Brandeis University

INTERNATIONAL BUSINESS SCHOOL

Titanic: Machine Learning from Disaster

Group Member:

Luhan Shen, Jiawei Zhang

Final Project 1

183BUS-211F-4: Analyzing Big Data I

Instructor: Prof. Ahmad Namini

December 19, 2018

1.Introduction

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean on 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. 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. We would like to summarize survivors' characteristics based on the data available and make predictions with machine learning model. This is a binary classification problem.

2.Data description

We collect data from Kaggle,Titanic: Machine Learning from disaster. The dataset contains information of passengers aboard.

Table1: variable description

Variable	Description
Pclass	Ticker class
Name	Name of Passenger's
Sex	Passenger's Sex
Age	Age of Passenger's
SibSp	Number of siblings/spouses aboard the Titanic
Parch	Number of parents/children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin Number
Embarked	Port of embarkation, C = Cherbourg, Q = Queenstown, S =Southampton
Survived	Survival, 0=No, 1=Yes

Kaggle provides train dataset with 891 observations and test dataset with 418 observations. There are missing values in age, fare and Embarked.

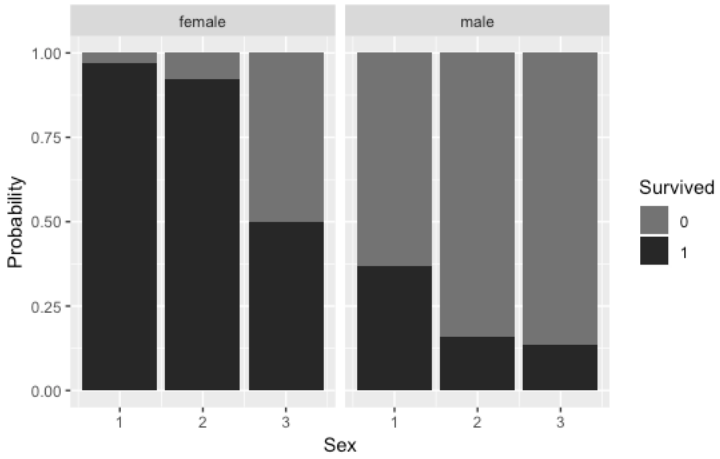
3. Data Analyzing

We present visualization of the most important variables in this part.

1. Sex and class

Female had high survival rate in both first and second class, nearly 90% of them got rescued. Half of women in third class survived. Both men and women are more likely to survive with higher class.

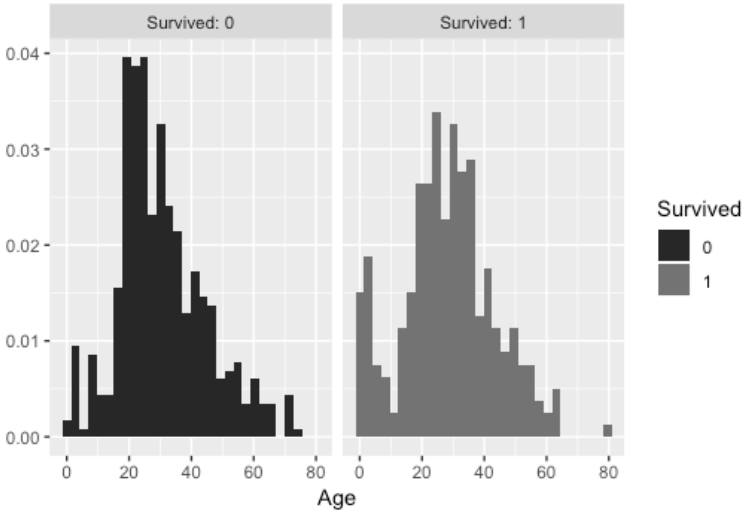
Figure 1: probability of survive/not under different sex and class



2. Age

Age has a tailed distribution. Babies under 5 years old and young people near 20 survived. Older than 40 have less chance to survive.

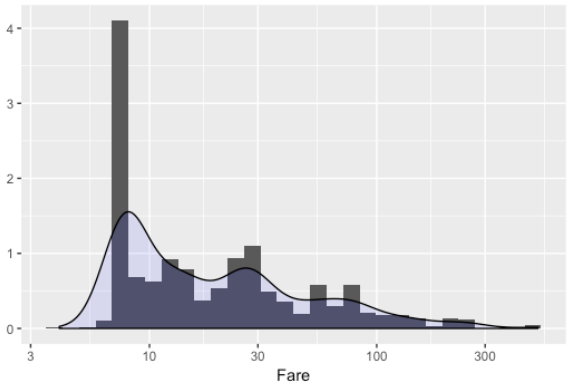
Figure 2: age histogram under survival/not



3. Fare

Though made log transformation, the fare is skewed, which may lead to overweight very high values in the model.

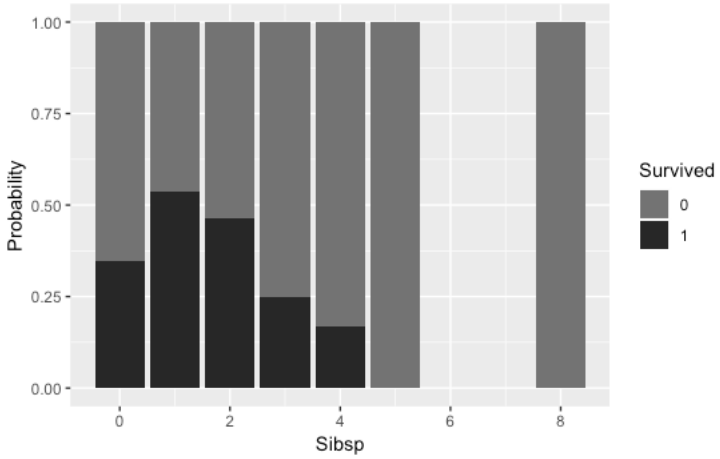
Figure 3: fare histogram and density plot



4. Sibsp

The family variables are similar, when the figure increase, both men and women have less chance to survive. As for Sibsp, single passengers or passengers with 1 or 2 siblings/spouses have more chance of survive. The more siblings and spouses on board, the less likely a person can survive.

Figure 4: probability of survive/not under different Sibsp



4. Feature Engineering

We find several facts interesting and decide to add them to the model to improve the accuracy.

1.Title

We also find there are lots of information inside the passengers' name, that is Suffix, representing social class, sex, age, marital status, etc. We aggregate all Suffix into six categories: (1) Mr. (2) Mrs.: Mrs., Mme. (3) Miss.: Miss., Mlle., Ms. 4. Master. 5.Royalty: Lady, Dona, the Countess, Don, Sir, Jonkheer 6.Ranked: Major, Dr, Capt, Col, Rev

2. Mother

We also find that mothers haver much higher survivor rate. We define woman older than 18 years old, have more than1 children, with suffix Mrs. as mother.

3. Age period

The babies and young adults have more survivor rate while older people are much less.

4. Family size

We find that family size has impact on survivor and split it into single, small, medium and large categories based on family member numbers.

Table2: Newly added variables

Variable	Description
Title	Suffix. Mr., Mrs., Miss., Master., Royalty, Ranker
Mother	Whether a passenger is woman.
Age period	Age. Babe: 1-5; Young 5-18; Adults: 19-25; Elder: 26-40, Old:40+
Family size	# of Family member. Single:1, Small: 2-3, Medium: 4-6, Large: >6

5. Data processing

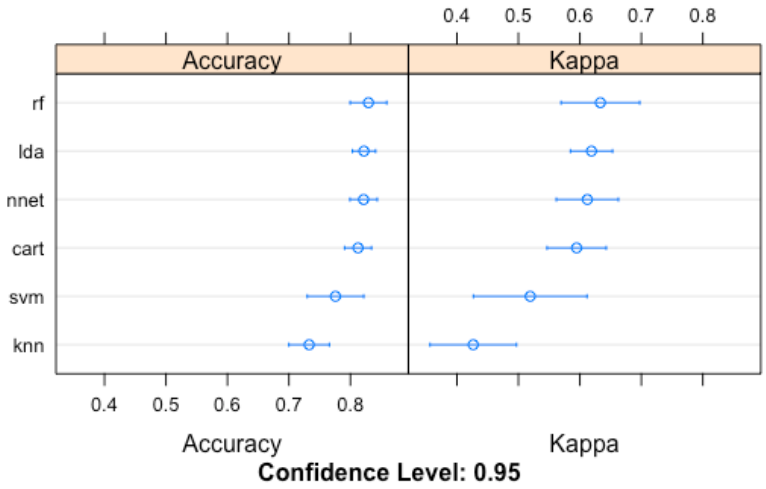
We mainly deal with missing data here.

Fare has one missing values, we fill it with the average ticket fare. Embarked has two missing values, we fill them with the port mode Southampton. Age has 177 missing values, we think it will cause biased if we just fill it with mean or median. We use recursive partitioning method (rpart) to predict age for NA rows which is binary tree process.

