International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 1

Analyzing Titanic Disaster using Machine Learning Algorithms

Dr.Prabha Shreeraj Nair
Dean Research, Tulsiramji Gayakwade Patil College of Engineering and Technology, Nagpur

ABSTRACT

Titanic disaster occurred 100 years ago on April 15, 1912, killing about 1500 passengers and crew members. The fateful incidents still compel the researchers and analysts to understand what could 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, we attempt 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, we use various machine learning algorithms namely Logistic Regression, Naïve Bayes, Decision Tree, Random Forest to predict the survival of passengers. In particular, we attempt to compare these algorithms.

Keywords: *titanic; prediction; classification; data mining; Rz; python; logistic regression; random forest; decision tree; naïve bayes*

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 was a British cruise liner that sank in the North Atlantic Ocean a 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. The dataset is publically available on a website called Kaggle.com [1]. This dataset has been studied and analyzed using various machine learning algorithms like Random Forest, SVM etc. Various languages and tools are used to implement these algorithms including Weka, Python, R, Java etc. Our approach is centered on R and Python for executing algorithms- Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest. The prime objective of the research is to analyze Titanic disaster to determine a correlation between the survival of passengers and characteristics of the passengers using various machine learning algorithms.

In particular, we will compare the algorithms on the basis of the percentage of accuracy on a test dataset.

II. DATASET

The dataset we use for our paper was provided by the Kaggle website. The data consists of 891 rows in the train set which is a passenger sample with their associated labels [1]. For each passenger, we were also provided with the name of the passenger, sex, age, his or her passenger class, number of siblings or spouse on board, number of parents or children aboard, cabin, ticket number, fare of the ticket and embarkation. The data is in the form of a CSV (Comma Separated Value) file. For the test data, we were given a sample of 418 passengers in the same CSV format. The structure of the dataset with a sample row has been listed in the three tables below:

Table I: Kaggle Dataset

Passenger ID	Survived	P class	Name
1	0	3	Braund, Mr.
2	1	1	Cunnings, Mrs.

TABLE II: KAGGLE DATASET (CONTD.)

Sex	Age	Sib Sp	Parch	Ticket	Parc
male	22	1	0	A/521171	
female	38	1	0	PC17599	

TABLE II: KAGGLE DATASET (CONTD.)

Fare	Cabin	Embarked
7.25		S
71.2833	C85	C

TABLE IV: ATTRIBUTES IN TRAINING DATASET

Attributes	Description
PassengerID	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)

Before building a model, we explored the dataset to determine what all factors or attributes can prove beneficial while creating the classifier for prediction. We started with few X-Y generic plots to get an overall idea for each attribute. Few generic plots have been shown below. From the age plot in Figure 1 we came to a conclusion that maximum or majority of the passengers belonged to the age group of 20-40.

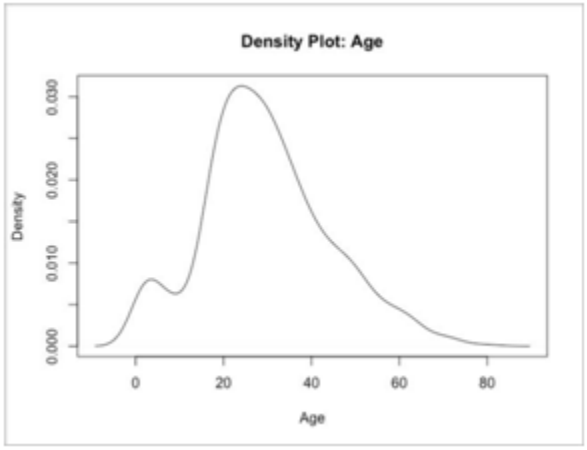

Fig. 1 Age plot

Similarly, we plotted a graph (Figure 2) and performed some calculations for the sex attribute and found out that the survival rate of the female is 25.67% higher to that of the male. Similarly, we explored each of the attribute to extract those attributes or features which we will use later for prediction. We also explored the dataset to determine the number of people survived vs. number of people who could 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 Figure 3.

