Life or Death on the Titanic

Michael Pilson mpilson

Due Wed, November 30, 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 remains one of the most notorious disasters in human history. The "unsinkable" ship collided an iceberg on April 15, 1912. What occurred after was more tragic, the ship slowly sank into the water and with it around 1500 died. Those that survived were forced to cling onto pieces of the wreckage or scramble onto the few life boats. However, many have speculated what factors determined if you were to survive or die. In this document, we have tried to answer this question through statistics. We will use basic machine learning classification techniques to put them into two categories, survive or did not survive.

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

Exploratory Data Analysis

Overview

To determine if a passenger survived or did not we will be looking at these contributing factors in the training dataset provided (titanic_train):

Explanatory Variables

Categorical

Pclass: ticket class (1 = 1st, 2 = 2nd, 3 = 3rd) Gender: male or female Embarked: Port of Embarkation (C=Cherbourg, Q=Queenstown, S=Southampton

Quantatative

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 (adjusted to equivalent of modern British pounds) Embarked: Port of Embarkation (C=Cherbourg, Q=Queenstown, S=Southampton

Response Variable Survived: survived (1) or dead (0)

Summary of Response Variable (Survived or Did not)

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

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

```
x=388/622
x
```

```
## [1] 0.6237942
```

```
y=234/622
y
```

```
## [1] 0.3762058
```

So out of the 622 passengers taken in this sample, 388 or 62.4% of the passengers died and 234 or 37.6% survived.

EDA of Individual Explanatory Variables with Response

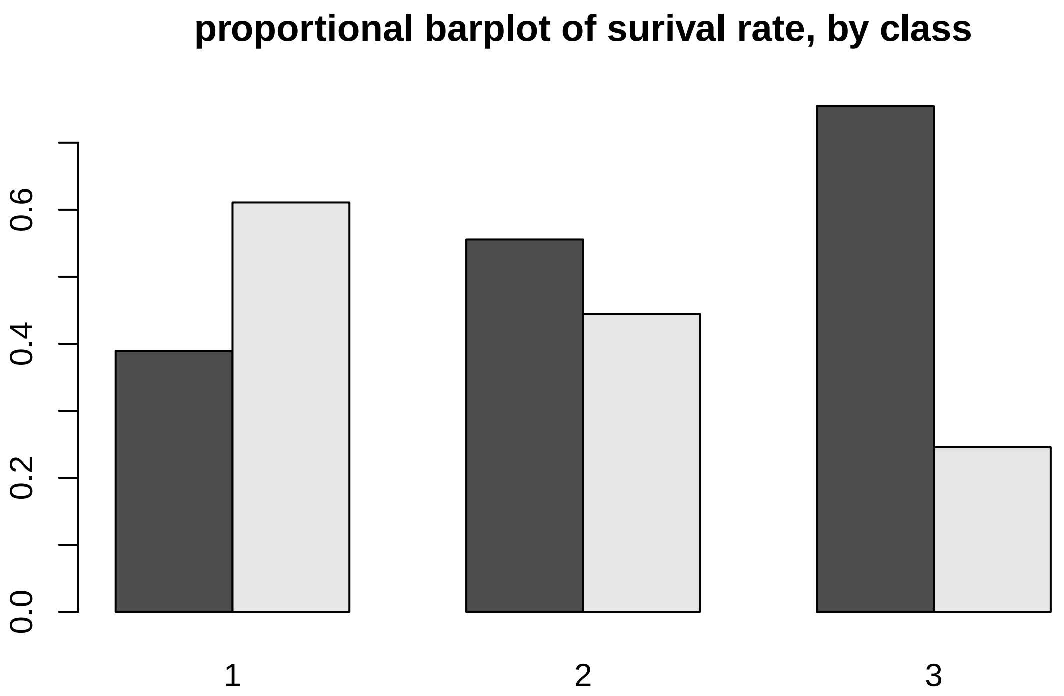
Now we will explore the relationships between the individual explanatory variables and the response survival.

