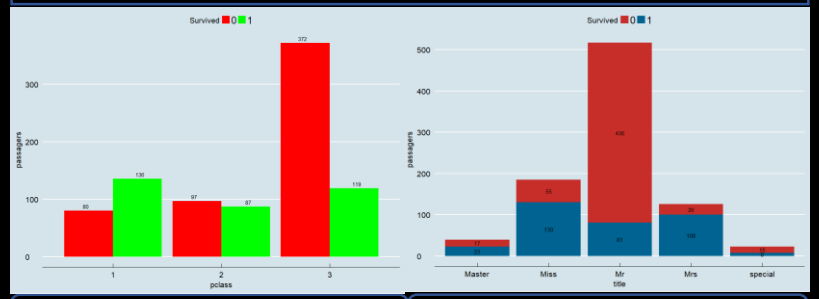
MATH 4432 Final Project: Titanic - Machine Learning from Disaster

Chow Wing Ho - 20279607

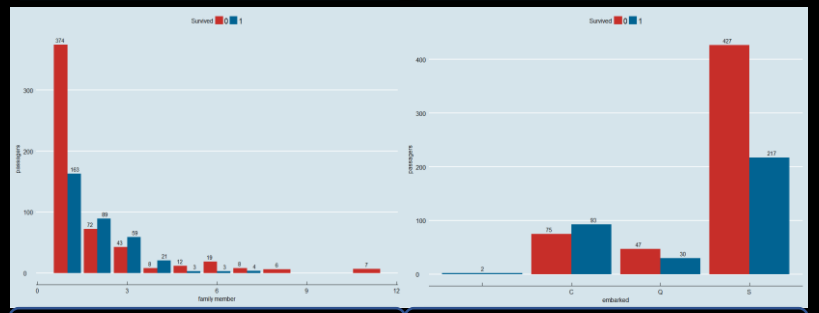
Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. Then, this project is going to find out the chance to survive in the Titanic with different factors.

Relationship between factor and survival



Impact for survival of Pclass | **Impact for survival of title**



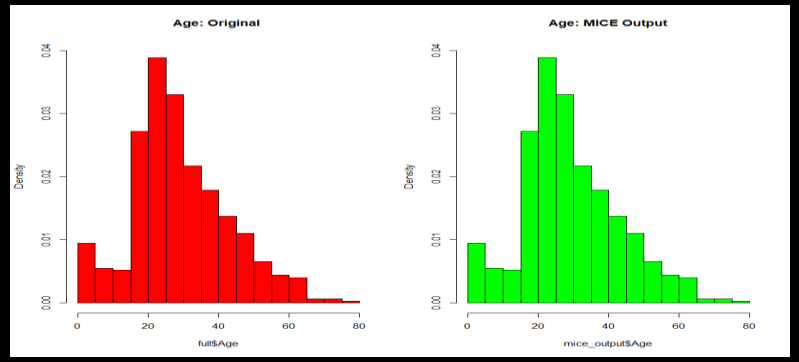
Impact for survival of family members | **Impact for survival of embarked**

The passages of Pclass = 1 are over 50% survived. The passages of title "Miss" are over 70% survived. The rate of survival is over 50% which family member is 2-3. The embarked is "S" that the dead rate is over 60%.

Data processing

Data cleaning for missing values

- There are four features with missing data which are **Fare, Cabin, Embarked and Age**
- For Fare, using the median fare of Pclass = 3 and embarked = "S" to replace the missing value.
- For Cabin, this factor doesn't consider as factor for random forest since it contains a lot of missing values that hard to replace.
- For Embarked, Since these two missing values were $80 for 1st class which are likely come from "C". Then, using "C" to replace the missing values.
- For Age, performing the mice imputation which exclude some useless variables to replace the whole Age variable.

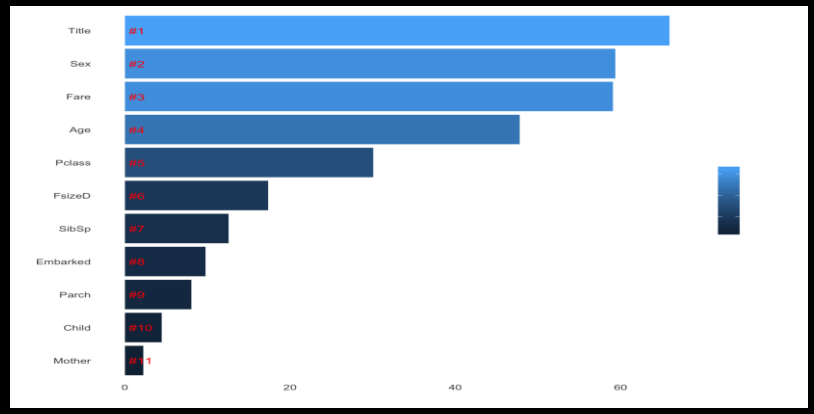


- For Title, using "special" to group "Dona", "Lady", "the Countess", "Capt", "Col", "Don", "Dr", "Major", "Rev" and "Sir", "Jonkheer". Use "Miss" for "Mlle" and "Ms" and "Mrs" for "Mne".
- For "SibSp" and "Parch", the method is creating the new variable "Fsize" to sum of "SibSp" and "Parch". Then, generating the variables "FsizeD" to define the family size is single, small and large.
- Creating the new variable "Child" to define the age of passages are child or adult.

First prediction

For 1st prediction, I use the random forest to predict the survival chance of test data set.

- rfmodel <- randomForest (Survived ~ Pclass + Sex + Age + SibSp+ Parch+ Fare+ Embarked+ Title+ FsizeD+ Child + Mother, data = train)
- the result is 0.80382



The histogram showed the variables "Title", "Sex", "Fare" and "Age" are key variables for prediction.

New variables

After the 1st prediction, there are two new variables which important to the prediction

- The 1st variable is "TicketCount" that the Ticket count is share may both survival or death. Then, the 1st group is using the unique count and the 2nd group is using the share count.
- The 2nd variable is "FamilyID" that the same surname passages may family which survived or dead together and surname may contain some identification.

Modification prediction

For 2nd prediction, I use the random forest to predict the survival chance of test data set.

- rf_model <- randomForest (Survived ~ Pclass + Sex + Age + SibSp+ Parch+ Fare+ Embarked+ Title+ FsizeD+ Child + Mother, data = train)
- The result is 0.82775 and ranked 214.