Titanic: Machine Learning From Disaster

Pratheek C Rajashekar
University of California
Riverside
[redacted]
pchin006@ucr.edu

Kamalika Poddar
University of California
Riverside
[redacted]
kpodd001@ucr.edu

ABSTRACT

We all know about the Titanic: The ship which drowned itself on its first sail. Our project is based on the data of the Passengers which can be used to find out the number of people survived and what was the reason behind their survival. Although, there exists some missing values in the dataset for features such as Age, Fare, Embarked and Cabin, there are many other features with zero missing values, such as Pclass, Sex, Title, Family Size and Ticket Size. An analysis is performed on these features by taking any of the feature as the splitting criterion. From the analysis, find out how these features affect the number of survivals. Since, the attribute Age has missing values, the median of the attribute Age is computed, which replaces the missing values in the dataset. Similarly, since the attribute Fare also has few missing values, there might be some possibility that these passengers got a free ticket on this maiden voyage. However, we compute the Fare for these missing values by performing Least Mean Square regression on the feature, Fare.

Furthermore, a Decision Tree model is trained using 80% of the training datasets. The other 20% of the training instances are used as a validation set for cross-validation of the model. Second, a Random Forest model is trained by removing the features such as Embarked, since it does not contribute much to the classification process. Since the ticket size is the same as the Family size, the Family size parameter is removed. This boosts the efficiency of the classifier by more than 1%. After analyzing these models, a comparison is made between the models. The model with the higher accuracy can be used in such disaster management incidents.

KEYWORDS

Data, Decision Tree, Random Forest.

1 INTRODUCTION

The Titanic incident that occurred on April 15,1912 lead to the loss of thousands of people on its maiden voyage. Though there was some amount of fortune for people on the ship in terms of survival, the probability of survival was more among the women, children and the upper-class. This project builds classifier models based on the entire passenger dataset. This classification facilitates to predict the survival of the passengers. The Titanic incident paved the way to improve the safety and security regulation of the ships.

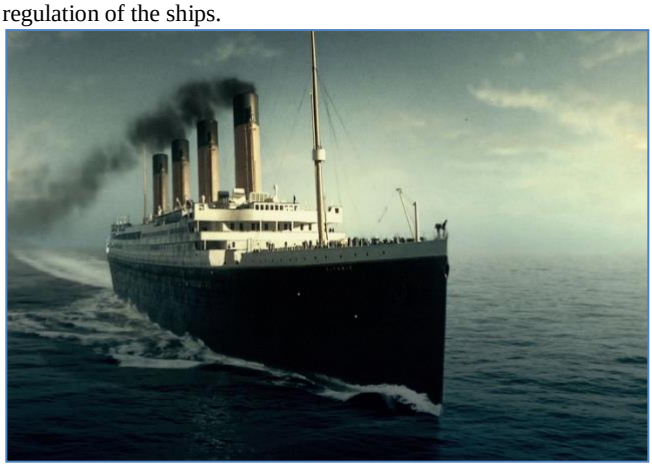


Figure 1: The historic Titanic which drowned itself, on its first sail, after getting struck by an iceberg.

Some of the conditions to consider during such calamities are:- a) Whether the person tries to save themselves first or gives priority to the spouse or to the children? Is the precedence given to the rich or the poor class? Who among the lower class of the ship survives first? Does the family size determine the rate of survival?

2 PROBLEM DEFINITION

The two problems which are to be addressed for the dataset are- a) The dataset obtained from Kaggle has missing values for some of the attributes such as Age, Fare, Embarked and Cabin. So, it is hard to classify the given dataset based on these features with missing values. These features cannot be used as the splitting criterion until the missing values are added back to the dataset. b) Given the training dataset, the classifier should predict whether the passengers from the validation or test dataset survived or not.

3 ANALYSIS

3.1 Exploratory Analysis

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often

with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.

3.2 Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

3.3 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

3.4 Least Mean Square Regression Analysis

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

4 RESULTS AND DISCUSSION

4.1 Passenger class relation to Survival of Passenger.

Among all the attributes the first attribute was Pclass and it was found that it was an important attribute in determining the survival rate

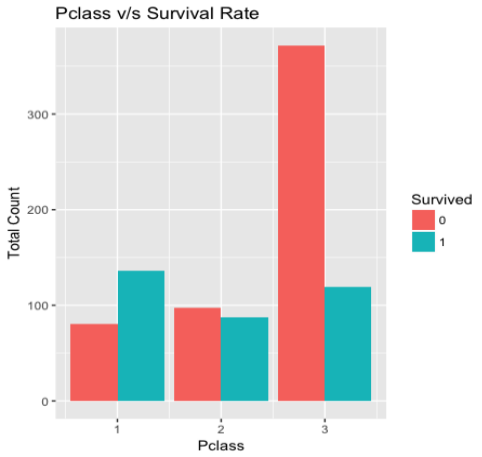


Figure 2: The ship had three classes, out of which the 3rd class has the least survival rate.

The Passengers belonging to the first class survived mostly and the passengers who belonged to the third class suffered the most death.

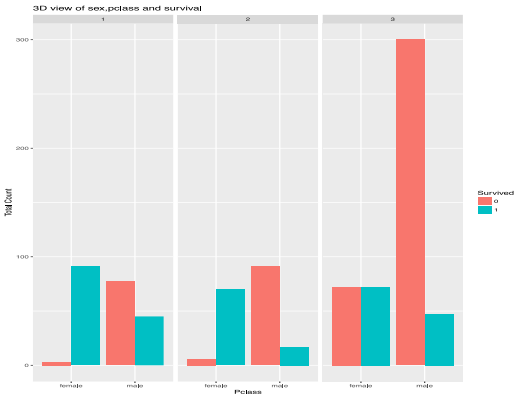


Figure 3: This figure illustrates the survival of a particular gender with respect to the class.

A 3D analysis was carried out to determine which Sex survived the most. It was found that females from the first class had the highest survival among all the passengers on the ship. Males from the third class were not able to escape the ship on time.

4.2 Survival Prediction with the relation of TITLE

To find out about different features, an analysis was performed on the designated “Title” that a person belonged to. There were 18 different title that could be observed from the dataset. These title could be a result of human mistakes as it was seen for Mlle,Mme. So for this project all 18 titles were grouped into five different tiltes- Mr, Mrs, Miss,Master, Officer.

Title as given in the data:

Capt	1	Miss	260
Col	4	Mlle	2
Don	1	Mme	1
Dona	1	Mr	757
Dr	8	Mrs	197
Jonkheer	1	Ms	2
Lady	1	Rev	8
Major	2	Sir	1
Master	61	The Countess	1

After merging the Title:

Title	No. of Entries
Officer	27
Master	61
Mrs	198
Miss	266
Mr	757

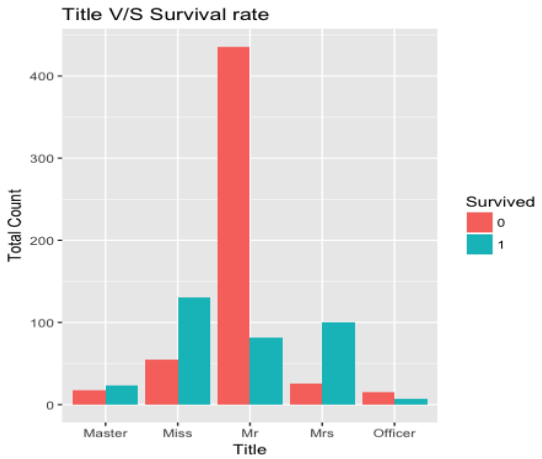


Figure 4: This figure shows the classification of data by the title of the passenger

It was observed that passengers with title Mr. survived the least among all the other Passengers.

Figure 5: The relationship between the passenger class and the title is plotted.

Plotting the designated Title against the three classes-1,2 and 3 gave us the information that the title with Mr. from the third class survived the least. All the passenger designated as Master from 1 and 2 class survived and all the Officer from second class died. These results pretty much prove the theory as we know what happened in history.

4.3 Family Size relation to Survival of Passenger

The next modified feature was Family size. This was used to find out how many people in the family survived or the survived people were linked to a family. We divide the passenger data into three groups- 1-Single, 2 to 5- Small, more than 5 as Big families.

Number of Members	Family size
1	Single
2-5	Small
5>=	Big

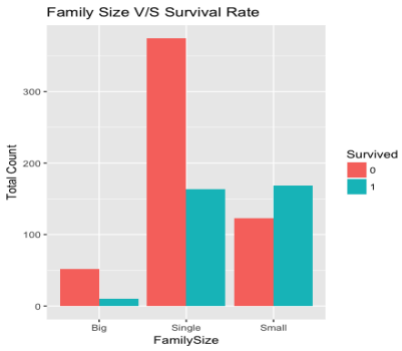


Figure 6: The figure gives the total number of survivals as per the family size.

The number people with small family survived the most and the Single had to suffer.

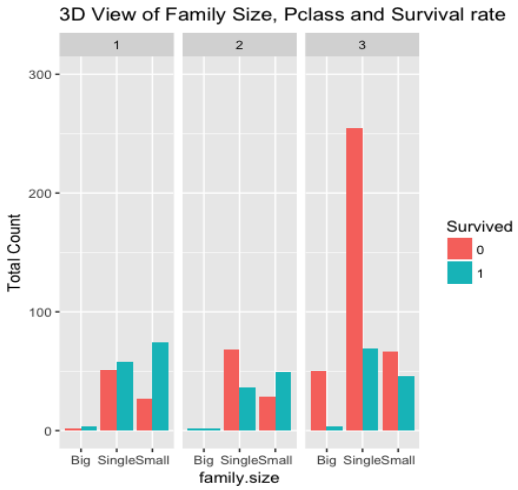


Figure 7: The relationship is built between the Family size and the Passenger Class.

After 3D analysis of the family size with Pclass, it was observed that small families from the first class survived the most and big family from the second class survived and as expected the single from the third class had to meet death.

4.4 Ticket Size vs Survival of Passenger

There was another change in the attributes and this time it was the tickets. This attribute helped us in learning how many tickets a person bought each time. The analogy taken was same as maintaining the family.

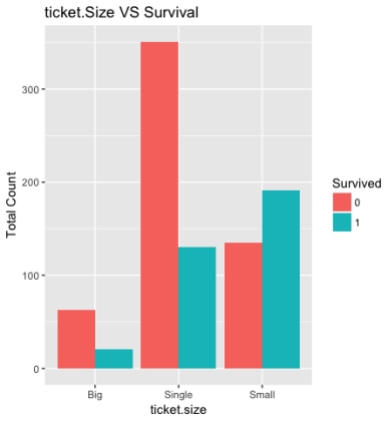


Figure 8: The relationship is built between the Ticket size and Survival Passengers.

The passengers who bought ticket for small families survived mostly.

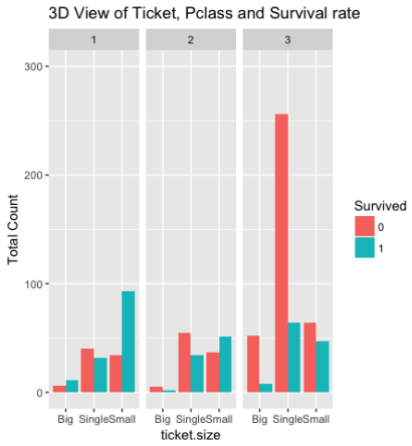


Figure 9: The 3D relationship is built between the Ticket size and the Passenger Class to estimate the survived passengers.

Single third class passengers survived the least whereas the small families with first class passengers survived the most.

4.5 Embarked vs Survival of Passenger

All the passengers boarded from three ports – Cherbourg, Queenstown and Southampton. The left for New York city from Southampton. There were two missing values for Embarked, we included those with Southampton and carried out the analysis.

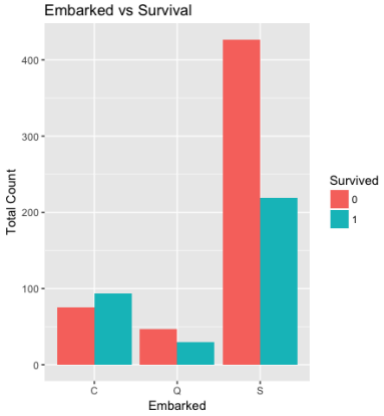


Figure 10: The relationship is built between the Embarked and Survival Passengers.

Mostly people who boarded from Cherbourg survived the most, and people boarded from Southampton surffered the most loss.

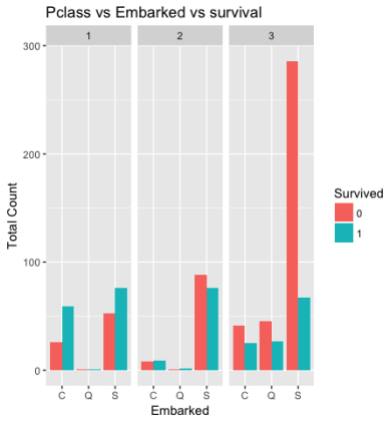


Figure 11: 3D plot between Embarked, Pclass and survival.

It was observed very few people boarded from Queenstown and most of the people survived was from first class and most of the people boarded from Southampton.

4.6 Decision Tree

After training the model with the provided data, the decision tree takes in the attribute which mattered the most as result it is seen that "title" had the highest rank among all and then it split into Pclass and at last Ticket size.

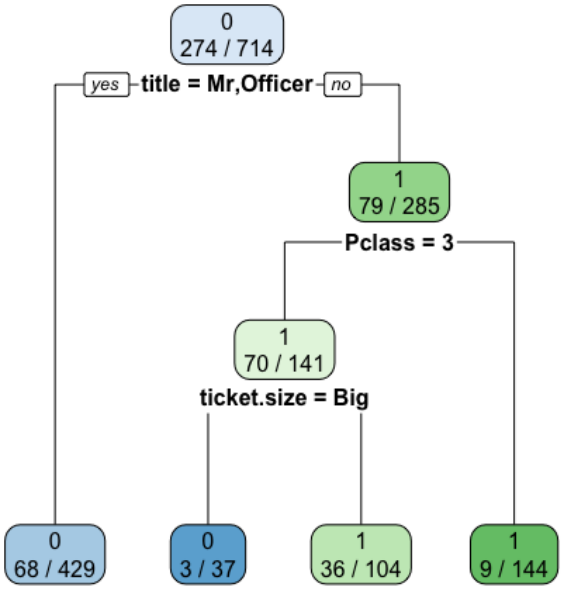


Figure 12: Decision Tree resulting an accuracy of 81.92%

The Decision tree results in 83.75% accuracy. After using 10-Fold Cross Validation technique it gives us an accuracy of 81.92%.

4.7 Random Forest

We applied Random Forest on the data.

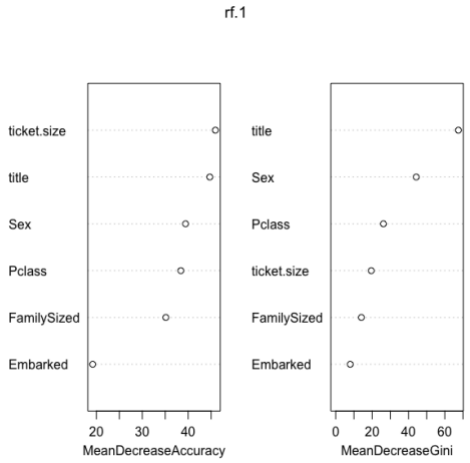


Figure 13: Random Forest with six attributes.

It was observed than title had the highest rank and we got an accuracy of 82.91% which was better than Decision tree.

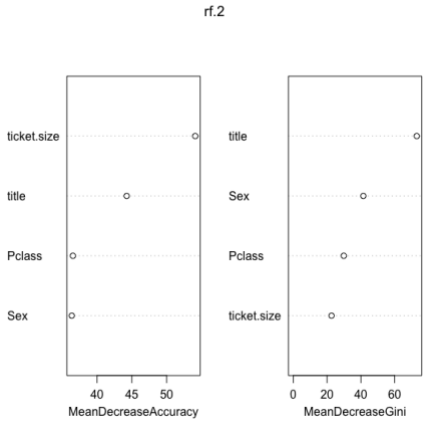
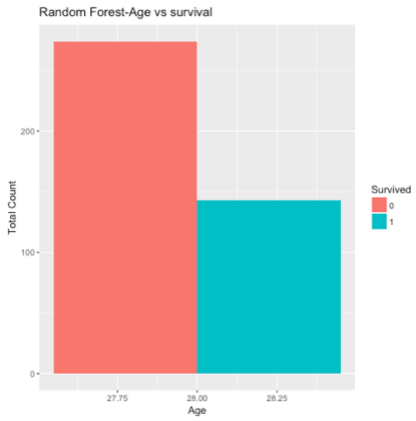


Figure 14: Random Forest with four attributes.

