Titanic Disaster
Logistic Regression

Devcharan K
Indian Institute of Technology, Madras

***Abstract*—The sinking of the Titanic ship that caused the death of hundreds of passengers and crew members is one of the most fatal disasters in history. The loss of lives was mostly caused due to a shortage of lifeboats on the ship. A logistic regression analysis of an extensive dataset on the Titanic passengers is presented which tests the likelihood that a Titanic passenger survived the accident based upon passenger characteristics. The main objective of the algorithm is to firstly find predictable or previously unknown data by implementing exploratory data analytics on the available training data and then apply different machine learning models and classifiers to complete the analysis. This will predict which people are more likely to survive.**

I. Introduction

The sinking of the RMS Titanic, a British passenger liner, is one of the most infamous shipwrecks in history. On April 15,1912, during her maiden voyage from Southampton to New York City, the Titanic sank after colliding with an iceberg. Of the estimated 2224 passengers and crew aboard, almost 1500 died, making this one of the most fatal accidents in history. The Titanic only carried enough lifeboats for 1178 people-slightly more than half of the number on board and one third of her total capacity.

Logistic Regression is a statistical analysis method used to predict a data value based on prior observations of a dataset. It is an important tool in machine learning as it allows an algorithm being used in a machine learning application to classify incoming data based on historical data. Logistic regression is most commonly used for algorithms with binary classification problems which are problems which have a prediction output of 'True' or 'False' and 'yes' or 'no' etc. The different characteristics of the *Train* dataset given is used to train the model to use said characteristics to predict the required result, which in this case is whether or not the given passenger survived the accident.

In this paper, the data about passengers of the Titanic such as Age, Sex, Passenger class, Fare of the ticket purchased, Embarking station etc. are considered and the model is used to determine whether a given passenger survived the accident or not. The goal is to find a correlation between the various characteristics of the passengers and their survival chances.

In Section 2, Logistic Regression and its applications are explained. In Section 3, the problem and the method solve it is shown. Section 4 is an Exploratory analysis of the cleaned data-set with the usage of Python packages namely Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn and visualizations of the same. In Section 5, the concepts used to prepare the model is evaluated. Finally the inferences of the analysis is presented in Section 6.

II. Logistic Regression

Regression is a type of supervised learning. Logistic Regression attempts to model the relationship between several classes of data to predict the pass/fail probability of a certain event based on the data. It can be used to handle binary classification problems where the outcome is one of two categories. It takes the features as input and predicts the probability of occurrence of a binary event using a logistic function. It is a special case of linear regression where the target variable is categorical in nature and therefore can be used for classification.

The linear regression function is given by

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n \tag{1}$$

Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric value. Below is an example logistic regression equation:

$$y = \frac{e^{(b_0 + b_1 x)}}{(1 + e^{(b_0 + b_1 x)})} \tag{2}$$

Where y is the predicted output, b_0 is the bias or intercept term and b_1 is the coefficient for the single input value (x). Each column in the input data has an associated b coefficient (a constant real value) that must be learned from the training data.

The goal of the logistic regression algorithm is to create a linear decision boundary separating two classes from one another. This decision boundary is given by a conditional probability.

$$P(y = 1|x; w)(probability for class 1) \tag{3}$$

$$P(y = 0|x; w)(probability for class 2) \tag{4}$$

Let us assume that there are two classes, one above the decision boundary (say 1) and one below the decision boundary (0). Logistic Regression calculates the probability of a particular set of data points belonging to either of those classes

given the value of x and w. The logic is that, say there is a set of values that is obtained from negative infinity to positive infinity based on a linear model that needs to be narrowed down to a score between zero and one. The link function, sigmoid function, takes care of this. The probability function is given by:

$$\frac{e^{(w_0+w_1x_1+w_2x_2)}}{1+e^{(w_0+w_1x_1+w_2x_2)}} \tag{5}$$

Now to get the probability of the alternate class, the value obtained above must be subtracted by 1 to obtain the sigmoid link function :

$$P(y=1) = \frac{1}{1+e^{(w_0+w_1x_1+w_2x_2)}} \tag{6}$$

$$P(y=0) = \frac{e^{(w_0+w_1x_1+w_2x_2)}}{1+e^{(w_0+w_1x_1+w_2x_2)}} \tag{7}$$

III. The Problem

The passenger details such as Age, Sex, Name, Passenger class, Number of siblings, Number of parents/children, Ticket fare, embarkment station etc of the Titanic passengers is given. The aim of this exercise is to predict whether a given passenger survived the disaster or not based on the aforementioned characteristics. The *Train* dataset has the information about whether the passengers survived or not along with the other information while the *Test* dataset does not have information about the survival. The *Train* dataset is to be used to train the model and the model should predict the survival chances of the samples in the *Test* dataset.

