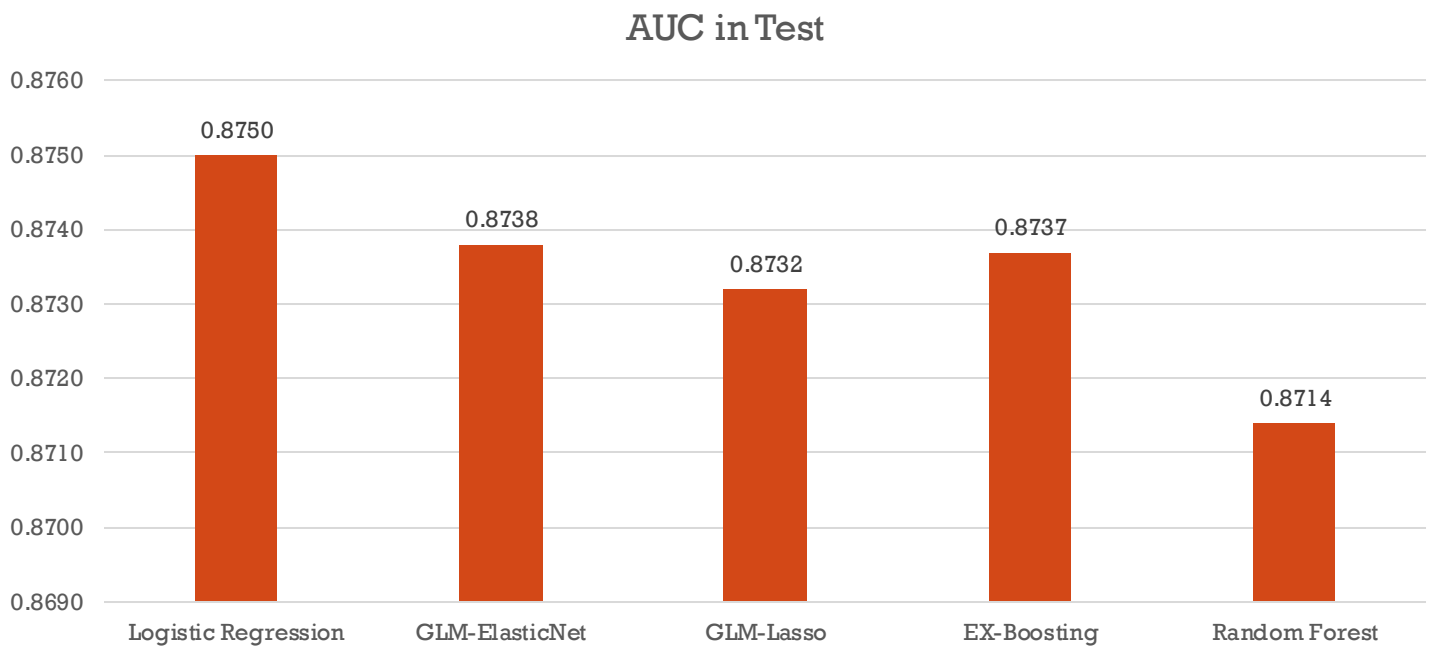
Explore and Setup Data

- Missing Values
  - Times_DQ_L12 has 5861 NAs, replaced by 0
  - Origination_LTV has 2692 NAs, imputed by median
  - TotalDeposit_Amount has 15 NAs, imputed by median
- Selecting Predictors
  - Dropped Loan_ID
  - Dropped Dates, including open_date, date_modified, last_date_active
  - Dropped Modification_reason, it is too sparse
  - Dropped status_asof_last_date, it directly correlated with charge off risk
  - Dropped FICO_tier, it is a post charge off indicator
  - Dropped DQ1_flag, DQ30_flag, DQ60_flag, DQ90_flag, DQ120_flag
  - Remained 27 predictors for charge off flag and charge off amount
- Preparing Data
  - Factorized categorical variables
  - One-Hot-Encoded

Predictive Models for Charge Off Flag

- Prediction for categorical variable
  - Randomly split sample into train and test by 80% and 20%
- The Benchmark model is logistic regression
- Lasso and Elastic-Net regularized generalized linear models
  - Preprocess: centered and scaled the numeric predictors
  - Tuned hyper-parameters by 10-fold cross-validation
  - Lass: lambda; Elastic Net: alpha and lambda
- Random Forest and XG-Boosting
  - Tuned hyper-parameters by 10-fold cross-validation
  - Random Forest: mtry, XG-Boosting: nrounds, max_depth and colsample_bytree
- Metric is accuracy for tuning hyperparameters
- Metric is AUC for evaluating goodness-of-fit

Predictive Models – Charge Off Flag Result in Test

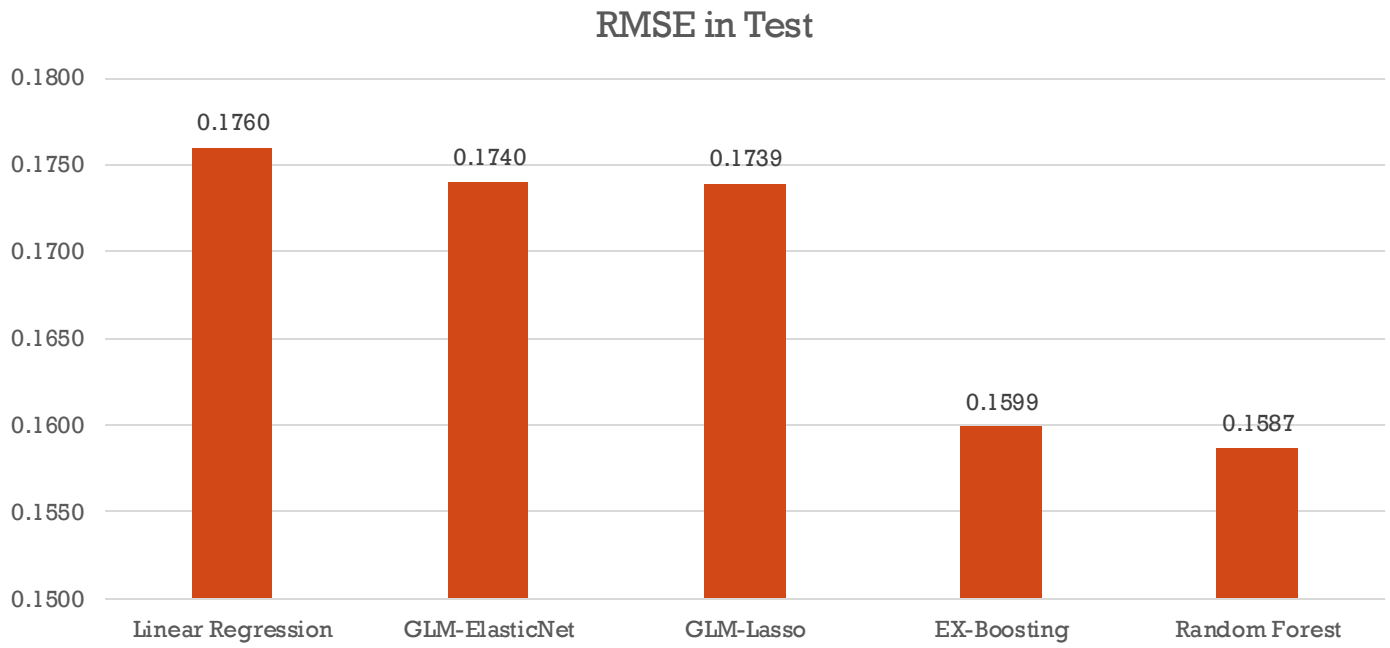


- Comparing the model performance in the test data, there isn't much difference between these models.

Predictive Models – Charge Off Amount

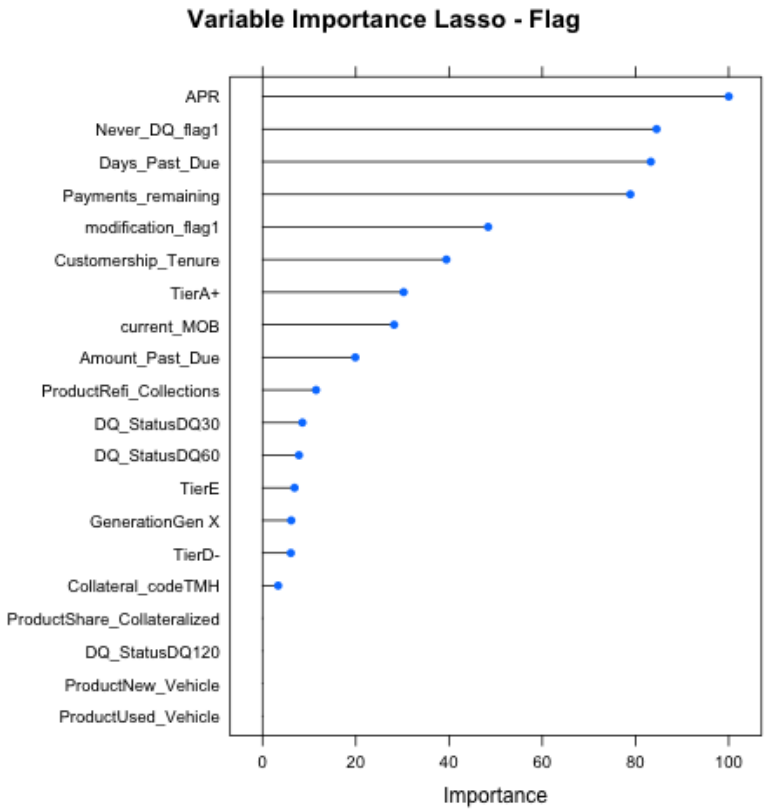
- Prediction for charge off rate (charge off amount / loan amount)
  - Filtered data by charge off flag equals 1
  - Randomly split sample into train and test by 80% and 20%
- The Benchmark model is linear regression
- Lasso and Elastic-Net regularized generalized linear models
  - Preprocess: centered and scaled the numeric predictors
  - Tuned hyper-parameters by 10-fold cross-validation
  - Lass: lambda; Elastic Net: alpha and lambda
- Random Forest and XG-Boosting
  - Tuned hyper-parameters by 10-fold cross-validation
  - Random Forest: mtry, XG-Boosting: nrounds, max_depth and colsample_bytree
- Metric is RMSE for tuning hyperparameters and evaluating goodness-of-fit

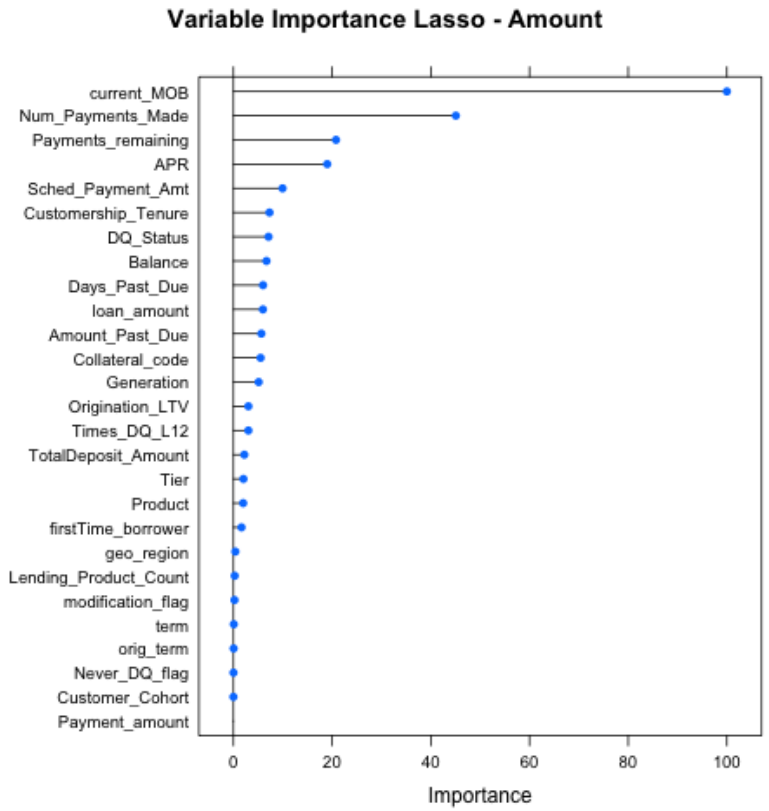
Predictive Models – Charge Off Amount Result in Test



- Comparing the model performance in the test data, Random Forest and EX-Boosting outperform all other models.
- This might be because ensemble method in general works better in small samples. We only have 643 observations that have charge off amount.

Understanding Risky Population – Variable Importance in Lasso





- In general, the risky loan population has a larger loan's current interest rate, has been delinquent at least once over life of loan, longer days past due, a shorter number of months since the loan originated, a fewer number of payments made, and larger payments remaining.