**Final Project**

[**Analyzing Titanic**](https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic) **Dataset for Survival**

**Submitted By -Radhika Hegde**

**Table of Contents**

[**1.** **Abstract** 3](#_Toc514017457)

[**2.** **Introduction** 4](#_Toc514017458)

[**3.** **Exploring Data** 5](#_Toc514017459)

[**4.** **Data Analysis** 5](#_Toc514017460)

[**5.** **Feature Engineering** 8](#_Toc514017461)

[**6.** **Relationship to Survival Rate** 12](#_Toc514017462)

[**6.1** **Sex vs Survival** 12](#_Toc514017463)

[**6.2** **Age vs Survival Boxplot** 14](#_Toc514017464)

[**7.** **Prediction** 15](#_Toc514017465)

[**7.1** **Using Conditional inference tree (ctree)** 15](#_Toc514017466)

[**7.2** **Using Decision tree** 16](#_Toc514017467)

[**7.3** **RandomForest Model** 17](#_Toc514017468)

[**8.** **Conclusion** 19](#_Toc514017469)

[**9.** **References** 20](#_Toc514017470)

[**Dataset** 20](#_Toc514017471)

[**Predicted data** 20](#_Toc514017472)

[**Source Code** 20](#_Toc514017473)

# **Abstract**

The sinking of the RMS titanic is one of the most infamous shipwrecks in history. On April 15, 1912 during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the entire world and the international community.

A policy of “women and child first” on the Titanic was clearly demonstrated but appears that having a large family was also not good for chances of survival. My research question would be ***“Who are the people most likely to survive in such an accident, where the survival was just a matter of luck and what sorts of people were likely to survive?”***

To build the model I will make Feature Engineering that could feed predictions or create additional variables. Where there are 891 observations (rows) in the dataset with 12 variables each. I have also thought of using Decision trees where that data inherently supports to the model. But going forward I will also use Random Forest model which will give more accuracy when compared to decision tree on the dataset and finally predict which sort of people were more likely to survive.

Firstly, I will try to figure out and take a look at the Sex and Age variables to see if any patterns are evident. I am still in the process of EDA (exploratory data analysis) which will help me explore the formal modeling and hypothesis testing task. I may also try using feature Engineering (Optional). This process attempts to create additional relevant features from the existing raw features in the data, and to increase the predictive power of the learning algorithm.

# **Introduction**

The sinking of the RMS titanic is one of the most infamous shipwrecks in history. On April 15, 1912 the largest passenger liner ever made during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the entire world and the international community and led to better safety regulations for ships. One of the reasons that the shipwreck resulted in such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others. The purpose of this analysis is to explore factors that influenced a person’s likelihood to survive. While there could hardly be a more chaotic event than frightened people scrambling to escape a sinking ship, the disaster is famous for saving “women and children first”.

The project is all about which sort of people traveling mostly survived and were likely to survive and we’ll try to predict who they were. Considering age, class, gender, family and some other important features of the passengers, the analysis is done. Most of the project is data exploration and basic model building. With this project, we will try to analyze which sort of people were given most importance in the process of evacuation after the wreck.

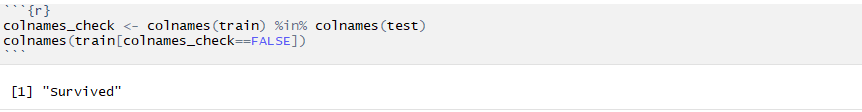
**Research Question:**

Who were the people most likely to survive in such an accident, where the survival was just a matter of luck and what sorts of people were likely to survive?

# **Exploring Data**

Initially the data is loaded into the R-Studio and then the packages that are required which are then imported for the project. Once the packages are imported, data is loaded and then divided into test and train datasets where we use train data for training the model and the test data to check the accuracy of the model built.

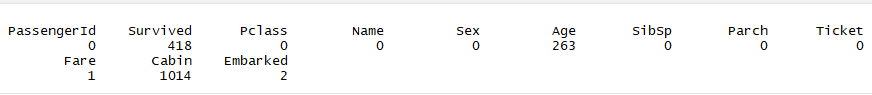
The training set has 891 observations and 12 variables and the testing set has 418 observations and 11 variables. The training set has 1 extra variable. We could see that in a very small dataset like this, but if its larger we want two compare them. As we can see we are missing the Survived in the test set.



# **Data Analysis**

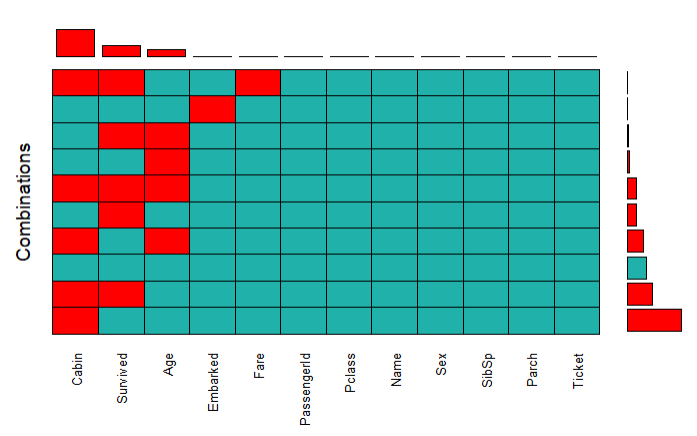
Analysis will be performed by considering age, class, gender, family and some other important features of the passengers. This document is more about training and evaluating the model. To do that, a dataset that has the complete passenger information is considered i.e. the train dataset. I will also try feature engineering where we will consider passenger name because we can break it down into additional meaningful variables which can feed predictions or be used in the creation of additional new variables.

Before feature engineering let’s look into missing values. For studying the complete data set lets join test and train data set. The below Fig1 is the combine data of train and test datasets and named as full. Here test data had 418 obs. and it didn’t have survival data. Whereas there are lot of other missing values like Cabin and Age variable too.



**Fig1: Missing value count per variables used in full**

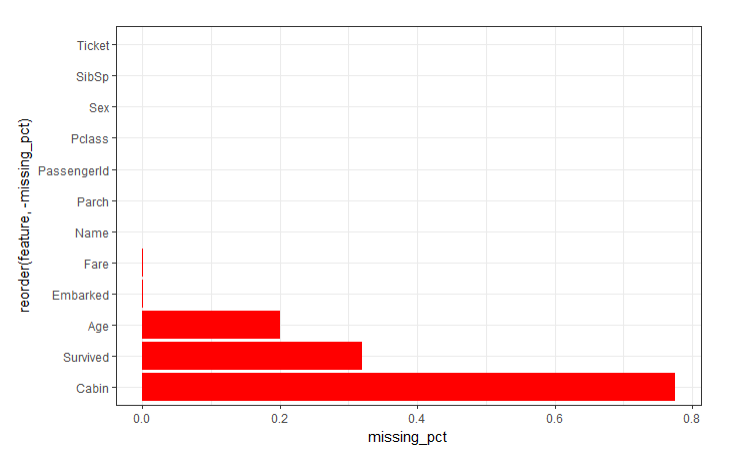
Below Fig2 displays all existing combinations of missing (red) and observed (lightseagreen) values. For example, this plot reveals that if variable Cabin are missing, they are also missing Age and Fare variables. However, when Embarked variable is missing, Cabin, Age and Fare variables have data. We are not concerned with Survival’s missing observations since they are mostly from test dataset.



**Fig2: Aggr function used to find the missing values**

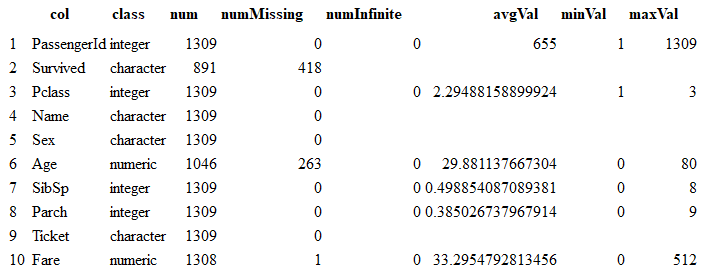
The data can be missing for many reasons. Generally, there are 3 steps required in dealing with missing data:

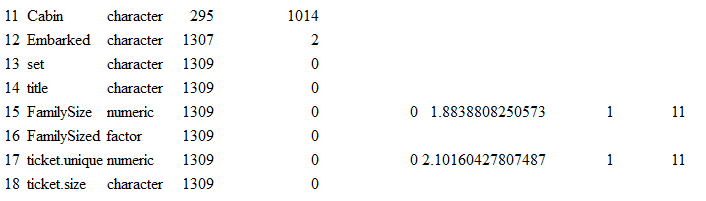
* Find the missing data
* Investigate the cause of missing data &
* Replace the missing values with sensible data or delete the missing data rows.



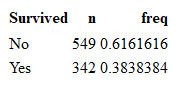
**Fig3: Missing values in percentage**

The below table shows useful data quality function for missing values





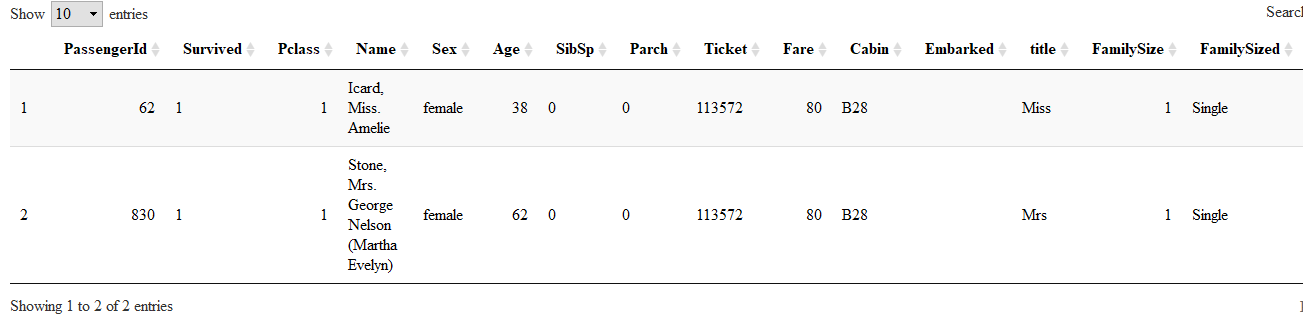
The independent variable, Survived, is labeled as a Bernoulli trial where a passenger or crew member surviving is encoded with the value of 1. Among observations in the train set, approximately 38% of passengers and crew survived. The below figure shows for train dataset (891 rows)

****

# **Feature Engineering**

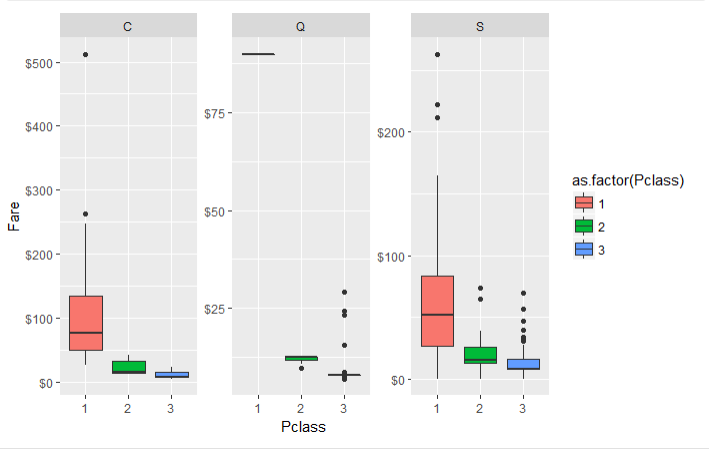
Missing data can’t be used in modeling, it has to be patched up with data. So now we can focus on preparing the data so that it can be used for study, such as exploratory data analysis and modeling fitting. Firstly, lets us look at missing values of Embarked variables. We know from our missing data plot [Missing values](http://rstudio-pubs-static.s3.amazonaws.com/241333_e3e176ecc7a44094a6bba75899a88f72.html#anchor) that there are no other variables missing when combined with Embarked variables.

The below Fig4 that there are two observations with missing data from Embarked variable. We also know that there are three Port of Embarkation -> C = Cherbourg; Q = Queenstown; S = Southampton. The table above shows that both tickets were from Pclass 1; Fare $80; Ticket no 113572 & Cabin B28.

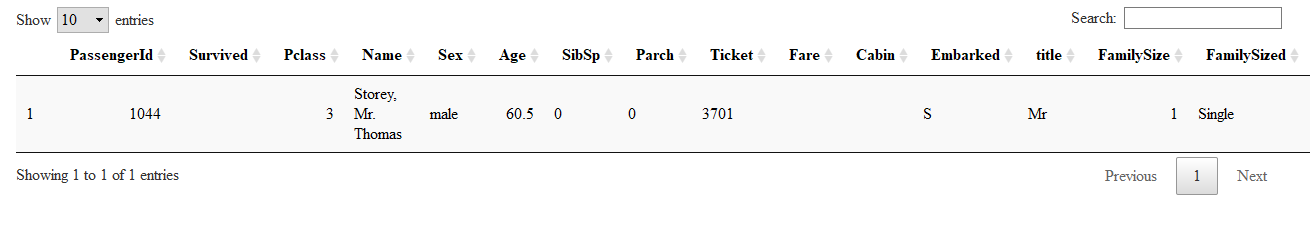


**Fig4: Missing Embarked**

We can see from the box plot that Pclass 1 median fare is approx $80 when embarked from C which is Cherbourg. Hence now I think it is sensible to assign the missing embarked with “C” for those two tickets whose Embarked values were missing.

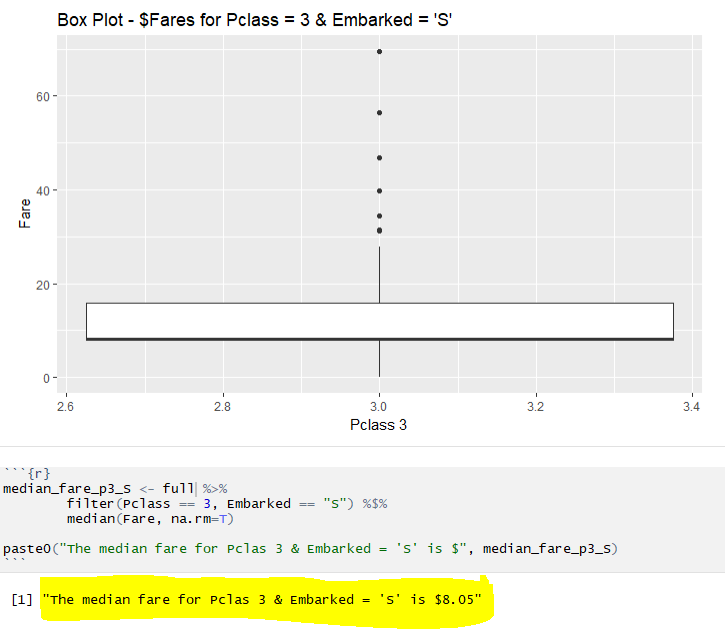


Now the PassengerId 1044 has Fare value missing as shown below in Fig5



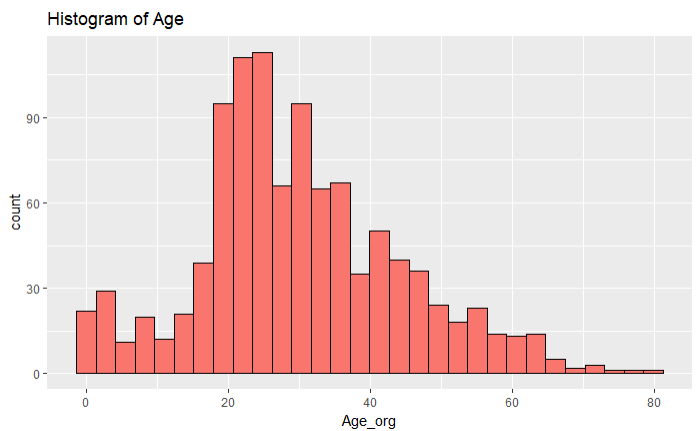
**Fig5: Missing Fare**

The density plot in below Fig6 shows the counts and the area underneath the polygon is one. This plot tells us approx. $8.05 accounts to around 15 percent of the fare in Pclass 3 and Embarked from S. I think it is sensible to replace NA with $8.05

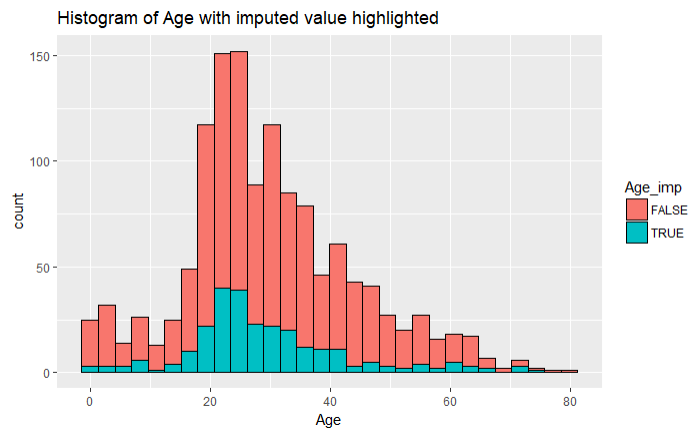


**Fig6: median fare for Pclass 3 and Embarked S**

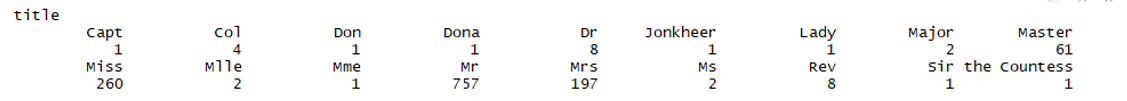
We’ll use MICE imputation method to impute missing Age variables. We’ll use Pclass, SibSp, Parch, Fare, Embarked, Title, FamilySize variables as predictors to impute missing values in Age.



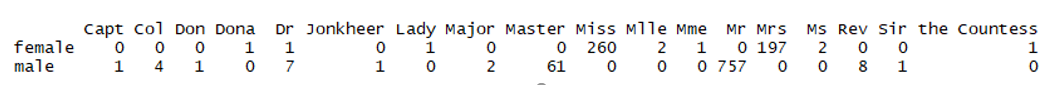
Now fill the missing values for age. The below shape of histogram has not changed significantly from original values and after imputing missing values. I think its sensible to use the imputed values.



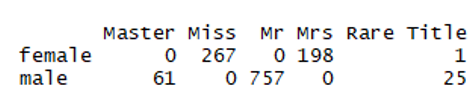
Titles also play a very import role. Sex variable doesn’t tell us whether a person is married or single but the name title does. Title helps us to predict if they are in family or not. Let’s extract title from names. So here we now have a nice new discrete column of titles as shown below



If sorted with respect to sex it looks as below



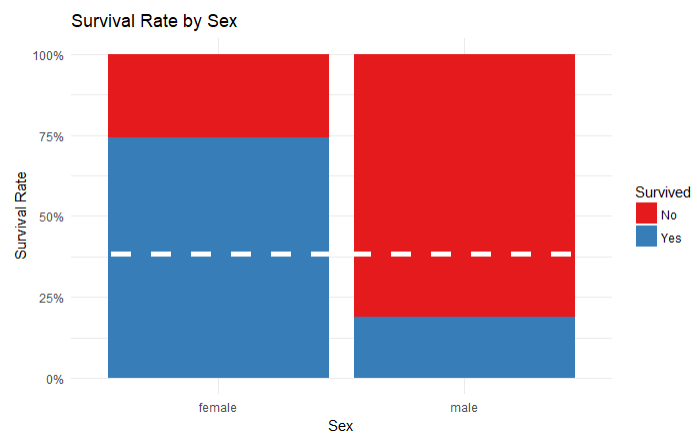
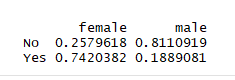
Here titles with very low cell counts can be combined to "Rare Title" level and I have also reassigned mlle, Ms, countess and Mme accordingly. The below title count shows by sex again



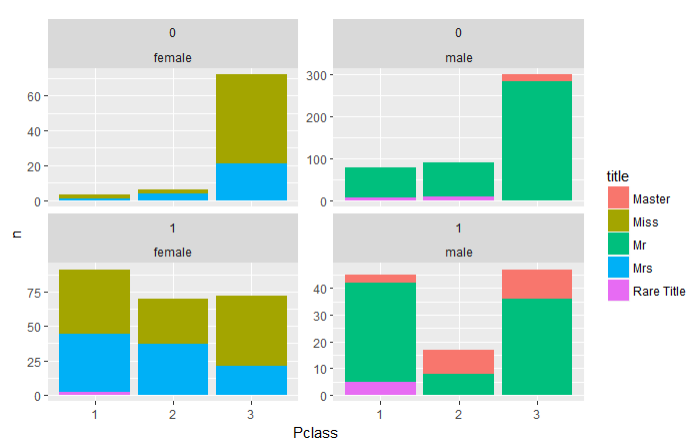
# **Relationship to Survival Rate**

## **Sex vs Survival**

74 % of all the females in train data survived whereas only ~18% out of all male population survived. That is huge differences!!!

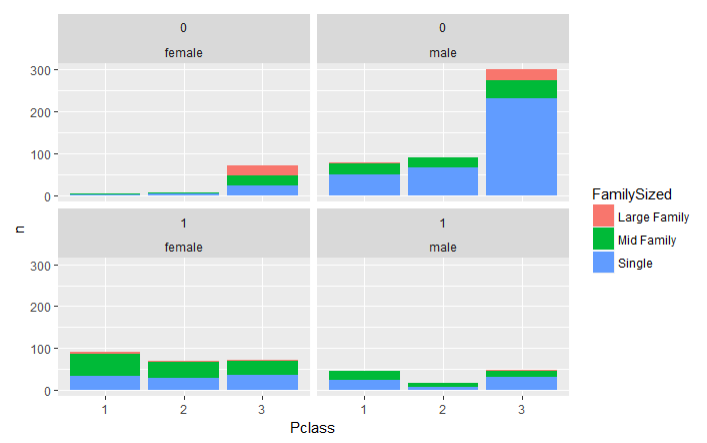
 

The below plot clearly shows that there are lot more female survivors then men.



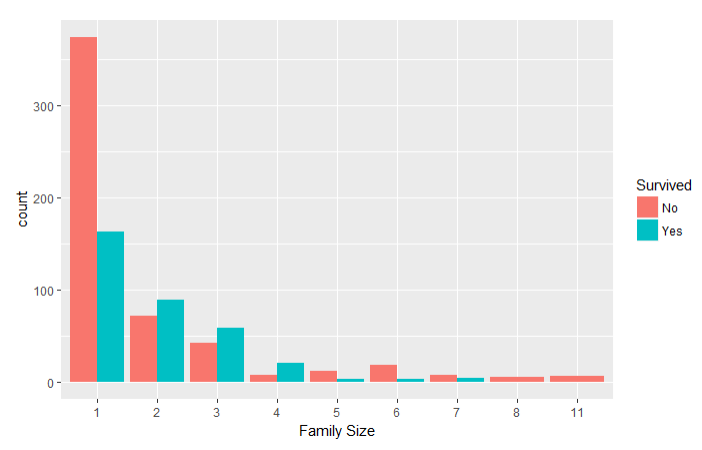
**Fig: Relation: Title | Pclass | Survival**

In all 3 classes female survivors outnumber male survivors. If you were female you had more chances of survival in all Pclass. You’ll have a good chance of survival if you were single or in small size family.

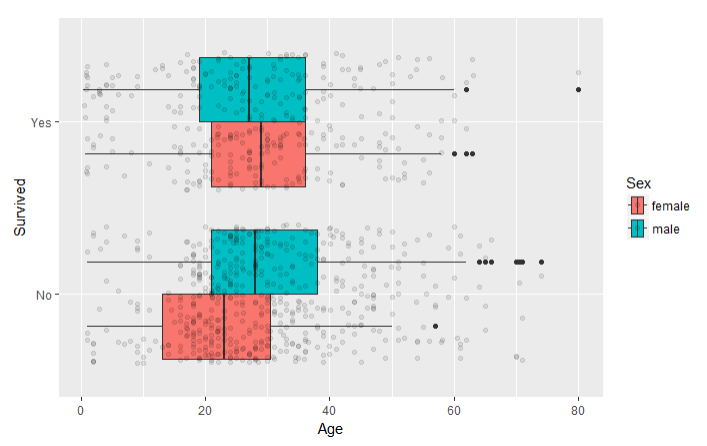


**Fig: Relation: Family Sized | Pclass | Survival**

And this confirm visually that small family have survived better than large family

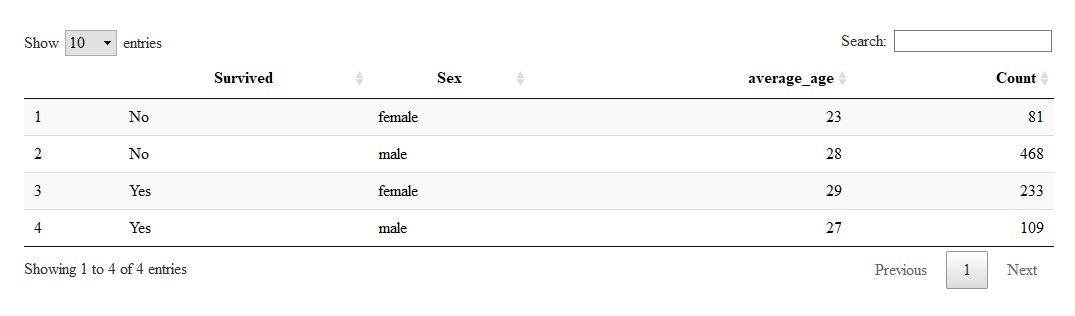


## **Age vs Survival Boxplot**



**Fig: BoxPlot for Age Vs Survival**

The below data table (train data) shows that those who did not survive, on average females were much younger than males.

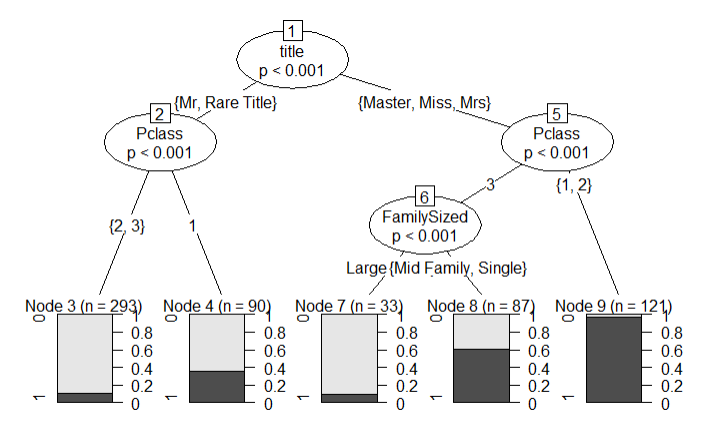


# **Prediction**

## **Using Conditional inference tree (ctree)**

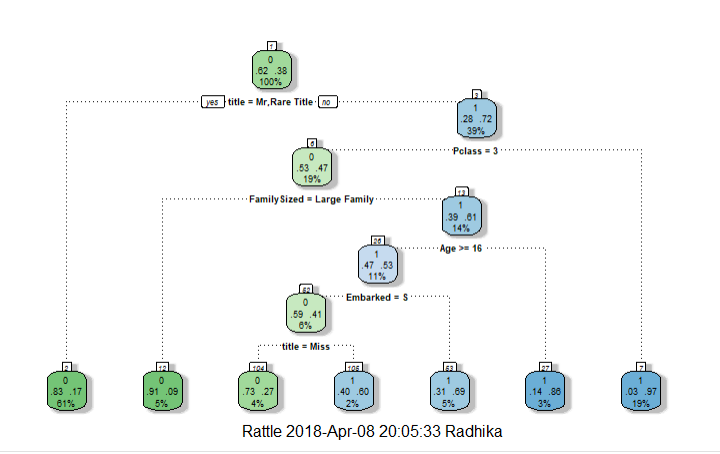
I feel conditional inference trees can be used as base learner for Random Forests. We’ll split the training set (891 rows)- train into trainset & testset using 70/30 ratio. And will build the ctree (conditional inference tree) model using:

**Survived ~ Pclass + Sex + SibSp + Parch + Fare + Embarked + Title + FamilySize + Age + FamilySized**



The above tree diagram has confirmed what we already knew from our data exploration. The tree diagram claims that the most important variable is Title and if you’re female and travel in Pclass 1 & 2 you have much chances of survival. If you are male the worst-case scenario would be to travel in Pclass 2 & 3. Since we had split the train data set into trainset and testset we know the survival status. Going forward we can use this information to check our model prediction power.

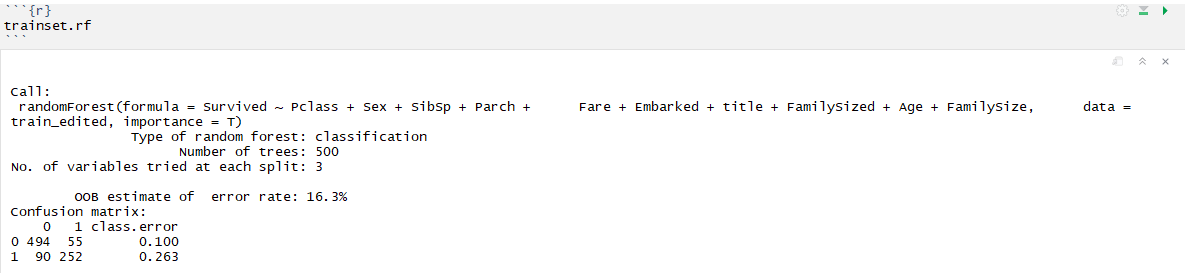
## **Using Decision tree**

The decisions that have been found go a lot deeper. Decisions have been found for the port of embarkation one that I didn’t even look at. The decision is made on the current node which appear to be the best at the time but are unable to change their minds as they grow new nodes.  

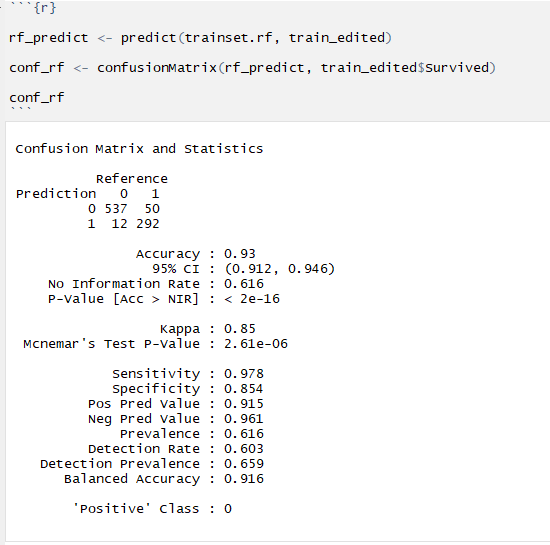
## **RandomForest Model**

Now will try using Randomforest. Randomforest is an ensemble learning method that grows multiple decision trees and each decision tree will produce its own prediction results. Will use the same variable to build the random model





Out of bag (OOB) error is the overall classification error in the above model and as we can see, OOB error still is more than 15%. This means that we still can reduce the OOB error in the model. The below Confusion matrix is created based on the train data setusing RandomForest - the overall confusion matrix accuracy is 93%



The below fig shows mean square error rate. The green line shows the error rate for survival whereas red lines shows for the dead. Plot is telling us that we are more successful predicting death than survival.

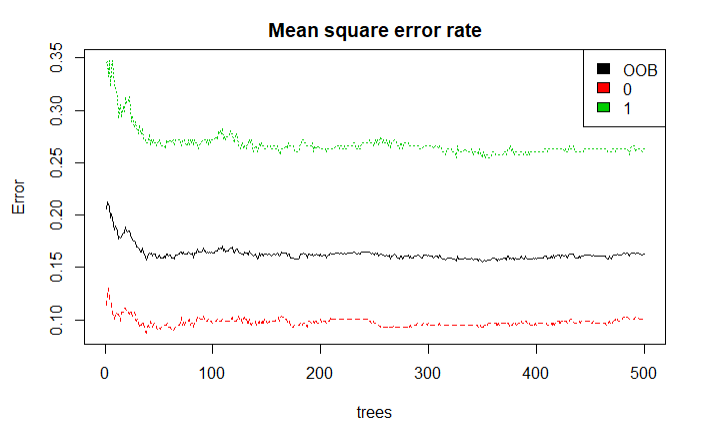
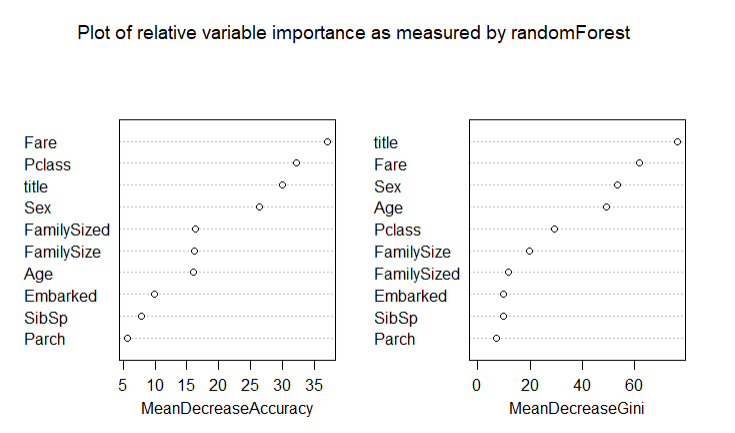
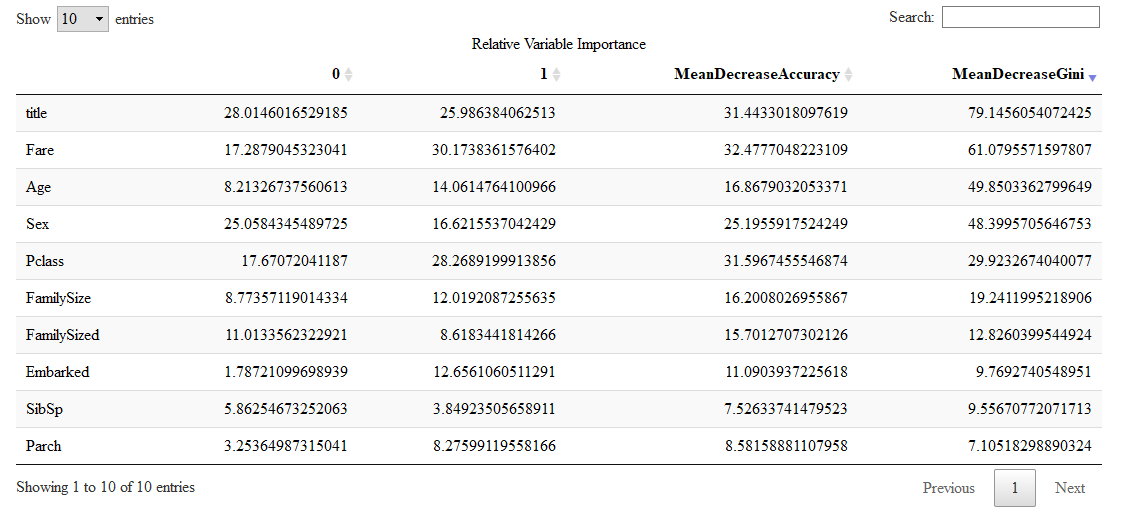


Fig: Error rate

Now we can move onto finding the important variable. The below plot validates what ctree predicted that Title is the most important variable. Sometime it is necessary to rank a variable based on its importance. Here I feel it is Title. Now we are ready for the final prediction using the training\_rf we have created on the train dataset and apply on the test dataset. But before that I would go through the extractor function for variable importance measures as produced by randomforest



We can try eliminating the least important variable (parch and SibSp) but the variable we have used here to build the model is what I think are the one which can predict the people most likely to survive in such an accident.

# **Conclusion**

One of the reasons that the titanic shipwreck led to such a tremendous loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

The analysis is based on variable pruning, subsetting and improving accuracy. Statistics are introduced through confusion matrix and plotting ROC curve. Model is strong as Random Forests with highest accuracy. Whereas by removing some parameters, we could increase the accuracy. At the end, we built model again with test dataset as we are interested in getting the survival.

# **References**

* R tutorial to learn coding in R

<https://www.tutorialspoint.com/r/>

* Understanding of Random Forests

<http://www.listendata.com/2014/11/random-forest-with-r.html>

* Udacity / Kaggle courses for machine learning

<https://www.udacity.com/course/machine-learning--ud262>

<https://www.kaggle.com/hiteshp/data-visualization-handbook/code>

## **Dataset**



## **Predicted data**



## **Source Code**

