STAT 542 Project 3_Jinran Yang

1. Computer System

MacBook Pro (Retina, 13-inch, Mid 2014)

Processor:2.6 GHz Intel Core i5

Memory: 8 GB 1600 MHz DDR3

2. Data pre-processing

For this project, I am provided with historical loan data issued by Lending Club. And the goal is to build a model to predict the chance of default for a loan. I have 30 features in total including the response 'loan_status'. The first step is pre-processing data for further analysis. The details are as follows:

- 'emp_length': First I Convert it to integers (<1 year->0, 1 year ->1, …, 9 years ->9, 10+ years-> 10) and then I replaced the NA's by median (6);
- 'home_ownership' : Replace the values ANY and NONE with OTHER;
- 'fico_range_low'&'fico_range_high': We only need to keep one of the FICO scores. I take the average of the two and call it 'fico_score';
- 'earliest_cr_line': I change the "earliest_cr_line" to the number of months from the January of its first year (1944);
- Replace NA's in 'dti', 'mort_acc', 'pub_rec_bankruptcies' and 'revol_util' with 0;
- Drop the 'grade', since the grade is implied by the subgrade;
- Drop the 'emp_title', since there are too many different job titles for this feature to be useful;
- Drop the 'title', since there are too many different titles, and the 'purpose' variable contains similar information;
- Drop the 'Zip'. Zip and state contain similar information, but there are too many different zip, so we just keep state.

3. Model

I try xgboost and randomForest which are suggested by the professor Liang. It seems that the performance of xgboost is very well. The parameters are as follow:

- 'nrounds' = 25;
- 'objective' = "binary:logistic" which implies binary classification using logistic regression;
- 'eval_metric' = "logloss";
- 'eta': 0.3 (default);
- 'max_depth': 6 (default);
- 'verbose': 1;

4. Performance

Since the performance of xgboost model already satisfy the requirement, I just present the result of xgboost model.

XGBoost Model	Test 1	Test 2	Test 3	Mean
Error	0.4520219	0.4530671	0.4523865	0.4524918
Running Time(min)	3.683879	3.412859	3.421459	3.506066