Project 1

After comparing several models, I use ExtraTreesClassifier model to fit the data and make a prediction. By using cross validation to test the accuracy of models, ExtraTreesClassifier is the best model I have tested.

ExtraTreesClassifier refers to the Extremely Randomized Trees, whose randomness goes further in the way splits are computed. This method is an enhanced version of Random Forests. In Random Forests, each tree in the integrated model is constructed with samples replaced by the training set. Furthermore, when node splitting is performed during the construction of the tree, the selected splitting point is no longer the best splitting point of all the features, but the best splitting point in the random subset of features. Therefore, the random subset of features is used in Random Forests, rather than look for the most discriminative thresholds, the thresholds here are randomly generated for each candidate feature, and the best of these thresholds is taken as the splitting rule. This method usually reduces the variance of the model a bit more, at the expense of slightly increasing the variance, which is used in Extremely Randomized Trees. In practice, I adjusted the parameter 'n_estimators' several times, and get the suitable value of this parameter, which is 175.

Because the training data contains the class label, thus, I adopted the supervised learning models to train the data set. Here I post the accuracy of several models selected.

```
The accuracy of ExtraTreesClassifier is:  0.954021547882
The accuracy of QuantileExtraTreesClassifier is:  0.954959023752
The accuracy of NormalExtraTreesClassifier is:  0.94594504283
The accuracy of ScaleExtraTreesClassifier is:  0.953098548173
The accuracy of DecisionTreeClassifier is:  0.906196134862
The accuracy of RandomForestClassifier is:  0.936324396094
The accuracy of SVMClassifier is:  0.939743471545
```

The method of evaluating estimator performance is cross-validation. In this project, the training set is split into 10 smaller sets, for each model, the score is computed 10 consecutive times with different splits each time, and the mean score stands for the corresponding model performance.

I tested four supervised learning models, from the picture posted, it is obvious that Extremely Randomized Trees is the best one to fit the data. In addition, I preprocessed the training data, and the second line in the picture shows the accuracy of Extremely Randomized Trees with quantile transformed data, which is the best among these seven results.

Consider about preprocessing, I tried several ways, including normalize function, scale function, quantiletransformer function.

QuantileTransformer belongs to non-linear transformation, it smooths out unusual distributions and is less influenced by outliers than scaling methods. The result is the best among these three methods of preprocessing data. Thus, it is inferred that the training data set is more suitable for non-linear transformation.

However, I found that the accuracy of models was different every time I ran the program. Therefore, it is inferred that the test results fluctuate within a narrow range. And what my picture shows is the overall situation.