(WPR) for the j^{th} stock and t^{th} week. $j = 1, 2, \ldots, 59$, and

$t = 1, 2, \ldots, 209$. The following Figure 7.1, provides the *data network diagram* for better understanding of the scenario.

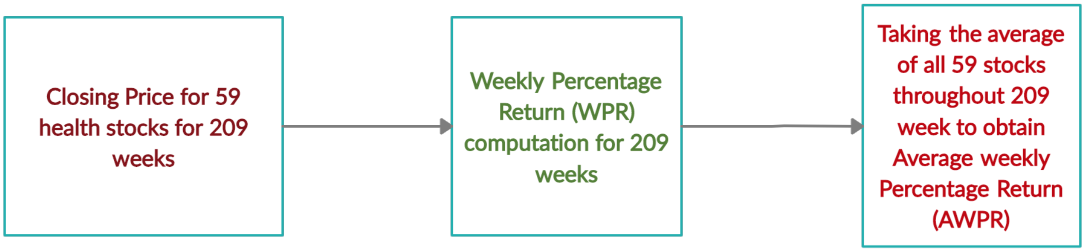

Closing Price for 59 health stocks for 209 weeks
Weekly Percentage Return (WPR) computation for 209 weeks
Taking the average of all 59 stocks throughout 209 week to obtain Average weekly Percentage Return (AWPR)

Figure 7.1: Network Diagram for Calculating AWPR

Table 7.1: Showing ten arbitrary AWPR for HBS

Date	AWPR
26-02-2016	2.02
03-06-2016	1.20
09-12-2016	1.92
13-10-2017	-0.91
01-12-2017	1.66
12-01-2018	2.24
16-02-2018	4.5
06-04-2018	-2.39
25-10-2019	0.64
06-12-2019	1

7.2.2 The Stochastic Process that Drives the Data

Our data can be thought as a stochastic realization at time t. It is the weekly percentage return for 59 stocks as a function of time. Mathematically, the stochastic process can be represented as $\{R_t\}_{t\in T}$ where R_t the set of AWPR values of HBS and T is a time index such that $T = [0, \infty)$. The following Table 7.1 illustrates ten random AWPR from 209 weeks for HBS.

In the next section, we proceed to describe briefly the method of developing the index-indicator ($\mathcal{I}$) of AWPR.

7.2.3 Developing the Index-Indicator ($\mathcal{I}$)

In the context of finance, different firms in the healthcare and pharmaceutical industry would be interested in having a method to monitor and evaluate the profitability structure as a function of real time. For instance, a healthcare firm in HBS would like to monitor the investment trend based on whether they are making a profit or loss with respect to specific period of time. Thus, it is crucial to monitor how an investment is progressing as a function of time. In this regard, We define *stochastic growth function (SGF)* that measures the rate of change of a profitability process as a stochastic realization of time. The analytical structure of the SGF function is:

$$\Omega(AWPR_t;\mathcal{I};\vartheta) = \frac{\mathcal{I}}{\vartheta}\left(\frac{AWPR_t}{\vartheta}\right)^{\mathcal{I}-1} \quad , \quad \mathcal{I} > 0, \vartheta > 0, t > 0 \quad , \tag{7.1}$$

Where $\mathcal{I}$ and ϑ are the shape and scale parameters, respectively, and $AWPR_t$ denotes return component which is the stochastic realization of time. [8] [11]. For n AWPRs, $AWPR_1 < AWPR_2 < \ldots < AWPR_n$, (where $AWPR_1 < AWPR_2 < \ldots < AWPR_n$ are the observed and successive), the joint probability density function, $f(AWPR_{t_1}, \ldots, AWPR_{t_n})$ can be expressed in terms of $\Omega(AWPR_t;\mathcal{I};\vartheta)$ as follows,

$$
\begin{aligned}
f(AWPR_{t_1},\ldots,AWPR_{t_n}) &= \prod_{i=1}^{n}\Big(\Omega(AWPR_{t_i}\Big)exp\Big[-\int_0^{AWPR_{t_n}}\Omega(y)dy\Big]\\
&= \prod_{i=1}^{n}\frac{\mathcal{I}}{\vartheta}\Big(\frac{AWPR_{t_i}}{\vartheta}\Big)^{\mathcal{I}-1}exp\Big[-\int_0^{AWPR_{t_n}}\frac{\mathcal{I}}{\vartheta}\Big(\frac{y}{\vartheta}\Big)^{\mathcal{I}-1}dy\Big]\\
&= \frac{\mathcal{I}^n}{\vartheta^{n\mathcal{I}}}\Big(\prod_{i=1}^{n}\Big)^{\mathcal{I}-1}exp\Big[-\Big(\frac{AWPR_{t_n}}{\vartheta}\Big)^{\mathcal{I}}\Big],\\
&where\ \ AWPR_1 < AWPR_2 < \ldots < AWPR_n.
\end{aligned}
\tag{7.2}
$$

Implementing the method of Maximum Likelihood Method (MLE) of parameter estimation, we can estimate the parameters $\mathcal{I}$ and ϑ from (7.3). The log-likelihood function is

$$
\mathscr{L} = L(AWPR_t;\mathcal{I};\vartheta) = nln\mathcal{I} - n\mathcal{I}ln\vartheta + (\mathcal{I}-1)\sum_{i=1}^{n}ln(AWPR_{t_i}) - \Big(\frac{AWPR_{t_n}}{\vartheta}\Big)^{\mathcal{I}} \tag{7.3}
$$

The parameter, $\mathcal{I}$ is a function of $AWPR_{t_n}$, the largest return value. We compute the estimate of $\mathcal{I}$ by equating the partial derivative of $\mathscr{L}$ with respect to $\mathcal{I}$ and setting it equal to zero, then solving for $\mathcal{I}$, given by,

$$
\frac{\partial\mathscr{L}}{\partial\mathcal{I}} = 0; \hat{\mathcal{I}} = \frac{n}{\sum_{i=1}^{n}log\big(\frac{AWPR_{t_n}}{AWPR_{t_i}}\big)} \quad . \tag{7.4}
$$

The parameter ϑ is a function of $\mathcal{I}$. In a similar way, as above, the estimate of ϑ is computed by equating the partial derivative of $\mathscr{L}$ with respect to ϑ to zero and then, substituting the estimate of $\mathcal{I}$, given by,

$$
\frac{\partial\mathscr{L}}{\partial\vartheta} = 0; \hat{\vartheta} = \frac{AWPR_{t_n}}{n^{\frac{1}{\mathcal{I}}}} \quad . \tag{7.5}
$$

In the context of financial modelling, we define the *index-indicator* ($\mathcal{I}$) as follows:

$$
\mathcal{I} = \frac{n}{\sum_{i=1}^{n}log\big(\frac{AWPR_{t_n}}{AWPR_{t_i}}\big)} \tag{7.6}
$$

where $AWPR_{t_n}$ is the largest AWPR for the HBS. Now, we will show that how $\mathcal{I}$ depends on the interpretation of the stochastic growth function $\Omega(AWPR_t)$.

- **Case 1:** $\Omega(AWPR_t)$ **is decreasing with time**

For $\Omega(AWPR_t)$ being a decreasing function of t, we have,

$$\begin{aligned}
&\Omega(AWPR_t) < \Omega(AWPR_{t-1}) \quad , for \quad t-1 < t \\
&\Rightarrow \frac{\mathcal{I}}{\vartheta}\left(\frac{AWPR_t}{\vartheta}\right)^{\mathcal{I}-1} < \frac{\mathcal{I}}{\vartheta}\left(\frac{AWPR_{t-1}}{\vartheta}\right)^{\mathcal{I}-1} \\
&\Rightarrow \left(\frac{AWPR_t}{\vartheta}\right)^{\mathcal{I}-1} < \left(\frac{AWPR_{t-1}}{\vartheta}\right)^{\mathcal{I}-1} \\
&\Rightarrow \left(\frac{AWPR_{t-1}}{AWPR_t}\right)^{\mathcal{I}-1} > 0
\end{aligned}$$

Replacing $AWPR_t$ with $AWPR_{t-1}$, the above inequality results in
$(\frac{AWPR_{t-2}}{AWPR_{t-1}})^{\mathcal{I}-1} > 0.$
Replacing $AWPR_{t-1)}$ by $AWPR_{t-2}$, in above inequality we have,
$(\frac{AWPR_{t-3}}{AWPR_{t-2}})^{\mathcal{I}-1} > 0.$
Proceeding in a similar manner, results in $(\frac{AWPR_{t1}}{AWPR_{t0}})^{\mathcal{I}-1} > 0$, where $AWPR_{t0}$ is the initial AWPR in the data.
Arranging all the above inequalities and expressing in a product form , we have,

$$\begin{aligned}
&\left[\left(\frac{AWPR_{t-1}}{AWPR_t}\right)^{\mathcal{I}-1}\left(\frac{AWPR_{t-2}}{AWPR_{t-1}}\right)^{\mathcal{I}-1}\cdots\left(\frac{AWPR_{t2}}{AWPR_{t1}}\right)^{\mathcal{I}-1}\left(\frac{AWPR_{t1}}{AWPR_{t0}}\right)^{\mathcal{I}-1}\right] > 0 \\
&\Rightarrow \left(\frac{1}{(AWPR_t)(AWPR_{t0})}\right)^{\mathcal{I}-1} > 0
\end{aligned}$$

Since, $AWPR_t, AWPR_{t0} > 0$ (With respect to our discussion previously, we made all the returns positive by adding a constant C, as returns can be positve and negative also, and the formula of $\mathcal{I}$ involves logarithm), in order to satisfy the above inequality, $\mathcal{I}$ must satisfy, $\mathcal{I} - 1 < 0 \implies \mathcal{I} < 1$. (Bad for investment).

- **Case 2:** $\Omega(AWPR_t)$ **is increasing with time**

For $\Omega(AWPR_t)$ being an increasing function of t, proceeding with the similar logic, we end up having $\mathcal{I} > 1$. (Good for investment).

- **Case 3:** $\Omega(AWPR_t)$ **is constant**

For $\Omega(AWPR_t)$ being an independent function of t, proceeding with the similar argument, we end up getting $\mathcal{I} \approx 1$.

We can now estimate the value of the stochastic growth factor $\Omega(AWPR_t)$ (given in (7.1)) used in the analytical modeling of the profitability structure as a function of t, given the estimates of parameters $\mathcal{I}$ and ϑ. $\Omega(AWPR_t)$ is a time-dependent measure of the rate of change in AWPR. A decrease in $\Omega(AWPR_t)$ implies that AWPR is decreasing as a function of time or an deterioration in the overall profit structure. This means that $\mathcal{I} < 1$. A rise in $\Omega(AWPR_t)$ suggests that there is an improvement pattern in AWPR as a function of time , implying that $\mathcal{I} > 1$. This means that the HBS is performing well as a whole, and no precautionary measures maybe required to boost the profit structure. When there is no change in $\Omega(AWPR_t)$, it implies that $\mathcal{I} \approx 1$ thus the profitability growth is constant as a function of time. Thus, the behavior of the change in the AWPR of a business segment of S&P 500 is dependent on $\mathcal{I}$ of the stochastic growth function (SGF) $\Omega(AWPR_t)$. That is, we can use $\mathcal{I}$ to evaluate and monitor the profit structure of HBS as a function of time. In the following section, we explain and demonstrate how $\mathcal{I}$ of the SGF can be used to monitor and evaluate the behavior of AWPR.

7.3 Using the index-indicator $\mathcal{I}$ to Monitor and Analyze the Behavior of Portfolio HBS Returns based on AWPR

In this section, we use the index-indicator $\mathcal{I}$ to monitor, assess and evaluate the Behavior of Portfolio Return Based on Average Weekly Percentage Return (AWPR). A portfolio return is concerned with how much profit or loss an investment portfolio incurred containing various investments, over a time period. Since, we have information of AWPR on different stocks for 209 weeks, we can think of this data as a stochastic realization of time. Our main goal for the study to evaluate and monitor if the investment is progressing or regressing with respect to time. Notice that, $AWPR_t$ in section 7.2.1 is a function of time.

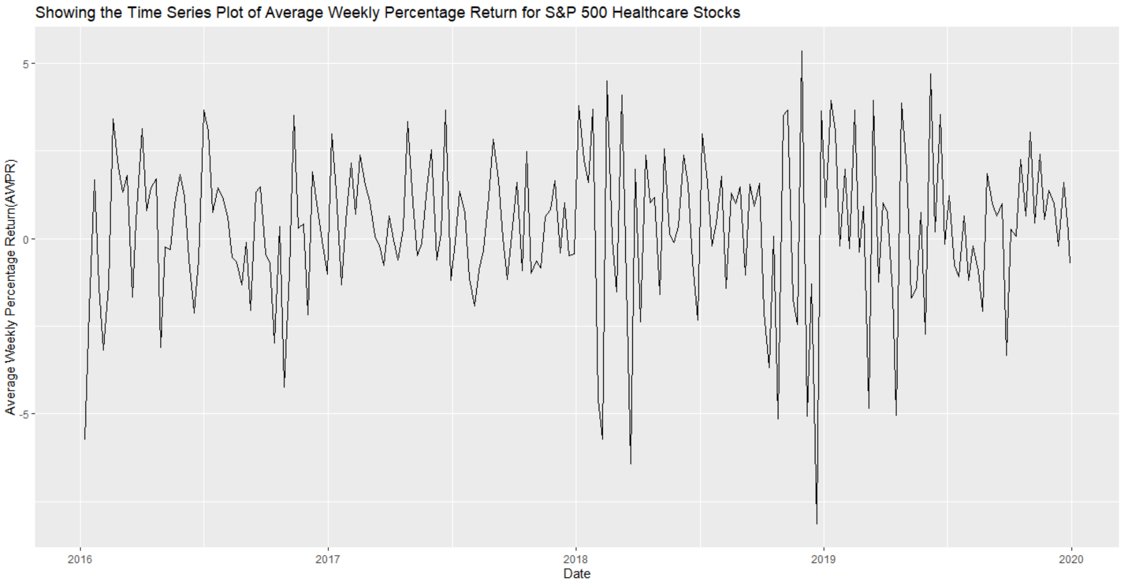


Figure 7.2: Showing the Behavior of AWPR as a function of time

In general, we can see that the AWPR of healthcare business segment (HBS) has high volatility over the 209 weeks period as the series wriggles back and forth. Now, given the nature of the AWPR, we ranked the AWPR from the smallest to the largest. This is given by Figure 2 with the week time index on the horizontal axis and weekly return on vertical axis. The probabilistic behavior of AWPR can be thought as a stochastic realization of time which has similarity with the non-homogeneous Poisson process (NHPP) [132]. Thus, the stochastic process is given by $AWPR_1 < AWPR_2 < \ldots < AWPR_{209}$. This allow us to compute the stochastic growth function of AWPR $\hat{\Omega}(AWPR_t, \hat{\mathcal{I}}, \hat{\vartheta})$, by estimating $\mathcal{I}$ and ϑ from equation (7.5) and (7.6). Since, there are some positive and negative values of AWPR (as shown in Table 7.2.1), we added a constant value $C = 8.15$ (the minimum value to make the maximum negative AWPR positive) to the numerator and denominator in the domain of $log(\frac{t_n}{t_i})$ in equation (7.6). We implemented this to make the domain of the $logarithmic function$ positive. Thus, the new modified formula for **index-indicator** ($\mathcal{I}$) becomes:

$$\begin{aligned}\mathcal{I}^* &= \frac{n}{\sum_{i=1}^{n} log\left(\frac{t_n+C}{t_i+C}\right)} \\ \mathcal{I}^* &= \frac{n}{\sum_{i=1}^{n} log\left(\frac{t_n^*}{t_i^*}\right)}\end{aligned} \tag{7.7}$$

It is also important to investigate if the largest observation t_n^* is an outlier. If it is an outlier, then we need to omit it from our analysis or adopt any other suitable mechanism. An alternate to remove data observation is to replace it by the average of last three observations. Since, the value of index-indicator ($\mathcal{I}^*$) is dependent on t_n^*, the largest observation, and our main analysis is dependent on $\mathcal{I}^*$, we need to make sure that t_n^* is not an outlier, otherwise, the analysis might be biased. We perform the outlier detection analysis in this regard.

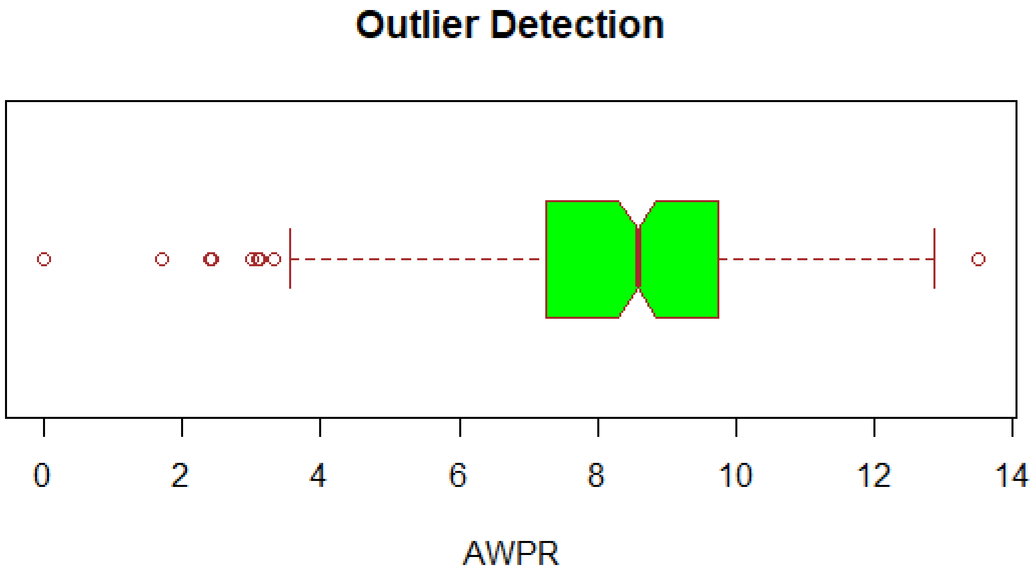


Figure 7.3: Showing the Initial Data Distribution

As the above Figure 7.3 shows, the largest observation (t_n^*) is an outlier. Next, we will apply the interquartile range (IQR) criterion to check validity of the above figure. The IQR criterion tells that all observations falling beyond the range $\mathcal{R} = [Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ are potential outliers, where IQR = third quartile (Q_3) - first quartile (Q_1).

Table 7.2: Summary Measure of the AWPR

Min	Q_1	Q_2	Q_3	Max
0	7.25	8.58	9.74	13.51

From the above table, we see that the largest observation is 13.51.

$$\mathcal{R} = [Q_1 - 1.5IQR, Q_3 + 1.5IQR]$$

$$= [(7.25 - 3.735), (3.515 - 3.735)]$$

$$= [3.515, 13.485]$$

We see that the largest observation t^*_n is falling beyond the range $\mathcal{R}$ and hence is an outlier. We now propose an analytical approach to solve the problem. We take the average of the last three observations and treat the average as our largest observation.

Also, we eliminate the minimum observation ($t^*_1 = 0$) from our data to apply the Equation (7.7) (otherwise (7.7) would be undefined). We have now information of 206 weeks and we can denote it by $t^*_1, t^*_2, \ldots, t^*_{206}$.

To obtain the new largest observation ($t^*_{largest} = t^*_{206}$), we take the arithmetic mean of t^*_{207}, t^*_{208}, and t^*_{209}.

$$t^*_{largest} = \frac{t^*_{207} + t^*_{208} + t_{209}}{3} = \frac{12.64 + 12.85 + 13.51}{3} = 13$$

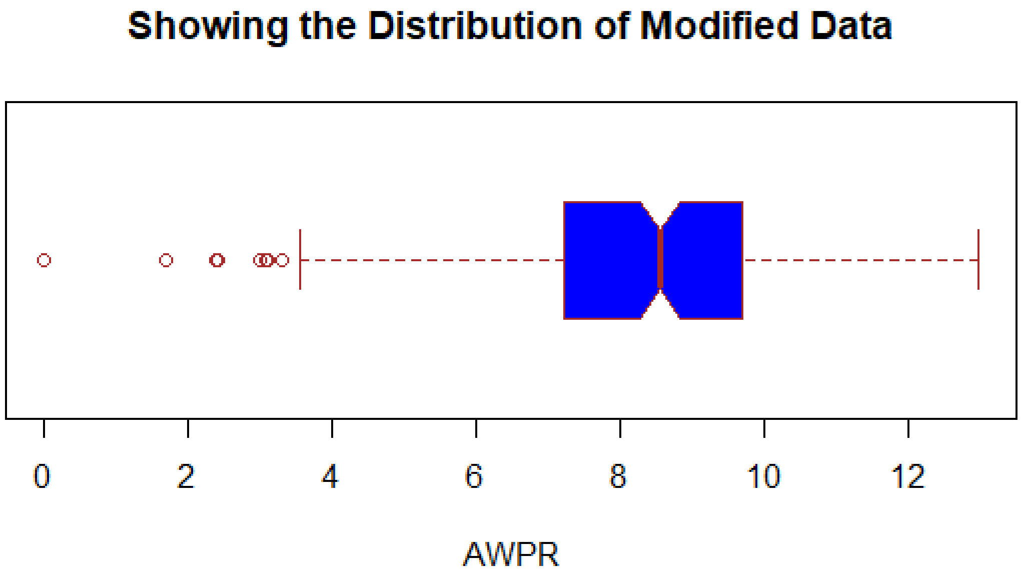


Figure 7.4: Boxplot Showing That Largest Observation is not an Outlier

As the above figure shows, the new largest observation $t^*_{206} = 13$ is not an outlier anymore. We can now analyse the profit structure by looking at the behavior or changes in $\mathcal{I}*$. Now, we estimate the $\mathcal{I}^*$ of the 206 weeks of AWPR, given by Table 7.3, from Equation (7.7). We see that $\mathcal{I}^* = \frac{206}{97.16} = 2.12 > 1$, which indicates that the stochastic growth function of $AWPR_t$, $\Omega(AWPR_t, \hat{\mathcal{I}}^*, \hat{\vartheta})$ is increasing or the HBS is performing well as a function of time.

Table 7.3: Evaluating the Profit Structure for HBS based on $\mathcal{I}$

Estimates	Values
$\hat{\mathcal{I}}$	2.12
$\hat{\vartheta}$	1.05

This finding suggests that the AWPR of heath care business segment are increasing with respect to time and there is no need to make the necessary adjustments/changes in business strategies. This justifies the high quality and efficiency of our analysis of monitoring the AWPR of HBS using the analytic procedure. Given the values of $\hat{\mathcal{I}}$ and $\hat{\vartheta}$, from Table 7.3, we can estimate the intensities $\hat{\Omega}(AWPR_t, \hat{\mathcal{I}}, \hat{\vartheta}))$ using equation (7.1).

$$
\begin{aligned}
\hat{\Omega}(AWPR_t, \hat{\mathcal{I}}, \hat{\vartheta}) &= \frac{\hat{\mathcal{I}}}{\hat{\vartheta}}\left(\frac{AWPR_t}{\hat{\vartheta}}\right)^{\hat{\mathcal{I}}-1} \\
&= \frac{2.12}{1.05}\left(\frac{AWPR_t}{1.05}\right)^{2.12-1} \\
&= 2.02\left(\frac{AWPR_t}{1.05}\right)^{1.12} \quad , \quad AWPR_t \geq 0
\end{aligned}
\tag{7.8}
$$

The following Figure 7.5, describes the behaviour of the SGF $\hat{\Omega}(\cdot)$ as a function of time.

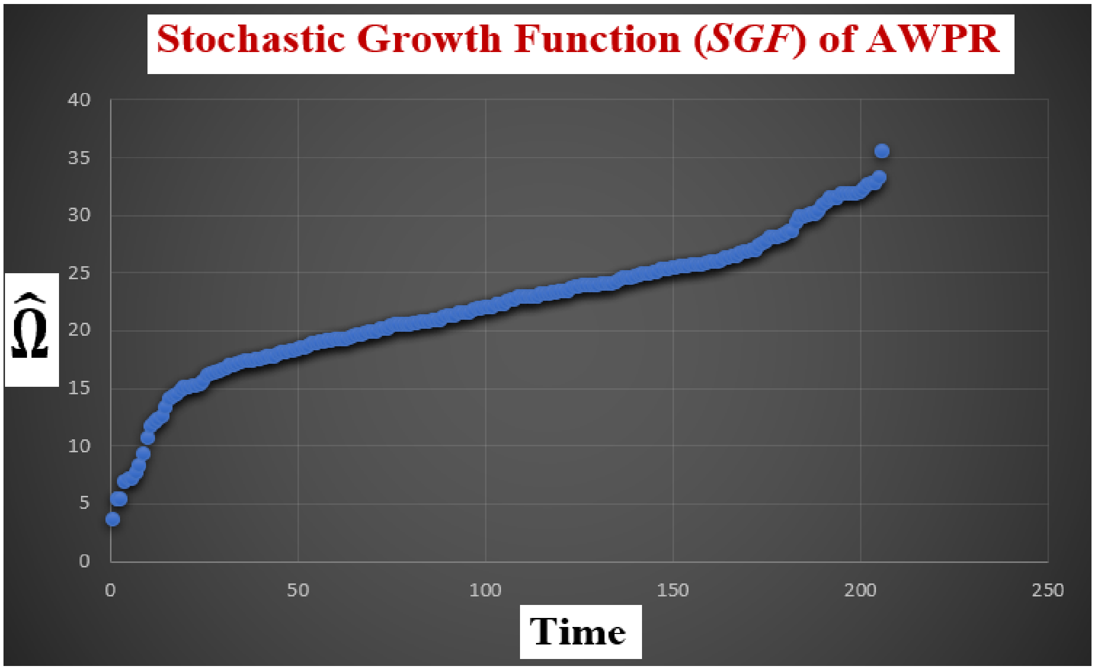


Figure 7.5: Stochastic Growth Function Curve of AWPR with Respect to Time

It is clearly shown that the SGFs $\Omega(A\hat{W}PR_t)$ of $AWPR$ is increasing with respect to time implying that the profit margin is increasing.

Table 7.4 illustrates the arrangements of $AWPR_t$ and the SGFs $\hat{\Omega}(\cdot)$ as a function of time.

Table 7.4: Describing the AWPR and SGFs as a Function of Time

Time (Week)	$AWPR_t$	$\hat{\Omega}(\cdot)$
1	1.71	3.66
2	2.4	5.35
3	2.43	5.43
4	3.01	6.9
$\vdots$	$\vdots$	$\vdots$
206	13	35.49

Now, we have the point estimate of for our index-indicator ($\mathcal{I}$), we proceed compute the 95% confidence interval of $\mathcal{I}$. $\mathcal{I}$ has a chi-square distribution with $2(n-1)$ degrees of freedom [132]. Which implies,

$$\hat{\mathcal{I}} \sim \frac{2n\mathcal{I}}{\chi^2{}_{2(n-1);\frac{\alpha}{2}}}$$

. From the above result, an exact $100(1-\alpha)\%$ confidence interval for $\mathcal{I}$ is given by:

$$\left[\frac{\hat{\mathcal{I}}}{2n}\chi^2{}_{2(n-1);\frac{1-\alpha}{2}}, \frac{\hat{\mathcal{I}}}{2n}\chi^2{}_{2(n-1);\frac{\alpha}{2}}\right] \tag{7.9}$$

where $\chi^2_{\gamma;v}$ is the $1-\gamma$ percentile of the chi-square distribution with v degrees of freedom. Plugging the estimate of $\mathcal{I}$ from Table 7.3, we obtain the 95% confidence interval of $\mathcal{I}$ to be [1.93, 2.47] which is precise. It implies that if we select observations randomly from our data for sufficiently large number of times, at least 95% of the cases, the interval [1.93, 2.47] contains the true index-indicator $\mathcal{I}$. To validate our results, we have considered values of AWPR for five random consecutive weeks and computed the estimates of $\mathcal{I}$ and ϑ, which is given in the following Table 7.5.

Table 7.5: Evaluating Five Consecutive Weeks of AWPR for Health-care Business Segment (HBS) based on $\mathcal{I}$

$\mathcal{I}$	$\hat{\vartheta}$
3.04	2.46
3.56	3.04
2.7	2.62
2.97	2.9
3.17	3.09

From the above Table 7.5, we see that for all of the five weeks, $\mathcal{I} > 1$ implying that the SGFs are increasing with respect to time.

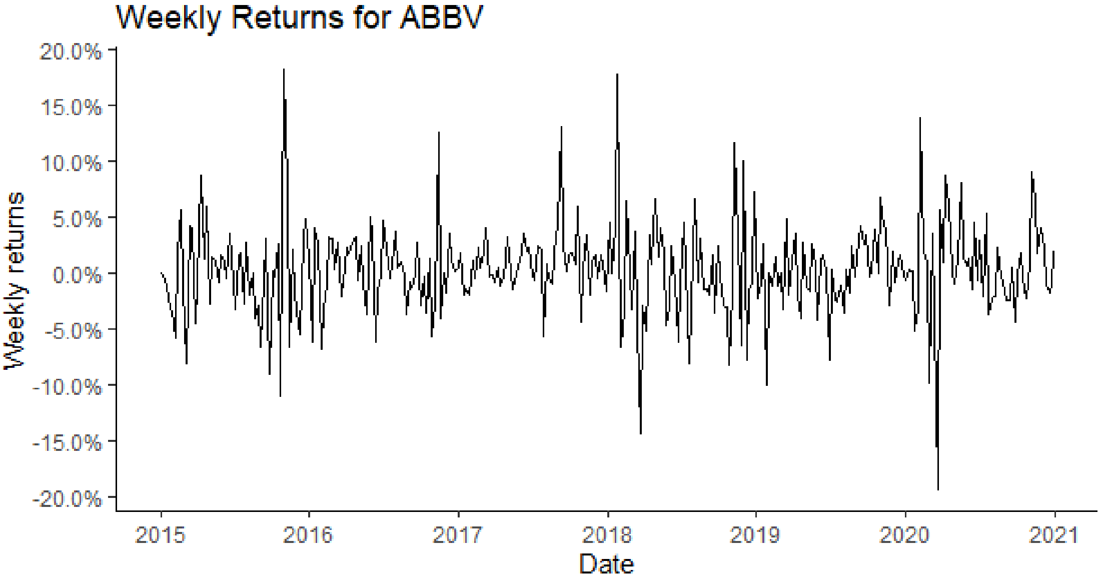


Figure 7.6: Showing Weekly Returns for ABBV

7.4 Monitoring The Behavior of a Particular Stock

In Section 7.3, we have analyzed the overall average weekly returns for the healthcare business segment (HBS) of S&P 500 and have shown that the performance is falling apart in terms of the index-indicator $\mathcal{I}$. Chakraborty & Tsokos (to be published) developed a data-driven analytical model for a particular healthcare stock (AbbVie Inc.) which was selected based on high dividend yield, high return, and low beta risk. The non-linear analytic model is based on three financial and three economic indicators. In this section, we will discuss about the process of monitoring the stock (ABBV) based on $\mathcal{I}$ as a function of time.

As the above Figure 7.6 shows, the weekly returns of ABBV has high volatility in some particular years.

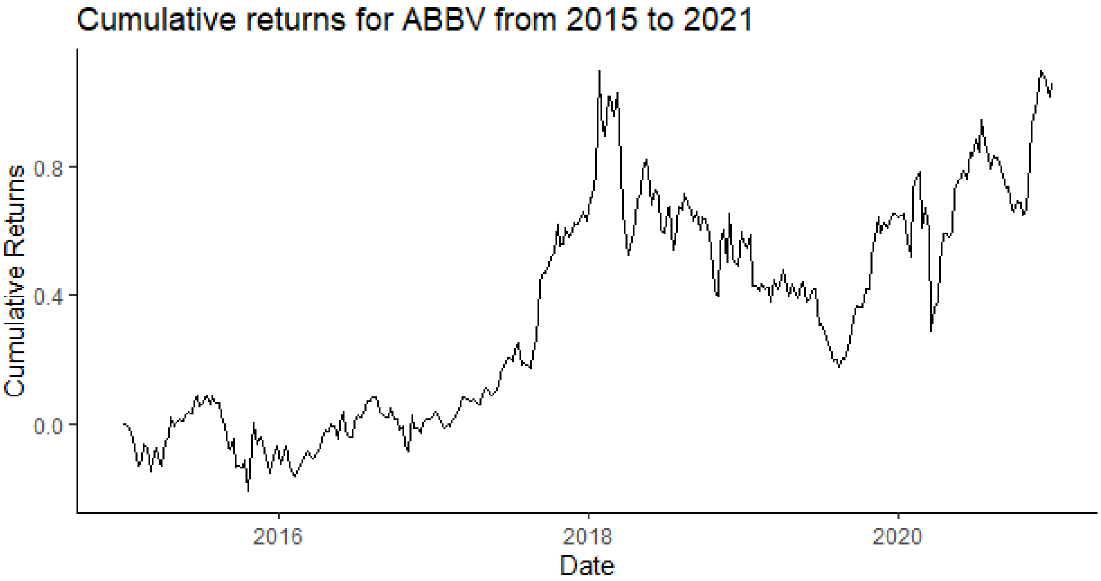


Figure 7.7: Cumulative Weekly Return of ABBV

The above Figure 7.7 illustrates that the cumulative weekly return has an increasing pattern on an average. However, there is a decreasing pattern from the beginnings of 2019 until the end of 2020. After that it increases periodically.

The stock return data for ABBV is a stochastic realization of time and we now proceed to estimate the parameters of the stochastic growth function (SGF), $\mathcal{I}$ and ϑ in a similar manner. The following Table 7.6 provides the estimates of $\mathcal{I}$ and ϑ. We see that $\mathcal{I} \approx 1$, implying that the stock ABBV is performing approximately the same/constant as a function of time.

Table 7.6: Evaluating Profit Structure of ABBV based on $\mathcal{I}$

Estimates	Values
$\mathcal{I}$	.97
$\hat{\vartheta}$	.08

Given the values of $\mathcal{I}$ and $\hat{\vartheta}$, from Table 7.6, we can estimate the SGFs for ABBV ($\hat{\Omega}(ABBV_t, \hat{\mathcal{I}}, \hat{\vartheta})$) using equation (1).

$$\begin{aligned} \hat{\Omega}(ABBV_t, \hat{\mathcal{I}}, \hat{\vartheta}) &= \frac{\hat{\mathcal{I}}}{\hat{\vartheta}}\left(\frac{ABBV_t}{\hat{\vartheta}}\right)^{\hat{\mathcal{I}}-1} \\ &= \frac{.97}{.08}\left(\frac{t}{.08}\right)^{.97-1} \\ &= 12.125\left(\frac{t}{.08}\right)^{-0.03} \quad , \quad t \geq 0 \end{aligned} \tag{7.10}$$

7.5 Conclusion

In this article We proposed an analytical method for tracking, assessing, and evaluating the performance of health care business segment (HBS) utilizing the information of S&P 500, as a stochastic realization of time. Our analytical model is based on the average weekly percentage return (AWPR) of 59 health care stocks. We have shown that our analytical method is functionally efficient and productive to monitor the ups and downs of the particular business segment. The stock return monitoring process based on the AWPR of HBS is a more robust version of the tradition stochastic models as the analytical method can model the stochastic growth function (SGF) of the profit structure of a business segment. As discussed in Section 7.3, the stochastic growth function $\Omega(\cdot)$ is a function of index-indicator ($\mathcal{I}$) that decides if a particular business segment/firm is growing in business, performing the same , or deteriorating as a function of time. In our study, we found ($\mathcal{I}$) to be 2.12 > 1 implying that the whole health care business segment is performing well (average percentage returns of the profit structure is declining) as a function of time. In other words, the SGF of AWPR of HBS is increasing. We also computed the 95% confidence interval for the index-indicator ($\mathcal{I}$) which happens to be precise in terms of capturing the true information of ($\mathcal{I}$) in long run. The resulting information of the behavior of the ($\mathcal{I}$) of proposed stochastic model is very important to the managers to make constructive and corrective decisions in monitoring the stock returns. For example, if ($\mathcal{I}$) < 1 , the average weekly returns are increasing, and hence change is required in the ongoing implemented business strategy; In our study, we have found that the pharmaceutical company AbbVie Inc. (ABBV). The $\mathcal{I} = .97 \approx 1$ for ABBV implies that the company is performing more or less the same in terms of returns, as a function of time . $\mathcal{I} = 1$ implies that the SGFs of AWPR remain unchanged (constant); and $\mathcal{I} > 1$, is the indication that the SGFs of AWPR are increasing with time. Managers may consider changes in the values of $\mathcal{I}$ that drives the profit structure to increase the revenue and necessary process adjustments are to be taken. The health care and pharmaceutical firms can be directly involved in the monitoring process since they control the significant financial indicators and the interactions. Thus, these firms under S&P 500 can determine what modifications and adaptations needs to be taken to increase the returns based on the estimated $\mathcal{I}$ of a particular time period. In a nutshell, our proposed analytical model can be implemented to any one of the *eleven* sectors of S&P 500, and any specific company of a sector, for a particular time period to monitor the profit pattern.

Chapter 8: A Real Data Driven Analytical Model to Predict Happiness

Journal Article : “A Real Data-Driven Analytical Model to Predict Happiness,” by Aditya Chakraborty & Chris P. Tsokos, Sch J Phys Math Stat, 2021 Mar 8(3): 45-61.CC-BY-NC 2021 by Copyright Holder. Used with Permission

8.1 Introduction

When we think about Happiness in modern life, we might be referring to the feeling we get after the first lick of a delectable ice cream cone or when spending quality time with some of our wonderful friends. This way of thinking about Happiness as satisfaction or amusement suggests that it is a subjective, emotional state, susceptible to the moment-to-moment experience that we are having.

Even though feeling good is a part of Happiness, some old lines of thought have defined Happiness more extensively. Specifically, Aristotle believed that the ultimate goal of human life was a notion of ancient Greeks called *eudaimonia*. The word is often translated as *Happiness*, but more likely means ”human flourishing” or ”a good life.” Being happy is not only associated with personal well-being but also with productivity on a large scale. Studies have been performed to understand the association between Happiness and productivity[9]. A happy mind is also associated with sound mental health. Health and Happiness are essential and possibly related to the pursuits of mankind. Sound health may play a vital role in determining the Happiness or, morbidness/sickness may cause unhappiness. Conversely, a feeling of Happiness may strengthen health conditions[103]. Numerous studies on Happiness has been done by social researchers concentrating on psychological and social causal and cognitive factors of Happiness. For instance, Happiness is routinely keep under surveillance in sociological surveys[95], and levels of Happiness have been connected to individual personality and idiosyncrasy[41], living conditions, dignity and morale , love, democracy [103] [104] [21] [49], and also with brain activity of specific individual[53]. Some studies have investigated Happiness concerning health in a widespread population. In an epidemiological study of Finnish men, it has been found that life satisfaction (measured through four items assessing whether life is interesting, happy, easy, or lonely) predicts lower mortality[83], but the specific contribution of Happiness was not reported. In the medical literature, the Happiness is often considered a contributing factor of good mental health. For instance, the mental health scale embedded in the Short Form-36 (SF-36) questionnaire includes an item on Happiness, [13], one item from the Bradburn scale of well-being asks whether the respondent is 'depressed or very unhappy' [17], and the validity of the Happiness-Depression scale was tested against a mental health questionnaire[136]. Some studies also found that the effect of the nationality of levels of Happiness may capture the impact of cultural integration on people's well-being[103]. Using an international cross-section of 28 countries, researchers have found a highly significant impact of democracy on the subjective well-being of people[44]. Thus, Happiness and democracy, as one would expect, are highly correlated.

In general, personal Happiness and well-being seem to the principal objective of human life. Throughout history, the virtue of Happiness has been considered as the ultimate end of temporal existence. Aristotle's ancient view about Happiness was *"Happiness is so important, it transcends all other temporal considerations"*. Aristotle's prescription for spending a good life was to exercise virtues like being kind, humble, wise, and honest in our actions consistently. In other words, accomplishing different physical and emotional needs, is the recipe for a happy life.

From our study, we found **Finland** being number one, followed by Denmark. The U.S is **fifth** and Romania being **54th**. The proposed model offers other useful information on the subject area. Our analytical model has been validated and tested to be of high quality, and our prediction of happiness is with a high degree of accuracy.

While we build the analytical model, we have the national average of happiness score as the response variable; hence, we proceed to develop an analytical model containing significant risk factors and other significant interactions. The proposed non-linear statistical model is based on several assumptions, such as linearity, multicollinearity, homoscedasticity, and different assumptions concerning statistical methodology. The dataset shows that some of the risk factors are highly correlated, as shown in Figure 8.3. The parameters of the models become difficult to interpret under the influence of multicollinearity. The parameters also become very unstable when independent variables are highly correlated, which leads to over-fitting the model. Moreover, we use different penalization regression methods: Ridge Regression (L_2), Lasso Regression (L_1), and Elastic net (EN) [130]. These machine-learning techniques are vastly used in applied sciences to address several ill-factors of the model (such as over-fitting) . Our proposed statistical model is useful in predicting individuals' Happiness, given the values of the significant risk factors. Also, we ranked the risk factors in accordance with their percentage of contribution to the happiness score. The validation and quality of our proposed analytical model have been statistically evaluated using R square (R^2), R square adjusted (${R^2}_{adj}$), Mean absolute deviation (MAD), root mean square error (RMSE), and residual analysis. The advantages of using this model has been discussed in the conclusion section. To the best of our knowledge, no such statistical model has been developed under the proposed logical structure to predict Happiness for developing countries. Therefore, searching for an proper data-driven analytical model in the prediction of Happiness is important.

8.2 Methodology

8.2.1 The Data

The World Happiness Report is a landmark survey of the state of global happiness that ranks descriptively 156 countries by how happy their citizens perceive themselves to be. The data has been obtained from the World Happiness Report 2019 website[64], where they used the **Gallup Poll** to get the answers to specific questions. The data has been collected for a total of 156 countries from 2005 to 2018. However, in our study, we only considered the data of **developed countries**(sorted based on the human development index[**HDI**]) in the world. Individuals were asked specific questions, and as a result of their response as a whole, a score was produced, which is termed as the national average. In our data, the average scores of the developed countries from 2005 to 2018 were tabulated. One of the main goals of our study is to understand what attributable variables significantly affect the happiness of an individual. We have eleven attributable variables and the **Ladder** (which is also called subjective well being [SWB] or happiness score as a measure of response. **For example, let there be an imaginary ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life, and the bottom of the ladder represents the worst possible life of an individual. On which step of the ladder is an individual standing currently is reported. This measure is also referred to as the *Cantril life ladder* or just life ladder in our analysis.**

The attributable variables (risk factors) that the data was collected on are given below. The descriptions of the risk factors are the same as provided in the world happiness report 2019.

- **LOG_GDP**(X_1)(Log GDP): Per-capita gross domestic product(in logarithmic scale) in purchasing power parity(PPP).

- **SOC_SUPPORT**(X_2)(Social Support): This variable is defined as is the national average of the binary responses (either 0 or 1) to the GWP question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

- **LIFE_EXPECT**(X_3)(Life Expectancy): Healthy life expectancies at birth are based on the data extracted from the World Health Organization's (WHO) Global Health Observatory data repository.

- **FREEDOM**(X_4): Freedom to make life choices is the national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

- **Generosity**(X_5): Generosity is the residual of regressing national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.

- **PER_CORR**(X_6)(Perception of Corruption): The measure is the national average of the survey responses to two questions in the GWP: "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?" The overall perception is just the average of the two 0-or-1 responses.

- **POS_AFFECT**(X_7)(Positive Affect): Positive affect is defined as the average of three positive affect measures in GWP. These are happiness, laughter, and enjoyment in the Gallup World Poll.

- **NEG_AFFECT**(X_8)(Negative Affect): Negative affect is defined as the average of three negative affect measures in GWP. These are worry, sadness, and anger, respectively.

- **CONF_GOV**(X_9)(Confidence in Government): How much trust and confidence does one have in government when it comes to handling [International problems/Domestic problems] – a great deal, a fair amount, not very much or none at all?

- **DEM_QUALITY**(X_{10}): Democratic quality is the National average of the first two dimensions of World Governance Index **(WGI)**[78] namely,*voice and Accountability* and *Political Stability and Absence of Violence/Terrorism.*

- **DEL_QUALITY**(X_{11}):Delivery quality is the National average of the last two dimensions of World Governance Index **(WGI)** namely, *Government Effectiveness*, *Regulatory Quality*, *Rule of Law* and *Control of Corruption.*

The definitions of the above-mentioned measures under **DEM_QUALITY** and **DEL_QUALITY** (which are also the six dimensions of the World Governance Quality Index **(WGI)** are as follows:

1. Voice and Accountability: Voice and accountability captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.

2. Political Stability and Absence of Violence/Terrorism: Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically motivated violence, including terrorism.

3. Government Effectiveness: Government effectiveness captures perceptions of the quality of public services, the quality of the civil service as the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.

4. Regulatory Quality: Regulatory quality captures perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.

5. Rule of Law : Rule of law captures perceptions of the extent to which government agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence.

6. Control of Corruption : Control of corruption captures perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests.

From Figure 8.1 below, we see that there are some missing observations in the data set. However, the proportion of missing values is small; we used **predictive mean matching (pmm)** algorithm to perform multiple imputation to our dataset. Predictive mean matching (PMM) is a useful technique to perform multiple imputation [110] for missing data points in a plausible manner, especially for imputing quantitative variables that are not normally distributed.

While the development of proposed analytical model to predict happiness score as a function of several risk factors, one of the most important assumptions is the normality of response (dependent variable). That is, the response variable Ladder should follow the

Gaussian probability distribution. The mid-values of happiness score seem to be reasonably straight, but the ends are skewed to a certain degree, as can be seen from the Q-Q plot shown by Figure 8.3.

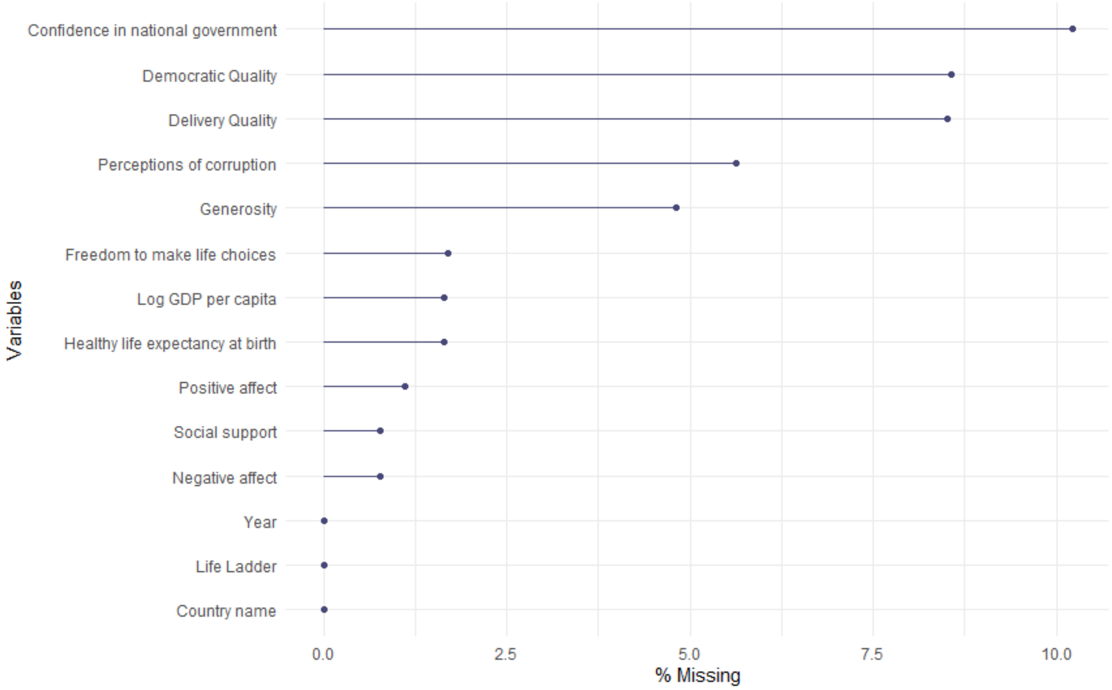


Figure 8.1: Showing The Distribution Of Missing Values in Happiness Data

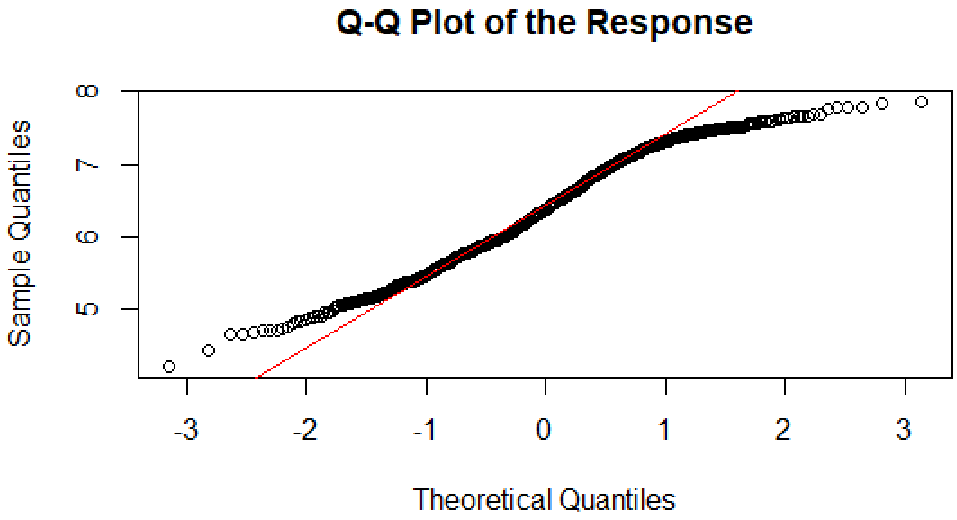


Figure 8.2: Q-Q Plot Of The Response Ladder

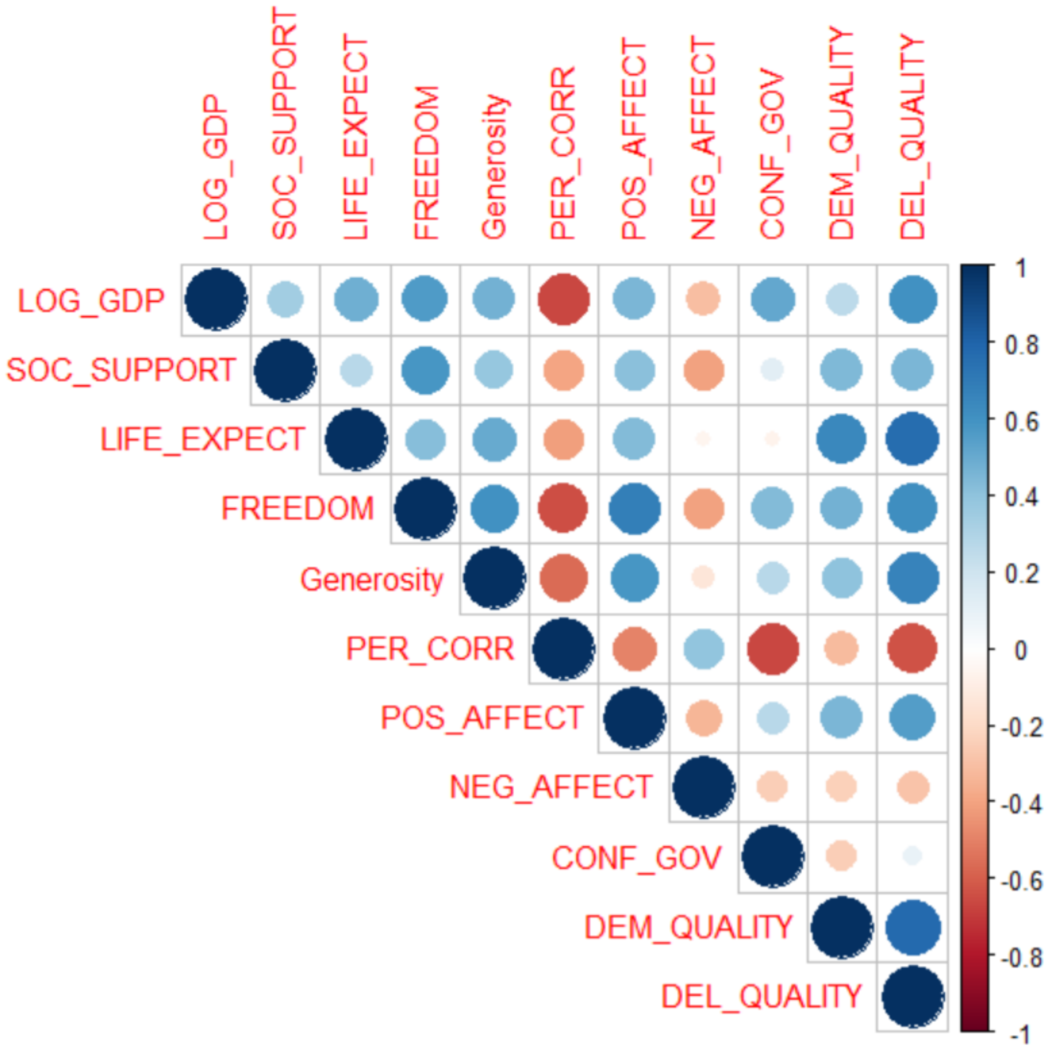


Figure 8.3: Correlation Matrix of The Attributable Variables

We have also shown through goodness-of-fit test (Shapiro-Wilk normality test, p-value $= 5.565 \times 10^{-10}$) that the response Ladder does not follow the normal probability distribution. Thus, the Q-Q plot supports that the national average of happiness scores, do not follow the Gaussian probability distribution. The correlation plot of of the risk factors is shown in Figure 8.2.3, where negative correlations are presented in red and positive correlations in blue color. The color intensity and the degree of relationship between each pair of risk factors are proportional to the correlation coefficients. From the following correlation matrix in Figure 8.4, we see that there are strong positive associations between the variables LIFE_EXPECT and DEL_QUALITY, Generosity, and DEL_QUALITY and DEM_QUALITY and DEL_QUALITY. Also, there is a strong negative association between the variables LOG_GDP and PER_CORR and PER_CORR and CONF_GOV. Thus, we would implement some regularization techniques such as Ridge Regression (L_1), Lasso Regression (L_2), and Elastic net regressions to take into account the over-fitting issue and compare their performance in terms of $RMSE$ and MAE.

