Analysis and Prediction of Survival of the Titanic

Brad Zhang
Summer Research
HKUST
bradzhang6@gmail.com

01. Introduction

One of the most infamous shipwrecks in the history of mankind is the sinking of the RMS Titanic. On April 15, 1912, the Titanic sank after colliding with a giant iceberg, killing 1502 out of the 2224 passengers and crew. This news-breaking tragedy shocked the international community and therefore led to a change in safety regulations for ships.

Out of the many reasons that the shipwreck led to such loss of life, one reason was that there were not enough lifeboats for the passengers and crew. Out of those who have survived, there was some element of luck involved. However, there were other factors which contributed to the surviving the shipwreck. For example, certain groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this report, the survival of certain crew members and passengers will be predicted based on various external factors of those on board.

02. The Titanic Dataset

The Titanic dataset contains 819 observations in training data and 418 observations in test data. The categorical features include each passenger's ID ('PassengerId'), their social class at that period of time ("PClass"), as well as their identification information such as their name, sex, age etc. Whilst the data may seem organized with categorical sets, as you go through the data there will be multiple issues that are uncovered. Such issues are missing values and unnecessary as well as badly written information. Due to these issues, my first step was to analyze and clean each categorical variable in order to determine which features are useful in the prediction of survived passengers.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C

Figure 1: Snap shot of the Titanic Dataset.

03. Cleaning The Data

In my first prediction, I didn't clean and properly analyze each feature in order to make a more accurate prediction and instead I had just considered all features into my prediction. This resulted in a poor prediction score of 0.61722.

- ➔ Within the features, there were four features which contained missing values: *Fare*, *Cabin*, *Embarked*, and *Age.*
- ➔ **Fare Analysis:** Inside the *Fare* feature, there was only one missing value out of all the given values. This resulted in two options: either leaving it the way it is or estimating the value through the given

data. I decided to estimate the missing value by using the median and the mean between the specific *Class* that this missing value lies. Using the median we can estimate the value and fully clean the *Fare* feature.

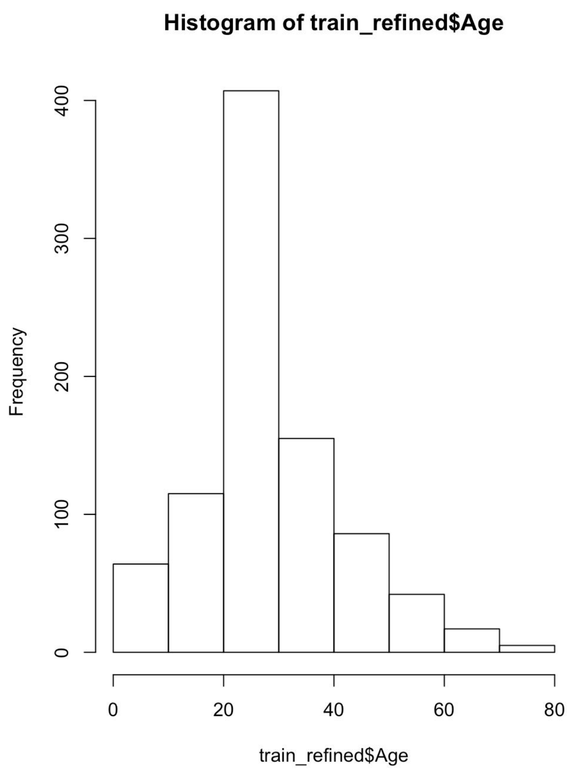
➔ **Cabin Analysis:** Within the *Cabin* feature, there are multiple missing values inside the dataset. As I used RandomForest in my second prediction, this feature can be withdrawn as a lot of values weren't present and the data could be questioned.

➔ **Embarked Analysis**: In the *Embarked* feature, there were only 2 missing values that were present, Passenger 62 and Passenger 830. Since both passengers purchased a ticket in first class that costed $80, we may assume they were most likely "C" in the embarked feature.

➔ **Age Analysis:** With *Age* being a potentially very important feature to include, the feature contains a great amount of missing values. Due to the potential of including *Age*, instead of discrediting the feature, the mean and median strategy was used to predict the missing values through the "PClass" and "Title" features (Figure 3). Once the values are predicted, the outliers needed to be removed which gave the overall summary below (Figure 2).

```
  PassengerId      Survived          Pclass          Sex           Age            SibSp            Parch
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   male  :577   1st Qu.:22.00   1st Qu.:0.000   1st Qu.:0.0000
 Median :446.0   Median :0.0000   Median :3.000                Median :29.70   Median :0.000   Median :0.0000
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                Mean   :29.70   Mean   :0.523   Mean   :0.3816
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                3rd Qu.:35.00   3rd Qu.:1.000   3rd Qu.:0.0000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                Max.   :80.00   Max.   :8.000   Max.   :6.0000
      Fare
 Min.   :  0.00
 1st Qu.:  7.91
 Median : 14.45
 Mean   : 32.20
 3rd Qu.: 31.00
 Max.   :512.33
```

Figure 2: The summary of the Titanic Data once cleaned and with outliers removed.



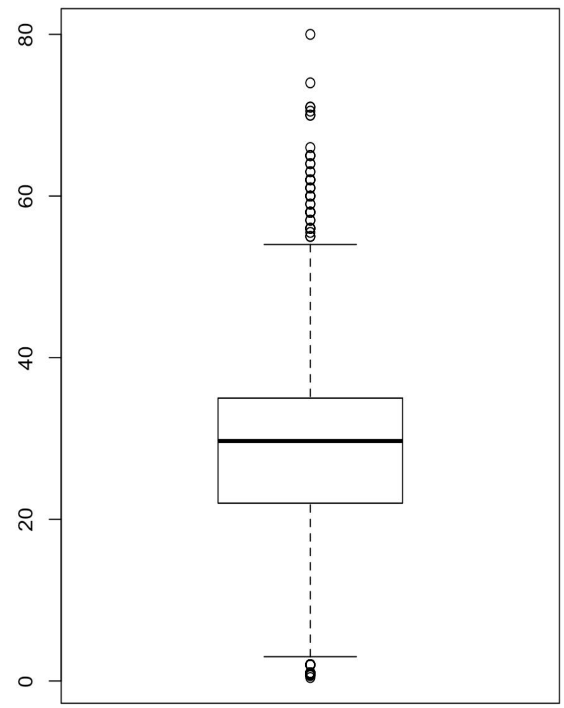


Figure 3: Predicting and identifying outliers in the *Age* feature

04. Standardization

As you may be wondering, the given training data is not the same in any relative scale, which leads up to an important aspect of predicting the survival of passengers, standardization. Standardization not only helped get rid of any potential skews, it helped in increasing the accuracy of data prediction from 72% up to 80%.

05. First Prediction

In my first prediction, I simply used the Generalized Linear Model (GLM) to predict the survival of the passengers without any consideration of the missing values as mentioned above. A major drawback of using GLM was that it suffered a substantial decrease in precision as the variables weren't aligned fully and correctly due to the missing values. This prediction gave me an overall score of 0.61722.

06. Modification

Although the first prediction yielded a score that was more accurate than guessing (0.5), there was definitely room for improvement. For the second prediction, I implemented Randomforest which enhanced the accuracy and performance of the algorithm. This helped improve the accuracy to 83.83% and an error rate of 0.09%.

07. Final Prediction

After using Randomforest to predict the test data set, it yielded a prediction score of 0.80334, a very big improvement from the first prediction. Although the score did not achieve full accuracy, it is satisfactory and there is definitely better and more accurate methods that will be explored in the future in order to further improve my results.

Submission and Description	Public Score
survival_prediction(final).csv	0.80334

Acknowledgements

This report was done by Brad Zhang

References

[1] Kaggle.com. "Titanic: Machine Learning from Disaster." *Kaggle*, 28 Sept. 2012, www.kaggle.com/c/titanic/data.