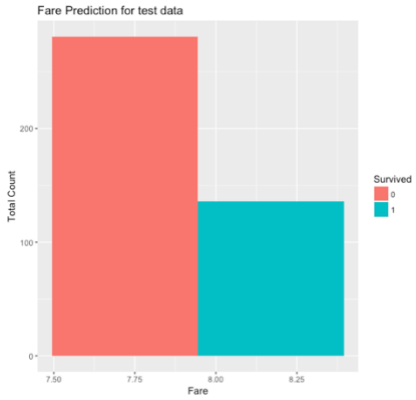
It was observed that removing Family size and Embarked, Random Forest resulting in ann accuracy of 84.03%. After verifying with 10-Fold cross Validation we received an accuracy of 83.923%.

4.8 Age and Fare analysis

There were many missing age in the provided dataset. The missing age were computed by finding the median and computing in each of the missing values.



The missing values for Age was computed as 28.



There was one missing value for Fare in the test data. The missing value for fare was found to be 7.945 after performing Least Mean Square Regression.

5 CONCLUSIONS

In conclusion, we observe that most people from first class survived. The total number of female survivals are higher than number of male survivals. Results show that when data is split into five titles- Master, Miss, Mr, Mrs, Officer, then, it is evident that the Master from first and second class has 100% survivals whereas officers from second class has 0% survivals. If the families are split as Single-1, Small-2 to 5, Big- more than 5, then the number of survivals for small families in first class is higher than other classes. We also try to split the number of ticket bought by the same person into ticket size, we observe that the output is same as family size and any one of the two can serve as an attribute. Since there are two missing values for Embarked, these values are merged with Southampton.

The decision tree takes three variables –Title (Mr, Officer), Pclass and ticket Size and predicts the accuracy after cross validation as 81.92%. Random Forest has accuracy of 82.91% with attributes such as – ticket size, sex, Pclass, Title, Family Size, Embarked. Considering only ticket size, sex, Pclass, Title -an accuracy of 83.923% is obtained after validation using 10-fold cross validation. The missing values of Age in the dataset are predicted by taking the median which is 28. Using Least Mean Square Regression Analysis, the missing Fare value of the row 1044 was predicted to be 7.9450.

A Contents

A.1 Introduction
A.2 Problem Definition
A.3 Analysis
A.3.1 Exploratory Analysis

A.3.2 Decision Tree

A.3.3 Random Forest

A.3.4 Least Mean Square Regression Analysis.

A.4 Results and Discussion
A.4.1 Passenger class relation to Survival of Passenger

A.4.2 Survival Prediction with the relation of TITLE

A.4.3 Family Size relation to Survival of Passenger

A.4.4 Ticket Size vs Survival of Passenger

A.4.5 Embarked vs Survival of Passenger

A.4.6 Decision Tree

A.4.7 Random Forest

A.4.8 Age and Fare analysis

A.5 Conclusions
A.6 References

ACKNOWLEDGMENTS

We thank Prof. Vagelis Papalexakis for his assistance throughout the project. Also, TA Ekta Gujral who helped us in coming up with ideas on the project.

REFERENCES

[1] Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll "An Introduction to Logistic Regression Analysis and Reporting" The Journal of Educational Research

[2] Leo Breman "Random Forests", Machine Learning 45, 5-32,2001.

[3] Akihito Hagihara,Manabu Hasegawa, Takeru Abe, Takashi Nagata, Yoshifumi Wakata, Shogo Miyazaki "Prehospital Epinephrine Use and Survival Among Patients With Out-of-Hospital Cardiac Arrest" JAMA, March 21, 2012 Vol 307, No. 11.

[4] Longyun Dong, Xibing Li, Gongnan Xie "Nonlinear Methodologies for Identifying Seismic Event and Nuclear Explosion Using Random Forest, Support Vector Machine, and Naive Bayes Classi cation" , Hindawi Publishing Corporation, Vol. 2014, Article ID 459137, 8 pages .

[5] https://www.kaggle.com/swamysm/beginners-titanic/notebook

[6] https://datasciencedojo.com

[7] https://stackoverflow.com

[8] https:// Wikipedia.com