Comparing Different Titanic Classifiers for Best Prediction

Yeonji Baek ybaek

Due Fri, April 21, at 11:59PM

Contents

Introduction **1**

Exploratory Data Analysis **1**
Dataset Background . . . 1
Univariate EDA on Response Variable (Training Set) . . . 2
Bivariate EDA on relationship between Survived and Quantitative Variable . . . 2
Bivariate EDA on relationship between Survived and Categorical Variable . . . 4
Bivariate EDA on classification pairs . . . 6

Modeling **7**
Linear Discriminant Analysis (LDA) . . . 7
Quadratic Discriminant Analysis (QDA) . . . 8
Classification Trees . . . 8
Binary Logistic Regression . . . 9
Final Recommendation . . . 10

Discussion **10**

Sources **10**

Introduction

[1] The 'unsinkable' luxury cruise liner Titanic famously sank on its first voyage. Because of this tragic incident, many people were sorrowful because they never knew that it was coming. It is important to figure out the trend and which type of passengers died due to this incident so that we can better develop the factors that caused passengers to die. In our project, we are going to be training with datasets and perform our machine learning classification skills to classify whether or not the passenger survived or not.

Exploratory Data Analysis

Dataset Background

From the titanic data provided, we observe that the training data provides 622 random samples of the Titanic's passengers along with 7 different variables. The titanic data is from Frank Harrell, professor in the department of Biostatistics at Vanderbilt University. [2] "The titanic data frame does not contain information from the crew, but it does contain actual ages of half of the passengers." Because we are interested in predicting the survival status with our different binary classifiers, we have to analyze the relationship of survival status (Survived), which is the response variable, with six different predictor variables: Class, Gender, SibSp, Parch, Fare, and Embarked.

The descriptions of each **predictor** variables are listed below:

Variable	Description
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)

The description of the **response** variable that we want to predict with different classifiers is presented below:

Variable	Description
Survived	Survived (1) or dead (0)

The first first few lines of data are presented below as an example:

```
## # A tibble: 6 x 7
##   Survived Pclass SibSp Parch  Fare Embarked Gender
##      <dbl>  <dbl> <dbl> <dbl> <dbl> <chr>    <chr>
## 1        0      3     0     2 20.2  S        female
## 2        0      1     0     0 52    S        male
## 3        1      3     1     0 11.2  C        female
## 4        0      3     0     0  8.71 C        male
## 5        0      2     0     0 10.5  S        male
## 6        1      2     0     1 33    S        female
```

Univariate EDA on Response Variable (Training Set)

We will first give a summary of our response variable in the training set, which includes 622 observations in the dataset:

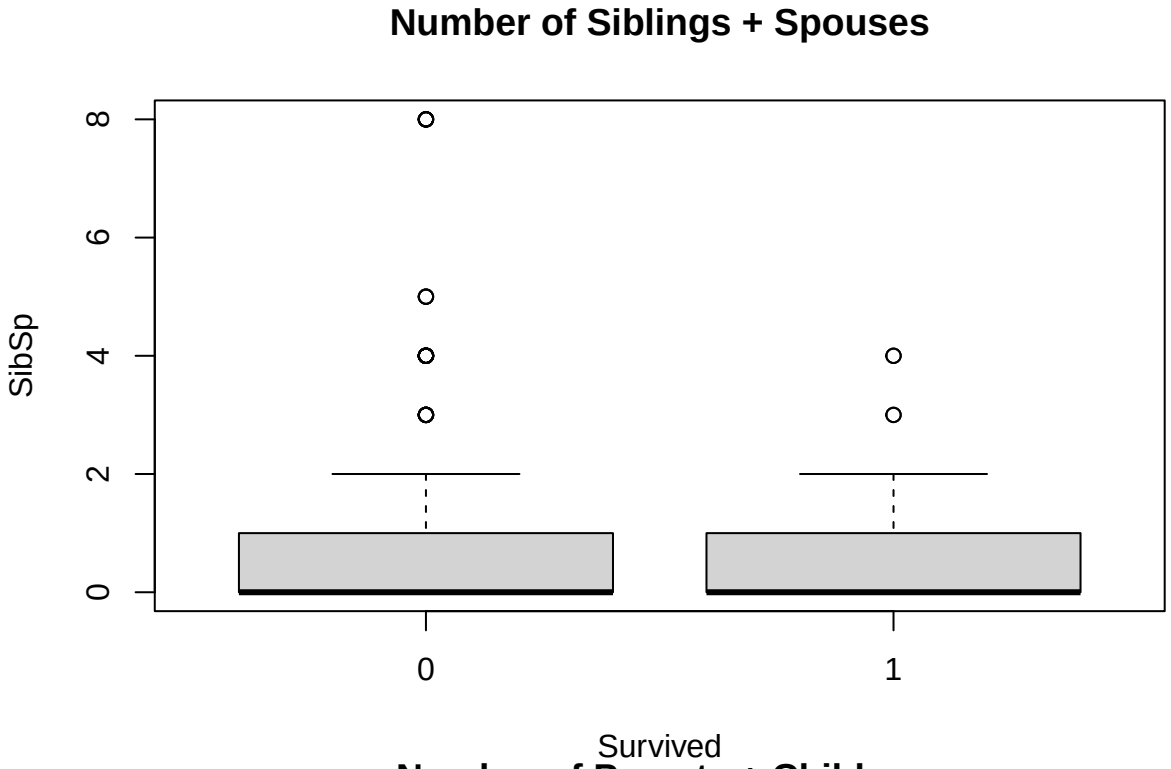
```
##
##   0   1
## 388 234
```

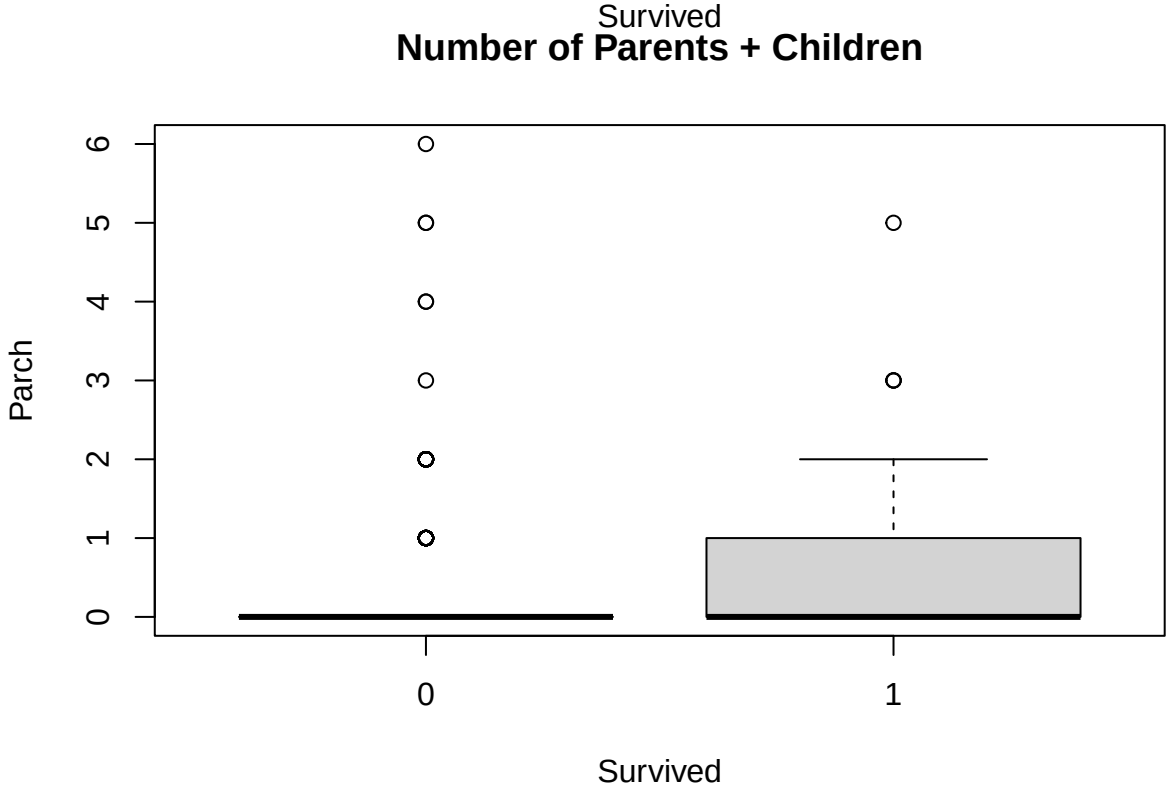
```
##
##         0         1
## 0.6237942 0.3762058
```

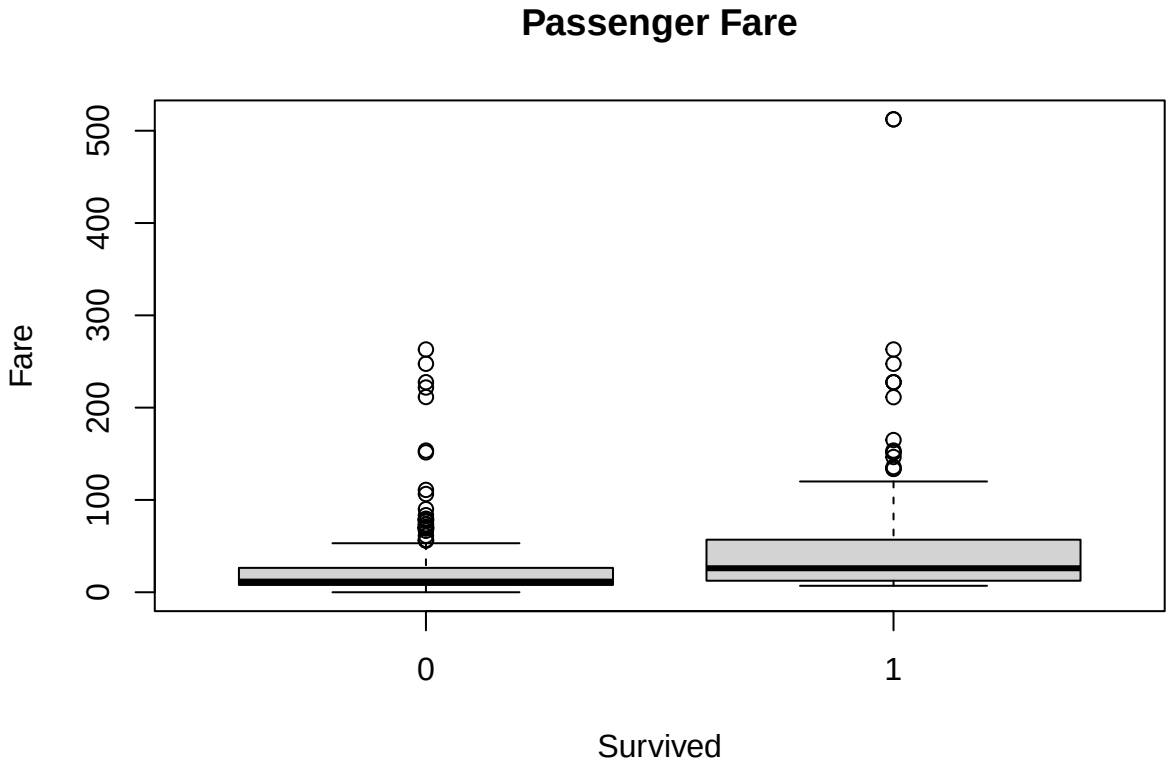
From the two tables, we see that there are 388 Titanic passengers who died (which makes up 62.38% of the passengers) and 234 Titanic passengers who survived (which make up 37.62% of the passengers).

Bivariate EDA on relationship between Survived and Quantitative Variable

We will now illustrate the relationship between Survived and the **quantitative** predictor variables we have. With quantitative variables, we will be utilizing boxplots since boxplots are the most helpful when it comes to visualizing the relationship:


Number of Siblings + Spouses
SibSp
0
2
4
6
8
0
1
Survived


Number of Parents + Children
Parch
0
1
2
3
4
5
6
0
1
Survived



From the boxplots above, we notice that there are not a lot of differences in SibSp (Number of siblings + spouse) other than the outliers. There are four different outliers from the dead status with individuals who had higher number of SibSp, going up to 8 SibSp. For Parch (Number of parents + children), the boxplot seems like it is implying that lower number of parents and children result in surviving. For passenger fare, we see that passengers who paid more fare had higher chance of survival. The survived (1) boxplot seems slightly higher than the dead (0) boxplot with the minimum, IQR, median, and even maximum. However, because of the multiple outliers in the dead boxplot, it is not clear since even many of the passengers who paid higher fare died.

Bivariate EDA on relationship between Survived and Categorical Variable

We will now illustrate the relationship between Survived and the **categorical** predictor variables we have. With categorical variables, we will be utilizing the conditional proportions of Survived. Each of the conditional proportions of Survived are conditioned on each categorical variable (Pclass, Gender, and Embarked):

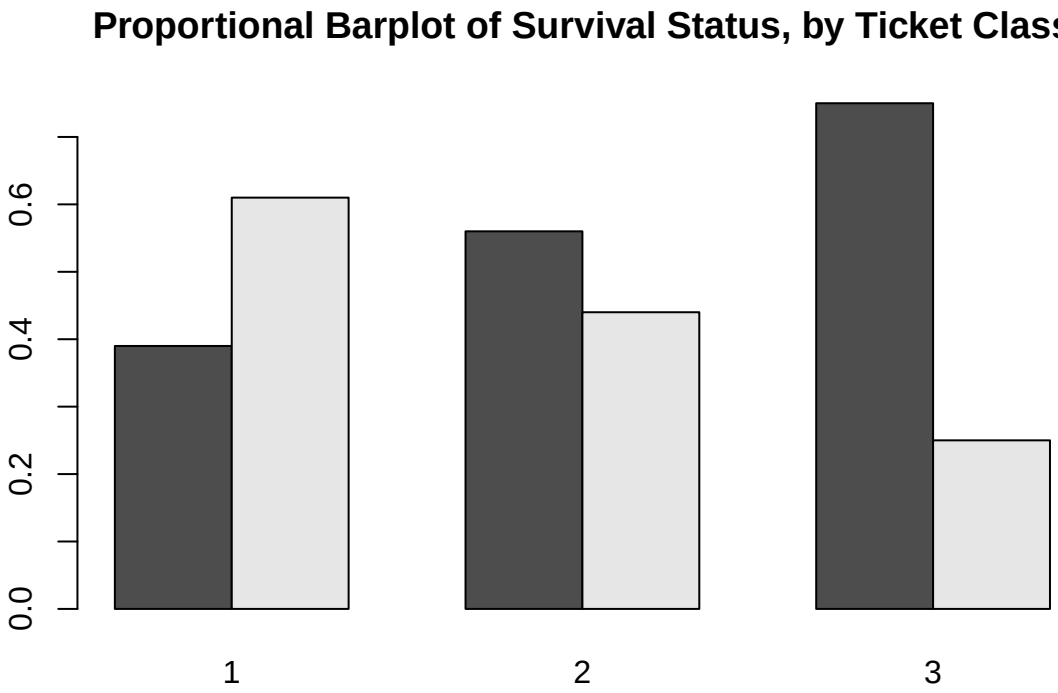
```
##
##        1    2    3
##   0 0.39 0.56 0.75
##   1 0.61 0.44 0.25
```

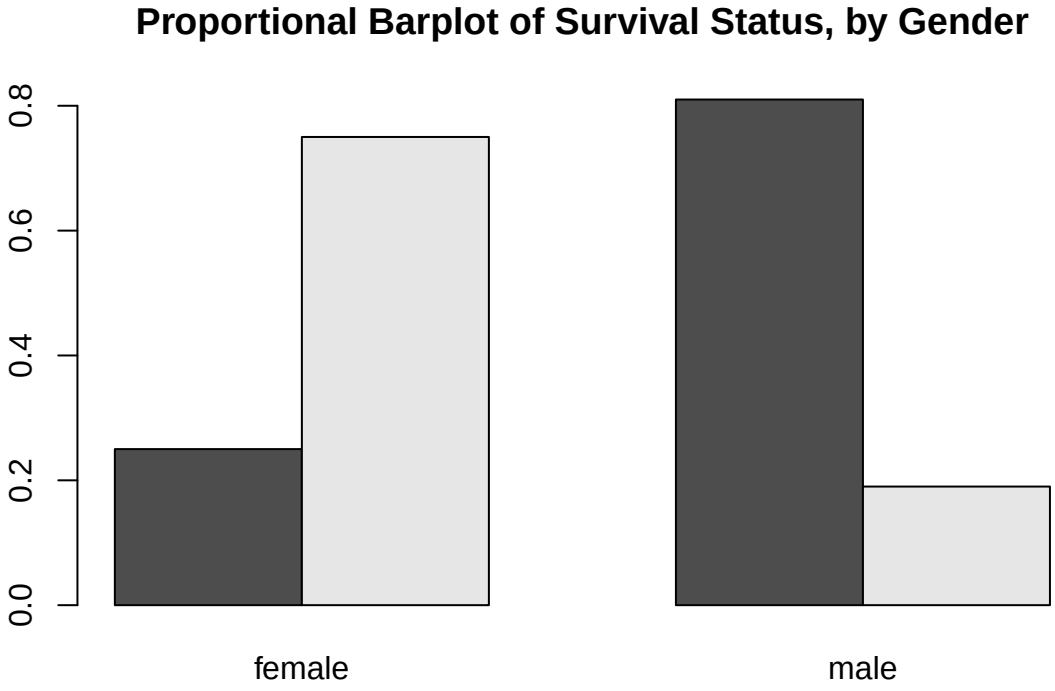
```
##
##     female male
##   0   0.25 0.81
##   1   0.75 0.19
```

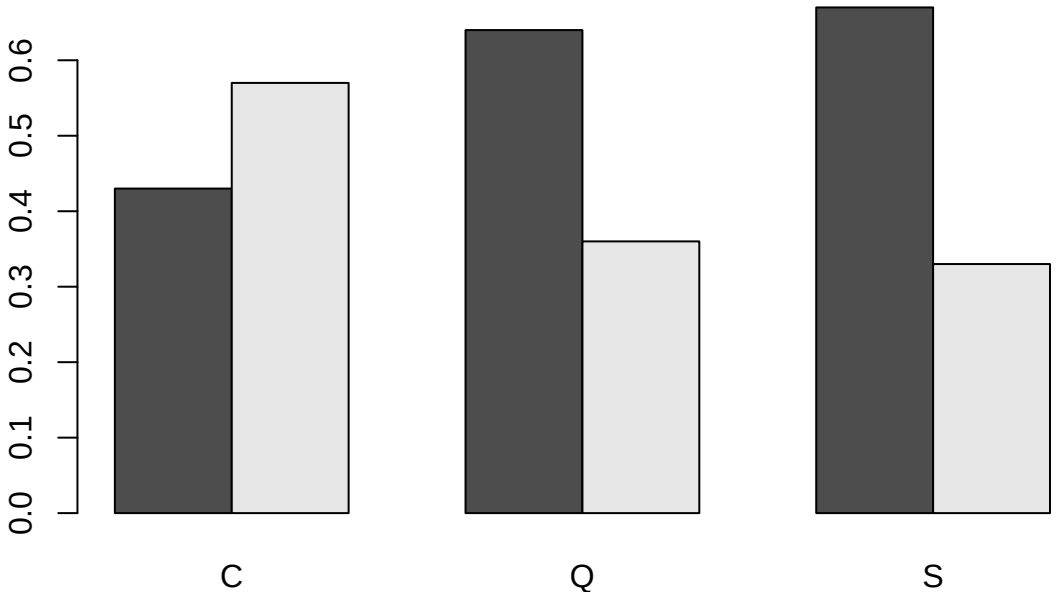
```
##
##        C    Q    S
##   0 0.43 0.64 0.67
##   1 0.57 0.36 0.33
```

