**MSCI 623**

**FINAL REPORT**

PREDICTION OF PASSENGERS SURVIAL ON THE TITANIC USING MACHINE LEARNING

**By**

Shahed Ali Khan (20762548)

Sudheer Uppalapati (20802432)



***ABSTRACT***

The sinking of the Titanic is the symbolism of Man vs Nature. Even 100 years after the Titanic met its fatal end, the story of the tragic wreck continues to fascinate people worldwide. In this report, the accuracy of machine learning models is explored in predicting the survival of passengers who were aboard on the RMS Titanic, which on its maiden voyage from Southampton to New York city sank in the North Atlantic Ocean. The intriguing observation which comes out from the sinking is that some people were more likely to survive than others.

The objective of this project is then to build a predictive model to predict which of the passengers survived the ship wreck. In particular, the response variable “Survived” will be modeled. On the data collected, we implemented classification algorithms like Logistic Regression, KNN, Naive Bayes and Decision tree using K-fold cross validation. We also implemented Association Rule Mining algorithm to find any common patterns like if-then using criteria of support and confidence. We found that Logistic Regression model had highest accuracy on the training data and so we implemented it in predicting the survival of the passengers on the testing data and through apriori we found two rules under chosen support and confidence.

***INTRODUCTION***

The most infamous disaster which occurred over a century ago on April 15, 1912, that is well known as sinking of “The Titanic”. The collision with the iceberg ripped off many parts of the Titanic. Many classes of people of all ages and gender where present on that fateful night, but the bad luck was that there were only few life boats to rescue. It was touted to be “Unsinkable”, yet it sank. It is virtually the only disaster that is perpetually remembered and commemorated. The disaster led to laws and treaties ensuring that enough lifeboats were to be carried for all aboard, to have emergency backup. From increased training and appropriate personal protection to standardizing requirements for emergency procedures, maritime safety has improved, and many lives have been saved.

The field of machine learning has allowed analysts to uncover insights from historical data and past events. The dataset of the Titanic survival is publicly available on a website called Kaggle.com. This dataset has been studied and analyzed using various machine learning algorithms like Logistic Regression, KNN, Naive Bayes and Decision tree. The approach is centered on Python for executing the machine learning algorithms. Our main focusing is to first explore hidden or previously unknown information by applying exploratory data analytics on the available data and then apply machine learning models to predict the survival of the passengers. We compared the accuracy of all the models and use the model with the highest accuracy on the dataset. Our project is interesting because we have many missing values in our data and we filled them using mean, median and mode techniques. We transformed some of our explanatory variables to suit our model. We dropped some of our explanatory variables which may not suitable for modeling.

***RELATED WORK***

Mikhael Elinder in his research, analyzed the relationship of social norms and sex with survival. It was concluded by him that on the Titanic, the survival rate of women is more than three times higher than the survival rate of men.

Trevor Stephens carried out the prediction of the survival of the passengers using Random Forest classifier and Decision tree algorithms using the following parameters such as Pclass, FamilyID, Family Size, Title, Fare, SibSp, Parch, Sex, Age and Embarked. The accuracy of the implemented algorithms is not mentioned.

Zeshi Zheng, Kunal Vyas and Lin Li suggested that dimensionality reduction and playing more with the dataset could improve the accuracy of the algorithms. The most important conclusion provided by them is that more features utilized in the models do not necessarily make results better.

Benno Torgler, Bruno S. Frey and David A. Savage concluded that people in their prime age died less often than older people. Passengers with high financial stability, traveling in first class, are better able to save themselves as are passengers in second class as compared to third class.

***DATA***

SUPERVISED LEARNING

**PROJECT FLOW**

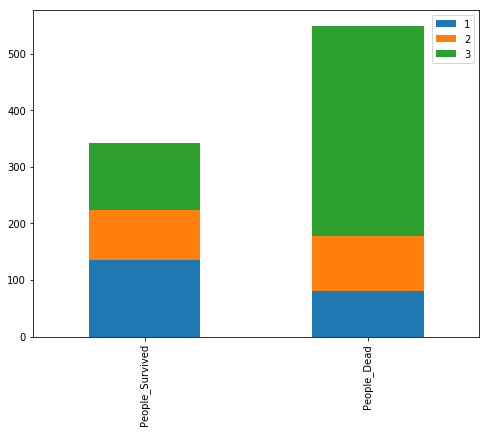
The dataset is provided by the Kaggle website. The training data set has 12 columns with 891 rows. The testing data has 11 columns with 418 rows. The data is a passenger sample with their associated labels. For each passenger, the name of the passenger, sex, age, his or her passenger class, number of siblings or spouse on board, number of parents or children aboard, cabin, ticket number, fare of the ticket and embarkation are provided. The description is provided below –

|  |  |
| --- | --- |
| Attributes | Description of the attributes |
| PassengerID | Identification number of the passengers. |
| Pclass | Passenger class (1, 2 or 3) |
| Name | Name of the passengers aboard |
| Sex | Gender of the passengers (male or female) |
| Age | Age of the passengers |
| SibSp | Number of siblings or spouse for a passenger on the ship |
| Parch | Number of parents or children for a passenger on the ship |
| Ticket | Ticket number |
| Fare | Price of the ticket |
| Cabin | Cabin number of the passenger |
| Embarked | Port of embarkation |
| Class Variable | Description of Class Variable |
| Survived | Target variable (values 0 for dead and 1 for survived) |

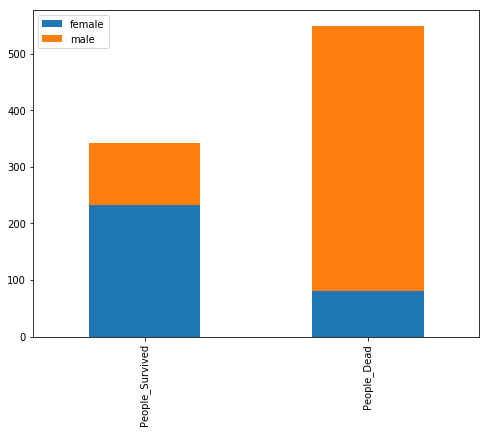
Before building a model to predict the survival of passengers, we have explored the data and found that Age and cabin column had NA values in them. Age column had 177 rows with NA values and cabin column had 687 rows with NA values. We found the impact of Pclass, Sex, SibSp and Parch on the class variable.

***DATA VISUALISATION***

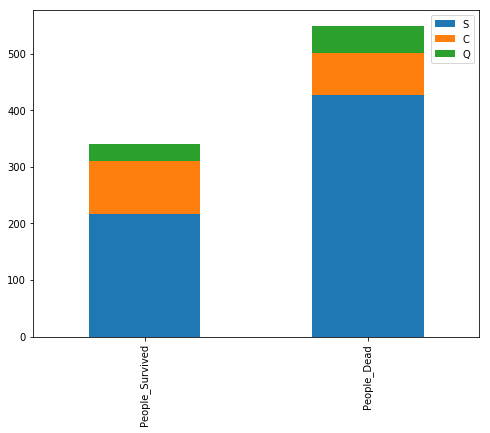
**Fig. 1. Bar chart for Pclass variable**



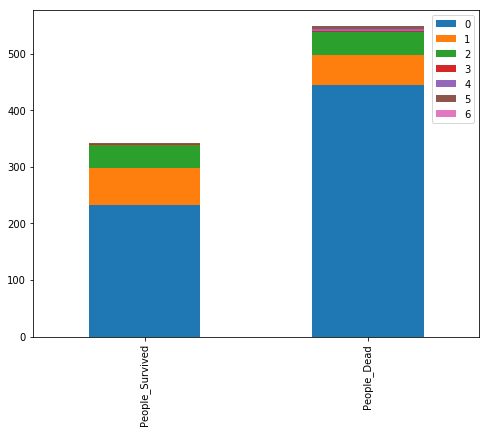
**Fig. 2. Bar chart for sex variable**



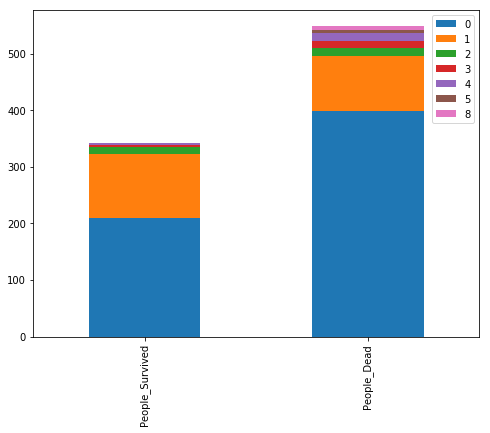
**Fig. 3. Bar chart for Embarked variable**



**Fig. 4. Bar chart for Parch variable**



**Fig. 5. Bar chart for Sibsp variable**



***DATA TRANSFORMATION***

1. Name - For the Name variable we have various titles before the names of the passengers. So, we first extract all the titles and inspect them. Later we map them to the four labels (category- 0,1,2,3) so that we can use the data in applying the machine learning algorithms.
2. Embarked – For the embarked variable we found missing values and so we proceed filling them by using the mode(S). Also we map the three categories to the three labels (category- 0,1,2) so that we can use the data in applying the machine learning algorithms.
3. Fare - For Fare we filled missing values using median of the fares.
4. Age - For Age we filled missing values using median of the Age. Also, we have mapped age into various categories for the ease of modeling.
5. Cabin – We have mapped cabin values to various levels. Then we filled missing values using median of those levels.
6. Sex – For sex we categorised male and female into 0 and 1.

