Author: Ajit Bhanot (abhanot2@illinois.edu)

CS 598: Practical Statistical Learning: Project 3

Project III: Lending Club Loan Status

1. Introduction

The data given to us is the historical loan data issued by lending club. The data is a cleaned version of the original data available on Kaggle. The focus is on the loan status where Class 1 (bad loans) and Class) are good loans that are paid off.

2. Methodology:

Some basic clean up of the data is done using the clean data function. Some of the main points are below:

- Most NA values are replaced with 0 (dti, mort_acc, bankruptcies).
- The income variable was changed to log
- The credit line was changed to year of the first credit line (this is an important variable in my knowledge and has a good impact on the credit rating of a person)

3. Model

Having used a logistic regression before in a previous project, I started with the same. After trying certain rudimentary algorithm, I used the xgboost function from the xgboost library. This is an optimum library to use since it is designed to be memory efficient. Another advantage are the various parameters that are available to us for tuning (https://xgboost.readthedocs.io/en/latest/parameter.html). I used the reference to understand the impact of the various params.

This model alone provided the needed accuracy and I stuck with this model alone. The various params used are as below:

1. Used binary logistic objective as it is fit for purpose.
2. The evaluation method has been given to us as log loss and is used.
3. The defaults for learning rates and maxdepth are used
4. The maximum depth for the trees was experimented with along with the rounds, while ~ 60 rounds does produce superior results, the gains are minimal over the ~ 40 rounds with max depth of 8. The time taken for the 40 rounds is ~ 33% less than the 40 rounds and therefore was sufficient for the problem at hand.
5. Printing is done every so often to ensure that the trend can be ascertained.

The execution was done on a Windows 10, Intel i7-7700 CPU @3.60GHz with 16GB RAM.

Measure	Split 1	Split2	Split3
Log Loss	0.448947569868316	0. 449732433110724	0.449460811821652
Timing	1.81142514944077	1.64957668383916	1.66306233406067