CSC8111- Machine Learning-Report:

Before building the classifier model or predict model, it is necessary to clean the dataset by python. First import some important package for preparing, showing in figure 1.

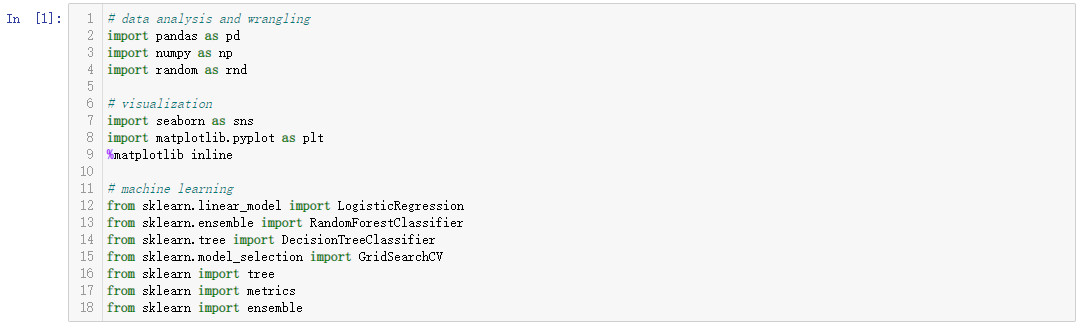


Figure 1

Then using ‘Pandas’ to input the source file test.csv and train.csv. After import the source file the original static is shown in figure 2 with its fist five line.

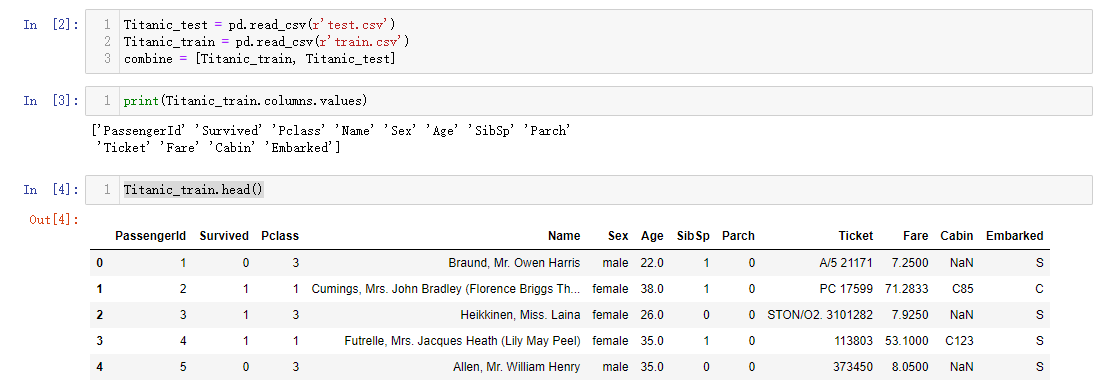


Figure 2

Next step is deleting meaningless variables, which can not affect the final result. Like ‘Name’, ‘Ticket’, ‘Cabin’, this kind of variables is only link to one specific passenger. No matter How it changed, there would not have any changes on the model. Also, the ‘Sex’ column should be converted from strings to numerical values which is required by most model algorithms. After this sort of operation the dataset is shown as figure 3.

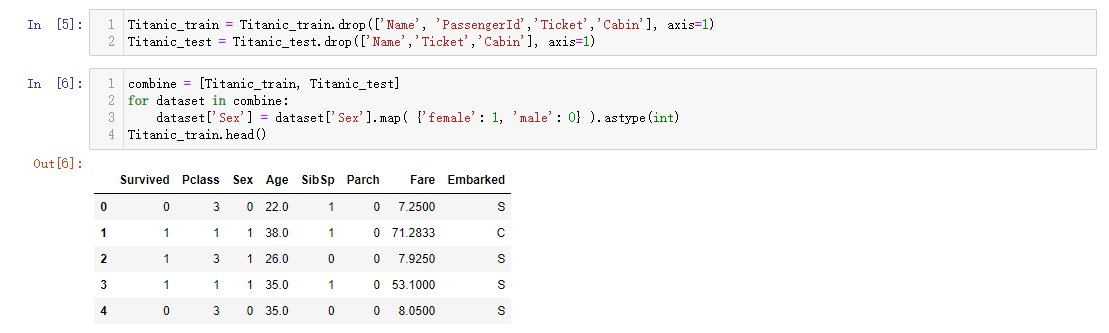


Figure 3

The ‘Age’ value should be estimated and completed with missing or null values.First preparing an empty array to contain guessed “Age” values based on “Pclass” \* “Gender” combinations. Meanwhile iterate over “Sex” (0 or 1) and “Pclass” (1, 2, 3) to calculate guessed values of “Age” for the six combinations. Then create Age bands and determine correlations with Survived, show in figure 4.

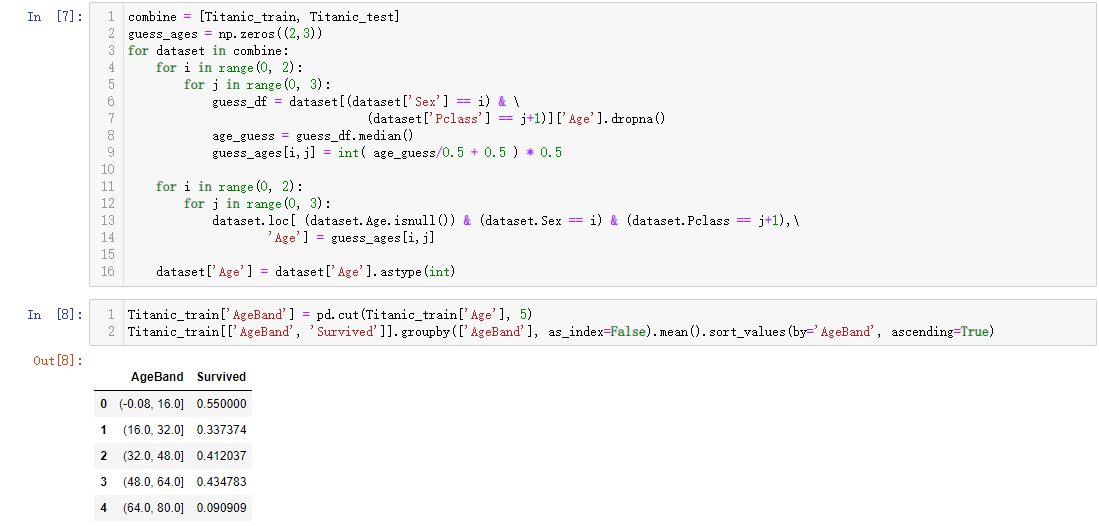


Figure 4

Replacing “Age” with ordinals based on these bands and then remove the “AgeBand” feature, the final result after deal with “Age” is shown in figure 5.

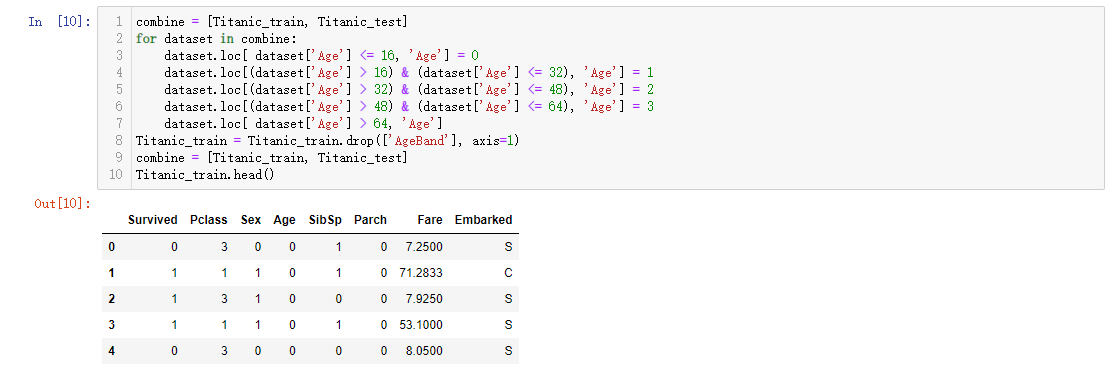


Figure 5

To make the results more accurate and simplify “Parch”, “SibSp”, create a new feature for “FamilySize” which combines “Parch” and “SibSp”. This will enable us to drop “Parch” and “SibSp” from the datasets. The result after operation is shown in figure 6.



