MATH4432

Result Report of Mini-Project 2, Topic 1

LAU, Wing Shing – 20342662

1 Introduction

This report is written to summarize the result of the findings in the competition named Titanic: Machine Learning from Disaster under the organization by Kaggle. The competition is aimed to analyze the type of people that is likely to survive from the disaster through studying the training dataset and predict the survival of people in the test dataset. The data fields of the training dataset are shown as below.

Variable	Definition	Key
Survived	Survival	0 = No, 1 = Yes
Pclass	Ticket Class	1 = 1st class, 2 = 2nd class, 3 = 3rd class
Name	Name of the person	
Sex	Sex	
Age	Age in years	
SibSp	Number of siblings or/and spouses aboard the Titanic	
Parch	Number of parents or/and children aboard the Titanic	
Ticket	Ticket Number	
Fare	Passenger Fare	
Cabin	Cabin Number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Table 1: Data Field of Training Dataset

2 Method

2.1 Preparation of Data

After the examination of all the data in training set and test set, there are missing data in both datasets.

Variable	Number of missing data
Survived	418
Age	263
Fare	1
Cabin	1024
Embarked	2

Table 2: Missing data in both datasets

Among the categories above, the missing data in Cabin is not considered to be filled in because there is no information about how the cabins have been named. Therefore, due to the insufficient information of data, the data column Cabin will not be considered in this finding.

There are 418 missing data in Survived, which is equal to the number of rows of testing data. These missing data is going to be predicted in later sections.

Regarding to other data columns with missing values, different filling methods have been applied for each of the columns.

For the data column Age, as there are hundreds of missing data and there are not enough data columns that are correlated to the prediction of Age, k-nearest neighbours method is not considered in this case. Instead, the missing values in Age is filled by the **complete** function inside the **mice** package. The model that has been applied in the **complete** function is the **random forest regression model**. The following data columns are used for the predictor variables.

Pclass, Sex, Age, SibSp, Parch, Fare and Embarked

For the data column Embarked, intuitively the embarkment of the people is related to the ticket class and the fare of the tickets they have paid. Under investigation, the 2

customers with missing values in Embarked are having 80 in both of their Fare, and 1 in both of their Pclass. From the observation of the plot below, Fare with the value 80 is relatively close to the median of the Pclass 1 in Embarked C. Therefore, both of the missing values in Embarked are filled with C.

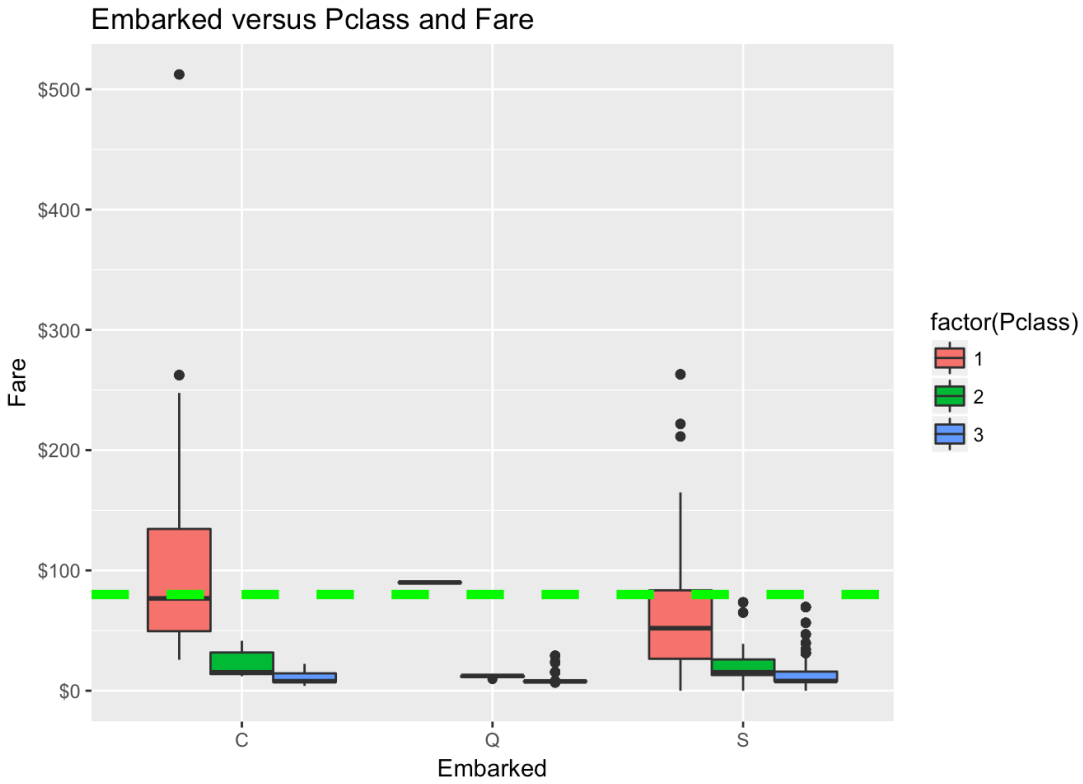

Figure 1: Embarked versus Pclass and Fare

For the data column Fare, the only customer with missing value is found to have Pclass 3 and Embarked S after investigation. Therefore, the missing value is filled with the **median** of the Fare of all the customers with Pclass 3 and Embarked S excepting that customer.

2.2 Observation of Data

There are some of the questions that have been interested in the finding.

- Are names of the people related to the survival rate?
- How are the ages of people related to the survival rate?
- How do family relatives affect the survival rate of a person?
- Are ticket fares related to the survival rate?

In purpose of answering these questions and help deciding the prediction model of the survival, several plots have been generated to make some observations on the data.

2.2.1 Are names of the people related to the survival?

After the observation of the raw data in both datasets, we found that the data column Name is consisted of the surname, first name and title of a person. Among these three parts, title of a person is important for the finding as it classifies a person into five main groups, including **"Master"**, **"Mr"**, **"Miss"**, **"Mrs"** and other **Rare Titles**, where **"Miss"** and **"Mrs"** reveal the marital status of a woman.

When the titles and sex of people are compared to the survival of people by using training set, the mortality of males is significantly greater than the female, and the mortality of people with **"Miss"** is greater than people with **"Mrs"**.

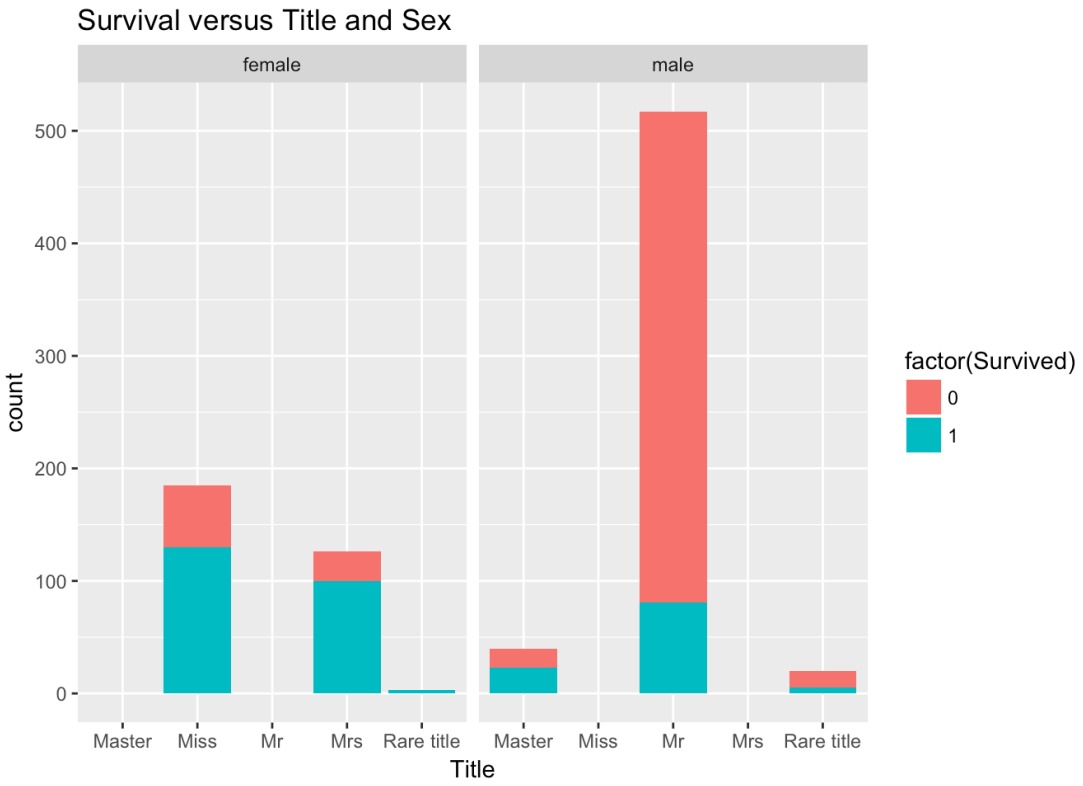

Figure 2: Survived versus Title and Sex

2.2.2 How are the ages of people related to the survival rate?

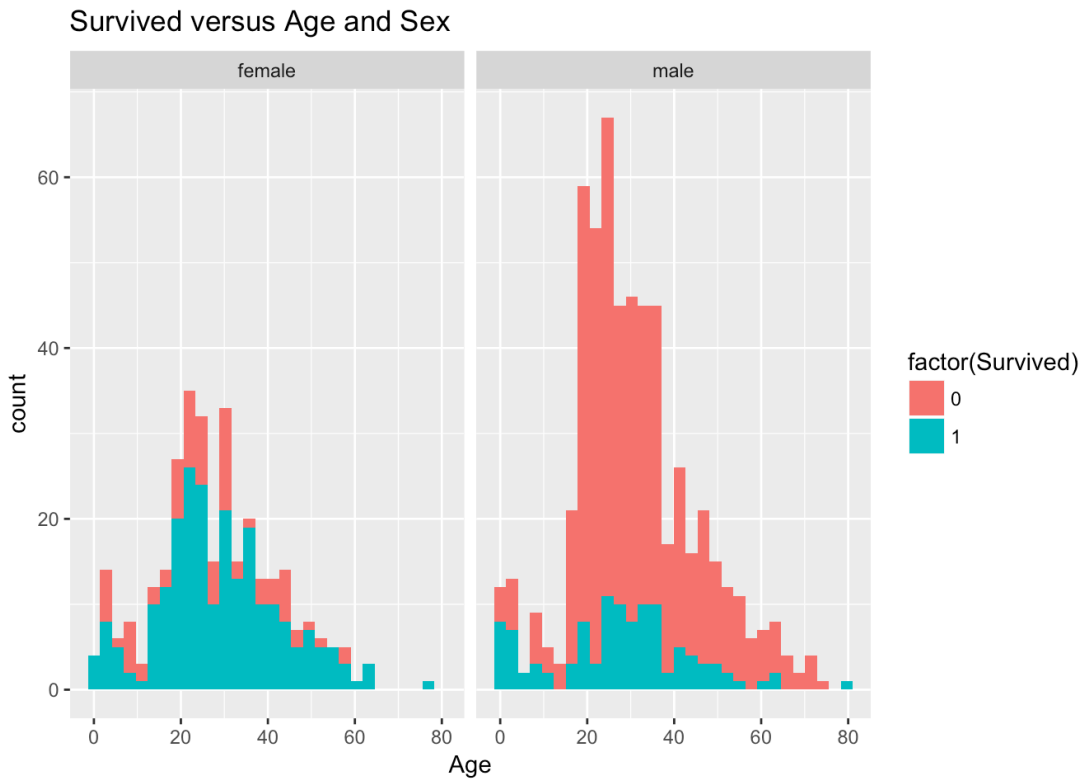

Figure 3: Survived versus Age and Sex

From the figure above, we can see that the people with age around 20 to 30 have a higher survival rate than other groups of age.

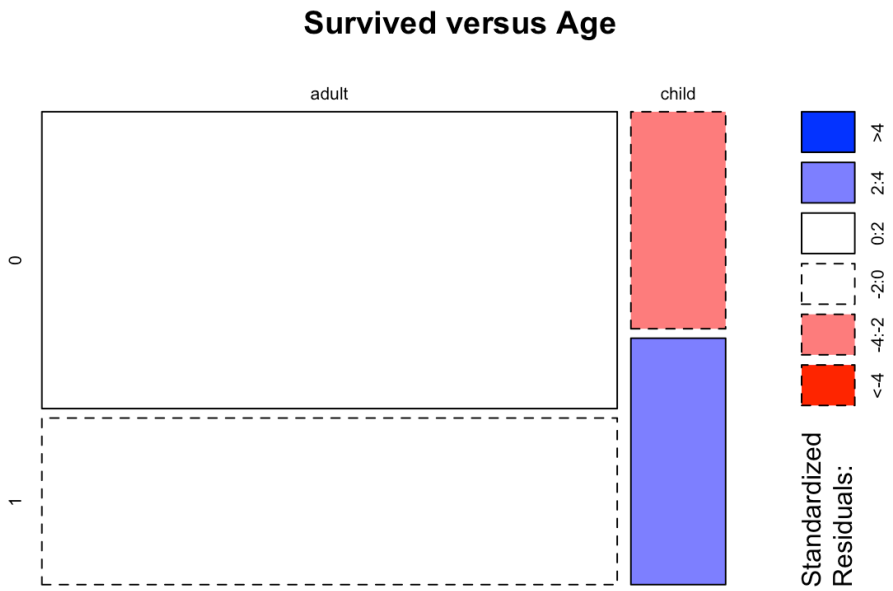

Figure 4: Survived versus Adult and Children

If we divide the group of ages into **child** and **adult**, we can conclude that children have a higher survival rate than the adults.

2.2.3 How do family relatives affect the survival rate of a person?

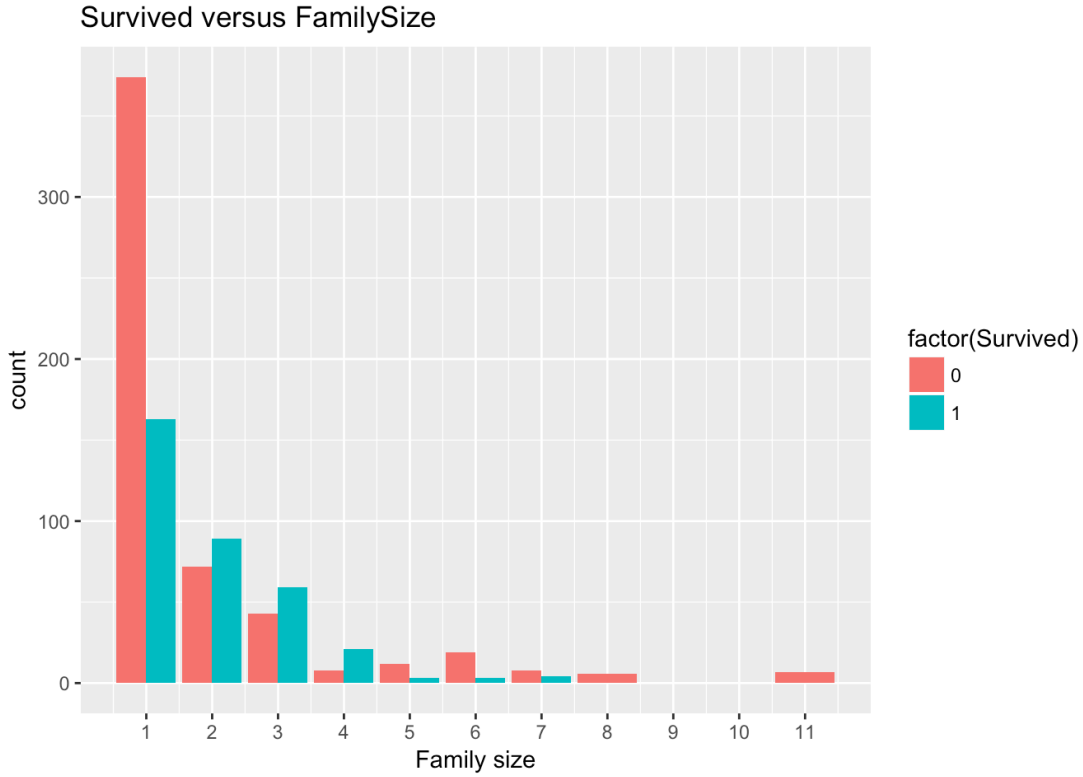

Figure 5: Survived versus Family Size

Before the plot is produced, the size of a family of a person is considered as

$$\sum(SibSp + Parch + 1)$$

where 1 stands for himself/herself.

