

Future loan payment prediction

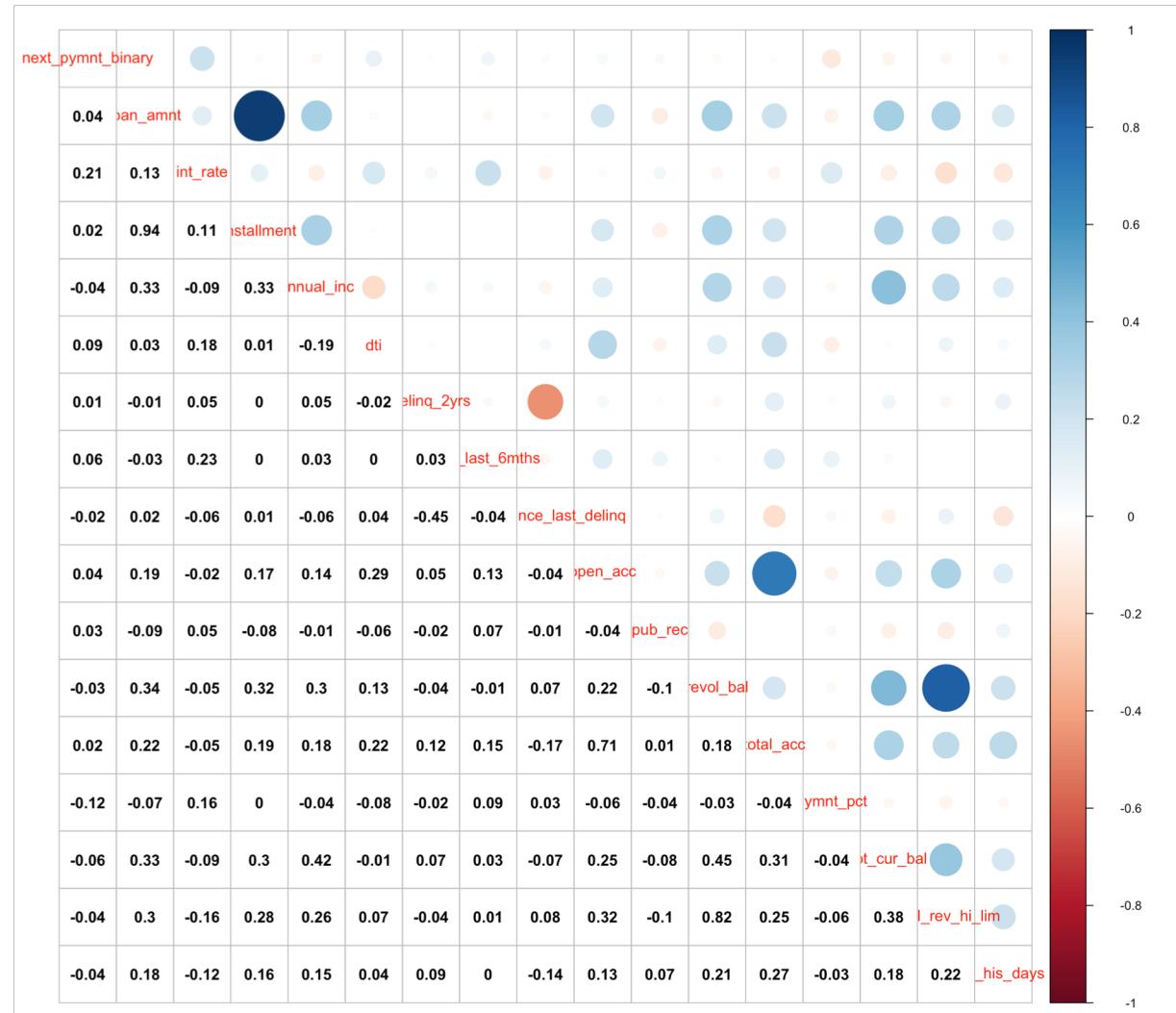
