Exploring the Titanic survivors’ data set - a machine learning approach

Juan De Dios Santos

I. Introduction

On April 15th 1912 the British passenger line, RMS *Titanic* sank during her maiden voyage after colliding with an iceberg, resulting in the deaths of over 1,500 people [1]. The outcome of the lives of the passengers, including the crew, has become a popular phenomenon due to the priority given by the captain to save the lives of the women and children, making famous the phrase *”women and children first”*.

In this report we will analyze a data set that provides information about the fate of some of the passengers of the Titanic, with the intent of predicting their outcome, based on the characteristics of the person, and facts about their boarding pass and reservation.

II. The data

The data set used on this report is made of 714 observations and 12 columns.

- `PassengerId`: ID of the passenger.
- `Survived`: Outcome of the passenger; `0` if it did not survived, `1` if it did survived.
- `Pclass`: Passenger class; `1` if upper or 1st class, `2` if middle or 2nd class, and `3` if lower or 3rd class.
- `Name`: Name, and title of the passenger.
- `Sex`: Sex of the passenger.
- `Age`: Age of the passenger in years, fractional if it is less than 1, and if the age is estimated, it follows this format `xx.5`
- `SibSp`: Number of siblings/spouses aboard related to the passenger.
- `Parch`: Number of parents/children aboard related to the passenger.
- `Ticket`: Ticket number.
- `Fare`: Passenger fare.
- `Cabin`: Cabin of the passenger.
- `Embarked`: Port of embarkation, `C` if Cherbourg, `Q` if Queenstown, or `S` if Southampton.

Table 1 is a summary of the `Survived`, `Pclass` and `Sex` attributes. The first column shows the death toll, which is 424 (59.3%), and the number of survivors, which is 290 (40.6%). The second attribute of the table is the passenger class. On the data we found that most of the passengers, 355 (49.71%) were staying in the third class, and the second class is the less common one, with 173 passengers (24.22%). Lastly, the third column present that there were 261 females aboard, and 455 males.

Survived	Pclass	Sex
Deaths: 424 Survived: 290	Min. : 1.000 1st Quartile: 1.000 Median: 2.000 Mean: 2.237 3rd Quartile: 3.000 Max. : 3.000	Female: 261 Male: 453

TABLE I. Summary and statistics regarding the survival rate, the passenger class, and the gender ratio

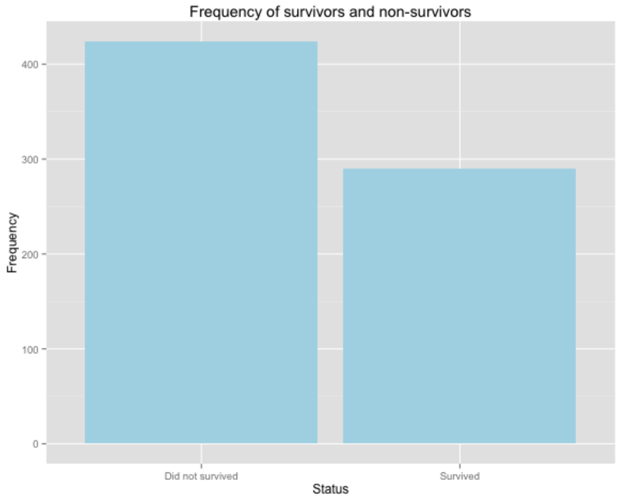

Fig. 1. Frequency of survivors and non-survivors

Most of the passengers of the Titanic were young adults, between the ages of 20 and 40. The younger passenger was 0.42 years old, and the oldest one was 80. The mean of the age range is 29.70, and the most common age, or median reported was 28.00. The next two figures are related to the ages. The first one is a histogram with a density line that shows the most common age group, and the second one is a boxplot.

Two other significant features present on the data set are the number of siblings/spouses, and number of parents/children aboard that are related to the person. According to the information encountered on the data set, most of the passengers were traveling alone, or with one companion.

The following two tables shows that most of the passengers were traveling alone, or with 1 companion, who could be either a sibling, spouse, parent or children.

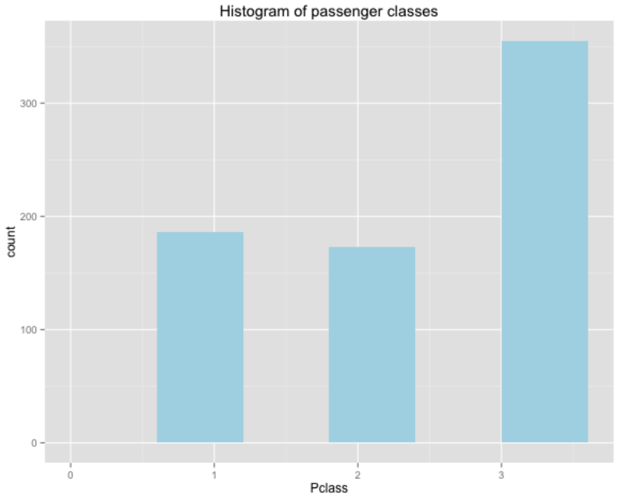

Fig. 2. Frequency of passenger classes

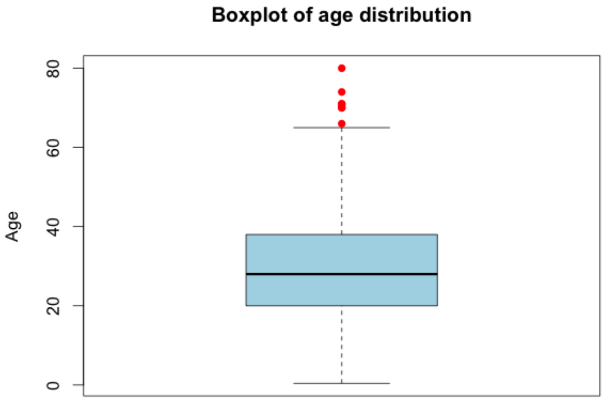

Fig. 4. Boxplot of ages

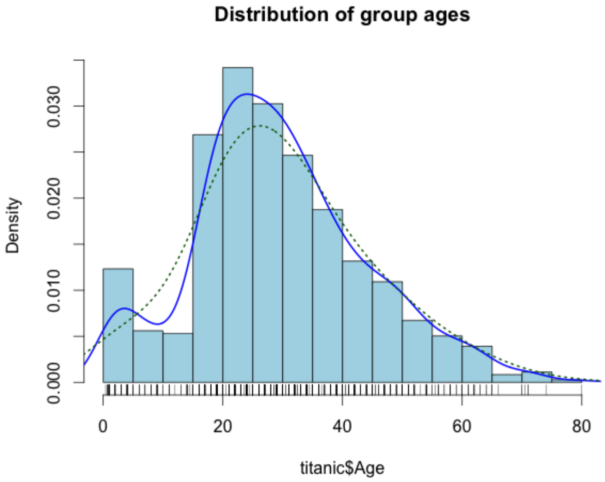

Fig. 3. Histogram with density line of the group ages

A. Association Rules

As part of the overview and exploration of the data, we performed an association rules analysis to test the assumption that women and children are most likely to survive. Moreover, we were also interested to find out any other frequent item set linked to the attribute of age and sex. To achieve this, we created a new attribute for the data set named `Stage`, that is based on the age of the passenger - if the passenger is 5 years old or less, then `Stage = 0`, otherwise `Stage = 1`.

0	1	2	3	4	5
471	183	25	12	18	5

TABLE II. TOTAL OF SIBLINGS/SPOUSES TRAVELING WITH THE PASSENGERS

0	1	2	3	4	5	6
521	110	68	5	4	5	1

TABLE III. TOTAL OF PARENTS/CHILDREN TRAVELING WITH THE PASSENGERS

The association rules model built applies the Apriori method, a technique that counts all the transactions, which in this case is the stage, sex and outcome of the person, and derive rules from them [2].

Our model was built using a minimum confidence level of 0.8, meaning that if a rule does not have said level, it will not be considered as frequent.

The Apriori model found one frequent association rule. Said rule is `{Sex=male,Stage=1} => {Survived=0}`, and has a confidence level of 0.8186, implying that 81.86% of all the men older than 5 years old died in the accident. The support of this rule is 0.4929, with a lift measure of 1.3784.

III. MODEL DEVELOPMENT AND PREDICTION

A. The classification method

The selected classification method for performing the predictions is the random forest technique. A random forest is an ensemble learning method used mostly for classification and regression, that is made of a number of decision trees. It works by constructing a large number of decision trees during the training phase, followed by selecting the most common class produced by the trees [3].

B. Feature creation

In the first section of this report we introduced a new feature, named `Stage` that is based on the age of the passenger. Besides this feature, we created another two - `Total.Family`, which is the sum of the `SibSp` attribute plus `Parch`, and `Title.Prefix` which is the title of the person, e.g. Ms.

C. Prediction

Our random forest model consisted of 290 trees, and it performed the prediction using the following attributes of the

Number of test	Training error	Testing error
1	18.16	17.82
2	18.48	20.26
3	21.12	16.51
4	17.84	19.07
5	19.02	14.71
Average	18.92	17.67

TABLE IV. ERRORS FROM THE TRAINING, AND TESTING PHASE

Feature	Importance
Title.Prefix	0.00
SibSp	3.84
Total.Family	8.97
Pclass	19.93
Stage	25.60
Sex	39.61

TABLE V. FEATURE IMPORTANCE

data set: `Stage`, `Sex`, `Total.Family`, `SibSp`, `Pclass`, and `Title.Prefix`.

70% of the data set was used as the training set, and the remaining 30% as the test set.

A total of five tests were performed. At each test, a new data sample was taken, followed by training the model, and testing it. The training errors reported during the five tests were: 18.16%, 18.48%, 21.12%, 17.84%, and 19.02% - an average of 18.92%.

The respective test errors were: 17.82%, 20.26%, 16.51%, 19.07%, and 14.71% - an average of 17.67%.

During the last test, the importance of each feature was calculated. As table V shows, `Title.Prefix`, one of the attributes we created, has an importance of 0.0%, and on the other hand, the sex of the passenger achieved an importance of 39.61%.

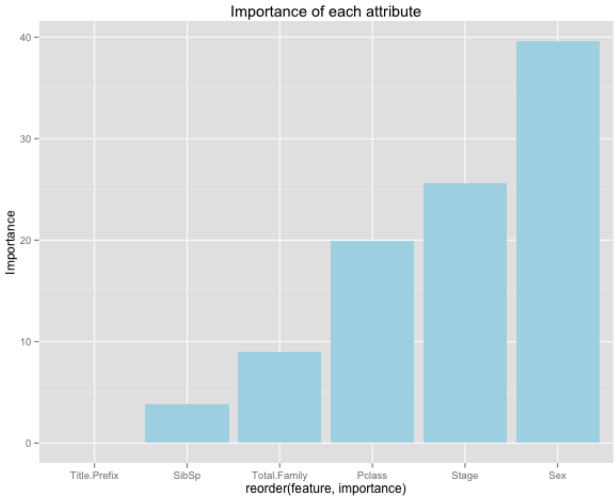

Fig. 5. Feature importance

Figure 6 displays the trend of the training error against the number of trees. As we can see on the black line, the OOB or out-of-bag error, was unstable during the beginning of the training step. Eventually, after more trees were being added to the forest, the error converged.

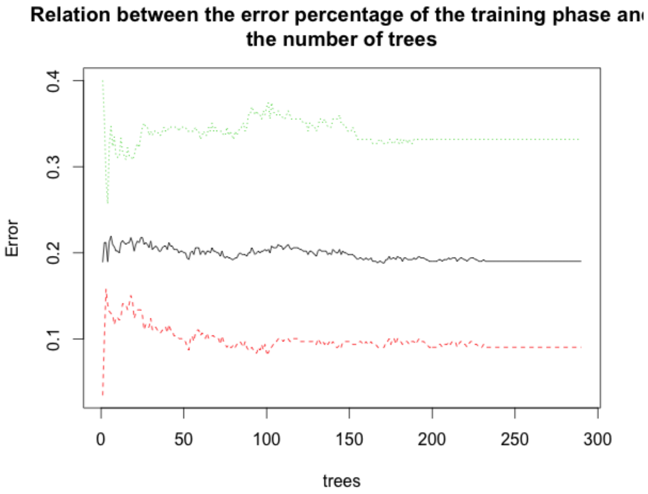

Fig. 6. Error percentage and the number of trees

IV. CONCLUSION

The results obtained during this analysis were satisfactory. The random forest classifier performed with an average training error of 18.92%, and a test error of 17.67%. On the other hand, we were expecting a more significant importance value for the `Title.Prefix` attribute. However, even though its importance was 0, we noticed an improvement during the training phase while using it.

REFERENCES

[1] W. Lord, *A night to remember*. Macmillan, 1955.

[2] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[3] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.