TITANIC Project

Introduction

In 1912, there was one of the most devastating maritime disaster in history. RMS Titanic sank in the Atlantic Ocean during its maiden voyage from UK to New York City after colliding with an iceberg. Sinking of Titanic claimed lives of 1,514 passengers. For this project, survival of Titanic ship will be predicted by using passengers' information.

Overview

For this project, Titanic dataset from Kaggle competition will be used. Titanic dataset consist of 891 observations and 12 variables

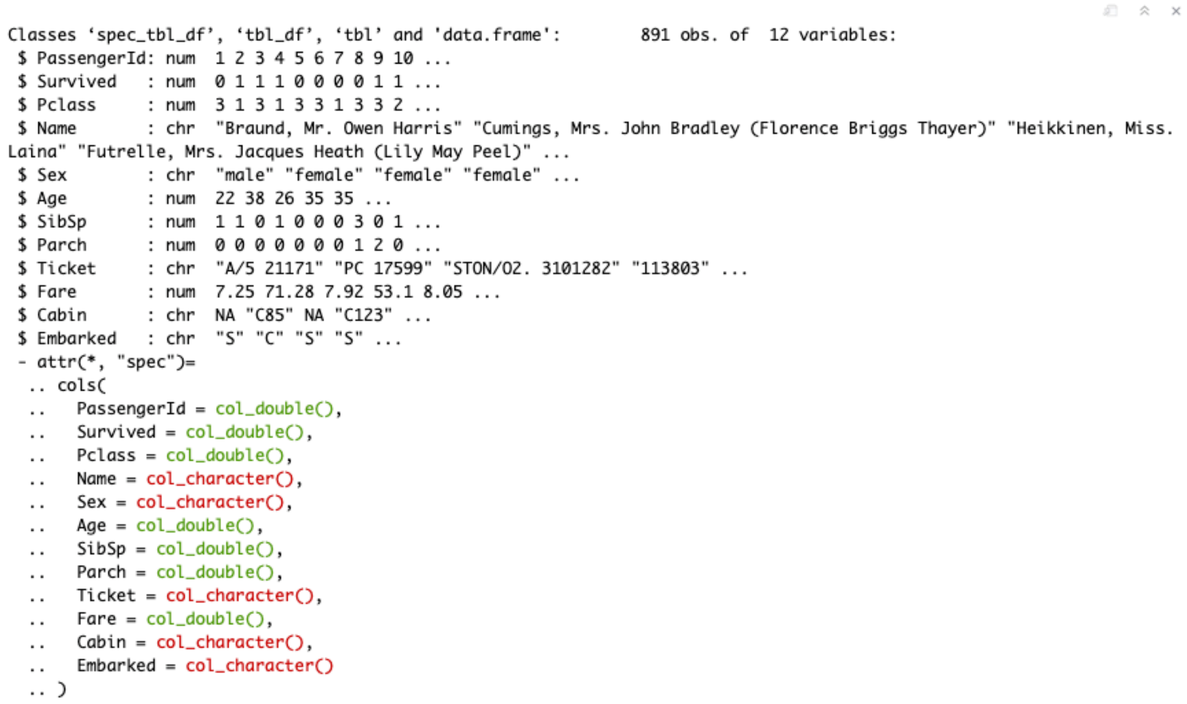
1.PassengerId – ID number of passenger
2.Survived – flag shoe that who are survival, this flag will be used as the target for prediction
3.Pclass – Class of ticket
4.Name – Name of passengers
5.Sex – Gender of passengers
6.Age - Age of passengers
7.SibSp - Number of siblings on board
8.Parch - Number of Parents on board
9.Ticket -Ticket Number
10.Fare - Price of Ticket
11.Cabin - Cabin Number
12.Embarked - Port of Embankment

Executive Summary

For comparing performance of predictive algorithm, basic model such as the gender-class model will be base model performance and then model performance will be improved by using more sophisticated algorithm. Titanic survival prediction model will be built by using passengers' information. First, the gender-class model will be used. Then, Decision Tree Model will be used to improve model performance by including factors that has potential relationship to the target. Finally, Random Forest, which is a sophisticated algorithm, will be used to boost model performance.

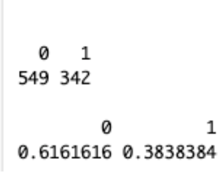
Data Exploration & Visualization

>str(df)- Check Structure of dataset



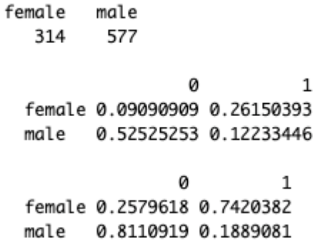
```
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame':       891 obs. of  12 variables:
 $ PassengerId: num  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : num  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : num  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. 
Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 ...
 $ SibSp      : num  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : num  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  NA "C85" NA "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
 - attr(*, "spec")=
  .. cols(
  ..   PassengerId = col_double(),
  ..   Survived = col_double(),
  ..   Pclass = col_double(),
  ..   Name = col_character(),
  ..   Sex = col_character(),
  ..   Age = col_double(),
  ..   SibSp = col_double(),
  ..   Parch = col_double(),
  ..   Ticket = col_character(),
  ..   Fare = col_double(),
  ..   Cabin = col_character(),
  ..   Embarked = col_character()
  .. )
```

>table(df$Survived), >prop.table(table(df$Survived)) - Check how many survival



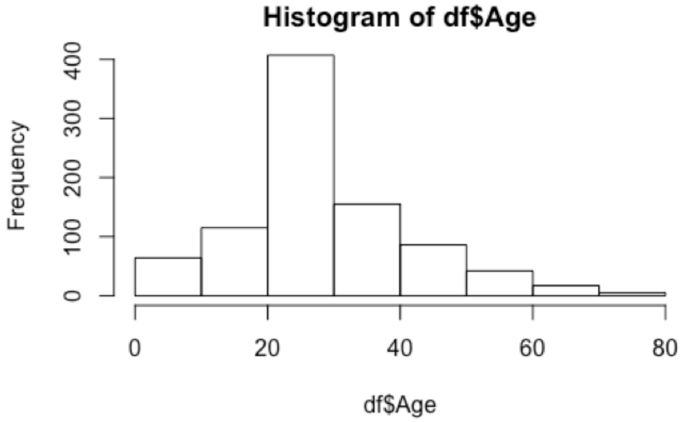
```
  0   1 
549 342 

        0         1 
0.6161616 0.3838384 
```

>table(df$Sex), >prop.table(table(dfSex, dfSurvived)), >prop.table(table(dfSex, dfSurvived),1) - Check passengers' gender and survival rate of each gender

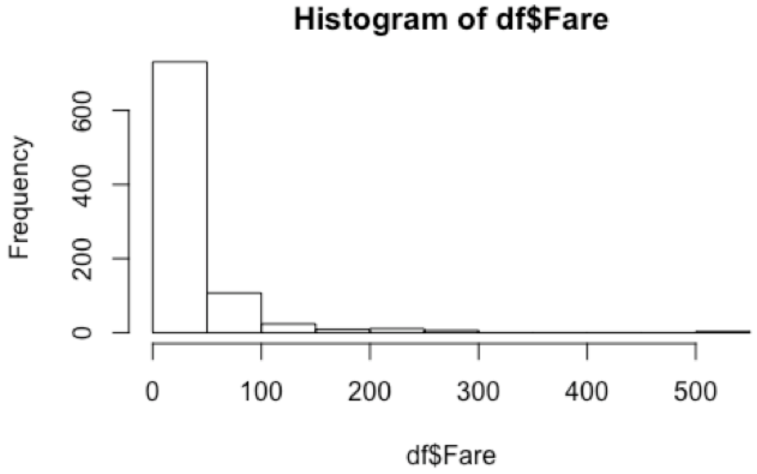


```
female   male 
   314    577 

                 0          1
  female 0.09090909 0.26150393
  male   0.52525253 0.12233446

                0         1
  female 0.2579618 0.7420382
  male   0.8110919 0.1889081
```

>hist(df$Age) - Check passengers' age distribution



>hist(df$Fare) - Check passengers' Fare

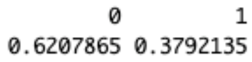


Predictive Model

0.1 Create Train-Test set

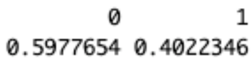
Titanic dataset will be separate into two datasets which are train_set and test_set. train_set will be used to train model and test_set will be used to evaluated model performance on unseen data.

Survival Rate : train_set



```
        0         1 
0.6207865 0.3792135 
```

Survival Rate : test_set



```
        0         1 
0.5977654 0.4022346 
```

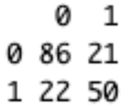
0.2 Performance Evaluation

For simplicity, model performance will be evaluated by using accuracy.

1. The Gender-Class Model

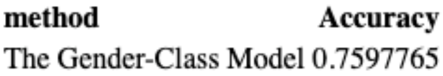
According to data exploration, 74.20% of female are survival and 81.12% of male died. Thus, for the gender-class model,
all female will be predicted to be survival.

The Gender-Class Model Confusion Matrix



```
   0  1
0 86 21
1 22 50
```

The Gender-Class Model Accuracy

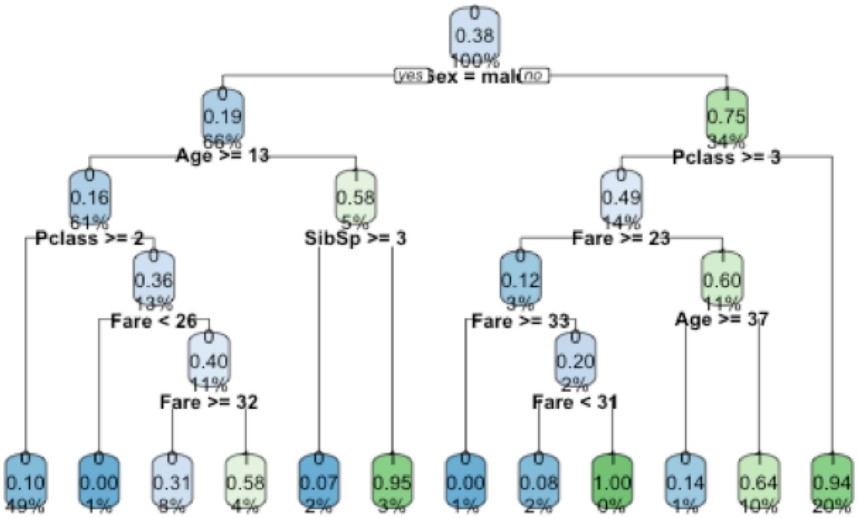


method	Accuracy
The Gender-Class Model	0.7597765

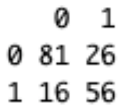
By using The Gender-Class Model, Accuracy of prediction is 75.98%. Next, Decision Tree which is more sophisticated
technic will be used to improve model performance.

2. Decision Tree Model

For decision Tree, six factors which are Pclass, Sex, Age, SibSp, Parch and Fare will be used to predict Survival. Decision Tree is basic classification model and this model can be plot as a tree as plot below.

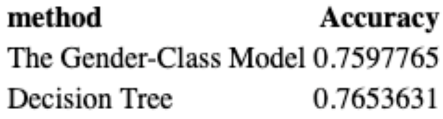


The Decision Tree Model Confusion Matrix



```
   0  1
0 81 26
1 16 56
```

The Decision Tree Model Accuracy



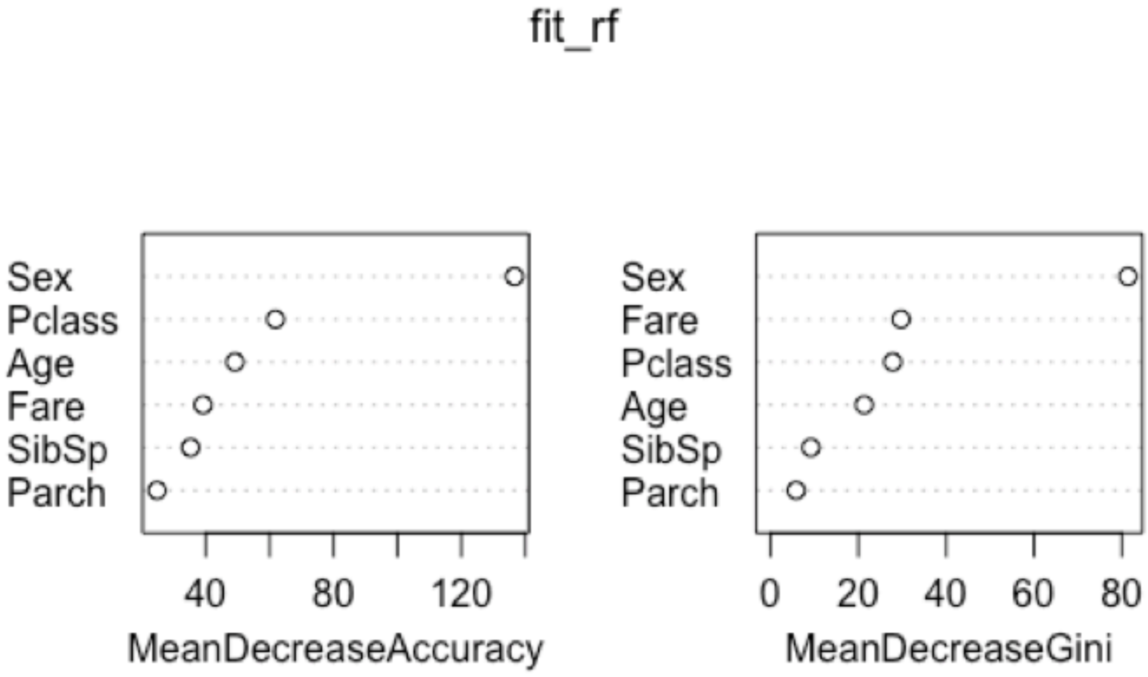
method	Accuracy
The Gender-Class Model	0.7597765
Decision Tree	0.7653631

By using Decision Tree Model, Accuracy of prediction is 76.54%, improved 0.56% from Decision Tree Model. Next, more sophisticated model called 'Random Forest' will be used to improve model performance.

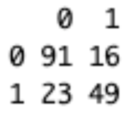
3. Random Forest Model

For random forest, same factors as decision tree model will be used. However, performance can be improved because of random forest sophisticated technique. Random Forest will make many decision trees and each tree can use only some factors for prediction. By combining, all tree together, variance of model will be decrease which improve model performance, this method call "Bagging".

Random Forest Feature Importance

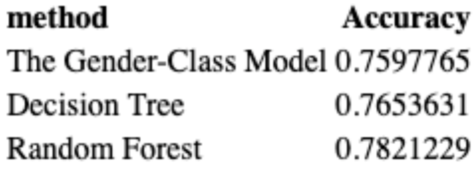


Random Forest Confusion Matrix



```
    0  1
 0 91 16
 1 23 49
```

Random Forest Accuracy

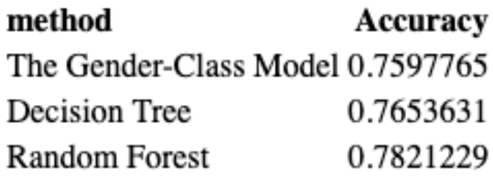


method	Accuracy
The Gender-Class Model	0.7597765
Decision Tree	0.7653631
Random Forest	0.7821229

Model Performance Summary

Start from base model called 'The Gender-Class Model', model performance was imporved my using 'Decision Tree Model' and 'Random Forest Model' which are more sophisticated model.

Table below show performance of each recommendation model.



method	Accuracy
The Gender-Class Model	0.7597765
Decision Tree	0.7653631
Random Forest	0.7821229

Conclusion

In this project, many classification models is used to solve TITANIC's survival prediction problem. First, base model called 'The Gender-Class Model', model performance was improved my using decision tree model which is basic model for classification task. Then, Random Forest Model, which is more sophisticated model, was used to further improve model performance.