8.2.2 Development of Statistical Model

We now start developing the non-linear analytical model, which is driven by the national average of happiness score as a function of the eleven risk factors and all possible interactions, as discussed previously. The general structure of our non-linear model with all possible interactions and additive error structure, is given by:

$$Ladder = \beta_0 + \sum_i \alpha_i x_i + \sum_j \gamma_j k_j + \epsilon \quad , \tag{8.1}$$

where β_0 is the intercept term of the model, α_i is the coefficient of i^{th} individual risk factor x_i, γ_j is the coefficient of j^{th} interaction term k_j and ϵ is the random error term of the model, that follows a normal distribution with zero mean and constant variance.

One of the main suppositions to develop the above model is that the response variable should follow the Gaussian probability distribution. As we have shown above, the dependent variable Ladder does not follow the Gaussian probability distribution initially. Therefore, we utilize a non-linear transformation to filter our happiness data so that it follows the normal probability distribution. We used Johnson transformation for our response, which is given by:

$$z = \gamma + \delta ln\left(\tfrac{x-\epsilon}{\lambda+\epsilon-x}\right) \quad , \qquad \epsilon < x < \epsilon + \lambda$$

and

$$TLadder = -0.43 + 0.87ln\Big(\frac{x - 4.2}{3.72 + 4.2 - x}\Big) \quad . \tag{8.2}$$

Here, *T Ladder* denotes the new response variable(transformed) after the use of Johnson S_U transformation to our old response. We now estimate the coefficients (weights) of the risk factors for the processed data in equation 8.2. To develop our analytical model, we initially proceed with the full statistical model, including all eleven risk factors and ten possible interactions between each pair. Thus, initially, we start structuring our model with $\binom{n}{k} = 55(n = 11, k = 2)$ terms that include the primary contribution of the risk factors and every possible interactions.

As we began with the full statistical model (fifty-five terms), as mentioned, we have applied the backward elimination method to identify the most significant contributions of both the individual attributable variables and interactions by eliminating the less important risk factors gradually.

Furthermore, backward elimination is deemed one of the best traditional methods for a set of feature vectors to encounter the problem of overfitting and carry out feature selection.

Though, the statistical estimation method of our data analysis has indicated that only seven out of the eleven risk factors significantly contribute and twenty-eight interaction terms, we can not omit the risk factors that are not significant, and simultaneously include any risk factor interacting with it in the model.

Thus, the best proposed statistical model with all risk factors and significant interactions that estimates the average happiness score accurately are eleven risk factor individually, and the twenty-eight interaction term, which is given by:

$$\widehat{TLadder} = \begin{cases} -0.45 + 0.42X_1 + .12exp(X_2) + .04X_3 + .27X_4 + \\ .16X_5 + .03exp(-X_6) + .12exp(X_7) - .07X_8 + .03X_9 \\ -.07X_{10} - .1X_{11} - .17X_1X_3 + .14X_1X_4 - .13X_1X_5 + \\ .27X_1X_6 - .03X_1X_7 - .11X_1X_8 + .22X_1X_{11} - .1X_2X_5 \\ +.19X_2X_6 + .14X_2X_8 + .22X_2X_9 + .16X_2X_{10} + .15X_2X_{11} + \\ .07X_3X_7 - .06X_3X_{10} - .43X_4X_6 - 0.19X_4X_8 - 0.29X_4X_9 + \\ 0.10X_4X_{10} - 0.30X_4X_{11} + 0.18X_5X_6 + 0.11X_5X_9 - .2X_5X_{10} + \\ .32X_5X_{11} - .02X_6X_{11} + 0.1X_7X_8 + 0.1X_7X_9 + 0.1X_8X_{11} \end{cases} \tag{8.3}$$

The $\widehat{TLadder}$ can be computed from equation (3) and is based on the Johnson transformation[106] of the data. We now proceed to utilize the anti-transformation on equation (8.3) to estimate the actual estimate national average of happiness score $\widehat{Ladder}$ as follows:

$$\widehat{Ladder} = 4.2 + \frac{3.72}{1 + exp\left(\frac{\widehat{TLadder}+0.43}{0.87}\right)} \quad . \tag{8.4}$$

The proposed analytical model will assist social scientists and economists acknowledge how the happiness score changes when any of the eleven risk factors is varied by keeping the other risk factors fixed at the same time.

Likewise, with the variation of significant interaction. Anyone, interested to know the optimum levels of the risk factors at which the happiness score is maximized, can do the same by using any analytical optimization technique.

We now illustrate the percentage of contributions of the risk factors and the interactions to the happiness score as shown below in Table 8.1.

Table 8.1: Ranking of Individual Risk Factors and the Interactions With Respect to The Percentage of Contribution to The Response

Rank	Risk Factors	Contr.(%)
1	LOG_GDP	7.15
2	$FREEDOM \cap PER\_CORR$	5.58
3	$LOG\_GDP \cap PER\_CORR$	5.00
4	$FREEDOM$	4.94
5	$EXP(POS\_AFFECT)$	4.63
6	$FREEDOM \cap CONF\_GOV$	4.46
7	$FREEDOM \cap NEG\_AFFECT$	4.13
8	$CONF\_GOV \cap SOC\_SUPPORT$	3.89
9	$NEG\_AFFECT \cap SOC\_SUPPORT$	3.72
10	$GENEROSITY$	3.72
11	$FREEDOM \cap DEL\_QUALITY$	3.59
12	$GENEROSITY \cap DEL\_QUALITY$	3.45
13	$EXP(SOC\_SUPPORT)$	3.30
14	$GENEROSITY \cap DEM\_QUALITY$	3.16
15	$LOG\_GDP \cap DEL\_QUALITY$	2.96
16	$PER\_CORR \cap SOC\_SUPPORT$	2.87
17	$LOG\_GDP \cap LIFE\_EXPECT$	2.55
18	$POS\_AFFECT \cap NEG\_AFFECT$	2.45
19	$LOG\_GDP \cap NEG\_AFFECT$	2.44
20	$CONF\_GOV \cap POS\_AFFECT$	2.38

To evaluate the quality of the proposed analytical model , we use both the coefficient of determination, R^2, and adjusted R^2, which are the basic criteria to evaluate the model performance. The sum of squares due to regression(SSR) is the squared sum of the differences between the predicted response and the mean response. It captures the observed variability of the model. The sum of squared errors (SSE), also termed as the residual sum of squares, is the variation that remains unexplained. We always try to minimize this error in a model The total sum of squares (SST) = SSE + SSR. R^2, the coefficient of determination, is defined as the proportion of the total response variation that is explained by the proposed model, and it measures how well the regression process approximates the real data points. Thus, R^2 is given by

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad .$$

However, R^2 itself does not consider the number of variables in the model. Also, there is the problem of the increasing R^2 with addition of variables in the model. To address these issues, we have the adjusted R^2, which considers the number of parameters and is given by

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \quad ,$$

where n is the number sample data points, and k is the number of independent risk factors used in the model, excluding the constant. For our final statistical model, the R squared is 88.8%, and R squared adjusted is 87.8%. Both R squared and R squared adjusted are very high and very close to each other. That is, the developed statistical model explains 88.8% of the variation in the response variable, a very high-quality model. Similarly, the risk factor that we included in the model, along with the relevant interactions, estimates almost 89% of the total variation in the happiness score.

8.2.3 Verifying Model Assumptions

Once the statistical model has been developed, it is necessary to check the model assumptions (if any). In our case, we have proposed a multiple non-linear regression model, which is very useful and conveys to us accurately some important information on the subject matter. However, multiple linear regression has some important assumptions which must be satisfied with the correctness of the proposed model. In this section, we will verify the important model assumptions.

8.2.3.1 Mean Residual should be Close to Zero

When one performs multiple linear regression (or any other type of regression analysis), one obtains a linear function that best fits the data. The entire data points usually don't fall exactly on this regression plane, but they are scattered around it. The residual(error)$\hat{\epsilon}$ is defined as:

$$\hat{\epsilon} = \text{residual} = \text{observed value-predicted value} = y - \hat{y} \quad ,$$

where y and $\hat{y}$ are the observed and predicted response. $\hat{e}$ is the estimated residual error from the linear fit. The sum of the residuals equals zero, assuming that the regression function is actually the "best fit."In our case, the mean residual is $-1.56 * 10^{-18}$, implying that it is almost zero. Figure 8.4 below illustrates the behavior of the residual estimator.

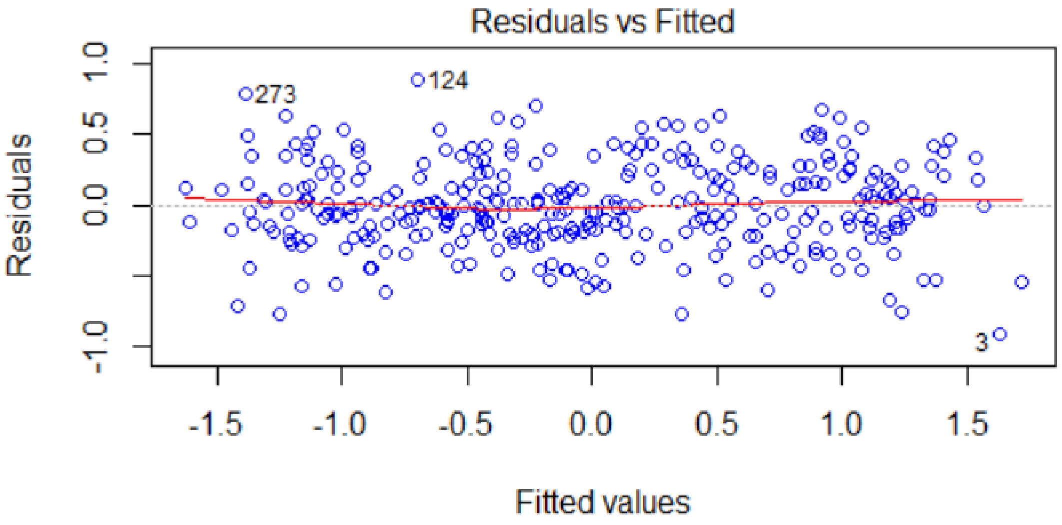


Figure 8.4: Fitted Vs. Residual Plot

8.2.3.2 Homoscedasticity of residuals

One of the main assumptions of the linear regression model is the homoscedasticity of the residuals or equal variance. That is, $Var(\hat{e}) = \sigma^2$ which is constant. From the above Figure 8.4, we see that residuals vary as the fitted values increase. It seems that the pattern is more or less uniform, which is shown by the red line. There is no increasing or decreasing trend. Hence, the assumption of the constant variance of the residuals has been satisfied.

Breusch-Pagan Test: Breusch-Pagan (BPG) test is used to test for heteroskedasticity of the error terms in a regression model. We obtained a p-value of .35173 by testing the null hypothesis of constant error variance against the alternative that the error variance changes with the level of the response (fitted values) or with a linear combination of predictors. Hence, we have significant reason to believe the error variance is constant.

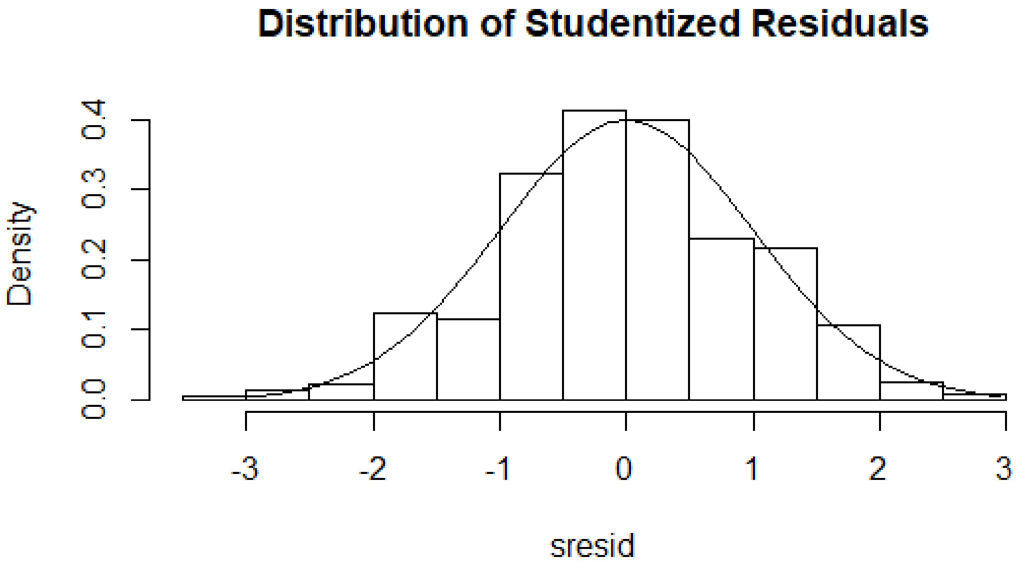


Figure 8.5: Normality of Studentized Residual(sresid) Plot

8.2.3.3 Normality of residual

One important assumption of linear regression is normality of residual. From Figure 8.6, and Figure 8.7, we see that the studentized residual follows a normal pattern.

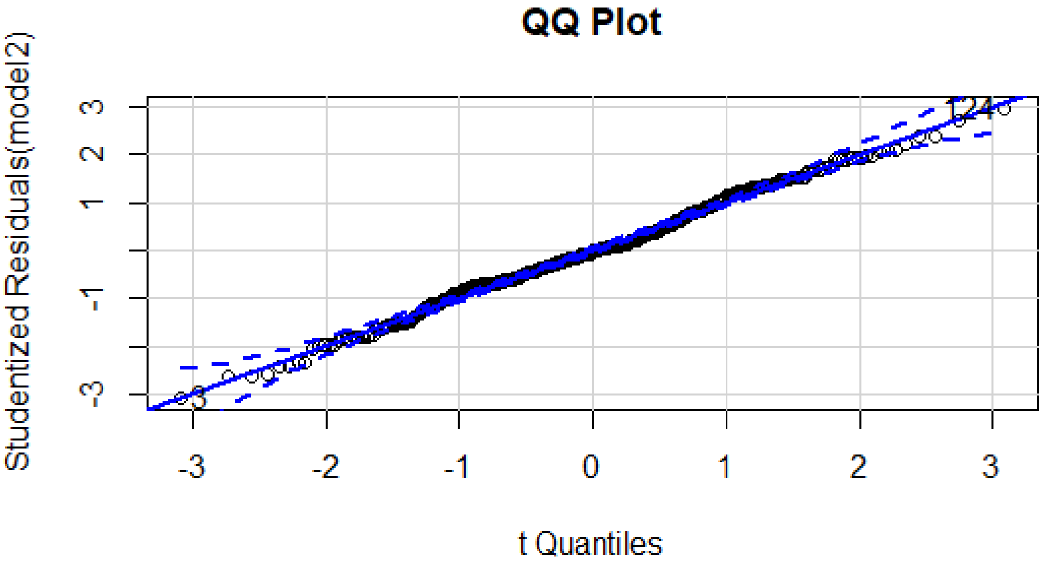


Figure 8.6: Q-Q Plot of Studentized Residuals

8.2.3.4 No auto-correlation between the residuals

We proceed to test the auto-correlation between the error terms of our model. The correlation between two error terms is defined as,

$$corr(\hat{e_i}, \hat{e_j}) = \begin{cases} 0, & \text{if } i \neq j. \\ 1, & \text{if } i = j. \end{cases} \tag{8.5}$$

where $\hat{e_i}$ and $\hat{e_j}$ are the i^{th} and j^{th} error terms in the model.

The following Figure 8.7 shows the autocorrelation of the residuals vs. lag plot. The X-axis corresponds to the lags of the residuals. The first line to the left shows the correlation of residuals with itself (Lag0); therefore, it will always be equal to 1. If the residuals were **not auto-correlated**, the correlation (Y-axis) from the immediate next line onwards would drop to a near-zero value below the dashed blue line (significance level). Hence, there is no auto-correlation between residuals in our model.

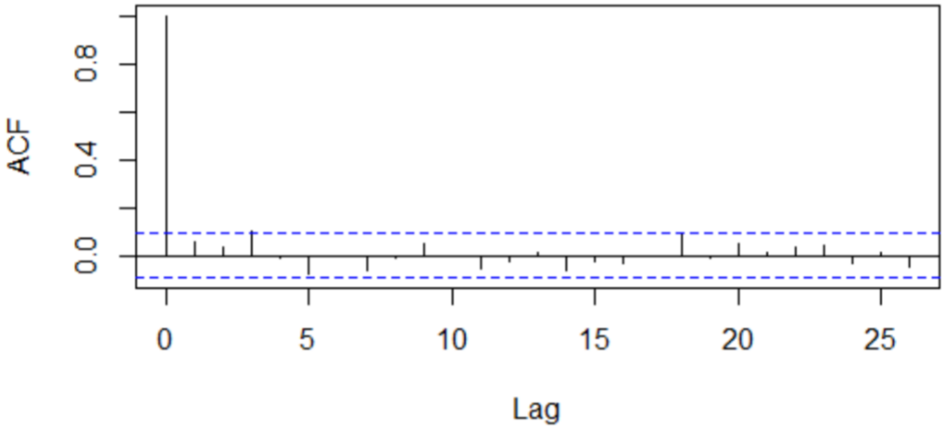


Figure 8.7: Showing The Auto-Correlation Plot of Residuals

RUN TEST: Also, we can verify the no auto-correlation case of the residuals by Run test (Wald, A. and Wolfowitz, J. (1940)[135]. Runs test examines the randomness of a numeric sequence by studying the frequency of runs. We obtained a p-value of 0.9264, which implies that we fail to reject the null hypothesis that residuals are random. Hence, there is no pattern.

Durbin-Watson test: The Durbin Watson Test[137] is a measure of auto-correlation (also called serial correlation) in residuals from the regression analysis. Auto-correlation is the similarity of a time series over successive time intervals. It can underestimate the standard error and can cause one to believe that the predictors are significant when they are not. The Durbin–Watson test statistic is used to detect the presence of autocorrelation at lag 1 in the residuals (also termed as prediction errors) in regression analysis. The test statistic for this test is given by:

$$DW = \frac{\sum_{t=2}^{T} \left(\hat{e_t} - \hat{e_{t-1}}\right)^2}{\sum_{t=2}^{T} \hat{e_t}^2} \quad ,$$

where $\hat{e_t}$ and $\hat{e_{t-1}}$ are the residuals at time points t and $t-1$, respectively.
A rule of thumb is that the test statistic values in the range of 1.5 to 2.5 are relatively normal. Values outside of this range could be a cause for concern. Field(2009) suggests that values under 1 or more than 3 are a definite cause for concern. The value we obtained for the test statistic is 1.89 with a p-value of .109, implying that there is insufficient sample evidence to reject the null hypothesis that the true auto-correlation in zero.

5. The regressors and the residuals are nor correlated: We calculated the Pearson's product-moment correlation coefficient between each regressor and the residuals. As expected, every time we obtained an insignificant p-value implying that the true correlation is zero.

We further studied the fact that if there are other statistical models that give better useful results than the proposed nonlinear regression model. Thus, we developed some penalized regression models and compared those with our proposed model. These models are given in the following section.

8.3 Penalized Regression Models

Penalized regression methods have proven to be a high-yielding area of research in statistics and data sciences. The key idea is to add a 'penalty' to regression to encourage desirable behavior in the model. Often this is done to reduce variability in estimating the parameters. While developing the proposed statistical model for happiness, we used OLS, the ordinary least square technique to obtain an approximate estimate of the coefficients (weights) of the attributable variables. To address the multicollinearity problem (since in our data set, some variables are strongly correlated), the Regularization methods are used. Since these methods are based on adding the regularization parameters(lambda and alpha) to the regression coefficients of the individual risk factors, these the model generalizes the data and prevents over-fitting. To further illustrate our proposed model's quality, we will discuss three machine learning regularization methods and our proposed non-linear analytical model.

8.3.1 Ridge Regression

For multiple linear regression, the ordinary least squares fitting procedure of the coefficient estimates(weights) $\beta_1, \beta_2, ...\beta_p$ that minimizes the cost function RSS (Residual Sum Of Squares), is given by,

$$RSS = \textstyle\sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 \quad .$$

Ridge regression is very similar to least square regression, except that the ridge coefficients are estimated by minimizing a slightly different quantity. In particular, ridge regression coefficient estimates $\hat{\beta^R}$ are the values that minimizes the following function:

$$\sum_{i=1}^{n} \big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\big)^2 + \lambda \sum_{i=1}^{p} {\beta_j}^2 \quad , \tag{8.6}$$

where $\lambda \geq 0$ is a tuning parameter (sometimes called a penalty parameter that controls the strength of the penalty term in ridge regression) to be determined via cross validation.

8.3.2 LASSO (Least Absolute Shrinkage and Selection Operator) Regression

The LASSO regression model appends an absolute value of magnitude of a coefficient as penalty term to the loss function that is given by:

$$\sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\Big)^2 + \lambda \sum_{i=1}^{p} \mid \beta_j \mid \quad . \tag{8.7}$$

Comparing (8.6) to (8.7), we see that the LASSO and Ridge regression have similar formulations. The only difference is that the β_j^2 term in the ridge regression penalty in (6) has been replaced by $\mid \beta_j \mid$ in the LASSO penalty (6). In statistical literature, the LASSO uses an L_1 penalty where the Ridge uses L_2 penalty.

8.3.3 Elastic Net

Elastic Net regression model is the combination of Ridge and LASSO regression methods. The loss function of elastic net model can be defined by:

$$\sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\Big)^2 + \lambda\Big[(1-\alpha)\sum_{i=1}^{p}{\beta_j}^2 + \alpha\sum_{i=1}^{p} \mid \beta_j \mid\Big] \quad . \tag{8.8}$$

However, in the above equations (8.6, 8.7, and 8.8) the constructions of the three models will be the same structure as our proposed model in equation (1) with only the coefficient estimation will be different because of the randomness in selecting the training data set.

8.4 Comparison among different Models

We now proceed to compare the performance of the proposed model with the other three models using the following two matrices.

8.4.1 Root Mean Squared Error (RMSE)

After each repetition of the cross-validation, the model assessment metric RMSE is computed, which is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\Big(y_i - \hat{y_i}\Big)^2}{n}} \quad ,$$

where y_i and $\hat{y_i}$ are the observed and predicted responses.

8.4.2 Mean Absolute Deviation (MAE)

The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction which is given by

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|}{n} \quad ,$$

where y_i and $\hat{y_i}$ are the observed and predicted responses.

While comparing the proposed model with the three regularization methods Ridge, LASSO, and Elastic Net, we have found that our proposed analytical model performs better in terms of validations matrices RMSE and MAE, as described above. Table 8.2 below provides multiple comparisons among the different models in terms of training and testing accuracy.

Table 8.2: Comparison Among Different Models in terms of RMSE and MAE

Table of Comparison				
	RMSE		MAE	
	Training	Testing	Training	Testing
Proposed Model	**.31**	**.43**	**.24**	**.31**
RIDGE	.38	.5	.3	.35
LASSO	.36	.52	.27	.37
EN	.36	.52	.29	.37

From the above Table 8.2, we see that our proposed nonlinear statistical model gives minimum testing error in terms of RMSE and MAE when compared with the penalized regression models. Thus, our analytical model outperforms the other three models for our happiness data.

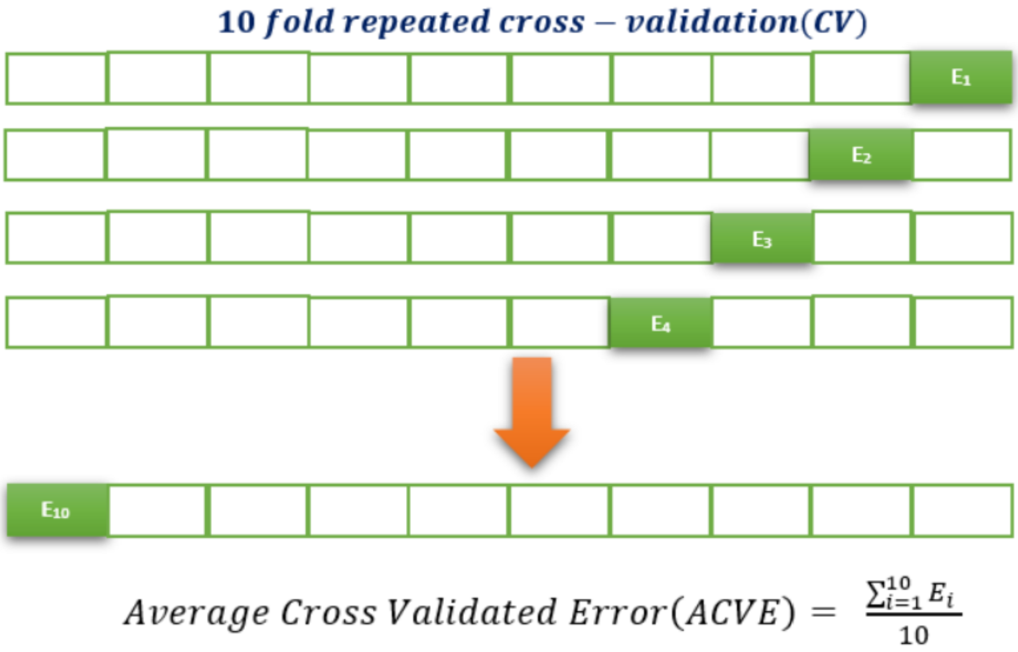

10 fold repeated cross – validation(CV)
E1
E2
E3
E4
E10
Average Cross Validated Error(ACVE) = $\frac{\sum_{i=1}^{10} E_i}{10}$

Figure 8.8: Brief Illustration Of Repeated Ten Fold Cross Validation

8.5 Validation and Prediction Accuracy of The Proposed Model

We developed our analytical model on 80% training data and validated the model based on 20% testing data. In the testing data (validation data), the test error is the average error that occurs from using the analytical method to predict the response on a new set of observations. That is a measurement that was not used in training the method. The test error gives an idea about the consistency of the analytical model. Moreover, we performed repeated ten-fold repeated cross-validation (10 times) for our validation testing. The primary objective is that we will use 10-fold cross-validation, then we repeated cross-validation ten times, where each of the repetition folds are split differently. In 10-fold cross-validation, the training set is divided into ten equal subsets. One of the subsets is taken as the testing set in turn, and (10-1) = 9 subsets are taken as a training set in the proposed model. The error mean square error E_1 is computed for the held out set. This procedure is repeated ten times; each time, a different group of observations is treated as a validation set. This process results in 10 estimates of the test error, $E_i, \quad i = 1, \ldots 10$. The average error of each set throughout the cross-validation process is termed as a cross-validated error. The following Figure 8.8, illustrates briefly the idea of 10 fold repeated cross-validation, where $E_i, \quad i = 1, \ldots 10$ is the mean square error (MSE) in each iteration and ACVE is the average cross-validated error.

Now we employ the following three methods to illustrate the prediction accuracy of the proposed model.

8.5.1 Min-Max Accuracy

Min-Max-Accuracy is the average of the ratio of minimum value between the actual observation and predicted observation and maximum between actual observation and predicted observation. Mathematically, it can be expressed as follows:

$$\text{Min} - \text{Max} - \text{Accuracy} = \text{mean}\left[\frac{min(y_i, \hat{y_i})}{max(y_i, \hat{y_i})}\right] \quad ,$$

where y_i and $\hat{y_i}$ are the observed and predicted response.
It gives an idea about how far the model's prediction is off on an average. For a perfect model, this measure is 1.This can be taken as the accuracy of the proposed model. For our developed model, the Min-Max accuracy is **96.2%**, which is quite impressive.

8.5.2 Correlation Accuracy

A simple correlation between the original observations and predicted observations can be used as a form of accuracy measure. A greater correlation accuracy implies that the original and predicted observations have analogous directional movement, i.e., when the original observations increase, the predicted observations also increase and vice-versa. We obtained a correlation accuracy of **90.5%** in the test data, which implies that our statistical model is of high quality and should be useful for applied predictive analysis for real data.
Table 8.3 below provides the two measures of prediction accuracy for our proposed model.

Table 8.3: Prediction Accuracy for the Proposed Model

Min-Max-Accuracy	Correlation Accuracy
96.2%	90.5%

Thus, the above two methods attest to the high quality of our proposed model.

8.6 Discussions

After obtaining the significant risk factors along with their significant interactions, we rank them with respect to the percent of contribution to the happiness scores for the developing countries as shown in Table 8.1. The risk variable that has the largest contribution to the happiness score is the variable **LOG_GDP** which contributes 7.15% of the total variation to the happiness score. The next largest contribution is the combined effect of freedom and perception of corruption with a 5.58% contribution. Numbers 3, 4, and 5 are the combined interaction effect of LOG_GDP, FREEDOM, and exp(POS_AFFECT) with a contribution of 5%, 4.94%, and 4.63%, respectively. Hence, adding these risk factors up, we see that they explain almost 89% of the total variability in the national average happiness score for all developing countries. We can address the usefulness and importance of the proposed model in the subject area in **five** important categories.
These categories are given below.

1. We have identified and tested the individual attributable variables(risk factors) responsible for the change in happiness score across all the developed countries.

2. we have identified the significant interactions that influence the happiness score in our model.

3. we have ranked the individual risk factors and interactions as a percentage of contribution for the the response of the national average of happiness score (Ladder) or subjective well-being (SWB).

4. We can obtain excellent predictions of happiness of individuals given the values of the attributable variables from our analytical model with a high degree of accuracy.
5. Any particular country might use our non-linear statistical model to work on specific risk factors to increase their happiness score. For example, one can work on the variable SOC_SUPPORT (X2) if the value for a particular year is not satisfactory and work on other important aspects to increase the value so that the happiness score can be increased.

We have also ranked all the developed countries based on the **predicted happiness score** of the most recent observations (data) available for the year 2019. The following Table 8.4, illustrates the ranking of the countries.

It is interesting to note that Finland and Denmark possess the top happiness scores while the *United States* is fifth. Also, studies[44] has shown a significant influence of democracy on an individuals' subjective well-being (happiness). Finland and Denmark falling into the category of the top democratic countries of the world also validate the fact.

8.7 Conclusion

We have developed a real data-driven analytical model that very accurately identifies the following very useful findings concerning the happiness of the society of developed countries in the world:

- Identifies the significant attributable variables (risk factors) that drives the degree of happiness.
- Identifies the significant interactions of the risk factors that contribute to the degree of happiness.
- We rank the individual and interactions of the risk factors with respect to their percentage of contribution to the degree of happiness.

- The developed analytical model predicts the degree of happiness very accurately for a given response to a set of questions.

- The developed model can be used strategically to increase the degree of happiness by working with the identified risk factors.

- Furthermore, one can perform surface response analysis to identify the target values of the risk factors so as to be, say, 95 percent sure that we will maximize the degree of happiness based on the identified values.

The developed analytical model has been evaluated by several statistical methods that include the R^2 and $R^2_{adjusted}$ that attest to its high quality. The risk factor **LOG_GDP** is the highest contributor to the happiness score contributing 7.15%, while **DEL_QUALITY** contributes the least with 1.31% to the response. The findings of our study suggest that economists and other social scientists might need to pay more attention to emotional well-being as a causal force. Also, since individual happiness in an organization has a positive correlation with *productivity*, our proposed statistical model can be used for firms' promotion policies, and they may be useful for managers and human resources professionals. Human resource managers can use our model to predict the individual happiness score by using the questionnaire (Appendix B) . It will help the company to identify those individuals who need to be rewarded and those who need to improve their happiness score. Identifying those individuals are essential for the company as happiness is correlated with an increase in productivity. Our proposed statistical model is also highly useful for *decision making* and *strategic planning* on controlling the factors responsible for causing people to be unhappy and depressed. Finally, since happiness is the most crucial aspect of human life that we seek, controlling the most critical risk factors that significantly contribute to the happiness are essential to control the crime rate of a country, as there is a negative correlation between the individual country's happiness score(Ladder) and crime rate.

Table 8.4: Ranking of Developed Countries based on Predicted Happiness Score

Rank	Country	Score
1	Finland	7.67
2	Denmark	7.55
3	Sweden	7.54
4	Iceland	7.38
5	United States	7.35
6	Canada	7.29
7	Ireland	7.17
8	Switzerland	7.16
9	United Kingdom	7.09
10	Germany	7.03
11	Malta	6.98
12	Luxembourg	6.96
13	Oman	6.96
14	Estonia	6.92
15	Singapore	6.89
16	Qatar	6.82
17	France	6.76
18	Uruguay	6.73
19	Slovenia	6.62
20	Malaysia	6.61
21	United Arab Emirates	6.56
22	Saudi Arabia	6.54
23	Netherlands	6.53
24	Argentina	6.51
25	Australia	6.51
26	Austria	6.51
27	Bahrain	6.51
28	Belarus	6.51
29	Belgium	6.51
30	Czech Republic	6.46
31	Norway	6.43
32	Israel	6.38
33	Lithuania	6.35
34	Chile	6.34
35	Spain	6.31
36	Slovakia	6.24
37	Japan	6.24
38	Hungary	6.14
39	Poland	6.12
40	New-Zealand	6.11
41	Cyprus	6.09
42	Italy	6.06
43	Kazakhstan	6.05
44	Russia	5.99
45	South Korea	5.94
46	Kuwait	5.70
47	Turkey	5.60
48	Croatia	5.55
49	Portugal	5.45
50	Montenegro	5.38
51	Latvia	5.35
52	Bulgaria	5.34
53	Greece	5.20
54	Romania	5.04

Chapter 9: A Real Data-Driven Clustering Approach For Countries based on Happiness Score

Journal Article : “A Real Data-Driven Clustering Approach for Countries Based on Happiness Score,” by Aditya Chakraborty & Chris P. Tsokos, Amfiteatru Economic, 23 (Special Issue No. 15), pp. 1031-1045, CC-BY-NC 2021 by Copyright Holder. Used with Permission

9.1 Introduction

In machine learning and data science literature, Clustering is the task of dividing the observations (data points) into several categories in such a way that data points falling into one group are being dissimilar than the data points falling to the other groups such that the variation within a group is minimized and the variation between the groups is maximized. It falls under the class of *unsupervised learning* techniques. It is primarily a tool to classify individuals on the basis of similarity and dissimilarity between them.

Our present study utilizes the world happiness data of 156 countries collected by the Gallup World Poll. Our study proposes the most accurate (if not the best) clustering algorithm with a very high degree of accuracy to classify different countries of the world based on several economic and social indicators. The most appropriate clustering algorithm has been selected based on different statistical methods. We also proceed to rank the top ten countries in each of three clusters according to their happiness score. The three leading countries in terms of happiness from cluster 1 (medium happiness), cluster 2 (high happiness), and cluster 3 (low happiness) are Oman, Denmark, and Guyana, respectively, followed by United Arab Emirates, Finland, and Pakistan. Finally, we use four popular machine learning classification algorithms to validate our cluster-based algorithms and obtained very consistent results with high accuracy. Richard Easterlin (1974) was the first economist to make prominent use of happiness data when he reported that despite increases in personal income over time, people were not reporting an increasing level of happiness[43]. Being happy not only is associated with personal well being but also with the productivity at a large scale Throughout history, it has been seen as the ultimate end of temporal existence. Recently, social researchers are prone to use sophisticated machine learning techniques in applied sciences [28] [26] to answer specific types of questions regarding happiness of individuals and socio-economic conditions of a country as a whole. Chakraborty & Tsokos [27] developed a data-driven analytical model with high performance to predict the happiness of the developed countries based on different social and economic indicators. Yarkoni & Westfall [142] reviewed some of the basic concepts and methods of statistical machine learning and provides instances where these concepts have been implemented to perform important applied psychological research that focuses on predictive research questions. They also recommended that an increased focus on prediction, rather than descriptive explanation, might lead us to greater understanding on the unknown parameter of interest. Chaipornkaew & Prexawanprasut [25] developed a machine learning prediction model for human happiness using four popular methods, namely KNN, Multi-Layer Perceptron, Naïve Bayes, and Decision Tree. Their proposed model suggested that the Decision Tree with Random Over-sampler technique is the best prediction algorithm for the analyzed data.

Our main goal of the study is the following:

1. To perform different types of explanatory analysis clustering is a great tool. But it is necessary to check the quality of the data set, that is to verify if the data is clusterable or not.We plan to check if there is any random pattern in the data before performing clustering.

2. We want to develop an appropriate clustering algorithm by implementing different algorithms to choose from which will classify similar observation (countries) with high level of accuracy based on the indicators.

3. After selection of the accurate clustering algorithm, we proceed to perform analysis of individual clusters to choose the indicators which are most influential.

4.We rank top **ten** countries based on happiness score in each cluster.

5. For validation purpose of our clustering, we perform *machine learning classification* with four very popular classification algorithms.

6. Finally, for an overall picture, we plan to create a global map to show the position of clusters in the map. It provides an overall idea about the happiness and also other socio -economic conditions related to happiness of individual countries. Then, we proceed to compare the cluster map with Figure 9.1, where we have world map based on happiness scores. In the following figure, the light pink indicates the countries with the highest happiness scores. These countries include the United States of America and Australia. On the contrary, red indicates the countries with the least happiness score. We see that the countries with the least happiness scores are some countries in Africa and Asia. The world map has been produced based on our real happiness data.

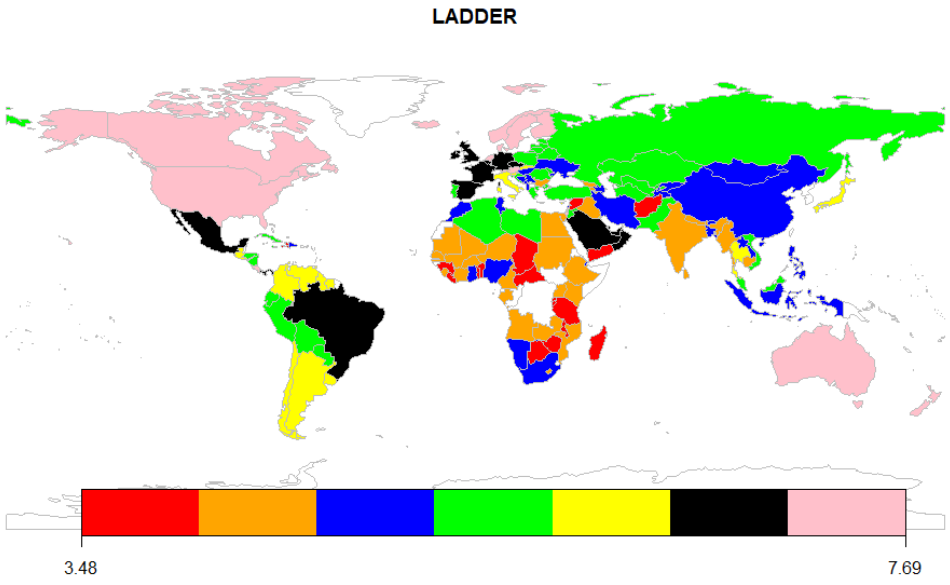


Figure 9.1: World Map Showing the Happiness Scores (LADDER) Of Different Countries

9.2 The Data

The description of the data has been provided in Chapter 9.

9.3 Investigating clustering pattern

Before performing any kind of cluster analysis, it is primitive to check if the data we are trying to analyze is clusterable. Analysis of non-clusterable data might produce misleading results if we proceed with clustering of the data.

9.3.1 Hopkin's Method

One important approach for testing clustering tendency is by using *Hopkins statistic*. The statistic computes the clustering pattern by computing the probability that a given data set is generated by a uniform distribution. More specifically, it tests the spatial randomness of the data. The problem of testing for clustering tendency can also be described as the problem of testing for spatial randomness[10]. The algorithm for computation of *Hopkins statistic* as follows.

1. Uniformly sample from $u_1, u_2, ..., u_n$ of a data set D

2. For each point $u_i \in D$, find it's nearest neighbor U_J and evaluate $d_i = dist(u_i, u_j)$.

3. Generate a simulated data set ($random_D$) drawn from a random uniform distribution with n points ($v_1, ..., v_n$) and the same variation as the original data set D.

4. For each point $v_i \in random_D$, find it's nearest neighbor v_j in D and calculate the distance $k_i = dist(v_i, v_j)$

5. Calculate the Hopkins statistic (H)(defined below) as the mean nearest neighbor distance in the random data set divided by the sum of the mean nearest neighbor distances in the real and across the simulated data set.

$$H = \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n} k_i + \sum_{i=1}^{n} d_i} \tag{9.1}$$

A value of H statistic close to 0.5 means that $\sum_{i=1}^{n} k_i$ and $\sum_{i=1}^{n} d_i$ are close to each other, and thus the data D is uniformly distributed. The null and the alternative hypotheses of the test are stated as follows:

Null hypothesis: The data set D is distributed uniformly (no meaningful clusters)

Alternative hypothesis: The data set D is not uniformly distributed (i.e., contains meaningful clusters)
If the value of H is close to 0, then we can say that there exists sufficient sample evidence to reject the null hypothesis and conclude that the data set D is significantly clusterable. Performing the testing problem, we have found that our happiness data is excellent for clustering (the H value = 0.27 from (1), far below the threshold 0.5).

9.3.2 Visual method

It is a good practice to express the data pictorially for getting a visual representation for the clustering assessment. The technique of deciding whether clusters are present as a step prior to actual clustering is called the assessing of clustering tendency. The visual assessment of cluster tendency (VAT) algorithm (Bezdek and Hathaway, 2002)[14] is the following.

1. At first, the dissimilarity (DM) matrix between the objects in the data set using the Euclidean distance measure is constructed.

2. The DM is then constructed so that identical objects are adjacent to each other. in this process, an *ordered dissimilarity matrix (ODM)* is produced.

3. The ODM is then displayed as an *ordered dissimilarity image (ODI)*, which is the visual output of VAT.

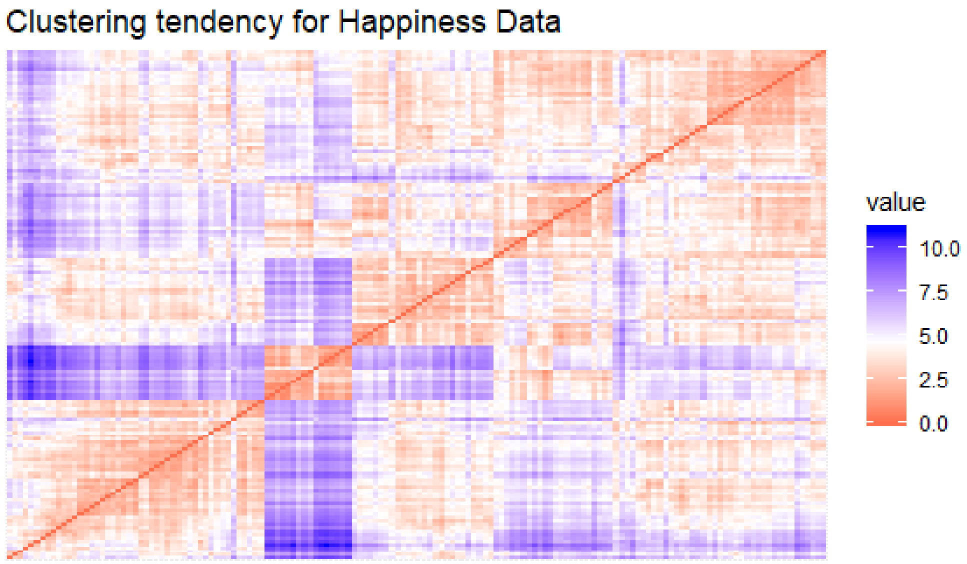


Figure 9.2: Showing The Clustering Tendency Of Happiness Data. Red: High Similarity , Blue: Low Similarity

For the visual evaluation of clustering tendency, dissimilarity matrix between observations has been constructed. In the above Figure 9.2, the color level is proportional to the value of the dissimilarity between observations: pure red if $dist(x_i, x_j) = 0$ and pure blue if $dist(x_i, x_j) = 1$.

Objects belonging to the same cluster are displayed in consecutive order. The dissimilarity matrix image in Figure 2 confirms that there is a clustering pattern in the happiness data.

9.4 Optimal number of clusters

Deciding the optimal number of clusters for a set of data is a basic challenge in clustering, such as k-means, k-medoids (PAM), and hierarchical clustering, which requires the user to select the number of clusters k to be determined. The method of selecting the number of clusters is somehow subjective and also dependent upon using the techniques for computing similarities and the parameters used for partitioning. However, are almost thirty methods (indices, see [29]) to decide the optimum number of clusters; the most popular methods include **elbow method, silhouette methods, Hartigan and Gap Statistic method**.

Elbow Method: The concept behind partitioning methods, such as k-means clustering, is to define clusters in such a way that the total intra-cluster variation [or total *within-cluster sum of square (WSS)*] is minimized. The total WSS is a measurement of the compactness of the clustering, and it is desired to be as minimal as possible. In the Elbow method plot, the total WSS is a function of the number of clusters. The number of clusters should be selected in such a way so that adding another cluster doesn't improve total WSS.

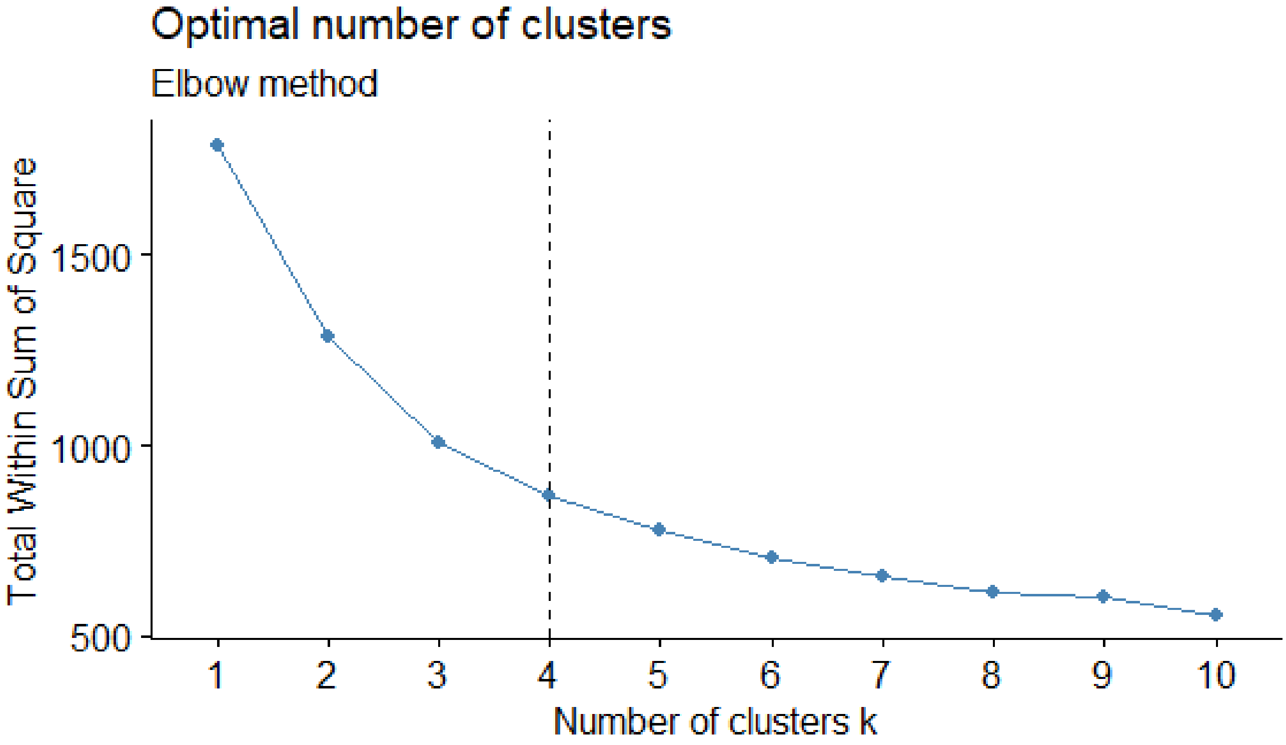


Figure 9.3: Elbow Method Showing The Optimum Number Of Clusters

However, the elbow method is a little bit subjective; we see from the above figure that the optimum number of clusters might be 4 as the elbow is approximately at 4.

Average silhouette method: This is a graphical display method suggested for partitioning techniques[109]. In the silhouette plot, each cluster is exhibited by a silhouette, which is based on the comparison of its tightness and separation. This silhouette represents the objects falling within their clusters and also the objects falling somewhere in between clusters. The whole clustering is displayed by incorporating the silhouettes into a unique plot, which allows understanding the relative quality of the clusters and an overview of the data configuration. The average silhouette width can be used to evaluate the validity of clustering and can be implemented to select an 'appropriate' number of clusters.

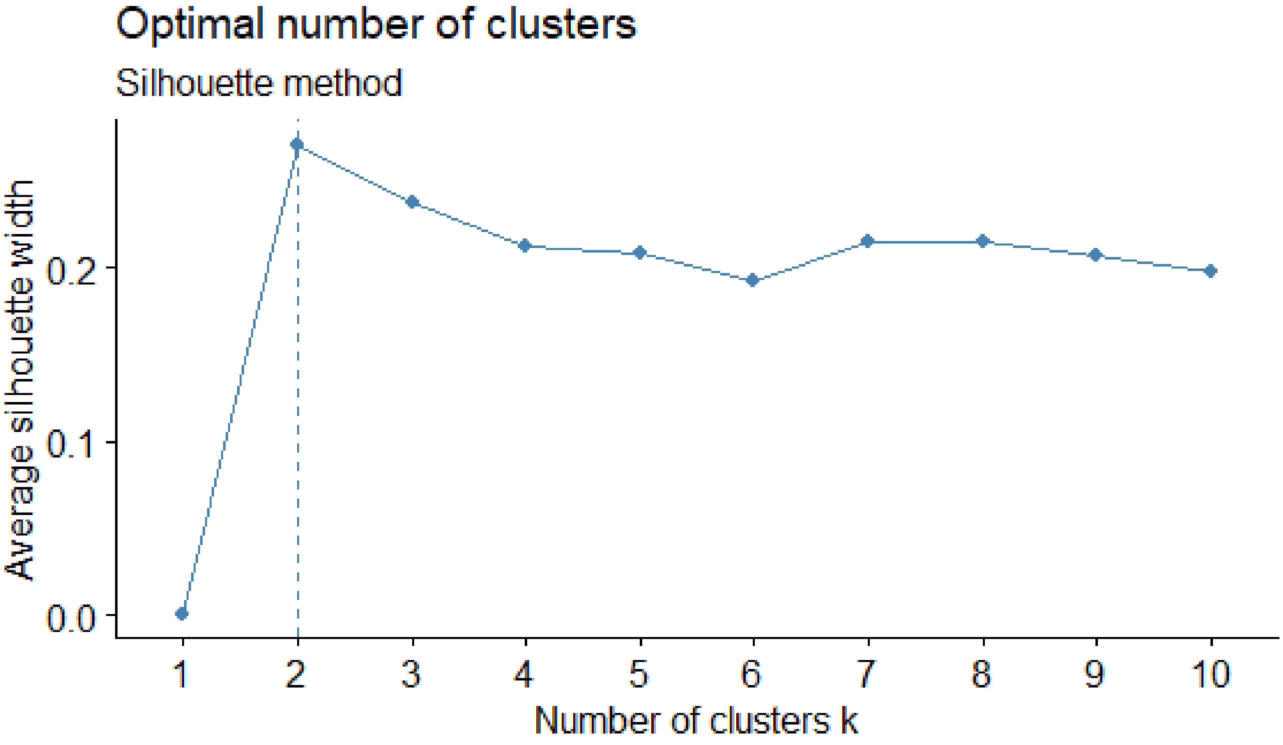


Figure 9.4: Average silhouette plot showing the optimum number of clusters

The above picture suggests that optimum number of clusters might be 2 by Average silhouette method approach.

Hartigan's Method: Hartigan's method[127] for k-means clustering is based on the following greedy heuristic: *select a point, and optimally reassign it.* This algorithm essentially compares the ratio of the within-cluster sum of squares for clustering with k clusters and one with $k+1$ clusters, accounting for the number of rows and clusters. If the number is greater than 10, then one might take $k+1$ as an optimum number of clusters.

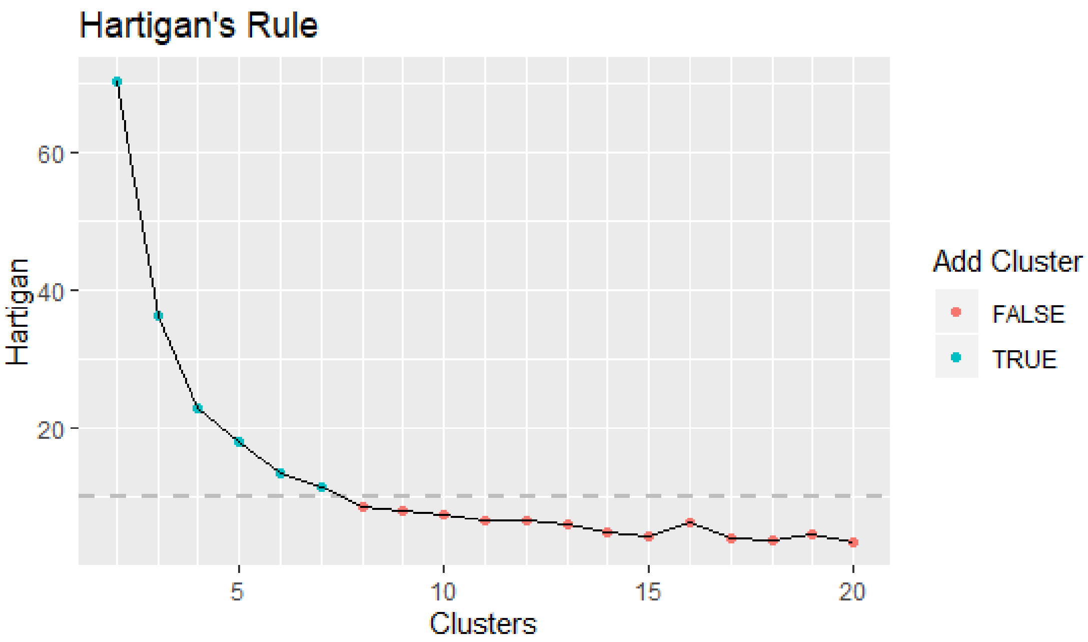


Figure 9.5: Hartigan's Plot Showing The Optimum Number of Clusters

From the above graph, we see that there are five complete joins if we join the blue points distinctly. Hence, this method suggests taking five clusters into account.

Gap Statistics: The gap statistic [62], published by R. Tibshirani, G. Walther, and T. Hastie (Standford University, 2001), might be applied to any data clustering technique. For different values of the cluster number k, the gap statistic gives a comparison between the total within intra-cluster variation with their expected values under null reference distribution of the data. The estimate of the optimal clusters is the value that maximizes the gap statistic (the value that yields the largest gap statistic), which implies that the clustering structure is significantly different from the random uniform distribution of points. The algorithm[77] works as follow:

1. The observed data is clustered, varying the number of clusters from $k = 1, ..., k_{max}$ and the corresponding total within intra-cluster variation W_k is calculated.

2. Afterwards, B reference data sets with a random uniform distribution are generated. Each of these reference data sets is clustered with varying number of clusters $k = 1, ..., k_{max}$, and the corresponding total within intra-cluster variation W_{kb} is computed.

3. The estimated gap statistic as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis: $Gap(k) = \frac{1}{B}\sum_{b=1}^{B} logW_{kb}^{*} - logW_k$ is computed. Also the standard deviation of the statistics is computed.

4. Select the minimum value of k such that the gap statistic is within one standard deviation of the gap at $k + 1 : Gap(k) \geq Gap(k + 1) - s_{k+1}$.

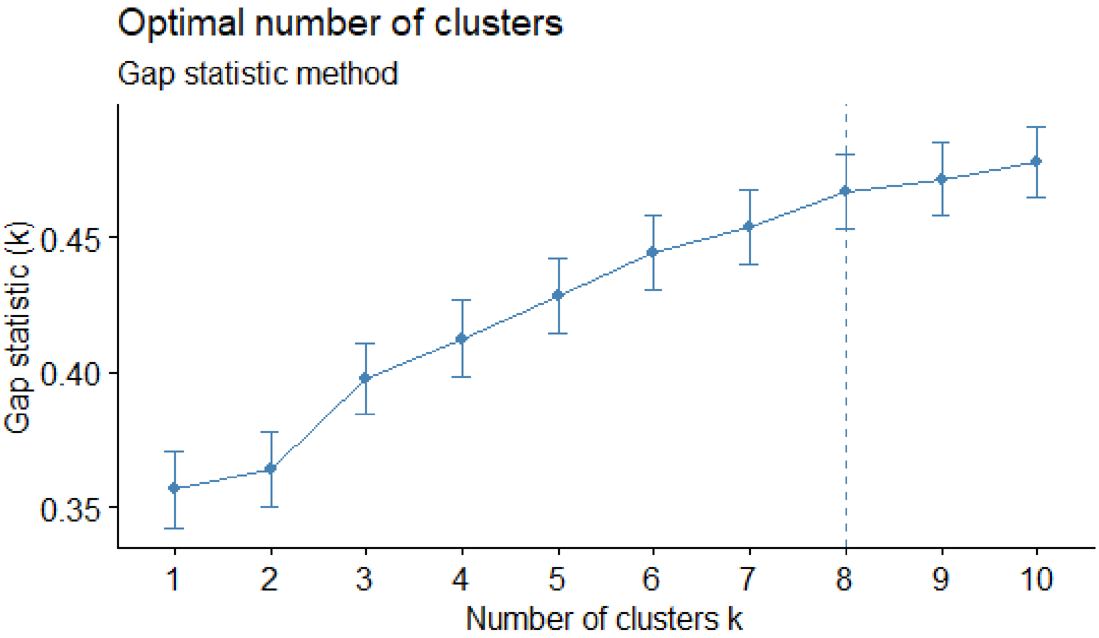


Figure 9.6: Gap Statistic Plot showing the optimum number of clusters

In the above Figure 9.6, Gap statistic vs. the number of cluster plot is based on 50 bootstrap samples shows that the algorithm selects number **eight** as an optimum cluster.

The ultimate cluster choice based on different Clustering validity indices: Several clustering algorithms[29] lead to different clusters of data. Moreover, even for the same algorithm, selecting different parameters or the presentation order of data objects may hugely influence the final clustering partitions. Thus, practical evaluation standards and criteria are critically essential to achieve desired clustering outcomes. At the same time, these evaluations also furnish some significant intuitions on how many clusters are inherent in the data. In fact, in most real-life clustering scenarios, the user faces the problem of choosing the number of clusters or partitions in the underlying data. There are several Clustering validity indices in the literature. All these clustering validity indices combine information about intra-cluster compactness and inter-cluster isolation, as well as other factors, such as geometric or statistical properties of the data, the number of data objects, and dissimilarity or similarity measurements.

By selecting the standard *euclidean* distance measure, we found the following result based on several Clustering validity indices.

Table 9.1: Comparison Among Several Clustering Validity Indices Based on Majority votes

Result showing the outputs based on Clustering validity indices
5 proposed 2 as the best number of clusters
12 proposed 3 as the best number of clusters
2 proposed 6 as the best number of clusters
4 proposed 10 as the best number of clusters

9.5 Selecting the best Clustering Algorithm

Selecting the best clustering algorithm can be a tough call for a research scientist. One of the rudimentary challenges of clustering is how to evaluate results without any supplementary knowledge beforehand. A usual approach for the evaluation of clustering results is to use validity indexes. Clustering validity approaches can use two criteria[108]: External criteria (evaluate the result with respect to a pre-specified structure), internal criteria (evaluate the result with respect to information intrinsic to the data alone). Hence, different types of indexes are used to solve different types of problems, and index selection depends on the kind of available information. Here, we will be using two types of clustering validation techniques, say stability criteria and internal criteria.

9.5.1 Internal Measures

It is based on intrinsic information in the data to assess the quality of the clustering. Internal measures include the connectivity, the silhouette coefficient, and the Dunn index (information intrinsic to the data). In order to perform internal validation, we selected measures that reflect the *compactness*, *connectedness* and *separation* of the clustering partitions. Connectedness talks about the extent to which observations are placed into the identical cluster as their nearest neighbors in the data space. It is measured by connectivity [Handl et al., 2005]. Compactness assesses cluster homogeneity via the intra-cluster variance, while separation measures the degree of disconnection between clusters by computing the distance between cluster centroids. Since compactness and separation demonstrate opposing trends, popular methods combine the two measures into a single score. The Dunn Index and Silhouette Width [Rousseeuw,1987] are both examples of non-linear combinations of compactness and separation[115].

9.5.1.1 Connectivity

Let N denote the total number of observations to be clustered. Define $nn_{i(j)}$ as the jth nearest neighbor of observation i, and let $x_{i,nni(j)}$ be zero if i and j are in the same cluster and $1/j$ otherwise. Then, for a particular clustering partition $C = \{C_1, ..., C_K\}$ of the N observations into K disjoint clusters, the connectivity is defined as

$$Conn(C) = \sum_{i=1}^{N} \sum_{J=1}^{l} x_{i,nni(j)}$$

where l is user-specified. The connectivity has a value between zero and ∞ and should be minimized.

9.5.1.2 Silhouette Width

The Silhouette Width is the average of each observation's Silhouette value. The Silhouette value measures the degree of confidence in the clustering assignment of a particular observation, with well-clustered observations having values near 1 and poorly clustered observations having values near -1. For observation i, it is defined as

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$$

where a_i is the average distance between i and all other observations in the same cluster

$$a(i) = \frac{1}{n(C(I) - 1)} \sum_{j \neq j \in C(i)} d(i,j)$$

The Silhouette Width $S(i)$ which lies in the interval $[-1, 1]$ should be maximized.

9.5.1.3 Dunn Index

The Dunn Index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It is computed as

$$D(C) = \frac{min_{C_k, C_1 \in C, C_k \neq C_1} \left(min_{i \in C_k, j \in C_1} \{dist(i,j)\} \right)}{max_{C_m \in C} \{diam(C_m)\}}$$

where $diam(C_m)$ is the maximum distance between observations in cluster C_m.The Dunn Index has a value between zero and ∞, and should be maximized.

9.5.2 Stability Measures

It is a special sort of internal measure criteria, which assess the uniformity in a clustering mechanism by comparing it with the clusters obtained after each column is removed, one at a time. There are four measures that fall into the stability measure. In all of these cases, the average is taken over all the deleted columns, and all measures must be minimized. Let N and M denote the total number of observations (rows) in a data set and the total number of columns, respectively, which are assumed to be numeric. We define the four measures following.

9.5.2.1 Average Proportion of Non-overlap (APN)

The APN [111] measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. Let $C^{i,0}$ represent the cluster containing observation i using the original clustering (based on all available data), and $C^{i,l}$ represent the cluster containing observation i where the clustering is based on the dataset with column i removed. Then, with the total number of clusters set to K, the APN measure is defined as

$$APN(C) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{l=1}^{M} \left(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n(C^{i,0})}\right) \quad ,$$

The APN lies within the interval $[0, 1]$, with values close to zero corresponds to highly consistent clustering results.

9.5.2.2 Average Distance (AD)

The AD [71] measure computes the average distance between observations placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. It is defined as

$$APN(C) = \tfrac{1}{MN}\textstyle\sum_{i=1}^{N}\sum_{l=1}^{M}\tfrac{1}{n(C^{i,0})n(C^{i,l})}\left[\textstyle\sum_{i\in C^{i,0},j\in C^{i,l}}\left\{dist(i,j)\right\}\right] \quad ,$$

The AD has a value between zero and ∞, and smaller values are preferred.

9.5.2.3 Average Distance between Means (ADM)

The ADM [40] measure calculates the average distance between cluster centers for observations that are placed in the identical cluster by clustering depending on the full data and clustering based on the data with a single column removed. It is defined as

$$ADM(C) = \tfrac{1}{MN}\textstyle\sum_{i=1}^{N}\sum_{l=1}^{M} dist(\overline{x}_{C^{i,l}},\overline{x}_{C^{i,0}}) \quad ,$$

where $\overline{x}_{C^{i,0}}$ is the mean of the observations in the cluster which contain observation i, when clustering is based on the full data, and $\overline{x}_{C^{i,l}}$‘ is similarly defined. Usually, ADM only uses the Euclidean distance. It also can take values between zero and ∞ and again smaller values are preferred.

9.5.2.4 Figure of Merit (FOM)

The FOM [54] measures the average intra-cluster variance of the observations in the deleted column, where the clustering is based on the remaining (undeleted) samples. This estimates the mean error using predictions based on the cluster average. For a particular left-out column l, the FOM is

$$FOM(l,C) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\sum_{i\in C_K(l)} dist(x_{i,l},\overline{x}_{C_k(i)})} \quad .$$

The FOM values can range between zero and infinity and smaller values are preferred. For the selection of the most accurate clustering algorithm, we compared among three popular clustering algorithm, namely k-means, hierarchical, and PAM, and we selected the *stability* as our measure of choosing the most appropriate clustering algorithm as it gives a robust result throughout the four above mentioned stability measures. The following Table 9.5.1 illustrates that k-means have been chosen by the four stability measures as an optimal algorithm.

Table 9.2: Comparison Four Stability Measures To Choose Appropriate Clustering Algorithm

Measure	Score	Algorithm
APN	.063	kmeans
AD	3.326	kmeans
ADM	.3	kmeans
FOM	.765	kmeans

9.6 K-means Clustering

The fundamental concept behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as a total within-cluster variation) is minimized. There are several k-means algorithms available. However, the most frequently used algorithm is the Hartigan-Wong algorithm (1979). It is defined as the total within-cluster variation as the sum of squared Euclidean distances between the items and the corresponding centroid:

$$W(C_k) = \sum_{x_i\in C_k}(x_i - \mu_k)^2 \quad ,$$

where x_i is the observation belonging to the cluster C_k and μ_k is the mean value of the points assigned to the cluster C_k. Each observation x_i is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers μ_k is a minimal. We define the total within-cluster variation as follows:

$$T_{Variation} = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad ,$$

where K is the total number of clusters. The total within-cluster sum of square measures the compactness (i.e goodness) of the clustering and we want it to be as small as possible. The K-means algorithm can be summarized as follows:

- At first, the number of clusters (k) to be specified by the researcher.
- Randomly k objects are selected from the data set as the initial cluster centers or means.
- Each observation is assigned to their closest centroid, based on the Euclidean distance between the object and the centroid.
- For each of the k clusters, the cluster centroid is updated, and the new mean values are calculated of all the data points in the cluster. The centoid of a k^{th} cluster is a vector of length p that contains the means of all of the variables for the observations in the k^{th} cluster, where p is the number of variables.
- The total within sum of square is then iteratively minimized. That is, iterating steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached.

The following Figure 9.7, gives a schematic diagram of K-means algorithm.

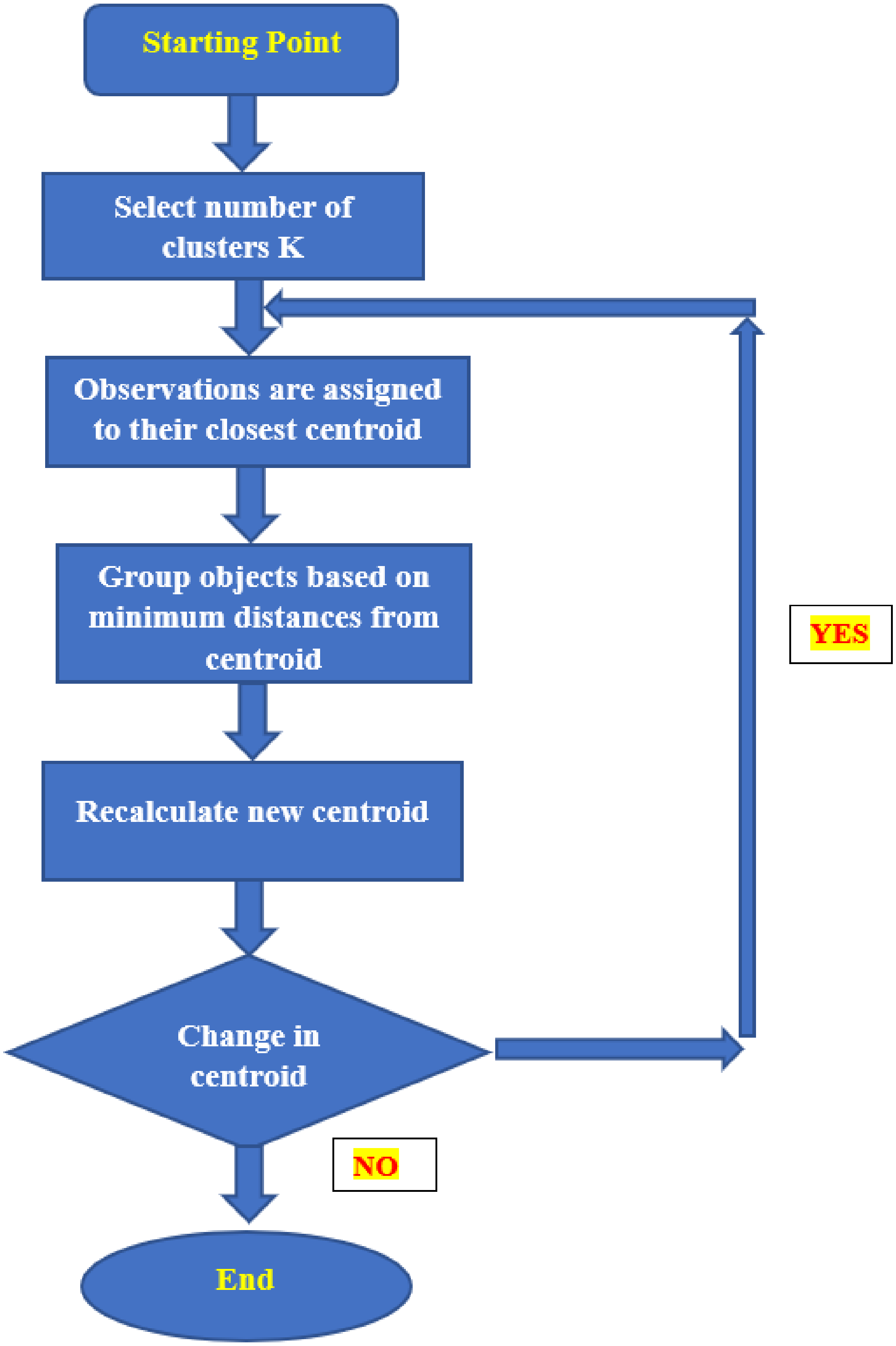

Starting Point
Select number of clusters K
Observations are assigned to their closest centroid
Group objects based on minimum distances from centroid
Recalculate new centroid
Change in centroid
YES
NO
End

Figure 9.7: Flowchart of K-Means Algorithm

One of the goals of our study to investigate which countries fall into similar groups (clusters) based on the happiness data. After we have selected the number of clusters and the appropriate clustering algorithm, we are in a position to perform the k-means clustering with clusters 3 (that is, group the data into three clusters/subgroups).

We see that our clustering resulted in 3 clusters with sizes 55, 69, and 26, respectively. The following figure projects the multi-dimensional data into three clusters that are formed as a result of k-means clustering.

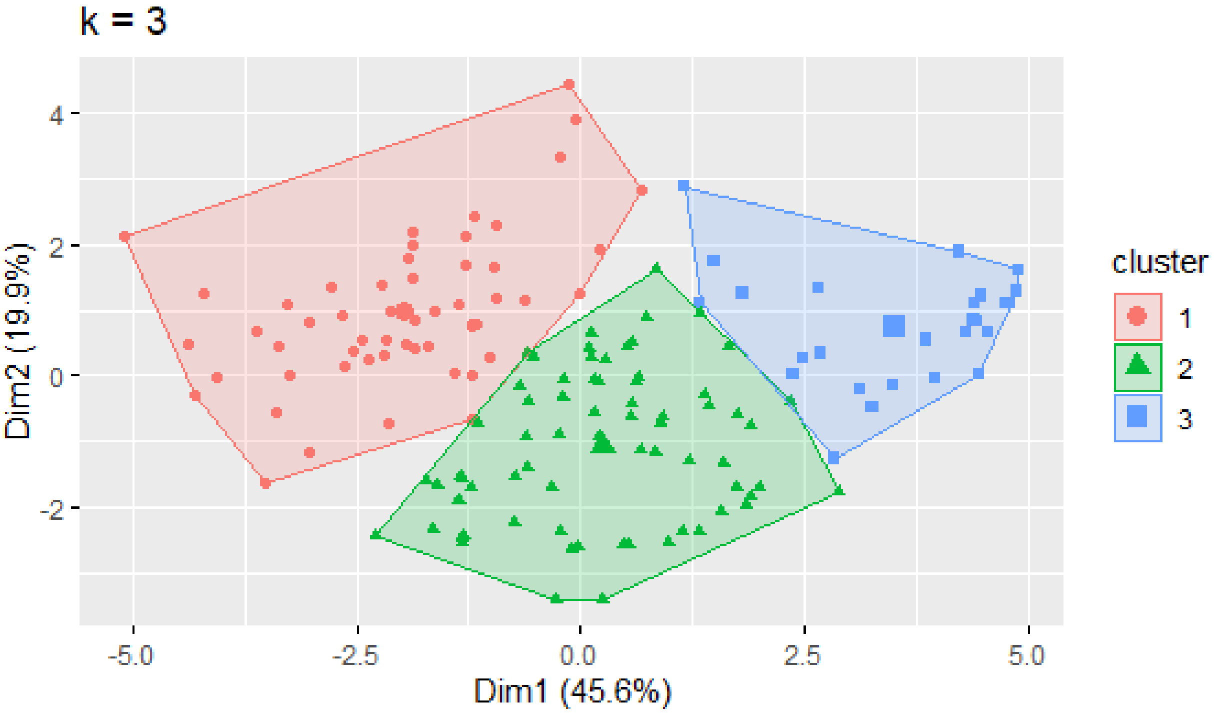


Figure 9.8: Showing The Three Clusters Obtained By k-means Clustering

The above Figure 9.8 provides a multidimensional scaling using principal component analysis (PCA), and the data was plotted based on the first two principal components. From the figure, we can see that the algorithm did a good job of clustering the data points. We can also plot the three clusters obtained by the k-means algorithm to investigate the behaviors of happiness throughout all three clusters.

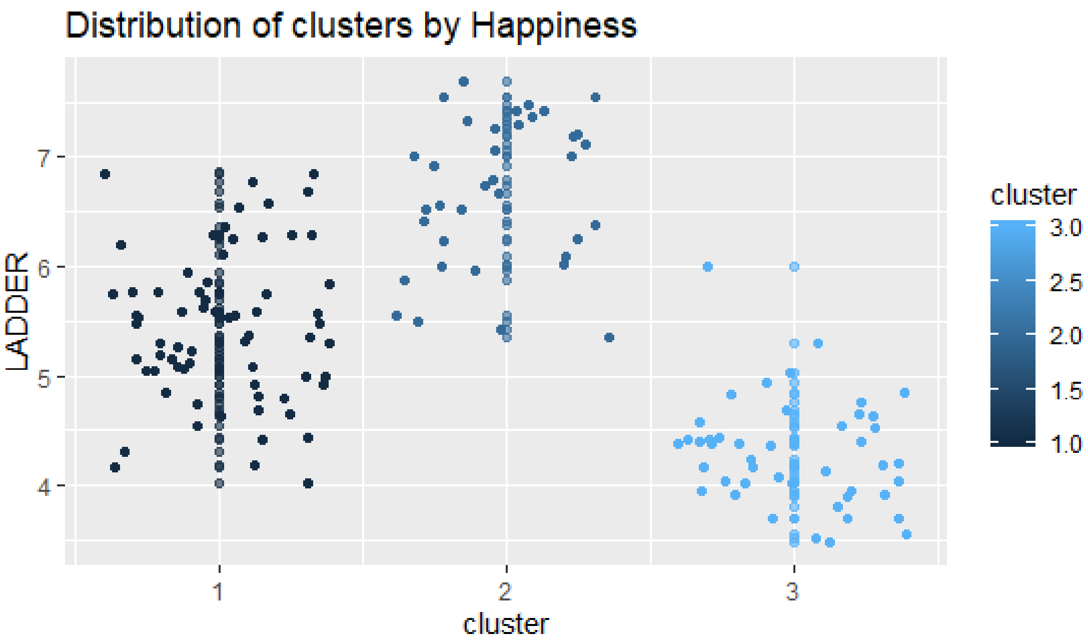


Figure 9.9: Showing The Distribution of Happiness Throughout Three Clusters

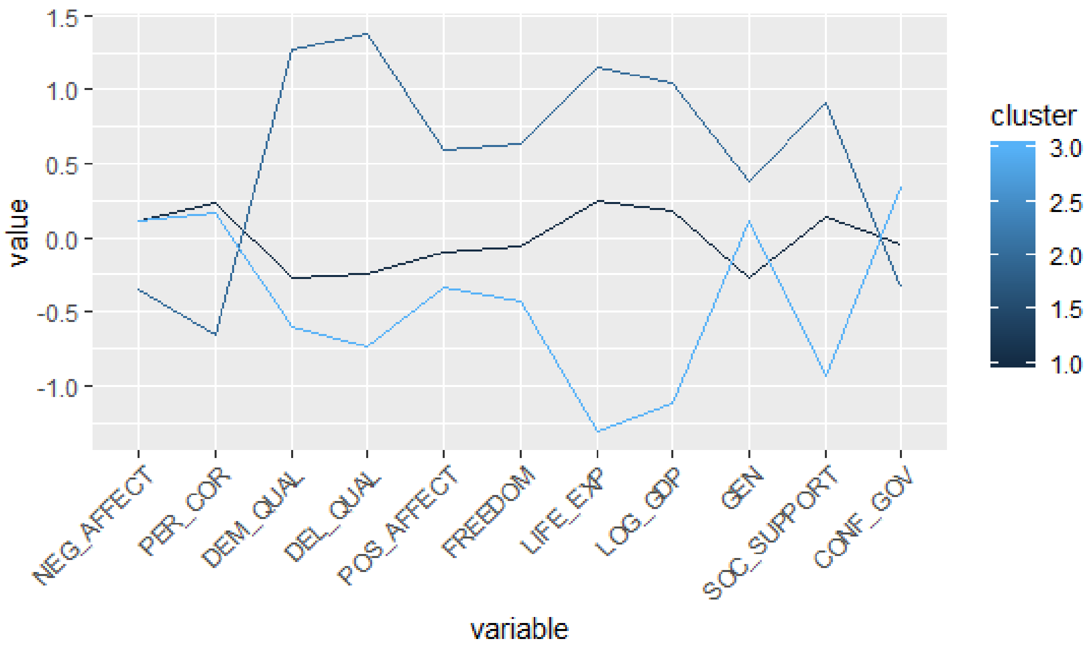


Figure 9.10: Showing The Distribution of Indicators Throughout Three Clusters

From the above Figure 9.9, we notice that countries falling within cluster 2 happens to be happiest among the three, followed by cluster 1 and cluster 3. Most probably, cluster 2 contains the most developed and democratic countries of the world, and cluster 3 contains most of the underdeveloped and developing, and less democratic countries. It seems Cluster 2 contains most of the developing countries followed by Cluster 1 and Cluster 3.

After implementing the k-means algorithm, we have computed the cluster means for all eleven indicators. We can plot the cluster means for all eleven indicators throughout the three clusters to compare how each attributable variable (indicators) behaves in these three clusters.

As the above Figure 9.10 illustrates, we notice that each average values of the indicators in cluster 2 are higher than that of cluster 1 and cluster 3 except **NEG_AFFECT** (X_8), **PER_COR** (X_6) and **CONF_GOV** (X_9). On the other hand, every average numerical measure of indicators in cluster 3 is worse than that of cluster 2 and cluster 1 except **Generosity** (X_5) and **CONF_GOV** (X_9). By studying the graphs, We see an almost opposite pattern between cluster 2 and cluster 3. For example, we see the variable **SOC_SUPPORT** (X_2) in cluster 3 has been placed into a completely opposite position when compared to cluster 2. By visualizing the pattern, we might tell that most developed countries are placed into cluster 2, and most underdeveloped countries are classified into cluster 3. Cluster 1 behaves pretty averagely when compared with cluster 2 and cluster 3.

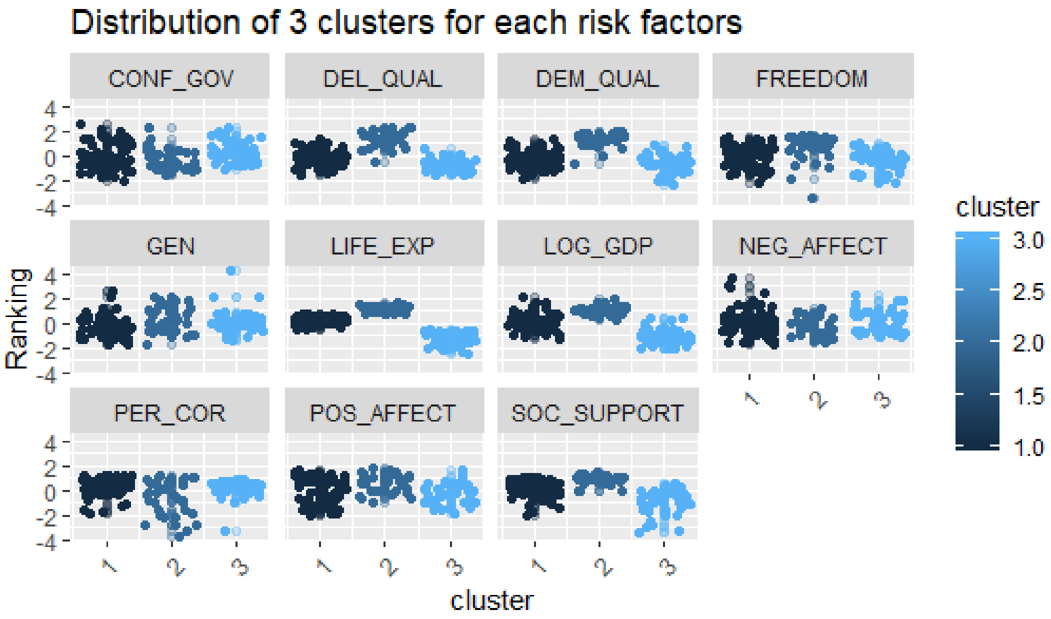


Figure 9.11: Showing The Distribution of Each Indicators Individually For Three Clusters

However, those countries classified into cluster 1 have the lowest average **Generosity (X_5)** values which need to be further analyzed. Also, it seems that there must be some correlation between cluster 1 and cluster 2 as they show a parallel trend for most of the indicators.

For a better understanding, we can visualize the variability of three clusters with respect to each indicators.

From Figure 9.11, we can obtain some very interesting facts about each indicators factor in each cluster. We see that countries belonging to cluster 3 has especially significantly low measures for indicators **LIFE_EXP (Life Expectancy)**,**SOC_SUPPORT (Social Support)**, and **LOG_GDP (logarithm of GDP)**. Since we got an idea from Figure 9.6.3 and Figure that underdeveloped countries belong to cluster 3, they can expect to have low GDP and life expectancy. One striking fact that these countries also have very low scores for social support(X_2) which means that on an average the citizens of those countries does not feel to get support from their relative, friends or Governments when they are in trouble. One interesting fact to notice that, however, there is not much difference among the **Generosity** (X_5) within the three clusters; some countries in cluster 3 outperforms some countries in cluster 2 when it comes to **Generosity (X_5)**.

Also, the countries belonging to cluster 1 and cluster 3 happen to have almost the same average measures of **perception of corruption (X_6)**.

One of the most important consequences of the clustering aspect is that we can rank the different countries in the world in each cluster based on happiness score. Since there is a positive correlation between happiness and democracy, by knowing the name of the happiest countries in the cluster, we might able to guess their socio-economic status.
Table 9.3 below shows the top 10 countries in cluster 3 based on happiness score.

It is important to note that, the top three countries, with respect to happiness in cluster 3 are Guyana, Pakistan, and Nigeria, respectively.

Table 9.3: Ranking Of Countries In Cluster 3 Based On Happiness Score

Country	Rank
Guyana	1
Pakistan	2
Nigeria	.3
Laos	4
South Africa	5
Djibouti	6
Ghana	7
Namibia	8
Mozambique	9
Zambia	10

Table 9.4: Ranking Of Countries In Cluster 2 Based On Happiness Score

Country	Rank
Denmark	1
Finland	2
Norway	3
Netherlands	4
Canada	5
Iceland	6
Sweden	7
New Zealand	8
Australia	9
Austria	10

The Table 9.4 below shows the top 10 countries in cluster 2 based on happiness score. Similarly, the Table 9.5 below shows the top 10 countries in cluster 1 based on happiness score.

Table 9.5: Ranking Of Countries In Cluster 1 Based On Happiness Score

Country	Rank
Oman	1
United Arab Emirates	2
Mexico	3
Brazil	4
Qatar	5
Saudi Arabia	6
Argentina	7
Kuwait	8
Colombia	9
Trinidad and Tobago	10

We have similarly listed the countries based on every three clusters based on the eleven indicators. A more detailed socioeconomic condition of the countries might be assessed out of it. Howell & Howell[68] has shown that there is a positive correlation with the subjective well-being (SWB) economic status of a country. We can further study how economic status correlates with the eleven indicators and which are the most important contributes to the economy of a country. In the next section, we briefly discuss some machine learning classification techniques to validate the performance of our clustering algorithm.

9.7 Validation of Clustering Algorithm

After clustering the observations into three clusters, the next important thing is to assess the validity of the clustering algorithm. To assess the performance of our k-means clustering algorithm, we have used four machine learning **classification algorithm** to see how well the data has been clustered with a high degree of accuracy. We chose **Decision Tree RPART**, **Decision Tree C5.0**, **Random Forest** , and **Extreme Gradient Boosted Tree (xgbTree)** classification algorithms for checking cluster validity. The reason behind selecting these two classification algorithms is that they have no distributional assumptions and are also very popular supervised algorithms that happen to work well with a large amount of data. If we get high classification accuracy by these two methods, we can conclude that the k-means clustering method we have implemented here is one of the good if not the best clustering algorithm for the data we have.

9.7.1 Decision Tree

In the literature of statistical data mining, decision trees [107] play a vital role in decision analysis. Tree-based learning algorithms are deemed to be one of the finest and frequently used supervised learning methodologies. Tree-based methods boost the predictive performance of the models with a high degree of accuracy and ease of interpretation. Unlike linear models, they address non-linear relationships quite perfectly. These models are flexible at solving any kind of data-driven decision-making problem. (classification or regression).

9.7.1.1 Recursive Partitioning And Regression Trees (RPART)

The RPART algorithm [128] works by splitting the data set recursively, which means that the subsets that arise from a split are further split until a predetermined termination criterion is reached. At each step, the split is done based on the independent variable that results in the largest possible reduction in the heterogeneity of the dependent (predicted) variable.

9.7.1.2 The C5.0 Decision Tree Algorithm

While there are numerous implementations of decision trees, one of the most widely known algorithms is the C5.0 [99] [100]. This algorithm was proposed by computer scientist J. Ross Quinlan as a modified version of his prior algorithm, C4.5, which itself is an improvement over his Iterative Dichotomiser 3 (ID3) algorithm. The C5.0 algorithm has become the industry standard for producing decision trees since it performs satisfactorily for most types of complex real-life data-driven problems. Compared to more advanced and sophisticated machine learning models (e.g., Neural Networks and Support Vector Machines), the decision trees under the C5.0 algorithm generally perform nearly as well but are much easier to understand and deploy.

9.7.2 Random Forest

Random forest or decision tree forest is an ensemble-based method that focuses only on ensembles of decision trees. This method was proposed by Leo Breiman and Adele Cutler [18] and combines the base principles of bagging with random feature selection to add additional diversity to the decision tree models. After the ensemble of trees (the forest) is formed, the model uses a vote to combine the trees' predictions. Random forests combine versatility and power into a single machine learning approach. As the ensemble uses only a small, random portion of the full feature set, random forests can handle extremely large data sets, where the so-called "curse of dimensionality" might cause other models to fail.

9.7.3 Extreme Gradient Boosted Tree (xgbTree)

Like Random Forests, Gradient Boosting [50] [96] is an *ensemble learner* which creates an ultimate model depending on a set of independent models. Usually, the performance of these individual models is low, and they are prone to overfit the data when implemented solely. However, combining many such low-performing models in an ensemble, in an iterative way, usually leads to an overall much improved and accurate result. In boosting, the individual models are built sequentially by putting more weight on instances with wrong predictions and high errors. The main logic behind this is that instances, which are hard to predict correctly ("difficult" cases) are targeted during the learning process so that the model learns from past mistakes. When we train each ensemble on a subset of the training set, we also call this *Stochastic Gradient Boosting*, which can help improve the performance of our model. Extreme Gradient Boosting (XGBoost) [31] is a more sophisticated implementation of the Gradient Boosting algorithm, which uses more hyper-parameters to find the best tree model by employing a number of useful adjustments to prevent overfitting and make the model exceptionally successful, particularly with structured data.

In the following section, we will evaluate the performance of the above-stated classification algorithms to access the quality of our k-means clustering algorithm.

9.8 Evaluation Of The Classification Algorithms

After we implement the four popular machine learning classification algorithms to judge the performance of our k-means clustering algorithm, it is important to check and evaluate the performance of the classification algorithms in terms of evaluation matrices. We will access the quality of the classification algorithms based on the following two evaluation matrices.

9.8.1 Accuracy

Accuracy is one of the most widely used metric for evaluating classification models. Conventionally, multi-class accuracy is defined as the average number of correct predictions as follows.

$$Accuracy = \tfrac{1}{N} \textstyle\sum_{K=1}^{|G|} \sum_{x:g(x)=k} \big(g(x) = \widehat{g}(x)\big),$$

where G is the number of classes, $g(x)$ and $\widehat{g}(x)$ are the classifier and estimated value of the classifier respectively and $I(.)$ is an indicator function which takes the value 1 if the classes match and 0 otherwise.

9.8.2 Cohen's Kappa

Cohen's Kappa or Kappa statistic is a very useful metric in machine learning when we deal with a multi-class classification problem. Basically, it suggests how better the desired classifier performs over the performance of a random classifier that simply makes arbitrary guesses according to the frequency of each class. Always Cohen's kappa [133] is less than or equal to 1 and values zero or less, implies that the classifier is not useful. It is defined as follows.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \quad ,$$

where p_0 is the observed accuracy, the number of instances that were classified correctly and p_e is the expected accuracy, the accuracy that any random classifier would be expected to achieve based on the confusion matrix.

It is very interesting to note that we got very high accuracy and Kappa(κ) values using the four machine learning classification algorithm implying the great performance of our k-means algorithm.

The following Table 9.6 illustrates the values of accuracy and Kappa(κ) generated by the four classification algorithm.

Table 9.6: Comparing The Performance Of Different Classification Methods

Method	Accuracy	Kappa
Rpart	96.6	94.7
C5.0	97.3	95.9
Random Forest	97.8	96.5
xgbTree	99.7	99.1

From the above Table 9.6, we see that the classification is pretty good and consistent with respect to accuracy and kappa measure based on the four popular classification algorithm.

The following Figure 9.12 shows a visual comparison of the four different classification methods based on accuracy and kappa.

It shows that the extreme gradient boosted tree performs the best, followed by random forest, decision tree C5.0, and decision tree RPART. For the first three algorithms, we see the median accuracy and kappa measures are pretty indistinguishable.

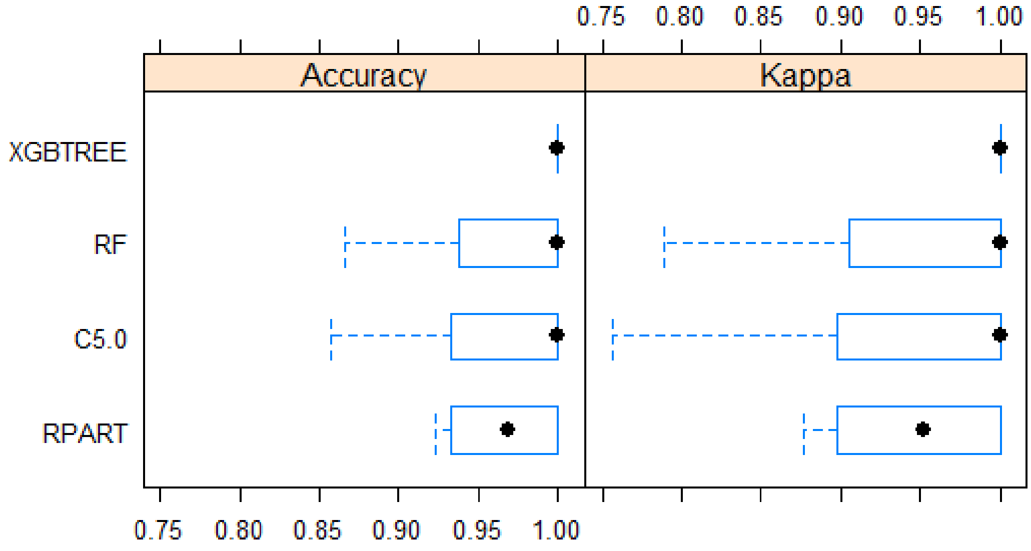


Figure 9.12: Visual representation of four classification methods in terms of accuracy and kappa

9.9 Contribution and Conclusion

By implementing appropriate clustering algorithm to our happiness data, we have successfully accomplished all the goals introduced in section 1.

- We have shown statistically and visually that there is a meaningful clustering pattern in our happiness data.

- We have implemented different methodologies to select an optimal number of clusters as three and the most appropriate clustering algorithm as k-means.

- We have compared the happiness scores for different clusters and have done some exploratory data analysis to understand which indicators contribute the most to each cluster.

- We have ranked top **ten** countries in each of three clusters according to their happiness score. The three leading countries in terms of happiness from cluster 1, cluster 2, and cluster 3 are **Oman**, **Denmark** and **Guyana** respectively, followed by United Arab Emirates, Finland, and Pakistan.

- For validation purpose of our clustering algorithm, we have selected four popular *machine learning classification algorithms* to compare with. We got outstanding classification accuracy, which was also pretty consistent throughout the four methods implying that our cluster has been instrumental.

- The following Figure 9.13 shows that our clustering has been very useful if we compare it with Figure 9.1 in section 9.1. As we have guessed earlier, the happiest countries are those which fall into cluster 2 (green), followed by cluster 1 (red) and cluster 3 (blue).

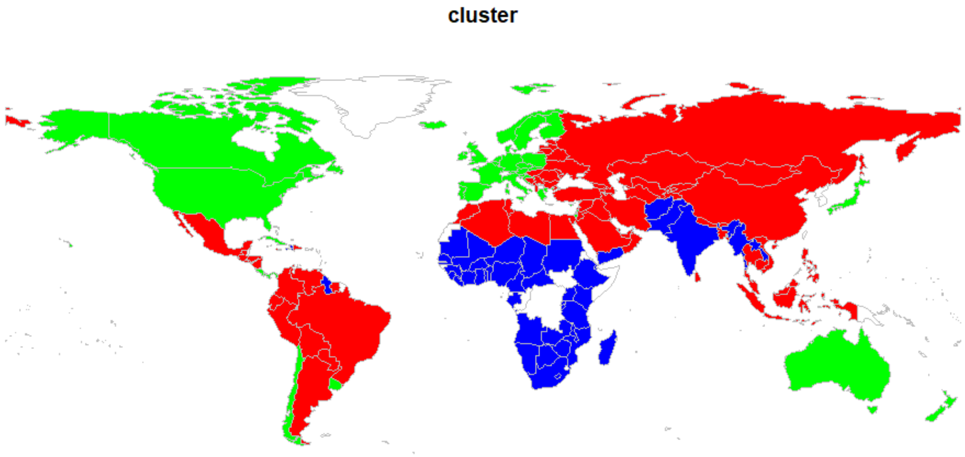


Figure 9.13: Showing The Distribution Of Three Clusters In World Map

The way we have ranked that countries in different clusters by happiness score, one can rank the countries based on all indicators. This will provide tremendous amount of information about the economic condition of individual countries and also at the same time those countries with low score would be able to understand on which indicators they are supposed to be working on.

One might also try different types of clustering algorithms such as PAM (Partition Around Medoids), Hierarchical clustering, etc., and can also evaluate their accuracy by using different classification algorithms. It would also be interesting to investigate the performance of dimension reduction techniques as PCA (Principal Component Analysis), PLS (Partial Least Square) and Factor Analysis techniques to use the components as potential indicators to predict happiness score in future.

Chapter 10: Conclusion and Future Work

10.1 Research on Bayesian Prior Selection Problem

One of the most important aspect of Bayesian analysis to select the probability distribution of the most appropriate prior. I am interested in performing Empirical Bayesian analysis by identifying the priors of the parameters via different re-sampling techniques to achieve this goal. If the probabilistic characteristic of the underlying distribution of the prior is found to be of multiparameter, Copula methods can be used to obtain the bivariate/multivariate empirical Bayesian estimates. Finally, we can compare the Empirical Bayesian estimates with the parametric (MLE/PWM, etc.) and non-parametric (KDE) methods, and perform sensitivity analysis.

10.2 Research on the clustering time-dependent data using non-parametric Kernel Density Estimation (KDE) Method

One of my future research goals is to perform research on non-parametric Kernel Density Estimation (KDE) Method for time-dependent data.

The following describes the research problem related to heart valve stenosis. Stenosis is the term for a valve that doesn't open properly. The flaps of a valve thicken, stiffen, or fuse. As a result, the valve cannot fully open. Thus, the heart has to work harder to pump blood through the valve, and the body may suffer from a reduced supply of oxygen. We want the valve fully opened to avoid a stroke heart attack. As a research question, one could ask what are the risk factors that cause the shrinking of the valve. Doctors can list multiple risk factors which can cause shrinkage of the valve but if we can get the ECG signal data (non-stationary signals) we can build a model where the response variable would be the diameter of the artery. For example, a person goes to a cardiologist with a cardiac problem. The doctor might do a stress test to see how well his heart handles physical activity. Some heart disorders are easier to find when the heart is hard at work. During a stress test, the heart will be checked while he exercises on a treadmill (walking, jogging, and running) or stationary bicycle. If someone is having trouble completing the stress test in a specified period, it may mean there is reduced blood flow to his heart. Reduced blood flow can be caused by several different heart conditions, some of which are very serious. During the stress test, different signals correspond to different diseases (which can be obtained in different stages, for example walking, jogging, and running or stationary). For example, a normal signal implies a person has a normal heart and no potential indication of coronary heart disease. A different signal might indicate disease A, another signal might indicate disease B, another signal might indicate A and B both (interaction), and so on. Given these non-stationary signals, we can perform a cluster analysis. If a signal falls in any particular cluster, it indicates that a person has a certain disease(s). In the absence of real data, if possible, a time-dependent simulation can be done, and the analysis can be performed.

10.3 Work on statistical methods for High Dimensional data

High-dimensional statistics is concerned with data sets in which the number of features is equal to or greater than the number of observations. Because classical theory and methodology can fail in surprising and unexpected ways, data sets of this type present a variety of new challenges, which is one of my future research goals.

10.4 Developing a highly accurate clustering algorithm based on the demands of the clients

Based on the most significant financial indicators that we have identified via analytical modeling, my future goal is to develop a clustering algorithm that can categorize every healthcare stocks of S&P 500 based on any specific criteria of a customer/client (for example, high dividend, low beta risk, and high price to earnings (PE) ratio.)

References

[1] Ahmad, L., Eshlaghy, A., Poorebrahimi, A., Ebrahimi, M., and Razavi, A. (2013). Using three machine learning techniques for predicting breast cancer prediction. *Journal of Health and medical informatics*, 4(2):1–3.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

[3] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769.

[4] Arnold, B. C. and Press, S. J. (1989). Bayesian estimation and prediction for pareto data. *Journal of the American Statistical Association*, 84(408):1079–1084.

[5] Asano, J., Hirakawa, A., and Hamada, C. (2014). Assessing the prediction accuracy of cure in the cox proportional hazards cure model: an application to breast cancer data. *Pharmaceutical statistics*, 13(6):357–363.

[6] Ascher, H. and Feingold, H. (1984). Repairable systems reliability: modeling, inference, misconceptions and their causes. *Journal of The Royal Statistical Society Series C-applied Statistics*.

[7] Avouyi-Dovi, S., Matheron, J., et al. (2006). Productivity and stock prices. *FSR FINANCIAL*, page 81.

[8] Bain, L. J. and Engelhardt, M. (1991). Statistical analysis of reliability and life testing models. marcel & decker. *New York*.

[9] Balaeva, A. Y. and Belyakov, A. A. (2020). Development of an economic and mathematical model for investing in personnel. *Vestnik of Samara University. Economics and Management*, 11(2):92–101.

[10] Banerjee, A. and Dave, R. N. (2004). Validating clusters using the hopkins statistic. In *2004 IEEE International conference on fuzzy systems (IEEE Cat. No. 04CH37542)*, volume 1, pages 149–153. IEEE.

[11] Bassin, W. (1969). Increasing hazard functions and overhaul policy. In *Proc. 1969 Ann. Symp. Reliability*, pages 173–178. IEEE Chicago.

[12] Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

[13] Berwick, D. M., Murphy, J. M., Goldman, P. A., Ware Jr, J. E., Barsky, A. J., and Weinstein, M. C. (1991). Performance of a five-item mental health screening test. *Medical care*, pages 169–176.

[14] Bezdek, J. C. and Hathaway, R. J. (2002). Vat: A tool for visual assessment of (cluster) tendency. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, volume 3, pages 2225–2230. IEEE.

[15] Bland, J. and Altman, G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 1(8476):307.

[16] Borg, M., Badr, I., and Royle, G. (2012). The use of a figure-of-merit (fom) for optimisation in digital mammography: a literature review. *Radiation protection dosimetry*, 151(1):81–88.

[17] Bradburn, N. M. (1969). The structure of psychological well-being.

[18] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). *Classification and regression trees*. Routledge.

[19] Brock, G., Pihur, V., Datta, S., and Datta, S. (2008). clvalid: An r package for cluster validation. *Journal of Statistical Software*, 25:1–22.

[20] Brody, T. (2016). *Clinical trials: study design, endpoints and biomarkers, drug safety, and FDA and ICH guidelines*. Academic press.

[21] Cammock, T., Joseph, S., and Lewis, C. A. (1994). Personality correlates of scores on the depression-happiness scale. *Psychological Reports*, 75(3_suppl):1649–1650.

[22] Case, L. D., Kimmick, G., Paskett, E. D., Lohman, K., and Tucker, R. (2002). Interpreting measures of treatment effect in cancer clinical trials. *The oncologist*, 7(3):181–187.

[23] Castillo, E. and Hadi, A. S. (1995). A method for estimating parameters and quantiles of distributions of continuous random variables. *Computational statistics & data analysis*, 20(4):421–439.

[24] Castillo, E. and Hadi, A. S. (1997). Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620.

[25] Chaipornkaew, P. and Prexawanprasut, T. (2019). A prediction model for human happiness using machine learning techniques. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 33–37. IEEE.

[26] Chakraborty, A. and Tsokos, C. (2021a). Survival analysis for pancreatic cancer patients using cox-proportional hazard (cph) model. *Global Journal of Medical Research*.

[27] Chakraborty, A. and Tsokos, C. P. (2021b). A real data-driven analytical model to predict happiness. *Scholars Journal of Physics, Mathematics and Statistics*.

[28] Chakraborty, A., Tsokos, C. P., et al. (2021). Parametric and non-parametric survival analysis of patients with acute myeloid leukemia (aml). *Open Journal of Applied Sciences*, 11(01):126.

[29] Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61:1–36.

[30] Chen, T., He, T., Benesty, M., and Khotilovich, V. (2019). Package â€˜xgboostâ€™. *R version*, 90.

[31] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

[32] Chen, Y., Jia, Z., Mercola, D., and Xie, X. (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and mathematical methods in medicine*, 2013.

[33] Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43(4):478–484.

[34] Chun, S. H., Chun, J., Lee, K.-Y., and Sung, T.-J. (2018). Effects of emergency cerclage on the neonatal outcomes of preterm twin pregnancies compared to preterm singleton pregnancies: A neonatal focus. *Plos one*, 13(11):e0208136.

[35] Cicchetti, D. V. (1992). Neural networks and diagnosis in the clinical laboratory: state of the art. *Clinical chemistry*, 38(1):9–10.

[36] Cochran, A. J. (1997). Prediction of outcome for patients with cutaneous melanoma. *Pigment cell research*, 10(3):162–167.

[37] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

[38] De Haan, L., Ferreira, A., and Ferreira, A. (2006). *Extreme value theory: an introduction*, volume 21. Springer.

[39] del Castillo, J. and Daoudi, J. (2009). Estimation of the generalized pareto distribution. *Statistics & Probability Letters*, 79(5):684–688.

[40] Della Mea, V., Demartini, G., Di Gaspero, L., and Mizzaro, S. (2006). Measuring retrieval effectiveness with average distance measure (adm). *Information Wissenschaft und Praxis*, 57(8):433–443.

[41] DeNeve, K. M. and Cooper, H. (1998). The happy personality: a meta-analysis of 137 personality traits and subjective well-being. *Psychological bulletin*, 124(2):197.

[42] Derringer, G. and Suich, R. (1980). Simultaneous optimization of several response variables. *Journal of quality technology*, 12(4):214–219.

[43] Di Tella, R. and MacCulloch, R. (2006). Some uses of happiness data in economics. *Journal of economic perspectives*, 20(1):25–46.

[44] Dorn, D., Fischer, J. A., Kirchgässner, G., and Sousa-Poza, A. (2007). Is it culture or democracy? the impact of democracy and culture on happiness. *Social Indicators Research*, 82(3):505–526.

[45] Du, X., Li, M., Zhu, P., Wang, J., Hou, L., Li, J., Meng, H., Zhou, M., and Zhu, C. (2018). Comparison of the flexible parametric survival model and cox model in estimating markov transition probabilities using real-world data. *PloS one*, 13(8):e0200807.

[46] Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, pages 826–838.

[47] Ferlay, J., Shin, H., Bray, F., Forman, D., Mathers, C., and Parkin, D. (2008). Globocan 2008, cancer incidence and mortality worldwide: Iarc cancerbase no. 10. *Lyon, France: International Agency for Research on Cancer*.

[48] Fernandez, L. (1986). Non-parametric maximum likelihood estimation of censored regression models. *Journal of Econometrics*, 32(1):35–57.

[49] Frey, B. S. and Stutzer, A. (2000). Happiness prospers in democracy. *Journal of happiness Studies*, 1(1).

[50] Friedman, J. H. (1999). Greedy function approximation: a gradient boosting machine 1 function estimation 2 numerical optimization in function space. *North*, 1(3):1–10.

[51] Garbin, C., Zhu, X., and Marques, O. (2020). Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79(19):12777–12815.

[52] Garner, C. A. et al. (2006). Should the decline in the personal saving rate be a cause for concern? *Economic Review-Federal Reserve Bank of Kansas City*, 91(2):5.

[53] George, M. S., Ketter, T. A., Parekh, P. I., Horwitz, B., Herscovitch, P., Post, R. M., et al. (1995). Brain activity during transient sadness and happiness in healthy women. *American Journal of Psychiatry*, 152(3):341–351.

[54] Giancarlo, R., Scaturro, D., and Utro, F. (2008). Computational cluster validation for microarray data analysis: experimental assessment of clest, consensus clustering, figure of merit, gap statistics and model explorer. *BMC bioinformatics*, 9(1):1–19.

[55] Gómez-Ríos, A., Luengo, J., and Herrera, F. (2017). A study on the noise label influence in boosting algorithms: Adaboost, gbm and xgboost. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 268–280. Springer.

[56] Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

[57] Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water resources research*, 15(5):1049–1054.

[58] Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized pareto distribution. *Technometrics*, 35(2):185–191.

[59] Halinski, R. S. and Feldt, L. S. (1970). The selection of variables in multiple regression analysis. *Journal of Educational Measurement*, 7(3):151–157.

[60] Harrington, E. C. (1965). The desirability function. *Industrial quality control*, 21(10):494–498.

[61] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

[62] Hastie, T., Tibshirani, R., and Walther, G. (2001). Estimating the number of data clusters via the gap statistic. *J Roy Stat Soc B*, 63:411–423.

[63] Hayward, J., Alvarez, S. A., Ruiz, C., Sullivan, M., Tseng, J., and Whalen, G. (2010). Machine learning of clinical performance in a pancreatic cancer database. *Artificial intelligence in medicine*, 49(3):187–195.

[64] Helliwell, J. F., Huang, H., and Wang, S. (2019). Changing world happiness. *World happiness report 2019*, 2:11–46.

[65] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[66] Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349.

[67] Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based boosting 2.0. *The Journal of Machine Learning Research*, 11:2109–2113.

[68] Howell, R. T. and Howell, C. J. (2008). The relation of economic status to subjective well-being in developing countries: a meta-analysis. *Psychological bulletin*, 134(4):536.

[69] Howrey, E. P. (2001). The predictive power of the index of consumer sentiment. *Brookings papers on economic activity*, 2001(1):175–207.

[70] Ilic, M. and Ilic, I. (2016). Epidemiology of pancreatic cancer. *World journal of gastroenterology*, 22(44):9694.

[71] Jackson, M. O. (2008). Average distance, diameter, and clustering in social networks with homophily. In *International Workshop on Internet and Network Economics*, pages 4–11. Springer.

[72] Jauhari, F. and Supianto, A. A. (2019). Building studentâ€™s performance decision tree classifier using boosting algorithm. *Indones. J. Electr. Eng. Comput. Sci*, 14(3):1298–1304.

[73] Jiang, W. (2002). On weak base hypotheses and their implications for boosting regression and classification. *The Annals of Statistics*, 30(1):51–73.

[74] Johnson, N. L. (1949). Bivariate distributions based on simple translation systems. *Biometrika*, 36(3/4):297–304.

[75] Kahng, S. E., Benayahu, Y., Wagner, D., and Rothe, N. (2008). Sexual reproduction in the invasive octocoral carijoa riisei in hawaii. *Bulletin of Marine Science*, 82(1):1–17.

[76] Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

[77] Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda.

[78] Kaufmann, D., Kraay, A., and Mastruzzi, M. (2010). Response to â€˜what do the worldwide governance indicators measure?â€™. *The European Journal of Development Research*, 22(1):55–58.

[79] Kermani, F. (2017). Validation of clustering methods for medical data sets. *Acta-HealthMedica*, 2(1):116–116.

[80] Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C. D., Biankin, A. V., Neale, R. E., Tempero, M., Tuveson, D. A., Hruban, R. H., et al. (2016). Pancreatic cancer. *Nature reviews Disease primers*, 2(1):1–22.

[81] Kleinbaum, D. G. and Klein, M. (2012). Kaplan-meier survival curves and the log-rank test. In *Survival analysis*, pages 55–96. Springer.

[82] Kleinbaum, D. G., Klein, M., et al. (2012). *Survival analysis: a self-learning text*, volume 3. Springer.

[83] Koivumaa-Honkanen, H., Honkanen, R., Viinamäki, H., Heikkilä, K., Kaprio, J., and Koskenvuo, M. (2000). Self-reported life satisfaction and 20-year mortality in healthy finnish adults. *American Journal of Epidemiology*, 152(10):983–991.

[84] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17.

[85] Kumar, R. (2015). *Valuation: theories and concepts*. Academic Press.

[86] Lajevardi, S. (2014). A study on the effect of p/e and peg ratios on stock returns: Evidence from tehran stock exchange. *Management Science Letters*, 4(7):1401–1410.

[87] Lemmon, M. and Portniaguina, E. (2006). Consumer confidence and asset prices: Some empirical evidence. *The Review of Financial Studies*, 19(4):1499–1529.

[88] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867.

[89] Li, D., Xie, K., Wolff, R., and Abbruzzese, J. L. (2004). Pancreatic cancer. *The Lancet*, 363(9414):1049–1057.

[90] Lu, H., Wang, H., and Yoon, S. W. (2019). A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. *Expert Systems with Applications*, 116:340–350.

[91] Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., and Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in biology and medicine*, 121:103761.

[92] McDOWELL, I. and Praught, E. (1982). On the measurement of happiness: an examination of the bradburn scale in the canada health survey. *American journal of epidemiology*, 116(6):949–958.

[93] Michaud, D. (2004). Epidemiology of pancreatic cancer. *Minerva chirurgica*, 59(2):99–111.

[94] Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology*, 14(1):1–13.

[95] Myers, D. G. and Diener, E. (1996). The pursuit of happiness. *Scientific American*, 274(5):70–72.

[96] Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.

[97] O'brien, P. C. (1988). Comparing two samples: extensions of the t, rank-sum, and log-rank tests. *Journal of the American Statistical Association*, 83(401):52–61.

[98] Palade, V. and Bocaniala, C. D. (2006). *Computational intelligence in fault diagnosis*. Springer Science & Business Media.

[99] Pandya, R. and Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 117(16):18–21.

[100] Pang, S.-l. and Gong, J.-z. (2009). C5. 0 classification algorithm and application on individual credit evaluation of banks. *Systems Engineering-Theory & Practice*, 29(12):94–104.

[101] Park, K., Ali, A., Kim, D., An, Y., Kim, M., and Shin, H. (2013). Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, 26(9):2194–2205.

[102] Perera, M., Tsokos, C., et al. (2018). A statistical model with non-linear effects and non-proportional hazards for breast cancer survival analysis. *Advances in Breast Cancer Research*, 7(01):65.

[103] Perneger, T. V., Hudelson, P. M., and Bovier, P. A. (2004). Health and happiness in young swiss adults. *Quality of Life Research*, 13(1):171–178.

[104] Pettijohn, T. F. and Pettijohn, T. F. (1996). Perceived happiness of college students measured by maslow's hierarchy of needs. *Psychological Reports*, 79(3):759–762.

[105] Pham, M. H., Tsokos, C., and Choi, B.-J. (2019). Maximum likelihood estimation for the generalized pareto distribution and goodness-of-fit test with censored data. *Journal of Modern Applied Statistical Methods*, 17(2):11.

[106] Polansky, A. M., Chou, Y.-M., and Mason, R. L. (1999). An algorithm for fitting johnson transformations to non-normal data. *Journal of quality technology*, 31(3):345–350.

[107] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[108] Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34.

[109] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

[110] Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94.

[111] Samundeeswari, E. and Kiruthika, M. (2018). Clustering similar images using various image. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 7(3).

[112] Sashegyi, A. and Ferry, D. (2017). On the interpretation of the hazard ratio and communication of survival benefit. *The oncologist*, 22(4):484–486.

[113] Schmid, M. and Hothorn, T. (2008). Flexible boosting of accelerated failure time models. *BMC bioinformatics*, 9(1):1–13.

[114] Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

[115] Sengupta, A. (2009). *Advances in multivariate statistical methods*, volume 4. World scientific.

[116] Sheha, M. A. A. and Tsokos, C. P. (2019). Statistical modeling of emission factors of fossil fuels contributing to atmospheric carbon dioxide in africa. *Atmospheric and Climate Sciences*, 9(3):438–455.

[117] Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., and Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 56(12):2353–2360.

[118] Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., and Chai, C. (2019). A feature learning approach based on xgboost for driving assessment and risk prediction. *Accident Analysis & Prevention*, 129:170–179.

[119] Siekierski, K. (1992). Comparison and evaluation of three methods of estimation of the johnson sb distribution. *Biometrical Journal*, 34(7):879–895.

[120] Singh, V. P. and Guo, H. (1995). Parameter estimation for 3-parameter generalized pareto distribution by the principle of maximum entropy (pome). *Hydrological sciences journal*, 40(2):165–181.

[121] Song, R., Chen, S., Deng, B., and Li, L. (2016). extreme gradient boosting for identifying individual users across different digital devices. In *International Conference on Web-Age Information Management*, pages 43–54. Springer.

[122] Soukissian, T. (2013). Use of multi-parameter distributions for offshore wind speed modeling: The johnson sb distribution. *Applied energy*, 111:982–1000.

[123] Strauss, D. J., Shavelle, R. M., and Ashwal, S. (1999). Life expectancy and median survival time in the permanent vegetative state. *Pediatric neurology*, 21(3):626–631.

[124] Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24:303–329.

[125] Svozil, D., Kvasnicka, V., and Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1):43–62.

[126] Tang, G. Y. and Shum, W. C. (2003). The conditional relationship between beta and returns: recent evidence from international stock markets. *International Business Review*, 12(1):109–126.

[127] Telgarsky, M. and Vattani, A. (2010). Hartiganâ€™s method: k-means clustering without voronoi. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 820–827. JMLR Workshop and Conference Proceedings.

[128] Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). Package â€˜rpartâ€™. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016).*

[129] Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.

[130] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[131] Tootell, G. M. (1998). Globalization and us inflation. *New England Economic Review*, page 21.

[132] Tsokos, C. P. (1995). Reliability growth: Nonhomogeneous poisson. *Recent Advances in Life-Testing and Reliability*, page 319.

[133] Vieira, S. M., Kaymak, U., and Sousa, J. M. (2010). Cohen's kappa coefficient as a performance measure for feature selection. In *International Conference on Fuzzy Systems*, pages 1–8. IEEE.

[134] Vincent, A., Herman, J., Schulick, R., Hruban, R. H., and Goggins, M. (2011). Pancreatic cancer. *The lancet*, 378(9791):607–620.

[135] Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.

[136] Walsh, J., Joseph, S., and Lewis, C. A. (1995). Internal reliability and convergent validity of the depression-happiness scale with the general health questionnaire in an employed adult sample. *Psychological Reports*, 76(1):137–138.

[137] White, K. J. (1992). The durbin-watson test for autocorrelation in nonlinear models. *The Review of Economics and Statistics*, pages 370–373.

[138] Winnett, A. and Sasieni, P. (2001). Miscellanea. a note on scaled schoenfeld residuals for the proportional hazards model. *Biometrika*, 88(2):565–571.

[139] Wong, W. H. (1986). Theory of partial likelihood. *The Annals of statistics*, pages 88–123.

[140] Xu, Y., Kepner, J., and P Tsokos, C. (2011). Identify attributable variables and interactions in breast cancer. *Journal of applied sciences*, 11(6):1033–1038.

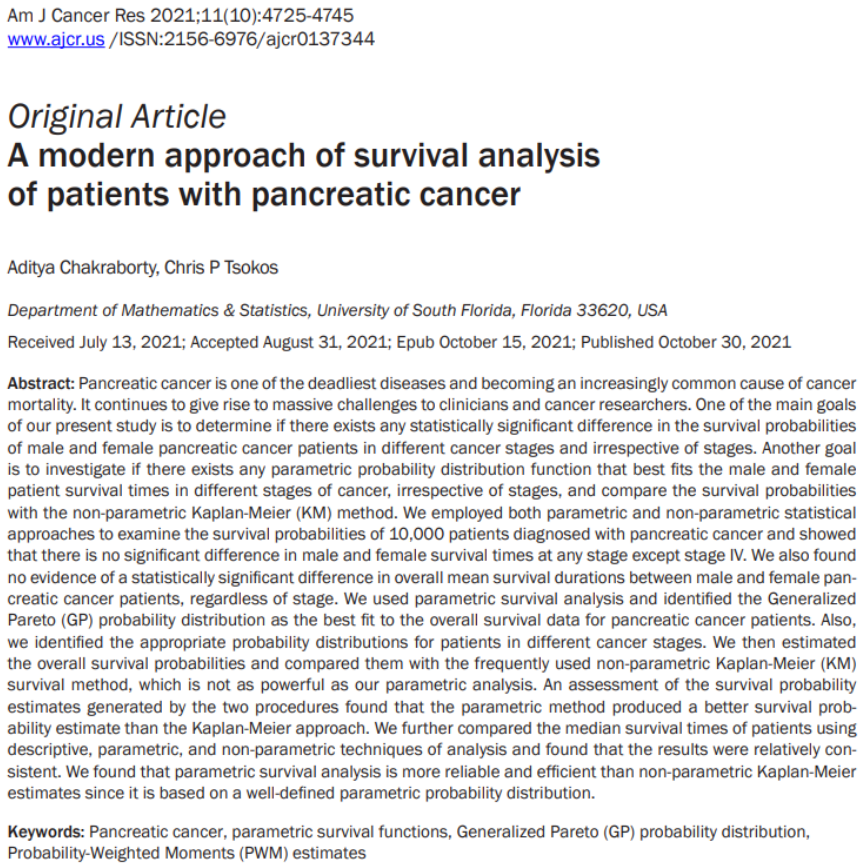
[141] Xu, Y. and Tsokos, C. (2012). Probabilistic survival analysis methods using simulation and cancer data. *Problems of Nonlinear Analysis In Engineering Systems, English/Russian*, 1(37):47–59.

[142] Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

[143] Zhang, J. (2007). Likelihood moment estimation for the generalized pareto distribution. *Australian & New Zealand Journal of Statistics*, 49(1):69–77.

[144] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Appendix A: Abstracts of Published Drafts Included in the Dissertation, and permission emails to publish



Am J Cancer Res 2021;11(10):4725-4745
www.ajcr.us /ISSN:2156-6976/ajcr0137344

Original Article

A modern approach of survival analysis of patients with pancreatic cancer

Aditya Chakraborty, Chris P Tsokos

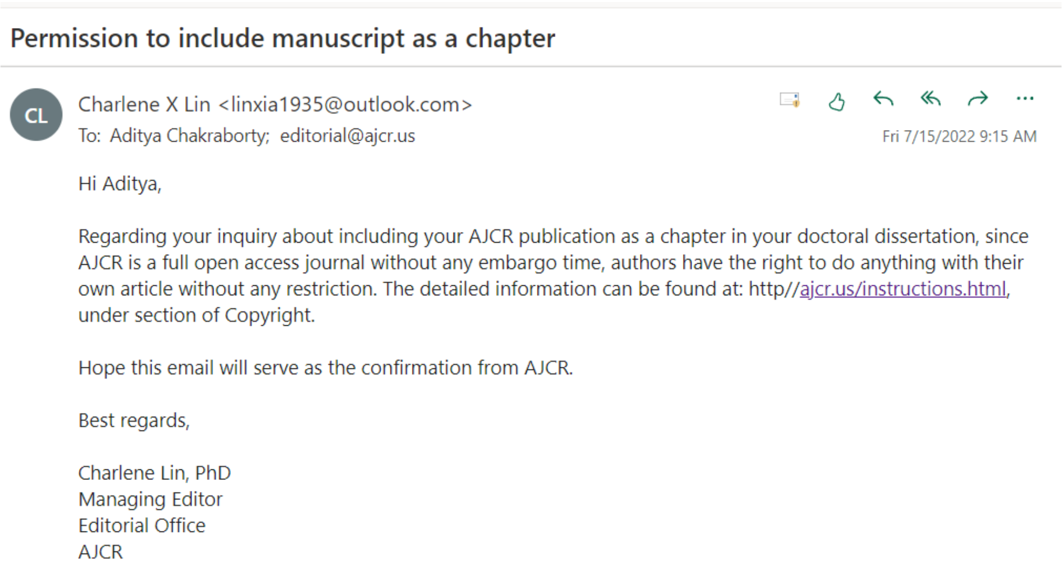
Department of Mathematics & Statistics, University of South Florida, Florida 33620, USA

Received July 13, 2021; Accepted August 31, 2021; Epub October 15, 2021; Published October 30, 2021

Abstract: Pancreatic cancer is one of the deadliest diseases and becoming an increasingly common cause of cancer mortality. It continues to give rise to massive challenges to clinicians and cancer researchers. One of the main goals of our present study is to determine if there exists any statistically significant difference in the survival probabilities of male and female pancreatic cancer patients in different cancer stages and irrespective of stages. Another goal is to investigate if there exists any parametric probability distribution function that best fits the male and female patient survival times in different stages of cancer, irrespective of stages, and compare the survival probabilities with the non-parametric Kaplan-Meier (KM) method. We employed both parametric and non-parametric statistical approaches to examine the survival probabilities of 10,000 patients diagnosed with pancreatic cancer and showed that there is no significant difference in male and female survival times at any stage except stage IV. We also found no evidence of a statistically significant difference in overall mean survival durations between male and female pancreatic cancer patients, regardless of stage. We used parametric survival analysis and identified the Generalized Pareto (GP) probability distribution as the best fit to the overall survival data for pancreatic cancer patients. Also, we identified the appropriate probability distributions for patients in different cancer stages. We then estimated the overall survival probabilities and compared them with the frequently used non-parametric Kaplan-Meier (KM) survival method, which is not as powerful as our parametric analysis. An assessment of the survival probability estimates generated by the two procedures found that the parametric method produced a better survival probability estimate than the Kaplan-Meier approach. We further compared the median survival times of patients using descriptive, parametric, and non-parametric techniques of analysis and found that the results were relatively consistent. We found that parametric survival analysis is more reliable and efficient than non-parametric Kaplan-Meier estimates since it is based on a well-defined parametric probability distribution.

Keywords: Pancreatic cancer, parametric survival functions, Generalized Pareto (GP) probability distribution, Probability-Weighted Moments (PWM) estimates

Figure A.1: Abstract 1



Permission to include manuscript as a chapter

CL Charlene X Lin <linxia1935@outlook.com>
To: Aditya Chakraborty; editorial@ajcr.us

Fri 7/15/2022 9:15 AM

Hi Aditya,

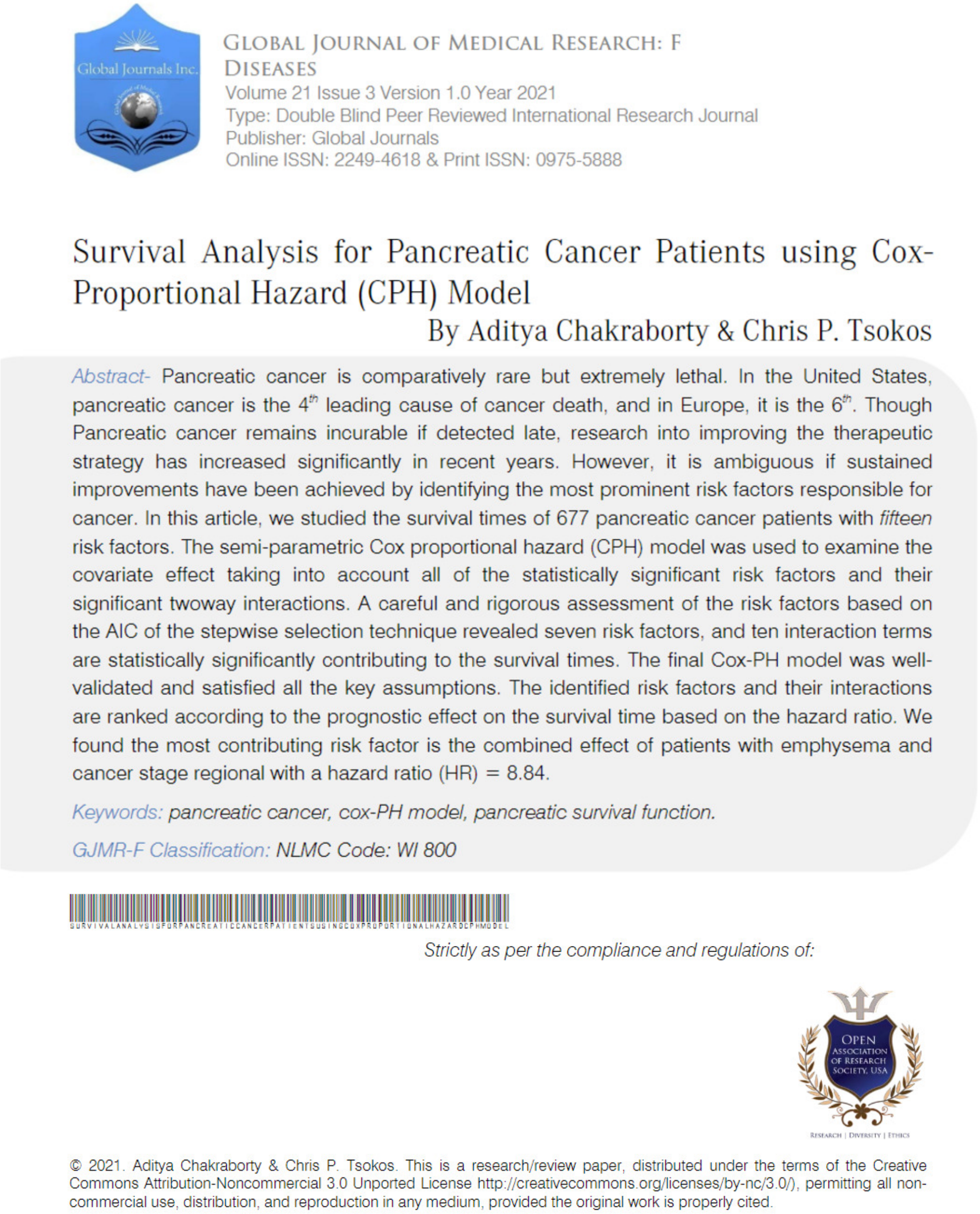
Regarding your inquiry about including your AJCR publication as a chapter in your doctoral dissertation, since AJCR is a full open access journal without any embargo time, authors have the right to do anything with their own article without any restriction. The detailed information can be found at: http//ajcr.us/instructions.html, under section of Copyright.

Hope this email will serve as the confirmation from AJCR.

Best regards,

Charlene Lin, PhD
Managing Editor
Editorial Office
AJCR

Figure A.2: Email AJCR



Global Journal of Medical Research: F
Diseases
Volume 21 Issue 3 Version 1.0 Year 2021
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4618 & Print ISSN: 0975-5888

Survival Analysis for Pancreatic Cancer Patients using Cox-Proportional Hazard (CPH) Model

By Aditya Chakraborty & Chris P. Tsokos

Abstract- Pancreatic cancer is comparatively rare but extremely lethal. In the United States, pancreatic cancer is the 4th leading cause of cancer death, and in Europe, it is the 6th. Though Pancreatic cancer remains incurable if detected late, research into improving the therapeutic strategy has increased significantly in recent years. However, it is ambiguous if sustained improvements have been achieved by identifying the most prominent risk factors responsible for cancer. In this article, we studied the survival times of 677 pancreatic cancer patients with *fifteen* risk factors. The semi-parametric Cox proportional hazard (CPH) model was used to examine the covariate effect taking into account all of the statistically significant risk factors and their significant twoway interactions. A careful and rigorous assessment of the risk factors based on the AIC of the stepwise selection technique revealed seven risk factors, and ten interaction terms are statistically significantly contributing to the survival times. The final Cox-PH model was well-validated and satisfied all the key assumptions. The identified risk factors and their interactions are ranked according to the prognostic effect on the survival time based on the hazard ratio. We found the most contributing risk factor is the combined effect of patients with emphysema and cancer stage regional with a hazard ratio (HR) = 8.84.

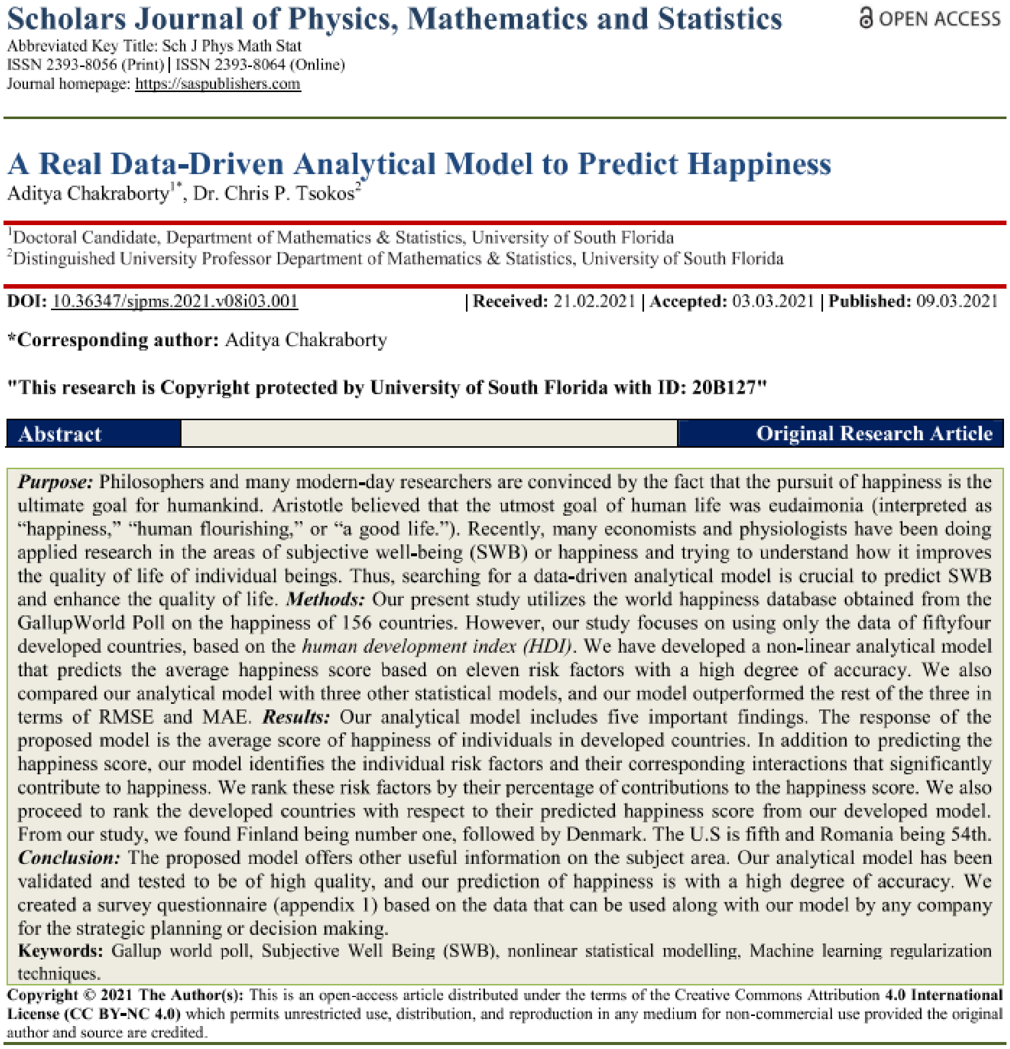
Keywords: pancreatic cancer, cox-PH model, pancreatic survival function.

GJMR-F Classification: NLMC Code: WI 800

Strictly as per the compliance and regulations of:

© 2021. Aditya Chakraborty & Chris P. Tsokos. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure A.3: Abstract 2 showing the creative commons licence statement at the bottom



Scholars Journal of Physics, Mathematics and Statistics

OPEN ACCESS

Abbreviated Key Title: Sch J Phys Math Stat
ISSN 2393-8056 (Print) | ISSN 2393-8064 (Online)
Journal homepage: https://saspublishers.com

A Real Data-Driven Analytical Model to Predict Happiness

Aditya Chakraborty[1*], Dr. Chris P. Tsokos[2]

[1]Doctoral Candidate, Department of Mathematics & Statistics, University of South Florida
[2]Distinguished University Professor Department of Mathematics & Statistics, University of South Florida

DOI: 10.36347/sjpms.2021.v08i03.001 | **Received:** 21.02.2021 | **Accepted:** 03.03.2021 | **Published:** 09.03.2021

***Corresponding author:** Aditya Chakraborty

"This research is Copyright protected by University of South Florida with ID: 20B127"

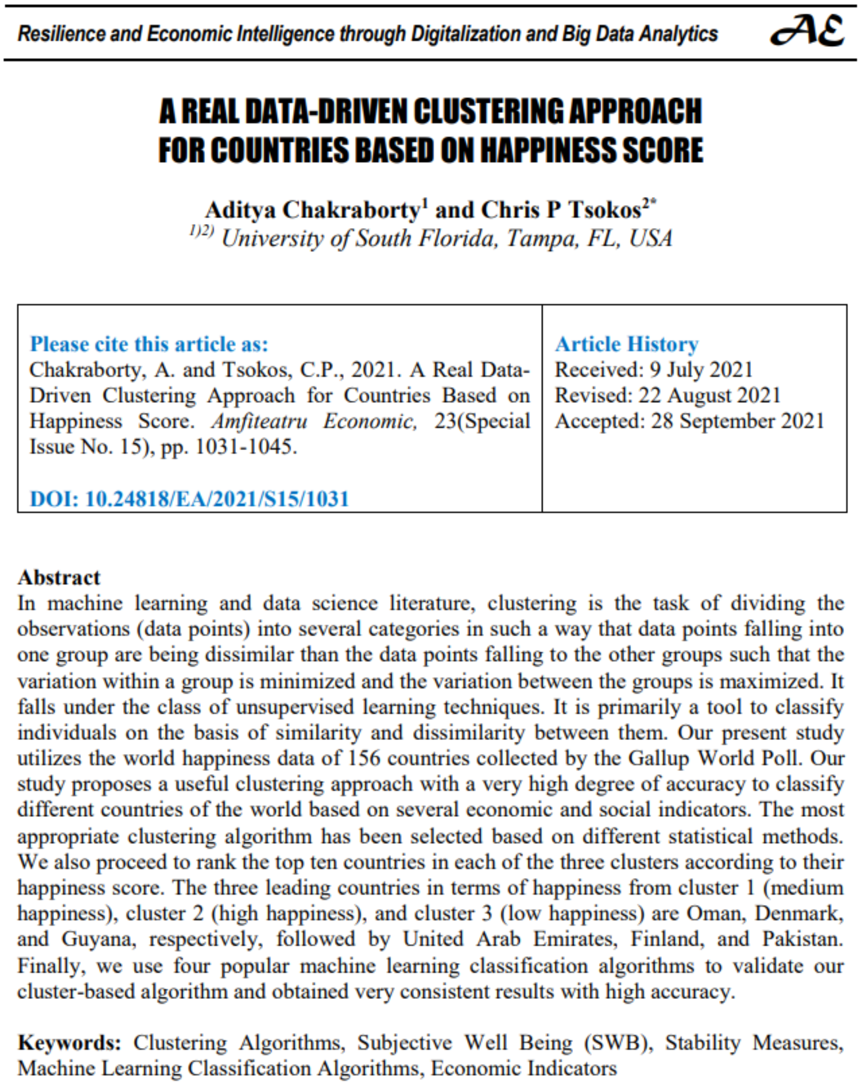
Abstract **Original Research Article**

Purpose: Philosophers and many modern-day researchers are convinced by the fact that the pursuit of happiness is the ultimate goal for humankind. Aristotle believed that the utmost goal of human life was eudaimonia (interpreted as “happiness,” “human flourishing,” or “a good life.”). Recently, many economists and physiologists have been doing applied research in the areas of subjective well-being (SWB) or happiness and trying to understand how it improves the quality of life of individual beings. Thus, searching for a data-driven analytical model is crucial to predict SWB and enhance the quality of life. ***Methods:*** Our present study utilizes the world happiness database obtained from the GallupWorld Poll on the happiness of 156 countries. However, our study focuses on using only the data of fiftyfour developed countries, based on the *human development index (HDI)*. We have developed a non-linear analytical model that predicts the average happiness score based on eleven risk factors with a high degree of accuracy. We also compared our analytical model with three other statistical models, and our model outperformed the rest of the three in terms of RMSE and MAE. ***Results:*** Our analytical model includes five important findings. The response of the proposed model is the average score of happiness of individuals in developed countries. In addition to predicting the happiness score, our model identifies the individual risk factors and their corresponding interactions that significantly contribute to happiness. We rank these risk factors by their percentage of contributions to the happiness score. We also proceed to rank the developed countries with respect to their predicted happiness score from our developed model. From our study, we found Finland being number one, followed by Denmark. The U.S is fifth and Romania being 54th. ***Conclusion:*** The proposed model offers other useful information on the subject area. Our analytical model has been validated and tested to be of high quality, and our prediction of happiness is with a high degree of accuracy. We created a survey questionnaire (appendix 1) based on the data that can be used along with our model by any company for the strategic planning or decision making.

Keywords: Gallup world poll, Subjective Well Being (SWB), nonlinear statistical modelling, Machine learning regularization techniques.

Copyright © 2021 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution **4.0 International License (CC BY-NC 4.0)** which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

Figure A.4: Abstract 3 showing the creative commons licence statement at the bottom



Resilience and Economic Intelligence through Digitalization and Big Data Analytics AE

A REAL DATA-DRIVEN CLUSTERING APPROACH FOR COUNTRIES BASED ON HAPPINESS SCORE

Aditya Chakraborty[1] and Chris P Tsokos[2*]
[1)2)] University of South Florida, Tampa, FL, USA

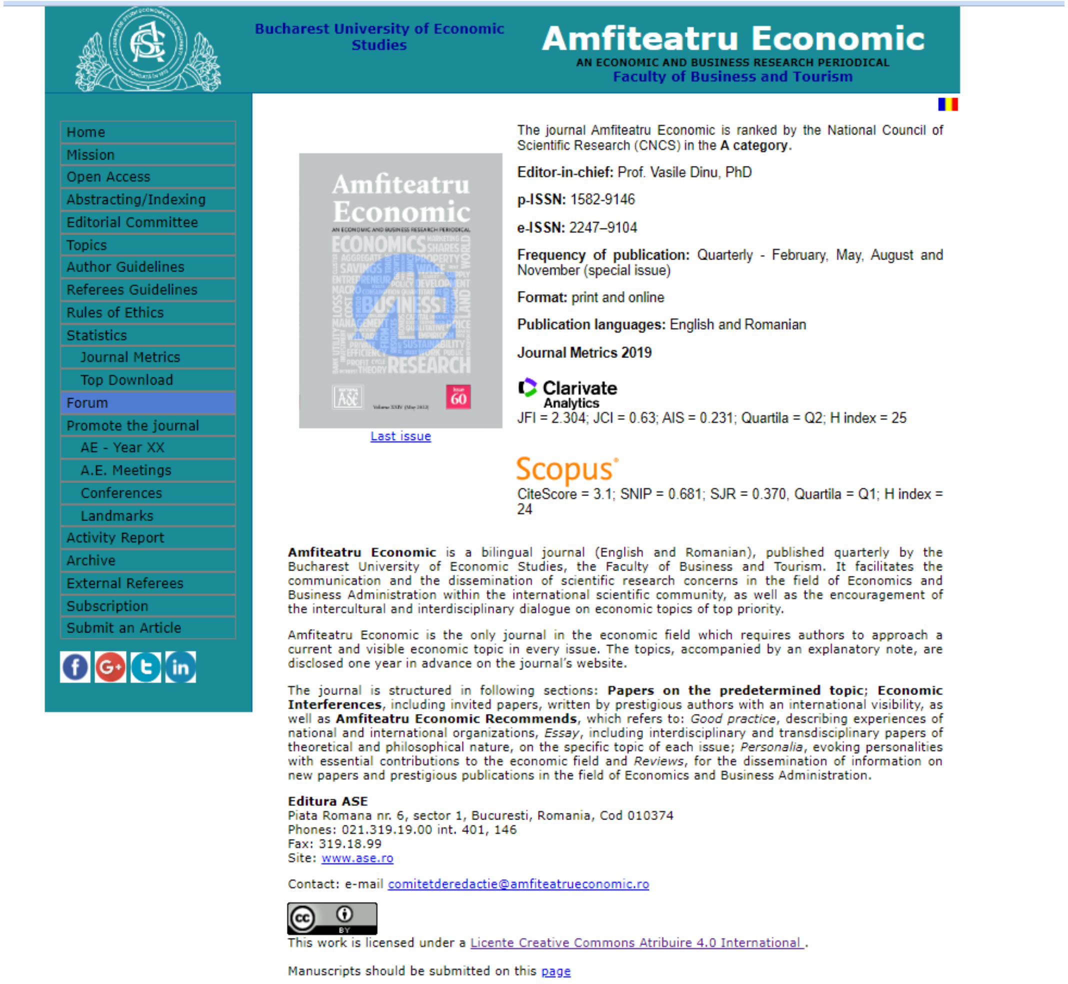
Please cite this article as:	**Article History**
Chakraborty, A. and Tsokos, C.P., 2021. A Real Data-Driven Clustering Approach for Countries Based on Happiness Score. *Amfiteatru Economic,* 23(Special Issue No. 15), pp. 1031-1045. **DOI: 10.24818/EA/2021/S15/1031**	Received: 9 July 2021 Revised: 22 August 2021 Accepted: 28 September 2021

Abstract

In machine learning and data science literature, clustering is the task of dividing the observations (data points) into several categories in such a way that data points falling into one group are being dissimilar than the data points falling to the other groups such that the variation within a group is minimized and the variation between the groups is maximized. It falls under the class of unsupervised learning techniques. It is primarily a tool to classify individuals on the basis of similarity and dissimilarity between them. Our present study utilizes the world happiness data of 156 countries collected by the Gallup World Poll. Our study proposes a useful clustering approach with a very high degree of accuracy to classify different countries of the world based on several economic and social indicators. The most appropriate clustering algorithm has been selected based on different statistical methods. We also proceed to rank the top ten countries in each of the three clusters according to their happiness score. The three leading countries in terms of happiness from cluster 1 (medium happiness), cluster 2 (high happiness), and cluster 3 (low happiness) are Oman, Denmark, and Guyana, respectively, followed by United Arab Emirates, Finland, and Pakistan. Finally, we use four popular machine learning classification algorithms to validate our cluster-based algorithm and obtained very consistent results with high accuracy.

Keywords: Clustering Algorithms, Subjective Well Being (SWB), Stability Measures, Machine Learning Classification Algorithms, Economic Indicators

Figure A.5: Abstract 4



Bucharest University of Economic Studies

Amfiteatru Economic

AN ECONOMIC AND BUSINESS RESEARCH PERIODICAL

Faculty of Business and Tourism

The journal Amfiteatru Economic is ranked by the National Council of Scientific Research (CNCS) in the **A category.**

Editor-in-chief: Prof. Vasile Dinu, PhD

p-ISSN: 1582-9146

e-ISSN: 2247–9104

Frequency of publication: Quarterly - February, May, August and November (special issue)

Format: print and online

Publication languages: English and Romanian

Journal Metrics 2019

Clarivate Analytics

JFI = 2.304; JCI = 0.63; AIS = 0.231; Quartila = Q2; H index = 25

Scopus

CiteScore = 3.1; SNIP = 0.681; SJR = 0.370, Quartila = Q1; H index = 24

Amfiteatru Economic is a bilingual journal (English and Romanian), published quarterly by the Bucharest University of Economic Studies, the Faculty of Business and Tourism. It facilitates the communication and the dissemination of scientific research concerns in the field of Economics and Business Administration within the international scientific community, as well as the encouragement of the intercultural and interdisciplinary dialogue on economic topics of top priority.

Amfiteatru Economic is the only journal in the economic field which requires authors to approach a current and visible economic topic in every issue. The topics, accompanied by an explanatory note, are disclosed one year in advance on the journal's website.

The journal is structured in following sections: **Papers on the predetermined topic**; **Economic Interferences**, including invited papers, written by prestigious authors with an international visibility, as well as **Amfiteatru Economic Recommends**, which refers to: *Good practice*, describing experiences of national and international organizations, *Essay*, including interdisciplinary and transdisciplinary papers of theoretical and philosophical nature, on the specific topic of each issue; *Personalia*, evoking personalities with essential contributions to the economic field and *Reviews*, for the dissemination of information on new papers and prestigious publications in the field of Economics and Business Administration.

Editura ASE
Piata Romana nr. 6, sector 1, Bucuresti, Romania, Cod 010374
Phones: 021.319.19.00 int. 401, 146
Fax: 319.18.99
Site: www.ase.ro

Contact: e-mail comitetderedactie@amfiteatrueconomic.ro

This work is licensed under a Licente Creative Commons Atribuire 4.0 International .

Manuscripts should be submitted on this page

Figure A.6: Amfiteatru Economic Journal Homepage showing the creative commons licence statement at the bottom

Appendix B: Our version of the survey questionnaire for World Happiness Report 2019 by Gallup Poll is posted which is the modified version and we request the same type of information

SURVEY QUESTIONNAIRE

Based on who is requesting the information for an individual that is associated with any Government, Company, Organization, Educational Institutions, etc.

GDP(X1): Per-capita gross domestic product of the country the individual resides (given information)

Social Support(X2): If you were in trouble, do you have relatives or close friends, you can count on to help you whenever you need them? A)YES □ B)NO □

Life Expectancy(X3): From the attached graph, identify your life expectancy. □Years.

Freedom(X4): Are you satisfied with your freedom to choose what you do with your life?
A) YES □ B) NO □

Generosity(X5): Have you donated money to a charity in the past month? A) YES, B) No. If the answer is YES, then how much?

Corruption Perception(X6): Is corruption widespread throughout your government, your company, or your organization? A) YES□ B) NO □

Positive Affect(X7): Happiness, laughter, and enjoyment.

7.1. On a scale of 1 to 10, how **happy** were you for the last five days?
7.2. On a scale of 1 to 10, how much did you **laugh** for the last five days?
7.3. On a scale of 1 to 10, how much did you **enjoy** for the last five days?

Negative Affect(X8): Worry, Sadness, and anger, respectively.

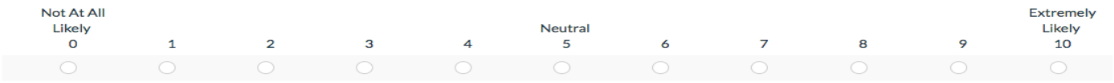
8.1. On a scale of 1 to 10, how **worried** were you for the last five days?
8.2. On a scale of 1 to 10, how **sad** were you for the previous five days?
8.3. On a scale of 1 to 10, how **angry** were you for the last five days?

Confidence in Government(X9): In your government, company, or organization, etc. how much trust and confidence do you have when it comes to handling [International problems/Domestic problems]?

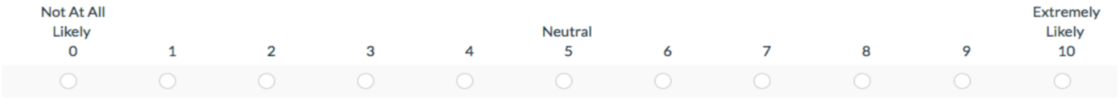
A) A great deal □ B) A fair amount□
C) not very much □ D) none at all □

Democratic Quality (X10):

10.1. On a scale of 1 to 10, how likely do you think that the country's citizens can participate in selecting their government, enjoy Freedom of expression, Freedom of association, and unprejudiced media coverage?



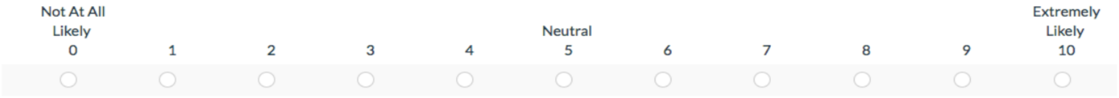
10.2. On a scale of 1 to 10, how likely you think people suffer consequences of political instability and politically motivated violence, including terrorism?



Not At All Likely					Neutral					Extremely Likely
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

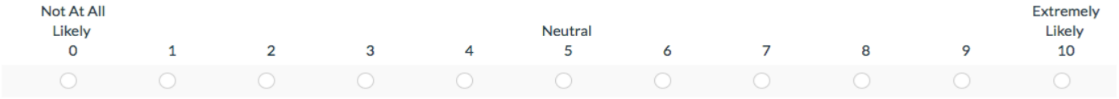
Delivery Quality(X11):

11.1. On a scale of 1 to 10, how likely do you think that your company/organization/government has maintained the quality of public services, the quality of the civil service, the quality of policy formulation and implementation, and the credibility of such policies?



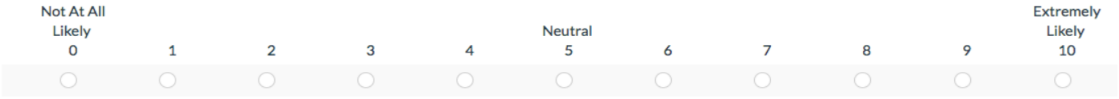
Not At All Likely					Neutral					Extremely Likely
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

11.2. On a scale of 1 to 10, how likely do you think that your company/organization/government can formulate and implement sound policies and regulations that permit and promote private sector development?



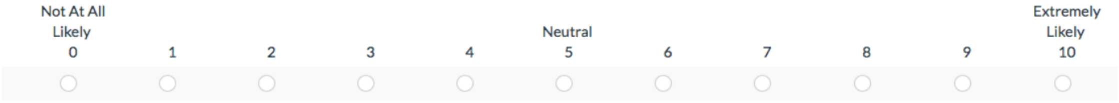
Not At All Likely					Neutral					Extremely Likely
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

11.3. On a scale of 1 to 10, how likely do you think that your company/organization/government agents and law enforcement agencies have confidence in the government and abide by society's rules?



Not At All Likely					Neutral					Extremely Likely
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

11.4. On a scale of 1 to 10, to what extent you think that public power is exercised for private gain, including both petty and grand forms of corruption by the elites for their individual interests?



Not At All Likely					Neutral					Extremely Likely
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○