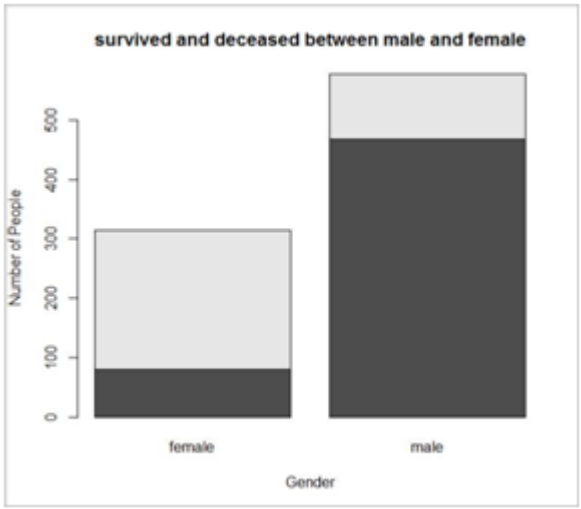

Fig. 2 Sex bar plot

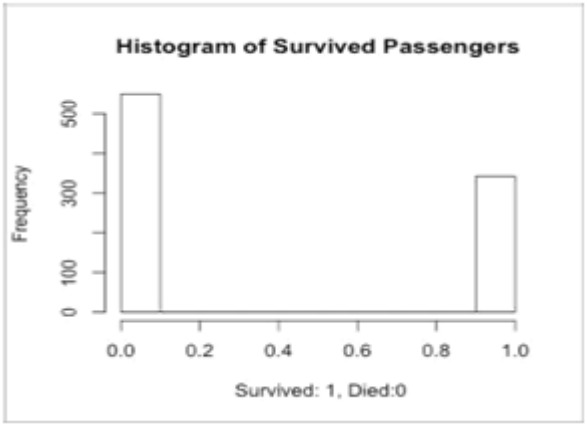

Fig. 3 Survival Histogram

We performed some data cleaning in order to deal with the missing values. We saw that the dataset is not complete. There are various rows for which one or more fields are marked empty (especially age and cabin). We think that age could an important attribute to predict the survival of passengers. So we used a technique to replace the NAs in the age column. The gender column has been changed to 0 and 1 (0 for male and 1 for female) to fit the prediction model in a better manner. We also introduced some new variables into the dataset to predict the survival more closely.

III. RELATED WORK

Many researchers have worked on the Titanic problem in order to compare various different machine learning techniques in terms of the efficiency of the algorithm to predict the survival of the passengers. Studies have tried to trade-off between different features of the available dataset to provide the best prediction results. Lam and Tang et al. used the Titanic problem to compare and contrast between three algorithms- Naïve Bayes, Decision tree analysis and SVM. They concluded that sex was the most dominant feature in accurately predicting the survival. They also suggested that choosing important features for obtaining better results is important. There were no significant differences in accuracy between the three methods they used [2]. Shawn Cicoria and John Sherlock et al. performed Decision tree classification and Cluster analysis to suggest that sex is the most important feature as compared to other features in determining the likelihood of the survival of passengers [3]. Kunal Vyas and Lin et al. suggested that dimensionality reduction and playing more with the dataset could improve the accuracy of the algorithms. The most important conclusion provided by them was that more features utilized in the models do not necessarily make results better [4]. Although many researchers have worked hard to determine the actual cause of the survival of some passengers and demise of others, we attempt to get better results and accuracy by utilizing various different combination of features and different machine learning methods.

IV. METHODOLOGY

The first step in our methodology is to clean the training dataset. We start with exploring different attributes for NA values. We see that age column has 177 rows with NA values and cabin 687 rows with NA values. As most of the data in the cabin column is missing, we decided to drop this column from our analysis. We assume that age is a very important attribute. Hence, we decided to keep the age column for the analysis. We attempt to establish a relationship between the title of the passengers and their age. We believe that Ms. A is younger than Mrs. A and we also assume that the people having same titles are closer in age. Titles of the passengers have been extracted form the name of the passengers and we have replaced the name column with the extracted titles. The missing entries has been replaced by the average age of the particular title-group i.e. if there is a missing age value for a woman with title Mrs. then the missing value gets replaced with the average age of all the women with title Mrs.(shown in Figure 4).

