Predictors of Survival for Titanic Passengers

Alden Chen, Birinder Singh

2018-11-17

Introduction

In this analysis, we look at data on passengers from the Titanic. Using seven features, we fit a decision tree model to find out which two features are the best predictors of survival for Titanic passengers.

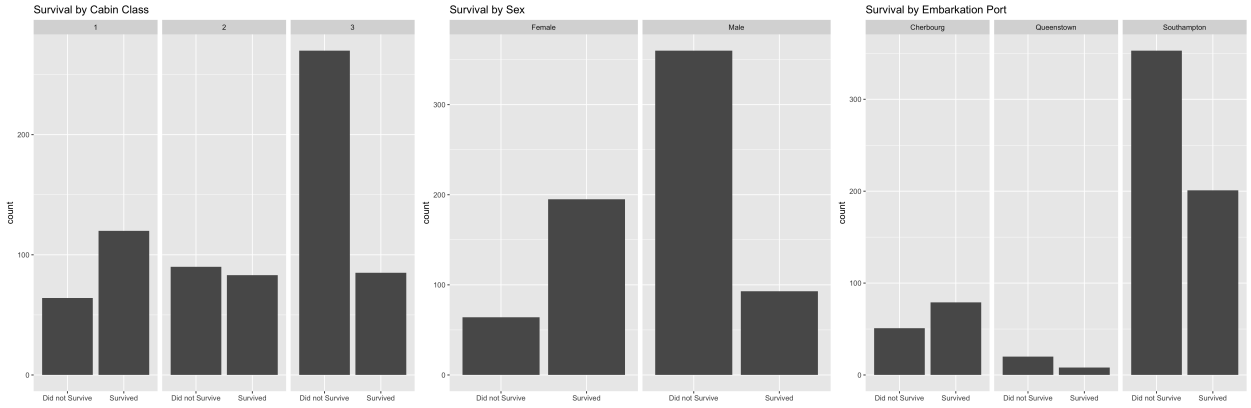
Exploratory Data Analysis:

Our target variable is `Survived`, a dummy variable that takes the value `0` is a passenger did not survive and `1` is he or she did survive. We have data on 712 passengers for these seven features

Feature	Description
Passenger Class	First, Second or Third Class
Sex	Male or Female
Age	Age of each passenger years
Siblings/Spouses	Number of siblings and spouses aboard the Titanic
Parents/Children	Number of parents and children aboard the Titanic
Fare	The amount each passenger paid for a his or her ticket
Port of Embarkation	Cherbourg, Queenstown, or Southampton

In this section, we look at some plots of the data and discuss their implications for prediction.

Figure 1: Bar Plots, Number of Survivors by Passenger Class, Sex, and Port of Embarkation

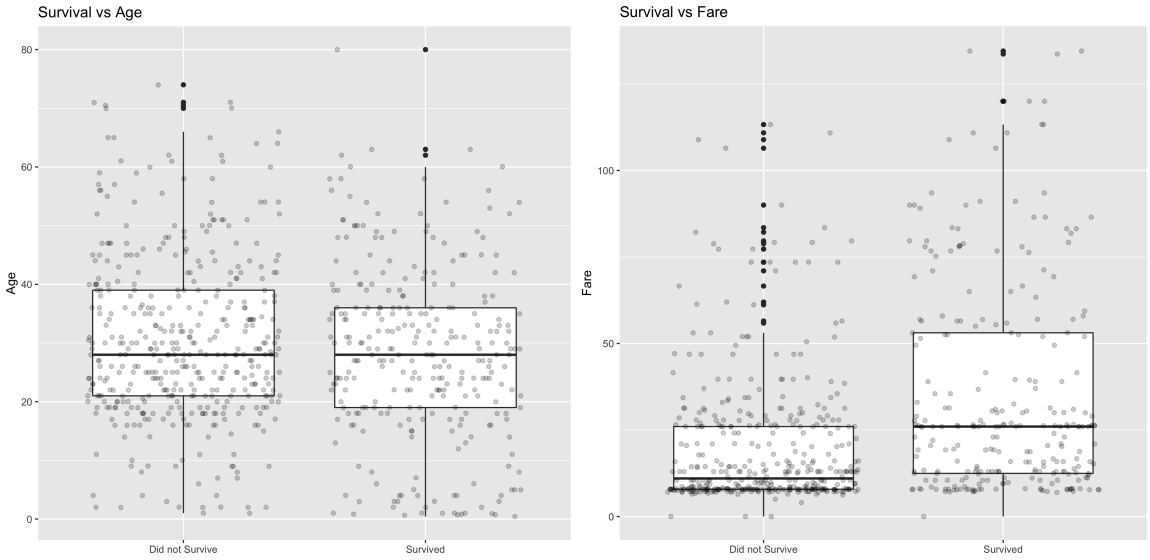


In the three bar plots above, we compare the number of people who survived based on passengers' sex, port of embarkation, and cabin class. We can see that there appears to be a large difference between in the number of passengers that survived based on sex. For female passengers, there are more passengers who survived than passengers than died, while the opposite is true for male passengers. We see a similar trend for cabin class. Among first class passengers, there are more survivors than non-survivors, while among third class passengers, there are many more non-survivors than survivors.This suggests that sex and passenger class could be important features for predicting survival.

For port of embarkation, the vast majority of passengers boarded at Southampton, which appears to have similar rates to Queenstown passengers. Cherbourg passengers seem to have a much higher survival rate

than passengers who boarded at other ports. Given the small number of Cherbourg passengers, it is unclear if port of embarkation will be a good predictor of survival.

Figure 2: Box Plots, Survival vs Age and Survival vs Fare



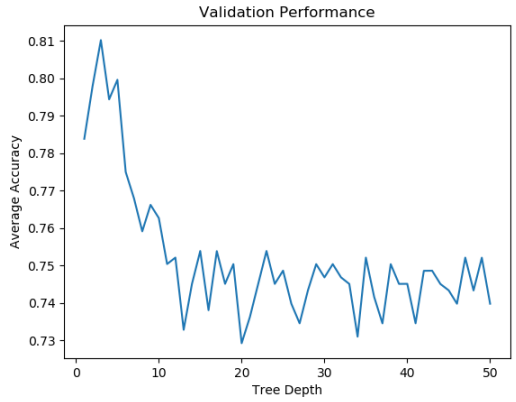
The box plots above show the distribution of ages and fares of the passengers based on whether or not they survived. There does not seems to be only a small difference in the ages of passengers who survived and those who did not, so age likely would not be a very important feature for predicting survival.

For fare, it appears that passengers who survived paid higher fares. This is likely related to the trend that there were more first class survivors than second or third class survivors. Presumably, first class passengers paid higher fares, so it is not surprising that survivors on average paid higher fares.

Model Development

To decide on the appropriate depth of the tree, we used five-fold cross validation. Below is a plot of the validation performance for different depths of trees. We can see a clear spike in the average accuracy at a depth of 3.

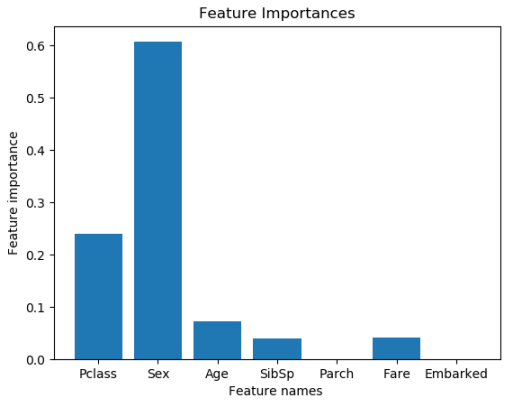
Figure 3: Validation Performance, 5-Fold Cross Validation



Results

After fitting a model of depth three, we looked at the feature importances of our model. The plot below summarizes our results. We can see that the two best predictors of survival for Titanic passengers are sex and passenger class. This supports our initial observations from the exploratory plots that we produced.

Figure 4: Feature Importances



References

Kaggle.(2012). *Titanic: Machine Learning from Disaster* [Data files and description]. Retrieved from: https://www.kaggle.com/c/titanic

scikit-learn developers. (2018). *sklearn.tree.DecisionTreeClassifier*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html