A. Data Cleaning

The analysis begins by cleaning the data by removing any non-integer values in integer columns and replacing missing values with an appropriate substitute. All missing values need to be filled. There are 177 missing values in Age data, 687 missing values in Cabin data and 2 missing values in Embarked data. Since 77.1% of the Cabin data is missing, no reasonable value can be imputed in the missing values' place as too many values are missing to draw any meaningful inferences from this data. Cabin data will be ignored in the analysis.

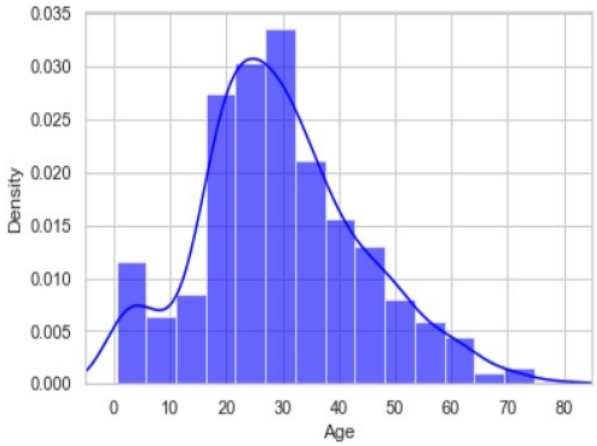

Fig. 1. Age Density of the Train Dataset

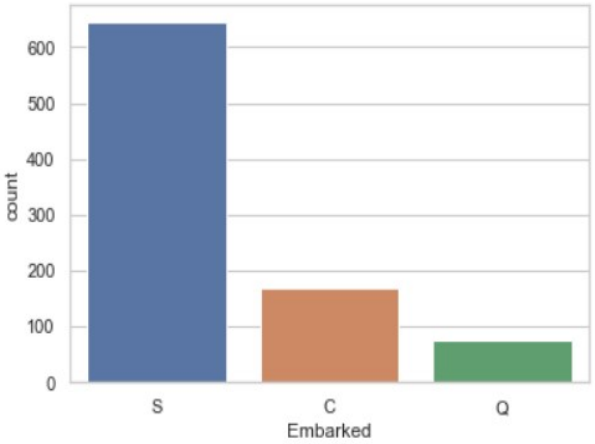

Fig. 2. Bar Plot of Embarked Ports

The missing data from Age cannot be imputed with the mean of the existing Age data as the density plot of the Age values is (right) skewed (Fig 1.). So, the median value will be imputed. The two missing Embarked data points can be imputed with the mode of the Embarked data which is the port where the most people boarded, which is Southampton (Fig 2.).

Additional categorical variables are created to replace the values in 'Sex', F and M with 0 and 1. The column PClass has values 1, 2 and 3. This is replaced with three columns namely PClass1, PClass2 and PClass3 each having values 0 when false and 1 when true.These are known as Hyperparameters. (This step helps to keep the analysis simple later on.)

The same cleaning techniques are applied to the Test dataset.

IV. Exploratory analysis

The aim is to see how the different attributes are related to survival rate in the Train dataset.

A. Age

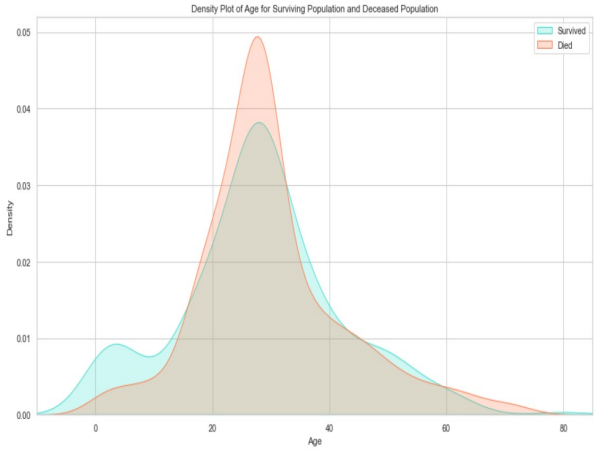

Fig. 3. Density plot of Age for Surviving and Deceased passengers

From Fig 3., the age distribution for survivors and deceased is actually very similar. A large proportion of the survivors were children, this shows that the people made an attempt to save the children by letting them get on the lifeboats.

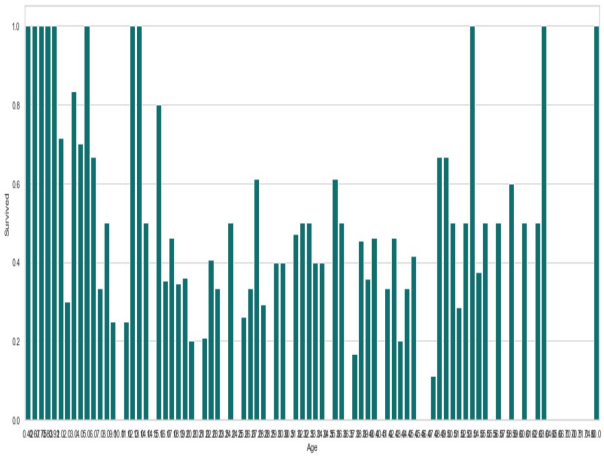

Fig. 4. Bar Plot of Survival Rate by Age

From Fig 4., upon inspection of the survival rates of under-16 passengers another variable can be made in place of age called 'Minor'. All passengers under the age of 16 can be categorized as a Minor and the remaining as adults.

B. Fare

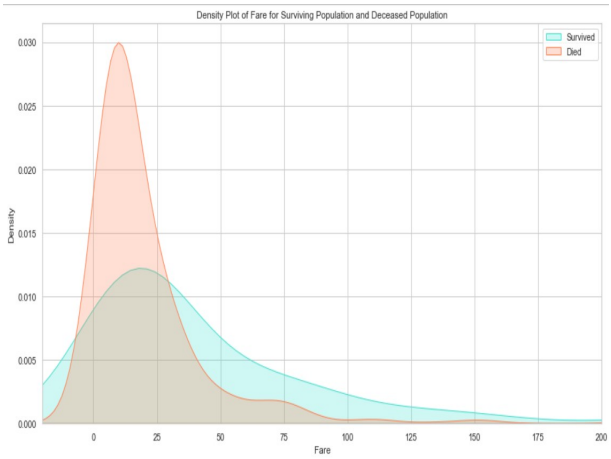

Fig. 5. Density plot of Fare Cost for Surviving and Deceased passengers

In Fig 5., as the distributions are clearly different for the fares of survivors vs. deceased, it's likely that this would be a significant predictor in the final model. Passengers who paid lower fare appear to have been less likely to survive.
A correlation between Passenger Class and Survival rates can also be expected as higher class passengers should be more likely to survive.

C. Fare and Age Correlation

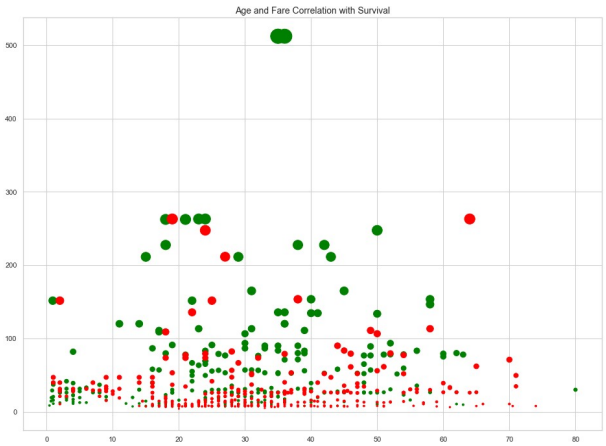

Fig. 6. Correlation Between Fare Cost and Age

Fig 6. shows the correlation between Age and Fare shows that for children (under 10 years of age) the fare paid does not matter but as the age increases passengers who paid more for their tickets were more likely to survive.

D. Passenger Class

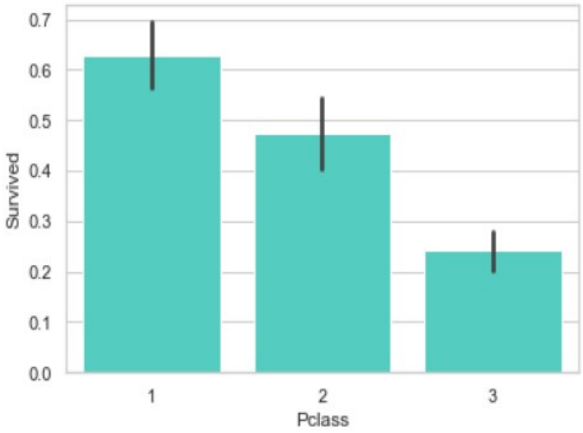

Fig. 7. Bar Plot of Passenger Class vs Survival rate

Unsurprisingly, being a first class passenger was safest.