Survived	Pclass	Name	Sex	Age	SibSp	Parch
0	3	Mr	0	22.00	1	0
1	1	Mrs	1	38.00	1	0
1	3	Miss	1	26.00	0	0
1	1	Mrs	1	35.00	1	0
0	3	Mr	0	35.00	0	0
0	3	Mr	0	32.37	0	0
0	1	Mr	0	54.00	0	0
0	3	Master	0	2.00	3	1
1	3	Mrs	1	27.00	0	2
1	2	Mrs	1	14.00	1	0

Fig. 4 Average age

In the past marine disasters the policy of Women Children First (WCF) has been used by the crew members giving women and children survival advantage over men [5]. Based on this social norm we decided to introduce some new attributes to strengthen our dataset and improve our analysis. A research study suggested [4] that titles such as ‘Dr’, ‘Col’, etc. from the Name column can be an important part of the analysis since it shows that these people are influential and respectable people. These attributes are listed in the table below:

TABLE V: NEW ATTRIBUTES AND THEIR DESCRIPTIONS

New Attributes	Description
Mother	Column value is 1 only if the title is Mrs. and value of parch is greater than 0. Otherwise, 2 is assigned.
Children	Column value is 1 only if age is less than or equal to 14. Otherwise, 2 is assigned.
Family	Z= X+Y+1 where X is the value of SibSp and Y is the value of Parch.
Respectable	Column value is 1 is the title is Dr, Rev, Capt., Col., Don. Or Lady. Otherwise 2 is assigned.

TABLE VI:
DATASET WITH NEW ATTRIBUTES

Survived	Pclass	Name	Mother
0	3	Mr	2
1	1	Mrs	2

TABLE VII:
DATASET WITH NEW ATTRIBUTES(CONTD.)

Age	SibSp	Parch	Sex
22.00	1	0	Male
38.00	1	0	Female

TABLE VIII:
DATASET WITH NEW ATTRIBUTES(CONTD.)

Children	Family	Respectable
2	**2**	**2**
2	**2**	**2**

Dataset after the addition of these attributes have been shown in the above tables (Table VI, VII, and VIII). We have dropped Name, Ticket, Cabin and Embarked out of our analysis as we believe these variables are not relevant to our analysis. Similarly, due to large variation in values, we also removed the fare attribute.

A. NAÏVE BAYES

Naïve Bayes is a classification algorithm that applies Bayes theorem to build a prediction model. There are different types of Naïve Bayes theorem. In this analysis, we used Gaussian Naïve Bayes algorithm. The algorithm starts by creating summaries of the train data which involves calculating mean and the standard deviation for each attribute, by class value. The prepared summaries are used for prediction. Prediction involved calculating the probability for a particular data sample to belong to each of the two class. The class with the largest probability value was considered as the predicted class. Classification accuracy was obtained by comparing the predictions to the class values of the test data [6]. We got a classification accuracy of 91.38755%.

Fig. 6 Prediction using Naïve Bayes

B. LOGISTIC REGRESSION

After Naïve Bayes classification we implemented Logistic Regression. Logistic Regression is a type of classification algorithm in which the target variable is categorical and binary [7]. In our dataset survived column is the dependent variable which is both binary and categorical (1 for survival and 0 for demise). We start with building a model including all the features i.e. Pclass, Sex, Age, SibSp, Parch, Mother, Children, Family and Respectable. After running the model, we observed that family is the least significant variable. Hence, we dropped family from our dataset and built the model again (Figure 7).

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8957  -0.5628  -0.4132   0.5842   2.5797

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.797105   2.191381   4.014 5.96e-05 ***
Pclass      -1.182083   0.123229  -9.593  < 2e-16 ***
Sex          2.746818   0.210892  13.025  < 2e-16 ***
Age         -0.028935   0.009296  -3.113 0.001853 **
SibSp       -0.457024   0.125210  -3.650 0.000262 ***
Parch       -0.383983   0.163159  -2.353 0.018601 *
Mother      -0.918969   0.486195  -1.890 0.058741 .
Children    -1.789409   0.455633  -3.927 8.59e-05 ***
Respectable -0.637123   0.810663  -0.786 0.431909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  766.49  on 882  degrees of freedom
AIC: 784.49
```

Fig. 7 Summary of logistic model

From the above results, we concluded that mother, parch and respectable are not stastically significant and they have high P-value. Pclass, sex and children are the most significant values as they have low P-values. We removed parch, mother and respectable from the dataset to build the logistic model again. The summary of the improved model is shown in figure 8.

```
Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
    data = c_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9999  -0.5777  -0.4164   0.5854   2.6326

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.932899   0.820313   6.013 1.82e-09 ***
Pclass      -1.171831   0.122143  -9.594  < 2e-16 ***
Sex          2.773348   0.197965  14.009  < 2e-16 ***
Age         -0.026753   0.008975  -2.981 0.002874 **
SibSp       -0.524347   0.121122  -4.329 1.50e-05 ***
Children    -1.468103   0.431380  -3.403 0.000666 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  773.43  on 885  degrees of freedom
AIC: 785.43