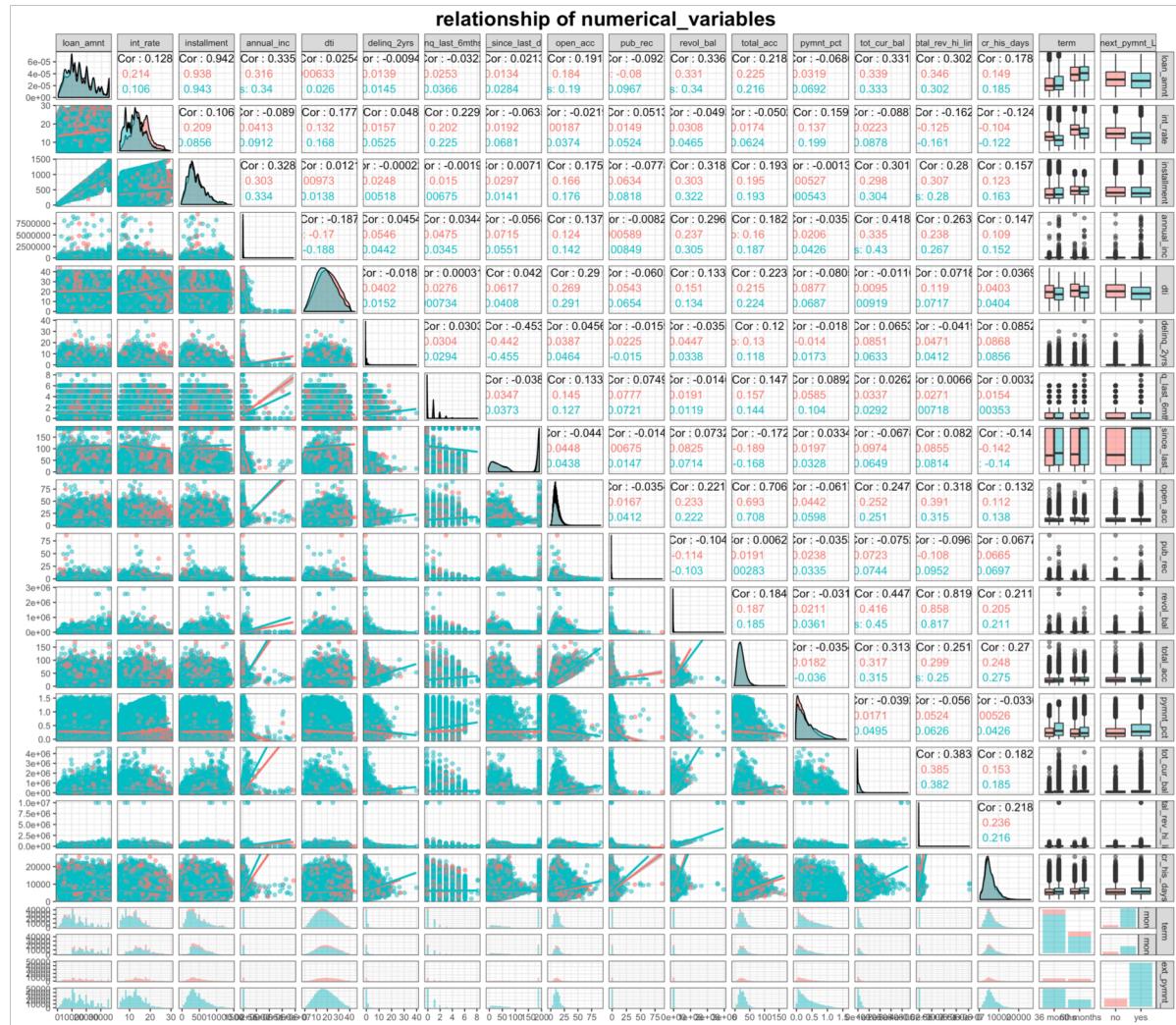
Lending club capstone project

Fan Yang

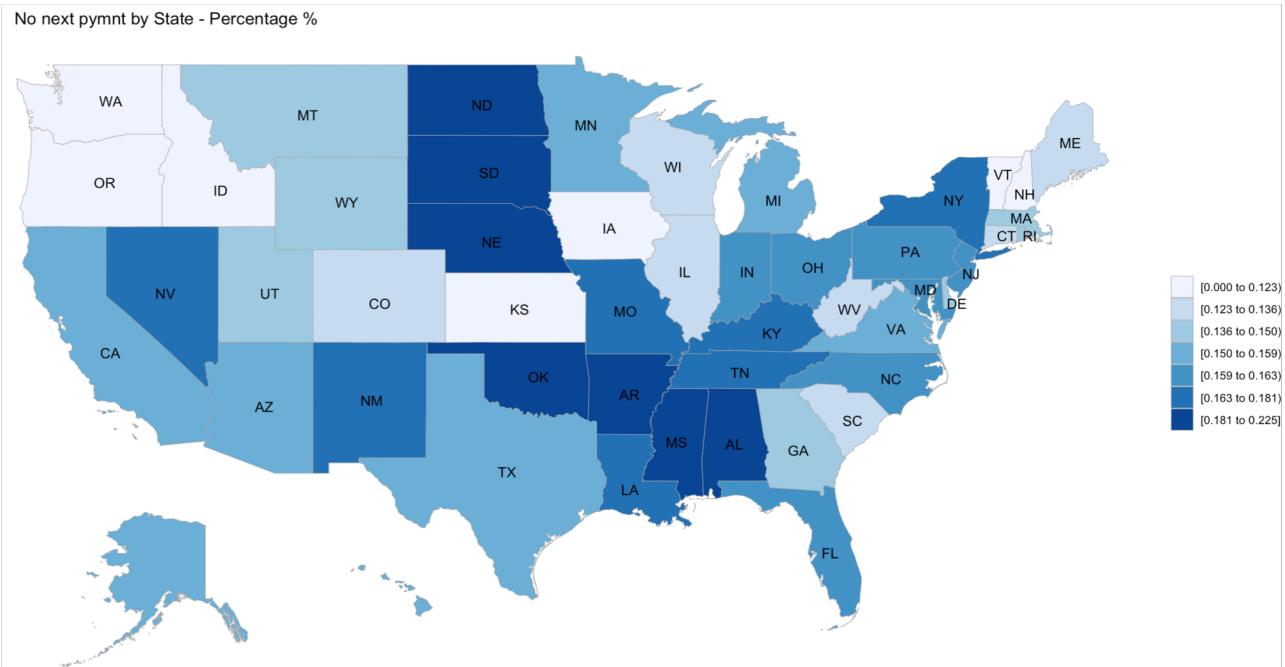
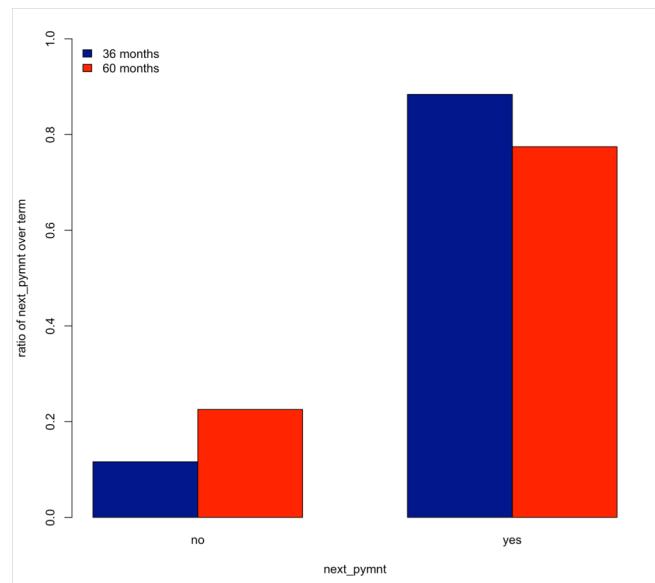
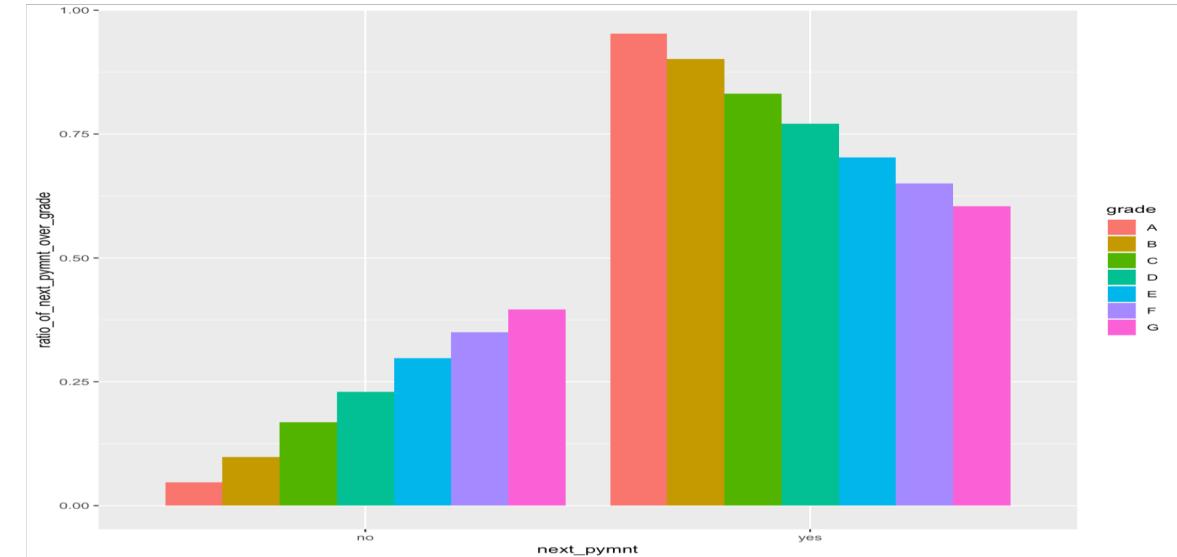
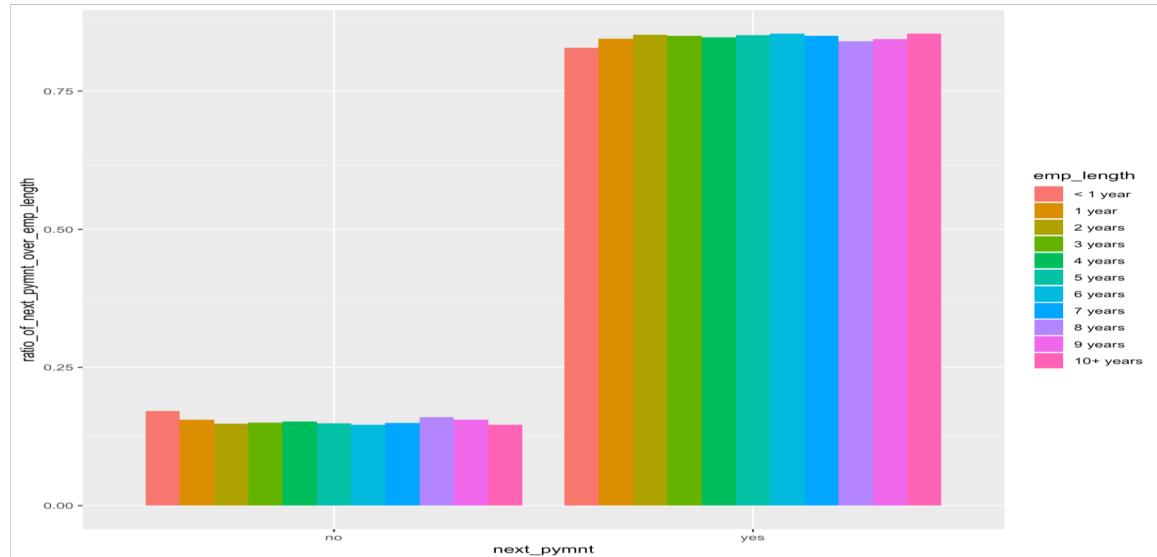
Project summary

- Lending club is a platform of person-to-person loan.
- Lending club data from Kaggle provides a complete record of loan issued between 2007-2015.
- A subset of loan were selected because they have no past due payment by January 1st 2016.
- However, it was found that 15.23% of these selected loan haven't completed payment on scheduled date by November 1st 2018, according to data from Lending Club Statistics.
- This project aims to predict whether customers, those are currently in healthy loan status, will fail to make next loan payment in the future.

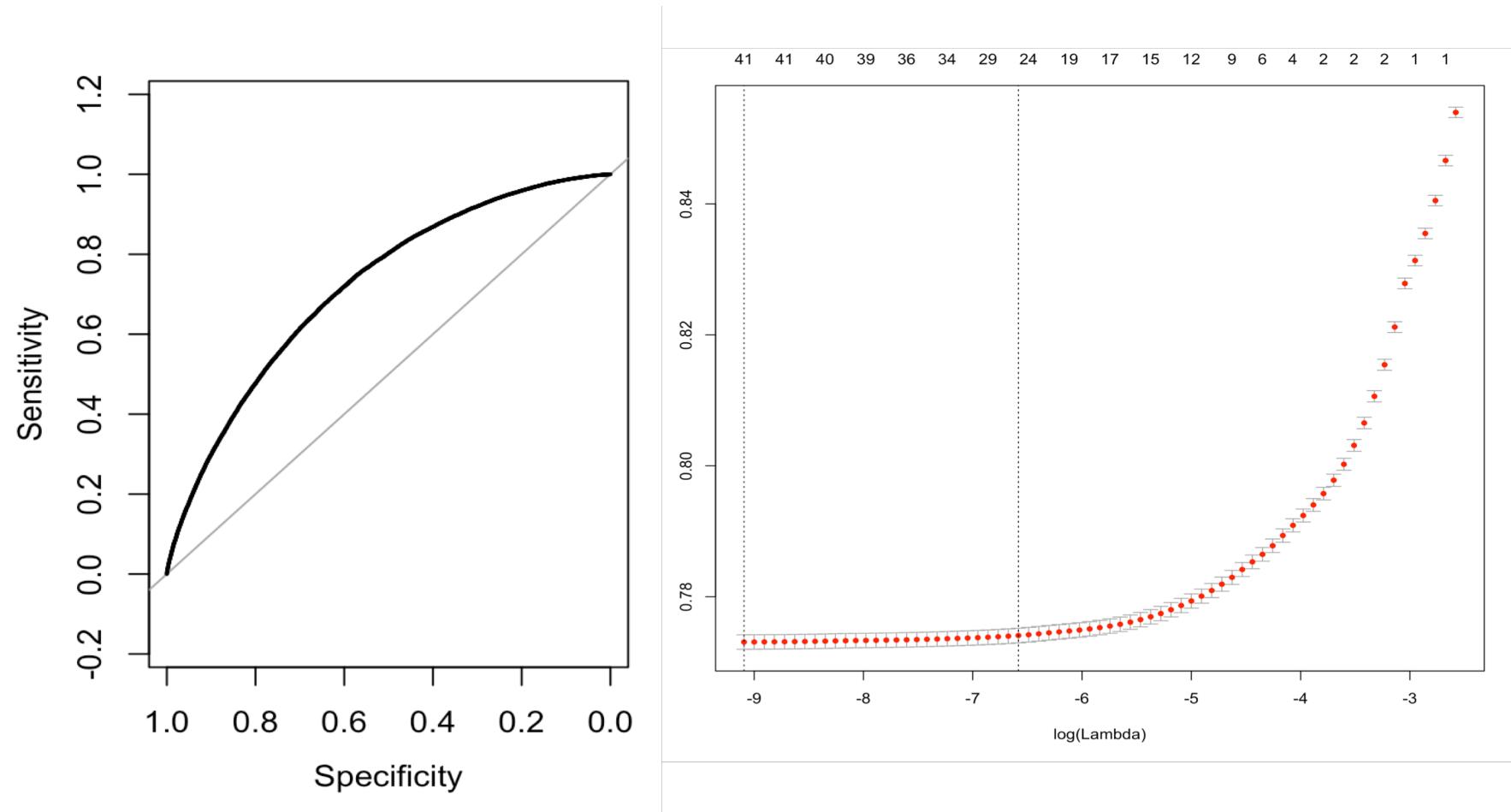
Visualizing relationship among numerical features



Exploring categorical feature are import for next_pymnt prediction

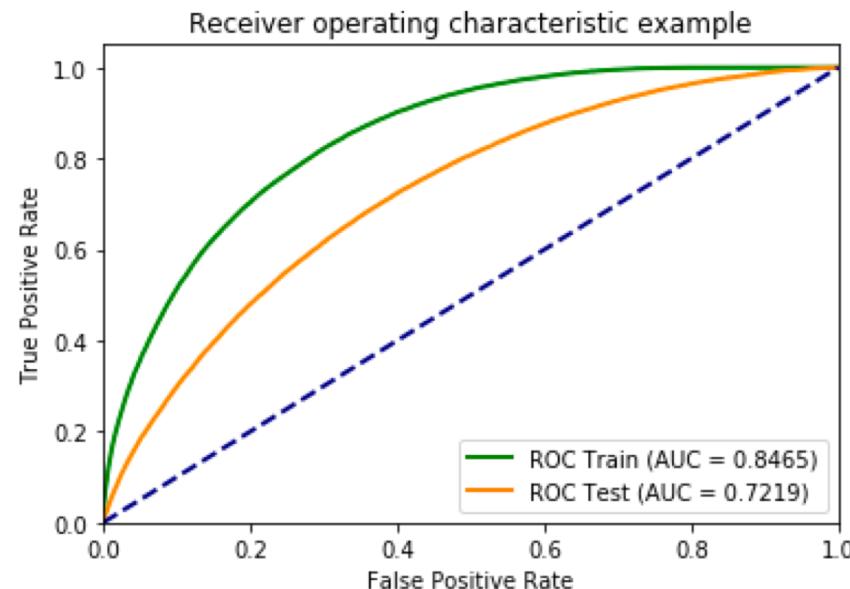


Glmnet build logistic regression model

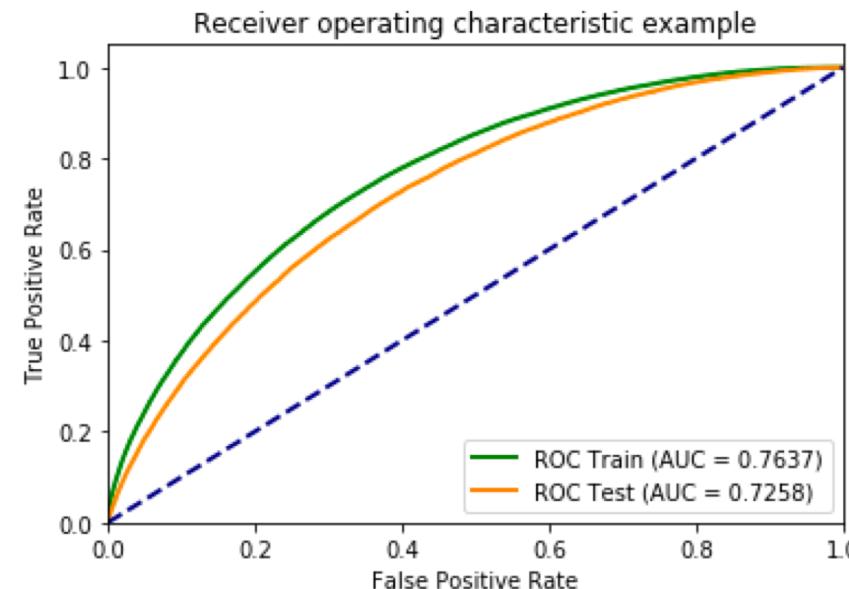


Area under the curve (AUC) 0.7187 on test dataset

Random forest and xgboost improve prediction performance



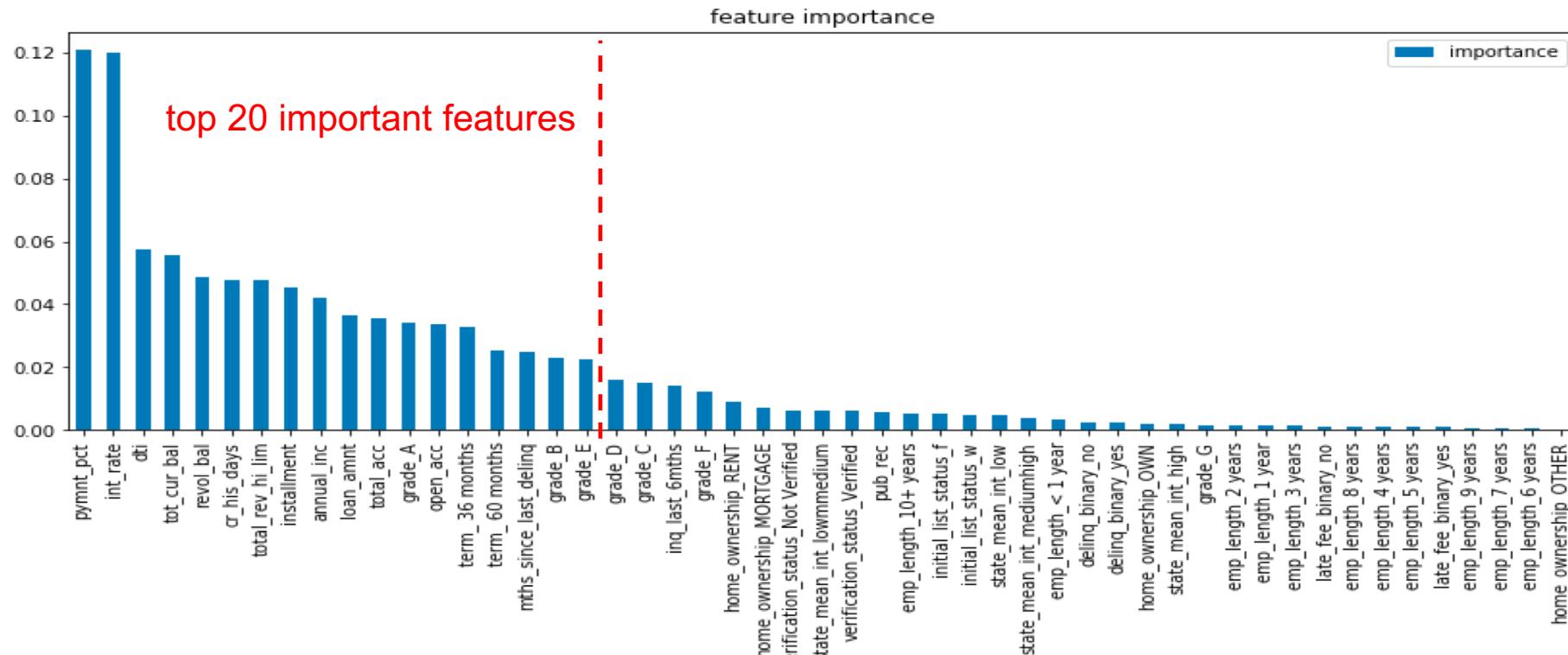
Random forest



xgboost

Since several numerical features that I selected for model have outlier, tree-based models show enhanced prediction capability.

Project conclusions



1. This project built three machine learning models to predict loan customers, who completed current payment, will fail to make next payment in the future.
2. For those customers, percentage of loan being paid, interest rate, debt-to-income (dti) ratio and total current balance play more important role than any other features in determining their subsequent payment behavior.