Prajwal Chinthoju (pkc3)

IE598 MLF F18

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\_estimators, report in-sample accuracies.

Correlation heatmap:

Chart

Description automatically generated with medium confidence

Although, there are a few features that are highly correlated, PCA is not used because we want to better understand the feature importances.

Trying out different N\_estimators

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Although, not visible clearly in the plot, the accuracies slightly increase with number of estimators. For this problem, we are using max\_depth=None i.e. we are forcing the individual trees to overfit and using ensembling via RF classifier to eliminate the overfitting. Therefore, we see a slight increase of accuracy while increasing the n\_estimators. The increase can be observed in the print outputs below.

mean accuracy with stratified k fold for n\_estimators= 10 is: 0.8063333333333332

mean accuracy with stratified k fold for n\_estimators= 20 is: 0.8094259259259259

mean accuracy with stratified k fold for n\_estimators= 30 is: 0.8113333333333334

mean accuracy with stratified k fold for n\_estimators= 40 is: 0.8123888888888888

mean accuracy with stratified k fold for n\_estimators= 50 is: 0.8131111111111111

mean accuracy with stratified k fold for n\_estimators= 60 is: 0.813716049382716

mean accuracy with stratified k fold for n\_estimators= 70 is: 0.8143015873015873

mean accuracy with stratified k fold for n\_estimators= 80 is: 0.8146759259259259

mean accuracy with stratified k fold for n\_estimators= 90 is: 0.8149465020576132

We can also see that the increase in accuracy is minimal while using high n\_estimators. Therefore, to reduce the time taken to fit the RF, n\_estimators=50 is used for the fitting the full training set. The out of sample accuracy observed for the final model is 0.815.

**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\_ }

Below is the plot for the feature importances of the final RF model (with n\_estimators=50)

Chart, bar chart, histogram

Description automatically generated

As expected, the most important features tend to be the previous months’ payment histories, followed by limit balance and billing amounts.

**Part 3: Conclusions**

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

1. What is the relationship between n\_estimators, in-sample CV accuracy and computation time?

The in sample CV accuracies as well as the computational time increase with increase in number of estimators. However this increase in accuracy tends to saturate at some point and hence the extra effort is not worth the increase in accuracy at high number of estimators.

1. What is the optimal number of estimators for your forest?

The optimal number of estimators seems to be 50. The computational time is very high beyond 50 estimators.

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

Previous payments seem to be the most important features.

1. 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 amount of influence each feature has on the target. In RF Classifier, this is computed as the averaged impurity decrease of all the decision trees in the forest w.r.t that feature.

**Part 4: Appendix**

Link to github repo: <https://github.com/chinthojuprajwal/IE517/blob/main/IE517_FY21_HW7/IE517_FY21_HW7_prajwal.ipynb>