My Name (shixuef2)
IE598 MLF F19
Module 7 Homework (Random Forest)

Part 1 Random forest estimators

N_estimator = 1

```
the train set score
[0.89881107 0.89784807 0.9004778  0.90099633 0.89981481 0.90111111
 0.89596682 0.89537425 0.8988556  0.89644828]
the test set score
[0.71209597 0.72509164 0.7230923  0.70976341 0.73033333 0.72266667
 0.72624208 0.74491497 0.73257753 0.73157719]
```

Run time: 5.474999999988128e-05 s

N_estimator = 2

```
the train set score
[0.9100337  0.90884848 0.90932998 0.90914478 0.90859259 0.90985185
 0.90752194 0.90855894 0.90874412 0.90855894]
the test set score
[0.78207264 0.78173942 0.78407198 0.77574142 0.79466667 0.79366667
```

Run time: 8.986399999955097e-05 s

N_estimator = 3

```
the train set score
[0.95329457 0.95340568 0.95425757 0.95455387 0.9532963  0.95407407
 0.95300174 0.95137217 0.95237213 0.95003889]
the test set score
[0.7664112  0.76407864 0.76741086 0.75608131 0.767      0.774
 0.78059353 0.7795932  0.77892631 0.76725575]
```

Run time: 0.00024202899999981042 s

N_estimator = 4

```
the train set score
[0.94659061 0.94629431 0.94603504 0.94762769 0.94796296 0.94822222
 0.94566868 0.94533536 0.94511314 0.94396504]
the test set score
[0.79440187 0.79306898 0.78807064 0.7844052  0.79766667 0.802
 0.80560187 0.80393464 0.80126709 0.79926642]
```

Run time: 0.00018161700000085546 s

N_estimator = 5

```
the train set score
[0.97214712 0.97236935 0.97144339 0.97122116 0.97137037 0.97125926
 0.96974186 0.96992704 0.97096404 0.96989   ]
the test set score
[0.78540487 0.78507164 0.78807064 0.77474175 0.783      0.78766667
 0.80426809 0.8036012  0.80026676 0.79026342]
```

Run time: 0.00028545100000165746 s

N_estimator = 6

```
the train set score
[0.96436905 0.96329494 0.96388755 0.96514686 0.96392593 0.96551852
 0.96248287 0.96340876 0.96303841 0.96244584]
the test set score
[0.80006664 0.79406864 0.79573476 0.79006998 0.79766667 0.805
 0.81527176 0.80793598 0.80793598 0.80426809]
```

Run time: 0.00020880199999950833 s

N_estimator = 7

```
the train set score
[0.98096226 0.98018445 0.9803326  0.98151783 0.98103704 0.98022222
 0.98000074 0.97900078 0.98063035 0.97977853]
the test set score
[0.7934022  0.78273909 0.78940353 0.78373875 0.79533333 0.799
 0.81827276 0.80460153 0.80926976 0.79893298]
```

Run time: 0.00048103800000021124 s

N_estimator = 8

```
the train set score
[0.97488796 0.97366569 0.97436942 0.9757028  0.97422222 0.97377778
 0.97288989 0.97344543 0.97414911 0.97322321]
the test set score
[0.80306564 0.79240253 0.79706764 0.78907031 0.80533333 0.812
 0.82260754 0.80860287 0.81293765 0.80826942]
```

Run time: 0.0005546669999993981 s

N_estimator = 9

```
the train set score
[0.98511056 0.98481425 0.98574021 0.98566614 0.98585185 0.98533333
 0.98477834 0.98470427 0.98529684 0.98559313]
the test set score
[0.79840053 0.78573809 0.79406864 0.78673775 0.79933333 0.80466667
 0.82127376 0.80793598 0.81293765 0.80526842]
```

Run time: 0.0006128140000001281 s

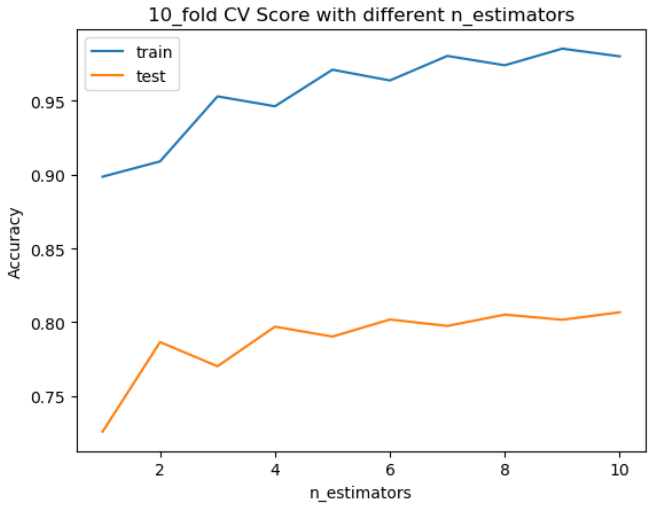
N_estimator = 10

```
the train set score
[0.97992518 0.98029557 0.98055484 0.98036964 0.98051852 0.97992593
 0.97963038 0.97922299 0.98018592 0.98029703]
the test set score
[0.80106631 0.79540153 0.80206598 0.79140287 0.80866667 0.81366667
 0.82494165 0.80893631 0.81460487 0.80626876]
```

Run time: 0.0005773210000015183 s

Mean Scores for Different N_estimators:

N_estimators	**1**	**2**	**3**	**4**	**5**
Train Score	0.8985704129197227	0.9089185332332977	0.9529666989879694	0.9462815054275213	0.9710333579928262
Test Score	0.7258355089483899	0.7865343253001102	0.7701350795594533	0.7969683374446301	0.7902355036854263
Running Time	5.474999999988128e-05 s	8.986399999955097e-05 s	0.00024202899999981042 s	0.00018161700000085546 s	0.00028545100000165746 s
N_estimators	**6**	**7**	**8**	**9**	**10**
Train Score	0.9637518715648936	0.9803666799509477	0.9740333515490169	0.985288892439191	0.9800925992922472
Test Score	0.8018018490594647	0.7974693931151178	0.8051356937113733	0.8016360823595647	0.8067021604965363
Running Time	0.00020880199999995950833 s	0.00048103800000021124 s	0.0005546669999993981 s	0.0006128140000000281 s	0.0005773210000015183 s



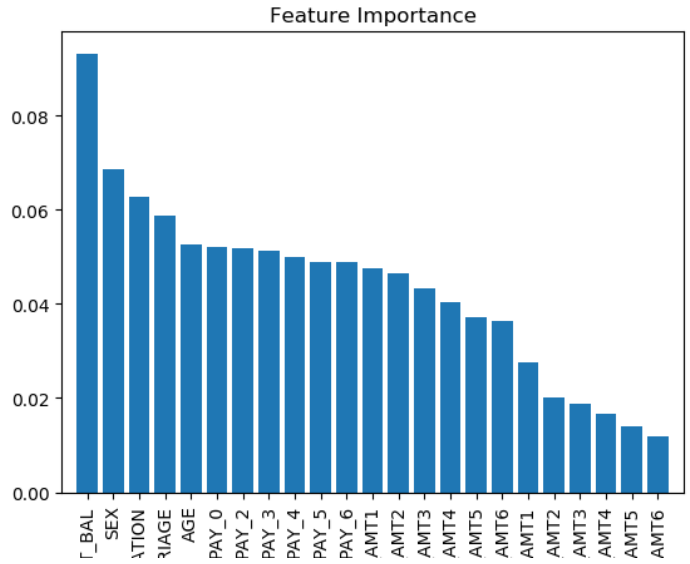
The Best N_estimator

```
Optimal n is:10
Optimal accuracy score is:0.8067021604965363
```

Part 2 Random forest feature importance

When n_estimator = 10, the features importances are as follows:

```
PAY_0 0.09324260165733629
AGE 0.068706274879237
BILL_AMT1 0.06280426922252177
LIMIT_BAL 0.05879430511505666
BILL_AMT6 0.05274585204563945
BILL_AMT2 0.05201928273277302
BILL_AMT3 0.05191002778791567
BILL_AMT5 0.051287985405491496
PAY_AMT1 0.050016568443232844
PAY_AMT2 0.0490142723295698
BILL_AMT4 0.04891678402865089
PAY_AMT3 0.04747646898265743
PAY_AMT6 0.046534731301291245
PAY_AMT4 0.043437359916984594
PAY_AMT5 0.0405242993469114
PAY_2 0.037127334507587664
PAY_3 0.03642932813855153
PAY_4 0.02751815178822547
EDUCATION 0.02025681165312073
PAY_6 0.01875614378091609
PAY_5 0.016645682969388786
MARRIAGE 0.014049011532518096
SEX 0.011786452434422081
```



Part 3 Conclusions

a) When the number of estimators increases, the running time increases, and the accuracy scores of test set and train set almost increases until the amount that nearly perfectly translates the data. So more estimators will give the model better performance, but also make code slower.
b) The optimal number of my estimators is 10.
c) PAY_0 contributes the most importance in your model according to scikit-learn function.
d) The feature importance is calculating the reduction of impurity in the node with the weight of probability of arriving at the node. The feature with higher value means it is more important.

Part 4 Appendix

Link to my code:
https://github.com/fengzixue96/IE598_F19_HW7/blob/master/IE598_F19_HW7.py

The screenshot: