Will Your Heart Really Go On? A Statistical Analysis of the Survivors on the Titanic

Manish Nagireddy
mnagired

Due Weds, November 25, at 8:00PM

Introduction

The Titanic is one of the most tragic events in history, with perhaps the only silver lining being that it brought us one of the greatest movies of all time. Nevertheless, what if there was a way to model who survived the crash based on a few arbitrary characteristics?

In this paper, we will train and evaluate machine learning classification techniques to predict whether someone survived or not, based on various metrics such as gender or ticket class.

[Data from Frank Harrell, Department of Biostatistics, Vanderbilt University, http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets]

Exploratory Data Analysis

Background and Variables

We have the following predictor variables:

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

and our response labels that we want to predict with our classifiers:

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

Summary of the Response Labels in the Training Dataset

We first note that in the training set, we have 622 observations, with 234 people surviving and 388 people dead. In other words, abour 38% of the people survived and 62% did not, as is shown in the following tables:

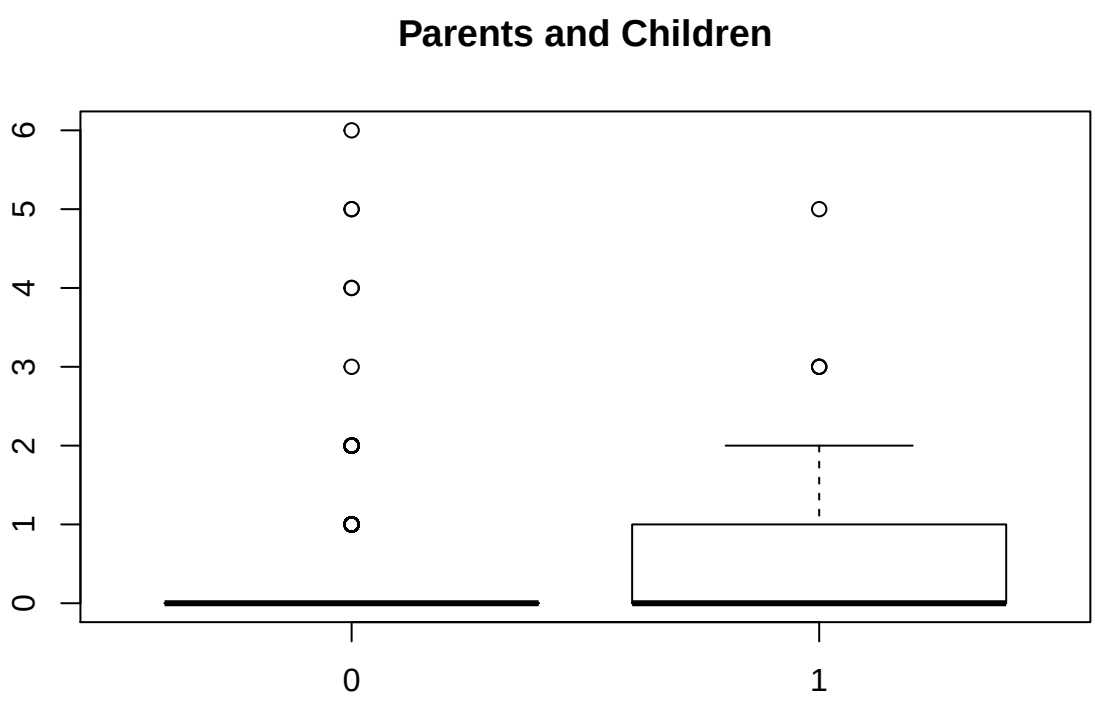
```
##
##   0   1
## 388 234
```

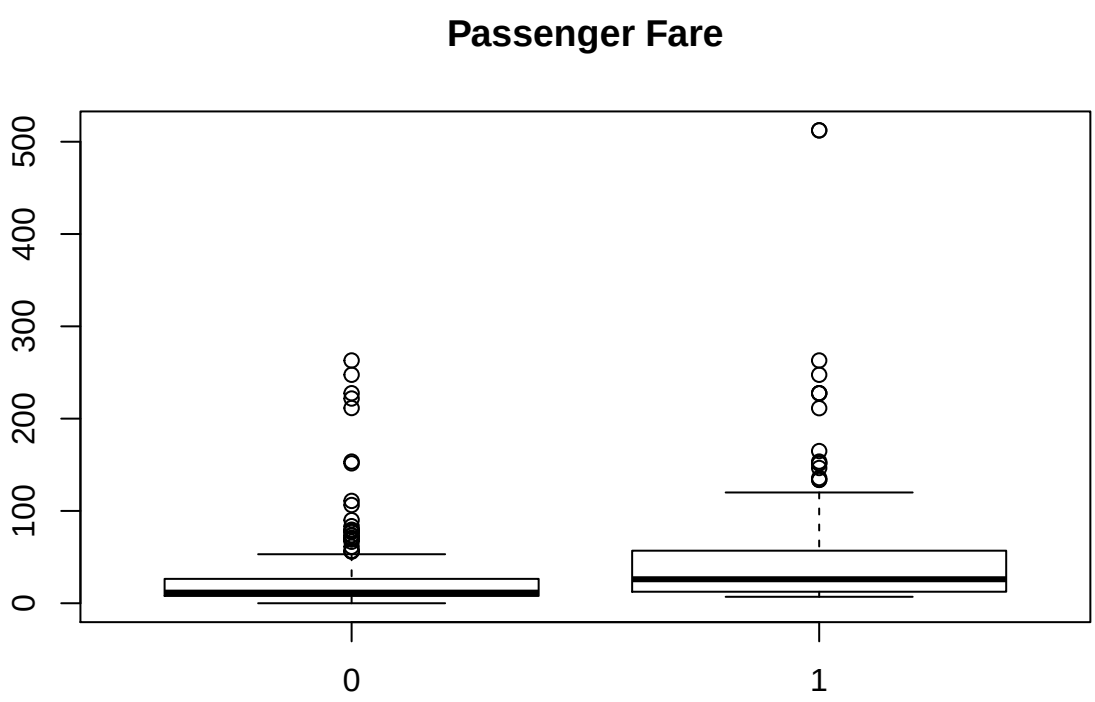
```
##
##         0         1
## 0.6237942 0.3762058
```

Some EDA on relationships between the response and the quantitative variables

Now, we will visualize the relationship between the response (`Survived`) and the various predictors (`Pclass`, `Gender`, `SibSp`, `Parch`, `Fare`, and `Embarked`).

In order to visually explore whether we expect the quantitative predictors to be useful in helping classify survivors, we show boxplots, which appear as follows:


Parents and Children
0
1
2
3
4
5
6
0
1



In the above boxplots, we note that if there are differences between those who survived and those who did not, we have evidence of a relationship and a variable that might be useful in our classifiers. To this end, we can see a slight difference in that the survivors appear to have a bit higher passenger fare. Also, the number of parents and children who survived is slightly larger than those who didn't. The number of siblings and spouses who survived appear to be roughly the same, however.

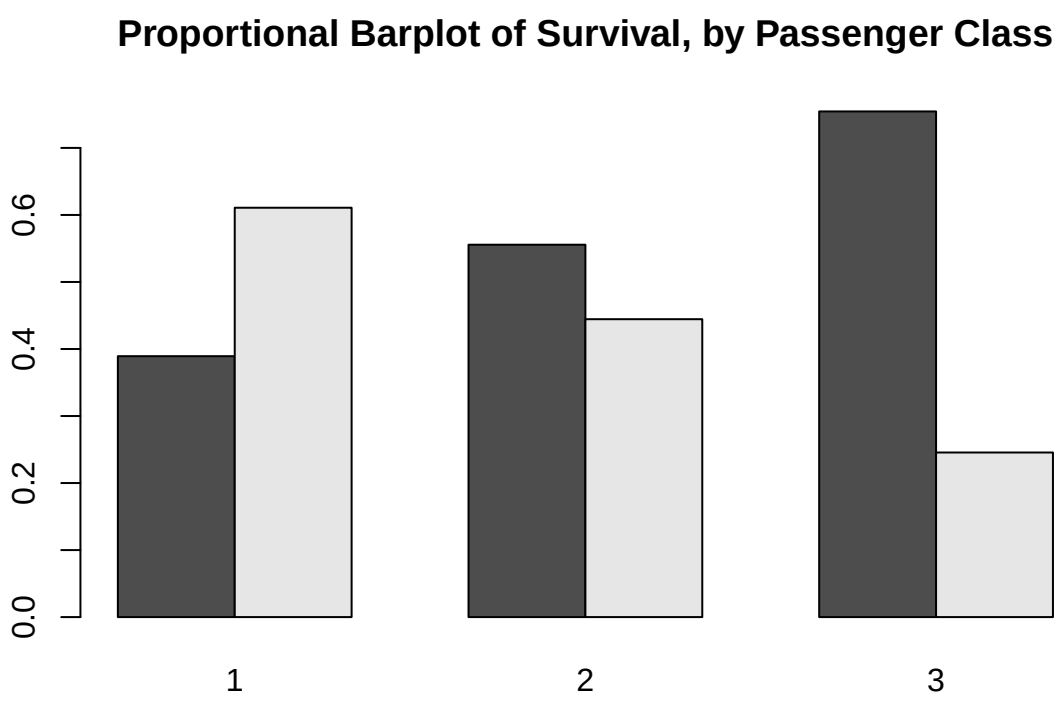
EDA on relationships between the response and categorical variables

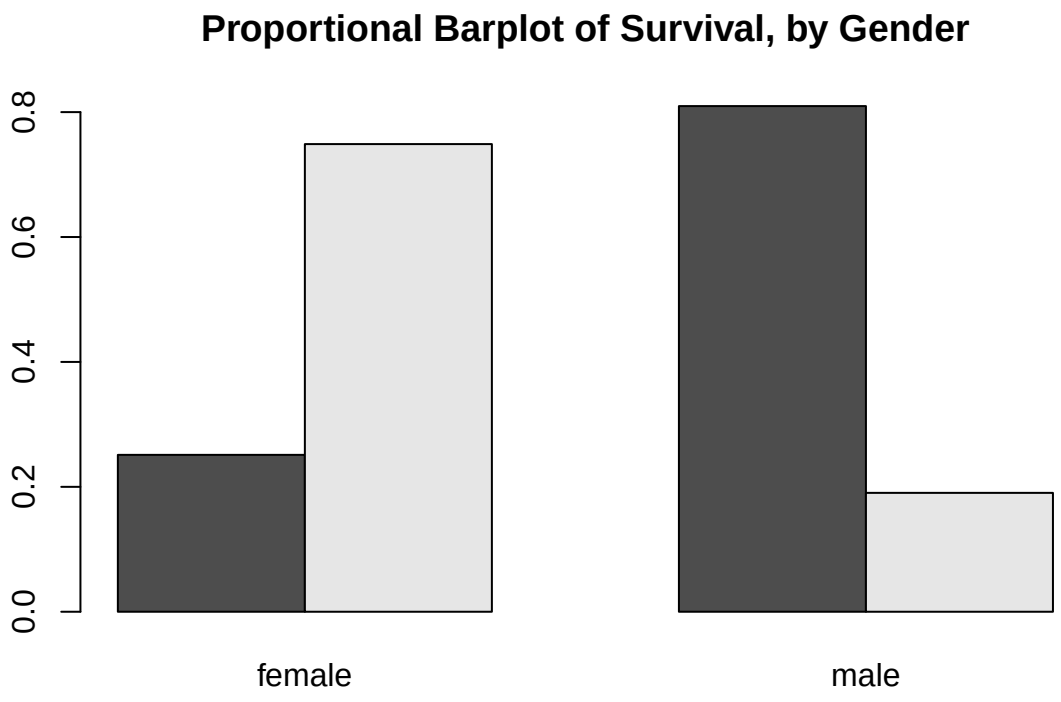
To explore the relationship between `Survived` and the categorical predictors `Pclass`, `Gender`, and `Embarked`, we can look at the conditional proportions of type, conditioned on quality, shown as follows:

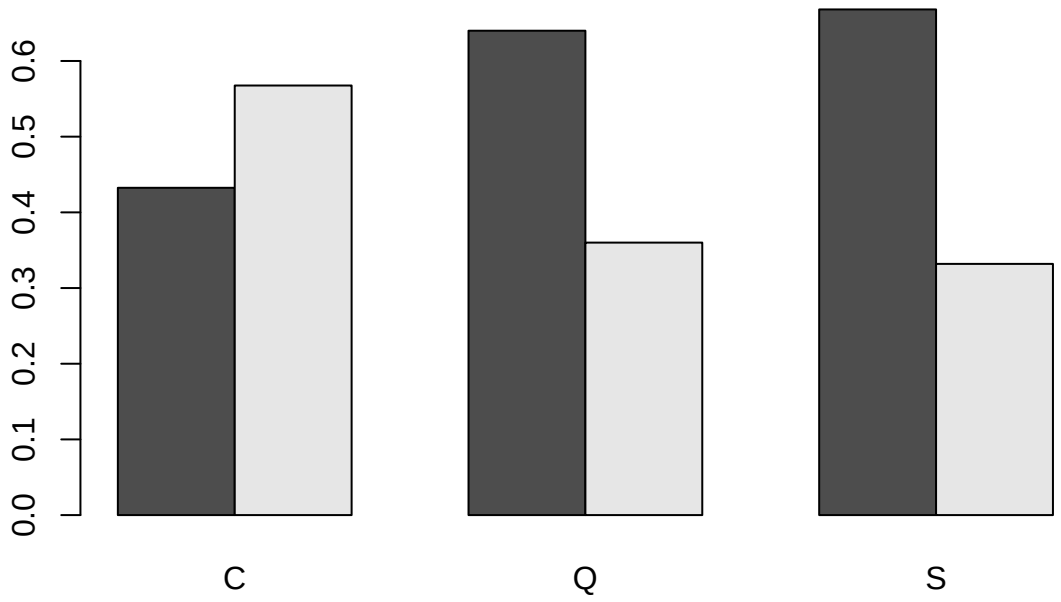
```
##
##            1         2         3
##   0 0.3892617 0.5555556 0.7544379
##   1 0.6107383 0.4444444 0.2455621
```

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

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


Proportional Barplot of Survival, by Passenger Class
0.0
0.2
0.4
0.6
1
2
3

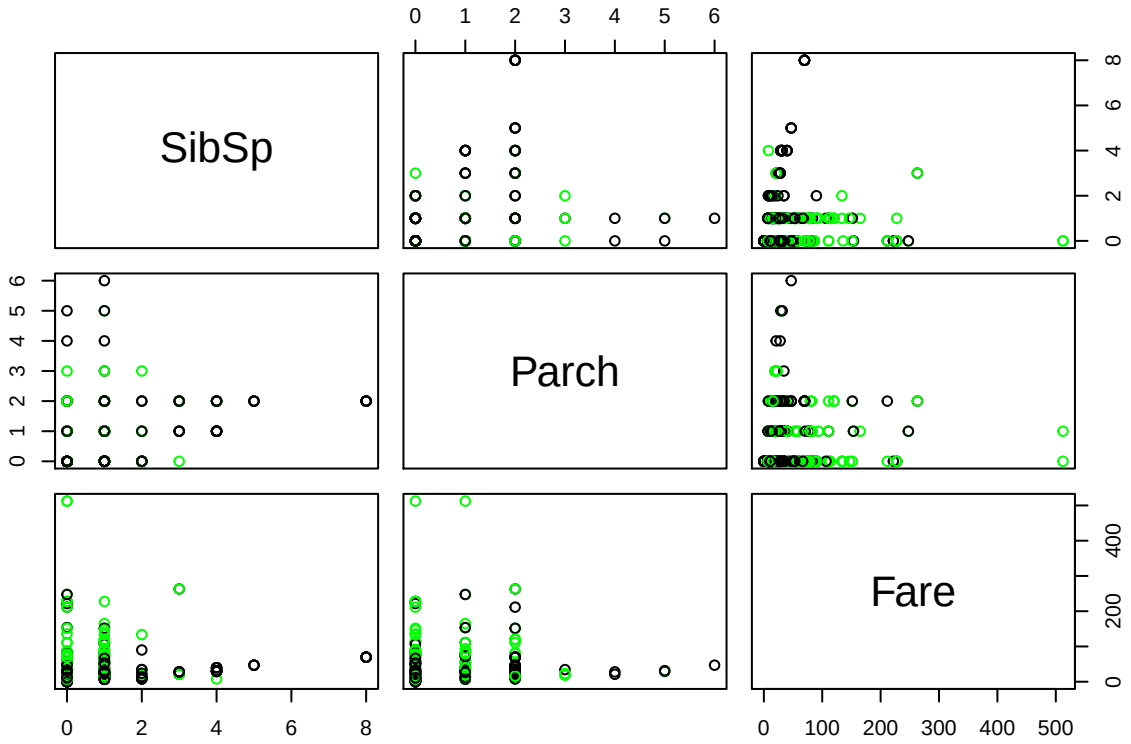

Proportional Barplot of Survival, by Gender
0.0
0.2
0.4
0.6
0.8
female
male



From the summaries above, we can see some patterns emerge. First, higher proportions of people survived as their passenger class got better (meaning they had a lower numerical value for passenger class). The inverse is also true: higher proportions of people died as their passenger class got worse (meaning they had a higher numerical value for passenger class). Also, a higher proportion of females survived and a higher proportion of males died. Finally, it appears that port Cherbourg had the highest proportion of survivors and port Southampton had the highest proportion of people who died.

Some visual EDA on classification pairs

Finally, to get a sense of which pairs of quantitative predictors might help classify type, we can inspect labeled bivariate plots. We do that in a pairs plot:



In the pairs plot above, we see a two-dimensional view of which combinations of variables might be useful in separating survivors (green circles on the plots) and those who died (black circles on the plot). There are some pairs of variables that do not show any reasonable separation; for example, the variable `Parch` does not seem to have any reasonable separation between the surivivors and non-survivors; however, we do see some potentially useful combinations, such as the `SibSp` and `Fare` variables.

Nevertheless, it is important to note that we have only looked at single or pairs of variables, and the true relationship in higher-dimensional space is likely more complicated.

Modeling

We now turn to building and assessing our classifiers for predicting whether or not someone survived. Our four classifiers are: linear discriminant analysis (lda), quadratic discriminant analysis (qda), classification trees, and binary logistic regression.

To ensure that our models are not overfitting to our sample, we randomly split our observations into training and test sets. All four models were built using the same training observations and assessed on the same set of test observations.

Linear Discriminant Analysis (LDA)

For our LDA and QDA models, we use the quantitative variables `Fare`, `SibSp`, and `Parch`.

The LDA classifier is built on the training data as follows:

Then we investigate the performance of the LDA classifier on our test data as follows:

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

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

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

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

On the test data, LDA gave an overall error rate of (12+83)/267 = 0.3558052, which is not too bad. Our LDA had an error rate of 0.7830189 in predicting the true survivors and an error rate of 0.07453416 in predicting true non-survivors. In short, our LDA is better at predicting people who didn't survive than predicting those who did.

Quadratic Discriminant Analysis (QDA)

Similarly, we use our quantitative variables for training a QDA classifier as follows:

And we investigate the performance of the QDA classifier on our test data as follows:

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

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

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

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

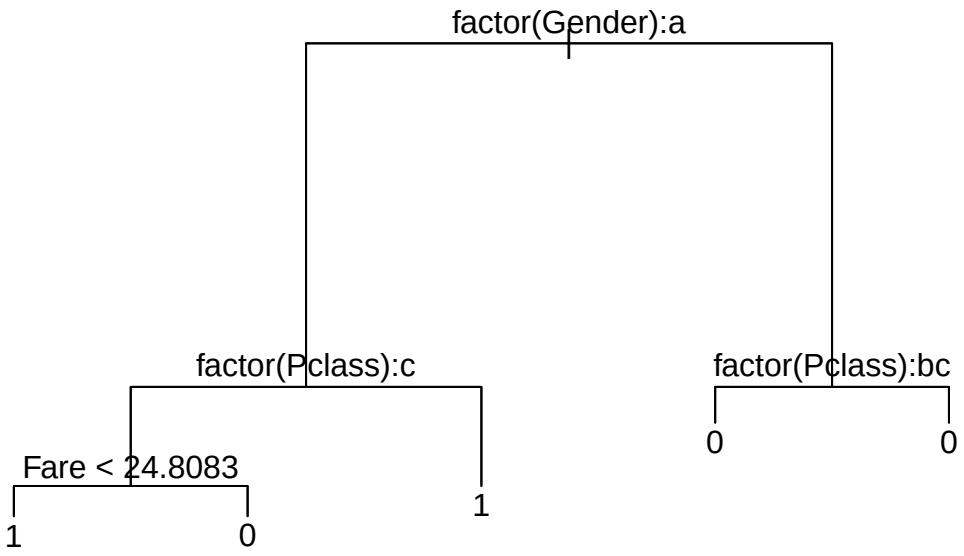
With QDA, we might expect slightly better performance than LDA, given that QDA is more flexible at finding nonlinear, curved decision boundaries.

Indeed, our results tabuled above show a slight decrease in overall error rate: (15+73)/267 = 0.329588 We do slightly better at properly classifying true survivors (error rate of 0.6886792) than with LDA (error rate of 0.7830189). We note, however, that the LDA model is better at idenitfying the non-survivors than the QDA model (the QDA has an error rate of 0.0931677 whereas LDA has an error rate of 0.07453416). It's possible that QDA is overfitting for the non-survivors.

Classification Trees

We can also account for the categorical variables (`Pclass`, `Gender`, and `Embarked`) with classification trees.

We fit a classification tree on the training data and plot it, as follows:



We note that the classification tree selected `Gender`, `Pclass`, and `Fare` to classify survivors. This means that these variables will be the "most important" for classification, with `Gender` being the "most" important and the others following respectively.

We then investigate the performance of the tree classifier on our test data as follows:

```
##
## titanic.tree.pred   0   1
##                 0 137  30
##                 1  24  76
```

```
## [1] 0.2022472
```

```
## [1] 0.2830189
```

```
## [1] 0.1490683
```

In this case, the classification tree ended up having an overall error rate of (24+30)/267 = 0.2022472. It had an error rate of 0.2830189 for predicting the true survivors. It also had an error rate of 0.1490683 for true non-survivors. The metrics for overall error rate and the error rates for the true survivors are better than the LDA and QDA. However, the LDA and QDA were better at predicting the true non-survivors.

Binary Logistic Regression

Finally, we consider binary logistic regression to model survivors. Similarly to the classification trees, a logistic classifier can use all the variables (`Pclass`, `Gender`, `SibSp`, `Parch`, `Fare`, and `Embarked`).

We train a logistic classifier on the training data, and then inspect the resulting confusion matrix from the test data, as follows:

We first fit a binary logistic regression to the data as follows:

We then apply the logistic model to the test data:

Since the logistic model applied to the test data yields probabilities (not red/white classification), we will convert the logistic probabilities into classification predictions by thresholding the probability, so that if prob>0.5 we will classify it as one type of wine (else, classify as the other type).

In order to associate the correct direction of probability with the appropriate wine type, we need to see how `Survived` is default ordered. We do that by running "levels" on the factored response variable, as follows:

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

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

We then evaluate how the the logistic classifier performed on our test data with a confusion matrix as shown:

```
##
## titanic.logit.pred   0   1
##                  0 135  30
##                  1  26  76
```

```
## [1] 0.2097378
```

```
## [1] 0.2830189
```

```
## [1] 0.1614907
```

The logistic model as a classifier (using threshold probability of 0.5) performs nearly as well as the classification tree, with overall error rate of only 0.2097378 ((26+30)/267). For true survivors, it gives an error rate of 0.2830189 (30/106), which is the same as the classification tree. For true non-survivors, it gives an error rate of 0.1614907 (26/161), which is comparable to the classification tree.

Final Recommendation

Of the four classifiers we tested, the classification tree performed the best. The logistic regression model's performance was comparable to the classification tree.

LDA and QDA performed relatively similarly to each other, with QDA performing slightly better.

It is worth noting that LDA and QDA are better at predicting the true non-survivors whereas the classification tree and logistic regression model are better at predicting the true survivors.

Our final recommendation is the classification tree because it had the lowest overall error rate (0.2022472) on the test data.

Discussion

Overall, our models did relatively well at classifying survivors. The classification tree proved to be the most effective in predicting survivors.

Other areas for future research that could be of greater interest to the public would be to see if there are any other variables, besides the ones we used, which could have influenced whether or not someone survived the crash.