Evaluating Classifier Performance for Predicting Passenger Survival Aboard the Titanic

Joseph Ouyang josephou

Due Monday, November 24, at 11:59PM

Contents

Introduction

The sinking of the RMS Titanic is one of the most well-known maritime disasters in history. Despite being advertised as an “unsinkable” luxury liner, the Titanic struck an iceberg and sank on its maiden voyage in 1912, resulting in the deaths of more than 1,500 passengers and crew. In this paper, we analyze demographic and travel-related characteristics of Titanic passengers and apply machine learning classification techniques to predict survival outcomes. Using physical, demographic, and ticket-related features (such as class, gender, number of relatives aboard, fare, and port of embarkation), we train and evaluate several classification models. The dataset used in this study originates from the Titanic passenger records curated by Frank Harrell (Department of Biostatistics, Vanderbilt University).

Exploratory Data Analysis

Description of Variables

Referencing the Titanic passenger records, we analyze a random sample of 889 passengers. In order to train and test our classification models, we split this random sample into a training dataset of 622 passengers and a testing dataset of 267 passengers. Summaries of the variables are listed below:

Below are the following predictor vairables:

- `Class`: ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- `Gender`: male or female
- `SibSp`: number of siblings and spouses of the individual who are aboard the Titanic
- `Parch`: number of parents and 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)

Below are the response labels we want our classifiers to predict:

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

Response Label Overview

Referencing our training set of 622 observations, we observe 234 passengers who survived comprising 37.62 percent of the passengers, and 388 passengers who died comprising 62.38 percent of the passengers, as shown in the following tables:

Counts Table:

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

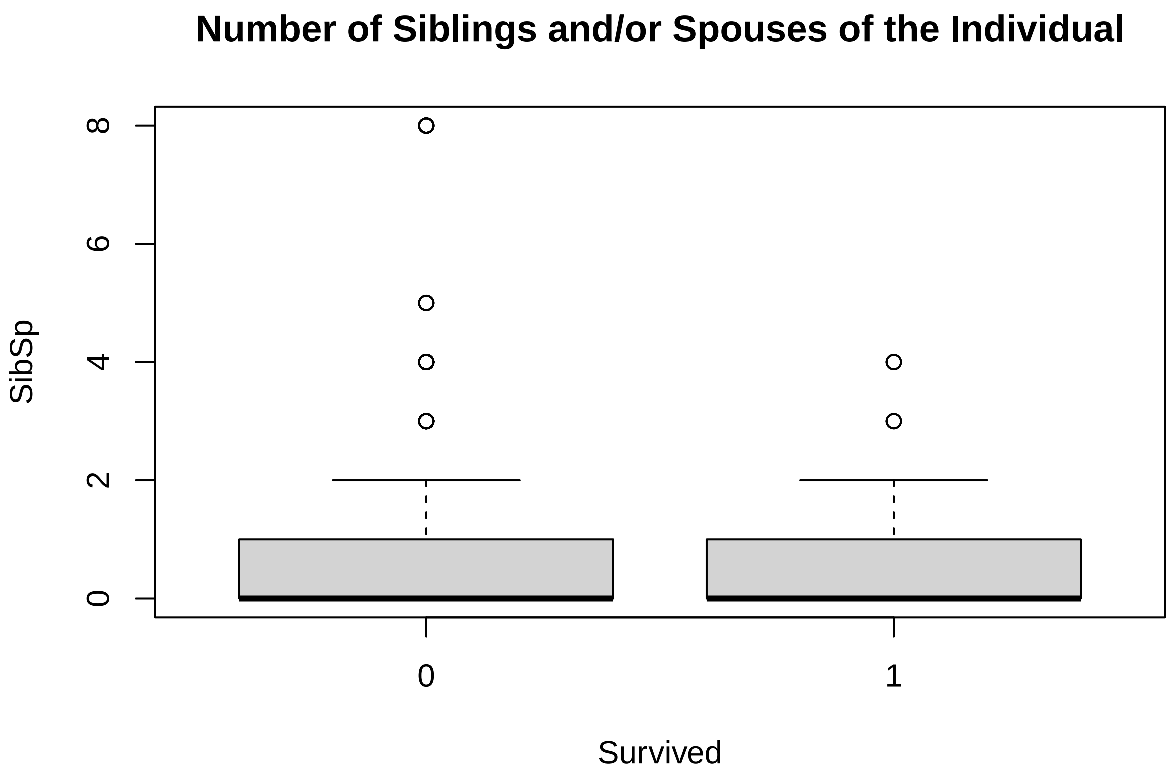
Proportions Table:

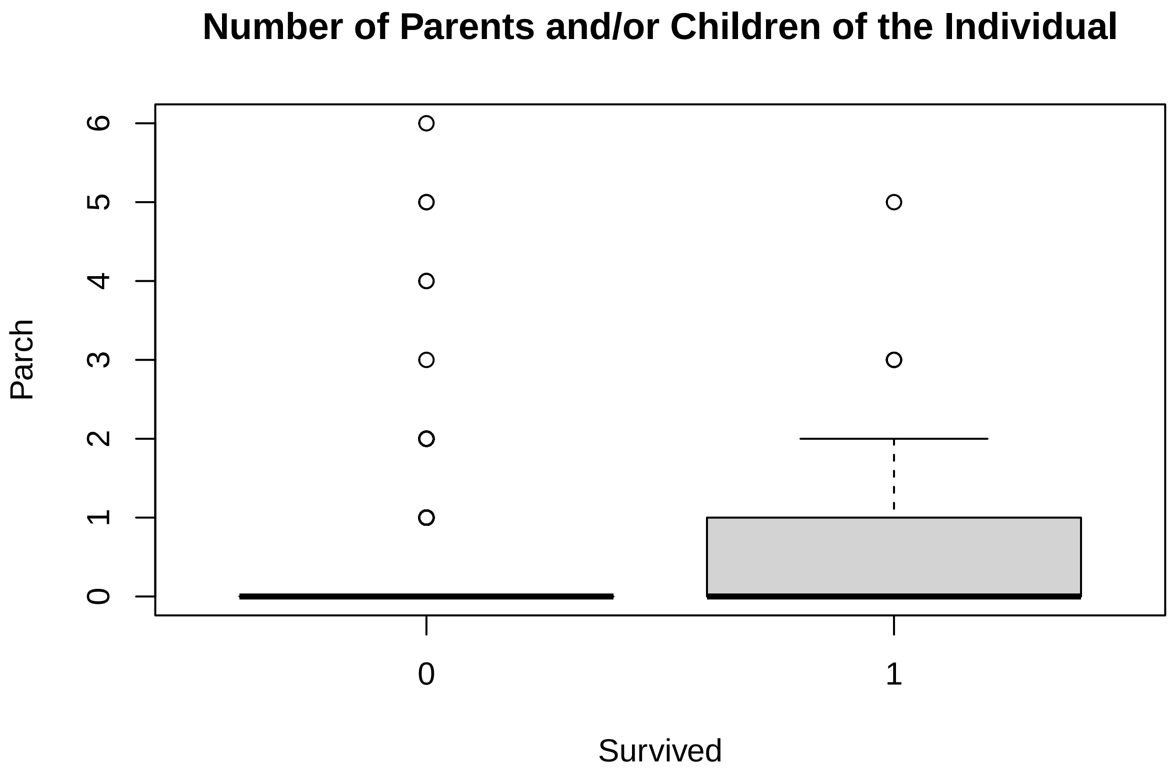
```
##
##         0         1
## 0.6237942 0.3762058
```

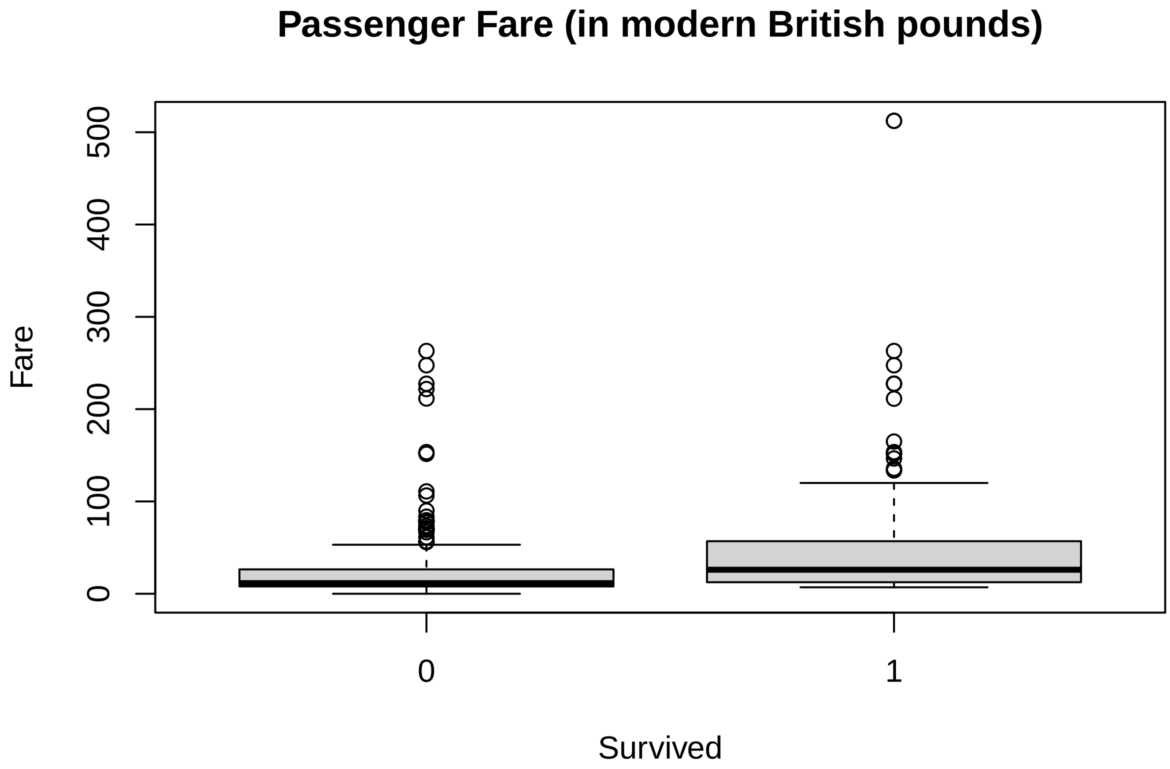
EDA on Relationships between `Survived` and the Quantitative Variables

After displaying the counts and proportions of each response label, we then move toward visualizing the relationship between the response, `Survived`, and the quantitative explanatory variables (demographic and ticket-related information).

The boxplots depicted below allow us to visually explore the variables and determine which predictors to be most useful in classifying whether or not each passenger survived.


Number of Siblings and/or Spouses of the Individual
SibSp
0
2
4
6
8
0
1
Survived


Number of Parents and/or Children of the Individual
Parch
0
1
2
3
4
5
6
0
1
Survived



