Explore and Setup Data

Missing values

Times\_DQ\_L12 has 5861 NAs, replaced them with 0.

Origination\_LTV has 2692 NAs, imputed by median.

TotalDeposit\_Amount has 15 NAs, imputed by median.

Exploring and selecting predictors

Dropped Loan\_ID, it contains no information.

Dropped Dates, including open\_date, date\_modified, last\_date\_active.

Dropped status\_asof\_last\_date, it directly correlated with charge off risk.

Dropped Modification\_reason, it is too sparse.

Dropped FICO\_tier, it is a post charge off indicator, directly correlated with charge off risk.

Dropped DQ1\_flag, DQ30\_flag, DQ60\_flag, DQ90\_flag, DQ120\_flag.

Remained 27 predictors for charge off risk and amount.

Cleaning and preparing data

Factorized categorical variables in R.

Normalized chgoff\_amt by dividing loan\_amount.

Part 1 charge off flag prediction

Random Forest

Accuracy in validation: 0.9426; in test: 0.9415

AUC in test: 0.8714

Logistic Regression

AUC in test: 0.875

Logistic Regression with Lasso

AUC in test: 0.8732

Accuracy in validation: 0.9407

Logistic Regression with ElasticNet

AUC in test: 0.8738

Accuracy in validation: 0.9413

Logistic Regression with Ridge

AUC in test: 0.8737

Accuracy in validation: 0.9413

XGboosting

AUC in test: 0.8737

Part 2 charge off amount prediction

Filtered data by chgoff\_flag = 1

Randomly split the sample to train and test.

Random Forest

Tuned hyperparameters by 10-fold cross-validation

rmse in validation: 0.1698; in test: 0.1587

Plot the variable importance and tuning process mtry.

Generalized Linear Model Regression with ElasticNet

Tuned hyperparameters by 10-fold cross-validation

rmse in validation: 0.1774; in test: 0.1740

Plot the variable importance and tuning process, alpha and lambda.

Generalized Linear Model Regression with Lasso

Tuned hyperparameters by 10-fold cross-validation

rmse in validation: 0.1661; in test: 0.1739

Plot the variable importance and tuning process, fraction.

OLS

rmse in test: 0.1760

XGBoosting

One Hot Encoded for categorical variables

Tuned hyperparameters by 10-fold cross-validation

rmse in validation: 0.1748; in test: 0.1599

Plot the variable importance and tuning process, nrounds, max\_depth, and colsample\_bytree.