COMP5310 Principles of Data Science, 2019sem2 Assignment 2 REPORT

Student name: Yonghe Tan, Yang Luo
Student ID: 490089255, 490087398
Unikey: ytan2413, yluo0006

Section 1a

Research Problem

About a century ago, a great cruise was under constructing. After few years, it was ready to sail. Millions of people considered it the dream ship and the apex of ocean technology. Thousands of people were lucky enough to get one ticket on board. 4 days later, it sank. It is Titanic. The worst maritime disaster ever.
We got a data set of passengers on board from Kaggle.com. It contains information of every passenger including name, sex, age, survived or not and so on. By unlocking the inner association of these information, it is possible to find out what factors could make a passenger to survive from an ocean disaster.
After research of stage 1, we have 3 basic knowledge of the data:

A. Gender, lady first rule, matters in this situation.
B. Age also matters in this situation. Young people are less possible to survive.
C. Pclass (ticket class, symbol of social class, power and so on) matters.

For Stage 2, we have 3 advance questions:

A. What about the importance of different features?
B. Given basic information of a certain person, can we struct a model to predict if he/she is alive or not?
C. Can we get a better prediction model?
D. Is our best model different from the baseline?

Section 1b

Evaluation Setup

We will evaluate importance level of different features by feature_importances_ class of Decision Tree and Random forest from sklearn library.
We will split data into train-test sets. By applying model to test set and passing the result, we can get accuracy and F1-score of a certain model. Due to imbalance distribution of our target data, we will mainly choose F1-score to judge whether a model work well or not and compare one model to other models with F1-score.
We will implement McNemar's test on the best model and the baseline. We can know that if there is significant difference between them.

Section 2

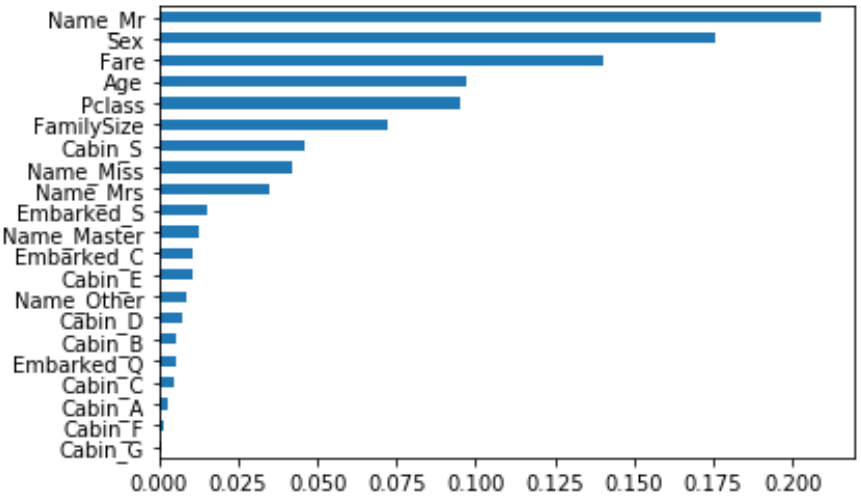
Approach Description

Our experiment will be implemented along the following steps:
Step 1: Define variable 'Survived' as the target variable.
Step 2: Data treatment. By this step, it is expected to eliminate factor which is not correlated or less correlated to the target variable and null data in valuable column will be filled in fair way. Dummy encoding for nominal data, data split for train/test set will be implemented.
Step 3: Model selection. The target variable is in Boolean type. Classification model such as Decision tree, Random forest and Logistic regression will be implemented. Naive Bayes method will be set up as baseline test. GridSearchCV will be implemented on each model to find the best parameter pair of each model in this data set. The best score and parameter pair of each model will be recorded.
Step 4: Ensemble model will be implemented; it consists of all classification models above. We will use 'soft' voting mode. The weight of each model will be adjusted by F1-score they hold. Also, accuracy and F1-score of ensemble model will be recorded.
Step 5: Conduct ROC and AUC of baseline model and the best model we got. McNemar's test on the best model and the baseline.

Section 3

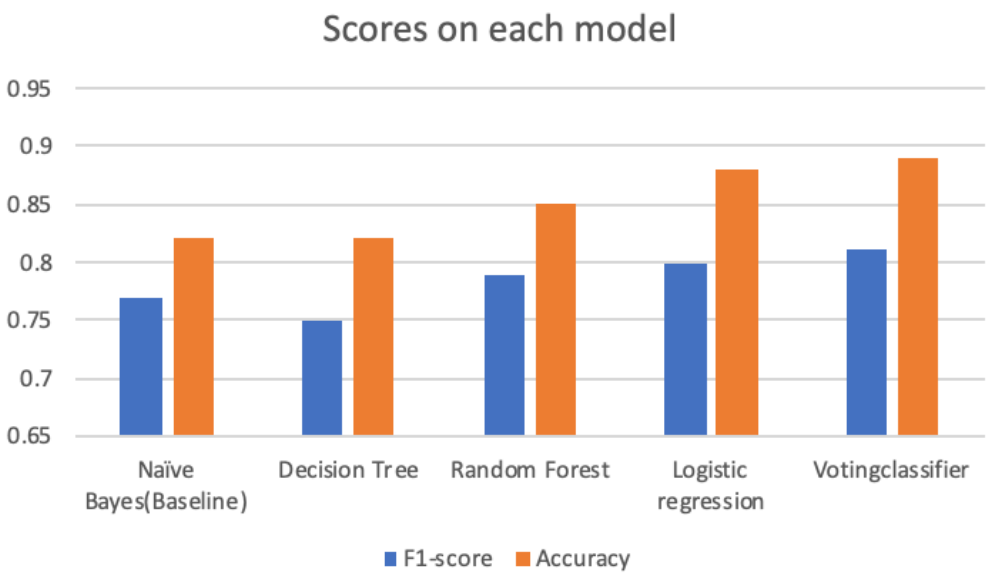
Result Presentation and analysis:

1. Feature importance from random forest:



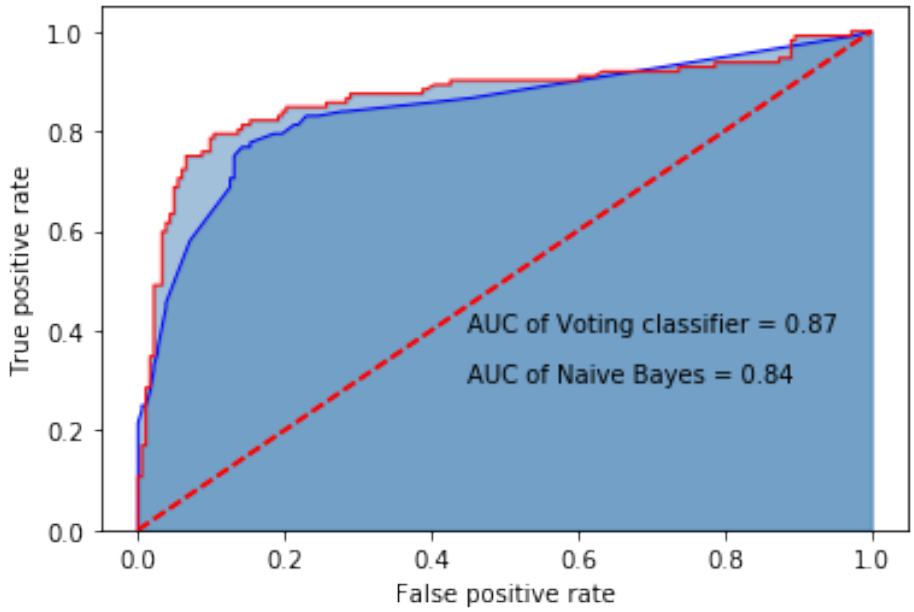
Analysis: Fare, age, FamilySize , gender and Pclass are the most important feature in the tree of random forest which satisfy our intuition and initial judgement. Because of the limited life boat, those who paid more fare and in high level Pclass may be in the upper class and more likely to survive. Women and children are also more likely to be rescued because of people's moral principle.

2. The scores of different models:



Analysis: As shown on table, we can get the information that all the models perform well and they can reach more than 80% on accuracy, and more than 70% on F1-score. Comparing to our baseline, all other classifiers are better. Especially, the voting classifier is the best model with 0.81 F1-score and 0.89 accuracy.

3. ROC and AUC on baseline and voting model:



Analysis: As shown on graph, the voting classifier performs better than baseline. Because its curve is closer to top left point(0,1) and the area under curve is larger than the baseline.

4. Hypothesis test between baseline and voting model:

   H0: there is no different between baseline and voting model.

   ```
   Can we reject H0?
    Yes, the p-value is too small
   P-value =  0.016156931261181322
   ```

   Analysis: As the result showing, the p-value is too small. So, we reject the null hypothesis, which means we favor that there is some improvement between voting classifier and our baseline. So, we choose the voting classifier.

Section 4

Conclusion:

1. The class of people, age and gender are the most important feature for people who are on titanic to survive after disaster.

2. Combining with stage 1 conclusion, we can draw a conclusion that women and children are more likely to survive and those people who buy expensive ticket are more likely to live. As for those who bought standing room ticket are more likely died in this disaster.

3. We recommend voting classifier as convincing model to predict whether people can survive in titanic event or not and there is efficient evidence that this model is better than our baseline.