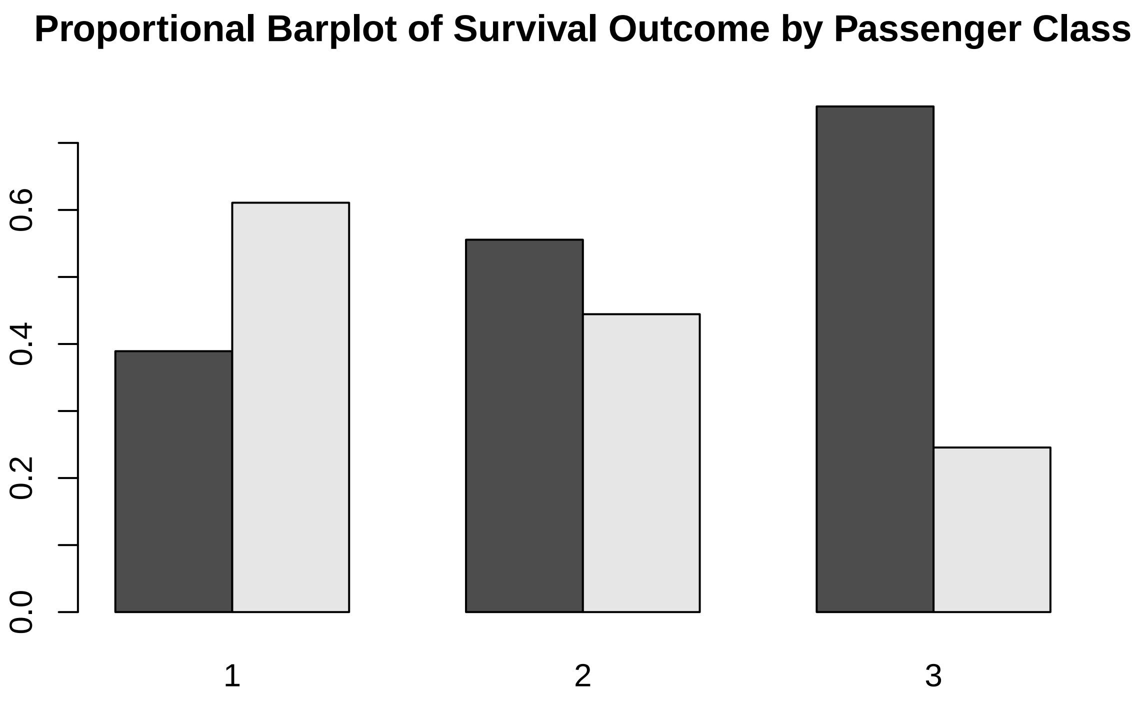
Referencing the above boxplots, we observe that some display differences between those who survived (1) and those who did not survive (0), and others do not display differences. For those that do, the differences tell us that we have some evidence of a relationship and a predictor variable that might be useful in our classifiers. Taking that into account, we note that those who paid a more expensive passenger fare had a higher survival rate, on average, when compared to those who paid a relatively cheaper passenger fare. On the other hand, there does not seem to be a relationship between number of siblings and/or spouses of the individual and survival outcome, and number of parents and/or children of the individual and survival outcome at first glance when referencing the respective boxplots. The medians of the two response labels are roughly the same for both relationships. However, it is important to note that a wider spread of number of parents and/or children of the individual (ranging from 0 to 2, median to Q4 respectively) exists for those that survived, which is in contrast to those who didn't survive being concentrated at 0 (median). This is also the case for passenger fare. Interestingly, when inspecting the first two boxplots displayed above, there seems to be more high outliers (in this context, those with more relatives) for the passengers who did not survive.

EDA on Relationships between `Survived` and the Categorical Variables

In order to visualize the relationship between `Survived` and the categorical variables of `Class`, `Gender`, and `Embarked`, we can analyze the conditional proportions of `Survived` conditioned on each of the categorical variables. The visualizations of these relationships are depicted below as barplots.

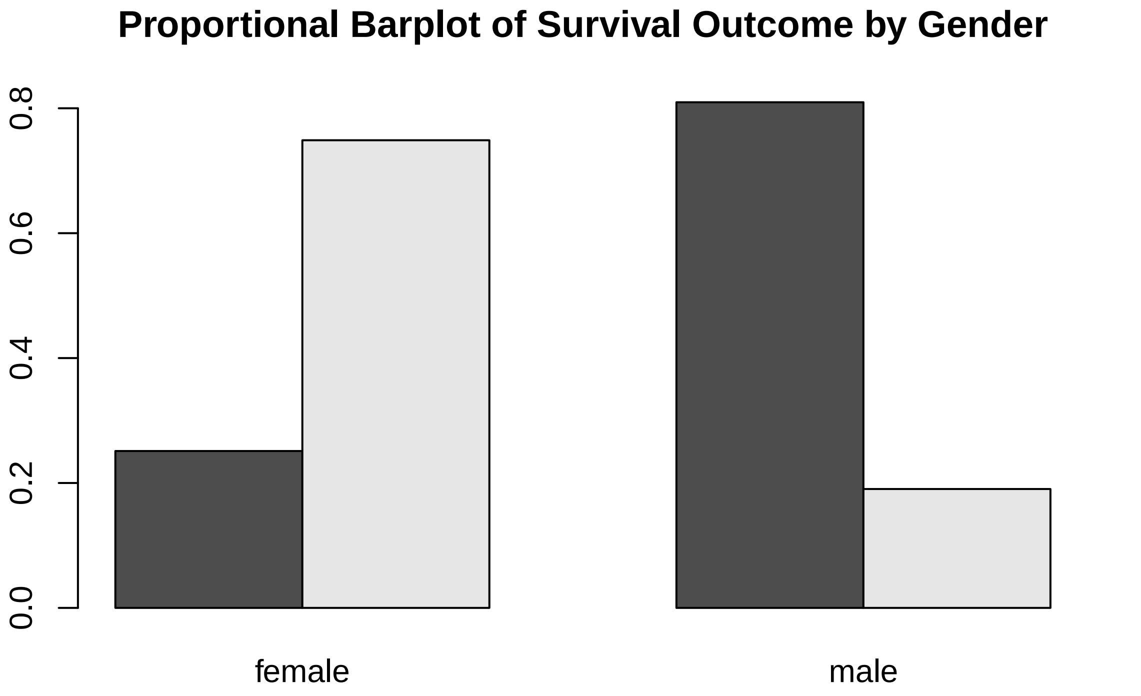
First, is the conditional proportions table for the variable `Class` followed by its respective barplot:

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



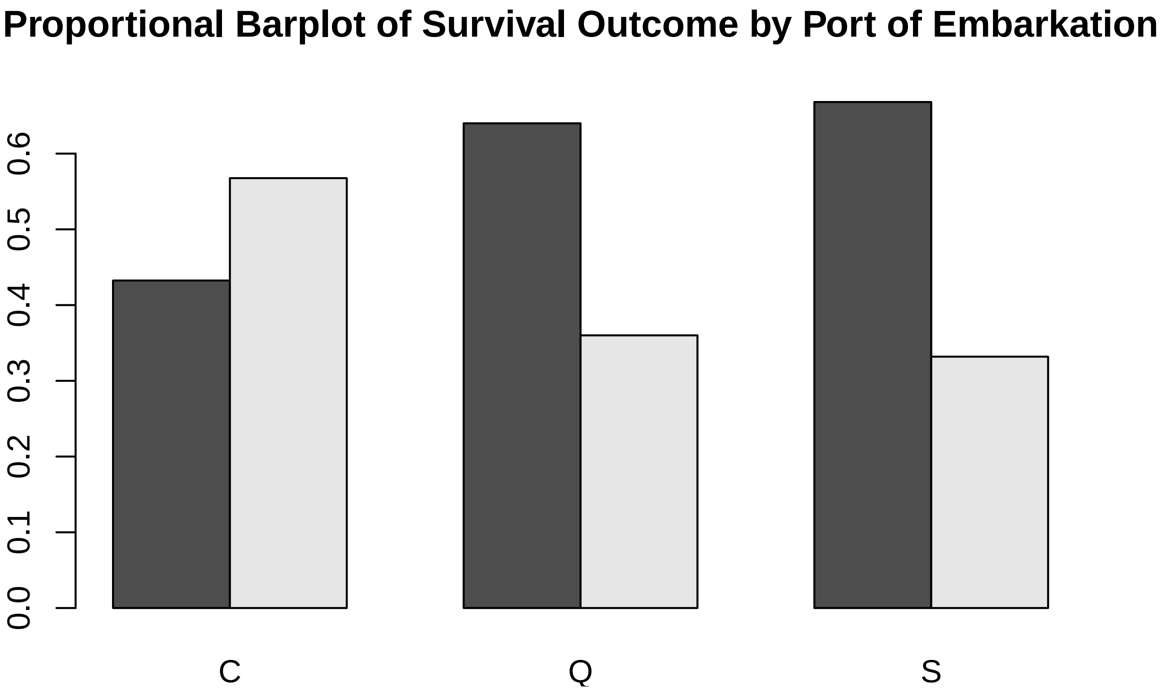
Second, is the conditional proportions table for the variable `Gender` followed by its respective barplot:

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



Third, is the conditional proportions table for the variable `Embarked` followed by its respective barplot:

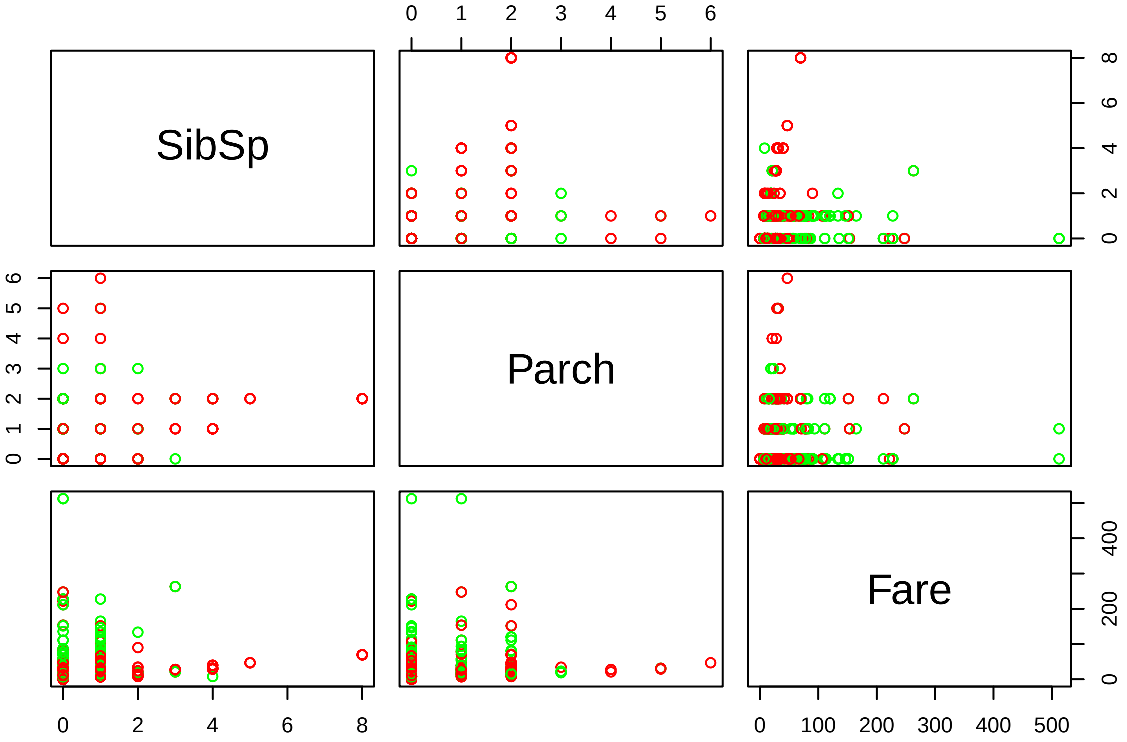
```
##
##             C         Q         S
##   0 0.4324324 0.6400000 0.6681128
##   1 0.5675676 0.3600000 0.3318872
```



Referencing the summaries above, it appears that the higher the passenger class (1st class being highest and 3rd class being the lowest), the likelihood of survival increases. The second barplot shows that females are more likely than males to survive the Titanic. Passengers who embarked from Cherbourg's port were shown to be more likely to survive than those who embarked from Queenstown and Southampton, with both showing roughly equal likelihoods of survival.

EDA on Classification Pairs

Finally, to explore how pairs of quantitative predictors might contribute to classifying survival outcome, we visualize their bivariate relationships in a labeled pairs plot shown below:



In the pairs plot above, we get a two-dimensional look at how different combinations of quantitative predictors might help distinguish survivors (green points) from non-survivors (red points). There are some pairs of variables that exhibit no relationship with each other; namely the pair `Fare` and `Parch` and the pair `Fare` and `SibSp`. One many expect that larger families pay higher total fares, but it is important to note that the dataset records individual fare, not family fare. Furthermore, most points cluster at low family sizes which makes finding clear separations between those who survived and those who did not survive difficult at first glance. However when we observe the variable `Fare`, those who survived tended to pay a higher fare with slightly more green points at the respective higher ends. It is also worth to note one individual high outlier that paid over 400 British pounds who survived probably because their wealth gave them priority to escape the Titanic safely. Overall, the pairs plot does not reveal any clear separations for `Parch` and `SibSp`, with the exception of those paying higher fares tending to survive more.

We should note, however, that these plots only show one or two variables at a time, and the true relationship between the predictors and survival likely exists in a higher-dimensional space that is more complex than what we can visualize here.

Modeling

We now shift our focus to building and evaluating classifiers for predicting passenger survival. The four methods we consider are: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), classification trees, and binary logistic regression.

