Paper 11682-2016

Survival Analysis of the Patients Diagnosed with Non-Small Cell Lung Cancer Using SAS® Enterprise Miner™ 13.1

Raja Rajeswari Veggalam, Akansha Gupta; SAS and OSU Data Mining Certificate

Dr. Goutam Chakraborty;

Oklahoma State University

ABSTRACT

Cancer is the second leading cause of deaths in United States. About 85%-90% of all lung cancers are non-small cell lung cancer and they constitute about 26.8% of cancer deaths. An efficient cancer treatment plan is therefore of prime importance for increasing the survival chances of a patient. The cancer treatments are generally given to a patient in multiple sittings and often doctors tend to make the decisions based on improvement over time. Calculating the survival chances of the patient with respect to time and determining the various factors that influence the survival time would help doctors make more informed decisions about the further course of treatment and help patients develop a proactive approach in making choices for the treatment. The objective of this paper is to analyze the survival time of patients suffering from non-small cell lung cancer, identify the time interval crucial for the survival and identify the effects of various factors such as age, gender and treatment type. The performances of the models built are analyzed to understand the significance of cubic splines used in these models. The dataset is from the CERNER database with 548 records and 12 variables from 2009-2013. The patient records with loss to follow up are censored. The survival analysis is performed using parametric and non-parametric methods in SAS Enterprise Miner 13.1. The analysis revealed that the survival probability of a patient is high within two weeks of hospitalization and the probability of survival goes down by 60% between weeks 5 to 9 of admission. Age and gender play a prominent role in influencing the survival time and risk. The probability of survival for female patients decreases by 70% and 80% during week 6 and week 7 respectively and for male patients' decreases by 70% and 50% during week 8 and week 13 respectively.

INTRODUCTION

Non-small lung cancer is a type of lung cancer in which cancer cells start to form in tissues of the lungs. It is lung cancer, which is not of small cell carcinoma type. Non-small cell lung cancer occurs mainly in smokers; however, this lung cancer is also seen in non-smokers as well. This lung cancer is more common in female patients than in male patients.

The main areas of research in this paper are as follows:

To build a non-parametric model for predicting survival time of cancer patients.

To build parametric models to identify the factors influencing the survival time.

To understand difference in survival probabilities with respect to important variables identified.

To evaluate the model performance in accordance to factors identified and to understand the significance of cubic splines in affecting model performance.

CENSORED DATA

The condition in which value of measurement or observation is known partially is called censoring in data mining. Missing values of the observed target variable mean that at the end of the period under consideration, we are still not sure about the outcome of the target. There are three types of censoring.

- Left Censored: Data point is below censored value.
- Right Censored: Data point is above censored value.
- Interval: The data point is between an interval of two values.

The Cerner data under analysis consists of right centered data; the follow up time of a few patients has been missing and hence these records are censored by changing the discharge date to Null

DATA COLLECTION AND PREPARATION

The data source is CERNER database. The data of interest is available in different tables. For example, patient demographics data is in the patient's table, treatment details are available in the medication table, appointment details of the patient are available in the encounter table. By using the patient ID as the primary key, the data of interest is merged from the entire table in to one dataset as shown in the Figure 1.

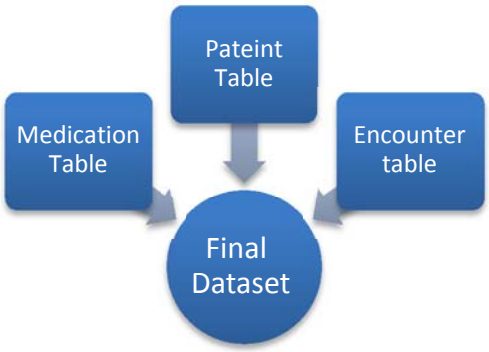


Figure 1: Schematic view of data consolidation

The data available is for the period 2009-2013 and contains 548 records with 12 variables. As the event is death due to non-small cell lung cancer, data of a patient's with status as not mapped or expired indicates incomplete records and censored data.

DATA FORMAT FOR SURVIVAL NODE USING SAS EM:

To facilitate the use of SAS EM and survival node the data should be in the format shown in Figure 2.

- An entity ID or the primary key of the data should be present. Primary key in current data is patient ID that uniquely identifies a patient.
- A start date and end date marks the start and end time of patients in analysis, thereby giving the length of stay at the hospital. Both the start and end dates should be mapped to TIME ID. The data has start date as admitted date and end date as discharge date.
- The target variable in the data set is 'Status', which is a binary variable containing two levels; Survived 1, Expired-0.

Figure 2: SAS EM Data Format

The final dataset has variables as shown in Table 1.

Num.	Variable	Type of Variable	Description
1	Age	Continuous	Age of Patient
2	Gender	Categorical	Gender of Patient
3	Race	Categorical	Race of Patient
4	Martial_Status	Categorical	Marital Status of Patient
5	Admitted_dt	Date	Patient admit date at hospital
6	Discharged_dt	Date	Patient discharged date from hospital
7	Procedure_Description	Varchar	Procedure performed on patient for cure
8	Diagnosis_Description	Varchar	Diagnosis given to patient
9	Age_Level	Categorical	Crossed target variable for age parametric survival analysis 0 – Expired 1 – Patients Surviving in age group 0-25 2 – Patients Surviving in age group 25-50 3 – Patients Surviving in age group 50-75 4 – Patients Surviving in age group 75-100
10	Gender_Level	Categorical	Crossed target variable for gender parametric survival analysis 0 – Expired patients 1 – Surviving male patients 2 – Surviving females patients
11	Status	Categorical	Target variable for non-parametric survival analysis 0 – Expired patients 1 – Surviving patients

Table 1: Final variables in the dataset

From the age wise (Figure 3) and gender wise (Figure 4) distribution, survival rate is highest among the patients within the age group of 63 to 82 followed by patients in the age group of 33 to 62 years. The survival rate is high in males compared to females.

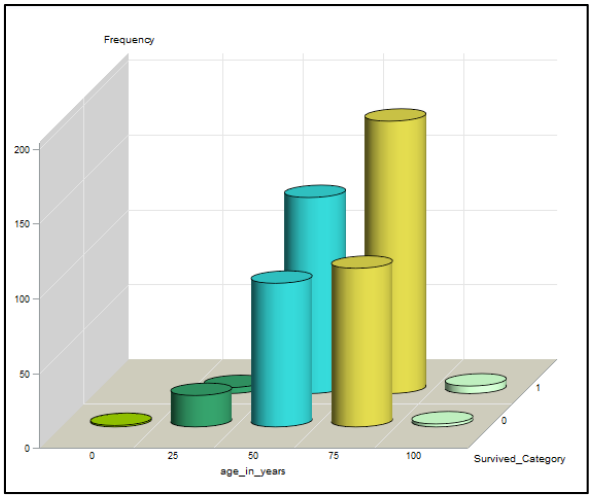


Figure 3 : Gender wise distribution of patients

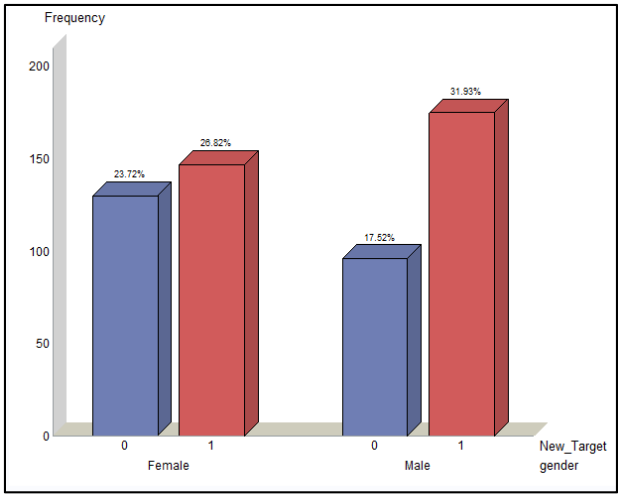


Figure 4: Age wise distribution of patients

CORRELATION

Correlation analysis is performed using PROC CORR for continuous variables: age and number of days admitted and results (Table 2) indicate there is no high correlation between them.

Pearson Correlation Coefficients, N = 548 Prob > |r| under H0: Rho=0		
	age_in_years	No_of_Days_Admitted
age_in_years	1.00000	0.03582
		0.4026
No_of_Days_Admitted	0.03582	1.00000
	0.4026	

Figure 5: Correlation Analysis

SURVIVAL FUNCTION AND HAZARD FUNCTION

Survival Analysis analyzes the time to event by calculating two functions:

The Hazard function H(x) gives the probability that an event that has not occurred at time ‘t-1’ will occur at time‘t’. It is also called failure rate and is given by the formula below:

$$H(x) = P(x)/S(x)$$

Where P(x) = Probability density function, S(x) = Survival function.

Survival function S(x) provides an estimated duration that gives the probability that observant under study will survive beyond specified time. It tells that the observant under study did not experience the event at time t-1 and will not experience the event at time t.

$$S(x) = 1-H(x)$$

Where h(x) = Hazard function.

The hazard function takes various shapes and the non-linear hazard functions are handled through cubic splines.

CUBIC SPLINE FUNCTIONS:

Splines are simplified functions of mathematics defined by polynomials. To model nonlinear shapes of hazard function, cubic splines are used. Cubic spline functions are segmented functions consisting of cubic functions joined together through knots represented in Figure 6.

$$csb(t, k_j) = \begin{cases} -t^3 + 3k_j t^2 - 3k_j^2 t & if \ \ t \le k_j \\ -k_j^3 & if \ t > k_j \end{cases}$$

Figure 6: Mathematical representation of cubic spline function.

Cubic splines are incorporated into the model:

- To smoothen the curves.
- To increase the flexibility of the model.
- To increase the model performance.

NON-PARAMETRIC SURVIVAL ANALYSIS

A non-parametric estimate of survival function models the survival probabilities as functions of time. The most common estimate under this category is the Kaplan Meier estimate and this algorithm is by default employed in the Survival Analysis node of SAS enterprise miner 13.1. It uses life test procedure.

The survival function and the hazard function of the cancer patients is estimated by running the survival node keeping the target as 'status' and time interval of 'week'.

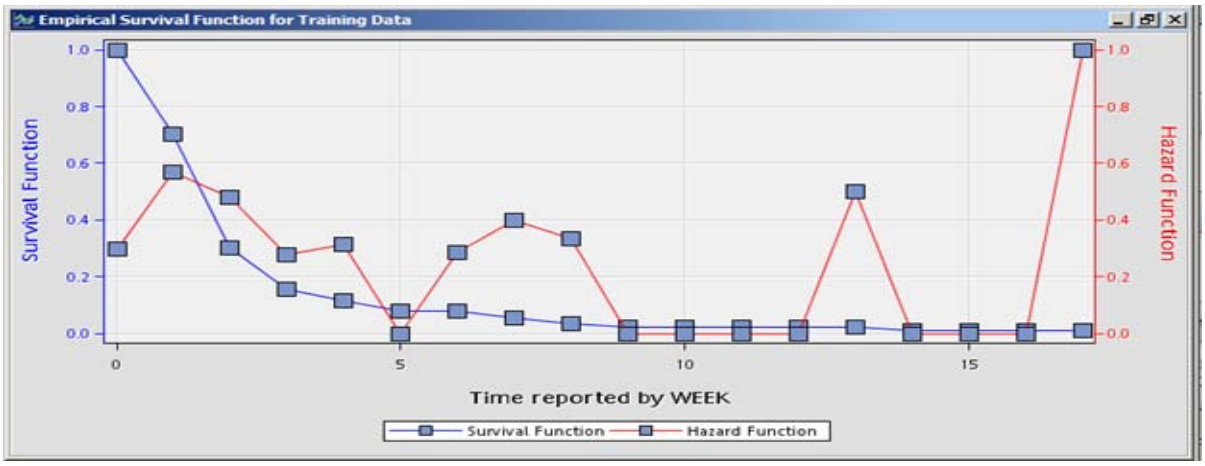


Figure 7: Empirical survival function and hazard function plot

The above plot indicates that the survival rate of the cancer patients is high in the first week but with subsequent weeks the probability decreases. From the hazard plot (Figure 7) the highest risk of surviving is observed at week 2, week 4, week 7, and week 13.

The LIFETEST Procedure

Life Table Survival Estimates

Interval [Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	PDF	PDF Standard Error	Hazard	Hazard Standard Error
												Evaluated at the Midpoint of the Interval			
0	1	162	202	447.0	0.3624	0.0227	1.0000	0	0	1.3699	0.0636	0.3624	0.0227	0.442623	0.033913
1	2	105	8	180.0	0.5833	0.0367	0.6376	0.3624	0.0227	0.8571	0.0639	0.3719	0.0269	0.823529	0.073239
2	3	34	8	67.0	0.5075	0.0611	0.2657	0.7343	0.0253	0.9853	0.1204	0.1348	0.0207	0.68	0.109672
3	4	8	5	26.5	0.3019	0.0892	0.1308	0.8692	0.0205	1.8514	0.4174	0.0395	0.0132	0.355556	0.123705
4	5	5	2	15.0	0.3333	0.1217	0.0913	0.9087	0.0184	2.8750	0.6778	0.0304	0.0127	0.4	0.175271
5	6	0	2	8.0	0	0	0.0609	0.9391	0.0166	2.7500	0.6187	0	.	0	.
6	7	2	0	7.0	0.2857	0.1707	0.0609	0.9391	0.0166	1.7500	0.6614	0.0174	0.0114	0.333333	0.232406
7	8	2	0	5.0	0.4000	0.2191	0.0435	0.9565	0.0158	1.5000	1.1180	0.0174	0.0114	0.5	0.342327
8	9	1	0	3.0	0.3333	0.2722	0.0261	0.9739	0.0134	5.5000	0.8660	0.00870	0.00840	0.4	0.391918
9	10	0	0	2.0	0	0	0.0174	0.9826	0.0114	5.0000	0.7071	0	.	0	.
10	11	0	0	2.0	0	0	0.0174	0.9826	0.0114	4.0000	0.7071	0	.	0	.
11	12	0	0	2.0	0	0	0.0174	0.9826	0.0114	3.0000	0.7071	0	.	0	.
12	13	0	0	2.0	0	0	0.0174	0.9826	0.0114	2.0000	0.7071	0	.	0	.
13	14	1	0	2.0	0.5000	0.3536	0.0174	0.9826	0.0114	1.0000	0.7071	0.00870	0.00840	0.666667	0.628539
14	15	0	0	1.0	0	0	0.00870	0.9913	0.00840	.	.	0	.	0	.
15	16	0	0	1.0	0	0	0.00870	0.9913	0.00840	.	.	0	.	0	.
16	17	0	0	1.0	0	0	0.00870	0.9913	0.00840	.	.	0	.	0	.
17	.	1	0	1.0	1.0000	0	0.00870	0.9913	0.00840	.	.	.	.	.	.

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
548	321	227	41.42

Figure 8: The Kaplan Meier estimates for survival analysis.

The Kaplan Meier estimate gives the survival rate and the failure rate for each interval generated. In the above result, for the interval 0-1 though the number of failed cases is 162. The survival probability still shows 1.00, this denotes that the censored observations (expired patients) doesn't change the survival probability.

Model Validation Statistics

Data Role	Beneft	Average Hazard Ratio	Depth	Lift	Kolmogorov-Smirnov Statistic	Gini Concentration Ratio
Train	0.4	0.75704	0.6	1.666667	0.5	-0.2

Figure 9: Model statistics for the survival model.

The benefit value of 0.4(Figure 9) suggests good model performance.

STEPWISE REGRESSION FOR SURVIVAL MODEL

To understand the effect of influencing variables over the survival time, stepwise regression is chosen in building the survival analysis model. The cubic splines are included in the regression model by enabling knot selection in the properties panel. The results of stepwise regression for survival analysis is as shown in Figure 10. The important variables excluding the cubic spline functions are gender and age.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	_csb1	1	2	35.6692		<.0001
2	_csb5	1	3	21.0629		<.0001
3	gender	1	4	7.2403		0.0071
4	age_in_years	1	5	6.4210		0.0113

Figure 10: Summary statistics of stepwise selection.

PARAMETRIC SURVIVAL ANALYSIS

Survival Probabilities of different groups can be estimated and compared by crossing the target variable with the variable of interest for comparison and modeling it through the survival node. Cubic spline functions are enabled in these cases.

EFFECT OF GENDER ON THE SURVIVAL RATE

From the research it is known that the survival time is greatly influenced by the gender of the patient. To understand and compare the survival rates and the hazard functions based on age, new target variable 'gender_level' is created with nominal interval measurement. The survived male patient is given the value 1, and the survived female patient is given the value 2 and the expired patient has a value of 0 (Figure 11). This new target variable is now modeled using the survival node with the time interval of week.

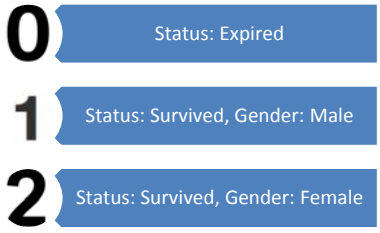


Figure 11: New Target variable

The sub-hazard functions related to gender and the results are shown in the Figure 12.

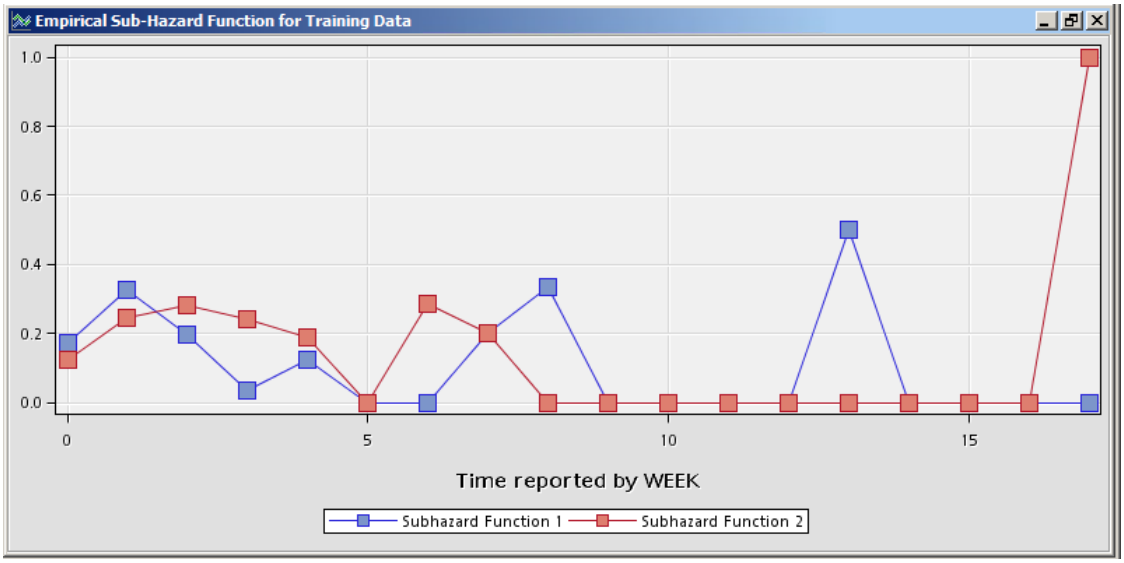


Figure 12: Sub-Hazard graph for males and females.

The peaks seen at week 6 and 7 highlighted in red above indicate that the risk of survival is high in week 6 and week 7 for female patients. The peaks seen at week 8 and week 13 highlighted in blue indicate the risk of survival is high in week 8 and week 13.

Model Validation Statistics

Data Role	Benefit	Average Hazard Ratio	Depth	Lift	Kolmogorov-Smirnov Statistic	Gini Concentration Ratio
Train	0.8	0.424821	0.2	5	1	0.6

Figure 13: Model statistics.

The increase of benefit value to 0.8 as compared to the basic model with a benefit value of 0.4 indicates model improvement. This suggests that gender plays a significant role in survival time and also cubic splines influence the model performance.

EFFECT OF AGE ON THE SURVIVAL RATE

Survival time in relation to the age of the patient has always been a curious point of research. To understand the difference in the survival times and the hazard function with respect to age, a new target variable by crossing the status of the patient with respect to age is created. The new target is 'age-level' is categorized as with patient falling between age 0-25 as 1, 26-50 as 2, 50-75 as 3, 75 to 100 as 4 (Figure 14). This new target variable is modeled using the survival node with time interval of week.

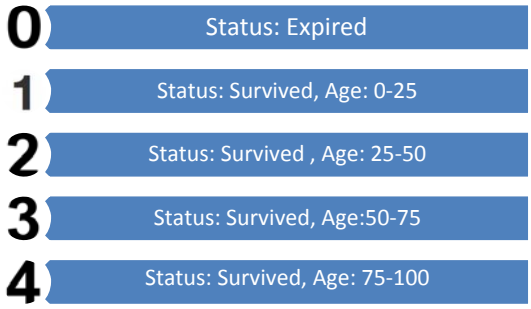


Figure 14: New Target Variable

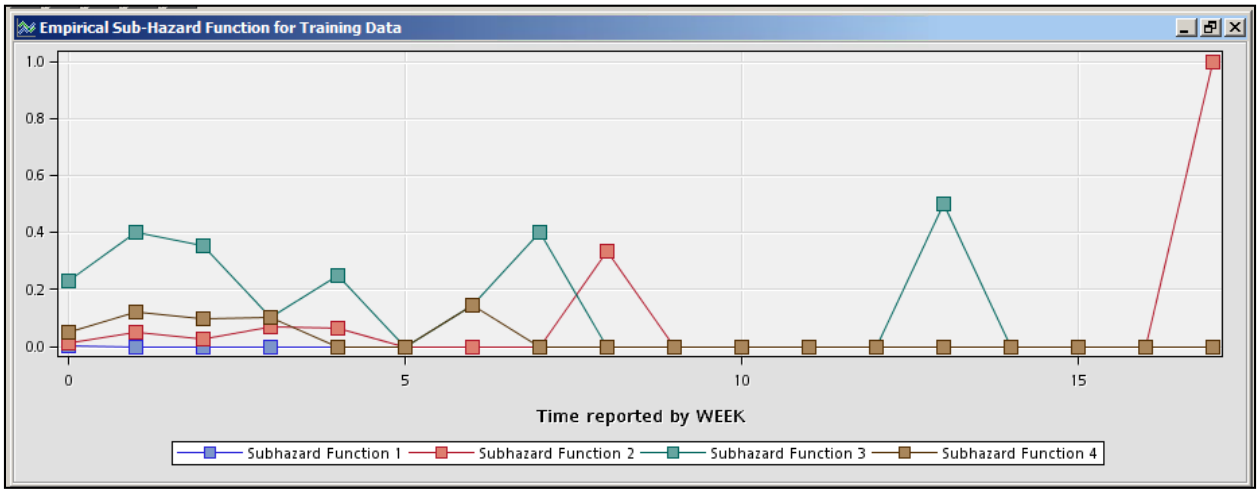
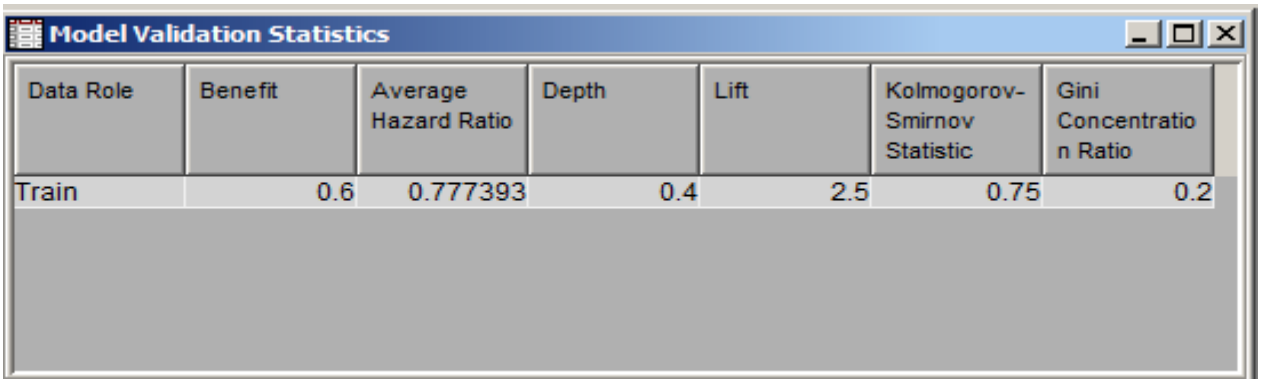


Figure 15: Sub-Hazard graph for different age groups.

The sub hazard plot (Figure 15) indicated above explains the risk of surviving in relation with age. The graph in blue belongs to the age group of 0-25 and there are no significant number of observations seen, hence the graph is flat. The probability of surviving for the patients within the age group 25-50 is low, shown by the red line. For this age group probability of survival is low in weeks 8 and 17. In addition, probability of survival is low for patients in the age group of 50-75, depicted by the green line, for the weeks 1, 2, 4, 6, and 13. Patients within the age group of 75-100 have low probability of survival for the first few weeks of admission in hospital till week 7. Flat graph for this group shows that if the patients are able to survive within 7 weeks, then their chances of survival are high after that.



Data Role	Benefit	Average Hazard Ratio	Depth	Lift	Kolmogorov-Smirnov Statistic	Gini Concentration Ratio
Train	0.6	0.777393	0.4	2.5	0.75	0.2

Figure 16: Model statistics.

The benefit value of 0.6 (Figure 16) and the average hazard ratio of 0.77 has improved compared to the basic model which indicates the significance of the age in survival time of the cancer patient and also cubic splines play a major role in influencing the model performance.

CONCLUSION

By performing the parametric and non-parametric survival analysis and analyzing the results, the following conclusions are drawn:

1. The survival rate is high in both males and females for the first two weeks after the treatment or admission to the hospital and probability of survival goes down by 60% between weeks 5 to 9 of admission.
2. Age and gender, are the factors that significantly influence the survival time and risk of surviving.
3. The probability of surviving for a male patient, given the condition that he survived till 7th week after the treatment, decreases by 70% and 50% in week 8 and 13 week respectively. The probability of surviving for a female patient, given the condition that she survives till week 5 after the treatment decreases by 70% and 80% in week 6 and 7 respectively. The doctors should give utmost care and should take necessary precautions around this time for efficient treatment.
4. The probability of survival for the patients in all of the age groups increases after week 3. The probability of survival is low after week 5, for the patients above age 25. The probability for survival for patients in age group 25-50 is low in week 8. For patients in age group 50-75 the probability is low for week 7 and week 13. Survival rate for old patients above the age of 75 is very low after week 6.
5. Cubic Splines help in improving the model performance significantly. There might be scenarios of over fitting of the model. This claim requires further exploration.

REFERENCES

Paper 3501 – 2015 Balamurugan Mohan and Dr. Goutam Chakraborty, Oklahoma State University. 2015. “Predicting transformer lifetime using survival analysis and modeling risk associated with overloaded transformers Using SAS® Enterprise MinerTM 12.1.”

Sascha Schubert, Susan Haller and Taiyeong Lee. 2012. “Paper 132-2012 It’s About Time: Discrete Time Survival Analysis Using SAS® Enterprise Miner™. “

Niosha Gunasekara. 2014. “Survival Analysis- An Applied Introduction.” Available at http://www.ats.ucla.edu/stat/sas/seminars/sas_survival/

Simona Despa.” What is Survival Analysis?”. Cornell University. Available at https://www.cscu.cornell.edu/news/statnews/stnews78.pdf

American Cancer Society. “Lung Cancer (Non-Small Cell).” 8/15/2014. Available at http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf

Centers for Disease Control and Prevention. “Leading Causes of Death.” 9/30/2015. Available at http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Raja Rajeswari Veggalam

Graduate Research Assistant- Management Information Systems,

Spears School of Business,

Oklahoma State University, Stillwater, OK – 74075.

E-mail: rajeswari.veggalam@okstate.edu

Work Phone: 216-650-3795

Raja Rajeswari Veggalam is a graduate student in Management Information Systems with SAS® and OSU Data Mining Certificate at Spears School of Business, Oklahoma State University. She has more than 2 years of experience working with IBM. She is passionate about data and her interests include predictive modelling, survival analysis, health analytics, text mining. She holds following credentials: SAS® Statistical Business Analyst, SAS® Base programmer and SAS® Certified Predictive Modeler.

Akansha Gupta

Graduate Student- Electrical and Computer Engineering with specialization in Data Analytics.

College of Engineering Architecture and Technology,

Oklahoma State University, Stillwater, OK – 74075.

E-mail: akansha.gupta@okstate.edu

Work Phone: 405-534-1834

Akansha Gupta is a graduate student in with Electrical and Computer Engineering with SAS® and OSU Data Mining Certificate at Spears School of Business, Oklahoma State University. She has more than 3 years of experience working with Accenture. She is a data enthusiast and her interests include data analysis and predictive modelling, data management and survival analysis. She holds following credentials: SAS® Base programmer and SAS® Certified Predictive Modeler.

Dr. Goutam Chakraborty

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU marketing analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

Oklahoma State University

Stillwater, OK, 74078

E-mail: goutam.chakraborty@okstate.edu