Predicting Passenger Survival for the Titanic

Taha Shakeel tshakeel

Due Monday, Novembrer 25, at 11:59PM

Contents


```
set.seed(151)
library("knitr")
library("dplyr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("MASS")
library("klaR")
library("tree")
library("rpart")
library("rpart.plot")
```

Introduction

The Titanic was designed to be an "unsinkable" luxury cruise liner. However, on its very first voyage, crossing the Atlantic ocean, the ship struck an iceberg that split the boat's hull hundreds of miles west of Canada. About only 37% of the passengers on board survived the incident. This paper will focus on training machine learning classification models to predict whether a passenger survived the Titanic based on various details of a passenger's trip.

Exploratory Data Analysis

Data Overview

The data that we will be using is from Frank Harrel, Department of Biostatistics, Vanderbilt University. It will be used to predict whether a passenger survived the Titanic based on 6 predictor variables: Class, Gender, SibSp, Parch, Fare, Embarked. The variables can be defined as so:

Pclass: ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)

Gender: male or female

SibSp: number of siblings + spouses of the individual who are aboard the Titanic

Parch: number of parents + children of the individual who are aboard the Titanic

Fare: Passenger fare in modern Great British Pounds (GBP)

Embarked: Port where passenger embarked (C = Cherbourg, Q = Queenstown, S = Southampton)

The classifiers will be used to predict the response variable: *Survived*: two categories: survived (1) or died (0)

Summary of Response Variable in the Training Dataset

The training set we will be using to create the classifications include 622 observations. Of those observations, 388 of the passengers died, making up about 62.4% of the observations and the other 234 observations making up about 37.6% of the passengers survived the Titanic. These values can be seem in the tables below.

```
##
##   0   1
## 388 234
```

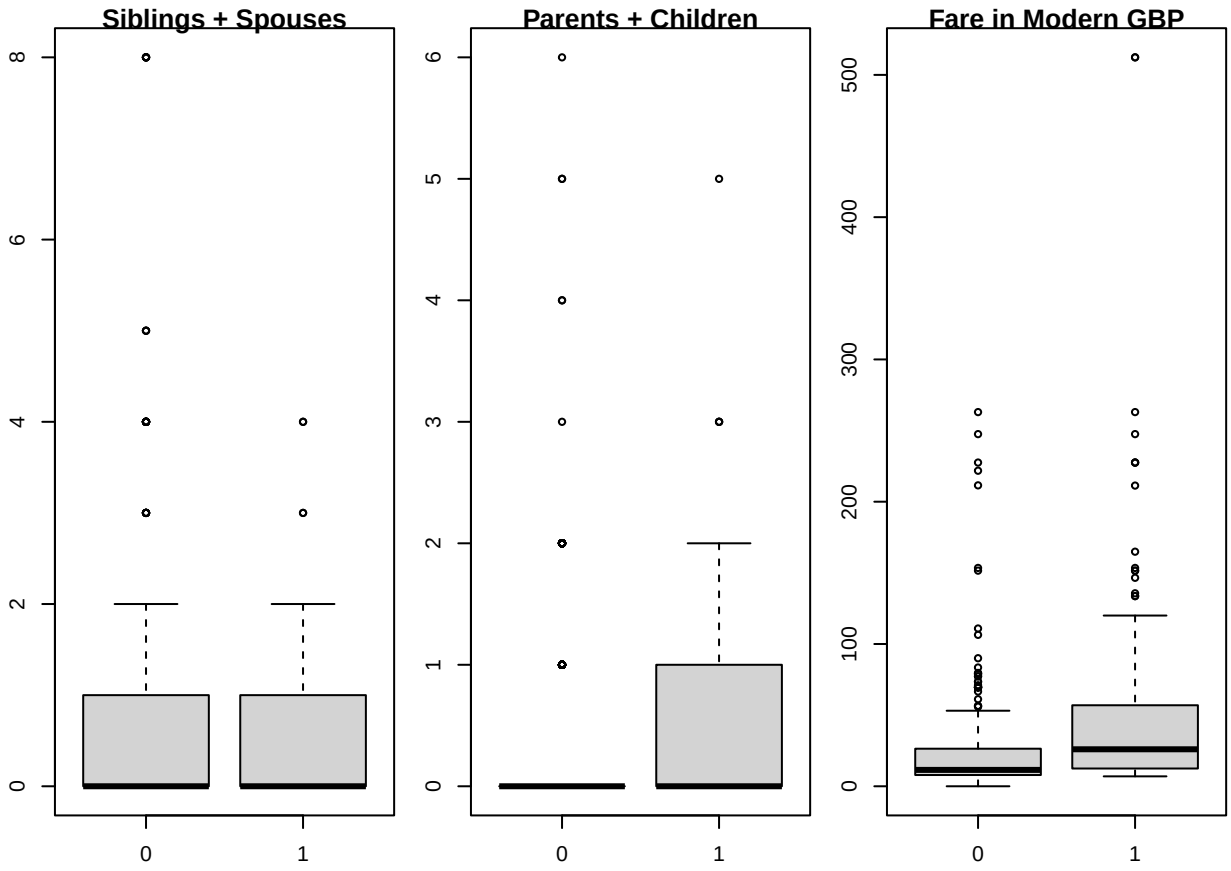
```
##
##         0         1
## 0.6237942 0.3762058
```

