Zhicheng Tang zt17
IE517
Module 7 Homework (Random Forest)

Part 1: Random forest estimators

Fit a random forest model, try several different values for N_estimators, report in-sample accuracies.

Cross validation of training data and gird search for n_estimators: [100, 200, 300] are conducted. The results are as follows:

```
Best parameter setting: {'n_estimators': 300}
Best cross-validation accuracy score: 0.8164
0.8153333333333334 {'n_estimators': 100}
0.8157777777777778 {'n_estimators': 200}
0.8164074074074074 {'n_estimators': 300}
```

From corss validation, random forest with 300 subtrees perform the best on in-sample accuray.

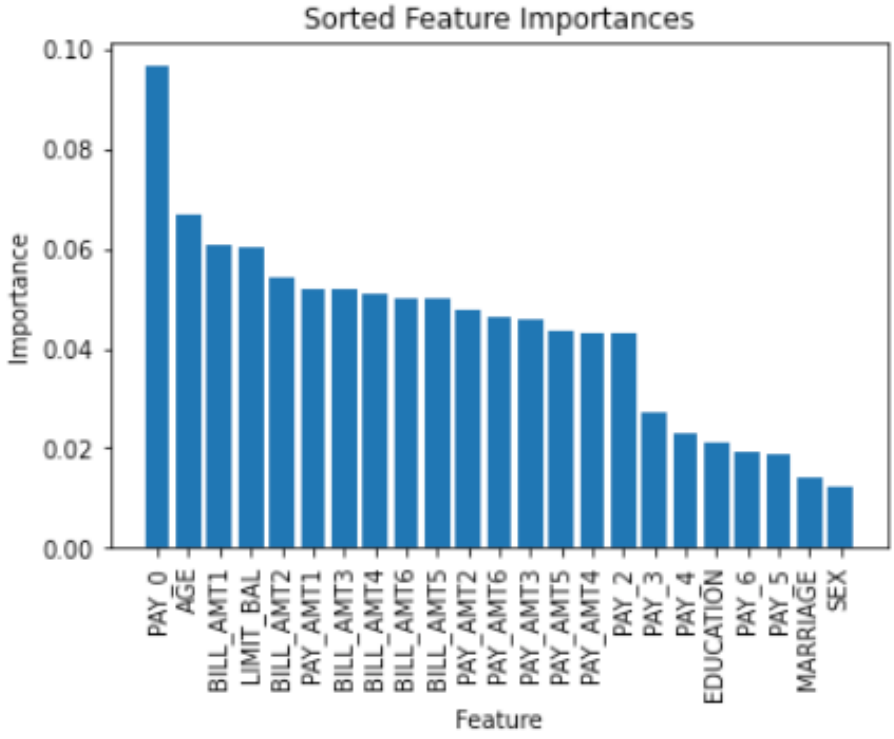
Part 2: Random forest feature importance

Display the individual feature importance of your best model in Part 1 above using the code presented in Chapter 4 on page 136. {importances=forest.feature_importances_ }

The followings are sorted features impotances.

```
{'PAY_0': 0.0966,
 'AGE': 0.067,
 'BILL_AMT1': 0.0606,
```

```
'LIMIT_BAL': 0.0604,
'BILL_AMT2': 0.0544,
'PAY_AMT1': 0.0518,
'BILL_AMT3': 0.0517,
'BILL_AMT4': 0.0509,
'BILL_AMT6': 0.0503,
'BILL_AMT5': 0.0502,
'PAY_AMT2': 0.048,
'PAY_AMT6': 0.0462,
'PAY_AMT3': 0.0459,
'PAY_AMT5': 0.0437,
'PAY_AMT4': 0.0432,
'PAY_2': 0.0432,
'PAY_3': 0.0272,
'PAY_4': 0.023,
'EDUCATION': 0.0209,
'PAY_6': 0.0192,
'PAY_5': 0.0189,
'MARRIAGE': 0.0139,
'SEX': 0.0124}
```



Part 3: Conclusions

Write a short paragraph summarizing your findings. Answer the following questions:
a) What is the relationship between n_estimators, in-sample CV accuracy and computation time?
b) What is the optimal number of estimators for your forest?
c) Which features contribute the most importance in your model according to scikit-learn function?
d) What is feature importance and how is it calculated? (If you are not sure, refer to the Scikit-Learn.org documentation.)

First from the test accuracy is 0.817 which is even higher than the cross validation average accuracy. It indicates that our model is not overfitting. We could probably increase the maximum depth or number of subtrees in the future model tuining.

a) n_estimators is positively correlative to CV accuracy and computation time.

b) The optimal number of estmator is 300.

c) PAY_0 contributes the most importance.

d) Since random forest is a little bit black-box, we can specifically see the causality of each features. The, we can only inspect how much a feature contribute to the impurity reduction of nodes. The impurity is calculated by some metrics for example entropy, etc. The feature importance score for a particular feature is calculated as the sum of the decrease in impurity over all node.

Part 4: Appendix

Github