Surviving the Titanic: An Analysis of Classifiers

Sibhy Rajesh
sibhyr@andrew.cmu.edu

4/28/2021

Contents

Introduction

The sinking of the Titanic quickly became infamous around the world. The massive, luxury liner was built to withstand, and yet fell on its first trip. And with only 37% of passengers surviving the wreck, the ship's end proved fatal for many.

Because of this low rate of survivability, it becomes of interest to understand what factors influenced a passenger's propensity to survive the Titanic. In this paper, we will train 4 different classifiers, and evalute their propensity to have survived using data from Frank Harrell of Vanderbilt University.(http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets)

Exploratory Data Analysis

Background and Variables

This dataset consists of information regarding passengers, including a number of characteristics, along with whether or not they survived the Titanic.

In our set we have access to the following explanatory variables:

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

Along with our response variable:

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

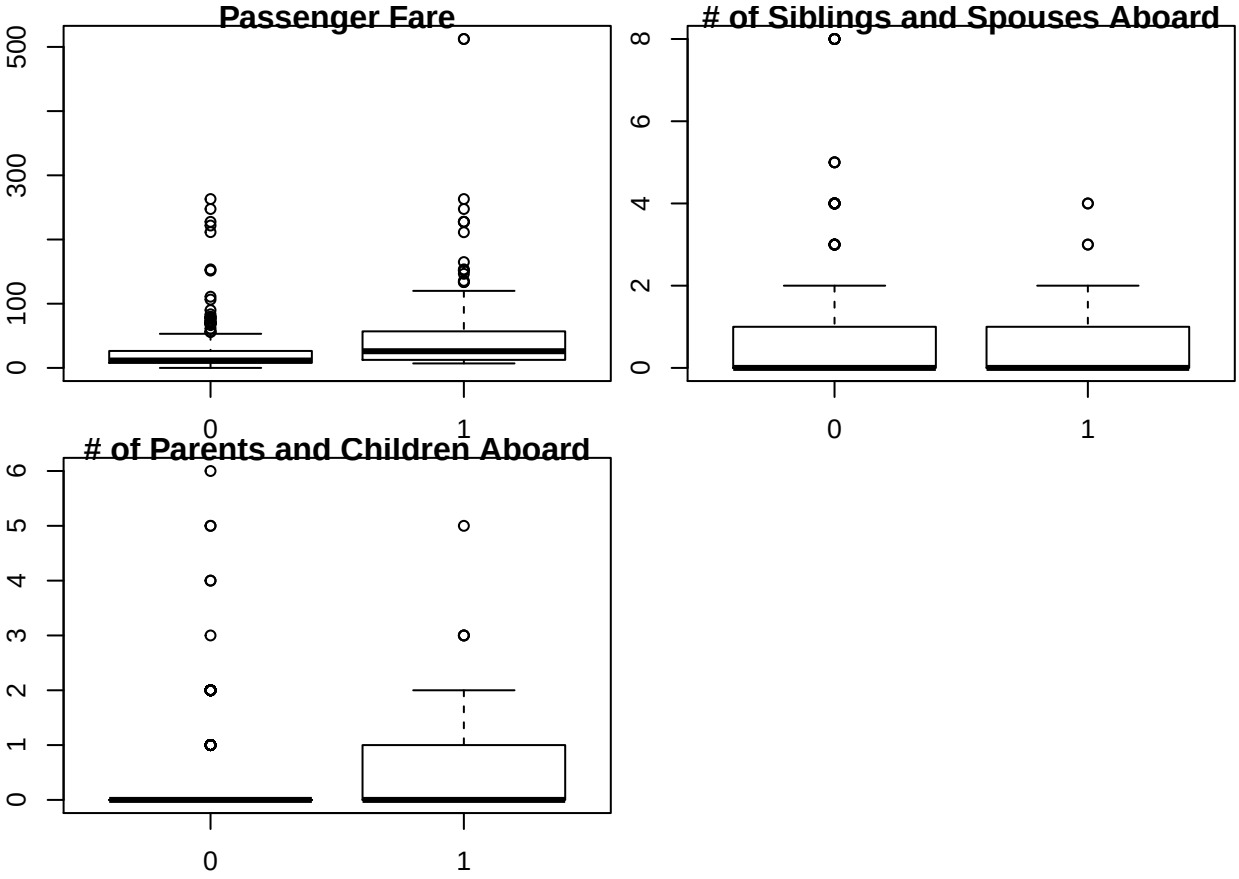
Training Dataset Summary

We note that our training dataset has 622 total observations. Of these observations, 234 are passengers who survived, and 388 are passengers who did not. This is seen in the table below.

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

EDA on Survival vs. Quantitative Predictor Variables

We can now visualize the relationship between passenger Survival and our different predictors. To do so, we use boxplots for each of our quantitative predictors to see how useful they may be in predicting passenger survival.



The above boxplots could help us gather evidence for a relationship between Survived and our quantitative predictors, which could be helpful when evaluating our classifiers. Starting with our boxplot comparing Fare, we see a slight difference in the median Fare between the two groups. It appears as though passengers who survived had, on average, slightly higher fares, but this observation is by no means concrete evidence. Moving to our boxplot comparing SibSp, we see no clear difference between the two groups except for outliers. It

seems as though there were a few passengers with large numbers of siblings and spouses(up to 5) who failed to survive the titanic. Finally, we look to our boxplot comparing Parch. Based on this boxplot, the number of parents and children held by passengers who survived

EDA on Survival vs. Categorical Predictor Variables

We now look to our categorical predictor variables to explore any possible relationships between them and our response variable, Survival. We can do so by creating proportion tables.

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

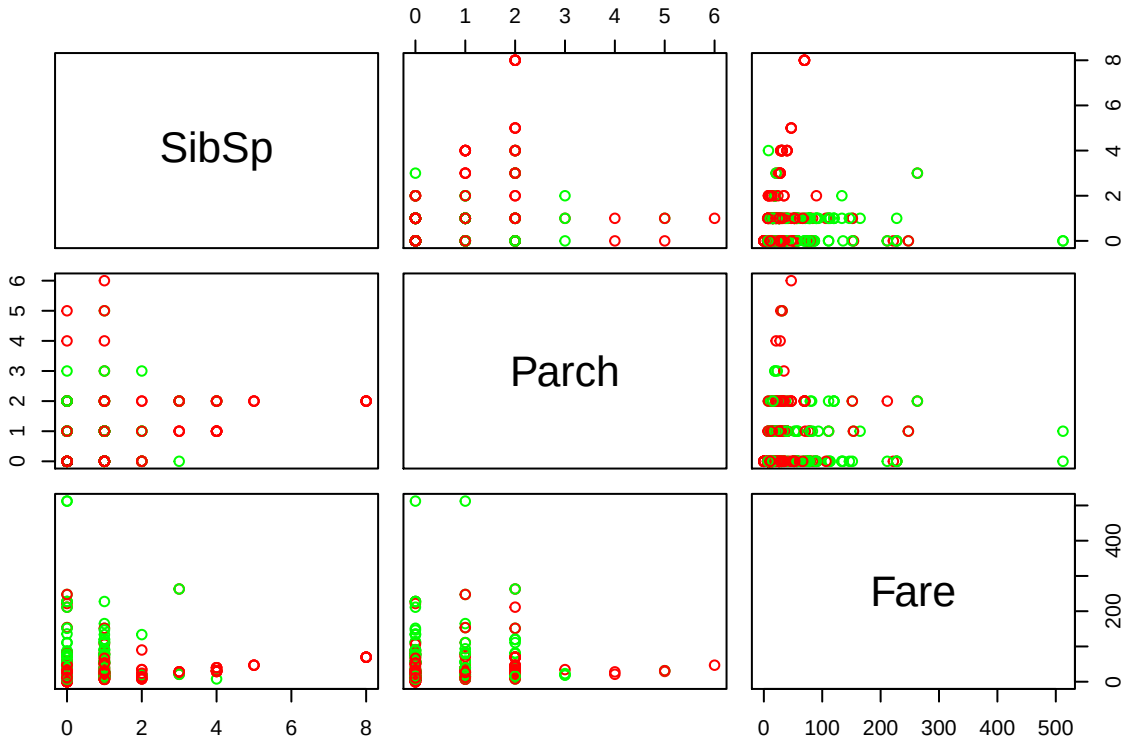
```
##
##            C         Q         S
##   0 0.4324324 0.6400000 0.6681128
##   1 0.5675676 0.3600000 0.3318872
```

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

Starting with the Pclass proportion table, we see that over 75% of passengers in the third class died with the titanic, whereas only 39% of passengers in the first class died with the titanic. This could be indicative of a relationship. Moving to the Embarked proportion table, we see around 33 - 36% of passengers who boarded at Queenstown and Southampton survived the titanic, whereas close to 57% of the passengers who boarded at Cherbourg survived the titanic. Finally we move to the Gender proportion table. Here we notice close to 75% of female passengers survived the titanic, whereas only 19% of male passengers survived the titanic. This difference could be relevant when evaluating classifiers.

EDA on Classification Pairs

Lastly, we can use a pairs plot to see which of our quantitative predictors will be useful in predicting whether a passenger survived.



The pairs plot above shows possible combinations of quantitaive predictors which could be helpful in distinguishing passengers who survived from those who didn't. Some combinations, such as Parch and Fare show little visible separation, indicating their lack of effectiveness However, we do see some separation in the plot between SibSp and Fare, as higher Fare values tend to show more green values(surviving passengers). This could be useful when doing our modeling.

We also note that while the pairs plot gives us a high level overview of some combinations, the actual relationship between Survived and our predictors is likely much more complicated.

Modeling

In this section we will create and then evaluate our classifiers for predicting whether a passenger survived. The four classifiers we will use are linear discriminant analysis, quadratic discriminant analysis, classification trees, and binary logistic regression.

1. Linear Discriminant Analysis (LDA)

Our LDA model will use the quantitative predictors SibSp, Parch, and Fare and is built as shown below.

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

We can now evaluate the LDA classifier on our test data and view the results below.

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

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

Based on these outcomes, we can calculate error rates to evalute the performance of the LDA classifier. The overall error rate is $(83 + 12)/(267) = 0.356$, which is relatively high. Looking at predicting Passengers who survived, we have a high error rate of $83/106 = 0.783$. On the other hand, the error rate for predicting Passengers who failed to survive is much lower than the other, coming out to $12/161 = 0.075$.

2. Quadratic Discriminant Analysis (QDA)

Just as we did with our LDA, we use the same quantitative predictors in our QDA classifier.

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

We then carry out a similar evalution as done with the LDA using a 2x2 table.

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

To preface our analysis, we might expect that QDA perform better than LDA because QDA is more built to handle non-linear decision boundaries.

Based on the outcomes of our table we can make observations regarding the error rate. The overall error rate is $(15+73)/267 = 0.330$. The error rate for predicting passengers who survived is $73/106 = 0.689$. For predicting passengers who failed to survive, the error rate comes out to $15/161 = 0.093$.

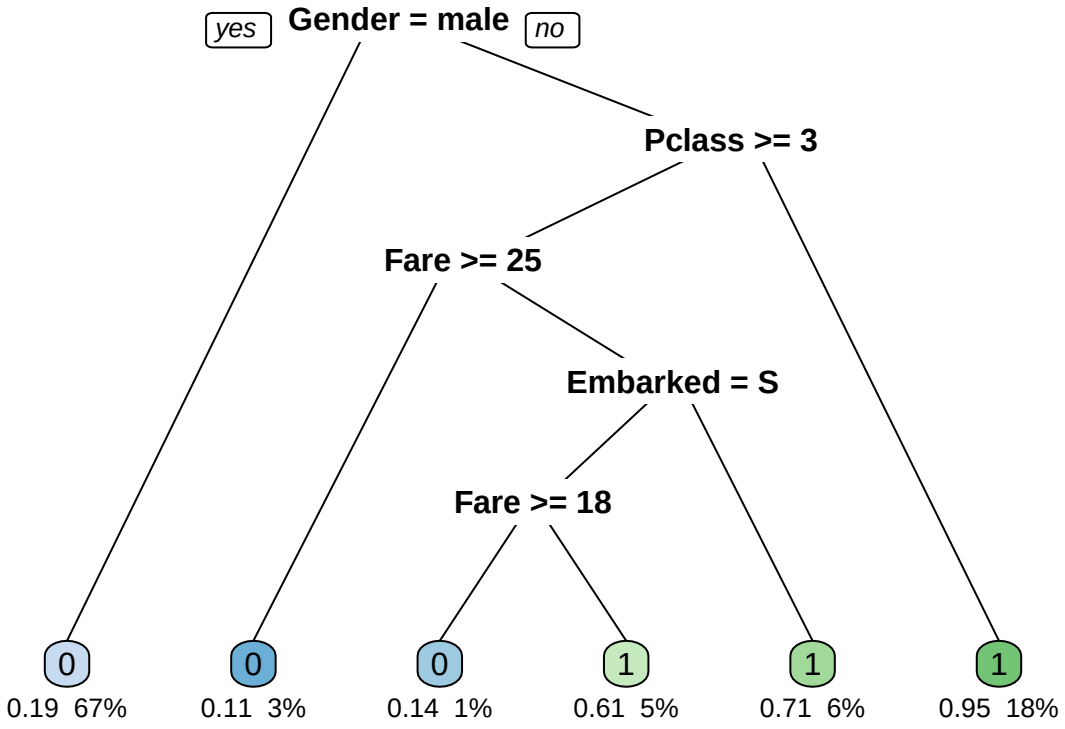
We observe a common pattern between our LDA and QDA: low error rate in predicting passengers who failed to survive, but high error rate in predicting passengers who survived. This could possibly hint at our categorical variables being relevant in predicting passengers who survived, as our LDA and QDA model didn't include these predictors.

3. Classification Trees

One benefit of using a classification tree is the ability to include categorical predictors which were omitted from our LDA and QDA. As a result, we build our model as follows:

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

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



We note that the tree prioritizes Gender as the most important predictor of whether or not a passenger survives. It then looks to the categorical variable of passenger class, and then other predictors like Fare and Embarked.

We can now explore the accuracy of our tree classifier.

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

We can now calculate the error rates for our tree classifier in predicting passenger survival. The error rate for predicting whether a passenger survived is $32/106 = 0.302$ and the error rate for predicting whether a passenger failed to survive is $20/161 = 0.124$. The overall rate is $(20 + 32)/267 = 0.195$.

4. Binary Logistic Regression

Next, we use a binary logistic regression to model whether or not a passenger survived based on our predictors. Like we did with our classification tree, we can include categorical predictors in this classifier.

We fit our model as follows:

```
titanic.logit <-glm(factor(Survived)~SibSp+Parch+Fare+factor(Gender)+factor(Embarked)+ factor(Pclass),da

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

Because our logistic model outputs probabilities, we will threshold the probabilities such that if it is greater than 0.5, we assign it to a 0 or 1 for our Survived variable. To make sure we are assigning to the correct

value for our Survived variable, we run "levels" on our response variable.

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

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

We can now obtain our classifier and evaluate its performance on our test_dataset.

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

Our logistic model classifier has an overall error rate of $(30 + 30)/(267) = 0.225$. For predicting passengers who survived the titanic, the model has an error rate of $30/106 = 0.283$. For predicting passengers who failed to survive the titanic, the model has an error rate of $30/161 = 0.186$.

Similar to our classification tree, our logistic model performed significantly better overall, likely due to the inclusion of our categorical variables.

Final Recommendation

This paper tests four unique classifiers and evaluates their ability to predict whether a passenger on the titanic survived or died based on certain predictor variables. Of these four models, the classification tree performed the best. The classification tree had the lowest overall error rate, and outperformed the logistic model classifier in predicting whether a passenger failed to survive.

We do note that the logistic model classifier had a lower error rate than the classification tree with regards to predicting whether a passenger survived, but this difference is outweighed by the aforementioned differences.

As expected, the LDA and QDA models suffered due to lack of inclusion of categorical predictors, which from our EDA we predicted may be important in the prediction of passenger survival.

Our final recommendation is the classification tree, as it has the lowest overall error rate, lowest error rate for predicting passengers who failed to survive, and is only marginally outperformed by the logistic model when predicting whether passengers survived.

Discussion

Overall, we find that of the four models tested, two stood out as more useful than the others. These were the classification tree and binary logistic model. We come to this conclusion because the LDA and QDA models had extremely high error rates when predicting whether or not a passenger survived the titanic. We do however acknowledge that despite this downfall, these two models were particularly effective at predicting whether a passenger failed to survive the titanic.

Some areas for future research could be comparing the outcomes of the titanic to other ship liner wrecks, and seeing if similar themes appear in the wrecks of other large ship liners. If so, it could be useful in building a more robust model for predicting the survival of passengers.