Loan Default Analysis and Prediction

Shiqing (Stacey) Wang

1

Introduction

Information about the background of the problem and business objectives

2

Data Preprocessing

Explanatory data analysis, data cleaning and feature selection

3

Modeling and Evaluation

Algorithm selection and comparison, evaluation of the selected models

4

Conclusions

Findings of the study, interpretation of results and recommendations for next step

01

Introduction

Introduce the background of the problem, and the business objectives of this study

Introduction

Background

- LendingClub is the world's largest peer-to-peer lending platform
- Analyze the loan default issue to make better decision
- Used loan data from 2007 to 2011 in this project

Objectives of the study

- Identify important predictor variables
- Build a model to predict which customers are most likely to default

02

Data Preprocessing

Explanatory data analysis, data cleaning and feature selection

Data Preprocessing

Missing data imputation

- Drop data that is ~100% missing
- Replace numerical missing data with median
- Assign categorical missing data a new level

Data cleaning

- Remove outliers
- Remove low probability (<5%) observations
- Convert data type as needed

Explanatory data analysis (EDA)

- Visualize overall default rate
- Explore default rate VS. different explanatory variables, such as loan amount, interest rate, annual income, grade, etc.

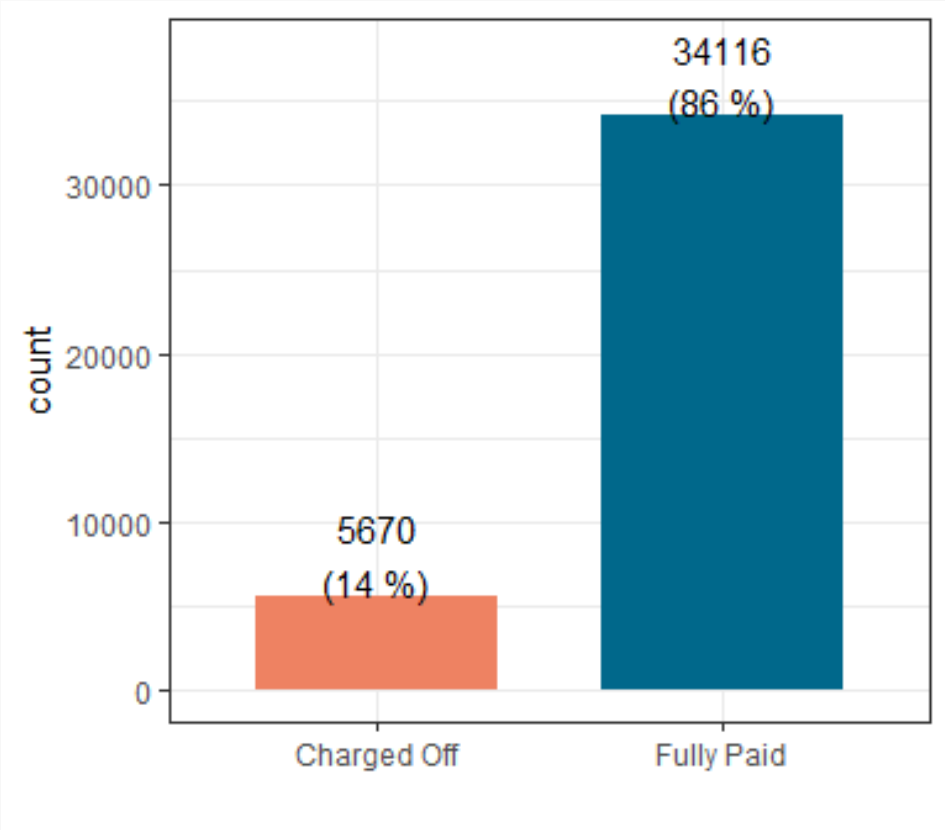
Feature Engineering and Selection

- Create new features with more predicting power
- Examine collinearity between numerical variables
- Examine feature importance

Explanatory Data Analysis

Overall default

- According to the training data, about 14% of the customers defaulted
- The loan default data is imbalanced

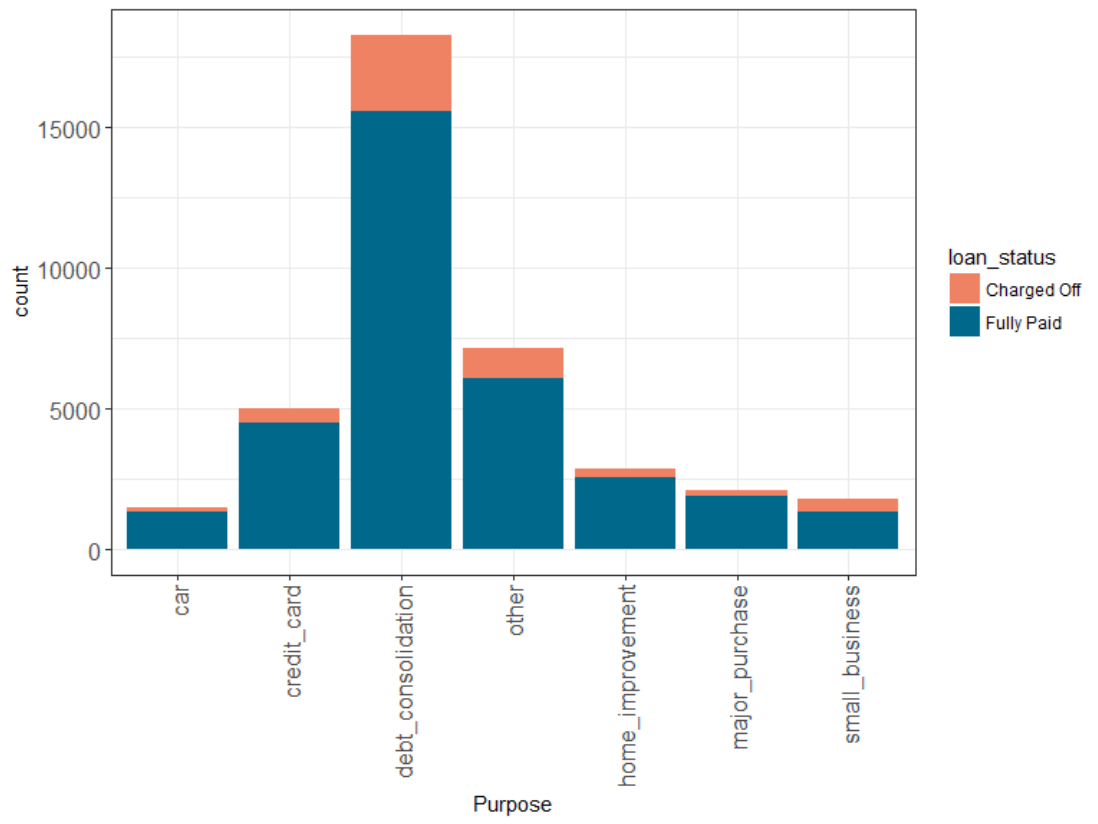


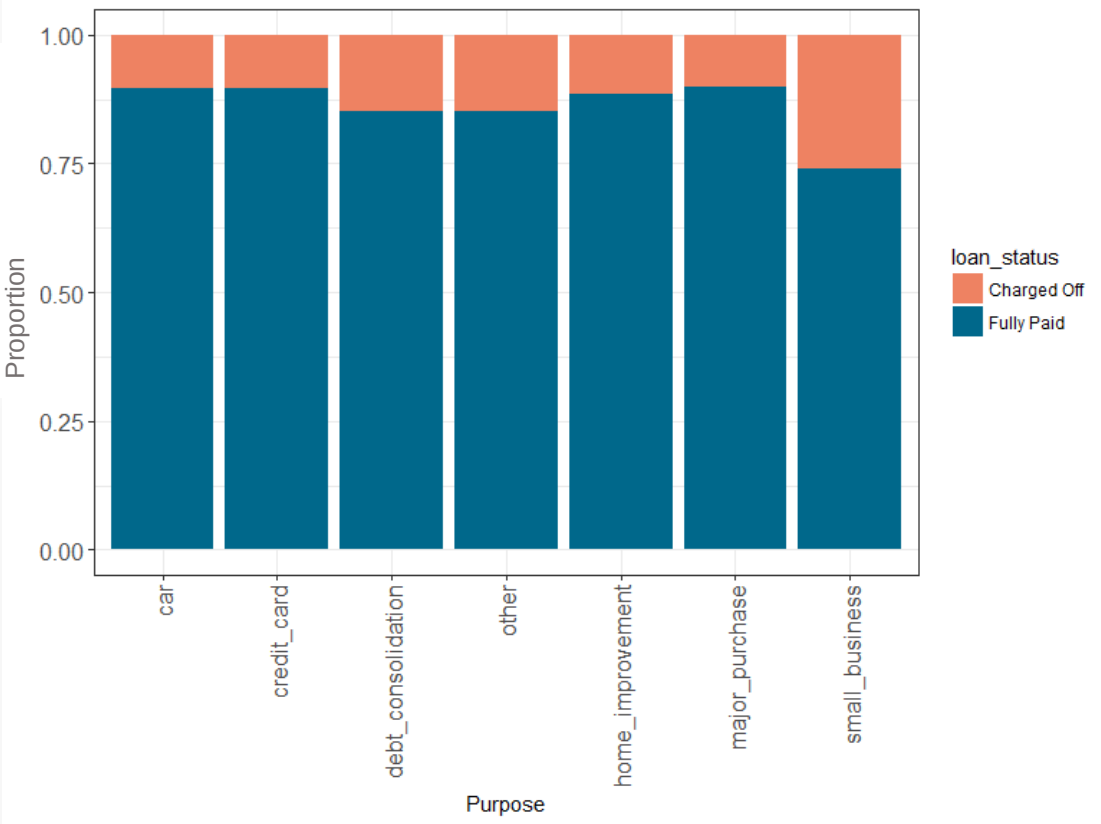
** Charge off is the declaration by a creditor that an amount of debt is unlikely to be collected.

Explanatory Data Analysis

Loan purpose VS. Default

- Small business borrowers are more likely to default

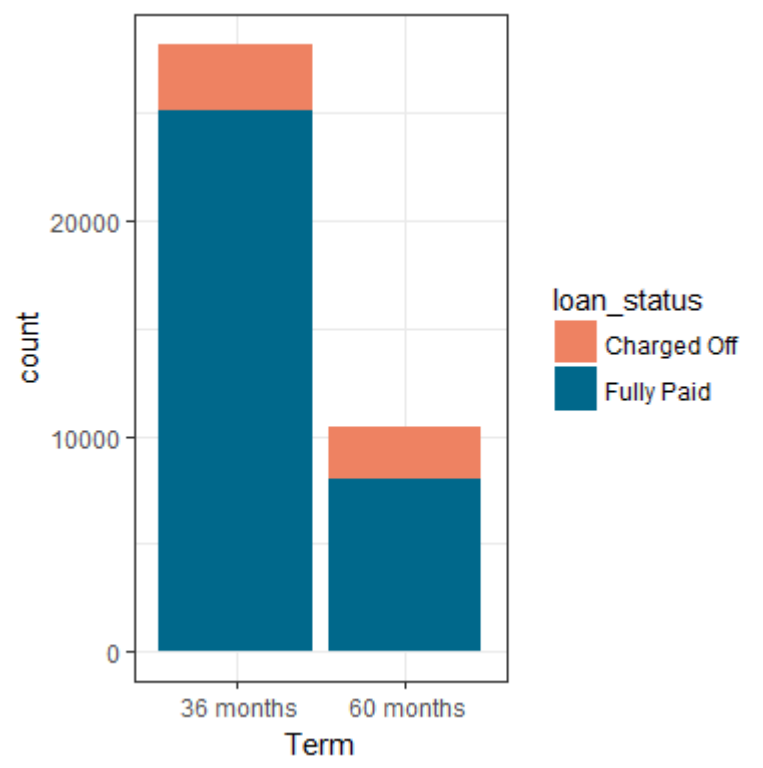


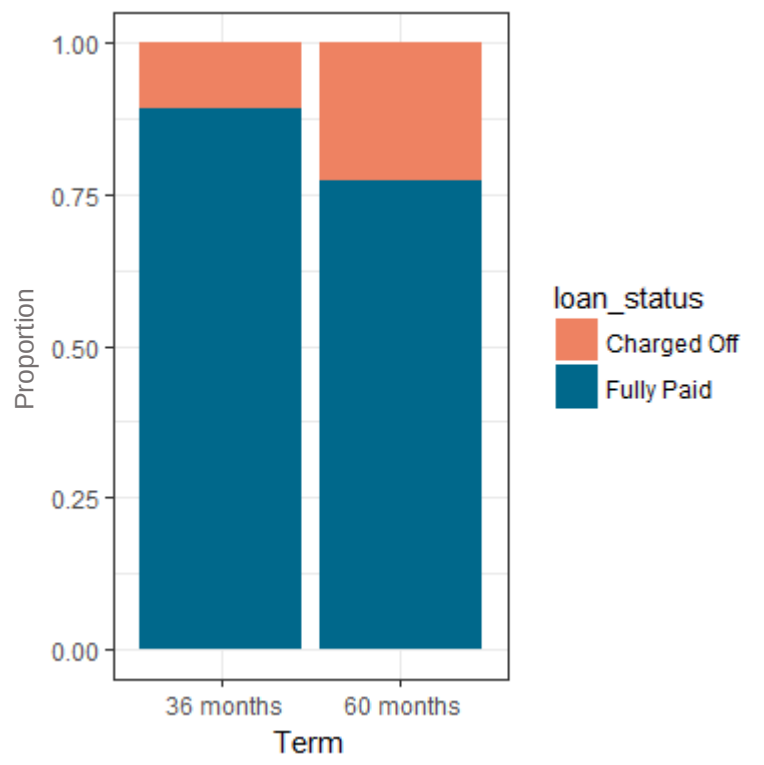


Explanatory Data Analysis

Loan term VS. Default

- Long term loan borrowers are more likely to default

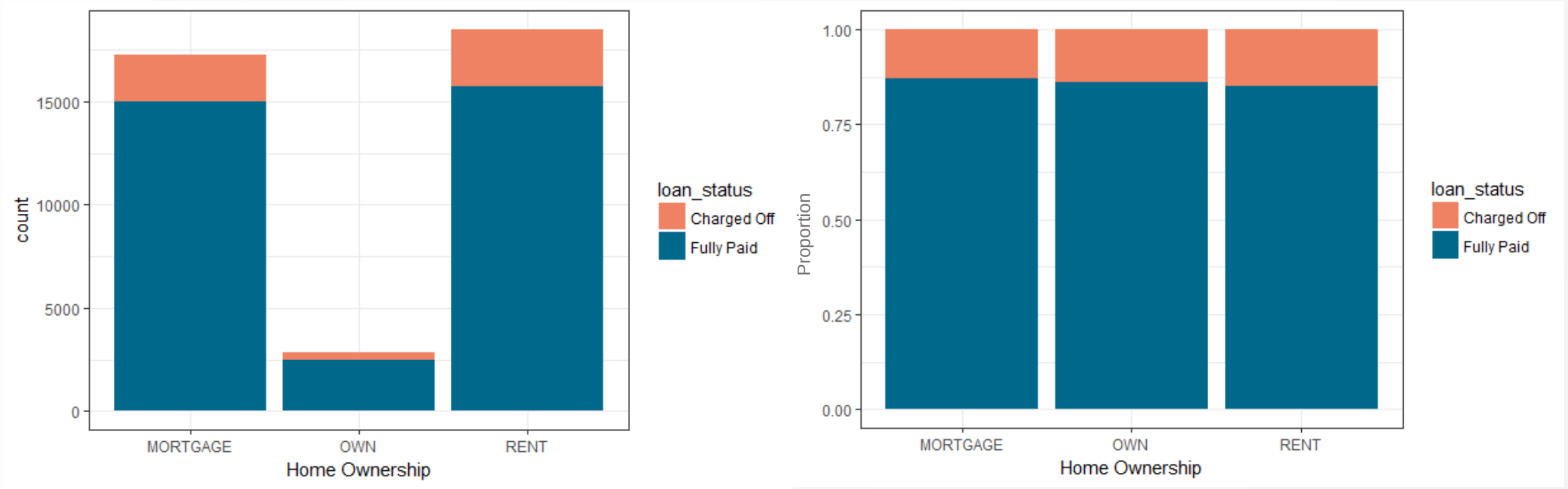




Explanatory Data Analysis

Home ownership VS. Default

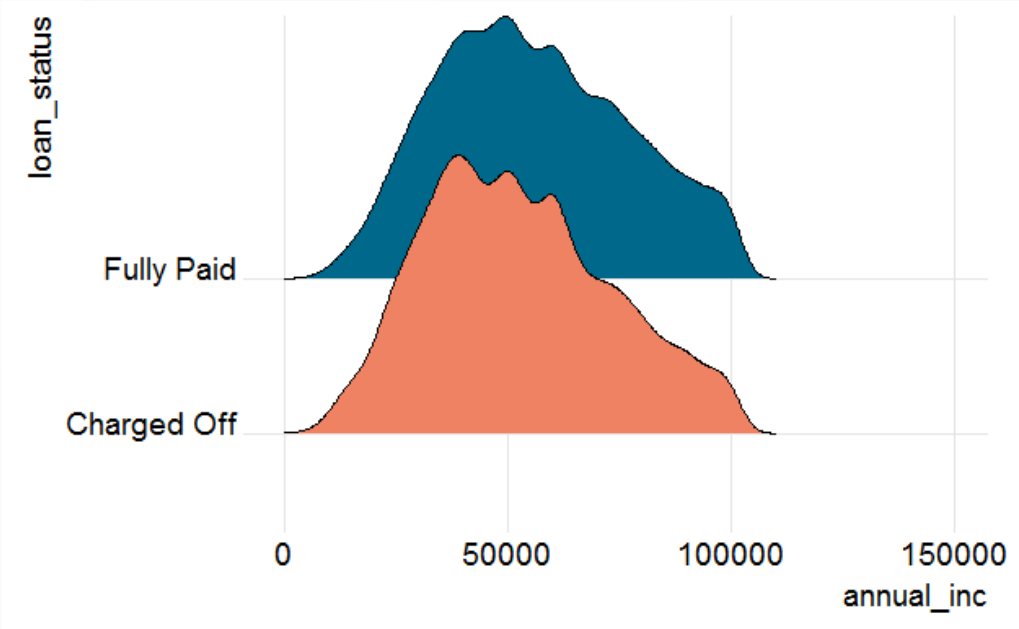
- Most borrowers are either under mortgage or renting
- Home ownership seems not have strong relationship with loan default

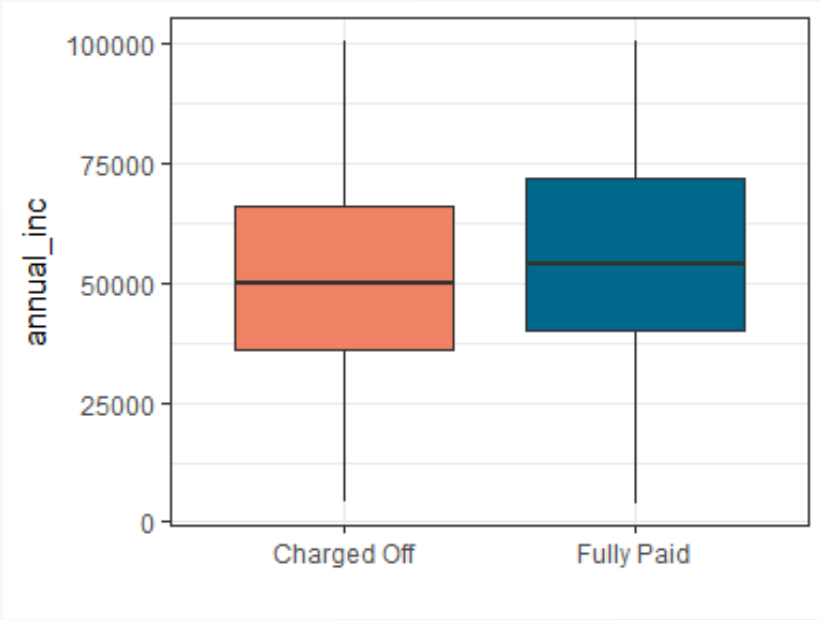


Explanatory Data Analysis

Annual income VS. Default

- Borrowers with lower annual income tend to default

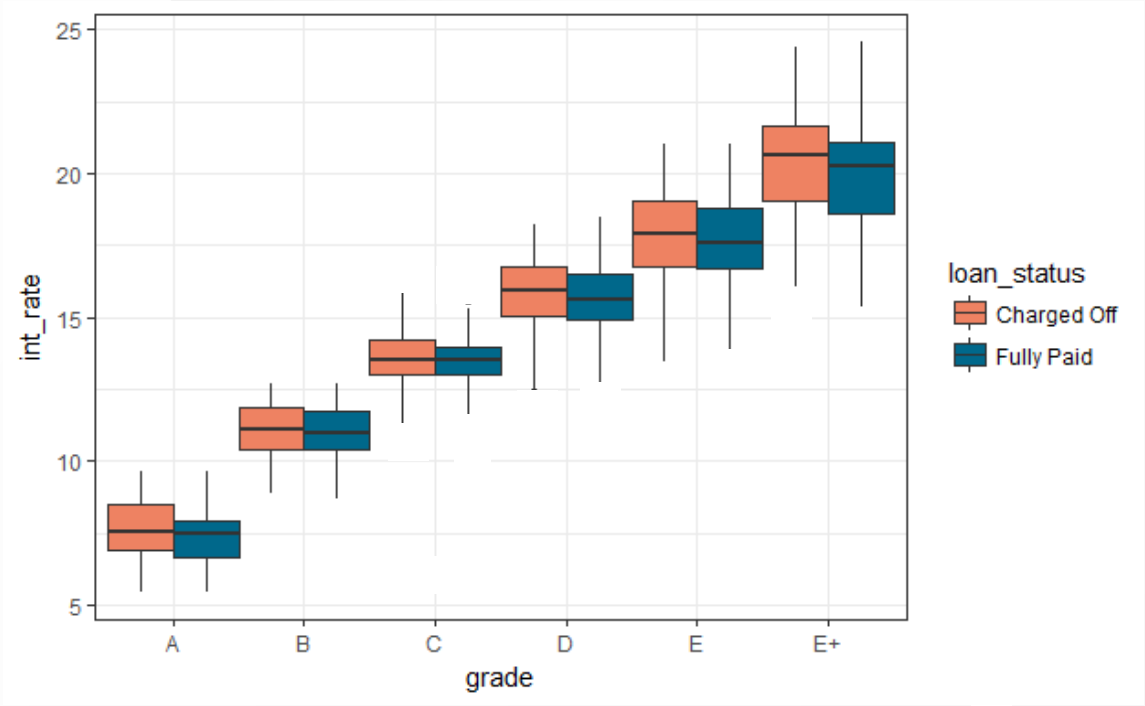


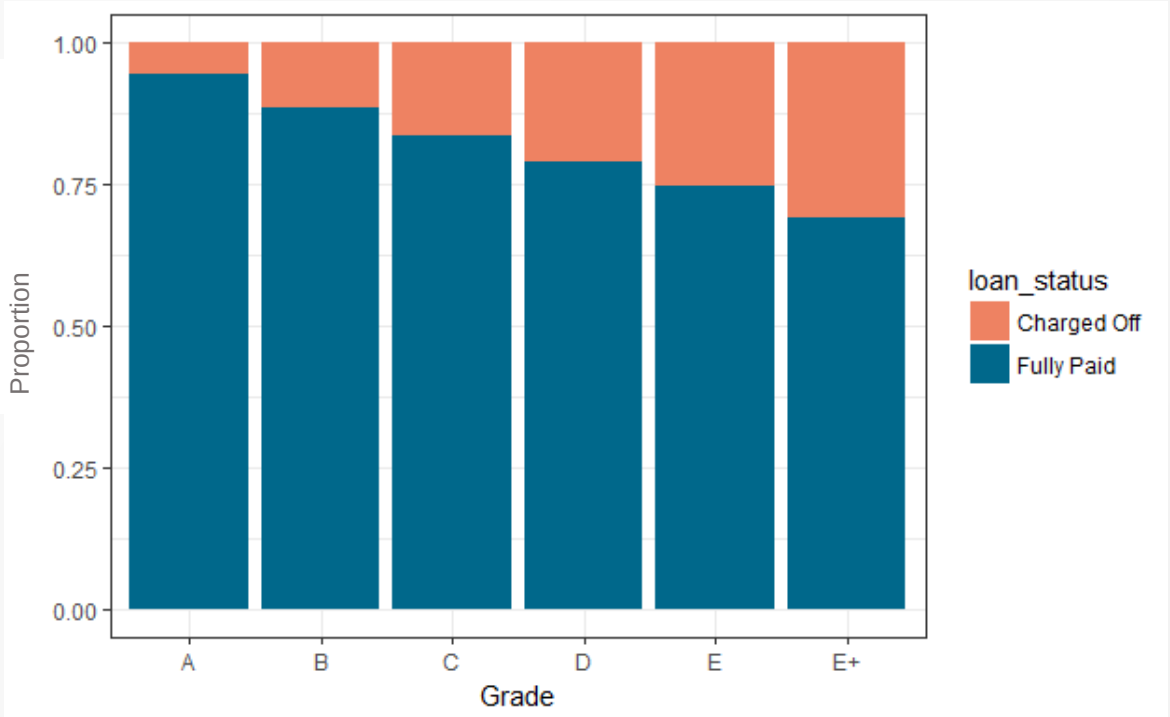


Explanatory Data Analysis

Interest rate, loan grade and default

- Higher loan grade corresponds to higher interest rate and higher default rate





Explanatory Data Analysis

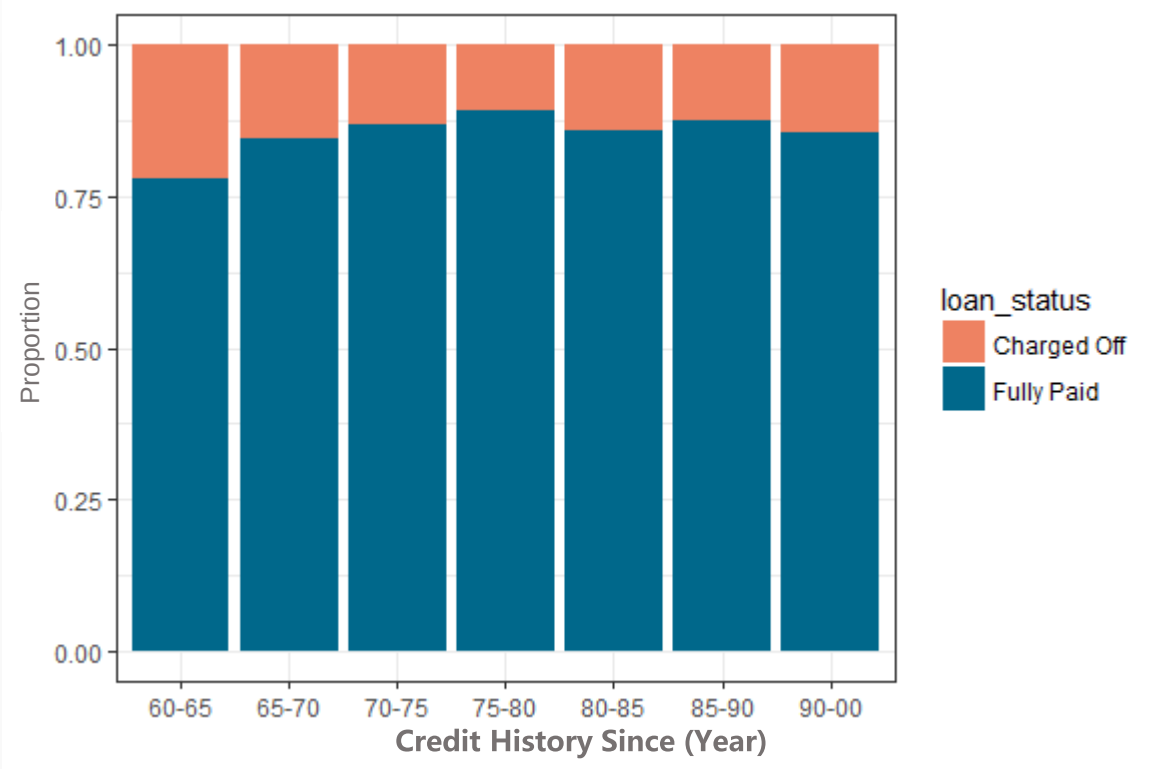
Loan amount VS. Default

- Generally, defaulted loan has wider range, and higher amount

Feature Engineering

New feature indicating borrower's age

- The loan data does not contain borrowers' age information
- Extract year from feature "earliest credit line", and group them into 5 year bucket
- Borrowers with credit history since 1960-1965 show highest default rate, which indicates those borrowers at their 50~60s might not have good financial management ability



Feature Engineering

New feature indicating region with higher default rate

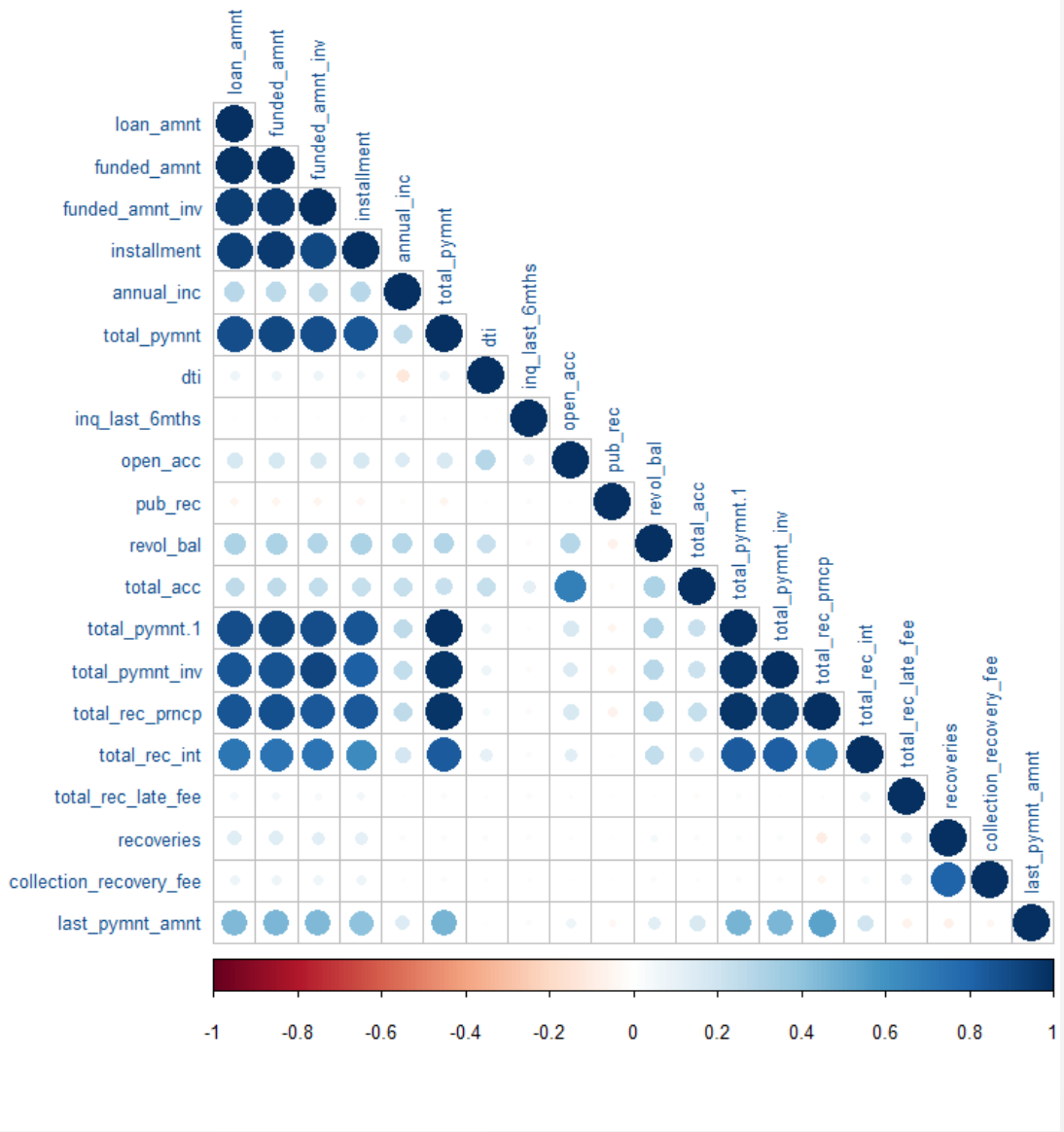
- Group data by state and check which states have highest default rate
- Create new variables indicating if a borrower is from one of those states with high default rate

State	Default rate (%)
CA	31.13
NY	16.88
FL	13.52
TX	12.73
NJ	7.75

Feature Selection

Collinearity between numerical variables

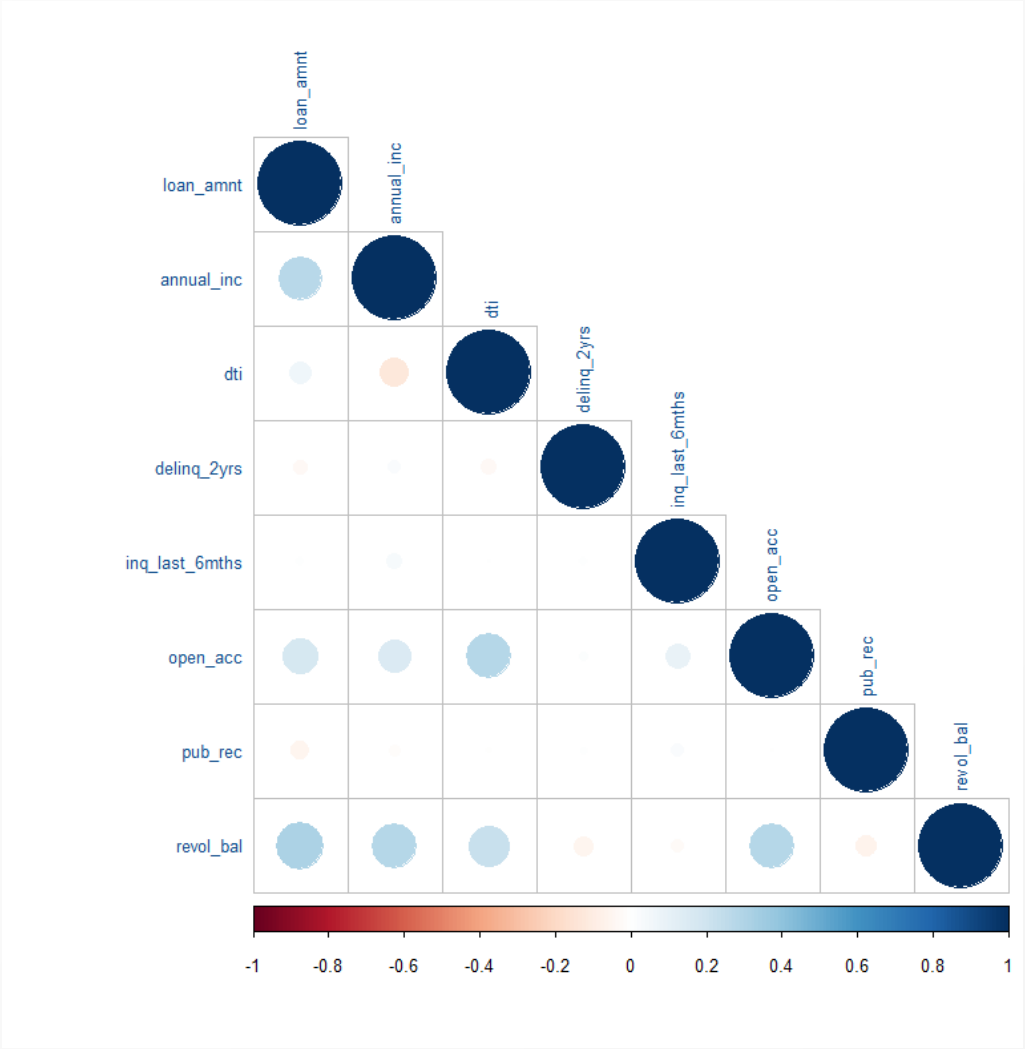
- Some highly correlated pairs
- Redundant variables need to be removed



Feature Selection

Collinearity check after removing redundant variables

- Now the correlation between predictor variables are very low



Modeling and Evaluation

Algorithm selection and comparison, evaluation of the selected models

Algorithm Selection

Logistic Regression

- A statistical method to predict whether an event will happen or not

Pros:

- Robust algorithm, easier to inspect and less complex

Cons:

- May over fit the data when the training set is high dimensional
- Sensitive to outliers and missing values

Random Forest

- Construct multiple decision trees and use the mode of results from individual trees to make prediction

Pros:

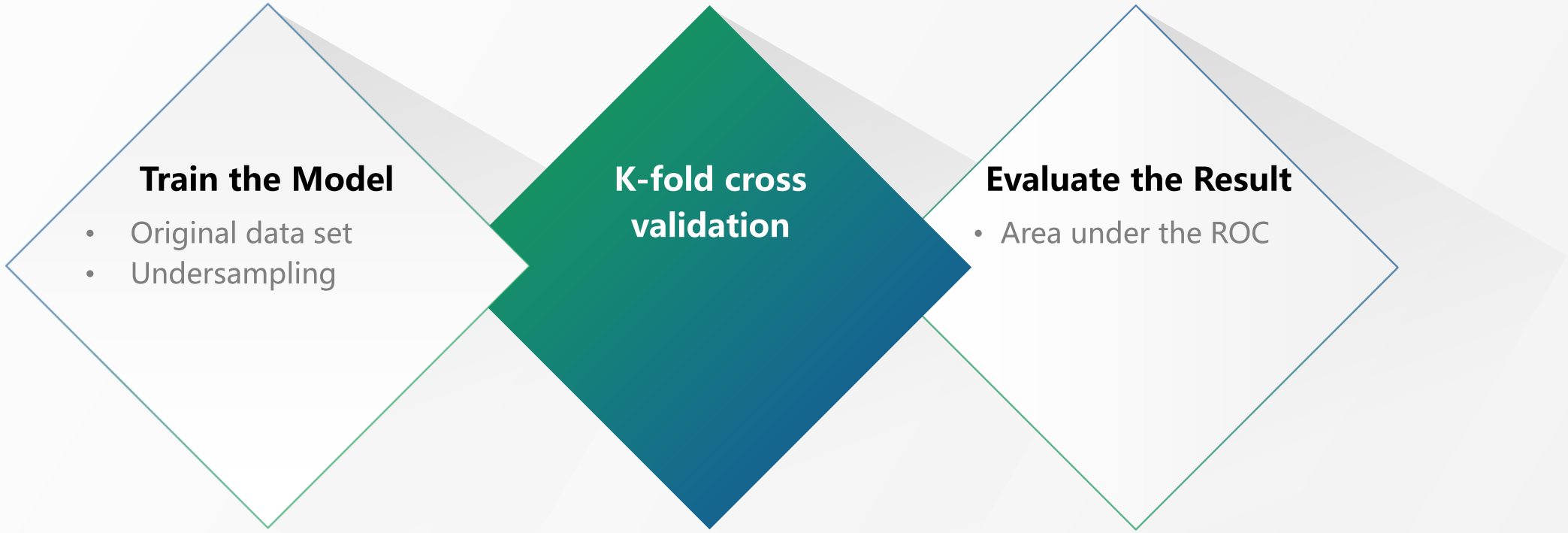
- Higher classification accuracy
- Less prone to overfitting

Cons:

- Difficult to analyze theoretically
- Large number of decision trees may slow down the algorithm

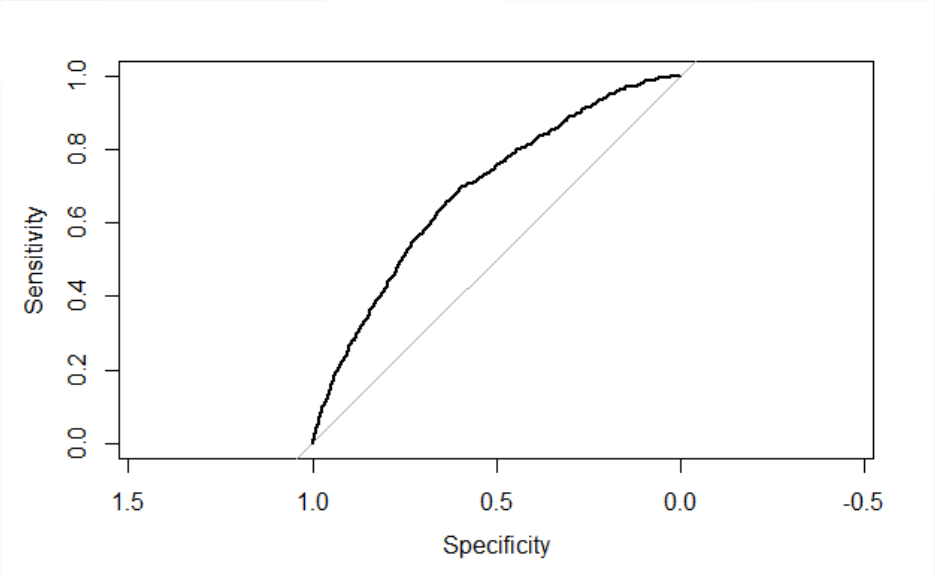
Modeling

The process can be summarized into 3 major steps:



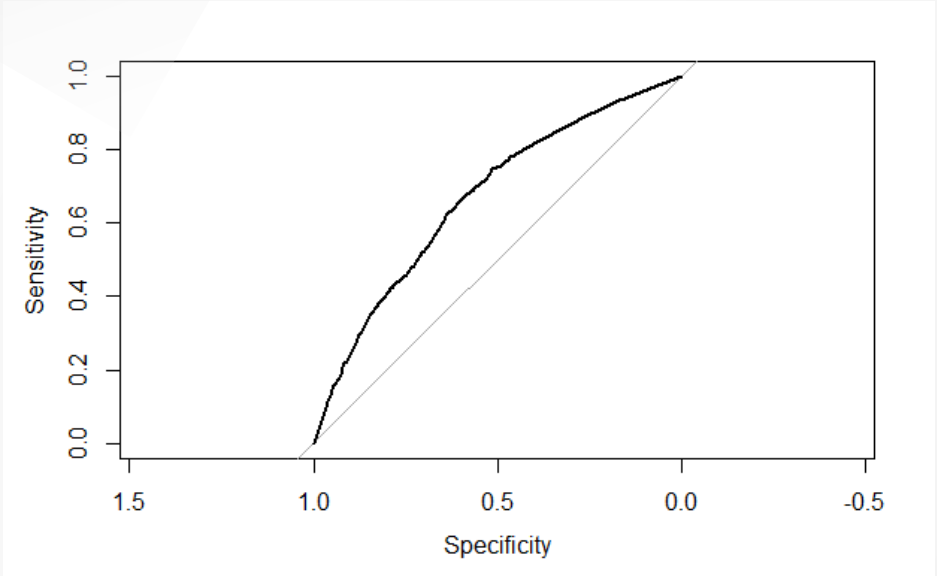
Modeling Results

GLM model



- AUC: 0.67

RF model



- AUC: 0.64

- Not a very high AUC
- Ways to improve:
  --Try 10 fold cross validation --Collect more data about borrowers who have defaulted

04

Conclusions and Recommendations

Findings of the study, interpretation of results and recommendations for next step

Conclusions and Recommendations

The most important predictors for default

- Annual income
- Loan purpose (small business)
- Interest rate
- State
- Number of derogatory public records

Recommendations for future study

- Try other machine learning algorithms to improve the model
- More feature engineering to have a better understanding of the data
- LendingClub should focus on checking borrower's infomation with most predicting power

THANK YOU

Shiqing (Stacey) Wang