Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms

B.Balakumar[1], P.Raviraj[2], K.Sivaranjani[3]

[1]*Assistant Professor, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India,*
balakumarmsu@gmail.com
[2]*Professor, Department of CSE, GSSS Institute of Engineering and Technology for Women,KRS Road,Metagalli,Mysore, Karnataka-570016,*
drpraviraj@gmail.com
[3]*PG Scholar, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India,*
ranjanishivak511@gmail.com

Abstract:

Titanic disaster occurred 100 years ago on April 15, 1912, killing about 1500 passengers and crew members. The fateful incident still compel the researchers and analysts to understand what can have led to the survival of some passengers and demise of the others. With the use of machine learning methods and a dataset consisting of 891 rows in the train set and
418 rows in the test set, the research attempts to determine the correlation between factors such as age, sex, passenger class, fare etc. to the chance of survival of the passengers. These factors may or may not have impacted the survival rates of the passengers. In this research paper, various machine learning algorithms namely Logistic Regression, Naive Bayes, Decision Tree, Random Forest have been implemented to predict the survival of passengers. In particular, this research work compares the algorithm on the basis of the percentage of accuracy on a test dataset.

Index Terms—Titanic; Prediction;
Python; Logistic Regression; Random Forest; Decision Tree;

I. INTRODUCTION:

The field of machine learning has allowed analysts to uncover insights from historical data and past events. Titanic disaster is one of the most famous shipwrecks in the world history. Titanic is a British cruise liner that sank in the North Atlantic Ocean few hours after colliding with an iceberg While there are facts available to support the cause of the shipwreck, there are various speculations regarding the survival rate of passengers in the Titanic disaster. Over the years, data of survived as well as deceased passengers has been collected. This dataset has been studied and analyzed using various machine learning algorithms like Random Forest. various languages and tools are used to implement these algorithms including Python .The approach of the research paper is centered on Python for executing algorithms-,Logistic Regression, Decision Tree, and Random Forest.

II. OBJECTIVE :

The dataset found on the Kaggle website has two perspectives website has two perspectives. One is a training data and the other is a testing data. The objective of the training data is to create a model which will help in predicting the outcomes of the test data. For the purpose of this research paper, the training data from the Kaggle website will be divided into two parts using three different ratios for training and creating the model and then predicting and the testing data from the Kaggle website will not be used. There will be application of different techniques for predicting whether a person survived the Titanic disaster or not. The data analysis will then be done and the prediction outcomes will be checked for accuracy. The accuracy will then be compared in order to suggest the

better performing algorithm with respect to used dataset.

DATASET:

The data consists of 500 rows in the train set which is a passenger sample with their associated labels. For each passenger, the name of the passenger, sex, age, his or her passenger class, number of siblings spouse on board, number of parents or children aboard, cabin, ticket number, fare of the ticket and embarkation were provided. The data is in the form of a CSV (Comma Separated Value) file. For the test data, the website provided a sample of 534 passengers in the same CSV format. The structure of the dataset with a sample row has been listed in Table. The attributes of the training set and their description have been mentioned in Table

Before building a model, data exploration is done to determine what all factors or attributes can prove beneficial while creating the classifier for prediction. To start the exploration, few X-Y generic plots are made to get an overall idea for each attribute. Some of the generic plots have been shown below. The age plot in suggested that maximum or majority of the passengers belonged to the age group of 20-40. Similarly, a graph Fig 2 is plotted and some calculations are performed for the sex attribute and the results suggested that the survival rate of the female is 25.67% higher to that of the male. Similarly, each of the attribute are explored to extract those attributes or features which would be used later for prediction. A survival histogram is generated to determine the number of people survived vs. number of people who can not survive. From the histogram it is clear that the number of people who survived is less than the number of people who could not survive. The survival histogram is shown in Fig 3. In order to deal with the missing values data cleaning is done. While observation it has been found that the dataset is not complete. There are various rows for which one or more fields are marked empty (especially age and cabin). But the age is an important attribute to predict the survival of passengers. Hence a technique to replace the NAs in the age column has been used. The gender column has been changed to 0 and 1

Attributes in the Training Data Set:

Passenger ID	Identification no. of the passengers.
Pclass	Passenger class (1, 2 or 3)
Name	Name of the passengers
Sex	Gender of the passengers (male or female)
Age	Age of the passenger
SibSp	Number of siblings or spouse on the ship
Parch	Number of parents or children on the ship
Ticket	Ticket number
Fare	Price of the ticket
Cabin	Cabin number of the passenger
Embarked	Port of embarkation (Cherbourg, Queenstown or Southampton
Survived	Target variable (values 0 for perished and 1 for survived)

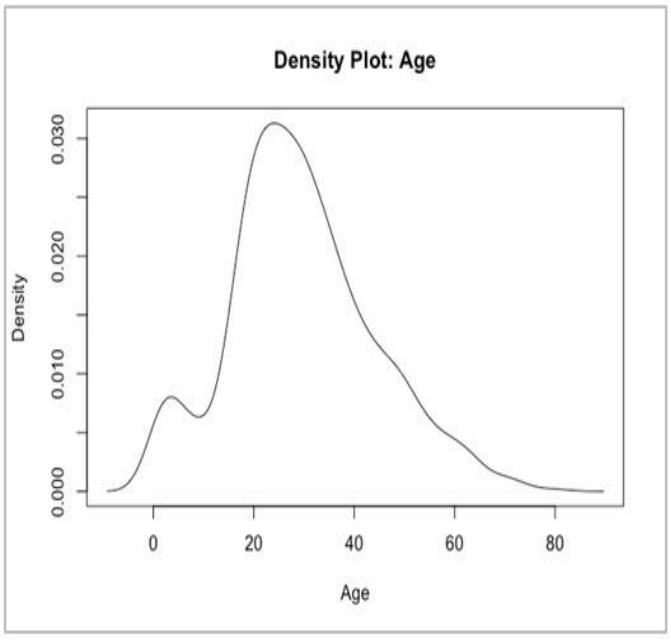

Fig: 1 Age plot

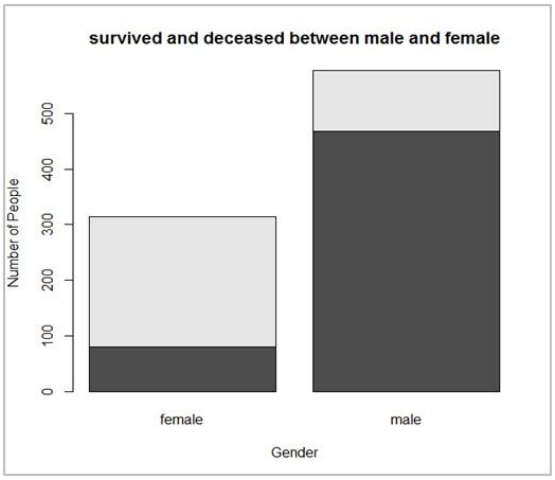

Fig. 2. Sex bar plot

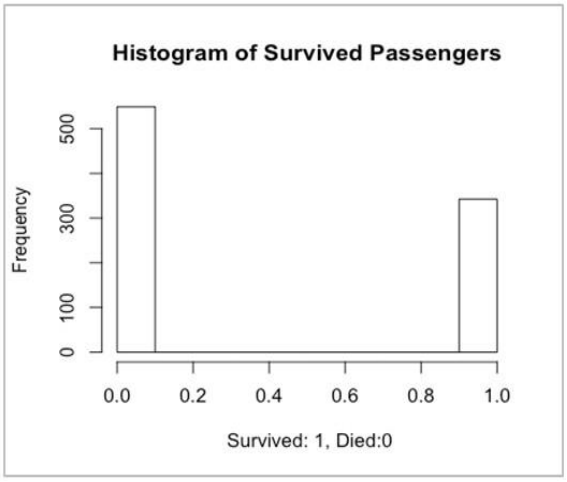

Fig. 3. Survival Histogram

III. METHODOLOGY:

The training data from the Kaggle website contains 892 rows and 12 columns. The first row of the dataset describes the different parameters for a passenger. The first column in the dataset gives the Passenger Id of a passenger and the second column of the dataset gives whether the person survived or not. The PClass attribute defines the class in which the passenger was travelling in the ill-fated ship. The following will be the roadmap for the research work: The 891 data will be divided into different ratios for training and then testing. The different algorithms which will be applied for the purpose of the research are Multiple Linear Regression and Logistic Regression. Apart from applying the above methods, there will be a trivial base case method which assumes all the women and children (whose age is below 18) survived the horrific incident and all the men died. This assumption is based on the records that initially women and children were allowed to come out of the ship.

Different segregation of the data for applying the above three methods are listed as follows:

1. Training Data- Passenger Id values from 1-446 Testing Data Passenger Id values from 447-891
2. Training Data- Passenger Id values from 447-891 Testing Data Passenger Id values from 1-446
3. Training Data- Passenger Id values from 1-600 Testing Data Passenger Id values from 601-891
4. Training Data- Passenger Id values from 601-891 Testing Data Passenger Id values from 1-600

ALGORITHM USED:

Prediction models are generated using four machine learning algorithms namely, Logistic Regression, Decision tree and Random forest. Each of these algorithms are compared to one another on the basis of the accuracy percentage. The attributes used in the test and train dataset for implementing these algorithms are- Pclass, Sex, Age, SibSp, Parch, Mother, Children, Family and Respectable.

Logistic Regression:

Logistic Regression is a type of classification algorithm in which the target variable is categorical and binary. In the dataset survived column is the dependent variable which is both binary and categorical (1 for survival and 0 for demise) The prediction model is built by including the features i.e.

Pclass, Sex, Age, SibSp, Parch, Mother, Children, Family and Respectable. After

running the model, it is observed that family is the least significant variable.

We have two possible outcomes 1 or 0 for N observations indicating a success or failure of an event. For N observations, a study of failures in a class is conducted for a particular subject. Following are the **outputs**: { 0, 0,0,1,1,1,1,1,1,1,1,0,0,0,0,0,……..} In which 1 indicates the pass in subject and 0 indicates the failure in subject. The likelihood principle says that all inference about a parameter should utilize observed data only through how it affects the likelihood function, the probability of observing the observed data given *p*. the likelihood is

$F(Y) = P(Y = (1, 1, 1, 0, 0, 1…1, 1, 0) |p) = p * p * p * (1 - p) * (1 - p) * p *…….p * p * (1-p)$

$F(Y) = \Sigma P\ y * (1 – P)\ 1- y$ ….

Eq 1.1 In our case P takes up the logistic distribution function

$K(x) = ez / (1+ez)$

$\ln F(Y) = \Sigma y \ln P + 1-y \ln (1 – P)$

$\ln F(Y) = \Sigma y \ln P - y \ln (1 – P) + \ln (1 – P)$

$\ln F(Y) = \Sigma y \ln (P / (1 – P)) + \Sigma \ln (1 – P)$

$\Sigma(P / (1 – P)) = ez / (1+ez) / 1/(1+ ez)$

$\Sigma(P / (1 – P)) = ez$ where $z = \beta 1 + \beta 2 * x$ $\Sigma (1 – P) = 1/(1+ ez)$ $\ln F(y) = \Sigma y (\beta 1 + \beta 2 * x) - \Sigma \ln(1 + e\ \beta 1 + \beta 2 * x)$ …. Eq 1.2

Random Forest:

Random forest algorithm is implemented for improving the accuracy of the classification model even further and determining the most significant features for survival. Random forest algorithm is a classification algorithm that constructs a multitude of decision trees at the time of training and it outputs the class which is the mode of the individual trees .The model has been built with all the variables of the cleaned train dataset, that are Pclass, sex, Age, Family, Children, SibSp, Mother, Parch and Respectable. In order to understand the significance of all these different variables in the classification process, an argument importance while building our model is used. From it is clear that Sex and Pclass play the most significant role in the classification model while Mother, Parch and Respectable are the least significant variables. This is in alignment with our analysis using logistic regression algorithm. The accuracy of random forest algorithm has been checked on the test data. After executing the Random forest analysis on test cases the model generated a confusion matrix as shown in Table VI. Using the confusion matrix we determined

that out of total 418 predictions, the model made 384 correct predictions, giving an accuracy of 91.8%.

RESULTS:

For comparing the four techniques used in this research work two metrics are used. First metric is accuracy and the second metric is false discovery rate. Both these metrics are computed using the confusion matrix. The structure of the confusion matrix is shown in table XI. Accuracy is the measure of how well a model predicts. Higher the accuracy the better. Accuracy is calculated using the formula TN+TP/ Total number of test set rows*100. False discovery rate is a method oconceptualizing the rate of type I (false positive) errors in null hypothesis testing when conducting multiple comparisons. For the problem used in the research paper, false discovery rate is an important metric as it would be dangerous if the system predicts a passenger would survive but in reality he does not survive. False discovery rate is calculated using the formula FP/FP+TP *100. Hence lower the false discovery rate the better. The accuracy and false discovery rate for each of the algorithm.

Comparison Of Algorithms:

Algorithm	Accuracy Score	False discovery Rate

Logistic Regression	94.26%	7.90%
Decision Tree	93.06%	9.26%
Random Forest	91.86%	10.66%

CONCLUSION AND FUTURE WORK:

Logistic Regression proved to be the best algorithm for the Titanic classification problem since the accuracy of Logistic Regression is the highest and the false discovery rate is the lowest as compared to all other implemented algorithms.The research also determined the features that are the most significant for the prediction. Logistic regression as well as Random forest suggested that Pclass, sex, age, children and SibSp are the features that are correlated to the survival of the passengers. Future work might include potentially validating more using pruning techniques that is to see if a shallower tree with same or improved accuracy can be achieved. Cross validation could also be used that is calculating accuracy based on different combinations of training and test data. It would be interesting to play more with dataset and introducing more attributes which might lead to good results. Various other machine learning techniques like SVM, K-NN classification can be used to solve the problem

REFERENCES

[1] Kaggle.com, 'Titanic:Machine Learning form Disaster',[Online]. Available: http://www.kaggle.com/. [Accessed: 10- Feb- 2017].

[2] Eric Lam, Chongxuan Tang, "Titanic Machine Learning From Disaster", LamTang-TitanicMachineLearningFromDisaster, 2012.

[3] Cicoria, S., Sherlock, J., Muniswamaiah, M. and Clarke, L, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster", Proceedings of Student-Faculty Research Day, CSIS, pp. 1-6, May 2014.

[4] Vyas, Kunal, Zeshi Zheng, and Lin Li, "Titanic-Machine Learning From Disaster", Machine Learning Final Project, UMass Lowell, pp. 1-7, 2015.

[5] Mikhael Elinder.(2012). 'Gender, social norms, and survival in maritime disasters', [Online]. Available: http://www.pnas.org/content/109/33/13220.full. [Accessed: 8- March - 2017].

[6] Frey, B. S., Savage, D. A., and Torgler, B, "Behavior under extreme conditions: The Titanic disaster", The Journal of Economic Perspectives, 25(1), pp. 209-221, 2011.

[7] Trevor Stephens. (2014), 'Titanic: Getting Started With R - Part 3: Decision Trees', [Online]. Available: http://trevorstephens.com/kaggletitanic-tutorial/r-part-3-decision-trees/. [Accessed: 11- March- 2017].

[8] Trevor Stephens. (2014). 'Titanic: Getting Started With R - Part 3: Decision Trees', [Online]. Available: http://trevorstephens.com/kaggletitanic-tutorial/r-part-3-decision-trees/. [Accessed: 8- March - 2017].

[9] Rex Morgan. (2016). Titanic [Online]. Available:http://www.because.uk.com/wp-content/uploads/Because-2016-03w.pdf. [Accessed: 9- March - 2017].

[10] Jason Brownlee. (2014). How to implement Nave Bayes in Python from scratch [Online]. Available: http://machinelearningmastery.com/naivebayes-