Predicting the survivals in the Titanic

Logistics Regression

Introduction

The main objective of this research is to come up with a predictive model which predicts whether a passenger who bordered Titanic ship that sank in April 15th 1912 in the North Atlantic Ocean will survive or die given particular parameters. It had only been four days since the ship had set sail from Southampton to New York. To achieve the objectives of this study, data was downloaded from Kaggle website. The main reason for predicting the survival chances is to learn to build a classification model with the highest accuracy of prediction. Having a good understanding of developing logistic regression with high accuracy, makes classification wide range of problems in real life becomes easy to solve, like in the Health sector, business industry, etc. For example logistic regression can be used in the health sector with the aid of Artificial intelligence to classify if the tumor is malignant or benign. In the business industry it can be used to predict if the customer is likely to come back to shop again or not. Also it can be used in the financial institutions to tell if the transaction is fraudulent or genuine. Logistic regression has a wide range of applications across different domains of professionalism. Through this study study we will predict the chances of a particular individual surviving given their details like Age, gender, Passenger class, Fare paid, Family size. This helps in predicting the possible outcomes of an event i.e., either if it will occur or not.

Methodology

The Titanic data used in this study is made up of 14 variables with 891 observations. The variables include:

PassengerId: Passengers Identifier.

Survived: which is a categorical variable either the passenger survived or died coded as 0 = died and 1 = survived, Pclass which is the passengers class with 1st, 2nd and 3rd which is Upper, Middle, lower respectively. Pclass is a ctegorical variable.

Name: which is a name of the passenger and it is a metadata.

Sex: which is categorical with male and female:

Age: Age in years(Continuous variable)

Sibsp: which is number of sibling or spouse aboard the ship.(Categorical)

Parch: which is the number of parents or children aboard the ship.(Categorical Variable)

Ticket: which is the ticket number for the passenger.(metadata)

Fare: which is the passengers fare (continuous variable)

Cabin: which is the Cabin Number.

Embarked: which is the port of embarkation; C for Cherbourg, Q for Queenstown and S for Southampton. It is categorical.

Frequencies and descriptive statistics such as the mean and the standard deviation with the aid of diagrams were employed in exploring the data. diagrams such as staked bar charts and histograms were used to visualize the interaction of the variables. Missing values in the data were removed form the data only to remain with the observations with complete response.

The statistical test carried out is binomial logistic regression which is used for classification of two possible outcomes of a categorical variable. In this case we are classfying the passengers as either survived or died.

Results

Summary Statistics

The data has 891 observations with 12 variables, namely **Survived**, **Pclass, Sex, Age, Sibsp, Parch, Fare, Embarked.** ,**Name, passengerId, Ticket,** and **Cabin.**.

The Summary statistics above shows that there are 177 missing Values in the column of the variable age. This missing values were removed by dropping them to remain with 714 complete observations.

selecting Variables of interest

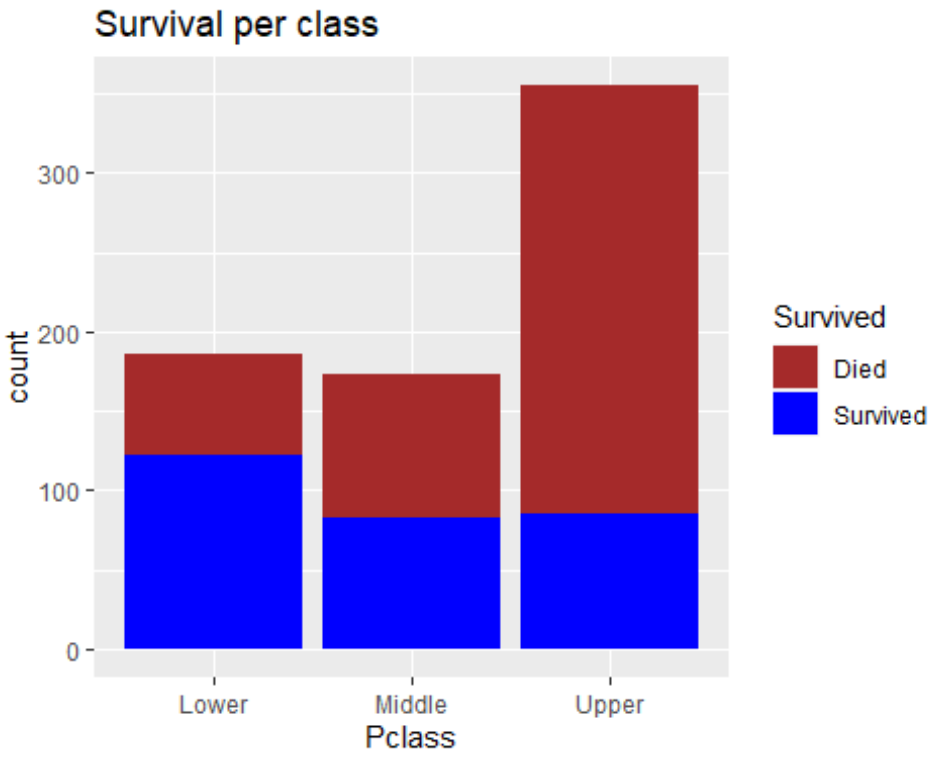
The variables of interest in this data are **Survived**, **Pclass, Sex, Age, Sibsp, Parch, Fare,** and **Embarked.** Variables that were dropped are **Name, passengerId, Ticket,** and **Cabin.**

All categorical variables were converted to be categorical by assigning ordinal groups values in order and nominal groups randomly.

Bar chart showing survval per class

The stacked bar chart below shows that 65.59% of the passengers who boardered in lower class survivied while 34.41% died. In the middle class 52.02% of the passengers died while 41.98% survived. Upper class was the most afected as 76.01% died while only 23.99% survived.

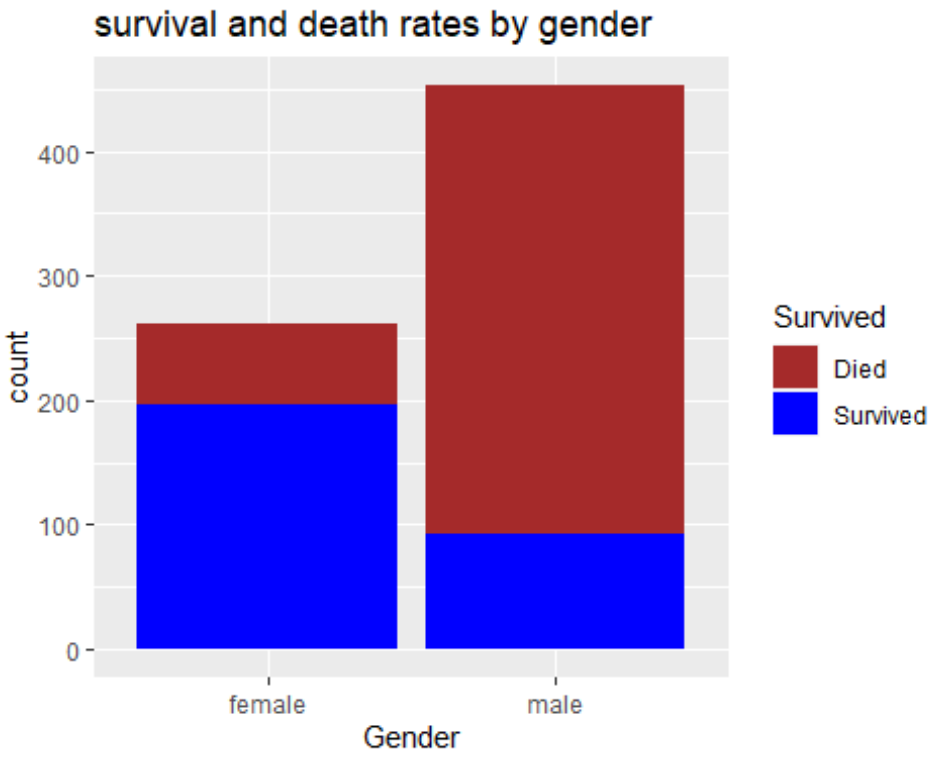
	Died	Survived
Lower	64	122
Middle	90	83
Upper	270	85



Bar chart Survival and death rates by gender

The graph below shows the survival and death rates by gender. Male were the most affected as 79.47% of those who had bordered perished in the titanic incident while 20.53% survived. 75.48% of the females survived while24.52% died.

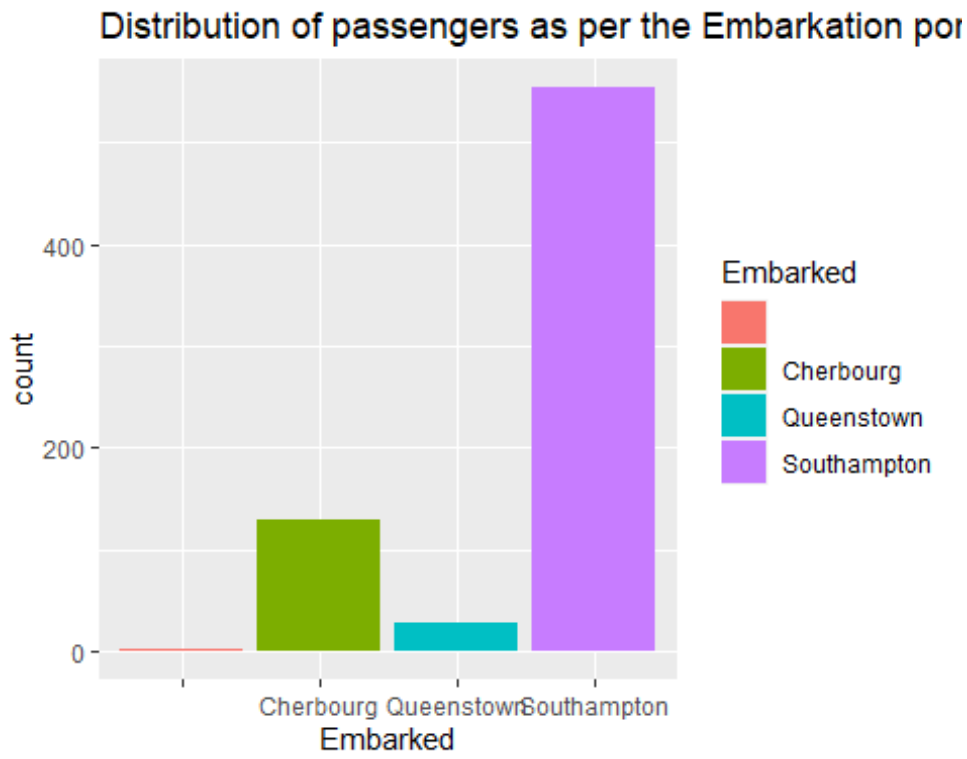
	female	male
Died	64	360
Survived	197	93



Bar chart for Embarkation

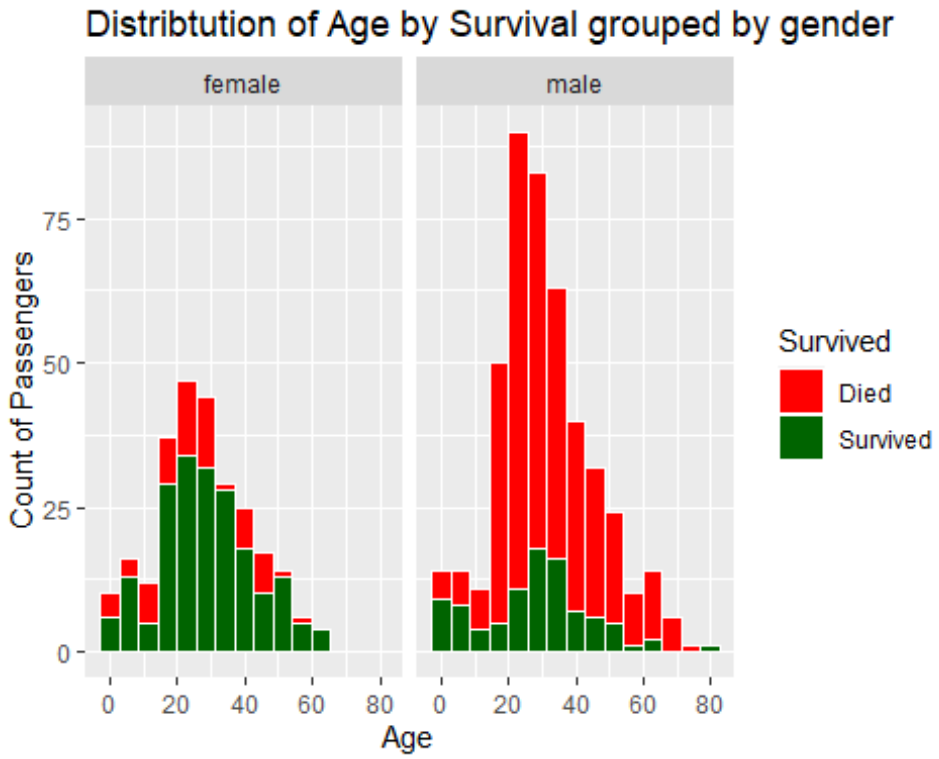
The majority of the passenger were from southampton making up 77.59%, followed by Cherbourg with 18.21% then Queenstown with 3.9% and the remaining with 0.28% were not traced.

	Died	Survived
	0	2
Cherbourg	51	79
Queenstown	20	8
Southampton	353	201



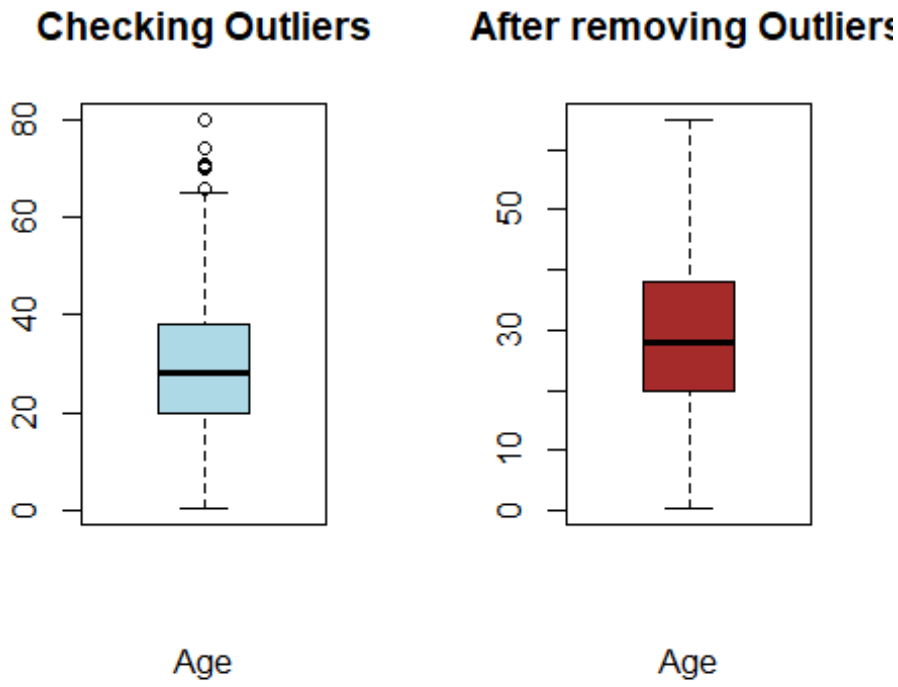
Histogram

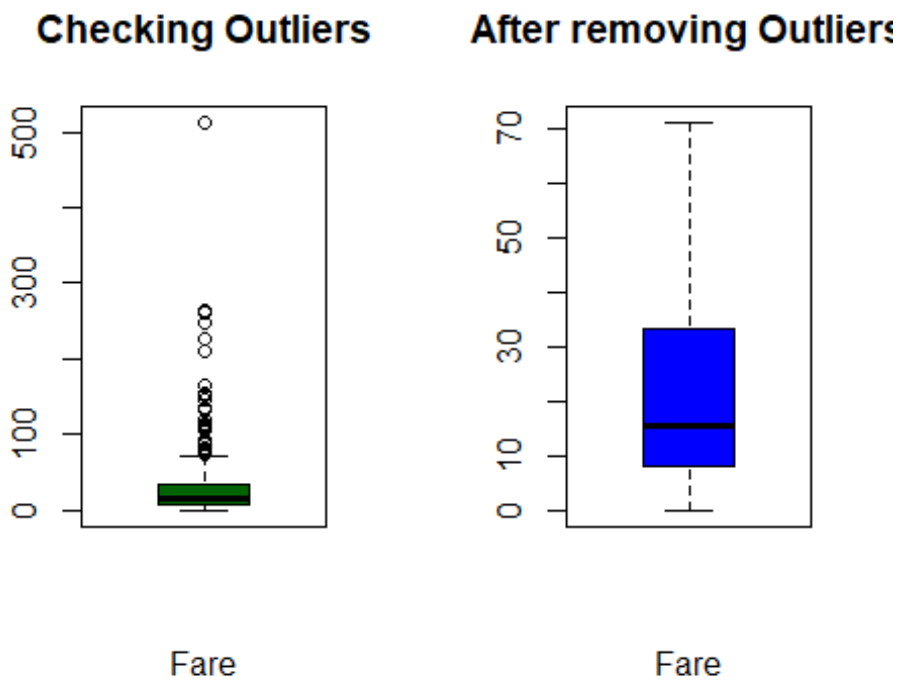
The Distribution of survival in the female histogram is heavily occupied those who survived while those who perished are few which is converse to that of male. The distribution of age for both groups are evenly distributed.



Logistic Regression Model For predicting Survival rates in titanic

Removing Outliers using Box-plot methods in Age and Fare





The variance inflation factors below shows that there is no variable to drop as they are all less than 5, which is the threshold.

```
## [1] 2 1 1 1 1 1 1 1
```

The model below shows the statistically significant coefficients for predicting survival;

```
## glm(formula = Survived ~ Pclass + Age + SibSp + female + embarked_c,
##     family = "binomial", data = train)

##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  1.64146268 0.48933381  3.354485 7.951300e-04
## Pclass2     -1.36151974 0.34548999 -3.940837 8.119779e-05
## Pclass3     -2.81721437 0.35139939 -8.017129 1.082451e-15
## Age         -0.04684709 0.00963842 -4.860453 1.171175e-06
## SibSp       -0.36253867 0.13636802 -2.658531 7.848203e-03
## female       2.81122880 0.25299190 11.111932 1.097587e-28
## embarked_c   0.70420624 0.31331993  2.247563 2.460408e-02
```

Confusion Matrix and the model accuracy on test data

Confusion matrix for the train data is as shown below:

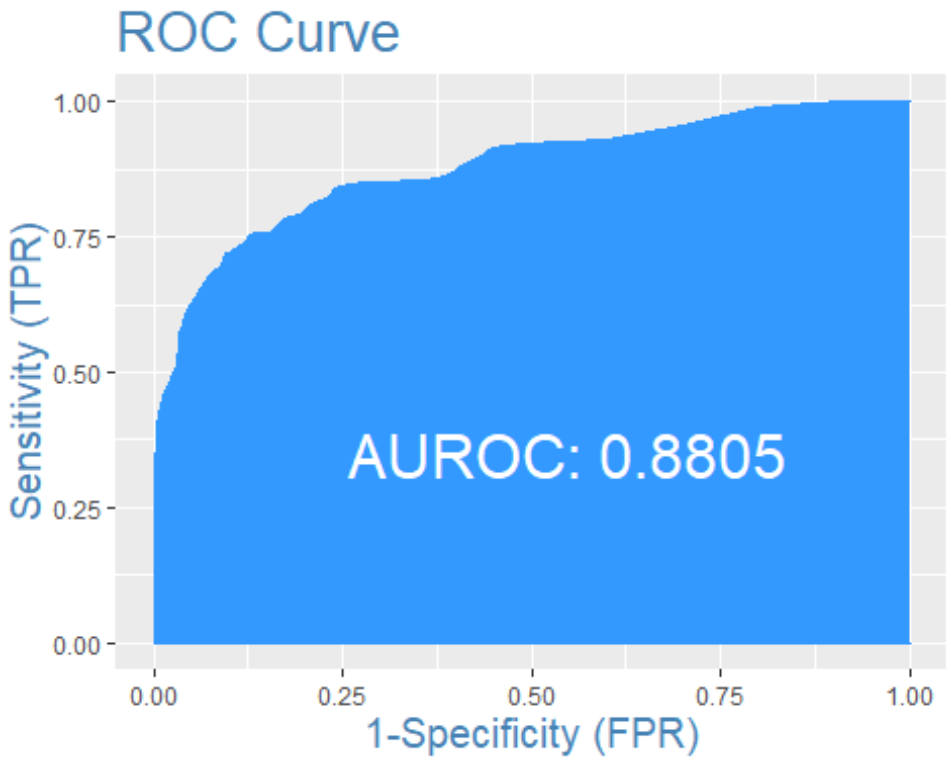
	Died	Survived
Died	305	71
Survived	30	165

The accuracy of the model is obtained from the confusion matrix by applying the formula below:

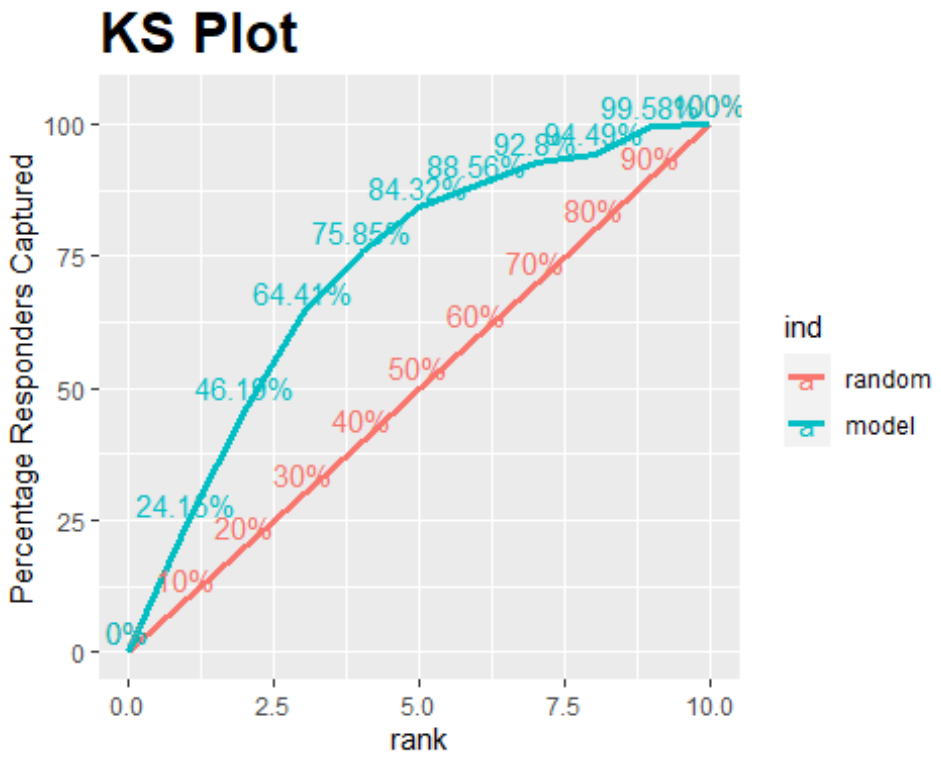
$$ModelAccuracy = \frac{TP + TP}{TP + TN + FP + FN} = \frac{305 + 165}{305 + 165 + 71 + 30} = 0.8031$$

The accuracy of the model on train data is 82.3117338 which is fit for predicting the survival in the Titanic incidence.

Below is an ROC curve showing an **AUROC** of 0.8805, which mean that the model quality is good for classification.



The KS plot below is showing the maximum difference between TP and FP;



Confusion Matrix and the model accuracy on test data

The confusion matrix for the test data is as shown below:

	Died	Survived
Died	79	23
Survived	10	31

on applying the formula below:

$$ModelAccuracy = \frac{TP + TP}{TP + TN + FP + FN} = \frac{79 + 31}{79 + 31 + 23 + 10} = 0.7692$$

we obtain the score of the model on test data as 76.92% which is good.

Conclusion

Through this study the objective of the research has been achieved as the model we did develop the model to predict the survival in the Titanic. The accuracy of the model on the training data was good as it was 80.31% as well as on the test data 76.92%. This high accuracy makes the model to be reliable in making prediction. The implication of the concept developed in this study can be applied to different domains of knowledge. The weakness of the research is that the data had a lot of missing observations which might affect model accuracy. Another weakness is that there is no future data to be used to test the model since the titanic incidence was just a one time accident.