From the plot of survival rate versus the family sizes of people, we can observe that people with family size of 1-4 have higher survival rate than the others. Therefore, family size of a person is negatively correlated to the survival rate.

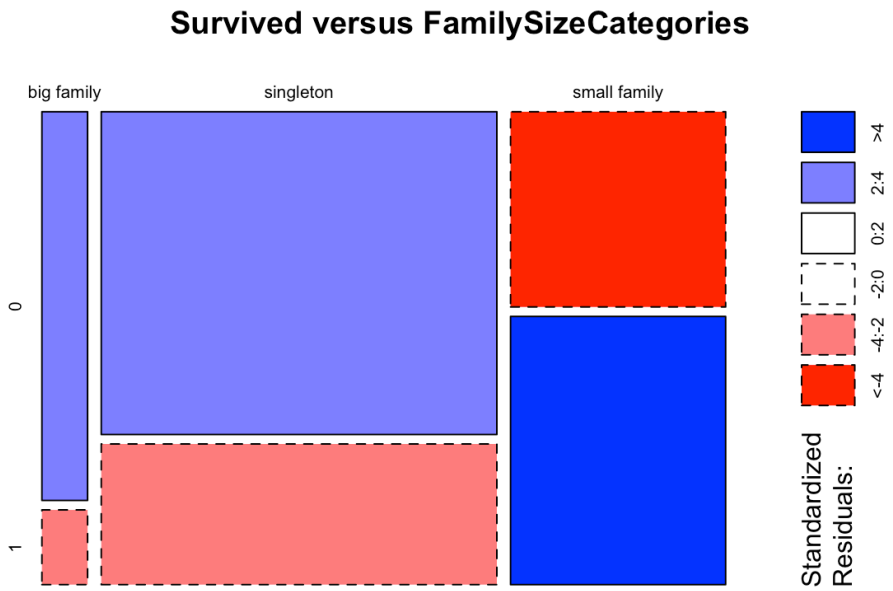

Figure 6: Survival Rate among different family groups

We define **singleton** as the people with family size 1, **small family** as people with family size 2-4, and **big family** as people with family size greater than 4. From the figure above, small family has relatively high survival rates.

2.2.4 Are ticket fares related to the survival?

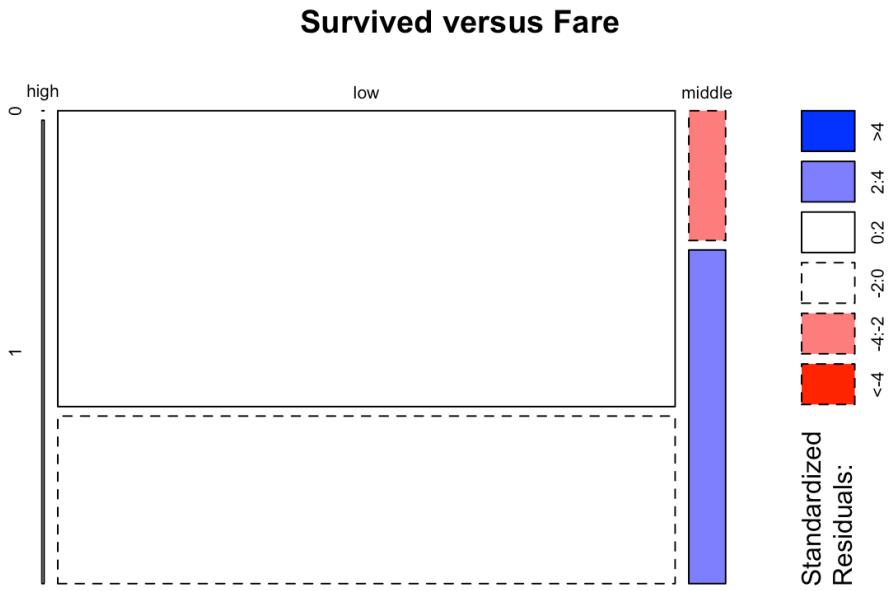

Figure 7: Survival Rate versus Ticket Fares

We define the ticket fares into three categories, namely high, middle and low. Fares with values higher than 300 is defined as high ticket fares. Fares with values in the range of 100-300 is defined as middle. Fares with values lower than 100 is defined as low.

From the figure above, we can observe that people with high fares are all survived, less than a half of people with low fares are survived and nearly three quarters of people with middle fares are survived. Therefore, we can conclude that people with lower ticket fares are likely not to survive.

2.3 Prediction on Test Data

Random forest classification is chosen as the model for predicting the survival rate in Test dataset. The response variable is the Survived data column in Train dataset. The predictor variables are chosen as

Pclass, Fare, Embarked, Title, Sex, Family Size and Child

in Train dataset.

Title is a new defined column for all data which contain five groups of titles specified in Section 2.2.1, including levels of **"Master", "Mr", "Miss", " Mrs" and "Rare title"**.

Family Size is a new defined column for all data which contain three groups of family size specified in Section 2.2.3, including levels of **"singleton", "small family" and "big family"**.

Child is a new defined column for all data which contain two groups of age classification specified in Section 2.2.2, including levels of **"adult" and "child"**.

After fitting the model using random forest, we would like to see the error rates in classification of the survival rate of people.

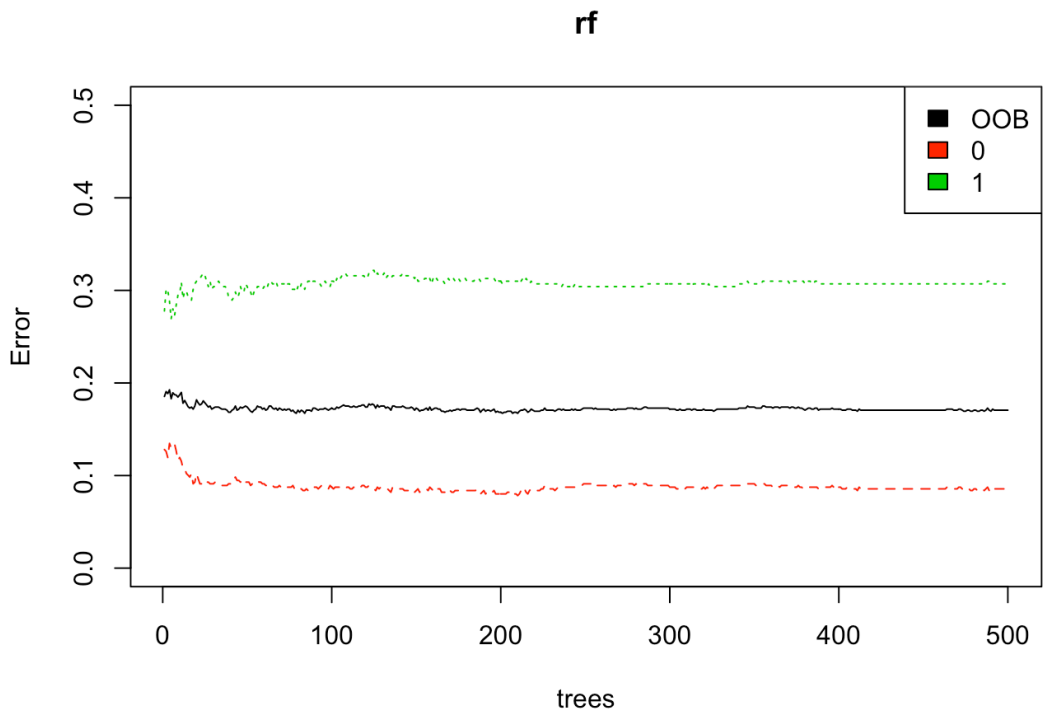

Figure 8: Error rate in classification of two survival levels

From the figure above, we can observe that as the number of trees is approaching to 500, the error rates for death and survival are 0.1 and 0.3 respectively. Therefore, the probability of correct decision of deaths would be higher than that of survivals.

3 Result of the prediction

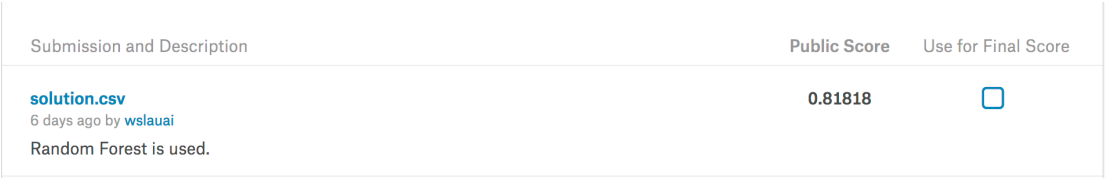

Figure 9: Score achieved after submission of the solution

4 Conclusion

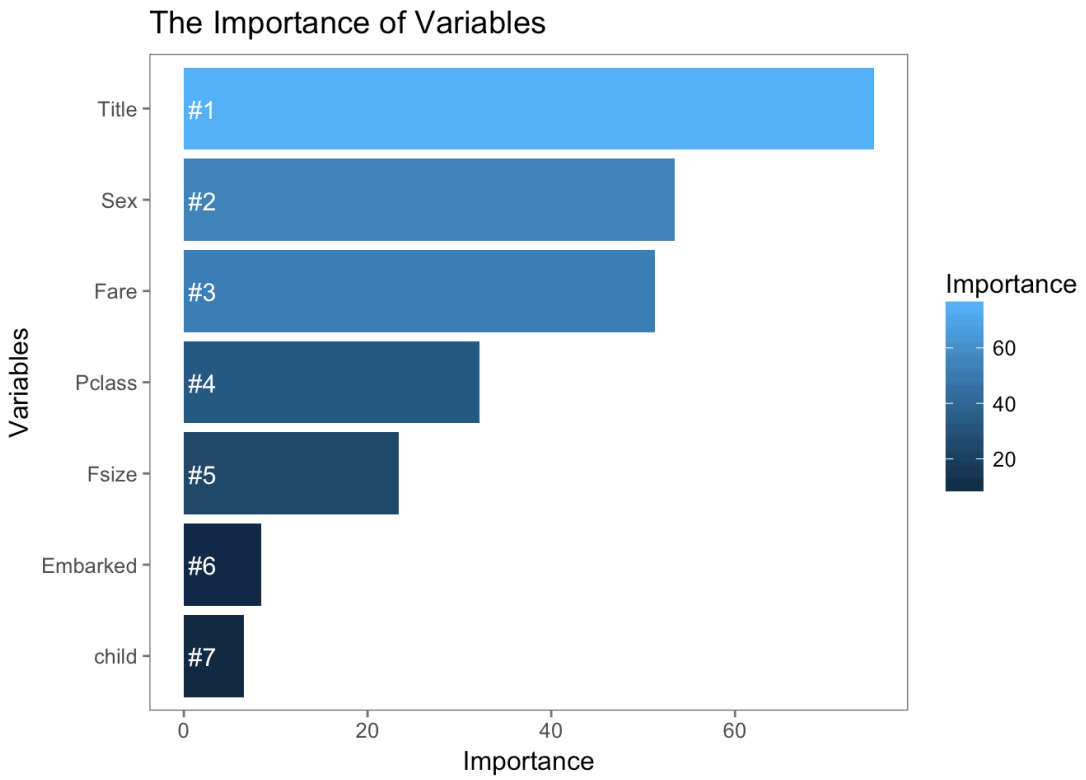

Figure 10: Importance of the predictor variables

Through the consideration of the importance of predictor variables in the random forest classification model, the classification of survival rate is mainly affected by the **Title, Sex and Fare** of a person.

From the observation of data in Section 2, we may conclude that (1) female has higher survival possibility than male, (2) Non-married females have higher survival possibility than married females and (3) Survival rate is positively proportional to the ticket fare paid by a person.

5 Reference

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017) *An Introduction to Statistical Learning with Applications in R*. Springer Science+Business Media New York.