EDA on Relationships Between Quantitative Variables and Survival

Next lets look at visualizing the relationships between the various quantitative variables and survival. To accomplish this we will look at boxplots and to make the comparisons easier the plots will be formatted on a singular grid shown below.

```
par(mfrow = c(1, 3),
    mai = c(0.3, 0.3, 0.1, 0.1))

boxplot(SibSp ~ Survived,
        data = titanic_train,
        main = "Siblings + Spouses")
boxplot(Parch ~ Survived,
        data = titanic_train,
        main = "Parents + Children")
boxplot(Fare ~ Survived,
        data = titanic_train,
        main = "Fare in Modern GBP")
```



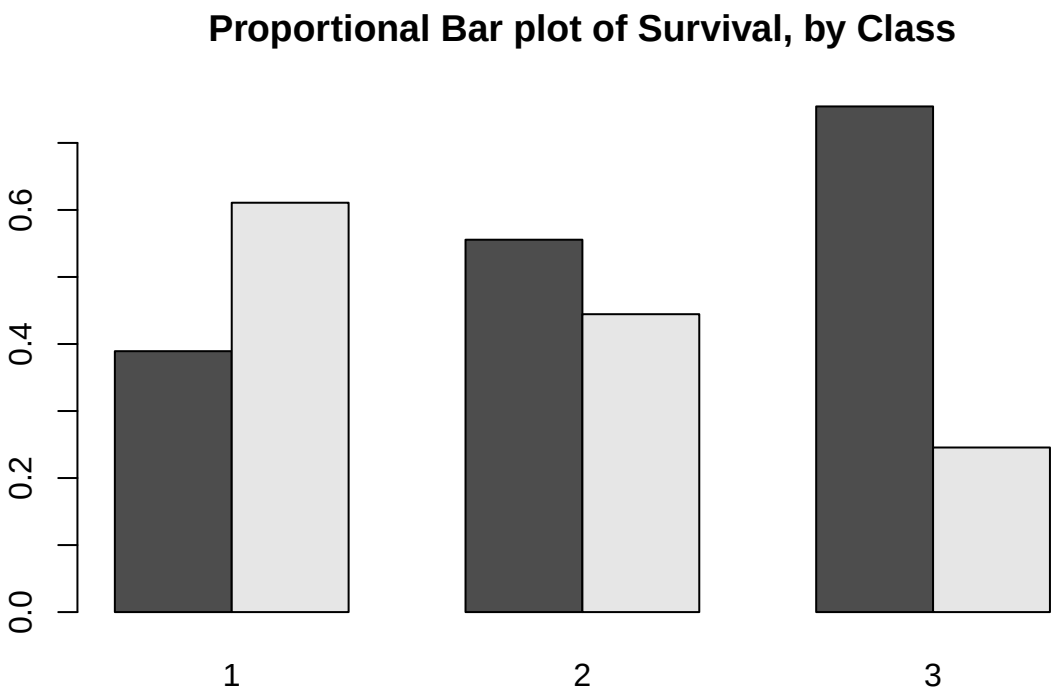
When looking at the Siblings + Spouses boxplot, we note that passengers who survived and died have roughly the same median and quartiles, however passengers who died have a larger spread and more outliers. The Parents + Children plot also shows a very similar median for both the categories in Survived. The passengers who survived has a larger IQR, but the passengers who died have several outliers creating a large range. The median and Q3 of fare the passengers pay is higher for passengers who survived, showing some evidence of a relationship.

EDA on Relationships Between Categorical Variables and Survival

To visualize the relationship between categorical variables and survival we can create a bar plot looking at the proportions of survival for the categories in the variables.

Passenger Class Visualization

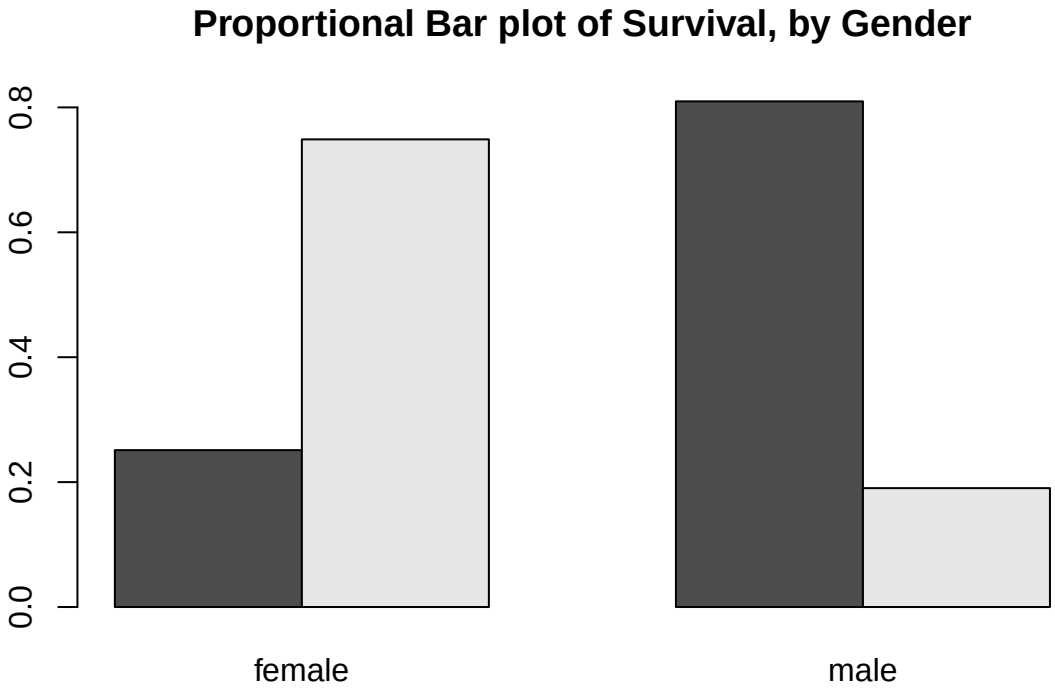
```
barplot(
  prop.table(
    table(titanic_train$Survived, titanic_train$Pclass),
    margin = 2),
  beside = TRUE,
  main = "Proportional Bar plot of Survival, by Class")
```



As the class number increases the proportion of passengers that didn't survive increase.

Passenger Gender Visualization

```
barplot(
  prop.table(
    table(titanic_train$Survived, titanic_train$Gender),
    margin = 2),
  beside = TRUE,
  main = "Proportional Bar plot of Survival, by Gender")
```



A large proportion of females survived compared to males where a majority died.

Passenegr Embarked Location Visualization

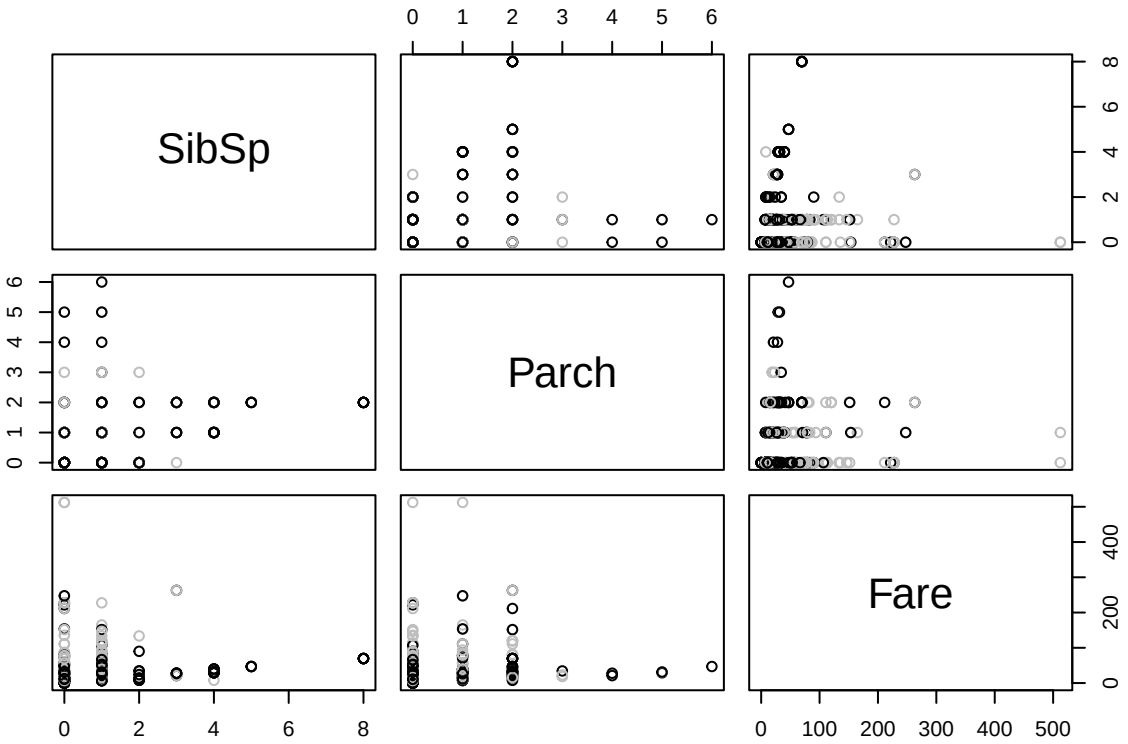
```
barplot(
  prop.table(
    table(titanic_train$Survived, titanic_train$Embarked),
    margin = 2),
  beside = TRUE,
  main = "Proportional Bar plot of Survival, by Embarked Location")
```

The passengers who embarked at Cherbourg had the largest proportion of survival (grey bar). The other two ports, Queenstown and Southhampton had a larger proportion of passengers who died.

EDA on Classification Pairs

To visualize how quantitative predictors will classify survival and their effectiveness, we can create a pairs plot.

```
pairs(titanic_train[ , c(3, 4, 5)],
      col = ifelse(titanic_train$Survived==0, "black", "grey"))
```



The pairs plot gives us a visual of the combinations of variables and how they separate surviving, the grey dots and dying, the black dots. The pair of variables that has the best separation of survival are SibSP (siblings + spouses) and Parch (parents + children). Having separation between the black and grey dots and being able to group individual colors is good for the modeling we will be doing in the next step. When looking at Fare the black and grey are grouped very close together which means it will be harder to group without error.

Modeling

Having completed the Exploratory Data Analysis, we can move to created and assessing classifiers for predicting the survival of passengers aboard the Titanic. The classifiers that will be used are linear discriminant analysis, quadratic discriminant analysis, classification trees, and binary logistic regression. We will first train the classifiers on a training data set and then asses their effectiveness on a testing data set, both of which were randomly created from the original data set.

Linear Discriminant Analysis (LDA)

A LDA model can only take quantitative variables, because of this for this model we will only be able to use SibSP, Parch, and Fare.

```
##
##        0   1
##   0 149  83
##   1  12  23
```

Overall the LDA had a error of (12+83)/267 = 0.356 which is alright. The error rate for predicting deaths was significantly lower at 12/161 = 0.075. However, predicting survival is not a strong suite for the LDA classifier as that has an error rate of 83/106 = 0.783.

Quadratic Discriminant Analysis (QDA)

Similar to the lda, a qda classifier only takes quantitative inputs. Creating a qda has the exact same process as a lda, just that we will now be creating groups quadratically which adds flexibility in the process of creating groups.

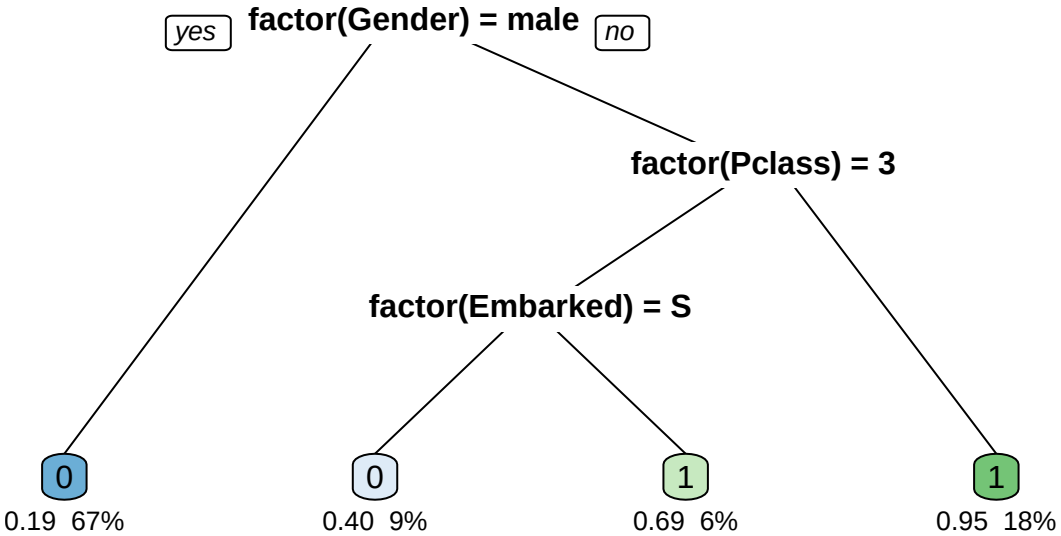
```
##
##        0   1
##    0 146  73
##    1  15  33
```

The overall error rate of the qda is $(15+73)/267 = 0.330$ which is less than the overall error rate in the lda which was 0.356, indicating that the qda classifier is more effective. The error rate for passengers that survived is $73/106 = 0.689$ which is lower than in the lda, but the error rate for passneger who didn't survive increased to $15/161 = 0.0931$.

Classification Tree

By using a classification tree we can now account for the categorical variables in addition to the quantitative variables.

```
titanic_tree <- rpart(factor(Survived) ~ factor(Pclass) +
                        factor(Gender) + factor(Embarked),
                      data = titanic_train,
                      method = "class")
rpart.plot(titanic_tree,
           type = 0,
           clip.right.labs = FALSE,
           branch = 0.1,
           under = TRUE)
```



The tree selected to use Gender as the first split in the tree, indicating that Gender is the best separator for survival. Next lets look at the performance of the tree with the test data set.

```
##
## titanic_tree_pred   0   1
##                 0 153  40
##                 1   8  66
```

The classification tree with the lowest error is the one that only has categorical variables as the predictors.

The overall error rate for this tree is (8+40)/276 = 0.18 which is lower than the overall with all the predictors which came out to be (20+32)/267 = 0.195. However, this tree is worse at predicting survived with an error rate of 40/106 = 0.377 compared to the rate of 32/106 = 0.302. On the flip side, this tree has a very low error in predicting passenger death of 8/161 = 0.05 compared to 20/161 = 0.124. Because of the lower overall error rate, I believe that the tree with only categorical predictors is better than the tree with six predictors.

Binary Logistic Regression

The last classification model we will be conducting is binary logistic regression which can account for both quantitative and categorical predictors, similar to the classification tree.

Logistic models provides probabilities, meaning we will have to convert the probability into predictions.

```
## [1] "0" "1"
```

```
##
## titanic_logit_pred   0   1
##                  0 131  30
##                  1  30  76
```

The binary logistic regression has an overall error of (30+30)/267 = 0.225, higher than the classification trees but better than LDA and QDA. The error when predicting passengers who survive is 30/106 = 0.283 and for passengers who didn't survive it is 30/161 = 0.186.

Final Recommendation

After training four different classifiers on a training data set and then evaluating them with a testing data set, we can now determine the best classifier in predicting survival. Overall the classification tree with only categorical predictors had the lowest error rate at 0.18. This is followed by binary linear regression, QDA, and then LDA.

In predicting if a passenger is not going to survive the titanic, the classification tree did the best followed by LDA, QDA, and then binary logistic regression.

In predicting whether a passenger survives the titanic, binary logistic regression has the lowest error of 0.283 followed by the classification tree, QDA, and then LDA.

With all this information in mind, our final recommendation is that the classification tree using only the categorical predictors be used to predict the survival of passengers aboard the titanic.

Discussion

The four classifiers we used to predict the survival of passengers aboard the titanic all did relatively well. The highest overall error of 0.356 came from the linear discriminant analysis. While the lowest error rate of 0.18 came from the classification tree.

The common trend between all the classifiers is a higher error rate when predicting if a passenger survived. We should also note that the best error rates came from models that incorporated categorical predictors. Highlighting the importance of socio-demographic factors like gender and class as being some of the best separators between surviving and not surviving the luxury ship.

For the future studies, it would be interesting to see the results on a larger scale as these data sets are relatively small. On a larger scale, incorporating additional factors such as age, nationality, etc would potentially increase the success of the classifier models. However, we do have to keep in mind including too many factors does lead to over fitting which is something we also have to keep in mind in this study as well, especially for the classification tree. Having this additional information may help create a more thorough and robust study.