Number of Fisher Scoring iterations: 5
```

Fig. 8 Summary of the improved model

The low P-values of Pclass and sex suggest that they have high correlation with the probability of having survived the disaster. The negative coefficient of Pclass implies that if all other variables are kept constant, people with higher Pclass value are less likely to survive. This basically means people with Pclass value 1 are more likely to survive than Pclass value 2 and people with Pclass value 2 are more likely to survive than Pclass value 3. Similarly, the positive coefficient of sex implies that if all other variables are kept constant people with higher sex value are more likely to survive. This translates to females (sex value 1) are more likely to survive the disaster than males (sex value 0). Lastly, we determined the accuracy of the model using the test data. We got an accuracy of 94.26%. We assumed the decision boundary to be 0.5. The class for which the probability was greater than 0.5, we considered that class to be the predicted class.

C. DECISION TREE

Next, we carried on the analysis by implementing Decision tree algorithm [8]. Decision tree gave us some useful insights. Some of the insights are if a passenger is female and she belongs to a passenger class of either 1 or 2, then the probability of survival is 0.95 and if a passenger is male and is greater than or equal to an age of 13, then the probability of his survival is 0.16. The generated decision tree is shown in figure 9.

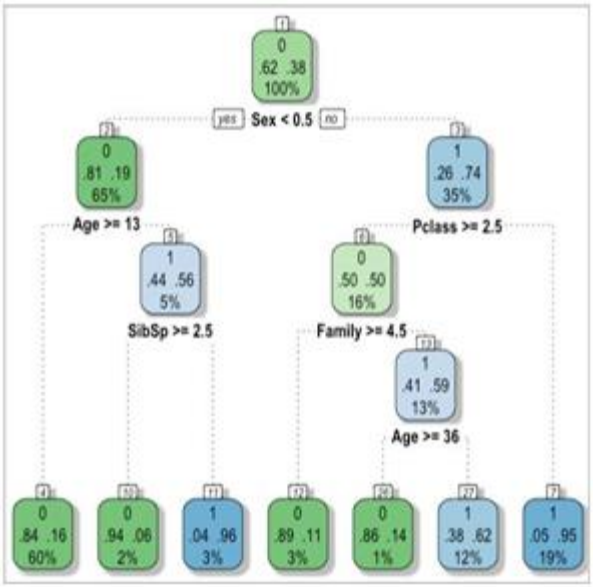

Fig. 9 Decision tree

From the confusion matrix in Table IX, we conclude that out of 418 predictions, our model made 389 correct predictions, giving an accuracy of 93.06%.

TABLE IX: CONFUSION MATRIX FOR DECISION TREE

	Actual	
Predicted	**Survived: NO**	**Survived: YES**
Survived: NO	252	15
Survived: YES	14	137

D. RANDOM FOREST

Next, we implemented Random forest algorithm for improving the accuracy of the classification model even further. 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 [9]. We built our model with all the variables of our 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, we used an argument "importance" while building our model.

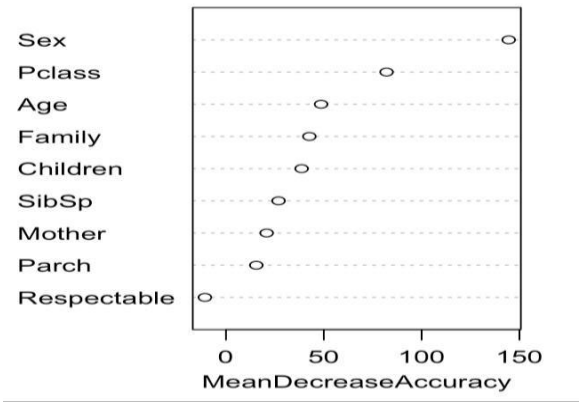

Fig. 10 Variable importance

From the above figure (Figure 10) we saw 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. We then checked the accuracy of the random forest algorithm on the test data.

TABLE X: CONFUSION MATRIX FOR RANDOM FOREST

	Actual	
Predicted	**Survived: NO**	**Survived: YES**
Survived: NO	250	18
Survived: YES	16	134

From the above confusion matrix (Table X), we can see that out of total 418 predictions, our model made 384 correct predictions, giving an accuracy percentage of 91.8%.

V. RESULTS

We choose two metrics to compare the four classification techniques we used in this research paper. First metric is accuracy and the second metric is false discovery rate. We calculate both this metrics using confusion matrix. Accuracy is the measure of how well our model predicts. Higher the accuracy the better. This is not the case for the second matrix. For the problem used in the research paper, false discovery rate is an important metric as it would be dangerous if we predict a passenger would survive but in truth he does not survive. Hence lower the false discovery rate the better. The accuracy and false discovery rate for each of the algorithm is listed in the table below:

TABLE XI:
COMPARISON OF ALGORITHMS

Algorithm	Accuracy	False Discovery Rate
Naïve Bayes	91.3%	15.47%
Logistic Regression	93.54%	8.60%
Decision Tree	93.06%	9.06%
Random Forest	91.86%	10.66%

VI. CONCLUSION/FUTURE WORK

We observed that 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. We also determined the features that were the most significant for the prediction. Logistic regression suggested that Pclass, sex, age, children and SibSp are the features that are correlated to the survival of the passengers.

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, Titanic: Machine Learning form Disaster [Online]. Available: http://www.kaggle.com/
2. Eric Lam, Chongxuan Tang. Titanic – Machine LearningFromDisaster.AvailableFTP: cs229.stanford.edu Directory: proj2012 File: LamTang-TitanicMachineLearningFromDisaster.pdf
3. Cicoria, S., Sherlock, J., Muniswamaiah, M. and Clarke, L, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster," pp. 4-6, May 2014
4. Vyas, K., Zheng, Z. and Li, L, "Titanic-Machine Learning From Disaster," pp. 6, 2015.
5. Mikhael Elinder. (2012). Gender, social norms, and survival in maritime disasters [Online]. Available: http://www.pnas.org/content/109/33/13220.full.
6. Jason Brownlee. (2014). How to implement Naïve Bayes in Python from scratch [Online]. Available: http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/
7. Wikipedia. Logistic Regression [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression.
8. Trevor Stephens. (2014). Titanic: Getting Started With R - Part 3: Decision Trees [Online]. Available: http://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/
9. Trevor Stephens. (2014). Titanic: Getting Started With R - Part 3: Decision Trees [Online]. Available: http://trevorstephens.com/kaggle-titanic-tutorial/r-part-3-decision-trees/