Lending Club Probability Prediction Report

Team Array3

Hongfei Li, Meng Du, Yiqin Wu

Pre-processing

1. Recode label

   We recode loan_status into a new column named "label", where 1 indicates "Default", and 0 indicates "Non-Default".

2. Missing value imputation
   a. Numeric values

      For missing values of numeric variables, we replace them by median.

   b. Date

      We recode missing values in date variable into a new level named "NoneDate". Specially, we regroup "mths_since_last_delinq" and "mths_since_last_record" to reduce their number of levels. We replace the missing value in earliest_cr_line by median.

3. Drop variables

   There are many variables that have too much missing values to impute. We simply delete these variables. In the end we have around 40 variables left.

Models

We find that the variable “recoveries” is very helpful in predicting default. If recoveries >0, the probability of default is 1. We label part of our test data by this rule and use the following predictive models to label the rest of our test data.

1. Random Forest

   Our first model is random forest with tree = 400 and mtry = 7.

2. Xgboost

   To apply Xgboost model, all the variables should be transformed to numeric type first. However, this may cause confusions if we process the train and test data separately. In order to avoid the mis-transformation, our team decided to use the numerical variables only.

   The weak classification model in this boosting model is gbtree. And the objective is set to be "binary:logistic". Other parameters are shown below:

Max.depth	eta	nthread
12	0.02	2
nround	alpha	gamma
1000	1	2
Min_child_weight	subsample	Colsample_bytree
1	0.5	0.5

Evaluation

We split loan.csv to evaluation our model. Our train contains 75% of loan and test contains 25% of loan. We do not add loan2016.

We run our code on AWS t2.2xlarge.

The runtime and logloss is as follows:

Model	Runtime(s)	Logloss
Random Forest	2521	0.08628121
Xgboost	1072	0.07317256