Assesment of Titanic Survival Factors

Ryan Wahler rwahler

Due Fri, April 19, at 11:59PM

Contents


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

Introduction

The titanic incident was one of the most deadly maritime disasters in which over one thousand died in the cold waters of the North Atlantic. The passenger ship kept extensive records of its passengers, of each of their characteristics and later, records were kept of which passengers survived and died in the accident. This paper will evaluate machine learning classifications in order to correctly classify passengers by their characteristics into the types of survived or dead.

[Data from Frank Harrell, Department of Biostatistics, Vanderbilt University, https://hbiostat.org/data/repo/titanic.html]

Exploratory Data Analysis

Variables

Each passenger was accompanied by the following characteristics:

- Class (ticket class (1 = first, 2 = 2nd, 3 = 3rd))
- Gender (male or female)
- SibSp (number of siblings + spouses of the individual who are aboard the Titanic)
- Parch (number of the parents + children of the individual who are aboard the Titanic)
- Fare (Passenger fare (adjusted to the equivalent of modern British pounds))
- Embarked (port of Embarkation (C=Cherbourg, Q=Queenstown, S=Southamptom))

The responses to be predicted by the classifiers

- Survived (survived (1) or dead (0))

Summary of Responses in the Training Dataset

There are 622 observations (passengers) withing the training data set and 267 passengers within the test data set for a total of 889 passengers within our working data. Of the test data set, 161 passengers, 60%, survived and 106, 40%, dead. While in the training data set, 234 passengers, 38%, survived and 388 passengers, 62% dead (visualized in the table below).

```
table(titanic_train$Survived)
```

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

```
prop.table(table(titanic_train$Survived))
```

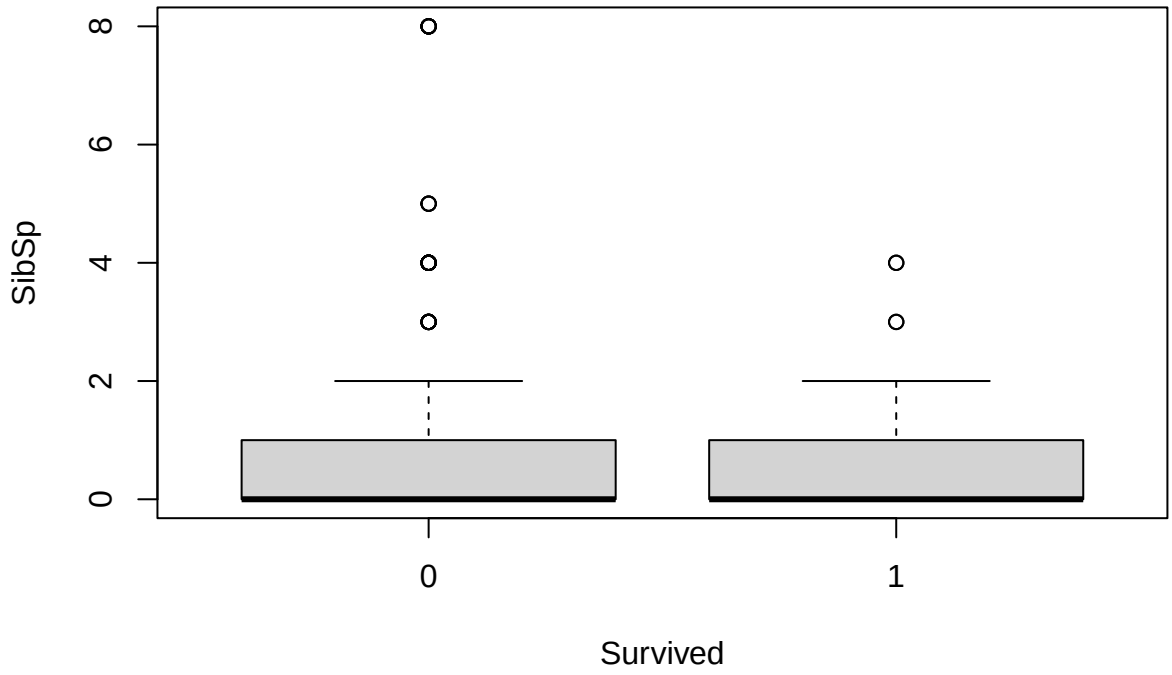
```
##
##         0         1
## 0.6237942 0.3762058
```

EDA on the relationships of Passenger Characterstics and Survival.

To visually explore the relationships of the response (survived) and the numerous predictors (characteristics of the passengers) boxplots will be constructed for the quantitative predictors and conditional proportions of type for the categorical predictors. Additionally a pairs plot of the quantitative predictors will give indication of what combinations of quantitative variables will be effective in predicting the survived response.

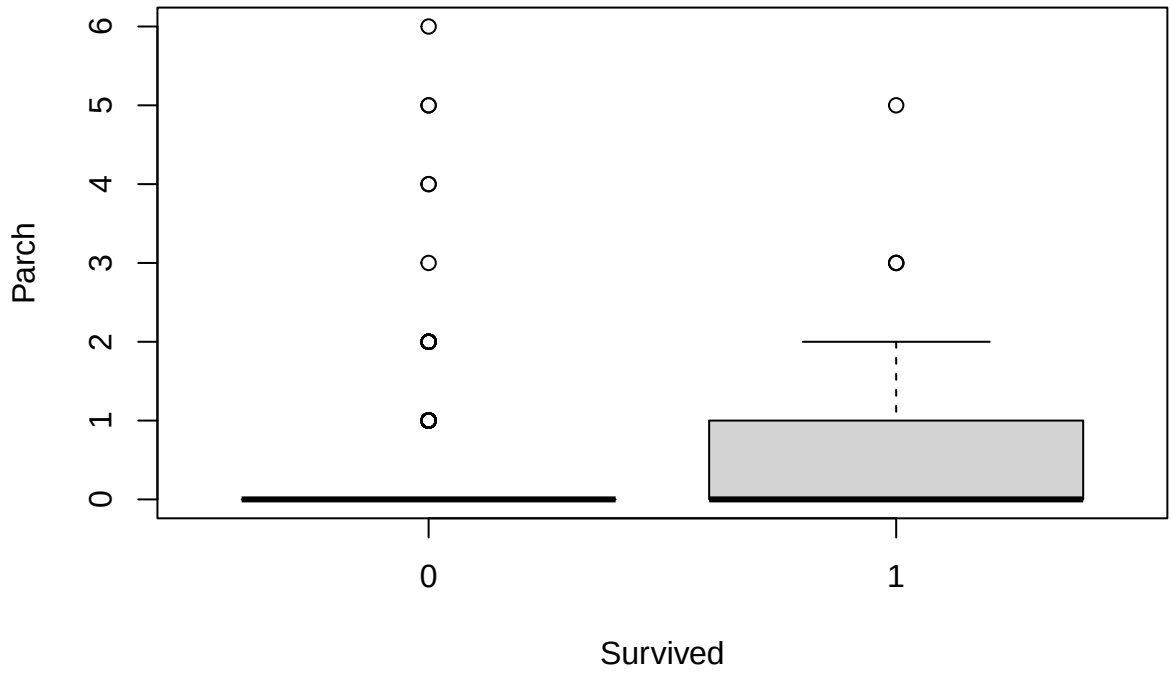
```
boxplot(SibSp ~ Survived,
        main = 'Number of siblings + spouses of the individual who are aboard the Titanic',
        data = titanic_train)
```

Number of siblings + spouses of the individual who are aboard the Tit



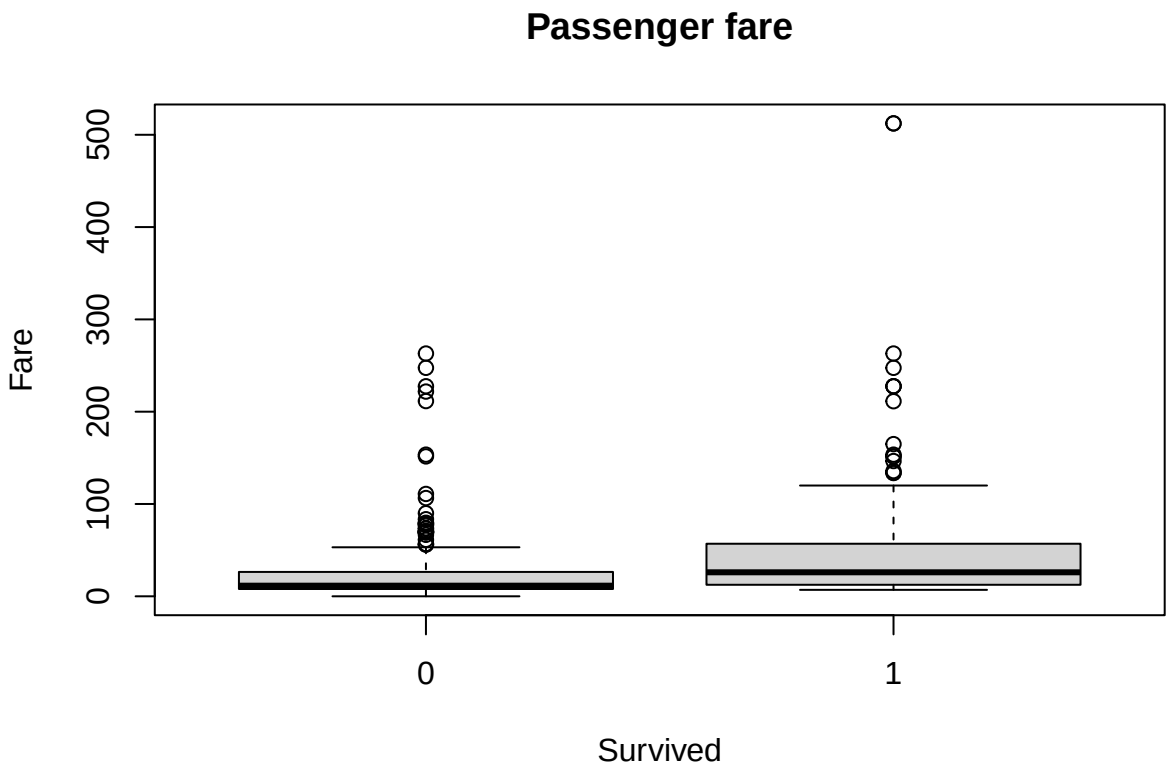

```
boxplot(Parch ~ Survived,
        main = 'Number of the parents + children of the individual who are aboard the Titanic',
        data = titanic_train)
```

lumber of the parents + children of the individual who are aboard the T



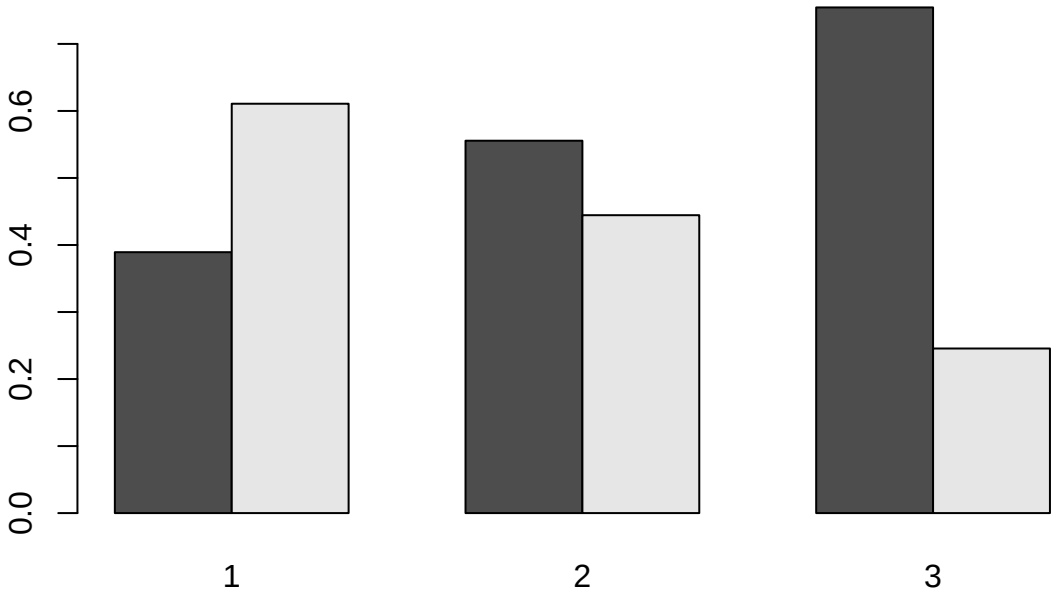

```
boxplot(Fare ~ Survived,
        main = 'Passenger fare',
```

```
        data = titanic_train)
```

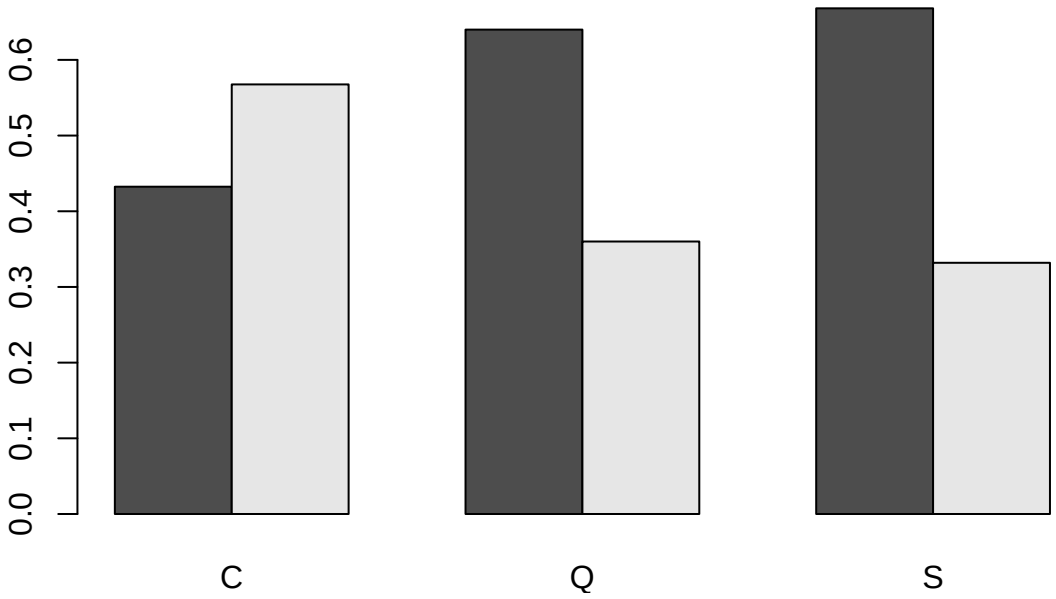


In the boxplots displayed, there would be evidence of a relationship between the response and the respective predictor that would require further exploration and may be significant to the classifiers to be constructed. Most noticed is that there is a higher concentration of passengers that survived with higher counts of parents/children also aboard the titanic. Passengers that survived also appeared to have paid more for their fare then passengers that did not. While the number of siblings and spouses of the passenger also on the titanic did not appear to have an association with that passenger's survival.

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

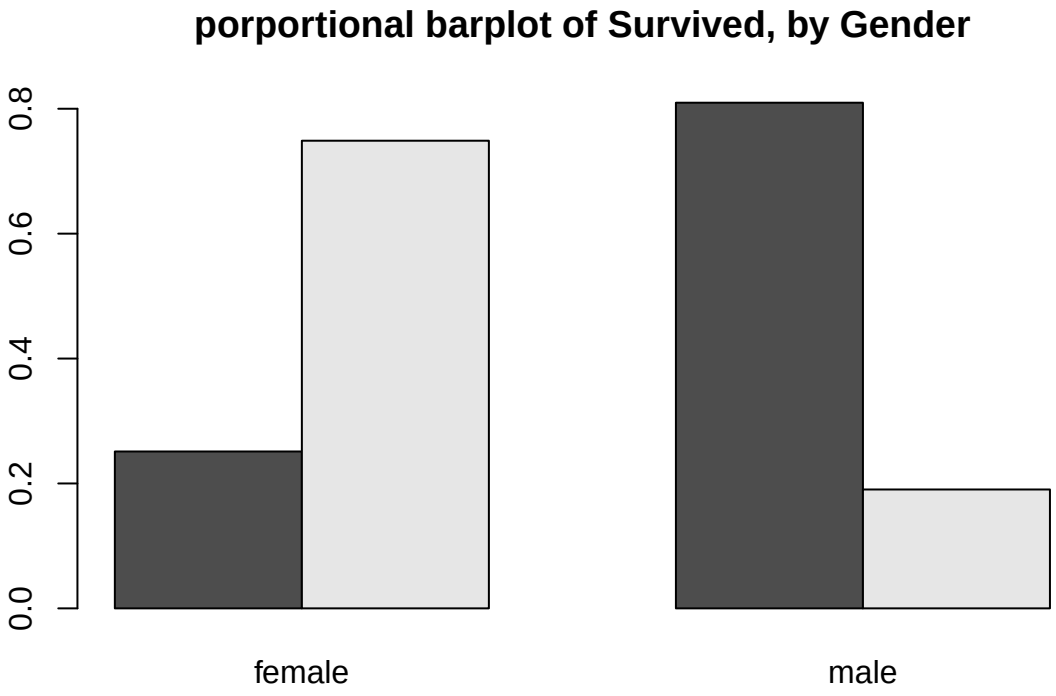



```
barplot(
  prop.table(
  table(titanic_train$Survived, titanic_train$Embarked),
            margin = 2)
  , beside = TRUE,
  main = "porportional barplot of Survived, by port of Embarkation")
```



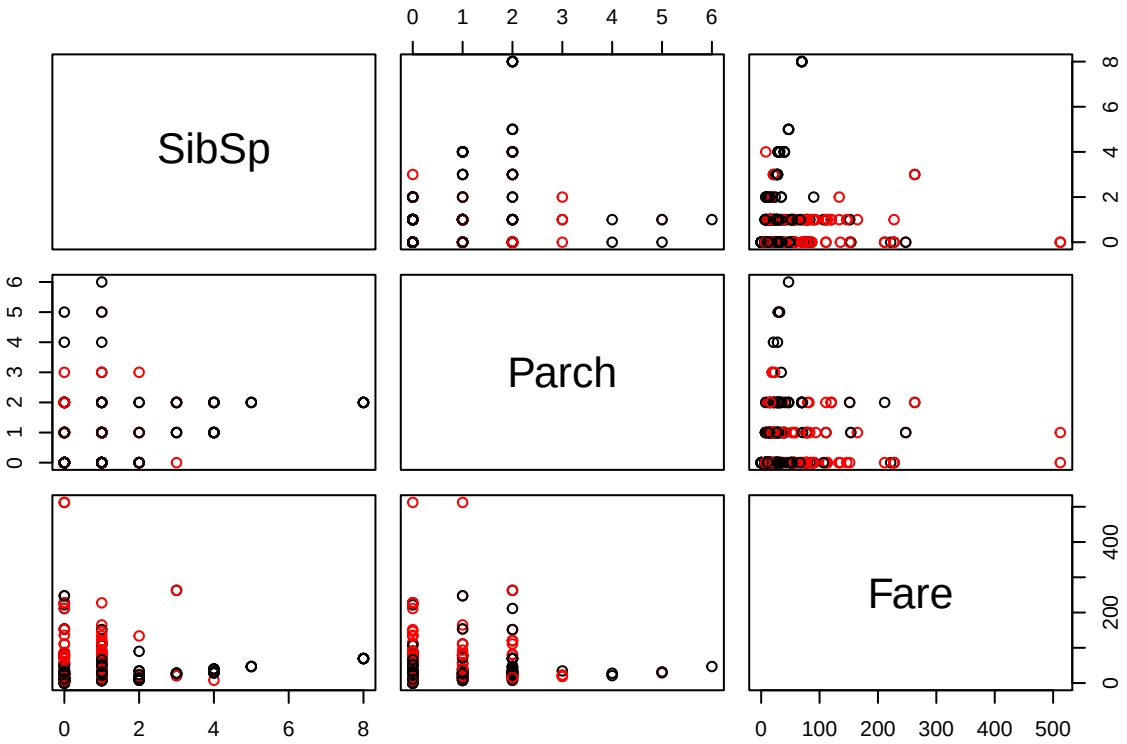

```
barplot(
  prop.table(
  table(titanic_train$Survived, titanic_train$Gender),
```

```
          margin = 2)
  , beside = TRUE,
  main = "porportional barplot of Survived, by Gender")
```



For our categorical variables, by the first barplot above, it appears a larger portion of passengers survived in first class than the subsequent classes, the higher class (First being the highest) the more likely the passenger is to survive. Additionally, the proportional barplot of Survived, by port of Embarkation, indicated that the passenger is most likely to have survived if they embarked in Cherbourg, while if they embarked in Queenstown or Southampton they would have a similar likelihood of survival.

```
pairs(titanic_train[ , c(3, 4, 5)],
      col=ifelse(titanic_train$Survived==1, "red", "black"))
```



The plot (of the pairs plot above) that shows the best separation of response variables (red = survived, black = dead) gives which quantitative predictors will be best suited for our classification models. The plot is not very promises with no clear separation above, measures such as SibSp and Parch only have integer values any many passengers directly over lap giving a poor indication (non-continuos variables). However SibSp (number of siblings + spouses of the individual who are aboard the Titanic) does appear to give the best separation across the predictors.

Of note in the Fare predictor there was one outlier that paid very well for the titanic and died. Additionally there was an outlier in SibSp that had eight number of siblings and/or spouses who are aboard the Titanic and died.

Modeling

We will construct four models by the classifiers; linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), classification trees and binary logistic regression. To appease the problem of over fitting, the data set was split into a training set and testing set, as was mentioned above. The EDA was and all Modeling will be constructed based on the training data and the test data will assess our functioning models to ensure there is not over fitting.

Linear Discriminant Analysis (lda)

```
titanic.lda <- lda(factor(Survived) ~ Fare,
                   data = titanic_train)
titanic.lda.pred <- predict(titanic.lda,
                            as.data.frame(titanic_test))
table(titanic.lda.pred$class, titanic_test$Survived)
```

```
##
##       0   1
##   0 155  84
##   1   6  22
```

To note, LDA models can only utilize continuous quantitative variables, the only characteristic of the passengers that fits these parameters is the fare of their ticket. While the other two quantitative variables, Parch and SibSp are based on the counts of people, and there are no fractions of people, only integer values. As a result our LDA model is solely trained on the passenger's respective fare, from passengers contained in the training data. After fitting the classifier by training data, running the test data, there was an overall error rate of ((84+6)/267 = 0.337) which is relatively poor. The overall error rate was composed of the error rates, classifying survived (84/106 = 0.792) and classifying dead (6/161 = 0.039). While the LDA classifier did quite well classifying the dead, it did very poor in classifying the survived.

Quadratic Discriminant Analysis (qda)

```
titanic.qda <- qda(factor(Survived) ~ Fare,
                   data = titanic_train)
titanic.qda.pred <- predict(titanic.qda,
                            as.data.frame(titanic_test))
table(titanic.qda.pred$class, titanic_test$Survived)
```

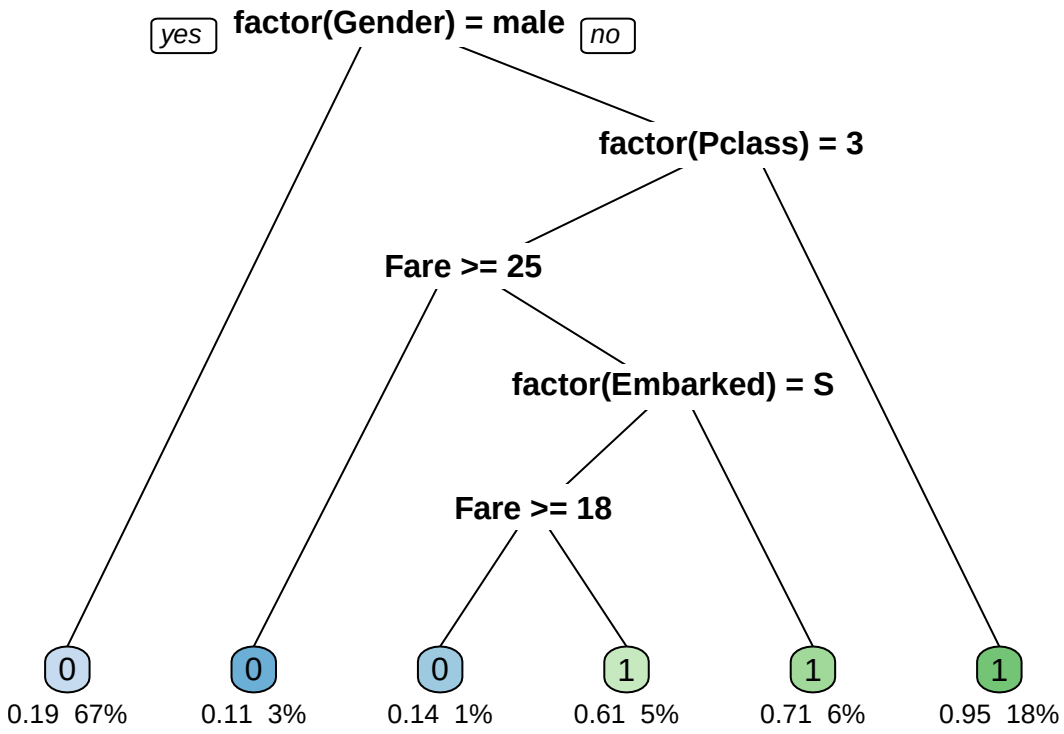
```
##
##       0   1
##   0 154  82
##   1   7  24
```

The same caveat as the LDA, QDA models can only utilize continuous quantitative variables, again the only characteristic of the passengers that is a continuous quantitative variable is the fare of their ticket. The QDA classifier was trained on our training data set, solely by the Fare characteristic. The table above displays the result of running the test data set through the QDA classifier compared to their true classifications. By the table, the QDA classifier had an overall error rate of ((82+7)/267 == .333). Which is composed of the error rates, classifying survived (82/106 == .77) and classifying dead (7/161 = .045). While the QDA classifier did well to classify the dead, much like the LDA classifier, it did very poor in classifying the survived. Actually in comparison to the LDA classifier, it did marginally better overall and in classifying the survived and slightly worse classifying the dead.

To note, there is an argument despite the status of the other quantitative variables (SibSp and Parch) being non-continuous they should be included in QDA and LDA. The sentiment being that variables are rarely every normal and truly continuous. Yet upon inclusion of these variables into QDA and LDA, an only marginally better overall error rate for QDA of (.330) and a significantly worse error rate for LDA of (.356).

Classification Trees

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




```
titanic.tree.pred <- predict(titanic.tree,
                             as.data.frame(titanic_test),
                             type = "class")
table(titanic.tree.pred, titanic_test$Survived)
```

```
##
## titanic.tree.pred   0   1
##                 0 141  32
##                 1  20  74
```

The primary advantage to a classification tree in comparison to the LDA and QDA classifiers that were just constructed is that a classification tree can account for all the quantitative and categorical characteristics of the passengers. Training the classification tree on the training data, it decided that the "most important" variable was the characteristic of gender of the passengers in discerning the survived and dead. The characteristic of gender was followed by passenger class, fare which had a threshold of 25, whether the embarked in South Hampton, and finally their fare now at a threshold of 18. The table above is the result of running the test data set through the QDA classifier and comparing to the respective passenger's true classifications. There was a lower overall error rate for the classification tree ((32+20)/267 == .195), signifying that the classification tree preformed better that the LDA and QDA classifiers. The overall error rate was composed of the error rate classifying the survived (32/106 == .302), better than both QDA and LDA, and the error rate classifying the dead (20/161 = .124), actually worse than those of LDA and QDA.

To note, in attempts to avoid over fitting the model to the training data and obtain a better result for the test data, the tree was pruned several times. But each level of pruning had a worse overall error rate than the original tree above. Indicating the tree was not over fitting to the training data and its classification can be generalized.

Binary Logistic Regression

```
titanic.logit <- glm(factor(Survived) ~ factor(Pclass) + Fare + SibSp + Parch + factor(Embarked) + facto
                     data = titanic_train,
```

```
                  family = binomial(link = "logit"))
titanic.logit.prob <- predict(titanic.logit,
                              as.data.frame(titanic_test),
                              type = "response")
levels(factor(titanic_test$Survived))
```

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

```
titanic.logit.pred <-ifelse(titanic.logit.prob > 0.5,"1","0")
table(titanic.logit.pred, titanic_test$Survived)
```

```
##
## titanic.logit.pred   0   1
##                  0 131  30
##                  1  30  76
```

Binary logistic regression comes with the same advantage as classification trees in comparison to QDA and LDA, where all characteristics of the passengers, quantitative and categorical can be used to classify the passengers. First training the logistic classifier on the training data, the logistic model returns probabilities given respective passenger's predictor variables (characteristics). In this case by the display above the confusion matrix, 1 (survived) is taken to be the 'yes' side where the probability is greater than .5. Otherwise, for probabilities less than .5 these are taken to be the predictions for the dead. These definitions are coded into the if-else logic so that any probability returned gives the proper response survived or dead for the associated probability. Taking the test data, running it through the binary logistic regression classifier and encoding each passenger to survived or dead based on their respective probability gives the confusion matrix. The confusion matrix operates just as the tables for the classification tree, QDA and LDA above, where the overall error rate ((30+30)/267 = .225) just slightly under preforms in comparison to the classification tree. The overall error rate is composed of the of the error rate classifying the survived (30/106 == .283). and the error rate classifying the dead (30/161 = .186).

Final Recommendation

Each model did notably better in classifying the dead than the survived. Out of these models LDA did the best in classifying the dead while the logistic classifier preformed best in classifying the survived. The classification tree had the smallest overall error rate and is the recommendation and given the similarly low overall error rate the logistic classifier would be an adequate alternative. In addition to the low overall error rate, the classification tree and the logistic model utilize all of the variables provided where LDA and QDA were limited to just one variable.

Discussion

The overall error rates of each of the models on the test data were not impressive, especially those of LDA and QDA, only when the classification tree and logistic classifiers were explored did more reasonable error rates come about. This can be likely attributed to the lack of permitted variables for LDA and QDA models, in this data set only fare was permitted because it was a continuous quantitative variable. The fact that the error rate of classifying the dead was so low for LDA and QDA suggests that fare is a good predictor for dead. While the other quantative and categorical variables of the passengers was needed to accurately predict the survived.

As this data comes from a historical event it would be difficult for more predictors to be collected and more instances of the response variable collected. However, a possibly characteristic to predict the passenger's response (survived or dead) would be their room location on the ship, (for example, how high above or below deck their room was). This comes with the caveat that it could be too highly correlated with fare or class. Further analysis on the titanic and the predictors of survival could aid maritime safety measures and procedures in the future and is something to pursue.