E. Embarked Port

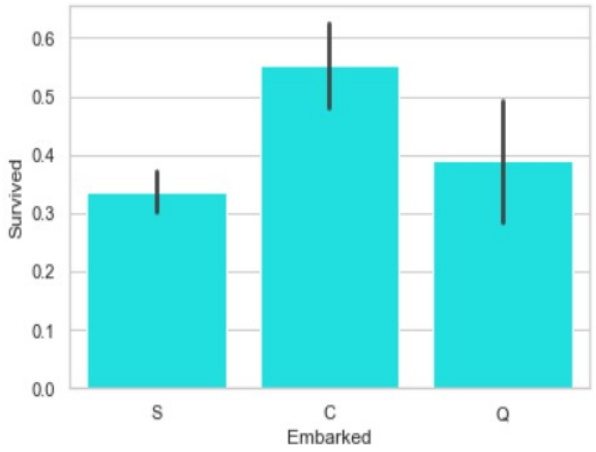

Fig. 8. Bar Plot of Embarked Port vs Survival rate

Fig 8. shows that passengers who boarded in Cherbourg, France, appear to have the highest survival rate. Passengers who boarded in Southampton were marginally less likely to survive than those who boarded in Queenstown. This is probably related to passenger class, or maybe even the order of room assignments (e.g. maybe earlier passengers were more likely to have rooms closer to deck). It's also worth noting the size of the whiskers in these plots. Because the number of passengers who boarded at Southampton was highest, the confidence around the survival rate is the highest. The whisker of the Queenstown plot includes the Southhampton average, as well as the lower bound of its whisker. It's possible that Queenstown passengers were equally unlucky, or unluckier than the Southampton passengers.

F. Traveling Alone

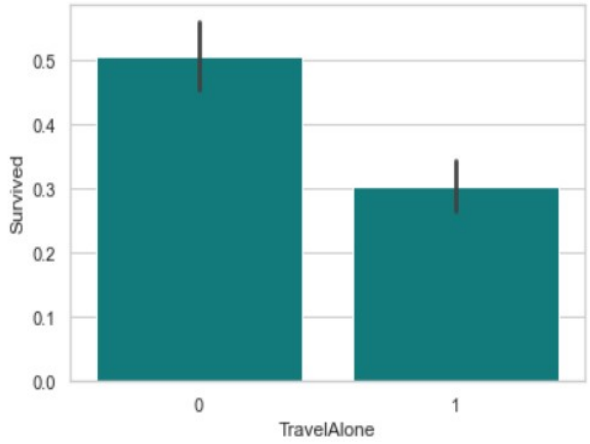

Fig. 9. Bar Plot of those Traveling alone vs Survival Rate

Individuals traveling without family were more likely to die in the disaster than those with family aboard. Given the era, it's likely that individuals traveling alone were likely male.

G. Gender

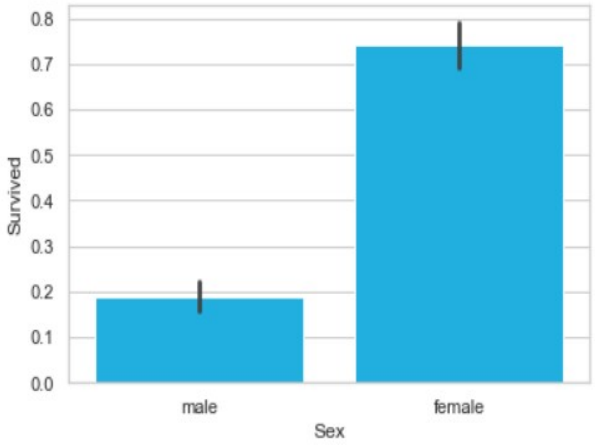

Fig. 10. Gender vs Survival Rate

Unsurprisingly, women were more likely to survive than men. The code of conduct, "Birkenhead drill" (Women and children first) seems to have been observed during the disaster.