6. Algorithms Evaluation

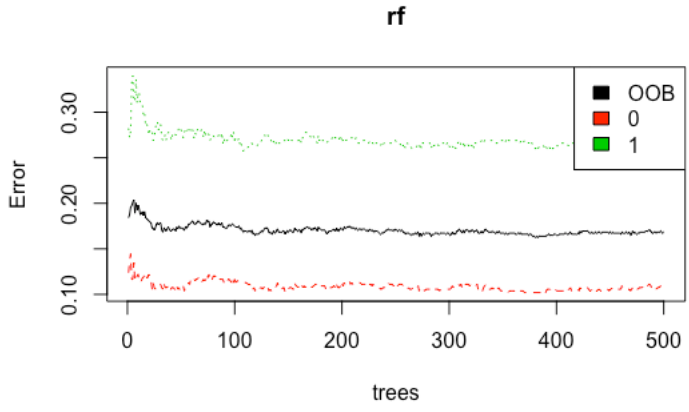
We run 6 algorithms (random forest, linear discriminant analysis, classification and regression tree, k-nearest neighbors, support vector machines and neural network) and use k-fold cross validation to estimate accuracy of models. The results show random forest is the best models under metrics accuracy or kappa and is much better than other algorithms.

Figure 5: algorithm accuracy and kappa



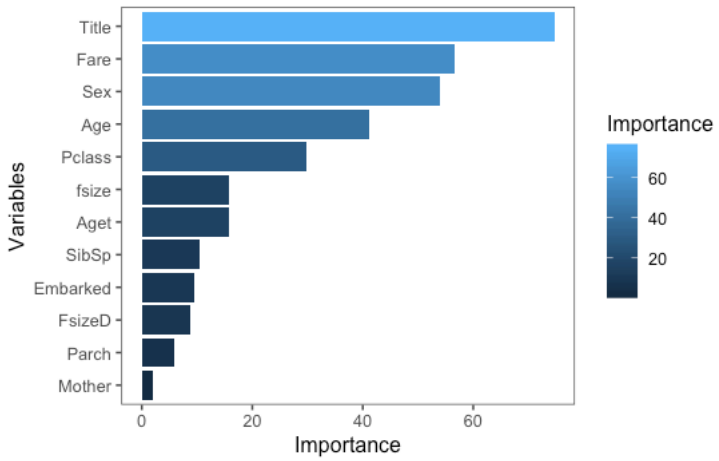
We look deeply into the random forest model and draw the model error plot. The black line shows the overall error rate which falls below 20%. The red and green lines show the error rate for 'died' and 'survived' respectively. We can see that the random forest algorithm is more successful predicting death than survival. The reason might be related with data itself which has more death samples (62%).

Figure 6: random forest error plot



We calculate the relative variable importance by plotting the mean decrease in Gini calculated across all trees. We can tell that Title, Fare, Sex, Age, Pclass are among the five most important factors which is in accordance with our guess.

Figure 7: variable importance plot



7. Model Validation

We use the test dataset Kaggle provided to test the accuracy of random forest algorithm. The result is 0.79, which ranks 3,500 in the titanic prediction competition.

8. Improvements

1. Missing values

We fill the NA of Age with the help of recursive partition (rpart) methods. However, the age distribution is different from previous one. It fills lots of NA will age between 20 to 25. It might be or might not be the case.

2. Accuracy

The accuracy is 0.79, far less than 1. We think there still are much improvements we can do such as trying different sampling methods, trying different parameters and algorithms. However, we should learn deeply into machine learning which can provide us with theory guide. In this way, we can get a better model quicker and more accurate.

APPENDIX:

Figure 8: family size and survival probability

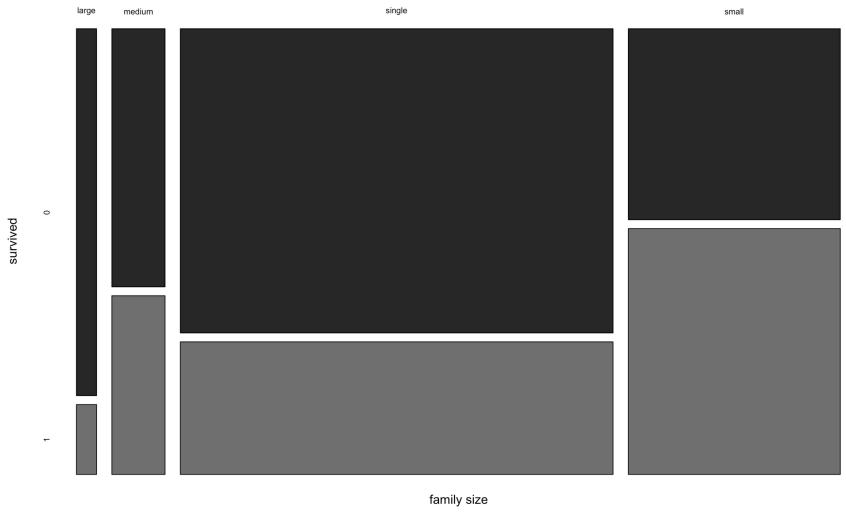


Figure 9: age density plot with survival/not