Name: Bao Van

Titanic Project

STAT 488

Date: April 4th, 2021

I Know Who Survived During the Titanic

1. Introduction

During the Titanic sinking event, there were a lot of people could not survive. However, there were plenty of people that survived based on some certain factors. As a data analytic/data scientist, I want to predict and find out which factors that will affect the chance of surviving, also the accuracy of my optimized method by using random forest classification with Python.

1. About the Data

The dataset contains 11 independent variables: Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked and one dependent variable that we want to predict, Survived. Within those independent variables: Pclass, SibSp, Parch, Age are integers, Fare is a float type, Name, Sex, Ticket, Cabin, Embarked are object types.

First thing, I want to look at our dependent variable Survived first to check if it is biased or not. From Figure 1, the number of 0s is a little higher than 1’s, but it doesn’t seem to be biased. When we do the splitting dataset, we can do it in a normal way of 80:20 splitting.

Next, I need to check whether there are any null values or not. In training dataset, Age has 177 nulls, Cabin has 687 nulls and embarked has 2 nulls. In testing dataset, Age has 86 nulls, Fare has 1 null and Cabin has 327 nulls. As we can see that, there is a large number of null values in both training and testing dataset for Cabin, so we might consider to drop this variable.

1. Method:

First, we need to due with the null values. For the Embarked in training dataset, since it is a categorical variable which has only 3 values: S, C, Q. It is reasonable that we can fill those 2 NAs with the most common port of embark. For the Fare in testing dataset, it has 1 null value, so we can fill mean value which will keep the characteristic of this Fare variable. For Age in both training and testing dataset, we can fill in the average age for each title and there are 6 title in this dataset: Miss, Mr, Mrs, Master, Dr and Ms. The last and the most difficult one to do with is Cabin variable. I noticed that, Cabin includes location and cabin number in the ship, so it is important to know where they are to predict better. I need to extract the first letter of Cabin column because the numbers are not really matter, and for those nulls I will label them as ‘U’ standing for ‘Unidentified’. After doing that, the number of ‘U’ is the highest in Cabin column in both training and testing dataset. However, we know that Cabin is associate with Fare, like buy a seat on an airplane. Now, we can check for the outliers for Cabin associate with Fare. From Figure 2, we can see that there are a handful of outliers for ‘U’. For those outliers, we can reassign cabin for each outlier, by simply find the mean for each cabin and say if the fare in ‘U’ is higher than a mean of a cabin, then we assign it to that cabin and we filter it in an order from highest mean cabin fare to lowest. From Figure 3, we can see the outliers are almost cleaned up.

When there are no null values anymore, we need to transform Sex, Embarked and Cabin into numerical datatype since random forest only works with numbers. Then I want to drop unimportant columns like Name, Ticket and PassengerId. To make our model predict more accurately, I normalized my dataset. In addition, I want to see which variables are important to predict the ‘Survived’, I used random forest to do feature importance. From Figure 4, we can see that Parch and Embarked are not very important and I don’t want to feed my model with too many variables, so I decided to drop them.

Finally, I applied random forest classifier to train my 80% of my training dataset with random state of 6 and stratified my Survived.

1. Result

After using random forest, my accuracy changes each time I rerun the cell, so I used k-fold to find the mean score which is 92% with a standard deviation of 0.07.

1. Future Work

If I have more time, I would be able to play around with the variables in our model or have time to find a better way to do with Cabin variable.