**Passenger Prediction Survival in Titanic Using R**

**Jinnian Shen**

**4/15/2020**

## Abstract

The main aim of this paper was to predict the Survival of the people who boarded the Titanic based on the Titanic data available on Github. Besides, the question of if someone would have survived the Titanic then was investigated using the decision tree model. The results were discussed and meaningful, thus leading to the arrival of meaningful conclusions.

**Key Words:** Results, Titanic, Survival, Passenger, and Data.

**Passenger Prediction Survival in Titanic Using** **R**

## Introduction

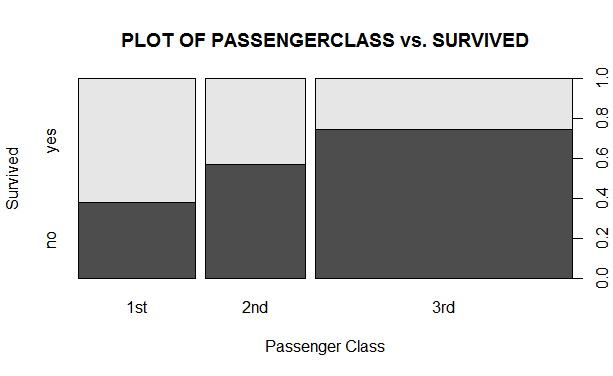
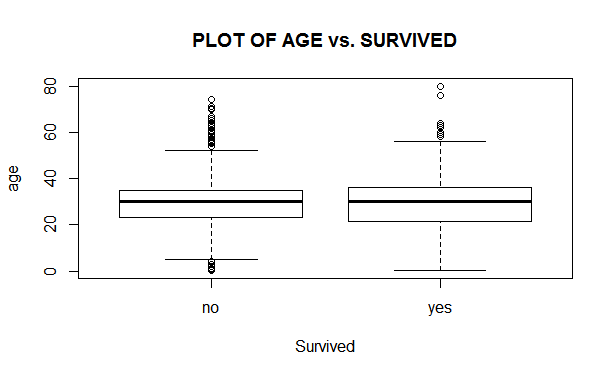
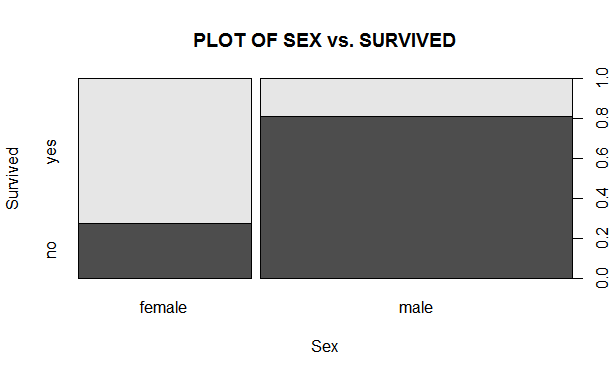
On April 15, 1912, the RMS Titanic ship collided with an iceberg and sank. Out of 2224 people on board, 1502 died. There were not enough lifeboats, and for this reason, better ship safety regulations were put in place. The Titanic tragedy forms this paper’s main aim as there was some element of luck in surviving the sinking for some groups. Some were more likely to survive than others, for example, women, children, and the upper class.

## Materials and Methods

The Titanic Survival data was available and downloadable on (https: //vincentarelbundock.github.io/Rdatasets/datasets.html.) Through the R software, data were numerically and graphically summarized, inspected, cleaned, and eventually analyzed. There were 1309 rows and five columns. The 5 column names were names of the passengers, their survival status, sex, age, and passenger class (Agrawal, 2018). All variables except age were categorical. Besides, the latter contained unique values, NAs, totaling to 263. As part of cleaning the data, the missing values were replaced with the mean of the age variable. The final data was ready for further analysis data after renaming it.

## Results

First, a visual representation of the data led to graphing of plots using the *plot ()* function in R. The plots below were for age, sex, and passenger class against survived. The passengers’ names column was not plotted because it was regarded unimportant for any further analysis.



As shown above in the first plot to the left, the portion of female passengers who survived was higher than the males who did. In other words, some fewer female passengers survived as compared to men (Agrawal, 2018). In the middle plot, the ages of the survivors were between 22 and 37. The majority of those who did not survive were aged between 22 and 35. For passengers in the 1st class, many survived while those in the third class were few. Almost half of those in the middle class survived while the rest drowned and died.

Before any data analysis, the data was partitioned into training and test datasets. The training set contained 891 rows, while the test had 418 of such. Then, using the decision tree model in the *rpart* R package, a regression model was fit on the data with the exclusion of passengers’ names variable. The response variable was survived while age, sex, and passenger class were predictors. The model was later used to predict the survivorship of passengers onboard. The prediction of the model occurred on the test data to form survivorship values, where 0 meant did not survive while one survived. The first six predicted values were:

892 893 894 895 896 897

0.8885714 0.8885714 0.8885714 0.2592593 0.5000000 0.8885714

Values were rounded off to the nearest 1000000th value. For instance, the 892nd passenger survived because 0.8885714 was rounded off to 1. For the 895th, they did not survive as 0.2592593 was a 0 when rounded off. As for the 896th, 0.5 was a 1 when rounded off. Lastly, the total number of passengers who survived the test data was totaled, and the survivorship status of the passengers in the test data was used to estimate the total number of those who survived and died. Below were the R outputs for the training (1) and testing (2) datasets.

1. 0.068 0.259259259259259 0.5 0.670454545454545

250 27 88 176

0.888571428571429

350

1. 0.259259259259259 0.5 0.888571428571429

16 128 274

In the training dataset, 277= (250+27) died while 614= (88+176+350) survived. The total number of passengers who died was 16, while survivors were 406 = (128+274). Therefore, the predicted total number of people who survived was 888, while 421 died.

## Discussion

The main aim of the paper was to investigate the prediction of the survivorship of passengers onboard the Titanic ship using the dataset provided. Therefore, to enable this, the data was partitioned and used to fit a decision tree model with the train data. The predicted values were generated in continuous form and between 0 and 1. Rounding off allowed classification of the survivorship data into 0 for did not survive and 1 for survived. The prediction of the survivorship status was conducted using test and training data and values totaled to get the predicted total number of people who survived and did not. The total number of people who boarded the Titanic ship in the predictive model was still 1309, similar to the real data.

## Acknowledgments

The whole concept of this paper owes to the knowledge of the prediction of house prices in Kaggle.com.

## References

Agrawal, S. (2018, July 4). Titanic dataset analysis. RPubs. https://rpubs.com/shivam2503/predictsurvival.

## Appendices

> titanic<-read.csv("Desktop/TitanicSurvival.csv")

> dim(titanic)

> colnames(titanic)

> summary(titanic)

> str(titanic)

> colSums(is.na(titanic))

> #removing NA values

> meang=mean(titanic$age, na.rm=TRUE)

> meang

> titanic$age.na<-is.na(titanic$age)

> titanic$age=ifelse(is.na(titanic$age),meang,titanic$age)

> summary(titanic$age)

> summary(titanic$age.na)

> mydata<-subset(titanic,select = -(age.na))

> View(mydata)

> #Plots

> plot(mydata$survived, mydata$age,xlab="Survived",ylab="age", main="PLOT OF AGE vs. SURVIVED")

> plot(mydata$sex, mydata$survived,xlab="Sex",ylab="Survived", main="PLOT OF SEX vs. SURVIVED")

> plot(mydata$passengerClass, mydata$survived,xlab="Passenger Class",ylab="Survived", main="PLOT OF PASSENGERCLASS vs. SURVIVED")

> mydata$survived<-as.numeric(mydata$survived)

> mydata$sex<-as.numeric(mydata$sex)

> mydata$passengerClass<-as.numeric(mydata$passengerClass)

> mydata$survived[mydata$survived=="2"]=0

> mydata$survived[mydata$survived=="1"]=1

> mydata$sex[mydata$sex=="1"]=0

> mydata$sex[mydata$sex=="2"]=1

> mydata$passengerClass[mydata$passengerClass=="1"]=1

> mydata$passengerClass[mydata$passengerClass=="2"]=2

> mydata$passengerClass[mydata$passengerClass=="3"]=3

> str(mydata)

> train=mydata[1:891,]

> test <- mydata[892:1309,]

> model <- rpart(survived ~ sex + age + passengerClass, data = train)

> prediction <- predict(model, test)

> head(prediction)

892 893 894 895 896 897

0.8885714 0.8885714 0.8885714 0.2592593 0.5000000 0.8885714

> solution <- data.frame(PassengerID = test$X, survived = prediction)

> write.csv(solution, file = "Solution.csv", row.names = F)

> table(solution$survived)

0.259259259259259 0.5 0.888571428571429

16 128 274

> predicted <- predict(model, train)

> head(predicted)

1 2 3 4 5 6

0.0680000 0.2592593 0.0680000 0.6704545 0.0680000 0.6704545

> solution <- data.frame(PassengerID = test$X, survived = predicted)

> write.csv(solution, file = "Solution.csv", row.names = F)

> table(solution$survived)

0.068 0.259259259259259 0.5 0.670454545454545

250 27 88 176

0.888571428571429

350