V. Model Evaluation

A. Recursive Feature Elimination

Recursive Feature Elimination is a feature selection algorithm that fits a model and removes the weakest features until the specified number of features is reached. Features are ranked by the model's *coef* or *feature_importance* attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

RFE requires a specified number of features to keep, however

it is often not known in advance how many features are valid. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features.

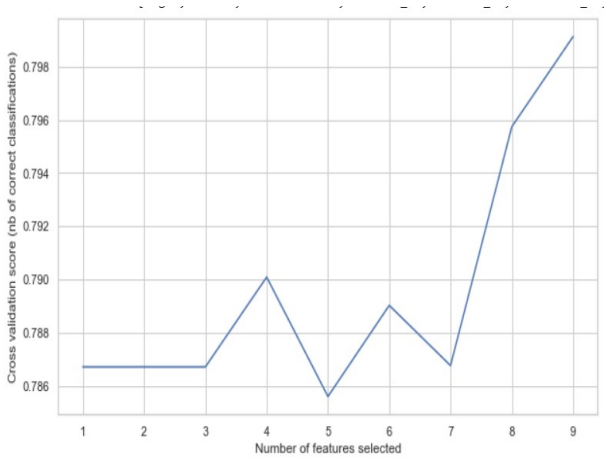

Fig. 11. Scores of RFE after Cross Validation

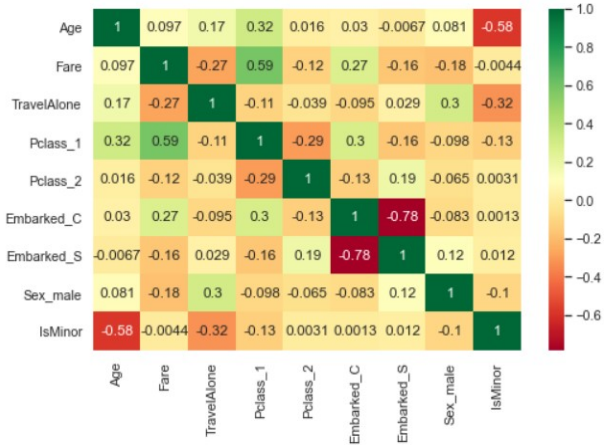

Fig. 12. Correlation strengths between the different features

9 attributes chosen for the model are as follows: *'Age', 'Fare', 'TravelAlone', 'Pclass_1', 'Pclass_2', 'Embarked_C', 'Embarked_S', 'Sex_male', 'IsMinor'*.

B. Training the Model

First the *train_test_split* method is used to train the model where the data within the *Train* dataset is divided into training and testing datasets so that the model can use the data to learn and compare its prediction with the real value of the event the model is trying to predict. This procedure is necessary before applying the model on the *Test* dataset to predict the survival rates.

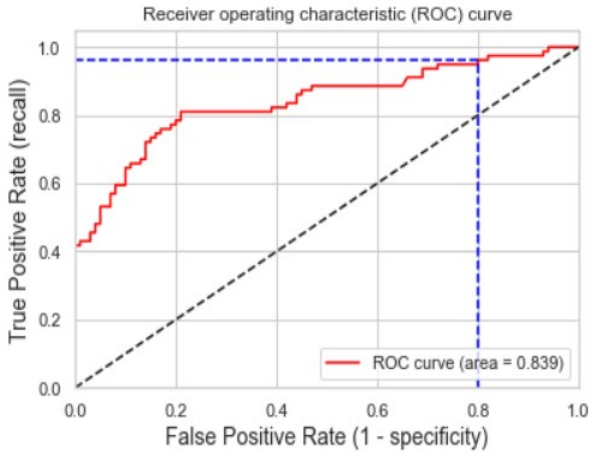

Fig. 13. ROC curve of the Model

Fig 11. is the ROC curve of the model and the Confusion Matrix is as follows. $\begin{bmatrix} 90 & 31 \\ 10 & 48 \end{bmatrix}$ Confusion matrix is used to show the performance of the algorithm. Accuracy of the model can be predicted using the confusion matrix.
GridSearch is a method to find the best hyperparameters for a specific model. Grid Search uses a different combination of all the specified hyperparameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters. This makes the processing time-consuming and expensive based on the number of hyperparameters involved. GridSearchCV performs cross validation along with grid search.

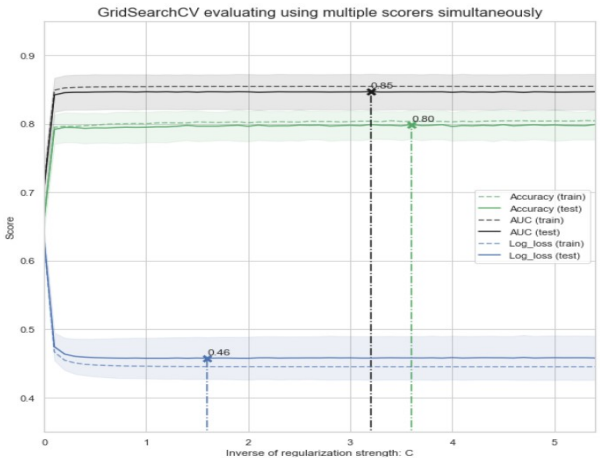

Fig. 14. K fold Grid Search CV using multiple scorers

In this model, applying Grid Search brought the model accuracy up from 77% to 80%.

The next step is to apply this model on the *Test* dataset to predict the survival rates.

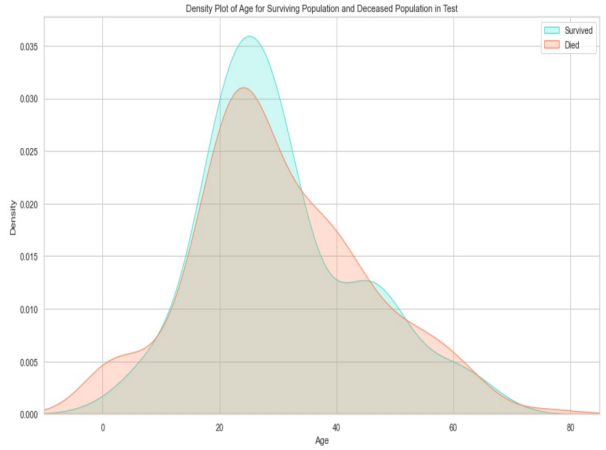

Fig. 15. Distribution plot for Age vs Survived and Deceased in Test Dataset

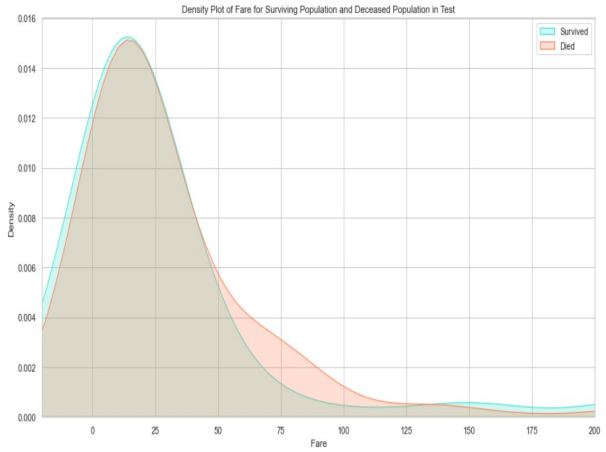

Fig. 16. Distribution plot for Fare cost vs Survived and Deceased in Test dataset

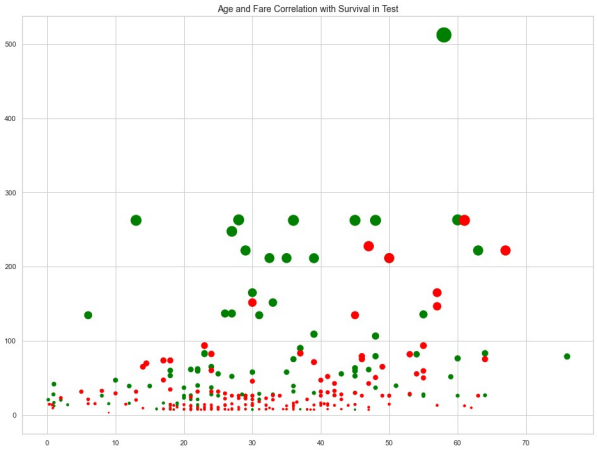

Fig. 17. Correlation between Age and Fare Cost in Test Dataset

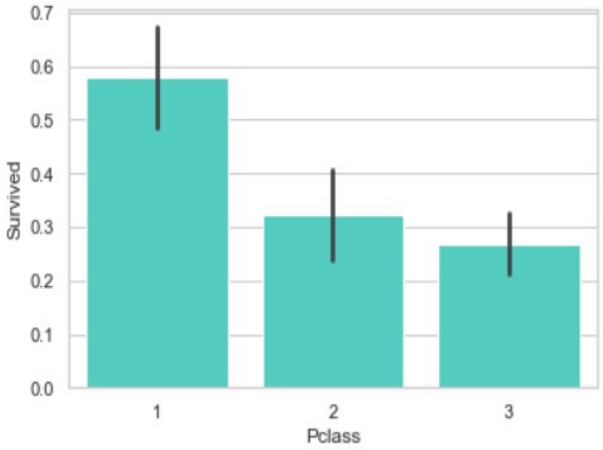

Fig. 18. Bar Plot of Passenger Class vs survival rate in Test Dataset

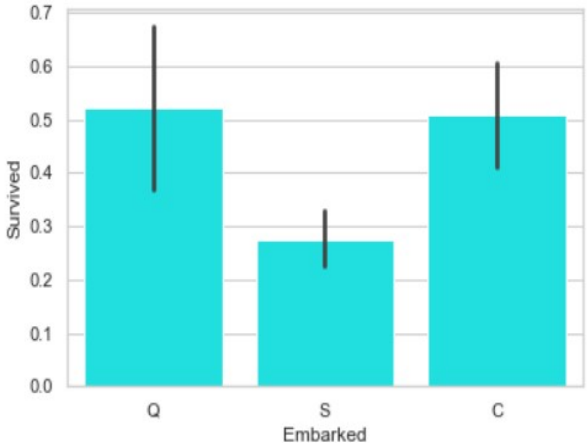

Fig. 19. Bar Plot of Embarked Port vs Survival Rate in Test Dataset

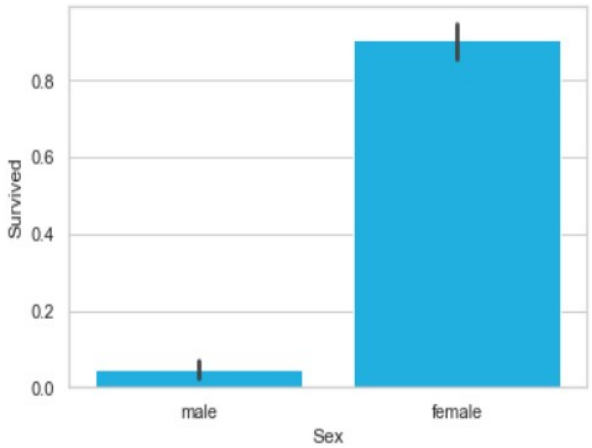

Fig. 20. Bar Plot of Gender vs Survival rate in Test Dataset

The above graphs, from Fig 15. to Fig 20. show a similar trend to the earlier graphs of the *Train* dataset.

Combining the *Train* and *Test* datasets to apply the same techniques reveals similar characteristics to the *Train* dataset as shown below.

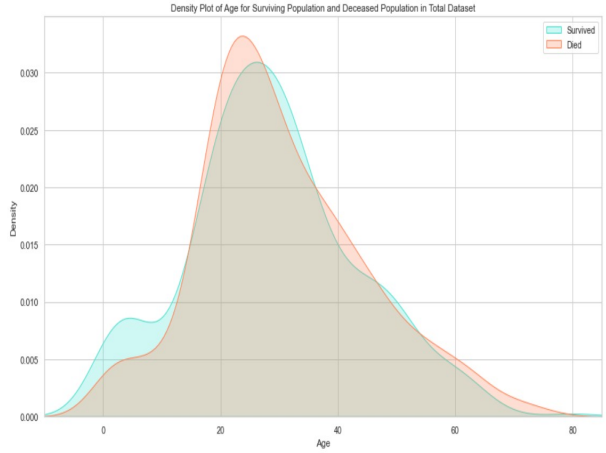

Fig. 21. Distribution plot of Age vs Survived and Deceased for the Total Dataset

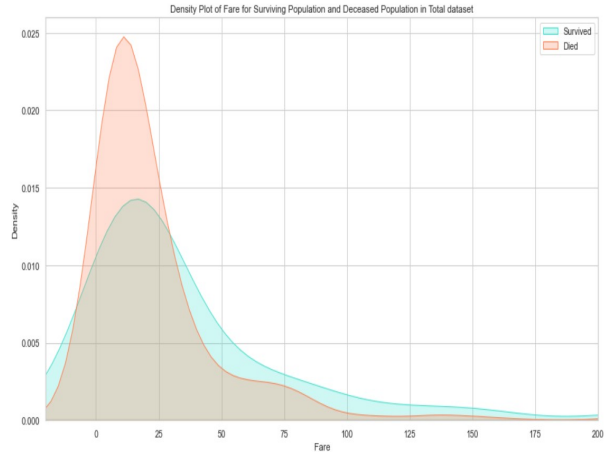

Fig. 22. Distribution plot of Fare Cost vs Survived and Deceased for the Total Dataset

