PA#3 https://github.com/yuxiaosun/capp-455136/tree/master/pa3 Yuxiao Sun 455136

1. Machine Learning Pipeline
https://github.com/yuxiaosun/capp-455136/tree/master/mlpipe

2. The problem
The goal is to predict "serious credit delinquencies" with individual level demographic and financial data. The data can be downloaded from https://www.kaggle.com/c/GiveMeSomeCredit/data.

3. Feature Engineering
I first normalized all the features. Two features "Monthly Income" and "Number of Dependents" have missing values in our dataset. Ideally I should split the non-missing data into a training and a validation set and loop through all the possible classifiers (mean/median, KNN, Bayes, Regression, EM. etc.) to find the best imputation method. However, for the focus of this homework I decided to simply drop all the missing values to avoid the influence of imputation method on modeling results. I will build a loop function for comparing different imputation method in my pipeline for the future.

As for feature selection, I ran a Random Forest to determine which feature provides better information to classify credit risk. I chose the top four features: Monthly Income, Age, Debt Ratio and Revolving Utilization of Unsecured Lines.

4. Modeling
I ran the following classifiers: K-Nearest Neighbors, Decision Tree, Support Vector Machine, Random Forest, Logistic Regression, Naive Bayes, Ada Boost and Gradient Boosting. For each of the classifiers I tested different combinations of parameters.

5. Evaluation
I derived the following five performance metrics: accuracy, precision, recall, F1 and AUC. I first used decision trees to compare these metrics over different threshold levels (Figure 1): 0.1-0.9. I chose AUC and F1 score as the main metrics and thus I decided to use the threshold 0.1 (0.2 would be also good) for the subsequent comparison of different classifiers.

For single classifiers (Figure 2), in average, NB and SVM work best in accuracy, whereas DT and KNN perform better in AUC, precision, recall and F1 score. The top 5 models in terms of F1 or accuracy score are all decision trees. The top 5 in terms of AUC have 3 KNN classifiers and 2 decision trees. The best single classifier seems to be a Decision Tree with "minimum samples split" = 10, "maximum features"=square root of total features, "maximum depth"=20 and "splitting criterion"=gini. It happens to perform best in almost all five metrics, with AUC=0.60, F1=0.24, accuracy=86.94, precision=0.20 and recall=0.29. The SVM ususally takes about 50 minutes to train, which is much slower than all other classifiers.

The ensembles of classifiers take much longer time to train. Here I used Ada Boosting with decision tree as base estimator. However, it seems that for this dataset this method does not increase performance in a significant way.

Based on the resource constraint and policy target, different classifiers should be used to predict "serious credit delinquencies". For instance, for a very risk-averse financial service provider, the recall success might be a more important metric than precision. Finally, it is usually a good idea to jointly use different classifiers by various ensemble methods.

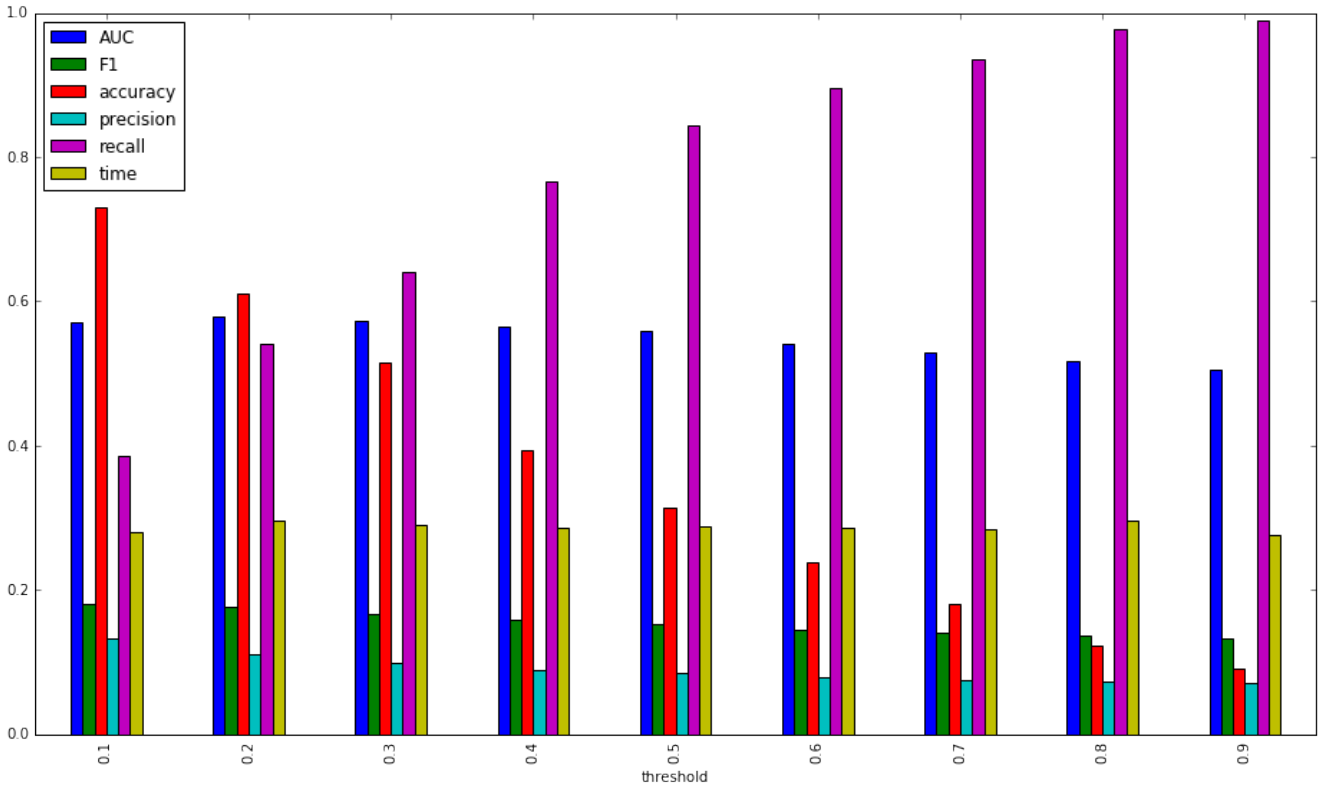

Figure 1 Average Performance of Different Thresholds (DT)

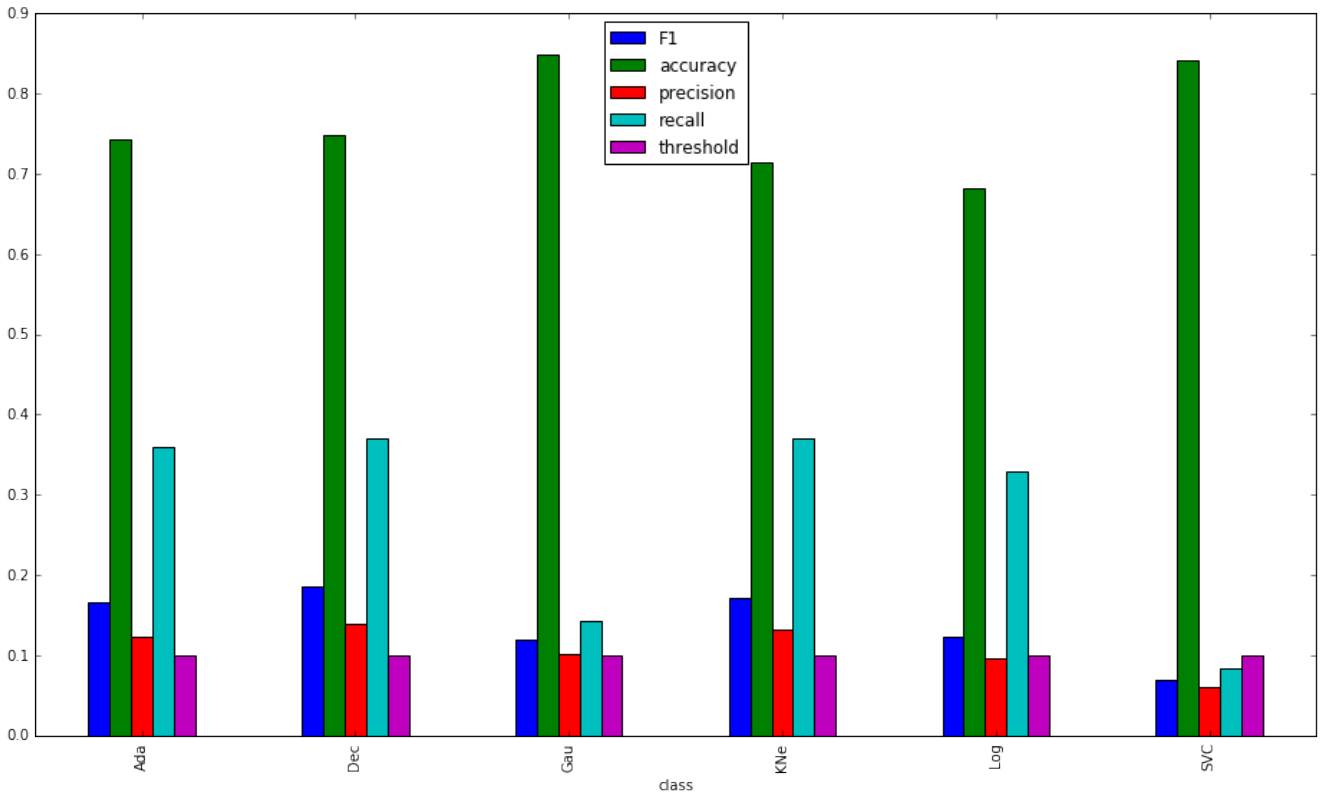

Figure 2 Average Performance of Different Classifiers