Assignment 3 – Machine Learning for Public Policy

Xin Ding

05.21.2018

Feedback from last assignment:

I think you misunderstood the assignment. A big chunk of the assignment was to write functions that you will be able to use for processing new datasets going forward. Yes, there was also a component for doing the same kind of data exploration you did for assigment 1 and for fitting a classifier to the data + evaluating it so I can give you some points.

In the sense that I got a very low score on assignment 2, I decided to give assignment 3 a fresh start.

Basically,

1. Based on the feedback, I thought about the assignment again and realize that I misunderstood it. Therefore I put all the steps in helper functions this time to give them a cleaner look and convenience for future data.

2. I plot the precision and recall for different models and compare the results for AUC, f1, precision, recall, and time collectively to show the advantages and disadvantages of different models.

3. I tried to incorporate all the recommended methods but have some trouble with the SVM. It takes forever to run.

Results:

It took about 70 minutes to run all the models. And here are the results.

For auc:

DT under {'criterion': 'entropy', 'max\_depth': 100, 'max\_features': 'log2', 'min\_samples\_split': 10} at 0.742263800432187 is the best overall.

For f1:

GB under {'learning\_rate': 0.5, 'max\_depth': 50, 'n\_estimators': 100, 'subsample': 1.0} at 0.5337663575985123 is the best overall.

For precision:

NB under {} at 0.7317073170731708 is the best overall.

For recall:BAG under {'max\_features': 5, 'max\_samples': 100, 'n\_estimators': 100} at 1.0

is the best overall.

This is the result for time: DT under {'criterion': 'entropy', 'max\_depth': 100, 'max\_features': 'log2', 'min\_samples\_split': 10} at 0.007062117258707683 is the best overall.

Below are the plots showing the precision and recall on each model.