Fig. 23. Correlation between Age and Fare Cost for the Total Dataset

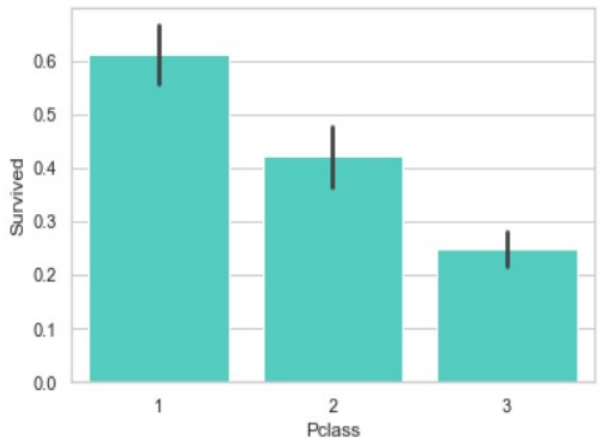

Fig. 24. Bar Plot for Passenger Class vs Survival Rate

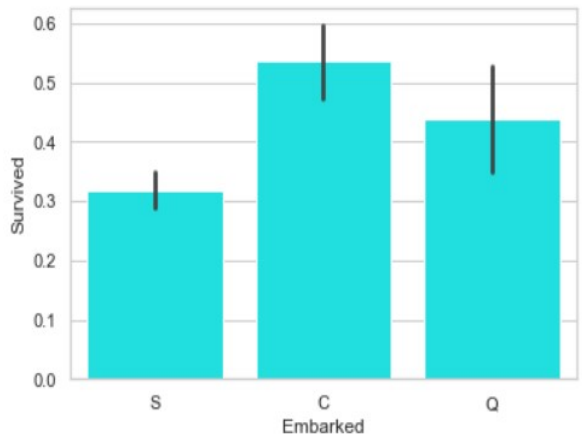

Fig. 25. Bar Plot for Embarked Port vs Survival Rate

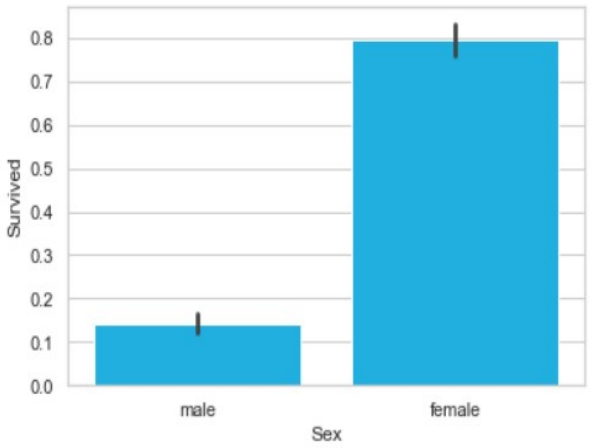

Fig. 26. Bar plot for Gender vs Survival Rate

VI. Inferences

Saving women and children was given a higher importance compared to saving men. Middle aged men had the lowest survival rate. Passengers who paid more for their tickets and higher class passengers were more likely to survive than lower class passengers. A slight correlation is found between passengers who boarded in Cherbourg, France, probably because of the order of room assignments.

Conclusion

Logistic Regression is a widely used machine learning technique especially in binary classification problems. In this paper Logistic Regression is used to understand what types of passengers were the likeliest to survive the Titanic sinking with respect to Age, Gender, Passenger class, Fare cost and other features.

References

[1] ”Logistic Regression” Wikipedia [Accessed on 3 October] Available: https://en.wikipedia.org/wiki/Logistic_regression
[2] Hyperparameter Tuning with GridSearchCV Available: https://www.mygreatlearning.com/blog/gridsearchcv/
[3] Math Behind Logistic Regression Algorithm Available: https://medium.com/analytics-vidhya/logistic-regression-b35d2801a29c
[4] Logistic Regression for Machine Learning Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/
[5] Recursive Feature Elimination (RFE) for Feature Selection in Python Available: https://machinelearningmastery.com/rfe-feature-selection-in-python/
[6] Logistic Regression Scikit-Learn Available: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression