

Emma Mayes (eemayes2)

IE517 MLF F21

Module 7 Homework (Random Forest)

Using the ccdefault dataset, and 10 fold cross validation described in Raschka;

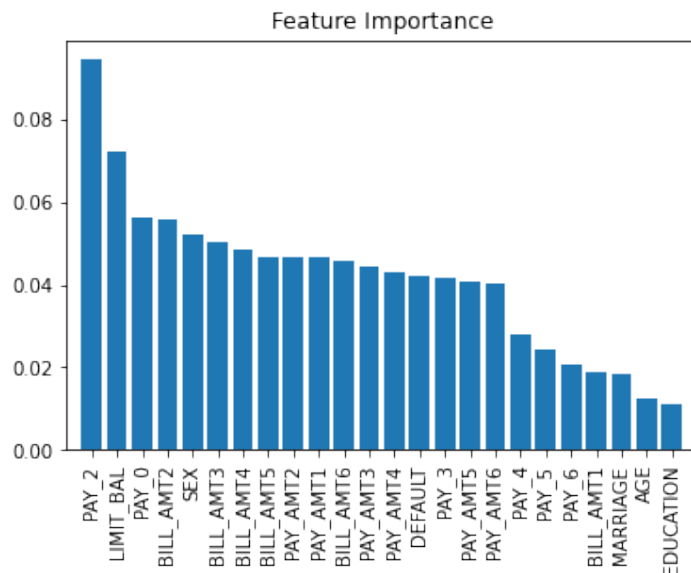
### Part 1: Random forest estimators

Fit a random forest model, try several different values for  $N_{\text{estimators}}$ , report in-sample accuracies.

*I tried  $n_{\text{estimators}} = 10, 20, 50, 75, 100, 150, 200, 400$*

### 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_ }`



### Part 3: Conclusions

Write a short paragraph summarizing your findings. Answer the following questions:

- What is the relationship between  $n_{\text{estimators}}$ , in-sample CV accuracy and computation time?  
*The greater the number of estimators, the greater the computation time, but also the greater the in-sample accuracy*
- What is the optimal number of estimators for your forest?  
 *$N_{\text{estimators}} = 75$ . Out of the range of  $n_{\text{estimators}}$  I ran, this gave the best in-sample and out-of-sample accuracy scores partnered with the shorted computation time, as more estimators were able to perform similarly, just with a longer time to train.*
- Which features contribute the most importance in your model according to scikit-learn function?

*Pay\_2 did, with it given 0.094387 as its feature importance*

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

*Feature importance is the mean and standard deviation based on how much decrease of impurity a given feature provides within each tree in the random forest*

#### **Part 4: Appendix**

Link to github repo: