Yuchen Duan (yuchend3)

IE598 MLF F18

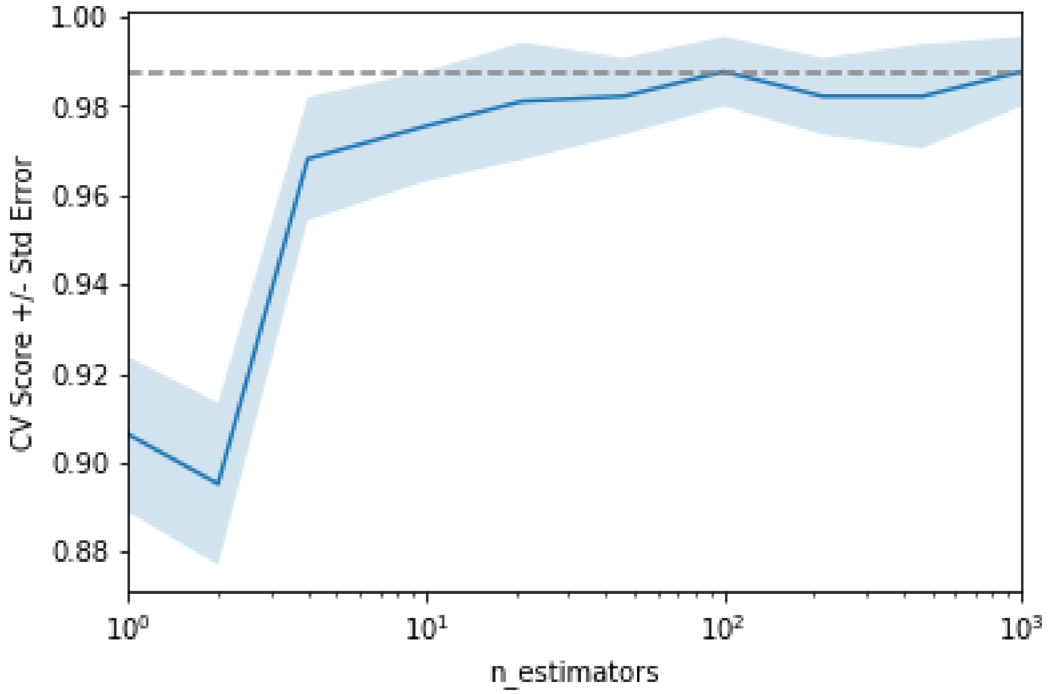
Module 7 Homework (Random Forest)

Using the Wine dataset, described in Raschka chapter 4 and 10 fold cross validation;

Part 1: Random forest estimators

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

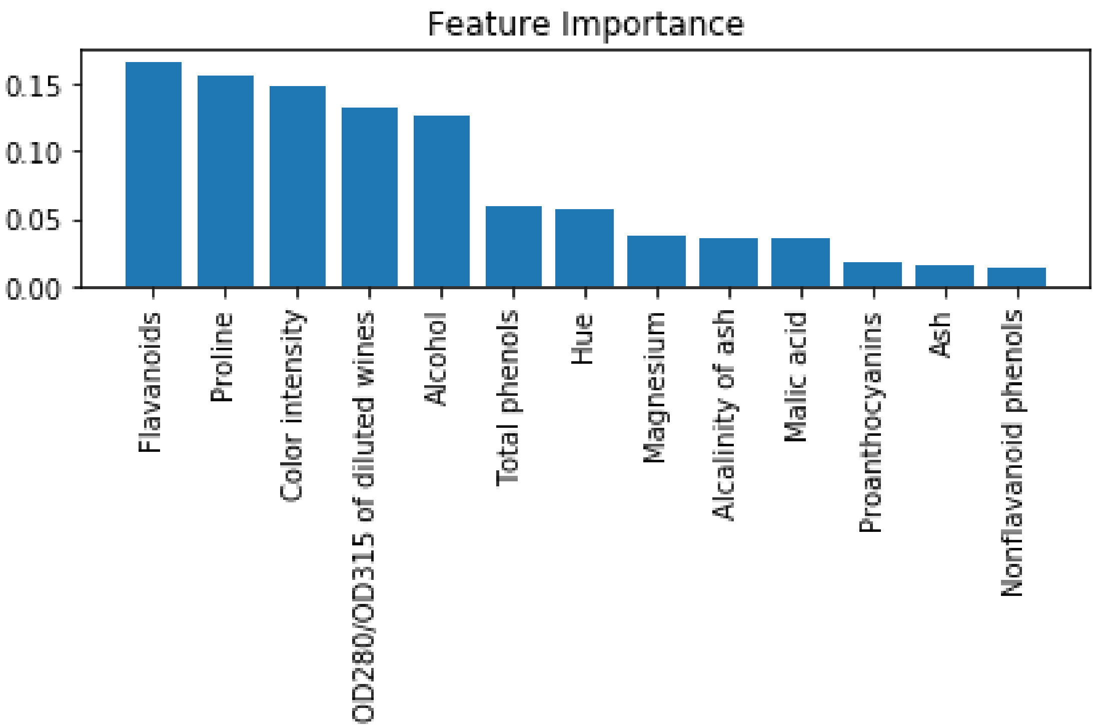
	1	2	4	10	21	46	100	215	464	1000
mean	1.0	0.944	1.0	1.0	1.0	0.933	1.0	1.0	1.0	1.0
std	1.0	0.944	1.0	1.0	1.0	0.933	1.0	1.0	1.0	1.0



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_ }

```
{'n_estimators': 100}

 1) Flavanoids                 0.166301

 2) Proline                  0.156037

 3) Color intensity            0.148261

 4) OD280/OD315 of diluted wines   0.131352

 5) Alcohol                  0.125929

 6) Total phenols              0.058346

 7) Hue                     0.056940

 8) Magnesium                 0.037776

 9) Alcalinity of ash          0.035967

10) Malic acid                0.035051

11) Proanthocyanins             0.018216

12) Ash                      0.015427

13) Nonflavanoid phenols         0.014398
```



Part 3: Conclusions

Write a short paragraph summarizing your findings.

The higher the estimators better accuracy but the grow increase most between first 100. After 100 it even decrease a little but the run time has increase exponentially.

What is the optimal number of estimators for your forest?

My optimal number of estimators is 100 due to it's run time and accuracy.

Favanoids have the most importance in my model each one time a ratio in computing the out put

Part 4: Appendix

Link to github repo

https://github.com/rainduan/IE598_F18_HW7