Figure 6

Embarked feature takes S, Q, C values based on port of embarkation and convert them to “0”, “1”,”2”. The training dataset has two missing values and fill them with the most common occurrence. The result after operation is shown in figure 7.



Figure 7

Now complete the “Fare” feature for single missing value in test dataset using mode to get the value that occurs most frequently for this feature. This operation replacing only a single value. The completion goal achieves desired requirement for model algorithm to operate on non-null values. We may also want round off the fare to two decimals as it represents currency. Now create “FareBand”. Convert the “Fare” to ordinal values based on the “FareBand”. This operation and final result is shown in figure 8.



Figure 8

The final train and test datasets are shown in figure 9.

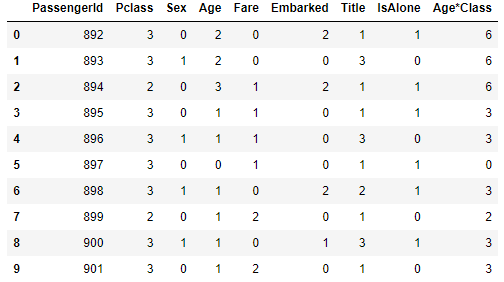
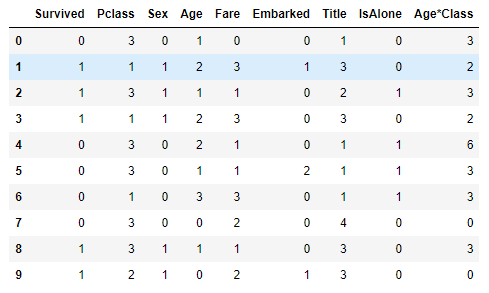


Figure 9(train) Figure 9(test)

In this competition, I used 9 models(Random Forest, Decision Tree, KNN, SVM, Logistic Regression, SVC, Linear SVC, Naïve Bayes, Perceptron, Stochastic Gradient Decent ) to fit, predict, and score the data set. The first fitting method used the model included in'sklearn', and the parameters were adjusted to the default parameters. The scoring method uses the ‘score’ function in the ‘sklearn’ library to predict the accuracy of each model.

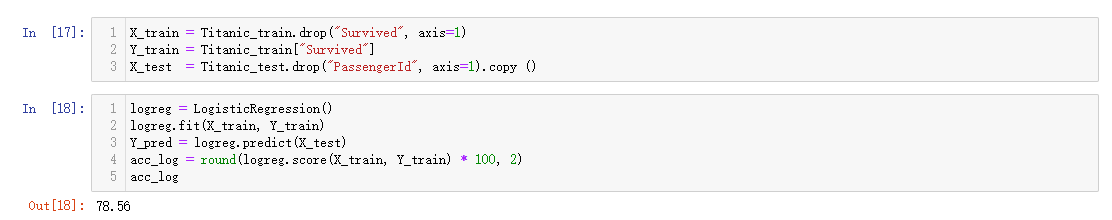


Figure 10

The accuracy of model prediction was ranked by model score function. The rank result in shown in figure 11.

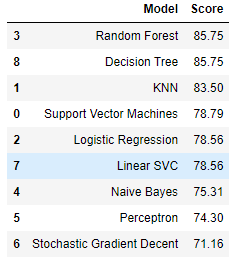


Figure 11

It shows that the prediction accuracy of Random Forest, Decision Tree and KNN ranked top three. But the score of Random Forest and Decision Tree are same. So, we have to do some optimization on them.

We use ‘GridSearchCV’ function to calculate the best ‘max\_depth’, ‘min\_samples\_split’, ‘min\_samples\_leaf’. Making some limit on tree’s growth by these parameters to avoid model overfitting.

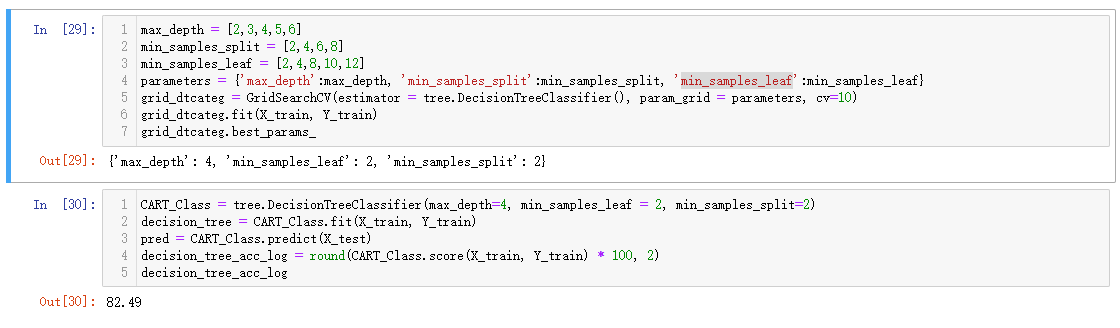


Figure 12

Also, we can draw the decision tree by ‘Graphviz’ which shown in figure 13 to get more information about the model and the forecasting process.

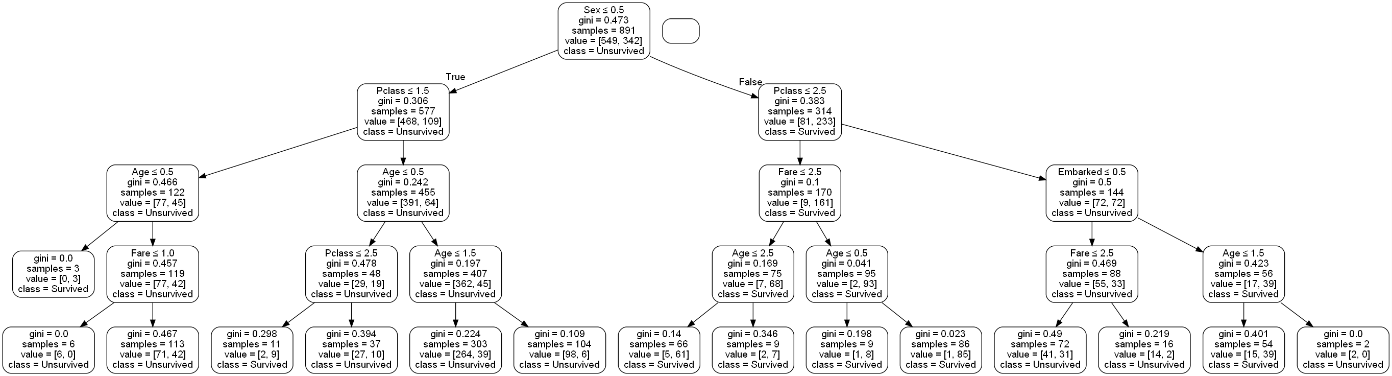


Figure 13

In order to obtain more accurate model prediction accuracy, ‘train’ can be divided into ‘train’ and ‘test’ again, the Proportion between them is 4:1. Take the Random Forest model as an example. The new test result is shown in figure 14.

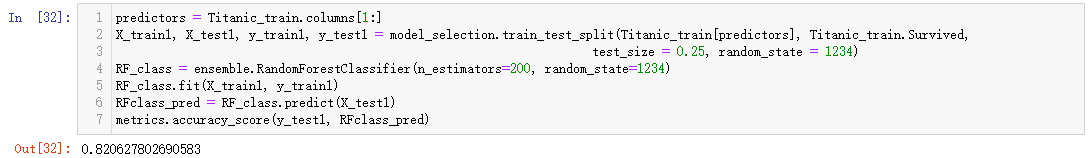


Figure 14

I also used the ‘receiver operating characteristic curve’ in order to get more model scoring information which shown in figure 15. The area value under the ‘ROC’ curve is 0.87, which shows the model has certain accuracy.

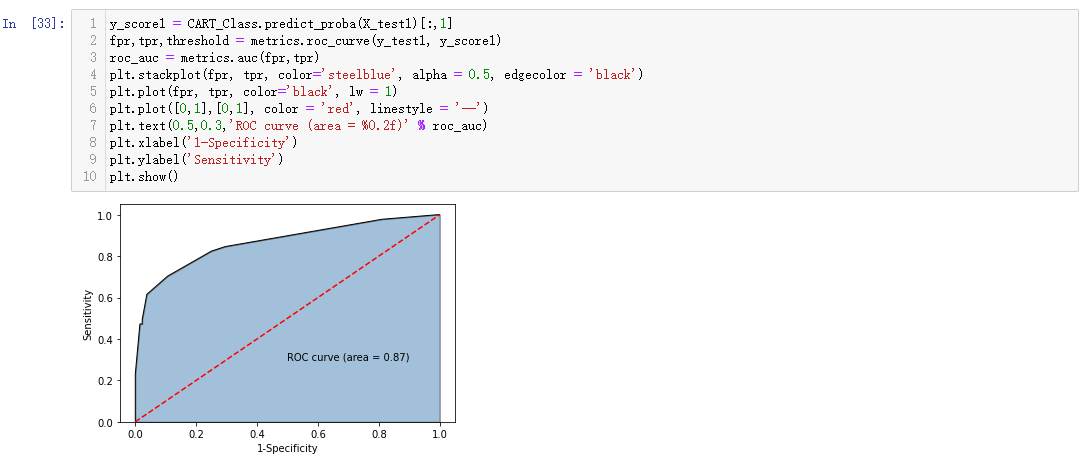


Figure 15

As we know in figure 11, Random Forest is the best model for doing predict task in ‘Titanic’ dataset. In order to better connect with the reality. We can use the ‘feature\_importances\_’ to make a judgment on the importance of all the ‘features’ in the data set. In order to determine which variables have the greatest impact on a person’s survival rate in a shipwreck event. The rank result in shown in figure 16.

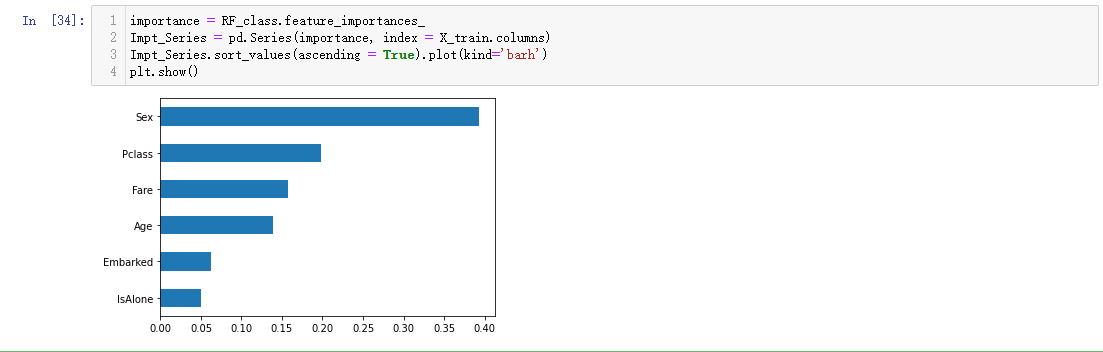


Figure 16

As can be seen from this table in figure 16, gender has the greatest impact on survival rate. Furthermore, we can see from the printed ‘decision tree’ that women are more likely to survive than men. This result may be because, from human relations, men naturally want to protect weaker person (Old man, woman, child). I think this result is correspond with the performance of social reality. In addition, ‘Pclass’ also had a great impact on the survival rate. This result may be because high-level cabins are usually located close to the deck. And around the high-level cabins, there will be more security personnel or staff to provide passengers. And safety-first aid measures for high-grade cabins may also be more than the lower cabins.