DSBA6156: Individual Project

Assigned: 10/06/2017

Due: 10/20/2017

100 points

This assignment has to be done individually by all the students. We are expecting a 5-6 page report to be submitted along with your codebase, instructions to execute your code, and intermediary data files that you may have generated.

The dataset Lendingclub2012to2013.csv and the corresponding data dictionary has been uploaded to Canvas. Go through the data dictionary to understand the significance of all the variables. Lending Club is a peer to peer lending service. The loans are given for a variety of reasons (want to reduce credit card debt, need money urgently, etc.). For more details, refer to this website www.lendingclub.com.

1) Perform a through feature analysis of the dataset. For instance, what is the distribution of each feature? If the feature is categorical, plot a histogram to show the feature distribution. If it is continuous, use appropriate plots to show the distribution. Include this analysis in your report. Use your intuition to also explain which features are best for classification. Propose new features not in the dataset that you think will be useful. This should be a good 2-3-page analysis of the dataset.

The exercise is to implement a suite of classifiers learned in the course thus far (decision tree, k-nearest neighbor, naïve Bayes, logistic regression, etc.). Split the data into 75-25 and hold out the 25. On the 75%, using an 80-20 split conduct your experiments and perform 10-fold cross validation

2) Which classifier gave the best results? Intuitively explain why?
3) Do a feature selection and analysis to come up with the optimal combination of features that yield highest accuracy. (Hint: You shouldn't be using all the features for this exercise!)
4) What is the classification accuracy for the classifiers (confusion matrix comprising of F1-score, recall, precision)?

Test your classifiers (all classifiers + optimal combination of features; essentially the best version of each classifier) on the hold out data (the 25%)

5) Report the accuracy scores (F-1, precision, and recall) of each classifier.

Please submit a PDF document (5-6-page report).