Survival Analysis of Hepatocellular Carcinoma

Atmaja Raman, Harsh Varddhan Singh, Abhishek Jaiswal, Rajyashree

Abstract: ***This paper performs the survival analysis for Hepatocellular carcinoma using two different algorithms-Random Forest model and Extreme Gradient Boosting (XGBoost) model. The models were used to perform binary classification. The patients were classified into two classes based on survival time > 10 months and <= 10 months. Results showed that the classification accuracy and misclassification rate of the random forest model was 0.66 and 0.34 respectively. The classification accuracy and misclassification rate of Extreme gradient boost model 0.61 and 0.39 respectively. The Random forest model performed better during testing.***

Index Terms: Survival Analysis, Hepatocellular carcinoma, Random forest, Extreme Gradient Boosting.

I. INTRODUCTION

Hepatocellular carcinoma is an customarily prevalent form of cancer and leads to cancer mortality.[1] Asian countries hold nearly 78% of hepatocellular carcinoma (HCC) cases which are reported every year globally. In Asia, exposure to aflatoxin, contaminated water with algal hepatotoxins, chewing betel nut and liver cirrhosis due to alcohol abuse are risk factors.[2] HCC incidence rates are different among different populations occupying a region. Males have higher rates of cancer as compared to females, with the ratios around 2:1 and 4:1(male: female). HCC's age distribution globally varies by sex, region, incidence rate etc. Female rates top in age range 5 years more than the maximum age group for males.[3] Hepatocellular carcinoma causes death in cirrhosis patients. Each patient may not only have a tumour but each tumour point may be genetically unique. Based on the liver disease and patients' history the genetics may differ. [4] Therapies for HCC are classified into four classes: surgical intervention, percutaneous intervention, trans arterial intervention, and drugs including gene and immunotherapy. Tumour removal, liver transplantation, and percutaneous intervention can cure and can improve survival in huge number of patients. In a few cases, trans arterial interventions produce good response rates and better survival.[5] HCC has bad prognosis as it is diagnosed at later stages. It is immune to radiotherapy and unresponsive to chemotherapy. In many cases, surgery along with liver transplant is the at best treatment option available. Hence, developing effective therapies is of utmost importance. [6] .Screening the blood for hepatitis C and B is a primary step. Vaccination for hepatitis B should be provided in enzootic countries. Liver ultrasound and AFP concentration levels have to be taken and checked frequently [2] In this paper, survival modelling is performed on data collected from patients affected by hepatocellular carcinoma. The patients are classified into two classes of survival time >10 months and <= 10 months. Random forest and Extreme Gradient boosting models are used for binary classification and a comparative study is performed between the models.

II.RELATED WORKS

In this section, a brief overview is given on previous studies on cancer and the methods used to predict survival time. Analysis of hepatocellular carcinoma using Bayesian network has taken the influence of various factors in a combined manner. Bayesian network is and importance measures are combined to highlight the main factors that have notable result on survival time.124 patients of more than 10 months and 77 patients of equal to or lesser than 10 months were classified perfectly. The model's accuracy was 67.2%. For those with survival more than ten months, true positive and false-positive rate was 83.22% and 48.67% respectively. PVTT was found to be the most noteworthy predictor of survival time for people who underwent hepatectomy using importance measure and BN.[7] For hepatocellular carcinoma, the 5-year mortality after surgery was predicted with an artificial neural network. The artificial neural network was implemented for predicting HCC's 5-year mortality and for comparing the prediction results with that of a logistic regression model. The mortality rate showed a notable correlation with gender, age, hospital volume, surgeon volume, CCI and LOS .It was seen that the ANN model surpassed LR model.[8] Patients who undergo liver transplants within the Milan criteria have excellent outcomes. Survival of patients once they cross this criteria remains uncertain. The aim was to observe surviving time of people with tumours that overshoot the Milan criteria using exploratory data analysis. Based on incremental values of size and number, the hazard ratios are 1·34 and 1·51 respectively. The outcome was linear with respect to size, where as it tended to plateau when number of tumours were over three. [9] Classification and regression tree (CART) model and an artificial neural network (ANN) were used to predict liver cancer patient survival. The results showcased the superiority ANN model. The AUC or area under receiver operating characteristic curve was found to be 0.915 for accuracy, 0.88 for sensitivity and 0.87 for specificity.

Revised Manuscript Received on May 07, 2019.

Atmaja Raman, (Student Btech) Computer Science Engineering, SRM IST, Ramapuram Campus, Chennai

Harsh Varddhan Singh, (Student Btech) Computer Science Engineering, SRM IST, Ramapuram Campus, Chennai

Abhishek Jaiswal, (Student Btech) Computer Science Engineering, SRM IST, Ramapuram Campus, Chennai

Rajyashree, (Faculty) Computer Science Engineering, SRM IST, Ramapuram Campus, Chennai

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

CART model was less accurate than ANN for predicting liver cancer survival.[10] A comparative study of support vector machine model and the random forest is done for cancer classification based microarray. Gene expression microarrays are crucial for making decisions clinically which leads to right prognosis and diagnosis in development of cancer. It is essential that the highly precise classification algorithms are applied for gene expression profile creation. The experiments indicate that support vector machines outshines random forests.[11] Survival rates of cancer patients are used to divide them into groups for treatment. A prognostic system is developed based on clustering approach. Groups of combinations are created using factor levels recorded in the data. Data partition sequences are generated by multiple clustering and the discrepancy measure between combinations is obtained from the data partition. This dissimilarity measure is used to obtain clusters of combinations with a hierarchical clustering.[12] Biological activity prediction based on the quantitative interpretation of molecular structure of compound was performed using Extreme gradient boosting model , an entity of Classification and Regression Tree .Xgboost outshines multiple machine learning algorithms in machine learning algorithms in performance .It performs notably on both low and high diversity datasets .[13]

III. SOFTWARE TOOLS USED

Exploratory.io is a tool that can operate on different data types and can discover patterns buried in the data using statistical methods and machine learning. For the purpose of coding Python version 3 along with packages such as seaborn, pandas and lifelines have been used.

IV. MODULE DESCRIPTION

Two models have been implemented for performing binary classification

A. Random forest model

B. Extreme Gradient boosting model

Random Forest: Random Forest is a versatile and straight forward machine learning algorithm, that mostly produces great results .It is also a frequently used algorithm because it is multipurpose and handles both the tasks which is classification and regression.

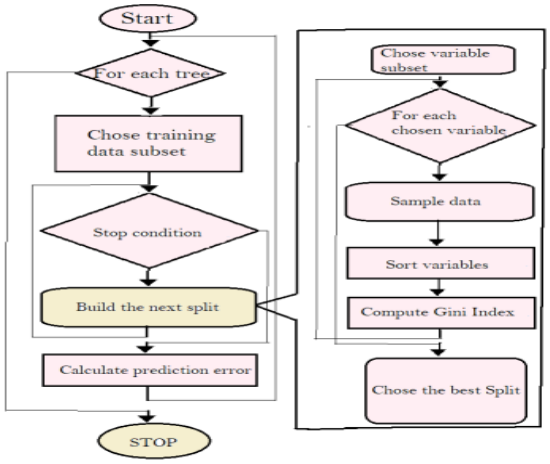

Fig 4.1.Random forest flowchart

The random forest consists of multiple decision trees which get merged together to get a more accurate prediction with more stability. In this algorithm, every node is divided using the optimum amidst a subgroup of predictors that are randomly selected at the node. RF model is like a Decision tree model with bootstrapping algorithm. For instance, there are a thousand observations in total with ten variables. Random forest constructs a multiple CART model with different primary variables and a distinct sample. Say, it will procure a part of hundred observations along with 10 randomly selected primary variables to construct a CART model. The process is done ten times repetitively and then an end prediction will be drawn on every observation. The end prediction will be relating to every other prediction. It is obtained by finding the mean of every prediction. Based on power of the each tree in the forest and the association among them a generalised error of a forest is obtained .[14]

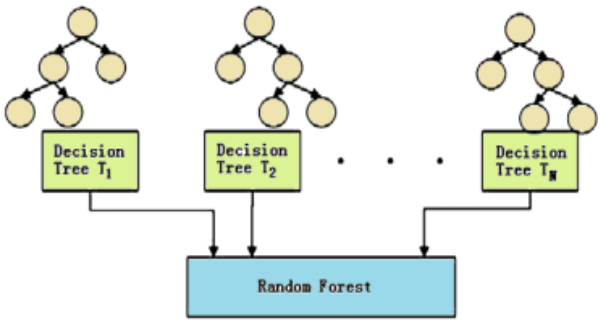

Fig 4.2.Random forest Decision tree

Extreme Gradient boosting: Extreme Gradient Boosting is based on preceding knowledge in gradient boosting. An additive mechanism is used for training purposes. If molecule I has a descriptor vector xi then to predict the result of tree model K additive functions are used. XGboost is an accurate and measurable application of extreme gradient boosting and it is shown to increase power of computation for boosted tree models .It was developed with the idea to enhance model performance and computational speed.

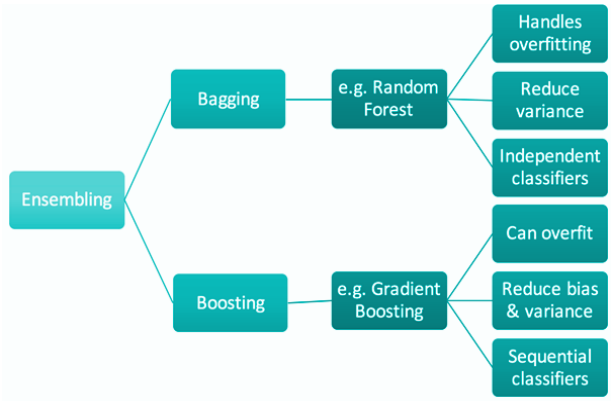

Fig 4.3. Extreme Gradient boosting

V. METHODOLOGY

This section describes the dataset used followed by plots and the implementation of the models.

A. DATASET DESCRIPTION Two hundred and ninety nine genuine medical records of patients were accumulated in the years 2006 to 2011 from Xi'an Jiaotong Health centre , Medical College at China. The dataset contained 16 columns, including gender, age, blood loss during operation, postoperative complication, portal vein tumour thrombosis (PVTT),trans catheter arterial chemoembolization (TACE), liver function, history of HBV, history of HCV, preoperative Alpha-fetoprotein, tumour size, clamping time of porta hepatis (TCPH), tumour number, methods of operations, growth or metastasis , and the survival time.

B. PLOTS The correlation map below shows the interdependence between variables . Every square in the map gives the connection between the variables. A correlation map is a great way to outline the data . White cells depict the high correlation between variables and the black cells depict low correlation.

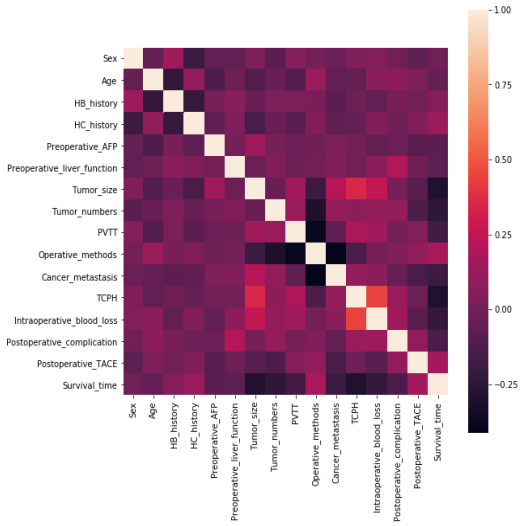

Fig 5.1.Correlation Map

The plot below depicts variable importance. Variable importance plot provides a list of the most significant variables in decreasing order. This is a difficult concept to describe as the variable's vitality is because of its interconnection with other variables. The random forest algorithm evaluates a variable's importance by considering the increase in error of prediction when the variable's data is permuted while others are left untouched. The top variables contribute more to the model than the bottom ones and also have high predictive power in classification.In the graph, Preoperative_AFP is the most significant variable.

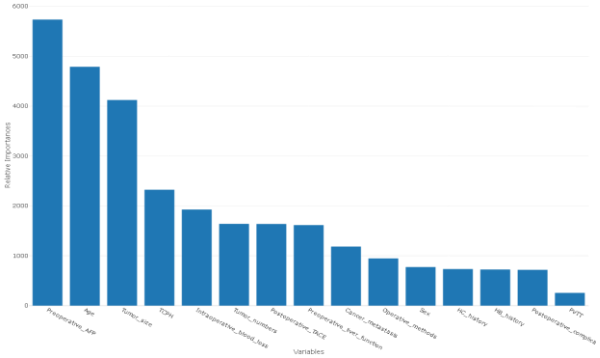

Fig 5.2.Variable Importance Plot

C. SURVIVAL ANALYSIS In cancer analysis, the time taken for episode of interest is given major importance. It is known as survival time and it can be referred to as the time taken from complete abeyance to reoccurrence or the time taken from prognosis until demise. The dataset contains the survival time of patient recorded after hepatectomy operation until the time of death. The task considered is to classify the patients based on survival time into two classes using Random forest model and extreme gradient boosting model and to compare the performance. Exploratory a data analytics tool is used to execute the models. The dataset is split in the ratio of 0.3:0.7 for testing and training respectively. There are 15 features that are considered in the model. Once trained the model is tested on the test dataset consisting of 90 records. This data is classified into two classes and the accuracy of classification is obtained. The predicted label and original class are displayed in the form of a pivot table. The model is evaluated further by finding confusion matrix values. From this precision, recall, AUC, etc are obtained.

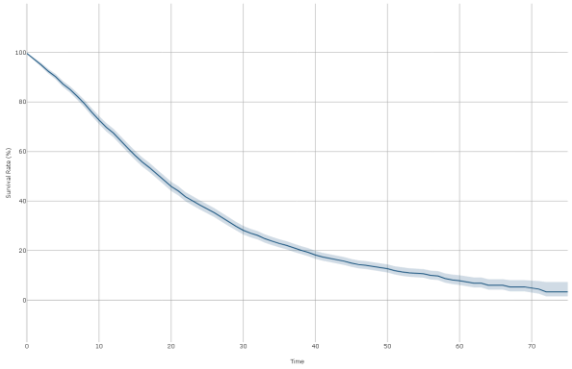

Fig 5.3.Survival Curve

A survival curve is a graph that displays the surviving percentage with time. The Y axis or the vertical line shows the proportion of people surviving. The X-axis gives the time duration once the observation starts. The graph is found to decline with time showing that the number of people surviving is decreasing as time increases. The optimal way to evaluate prognosis is to compare the ordeal of patients in an identical situation undergoing similar treatment and the same has been represented by a survival curve. The survival curve begins at time 0 with 100 % survival. After that it can reduce or remain at the same level but it can never increase.

VI. RESULTS

The results showed that the Random forest model has an accuracy of classification of 0.66 and rate of misclassification of 0.34.The Extreme gradient boosting model has an accuracy of classification of 0.61 and rate of misclassification of 0.39.Accuracy shows the comfort of the model in detecting the positive and negative class. It's computed as the addition of True Negative and the True Positive divided by total.

Accuracy = [TP + TN] / total eq (6.1)

Precision gives the probability of success for finding a correct positive class classification. It is computed as the True Positive number divided by the total positive count.

Precision = TP / [TP + FP] eq (6.2)

Recall shows the sensitivity of the model towards identifying the positive class. It is given by the True Positive value divided by addition of True Positives and False Negatives.

Recall = TP / [FN + TP] eq (6.3)

F-score also referred to as F-measure is the calculation of accuracy of a test. The F1 score is computed by finding the harmonic average of precision and recall. F1 score value is the best at 1 and worst at 0.

classes	predicted_label	(Number of Rows)
0	0	31
	1	14
1	0	16
	1	28
Total		89

Fig 6.4.Random Forest prediction

classes	predicted_label	(Number of Rows)
0	0	29
	1	16
1	0	19
	1	25
Total		89

Fig 6.5.XGBoost prediction

f_score # numeric	accuracy_rate # numeric	misclassification_rate # numeric
0.7891156462585	0.6555555555556	0.3444444444444

true_positive # integer	false_positive # integer	true_negative # integer	false_negative # integer
58	0	1	31

precision # numeric	recall # numeric
1	0.6516853932584

Fig 6.6.Random Forest Evaluation

f_score # numeric	accuracy_rate # numeric	misclassification_rate # numeric
0.7552447552448	0.6111111111111	0.3888888888889

true_positive # integer	false_positive # integer	true_negative # integer	false_negative # integer
54	0	1	35

precision # numeric	recall # numeric
1	0.6067415730337

Fig 6.7.XGBoost Evaluation

The study shows that the random forest model outperforms the extreme gradient boosting model. This is maybe because gradient boosting models are sensitive to overfitting.

VII. FUTURE SCOPE

Survival analysis is essential especially in clinical trials for diseases such as metastatic cancer. A subset of people going through some different treatment might hold an advantage as compared to other groups for survival. Such relationships should be studied to improve treatment options and drug discovery. The analysis can be improved by building better predicting models by using neural networks. Healthcare depends on vital decision tools in drawing inferences such as what type of drug medication is best suited for a disease cure or management and what is the best safety inclined procedure to perform.

In making conclusions like these, survival analysis plays a key role.

REFERENCES

1. Parkin DM. "Global cancer statistics in the year 2000" Lancet Oncology 2001;2:533–543.
2. Poon D, Anderson BO, Chen LT et al. Management of hepatocellular carcinoma in Asia: Consensus statement from the Asian Oncology Summit 2009. Lancet Oncol 2009;10:1111–1118.
3. El-Serag HB, Rudolph KL. "Hepatocellular carcinoma: epidemiology and molecular carcinogenesis". Gastroenterology. 2007 Jun;132(7):2557-76
4. Bruix J, Gores, G.J. & Mazzaferro, V. Hepatocellular carcinoma: clinical frontiers and perspectives. Gut 63, 844–855 (2014).
5. Blum H. Hepatocellular carcinoma: therapy and prevention. World J Gastroenterol 2005;11:7391-7400.
6. Aravalli RN, Steer CJ, Cressman EN (2008) Molecular mechanisms of
a. hepatocellular carcinoma. Hepatology 48:2047–2063. doi: 10.1002/hep.22580
7. Cai ZQ, Si SB, Chen C, Zhao Y, Ma YY, Wang L, Geng ZM. "Analysis of prognostic factors for survival after hepatectomy for hepatocellular carcinoma based on Bayesian network". PLoS ONE 10(3): e0120805, March 2015
8. Hung WT, Lee KT, Wang SC, Ho WH, Chang SC, Wang JJ, et al. (2012) Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma and performance comparison with logistic regression model: A nationwide Taiwan database study. In Proceedings of the Third International Conference on Innovations in Bio-Inspired Computing and Applications.IEEE: 241–245
9. Mazzaferro V, Llovet JM, Miceli R, et al. "Predicting survival after liver transplantation in patients with hepatocellular carcinoma beyond Milan criteria: a retrospective, exploratory analysis". Lancet Oncol 2009;10:35–43.
10. Chen CM, Hsu CY, Chiu H W, Rao HH. "Prediction of survival in patients with liver cancer using artificial neural networks and classification and regression trees".7th Natural Computation International Conference, Shanghai. Piscataway, IEEE: 811–815(2011).
11. A. Statnikov, L. Wang, C.F. Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, BMC Bioinformatics 9 (319) (2008) 1–10.
12. D. Chen, K. Xing, D. Henson, L. Sheng, A. Schwartz and X. Cheng, Developing prognostic systems of cancer patients by ensemble clustering, Journal of Biomedicine and Biotechnology 2009 (2009), 632786
13. Mustapha, I. B., and F. Saeed. 2016. "Bioactive Molecule Prediction Using Extreme Gradient Boosting." Molecules 21(8):1–12. doi:10.3390/molecules21080983.
14. L. Breiman, Random forests, Machine Learning 45(1) (2001), 5–32.

AUTHORS PROFILE

Author-1
Ph

Atmaja Raman(Student Btech)
Computer Science Engineering,
SRM IST, Ramapuram Campus, Chennai

Email- atmajaraman32@gmail.com

Author-2 Photo

Harsh Varddhan Singh (Student Btech)
Computer Science Engineering,
SRM IST, Ramapuram Campus, Chennai

Email- myoffice016@gmail.com

Author-3 Photo

Abhishek Jaiswal (Student Btech)
Computer Science Engineering,
SRM IST, Ramapuram Campus, Chennai

Email- abhishekjaiswal2103@gmail.com

Rajyashree (Faculty)
Computer Science Engineering,
SRM IST, Ramapuram Campus, Chennai
Email - rrajyashree123@gmail.com