UECS16322 I Semester V I 2018 I R

Data Analytics Project

Abhishek S 01FB16ECS018
Animesh N D 01FB16ECS058

Survival on the Titanic!

Predicting The Fate Of The Voyagers

Introduction

From statistical data obtained online (kaggel), we perform data analytics to come up with a reliable and precise model to predict the survival of a given passenger onboard the infamous RMS Titanic using R.

The Data

In this section we look into the data available and the different variables involved.

We are provided with two sets of data, a training data set (train.csv) and a testing data set (test.csv).

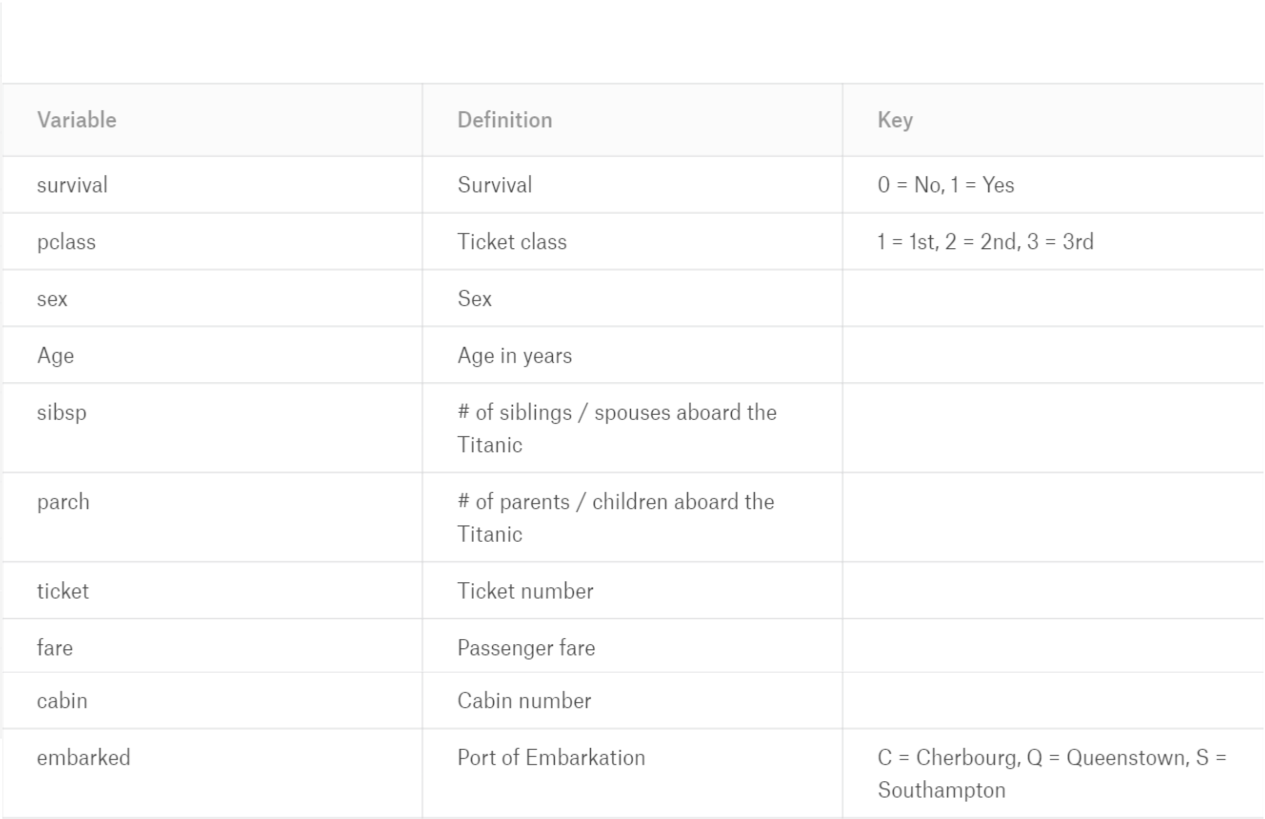
Given below is the data dictionary, where

pclass: A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

Data Dictionary



Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

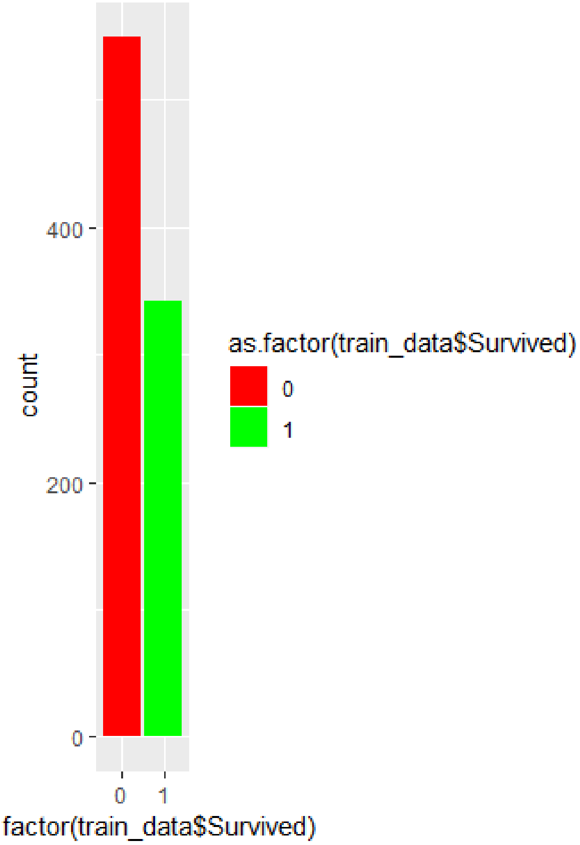
Now that we are familiar with the dataset provided, we start our analysis using R.

Cleaning the Data

We see that the data isn't consistent and does contain a lot of missing values, we shall tackle that first before moving onto building our predictive model.

Also we see how our initial training data looks, and it is noticed that the data is more biased towards non-survivors as indicated by the plot given alongside.

In addition to this, we take the columns SibSp and Parch, and replace it with a new Variable – 'FamilySize', as this is intuitively bound to be a more important criterion than those two variables separately, as we shall see later on.

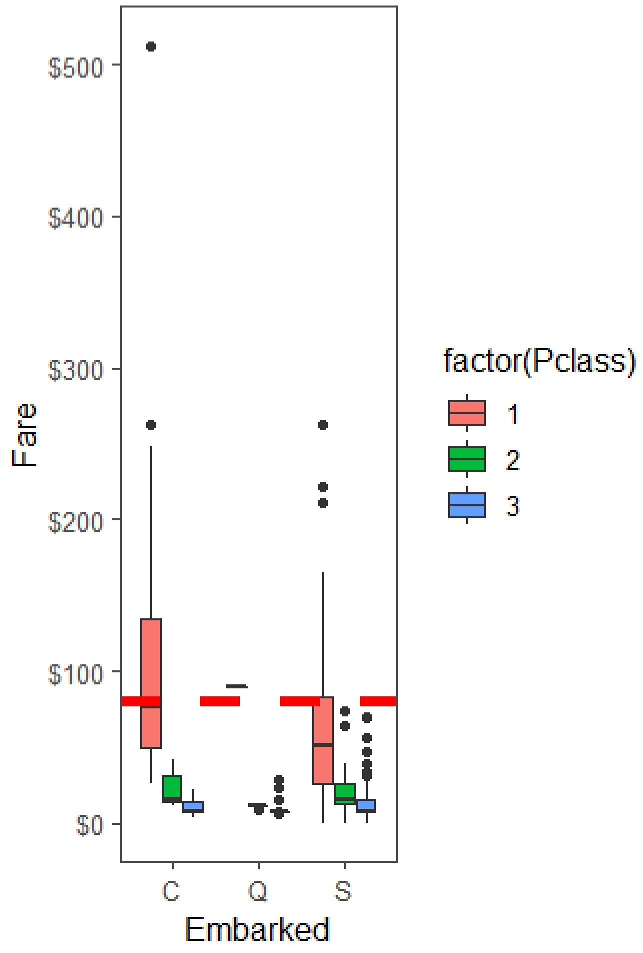


Imputing missing values

The columns with the missing values are : Embarked(2), Fare(1), Age(263)

Embarked:

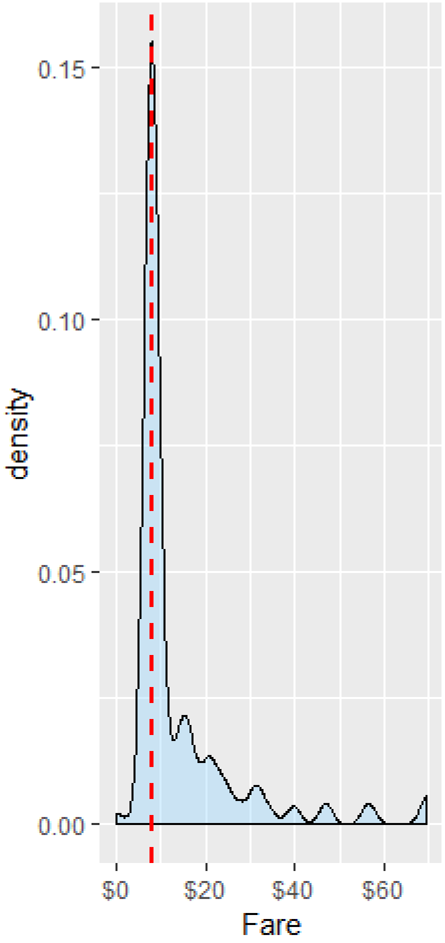
We calculate the distribution of the fare paid among the three different Embark points and try to find which median value is closest to $80 which is what they(the voyagers with missing embark values) paid. Both the Passengers also bought Class1 tickets.



The RED dotted line indicates the $80 cutoff. It is very evident from the scatter plots that the median of the fare paid in Embark point C is the closest to $80. Hence, we make a safe assumption that both these passengers boarded at point C.

Fare:

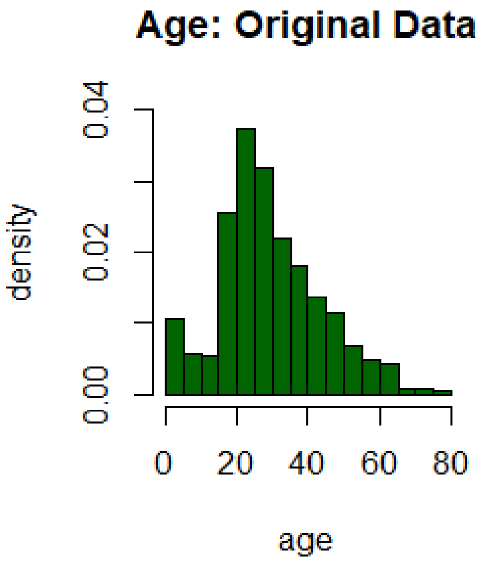
The fare value for row 1044 is NA. One way to fill this missing value is to replace it with the median value. (Since median is a better representation of the data than mean)

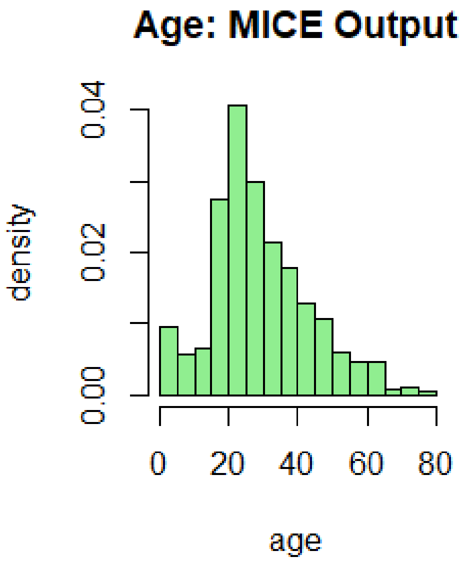


We plot the distribution of Fares of passengers who embarked at point S and bought a 3rd Class ticket. The density of the median value is high, hence it's a safe assumption to replace the value with the median.

Age:

There are 263 missing age values. This can drastically effect our model. All the previous methods to impute won't suffice here. Here we use predictive analysis to fill in the missing values. For this purpose we use the R function 'mice' , which stands for - Multivariate Imputation via Chained Equations. Mice uses a regression model based on the remaining variables to fill in these missing values.





To make sure we haven't tampered wrongly with the data we check the age distribution before and after imputing with mice.

Clearly there is minimal change.

Predictive Analysis

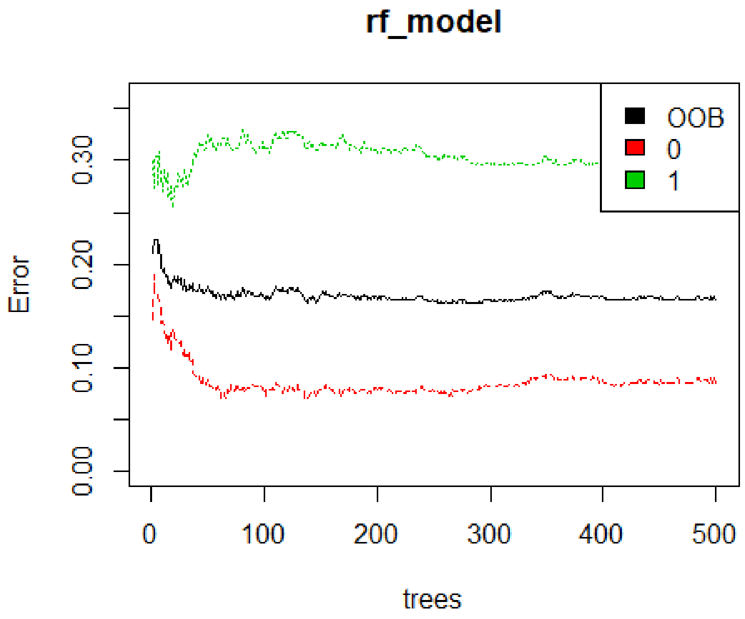
For comparison purposes, we are going to use two predictive models, namely, Random Forest Algorithm and the more common Logistic Regression. Since the end goal is to predict whether a given passenger survives or not, i.e. a binary output, it becomes a categorical problem. Hence, these two models are best suited to tackle such a situation.

Random Forest Model

This model runs on the principle of decision trees. It basically builds multiple decision trees and merges them in order to give a more accurate and stable outcome.

Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node.

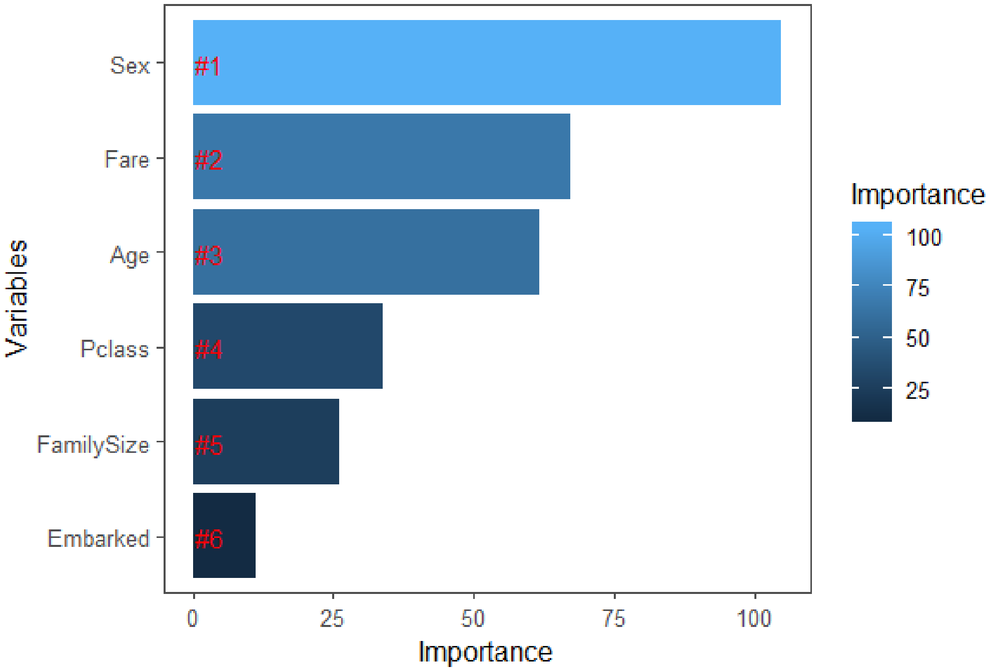
We build the model using the randomFores() function in R, and feed in our training data set, all the while only considering the following as predictor variables :
Pclass, Sex, Age, Fare, Embarked, FamilySize.



After building our model, we check for its consistency by calculating its error rates. The plot given above shows that our error rate for predicting deaths is lower than our survival prediction. And our overall error rate for the model is around 0.2

We also calculate the accuracy of our model with respect to the training data, it comes around to be 0.92031, which isn't bad at all!

Accuracy being - (True positive + True Negatives) / Total Observation



We also look at the importance/Information value for each variable used and it seen that clearly that sex plays a very large role followed by fare and age. The family size, a variable we introduced did have a fair amount of impact.

(This observation is in line with the famous James Cameron film ‘Titanic’ as well where Rose survives but Jack doesn’t! Considering the fact that Rose was a female, rich and hence would’ve paid a higher Fare, was Young , had an Upper class Ticket, Family Size was also large, she would’ve definitely been more favored to survive in contrast to poor Jack who only has his youth favoring him.)

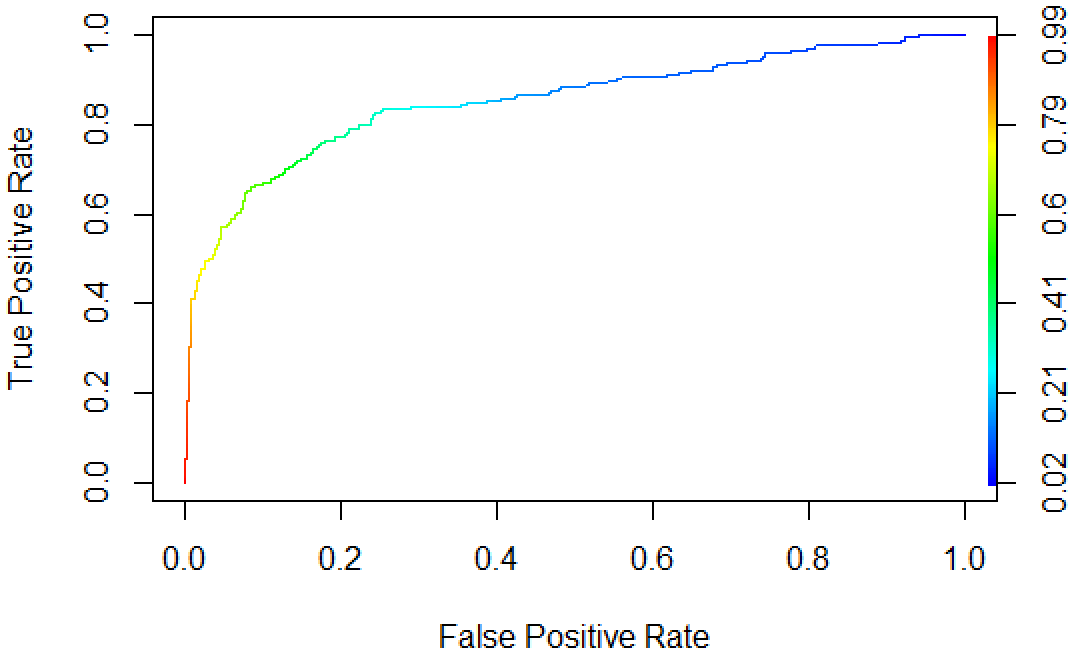
Logistic Regression Model

This is a fairly straightforward model, which uses principles of regression to classify, in this case into categories Survived or Not Survived.

Similarly to the previous method we take into consideration the following predictor variables : Pclass, Age, Sex, Embarked, Family Size

The training data was fed in, and the model is obtained using the glm() function.

The ROC Curve is plotted.
The Accuracy for this model is also calculated from the confusion matrix.
Accuracy comes out to be 0.80359



Conclusion

From Accuracy stand point, the Random Forest Model had a lead over the Logistic Regression Model by quite a margin.

When the models were run on the training data set however, this margin got narrower. The accuracy of the Random Forest Model was 0.77990, while for the Logistic Regression Model was 0.77033. The Random Forest Model, like any other Decision Tree based algorithm would’ve been a prey to Overfitting, hence the drop in accuracy.

Regardless, the Random Forest Model turns out to be slightly better than the Logistic Regression Model.