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

```
prop.table(table(titanic_train$Survived, titanic_train$Pclass),margin = 2)
```

```
##
##             1         2         3
##   0 0.3892617 0.5555556 0.7544379
##   1 0.6107383 0.4444444 0.2455621
```

```
barplot(prop.table(table(titanic_train$Survived, titanic_train$Pclass),margin = 2), beside = TRUE,
main = "proportional barplot of surival rate, by class")
```



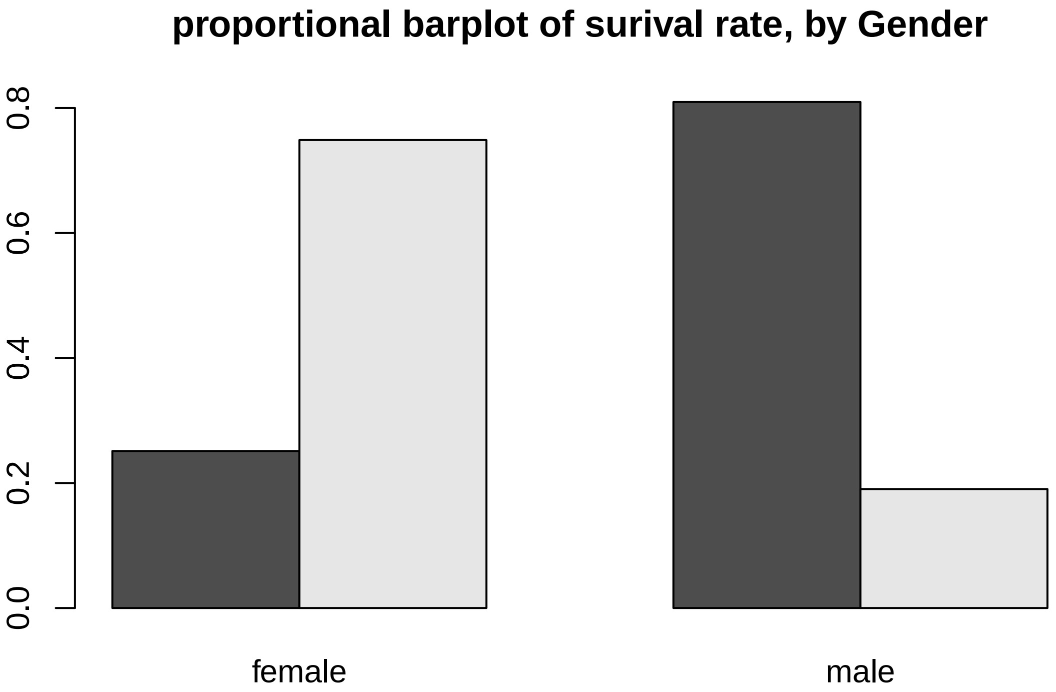
From this proportional barplot, one can see that there seems to be clear relationship between the class and if they survived. As the class progresses from 1st to 2nd to 3rd, the dark bar, or the percentage of deaths out of the people in the class, increases.

Gender: male or female

```
prop.table(table(titanic_train$Survived, titanic_train$Gender),margin = 2)
```

```
##
##        female      male
##   0 0.2512077 0.8096386
##   1 0.7487923 0.1903614
```

```
barplot(prop.table(table(titanic_train$Survived, titanic_train$Gender),margin = 2), beside = TRUE,
main = "proportional barplot of surival rate, by Gender")
```



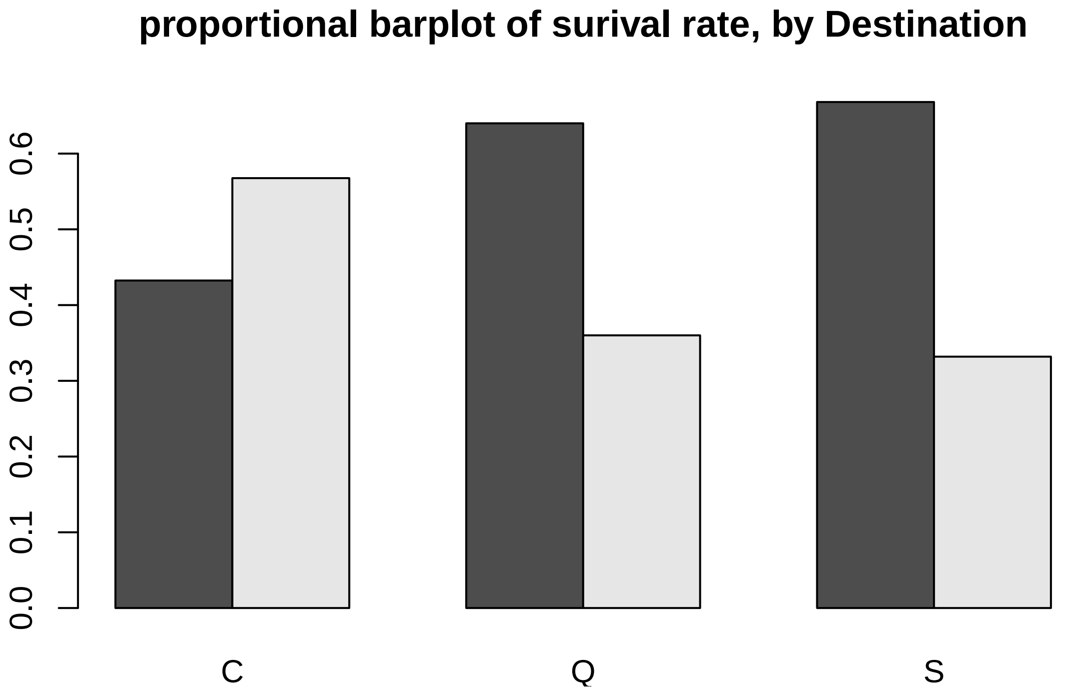
From this proportional barplot, one can see that there seems to be clear relationship between the gender and if they survived. For men the dark bar, or the percentage of men that died, is significantly higher than women.

Embarked: Port of Embarkation (C=Cherbourg, Q=Queenstown, S=Southampton

```
prop.table(table(titanic_train$Survived, titanic_train$Embarked),margin = 2)
```

```
##
##             C         Q         S
##   0 0.4324324 0.6400000 0.6681128
##   1 0.5675676 0.3600000 0.3318872
```

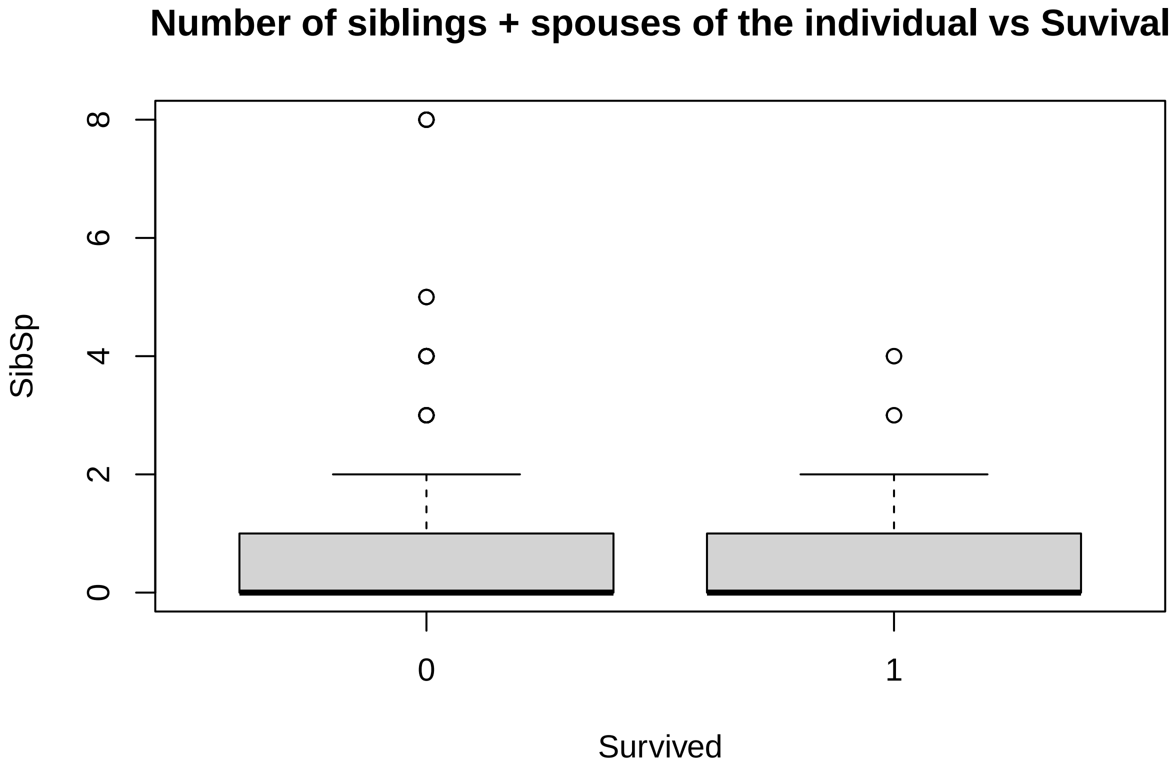
```
barplot(prop.table(table(titanic_train$Survived, titanic_train$Embarked),margin = 2), beside = TRUE,
main = "proportional barplot of surival rate, by Destination")
```



From this proportional barplot, one can see that there seems to be a correlation between the intended destination and if they survived. For those traveling to Queensboro or Southhampton the dark bar, or the percentage of those passengers that died, is higher than those traveling to Cherbourg.

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

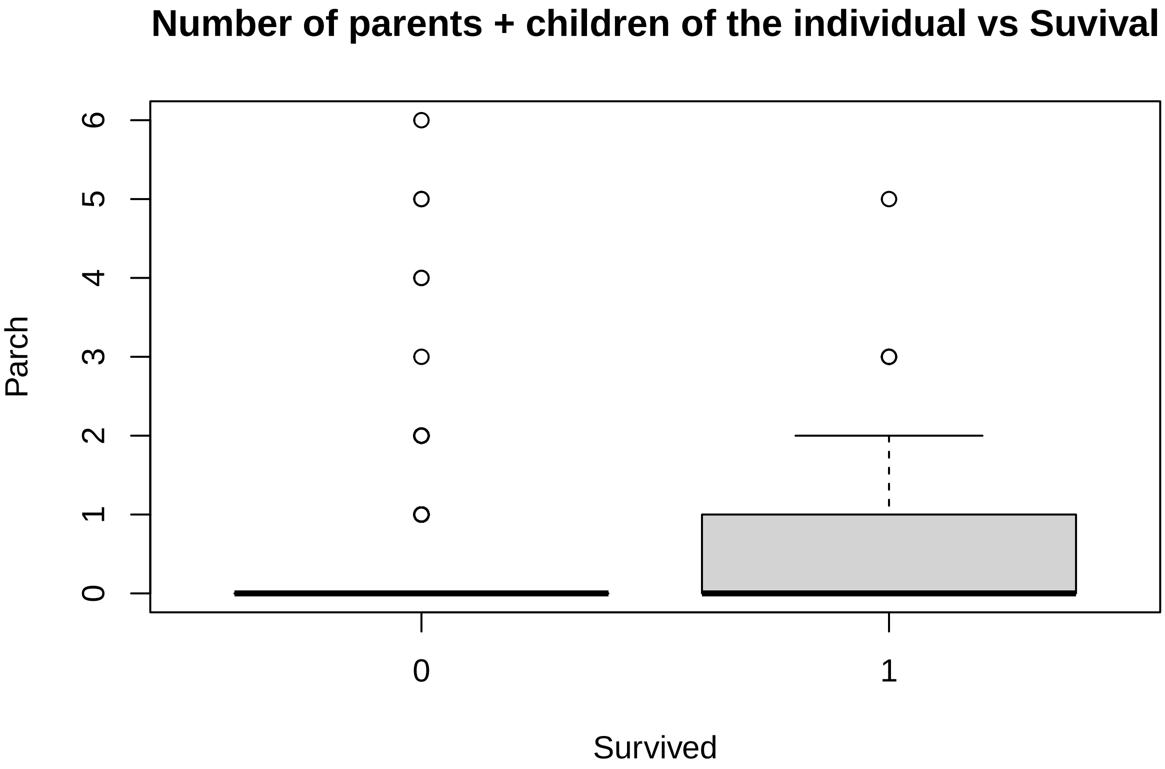
```
boxplot(SibSp ~ Survived,
main="Number of siblings + spouses of the individual vs Suvival",
data=titanic_train)
```



From this boxplot, one can conclude that the number of siblings + spouses of the individual who are aboard the Titanic, does not seem to affect your chances of surviving as the two seem to have identical distributions for those who lived or died.

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

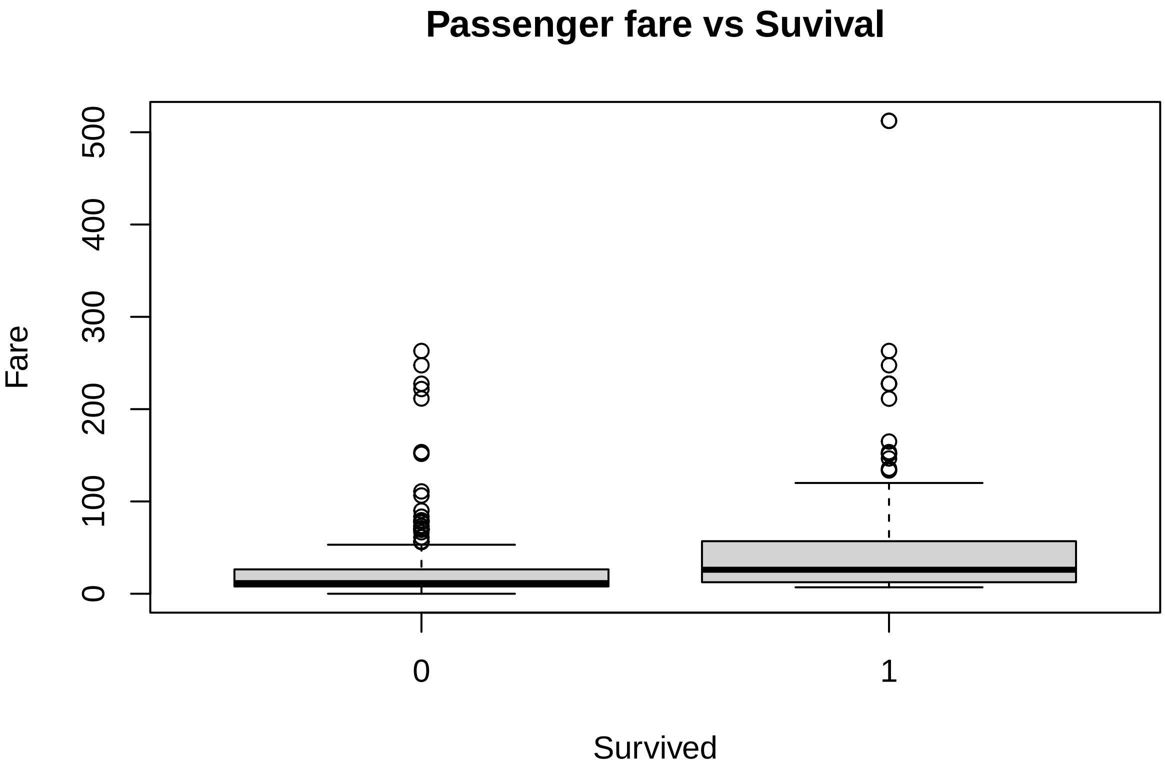
```
boxplot(Parch ~ Survived,
main="Number of parents + children of the individual vs Suvival",
data=titanic_train)
```



From this boxplot, one can conclude that the number of parents + children of the individual who are aboard the Titanic did seem to slightly affect your chances of survival. The data for those that survived was concentrated farther upward indicating that the more parents and children you had on aboard may have slightly increased your chances of surviving your survival.

Fare: Passenger fare (adjusted to equivalent of modern British pounds)

```
boxplot(Fare ~ Survived,
main="Passenger fare vs Suvival",
data=titanic_train)
```



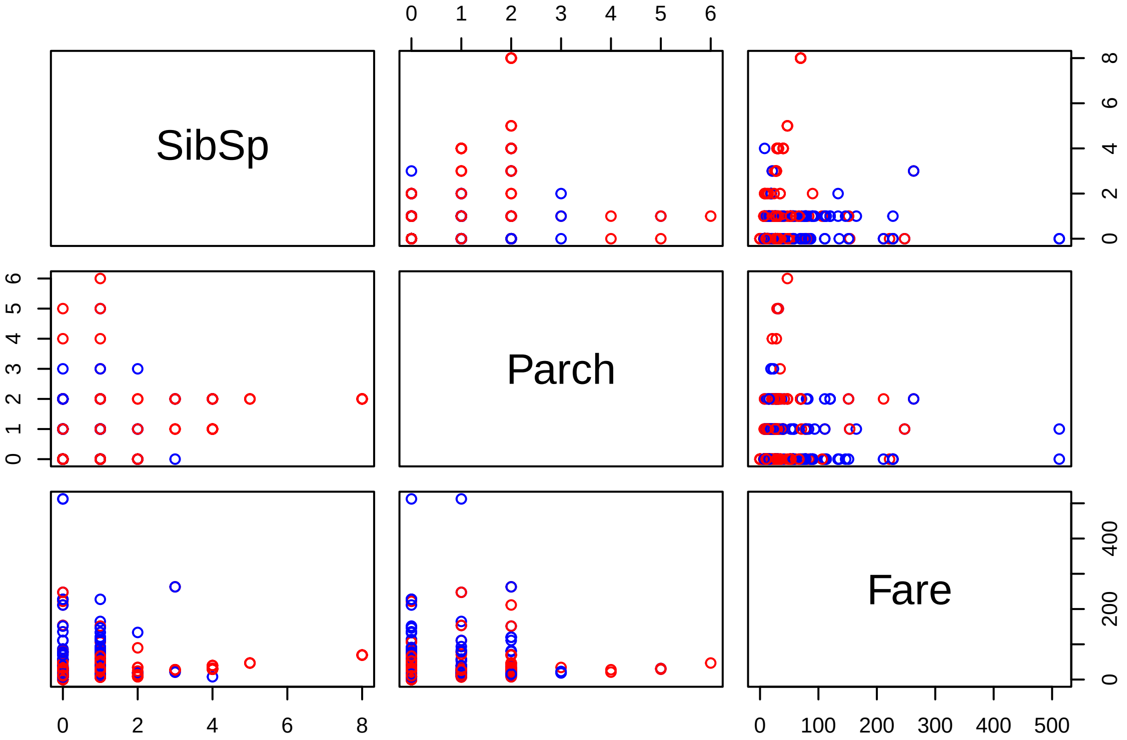
From this boxplot, one can conclude that the the fare of the passenger did seem to marginally affect your chances of survival. As for those who survived, the data was concentrated slightly farther upward indicating that maybe the more your ticket costs the higher chances of your survival.

So in conclusion, all predictors excluding than the number of sibilings and spouses of the passenger appear to affect your chances of surviving. Passengers who were women, in a better class, who payed more for their ticket, were traveling to Cherbourg, or had more parents and children on board with them seemed to have a better chance of surviving.

EDA on Classification pairs

Now we will look at the quantitative variables that might be useful in classifying if a passenger lived or died. To do this, we will analyze a pairs plot.

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



In the pairs plot above, we can see pairs of quantitative variables could be useful in determining for a random passenger if they lived (blue) or died (red). To determine if they are useful, we can look for clear separation between the red and the blue data points. Unfortunately, in any combination, there is no clear division between who survived and who died so it appears that the 3 quantatative variables will not be useful in our model.

Modeling

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)

For our LDA, we will only use the quantitative variables (Parch, Fare, SibSp). We will now build it on the training data:

```
titanic.lda <- lda(factor(Survived) ~ Parch + Fare + SibSp,
data = titanic_train)
```

Now we will test it on the test data:

```
titanic.lda.pred <- predict(titanic.lda,
as.data.frame(titanic_test))
table(titanic.lda.pred$class, titanic_test$Survived)
```

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

```
149+83+12+23
```

```
## [1] 267
```

```
95/267
```

```
## [1] 0.3558052
```

```
12/161
```

```
## [1] 0.07453416
```

```
83/106
```

```
## [1] 0.7830189
```

Out of the 267 passengers from the sample, the model overall incorrectly predicted 95 passengers, a (12+83)/267, 35.6% error rate. For those that died, the model produced an error rate of 12/161, 7.5%. For those that survived, the model produced an error rate of 83/106, 78.3% error rate. A 35.6% error is not desirable for a classification model so we will keep looking for a better model in predicting life or death.

For our QDA, we will only use the quantitative variables (Parch, Fare, SibSp). We will now build it on the training data:

```
titanic.qda <- qda(factor(Survived) ~ Parch + Fare + SibSp,
data = titanic_train)
```

Now we will test it on the test data:

```
titanic.qda.pred <- predict(titanic.qda,
as.data.frame(titanic_test))
table(titanic.qda.pred$class, titanic_test$Survived)
```

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

```
(15+73)/267
```

```
## [1] 0.329588
```

```
15/161
```

```
## [1] 0.0931677
```

```
73/106
```

```
## [1] 0.6886792
```

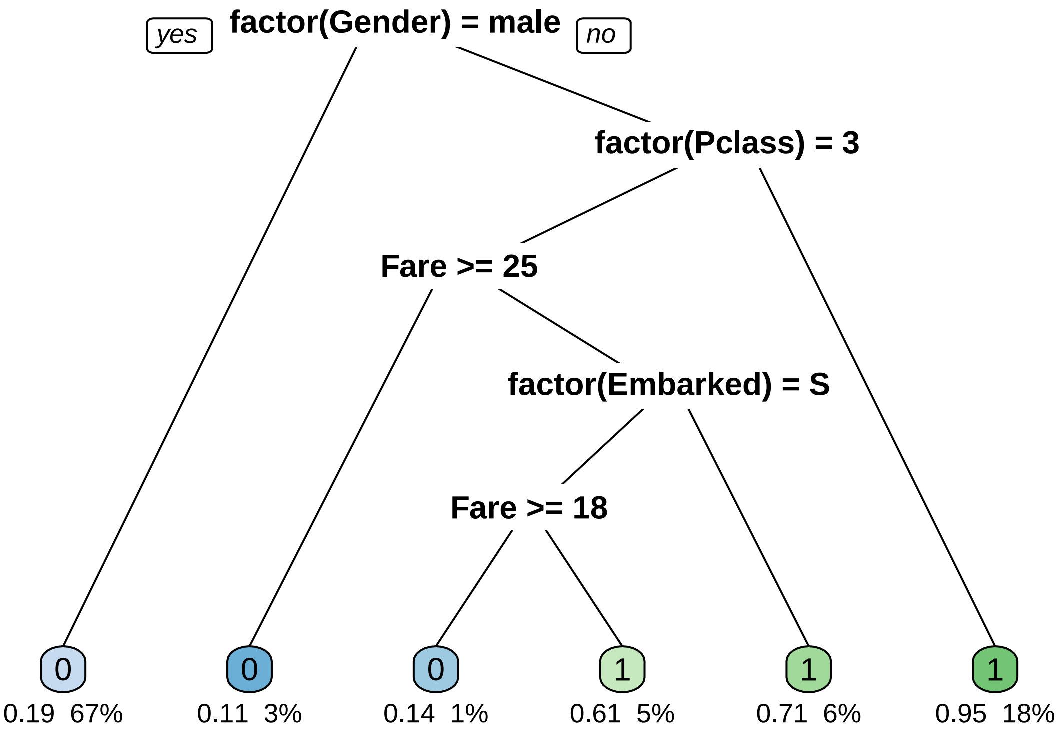
Out of the 267 passengers from the sample, the model overall incorrectly predicted 88 passengers, a (15+73)/267, 33% error rate. For those that died, the model produced an error rate of 15/161, 9.3%. For those that survived, the model produced an error rate of 73/106, 68.9% error rate. So, in conclusion, the QDA model preformed slightly better than the LDA. But still not that effective in predicting life or death.

All in all, the limitations of only using the quantatative variables seem to be very restrictive in accurately predicting survival. In both the models, the error rate for surviving was over 65% making it a very ineffective and inaccurate model.

Classification Trees

While the categorical variables are excluded in the LDA and QDA classifiers, using classification trees we can include them (Pclass, Embarked, Gender). We will use the training data to fit it:

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



Not surprisingly, the model takes full advantage of the categorical variables. As we discovered in our individual variable EDA the categorical variables seemed to have a stronger affect on the response variable. Fare also seemed to have a relationship with Survival. The model excluded SibSp (Which showed no relationship) and Parch (Which showed a slight relationship). It is no surprise that gender was the first classifier as the division between men and women was the most striking followed by the class.

We can now test it on the titanic testing data:

```
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
```

```
(20+32)/267
```

```
## [1] 0.1947566
```

```
20/161
```

```
## [1] 0.1242236
```

```
32/106
```

```
## [1] 0.3018868
```

Out of the 267 passengers from the sample, the model overall incorrectly predicted 52 passengers, a (32+20)/267, 19.4% error rate. For those that died, the model produced an error rate of 20/161, 12.4%. For those that survived, the model produced an error rate of 32/106, 30.2% error rate. So, in conclusion, the classification model preformed signifigantly better than the QDA and LDA due to inclusion of the cateogircal variables. The model appears to not have overfit as there is still some error. All in all, the model preformed well but still not exceptional in predicting life or death.

Binary Logistic Regression

Finally, our last hope will be use binary logistic regression to build an effective model which can include quantitative and categorical variables alike. We will first train on the training dataset.

```
titanic.logit1 <- glm(factor(Survived) ~ Parch + Fare + SibSp + factor(Gender) + factor(Embarked) + fac
data = titanic_train,
family = binomial(link = "logit"))

titanic.logit2 <- glm(factor(Survived) ~ Fare + factor(Gender) + factor(Embarked) + factor(Pclass),
data = titanic_train,
family = binomial(link = "logit"))
```

We can now test it on the other set of data:

```
titanic.logit1.prob <- predict(titanic.logit1,
as.data.frame(titanic_test),
type = "response")

titanic.logit2.prob <- predict(titanic.logit2,
as.data.frame(titanic_test),
type = "response")
```

Since the logistic regression model produces probabilities not sorting the data into life or death, we will use a threshold probability of .5 to classify them. So depending on if the data is > or < .5, the model will decide if the passenger would live or die. To make sure the probability is synced with the reponse categories we will run levels on the model and then convert it into a confusion matrix.

```
levels(factor(titanic_test$Survived))
```

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

```
titanic.logit1.pred <-ifelse(titanic.logit1.prob > 0.5,"0","1")
table(titanic.logit1.pred, titanic_test$Survived)
```

```
##
## titanic.logit1.pred   0   1
##                   0  30  76
##                   1 131  30
```

```
(131+76)/267
```

```
## [1] 0.7752809
```

```
131/161
```

```
## [1] 0.8136646
```

```
76/106
```

```
## [1] 0.7169811
```

From the confusion matrix, we can see that the model did not preform well. The model produced an overall error of 207 passengers, 77.5% error. For those that died, 131 passengers or 81.3% error rate, and for those that lived 76 passengers or 71.7% error rate. All in all, the model preformed the worst by far and would not be effective or accurate in predicting life or death.

```
levels(factor(titanic_test$Survived))
```

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

```
titanic.logit2.pred <-ifelse(titanic.logit2.prob > 0.5,"0","1")
table(titanic.logit2.pred, titanic_test$Survived)
```

```
##
## titanic.logit2.pred   0   1
##                   0  40  83
##                   1 121  23
```

```
(121+83)/267
```

```
## [1] 0.7640449
```

We also attempted at the logistic regression again with a new model which excludes Parch and SibSp, the two variables quantitative variables that appeared to have little or no relationship with the response, and again the overall error rate is 76.4%. So it is clear that logistic regression is not the best model.

Final Recommendation From the 4 models we tested (LDA, QDA, Classification Tree, and Logistic Regression), the Classification tree model was the most effective and accurate by far. It produced the lowest overall error rate of 19.4% which all things considered is not bad considering that sinking of the titanic was chaotic and killed those of all genders, backgrounds, and socioeconomic status. It also achieved the lowest error rate for survivors which the other models struggled to accomplish. Over fitting does not seem to be a concern as the error rate was still somewhat high so the model was no near a perfect fit. LDA and QDA models preformed similarly with an overally error rate of around 35% and the logistic regression model preformed horribly with around 77.5% error rate. Our final recommendation is the classification tree.

Discussion

Looking back, the models did not preform well in predicting life or death for passengers on the Titanic. LDA and QDA were forced to use only the quantitative variables which as we saw with our 1 variable EDA did not appear to be good predictors for survival. The classification tree faired better as it included those variables and used only the ones it saw fit.

We believe that if a more accurate model were to be created, a variable which I think would be possibly useful is age as younger people might have been prioritized compared older more middle aged passengers. We believe that the 19.4% error rate for our chose model, the classification tree, despite the fact that it is relatively high for predicting survival, the Titanic is a unique event in the sense that its tragedy did not discriminate. While there are obviously some factors that were significant in predicting survival such as gender or class, at the end of the day, people of all sorts were killed in the disaster.

We believe that if a more accurate model were to be created, a variable which I think would be possibly useful is age as younger people might have been prioritized compared older more middle aged passengers.