

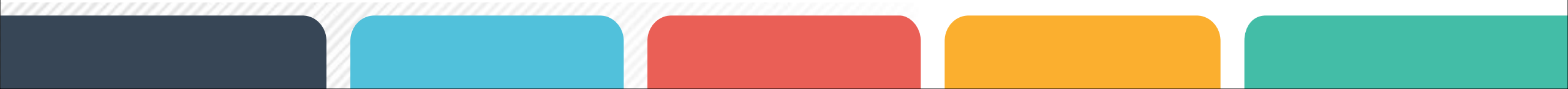


# Predict Default Loans with Lending Club Data



Hugo Yu

08-12-2020



# Intoruction



Build a tool to help the lending club investors to make better-informed dicesion on note choice



2007 to 2018 Lending Club historical data



Logistic regression, random forest, and gradient boosting tree

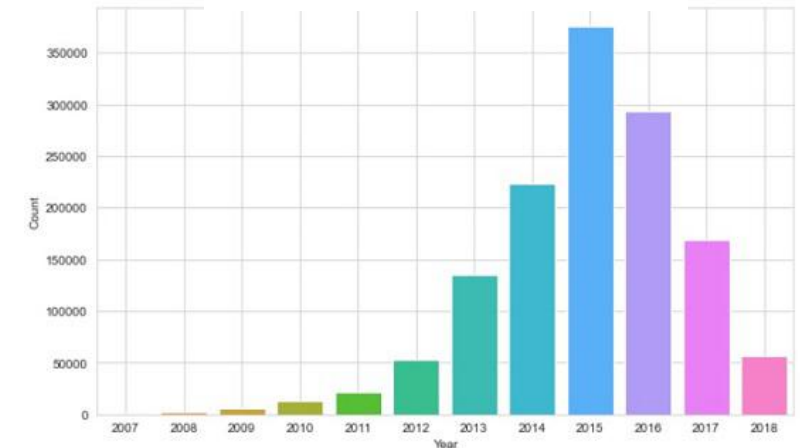


The model was able to predict the default loan with AOC score around 0.72 and the top important featres were identified

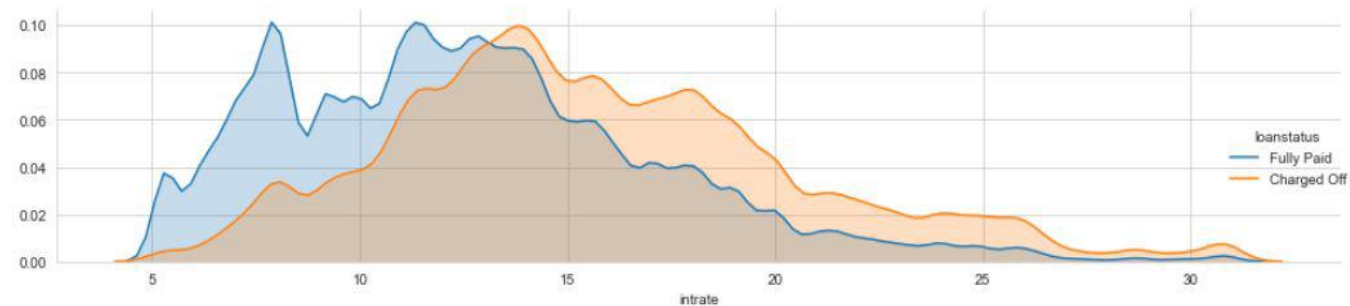
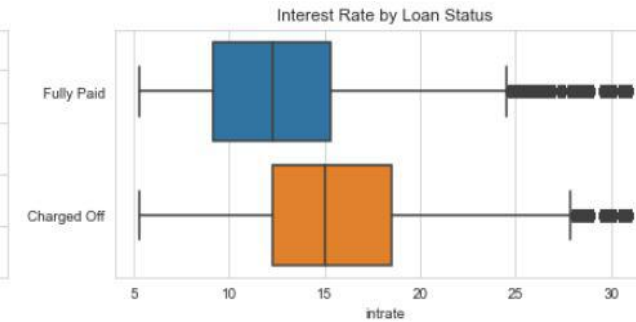
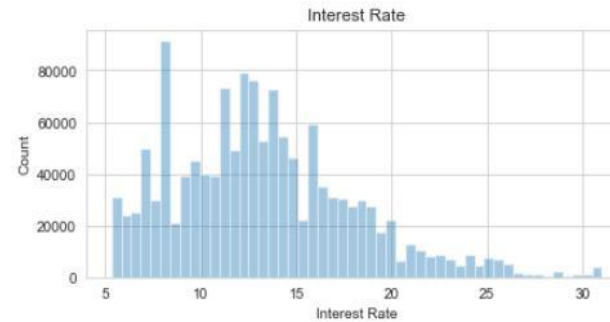
# EDA and Data Preprocessing

- Data description
  - historical 2007-2018
  - 2260701 loans
  - 150 associated features
  - latest note data by api
- Feature Selection
  - field match
  - missing data
  - highly correlated features
  - zipcode and text features
- Feature Engineering
  - label encoding
  - one-hot-encoding

accepted loans by year



Interest Rate



# Model Training and Evaluation

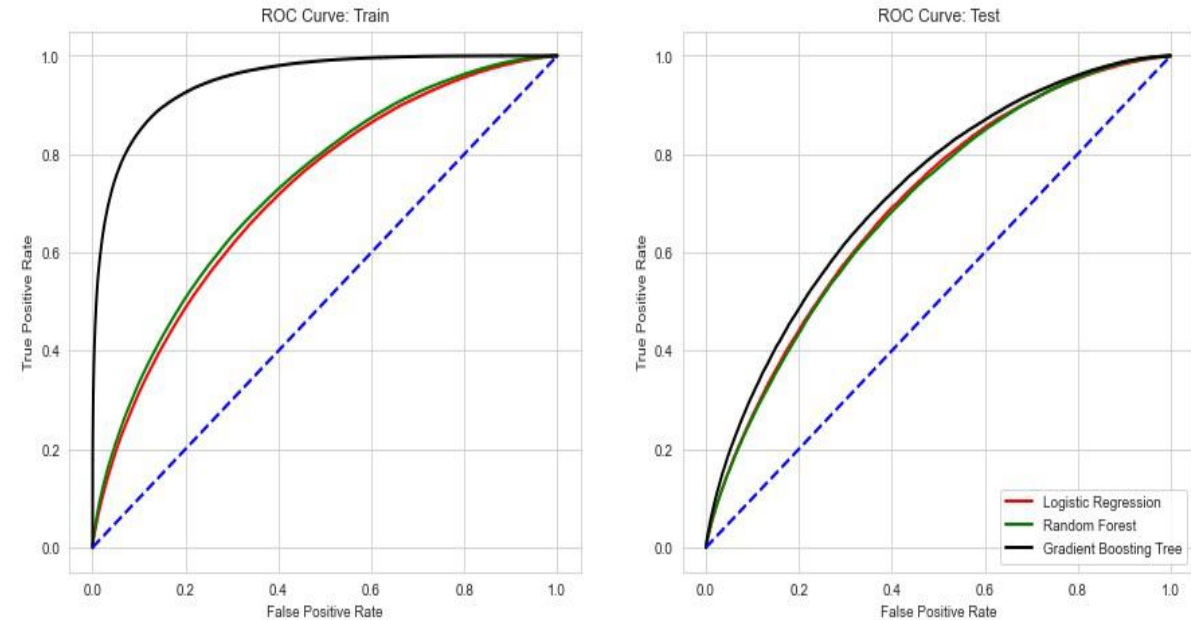
- Train/Test data split
  - train: test at 8:2

- Model Selection
  - Logistic regression
    - Recursive Feature Elimination
  - Random forest
  - Gradient boost tree
    - feature importance

## Top factors

|                 | importance |
|-----------------|------------|
| credithistory   | 0.101854   |
| mosinoldrevtlop | 0.073039   |
| mosinoldilacct  | 0.051138   |
| dti             | 0.045329   |
| annualinc       | 0.042021   |
| loanamnt        | 0.038239   |
| avgcurbal       | 0.031184   |
| intrate         | 0.029794   |
| revolbal        | 0.027767   |
| revolutil       | 0.026138   |

## Model Performance



In training dataset, GBT has the significantly higher AUC score than the other two. However, the GBT score is only slightly higher in test, indicating that the model is likely overfitting.

# Summary and Future Work

- The current model was able to predict the default loan at an AOC score around 0.72
- The top 5 factors:
  - credit score
  - months since the oldest revolving account
  - months since the oldest installment account
  - borrower's installment to debt ratio
  - and borrower's annual income
- Relate the zipcode feature to more meaningful information
- NLP analysis for text feature
- Fine tune the model
- Use FLASK to create an online tool