After data transformation when we inspect the data, we find ticket and passenger id not required for modeling. Therefore, we drop them from our data.

UNSUPERVISED LEARNING

For association using Apriori algorithm we have chosen 4 explanatory variables namely Pclass, Sex, Age, Survived and categorised those variables suitable for applying Apriori.

***RESULTS***

Prediction models are generated using four machine learning algorithms namely Naive Bayes, Logistic Regression, Decision tree and KNN.

Logistic Regression is a type of classification algorithm in which the class variable is categorical and binary. In our dataset the class variable, “survived” is the dependent variable which is both binary and categorical (1 for survival and 0 for dead). Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Cabin, Embarked and Title are the features used in building the logistic regression model.

Decision tree learning is the method of construction of a decision tree from class labeled training tuples. A decision tree can be considered as a flow-chart like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Cabin, Embarked and Title are the features used in building the decision tree model.

Nave Bayes is a classification algorithm that applies Bayes theorem to build a prediction model. It is based on some naive assumptions about the features. The assumption is that all the features are independent of each other. In other terms, the probability of value of one feature belonging to a class is independent of all other features. Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Cabin, Embarked and Title are the features used in building the decision tree model.

Comparison of the Models

|  |  |  |
| --- | --- | --- |
| S. No | Model | Accuracy of model using 10-fold cross validation |
| 1 | Naive Bayes Classifier | 0.78 |
| 2 | Decision Tree Classifier | 0.80 |
| 3 | KNN classifier | 0.76 |
| 4 | Logistic Regression | 0.81 |

Each of these algorithms are compared to one another on the basis of the accuracy percentage and finally on the basis of accuracy score of the 10-fold cross validation, the Logistic Regression model is used to predict the class variable for the test data.

For unsupervised learning algorithm, by setting confidence = 0.6 and minimum support = 0.7 we have found Adult, Dead, Male and (Adult, Male) having minimum support of 0.6. We have obtained only two rules.

Rule 1: Adult → male

We can infer that if we calculate the sex of adults based on confidence of 0.66. we can say that they are males.

Rule 2: male → Adult

We can infer that if we calculate the age of males based on confidence of 0.93. we can say that they are adults.

*CONCLUSION*

We know the very fact that the accuracy of the models may vary when the choice of feature modelling is different. Logistic Regression model was found to be the best among the other algorithms because of its high accuracy. Since the difference between the accuracy of the other models when compared with logistic regression was found to be low, probably there should no overfitting with the usage of logistic regression model. Also, we had determined the features that are significant for the prediction.

Future work might include potentially validating more using pruning techniques, grouping the data with respect to different conditions, predicting chances of survival in specific to an explanatory variable, for example survival chances of males in a specific compartment is higher than that of males in another compartment and so forth. Also, other machine learning techniques could be implemented to see if they have high accuracy on predicting the survival of passengers without overfitting. Much focussed conclusions can be drawn about the survival rate of the passengers by comparing and combining the results obtained using various machine learning algorithms.

***REFERENCES***

• "Titanic Passengers and Crew Listings". encyclopedia titanica. Retrieved 15 July 2011.

• ^ Lord, Walter (1976). A Night to Remember. London: Penguin Books. p. 197. ISBN 978-0-14-004757-8.

• ^ "Passenger List and Survivors of Steamship Titanic". United States Senate Inquiry. 30 July 1912. Archived from the original on 26 March 2012. Retrieved 15 July 2011.

• ^ Hall, Wayne (1986). "Social Class and Survival on the SS Titanic" (PDF). Social Science & Medicine. 22 (6): 687–690.

• ^ Barratt, Nick (2009). Lost Voices From the Titanic: The Definitive Oral History. London: Random House. p. 93. ISBN 978-1-84809-151-1.

• ^ Howells, Richard (1999). The Myth of the Titanic. United Kingdom: MacMillan Press. p. 18. ISBN 978-0-333-72597-9.