Titanic Survival Report

Problem Introduction: The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

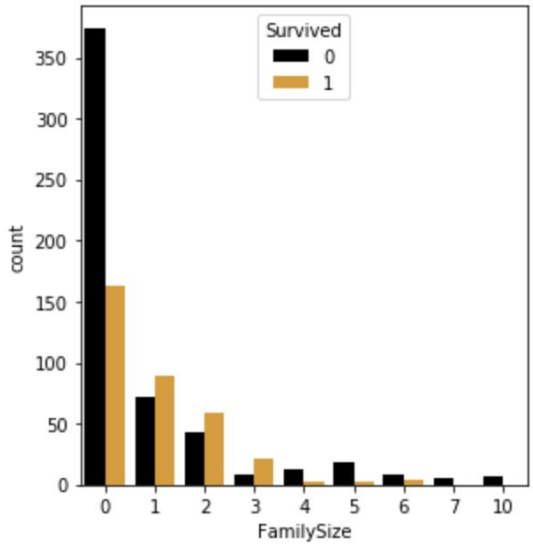
One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Overview: The titanic dataset is split into 2 sets, one for training and one for testing. The training set has 819 observations, the test data contains 418; they combine for a total of 1237 observations. The training data is organized into 12 categories; the test data has 11 categories, with the missing category of "Survived". There were many missing values in the training data, so the first step was to clean the data and fill in values.

Identifying Important Factors:

Obviously, luck is involved in each individual passenger's survival . However, there were many factors that correlated to the survival column. In order to better understand which factors affect survivability, I created bar graphs for each individual category to visually see the correlation. Shown below is a graph of the effect of family size on the chances of survival. As you can see, the smaller the family, the more likely to survive.



I also used the function of summary to see the R squared value of each individual factor. Using both the bar charts and R squared value, I found that gender, family size, age, and passenger class were the most relevant to whether the passenger survived or not. The place of embarkment had a much smaller impact.

Cleaning the Data (filling in missing values):

After identifying the important factors, it was time to clean the data. Because the original training dataset has so many missing values, using the raw data to predict survivability resulted in an accuracy of around 50%. Obviously, we would hope for better results, so some data cleaning had to be done.

I started with the category with the most missing values: age. The training data had 176 missing values. The most basic method is to calculate the median of the age values I already have, and then just fill in every missing value using the median. This improved the accuracy, but the model could definitely still be improved.

The method I used was fairly simple. First, in order to make sure my data wasn't skewed, I got rid of all the outliers (data points 3 standard deviations away from the median). Next, I found the factors that had the strongest correlation to age (SbSp and Pclass). This time I didn't create any boxplots; I simply found the factors with the lowest Pr value using the summary function. This model proved to be far more accurate than just using median. It can also be noted that I tried using randomforest to generate the missing values, but the accuracy at the end was worse, so linear model was the method I ended up choosing.

Because the embarkment classifier had little effect on survivability, I simply found the most common port "S" and filled in every missing value with that.

The fare category only had one missing value, so I just filled in with the median.

The final category that had missing values was cabin number. This category had a strong effect on the survival of passengers, with upper deck passengers having a much better chance. I used randomforest to fill in these missing values. This model might not have been the best, but it was hard for me to create a linear model to better predict cabin number.

Analysis/Conclusion: Once the data was properly cleaned, I used randomforest to predict whether each passenger survived or not. My training accuracy was higher than my testing accuracy, but that is to be expected. I ended with a 78% accuracy rate.

Through this project, I learned how to clean data using various methods. The most simplistic way was to just find the median of know values and then fill in every thing using that one value. The improved way was to create another predictive model that factored in other variables to better approximate the missing value. I think the most important aspect was choosing the correct predictive model. I originally tried just using a linear regression model, but the correlation wasn't very linear. Thus, I ended up choosing RandomForest.

Acknowledgements:
This report was done by Jeffrey Mi. Kaitlin Sun and Jonathan Mi helped contribute ideas as well.

Citations:

[1] Kaggle.com. "Titanic: Machine Learning from Disaster." *Kaggle*, 15 July. 2019, www.kaggle.com/c/titanic/data.

[2] James, Gareth. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2017.