To guard against overfitting and ensure that our results generalize beyond the sample, we randomly split the dataset into a training set and a test set. Each of the four models was trained on the same training subset and evaluated on the same test subset for a fair comparison.

Linear Discriminant Analysis (LDA)

For our LDA and QDA models, we use only the continuous variables, those being `SibSp`, `Parch`, and `Fare`. The LDA classifier is built on the training data which is shown below:

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

We next examine the test-set performance of the LDA classifier as shown below in the confusion matrix:

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

On the test data, LDA gave an overall error rate of (12+83)/267 = 0.3558, which is moderately high. More specifically, we observe the error rate at finding those who survived to be 83/(83+23) = 0.783, which is particularly high. Conversely, we observe the error rate at finding those who did not survive to be 12/(12+149) = 0.0745, which is relatively low.

Quadratic Discriminant Analysis (QDA)

Similarly, we use the same quantitative variables to train our QDA classifier as shown below:

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

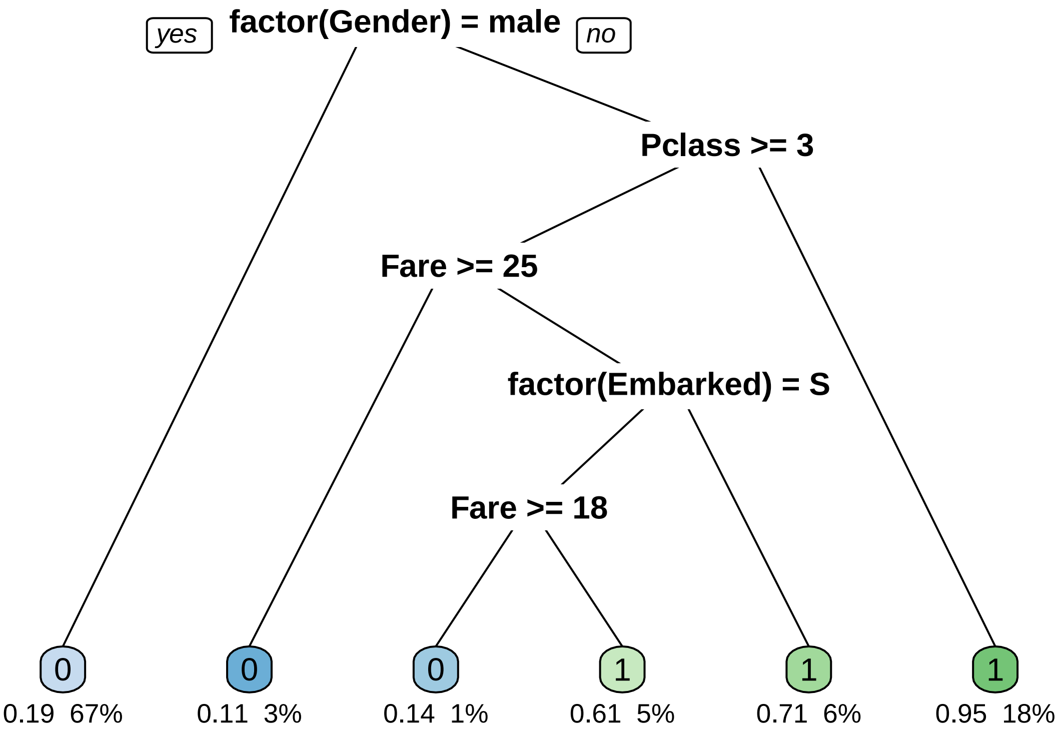
We next examine the test-set performance of the QDA classifier as shown below in the confusion matrix:

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

Because QDA can capture nonlinear, curved decision boundaries, we might expect it to perform somewhat better than LDA. As expected, the confusion matrix showed a slight decrease in error rate: (15+73)/267 = 0.3296. Interestingly, we observe the error rate for classifying those who survived decreased to 73/(73+33) = 0.689. On the other hand, the error rate for classifying those who did not survive increased to 15/(15+146) = 0.09317.

Classification Trees

While our LDA and QDA classifiers can only incorporate the quantitative predictors, a classification tree allows us to also include categorical variables (`Class`, `Gender`, and `Embarked`). We fit a classification tree on the training data and display its structure as follows:



The classification tree shows that gender is the main driver of survival, with men sent to a node where the estimated survival rate is only about 0.19, and this group represents roughly two thirds of the passengers. Among those who are not male, passenger class is the next key feature. Women in first or second class are grouped together and have a very high survival probability of about 0.95. For women in third class, both fare and port of embarkation matter. Those who did not embark from Southampton have relatively good outcomes, with survival around 0.7, while those from Southampton generally do worse, especially at the extremes of fare, where survival is closer to about 0.1 to 0.15. Overall, the tree highlights that being female and in a higher class is strongly associated with survival, and among third class women, both fare and whether they embarked from Southampton play an important role in their chances.

We next evaluate how well the tree classifier performs on our test data, as shown in the confusion matrix below:

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

On the test data, the classification tree gave an overall error rate of (20+32)/267 = 0.1948, which is lower than the overall error rates of the LDA and QDA models. More specifically, we observe the error rate at finding those who survived to be 32/(32+74) = 0.3019, which is a drastic decrease when compared to that from the LDA and QDA models. Conversely, we observe the error rate at finding those who did not survive to be 20/(20+141) = 0.1242, which is somewhat higher relative to that from the LDA and QDA models.

Binary Logistic Regression

Finally, we turn to binary logistic regression to model passenger survival. As with the classification tree, this method allows us to use all predictors, including categorical variables such as `Class`, `Gender`, and `Embarked`. We fit the logistic regression model on the training data and then examine its performance on the test set using the resulting confusion matrix, as shown below:

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

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

Because the logistic model applied to the test set produces probabilities rather than direct survival or non-survival labels, we turn these probabilities into class predictions by setting a cutoff. If the predicted probability is greater than 0.5, we classify the passenger as a survivor, and otherwise as a non-survivor. To make sure we interpret these probabilities in the correct direction, we first need to check how the response variable `Survived` is ordered as a factor. We do this by inspecting the levels of the factored response using levels().

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

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

We then obtain test classification from the logistic model using a threshold probability of 0.5, as shown below:

```r
titanic.logit.pred <-ifelse(titanic.logit.prob > 0.5,1,0)
```

To see how well our logistic classifier performed on our test data, we visualize a confusion matrix as shown below:

```
##
## titanic.logit.pred   0   1
##                    0 128  29
##                    1  33  77
```

The logistic classifier (threshold probability of 0.5) performed slightly weaker than the classification tree, with an overall error rate of (33+29)/267 = 0.2322. For classifying those who survived, the error rate was 29/(29+77) = 0.2736, which is slightly lower than that of the classification tree. For classifying those who did not survive, the error rate was 33/(33+128) = 0.2049, which is significantly higher than that of the classification tree.

We note that as for LDA, QDA, and the classification tree, the logistic classifier performed better on those who did not survive.

Final Recomendation

Of the four classifiers we tested, the best performance was shown through the classification tree model. A close second was the logistic classifier, which had a slightly higher overall error rate ($0.2322 > 0.1948$).

The performance of the LDA and QDA models were roughly similar as their overall error rates were 0.3558 and 0.3296 respectively, with QDA performing slightly better.

Our final recommendation is the classification tree, with the lowest overall error rate of 0.1948. More specifically, the error rate for classifying those who survived was one of the lowest (0.3019), with only the logistic classifier performing slightly better (0.2736). Given the low overall test error rate and the tree's clear, intuitive depiction of how the predictors relate to survival, we chose the classification tree as our final model.

Discussion

Overall, our classifiers were able to predict survival reasonably accurately using information on class, gender, family size, fare, and port of embarkation. Among the models we fit, the classification tree achieved a low test error and provided a clear and intuitive description of how these predictors relate to survival, highlighting patterns such as the strong disadvantage faced by men and lower class passengers. Because it balances predictive accuracy with interpretability, we selected the classification tree as our final model for making predictions on new Titanic passengers.

For future work, we could explore random forests as a way to address possible overfitting in a single tree and to stabilize predictions by averaging across many trees. It would also be interesting to apply clustering methods to see whether passengers naturally group into clusters with different survival patterns that are not fully captured by the existing variables. In addition, incorporating extra information such as age and cabin location could give a more detailed and stable picture of which combinations of characteristics most strongly influenced survival and might reveal important interactions that our current models cannot detect.