Project 3 Report

Tianqi Wu

11/24/2018

Preprocess

First, I changed the “loan_status” to 0/1 indicator where 0 refers to “fully paid” and 1 refers to “charge off”. Then, “id” is used as index. “Grade” is dropped since it is a subset of “sub_grade”. “emp_title” and “zip_code” are dropped since it has too many unique values. “emp_length” is converted to integer. “title” is dropped since “purpose” already has the information.“open_acc”, “evol_bal” and “total_acc” are dropped since there is no large mean difference between the two loan status. “10+ years” is replaced as “10 years” and “< 1 year” is replaced as “0 years”. “earliest_cr_line” only keeps the year. Average is taken between“fico_range_low” and “fico_range_high” and the new variable is named as “fico_score”. “ANY” and “NONE” are merged as OTHER" for “home_ownership”. “annual_inc” is log transformed to “log_annual_inc”.

Dummy variables are created for categorical variables. Missing value for numeric varibles are imputed by the mean and winsorization of 0.05 cut-off is used for both tails. Then, only variables with correlation 0.01 above to “loan_status” are kept.

Model

I used xgboost from python as model. The parameters are set as follows: objective =‘reg:linear’, subsample = 0.8, colsample_bytree = 0.8, learning_rate = 0.1, max_depth = 7, n_estimators = 100, min_child_weight = 8.

Result

From the table,the average log-loss is 0.45008. The overall running time of whole script for one split is around 5 minutes and the computer system is Macbook Pro 3.1GHz, 8GB memory.

Table 1: Error of Models

test1	test2	test3	average
0.449696	0.4507826	0.449765	0.4500815