We will graph each of the tables into a barplot to visualize the different proportions better. The barplots are shown below:

(*Note that black bars imply dead while grey bars imply survived.*)







After taking a look at the conditional proportions table and barplots, we notice these three things:

1) With lower class tickets (3rd class), more passengers are dead while more 1st class passengers survived (compared to lower class passengers).

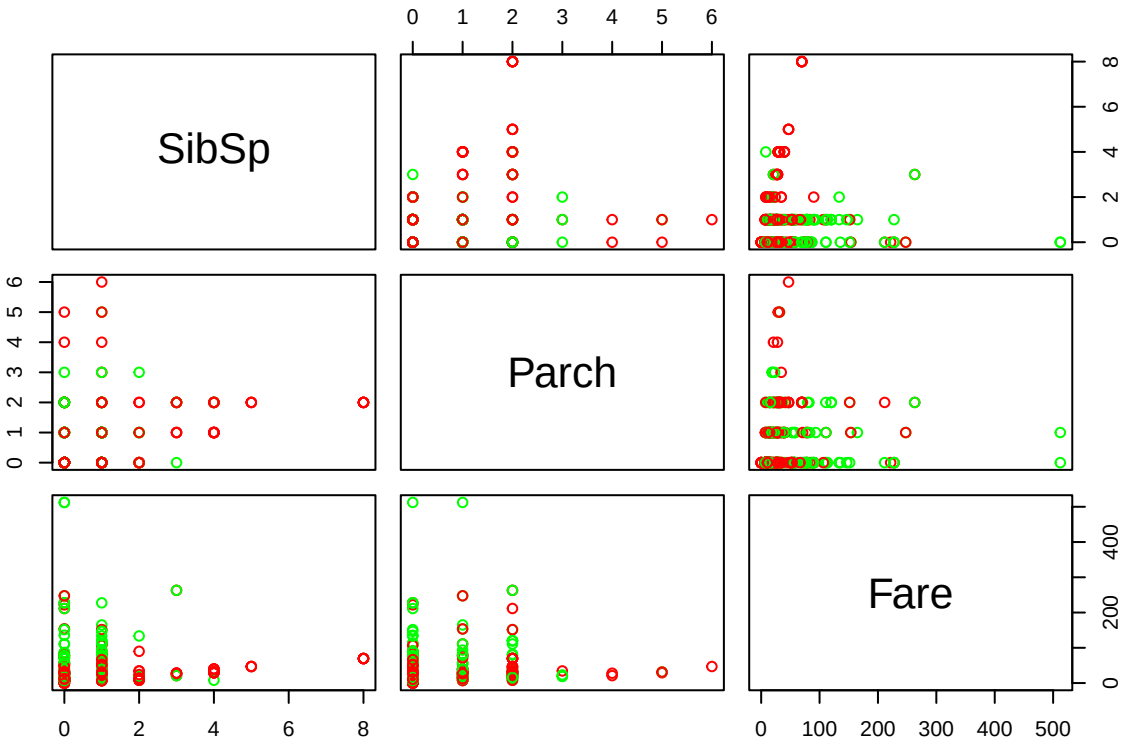
2) Male passengers were more likely to die while female passengers were more likely to survive.

3) More passengers who embarked from Port Southampton died compared to Cherbourg and Queenstown. On the other hand, more passengers who embarked from Port Cherbourg survived compared to other two port of embarkation.

Bivariate EDA on classification pairs

Lastly, we will be checking the different pairs of quantitative predictors to visualize which pairs might be helpful in classifying Survived. To do so, we will perform pairs plot:

Note that red means dead while green means survived.



From the pairs plot, we can slightly see which plots are mingled together and which groups can be separated into sections. For example, Parch and Fare are very mixed togther which do not illustrated a good separation. However, SibSp and Fare combination shows a plot that can seem easier to separate between survived and dead.

Modeling

Now that we examined all the variables and their relationships, it is time to build our classifiers for predicting the survival status of the Titanic passengers. We need to check the four different classifiers which are linear discriminant analysis, quadratic discriminant analysis, classification trees, and binary logistic regression.

Since we need to make sure that overfitting the dataset is not one of our problem, we are going to be splitting our observations into training sets and test sets randomly. Note that all the classifiers are utilizing the same training sets and testing it on the same test sets.

Linear Discriminant Analysis (LDA)

First, we will be only using quantitative variables, which are SibSp, Parch, and Fare, for our Linear Discriminant Analysis on the training data.

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

Now, we will examine how good our LDA classifier performs on the test dataset given to us.

```
survived.lda.pred <- predict(survived.lda,
                             as.data.frame(titanic_test))

table(survived.lda.pred$class, titanic_test$Survived)
```

```
##
##        0   1
```

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

From the test data, we can calculate the overall error rate of LDA, which is (12+83)/267 = 0.3558. We see that LDA has an error rate for classifying survived passenger of 83/106 = 0.783 while the error rate for dead passenger is 12/161 = 0.074. So, we see that LDA is best at finding dead passengers.

Quadratic Discriminant Analysis (QDA)

Second, we will be only using quantitative variables, which are SibSp, Parch, and Fare, for our Quadratic Discriminant Analysis on the training data.

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

Now, we will examine how good our QDA classifier performs on the test dataset given to us.

```
survived.qda.pred <- predict(survived.qda,
                             as.data.frame(titanic_test))

table(survived.qda.pred$class, titanic_test$Survived)
```

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

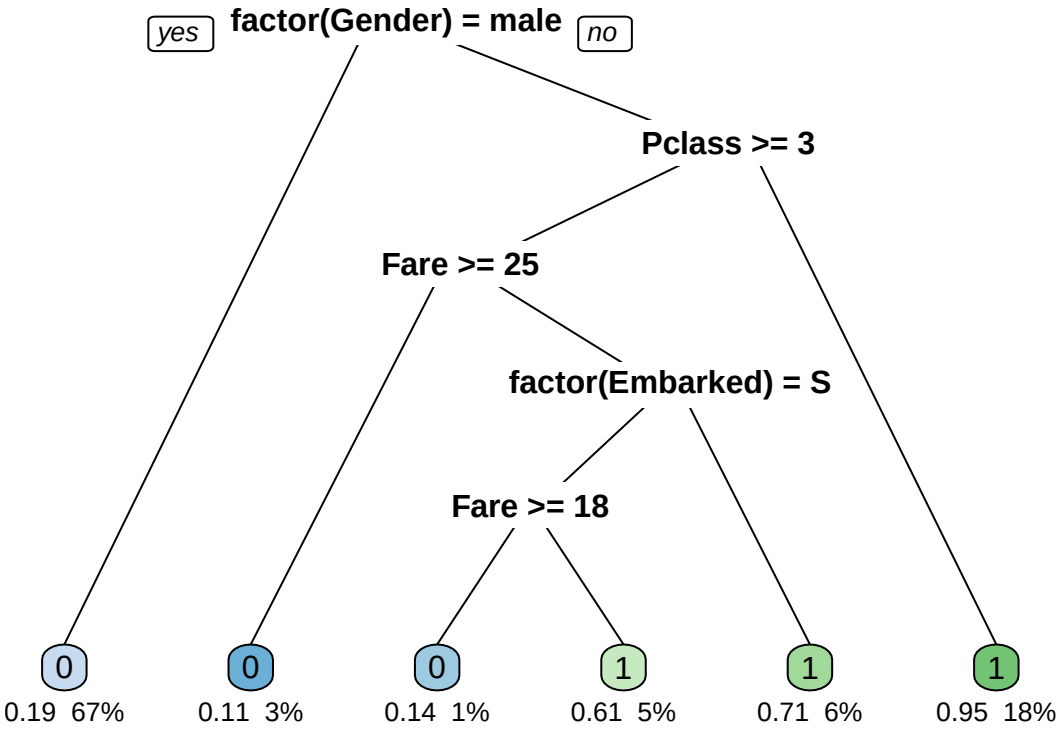
From the test data, we can calculate the overall error rate of QDA, which is (15+73)/267 = 0.3296, which is slightly better performance than LDA. For classifying specific survival status, QDA is also better at classifying dead passengers (15/161 = 0.093) compared to classifying alive passengers (73/106 = 0.6887).

Classification Trees

For Classification Trees, it is slightly different from LDA and QDA because we can also use categorical variables in this. We will be using all the variables for our Classification Trees on the training data.

```
survived.tree <- rpart(factor(Survived) ~ Pclass + SibSp +
                       Parch + Fare + factor(Embarked) + factor(Gender),
                       data=titanic_train,
                       method="class")

rpart.plot(survived.tree,
           type = 0,
           clip.right.labs = FALSE,
           branch = 0.1,
           under = TRUE)
```



In the above classification tree, we see that the the tree only selected Gender, Pclass, Fare, and Embarked to classify the survival status of the passengers. Now, we will examine how good our Classification Tree performs on the test dataset given to us.

```
survived.tree.pred <- predict(survived.tree,
                              as.data.frame(titanic_test),
                              type="class")

table(survived.tree.pred, titanic_test$Survived)
```

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

From the test data, we can calculate the overall error rate of Classification Tree, which is $(20+32)/267 = 0.1947$. For classifying specific survival status, Classification Tree is also better at classifying dead passengers $(20/161 = 0.124)$ compared to classifying alive passengers $(32/106 = 0.302)$.

Binary Logistic Regression

For Binary Logistic Regression, it is also slightly different from LDA and QDA because we can also use categorical variables in this. We will be using all the variables for our Binary Logistic Regression on the training data.

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

Now, we can use this to check the test data.

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

We will classify the passenger as dead if the probability is greater than 0.5 and classify as alive if less than 0.5.

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

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

```
survived.logit.pred <-ifelse(survived.logit.prob > 0.5,"1","0")

table(survived.logit.pred, titanic_test$Survived)
```

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

From the test data, we can calculate the overall error rate of Binary Logistic Regression, which is (33+29)/267 = 0.2322. For classifying specific survival status, Binary Logistic Regression is also better at classifying dead passengers (33/161 = 0.205) compared to classifying alive passengers (29/106 = 0.273).

Final Recommendation

After testing four different classifiers, we found out that the classification tree was performing the best with the overall error rate of 0.1947. We also found out that all of the classifiers were presenting better performance with classifying dead passengers compared to alive passengers. Both LDA and QDA were slightly higher than overall rate of 30% while Binary Logistic Regression was slightly higher than 20%.

We would recommend Classification Tree for classifying the survival status of Titanic passengers.

Discussion

In conclusion, we decided on classification tree as a recommendation. For future research, it would be nice if we can retrieve some other data on the passengers to see if there were other factors that impacted their survival status. It is important to know these things because we never know when it might occur, so we need to be prepared.

Sources

[1] https://canvas.cmu.edu/courses/32842/files/folder/project2-materials/data-and-story-prompts?preview=9490299 (Titanic dataset information on Canvas)

[2] https://hbiostat.org/data/repo/titanic.html