Saranpat Prasertthum (655667271)

IE517 ML in Fin Lab

Module 7 Homework (Random Forest)

Out[1]:

Click here to toggle on/off the raw code.

Load Data

Out[3]:

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	.
1	20000	2	2	1	24	2	2	-1	-1	-2	.
2	120000	2	2	2	26	-1	2	0	0	0	.
3	90000	2	2	2	34	0	0	0	0	0	.
4	50000	2	2	1	37	0	0	0	0	0	.
5	50000	1	2	1	57	-1	0	-1	0	0	.

5 rows × 24 columns

Part 1: Random forest estimators

Out[5]:

```
GridSearchCV(cv=10, estimator=RandomForestClassifier(),
             param_grid={'n_estimators': [10, 50, 100]},
             return_train_score=True, scoring='accuracy')
```

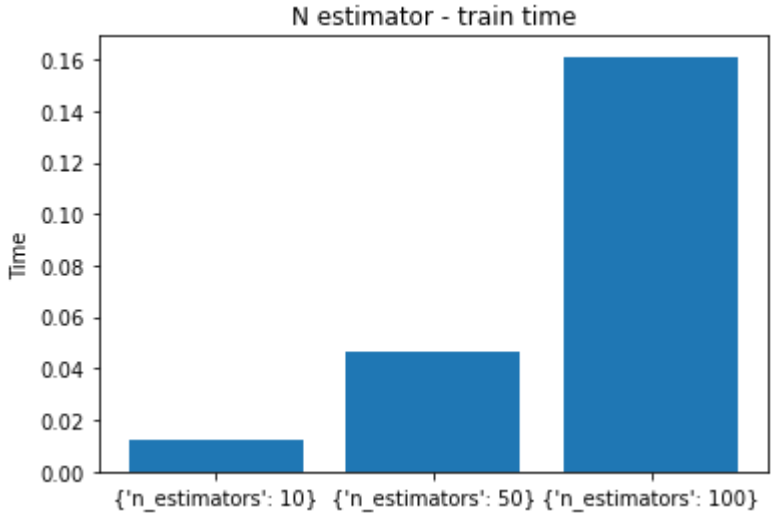
Out[6]:

	Params	Train Score	Test Score	Time
0	{'n_estimators': 10}	0.979881	0.807800	0.012062
1	{'n_estimators': 50}	0.998841	0.814600	0.046465
2	{'n_estimators': 100}	0.999344	0.815633	0.160956

Part 2: Random forest feature importance

```
1) PAY_0          0.0986534998307748
2) AGE    0.0675981322288534
3) BILL_AMT1      0.060794968463012564
4) LIMIT_BAL      0.05938585573161959
5) BILL_AMT2      0.05451331240022285
6) BILL_AMT3      0.05185449147076197
7) PAY_AMT1       0.05083125934847679
8) BILL_AMT6      0.05065825548621419
9) BILL_AMT5      0.050542074247345664
10) BILL_AMT4     0.04989948122514444
11) PAY_AMT2      0.04742839312450553
12) PAY_AMT6      0.04648475150914067
13) PAY_AMT3      0.04623896699943686
14) PAY_AMT5      0.04348973831161921
15) PAY_AMT4      0.042896326918097034
16) PAY_2         0.03944462464112495
17) PAY_3         0.029032476832814774
18) PAY_5         0.02402127551409982
19) PAY_4         0.022996783190280524
20) EDUCATION     0.020348760322455986
21) PAY_6         0.016465360500640916
22) MARRIAGE      0.014096163854901509
23) SEX           0.012325047848455981
```

Part 3: Conclusions



a) What is the relationship between n_estimators, in-sample CV accuracy and computation time?

ANS If N estimator increase it will also increase computational time.

Out[10]:

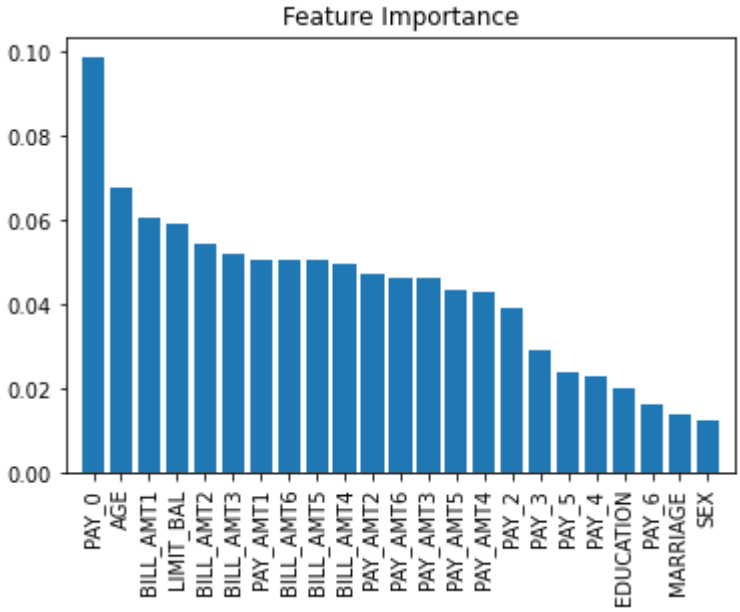
```
{'n_estimators': 100}
```

b) What is the optimal number of estimators for your forest?

ANS The best estimators is 50

Out[11]:

(-1.0, 23.0)



c) Which features contribute the most importance in your model according to scikit-learn function?

ANS The most importance feature in my model is "PAY_0"

d) What is feature importance and how is it calculated? (If you are not sure, refer to the Scikit-Learn.org documentation.)

ANS Feature importance in scikit-learn Random Forest models is calculated based on the decrease in node impurity caused by a feature, and the probability of reaching the node that uses the feature. This probability is determined by dividing the number of samples that reach the node by the total number of samples. The resulting values are then weighted and used to rank the importance of the features. Essentially, the higher the feature importance value, the more significant the feature is in predicting the target variable.

Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^{C} f_i(1-f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Variance / Mean Square Error (MSE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$
Variance / Mean Absolute Error (MAE) (Scikit-learn only)	Regression	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\mu\rvert$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

Credit: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3#:~:text=Feature%20importance%20is%20calculated%20as,the%20more%20important%20the%2

(https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3#:~:text=Feature%20importance%20is%20calculated%20as,the%20more%20important%20the%2

Part 4: Appendix

```
My name is Saranpat Prasertthum
My NetID is: 655667271
I hereby certify that I have read the University policy on Academic Integr
ity and that I am not in violation.
```