Loan Loss Reduction

By: Sergio Rodriguez

Outline

- Summary
- Business Problem
- Data
- Methods
- Results
- Conclusions

Summary

- Among five different classification machine learning models, Random Forest provides the best predictive accuracy at determining whether or not a prospective borrower will default on their loan.
- Top criteria when considering a potential loan applicant are:
  - Interest Rate
  - Debt to Income Ratio
  - Number of inquiries in the last 6 months
  - Term of the loan
  - Average FICO

Business Problem

- In order to consistently provide investor returns, Lending Club needs to appropriately vet their applicants to ensure that they indeed will pay off their loan in full with interest.
- To do so, Lending Club has asked us to analyze their prior loans over the period of 2007 - 2020 Q3 to provide them with insights on
  - 1. What features are most important when assessing a potential applicant’s creditworthiness
  - 2. Provide them with the best classification model that can accurately predict whether or not a potential applicant's loans will be charged off.
- Doing so, will allow Lending Club to make more sound decisions when offering lines of credit to future applicants.

Data

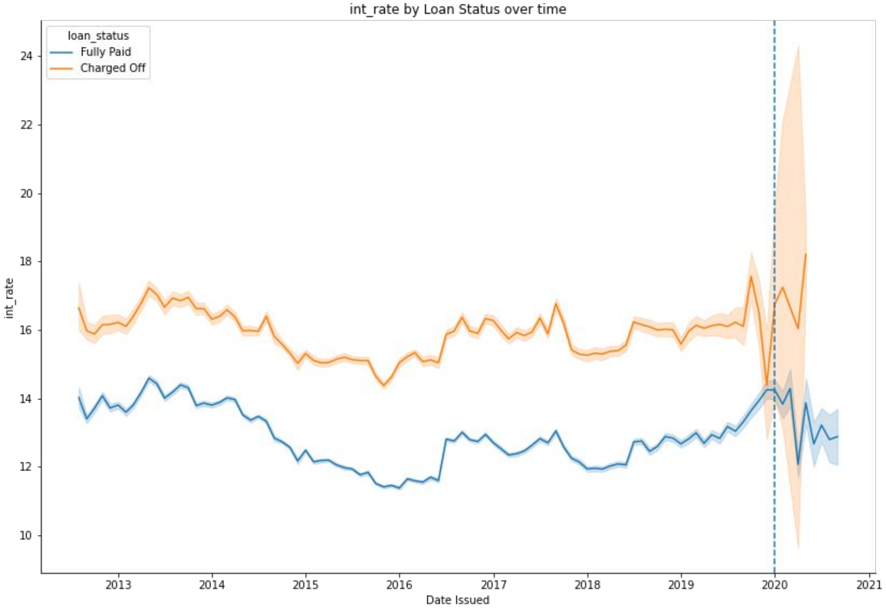
The data used in this project includes loan level data for accepted loan data from 2007 - 2020Q3 from Lending Club. This data was accessible via Kaggle link below. The features included in the data are general criteria included for applicants when applying for a loan, as well as active loan criteria. Below are some of the feature names and corresponding definitions:

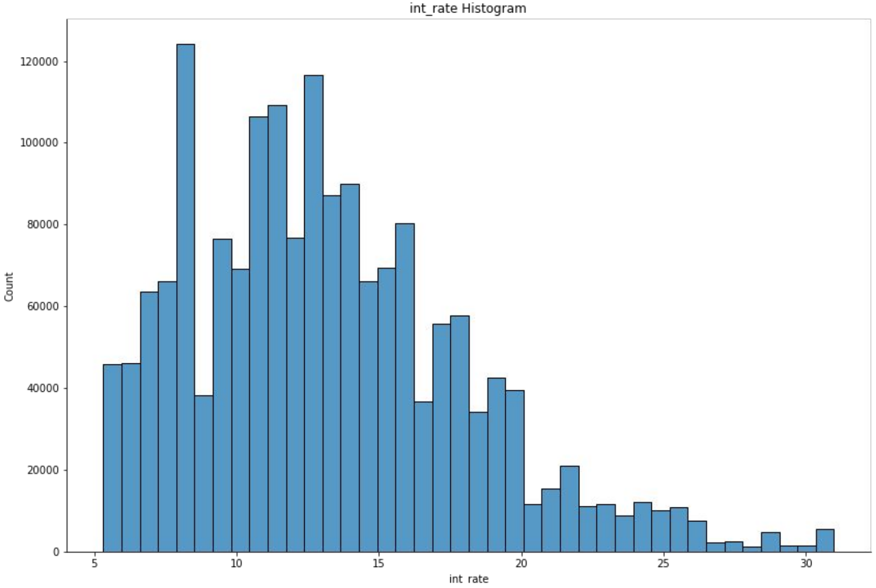
- loan_amnt : The listed amount of the loan applied for by the borrower
- term : The number of payments on the loan. Values are in months and can be either 36 or 60.
- int_rate : Interest Rate on the loan
- installment : The monthly payment owed by the borrower if the loan originates.
- emp_length : Employment length in years
- annual_inc : The self-reported annual income provided by the borrower during registration.
- inq_last_6mths : The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
- pub_rec : Number of derogatory public records

Link to Kaggle dataset:

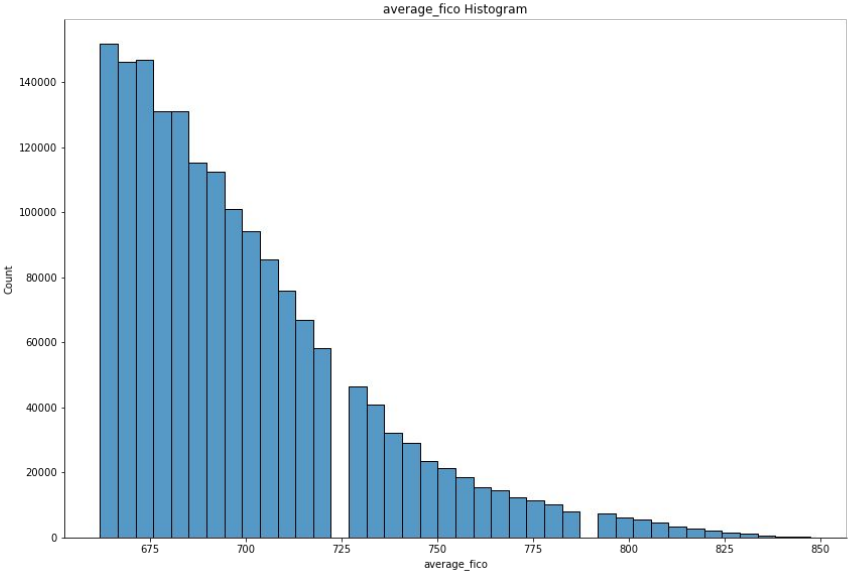
- https://www.kaggle.com/ethon0426/lending-club-20072020q1

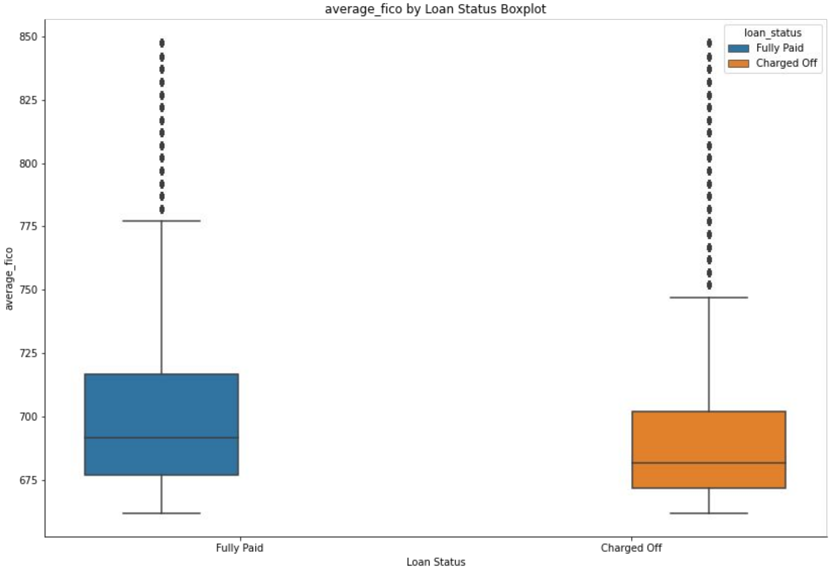
Methods - EDA: Interest Rate





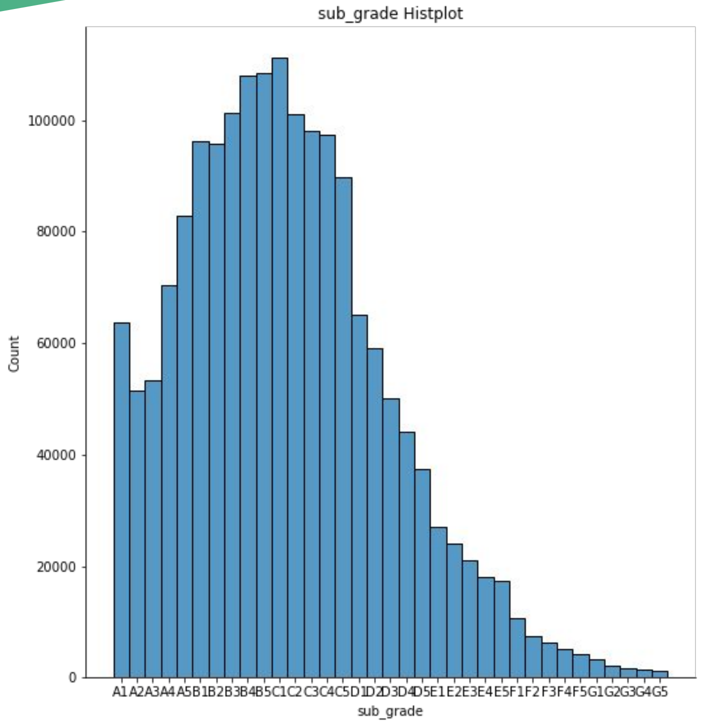
Methods - EDA: Avg FICO Score



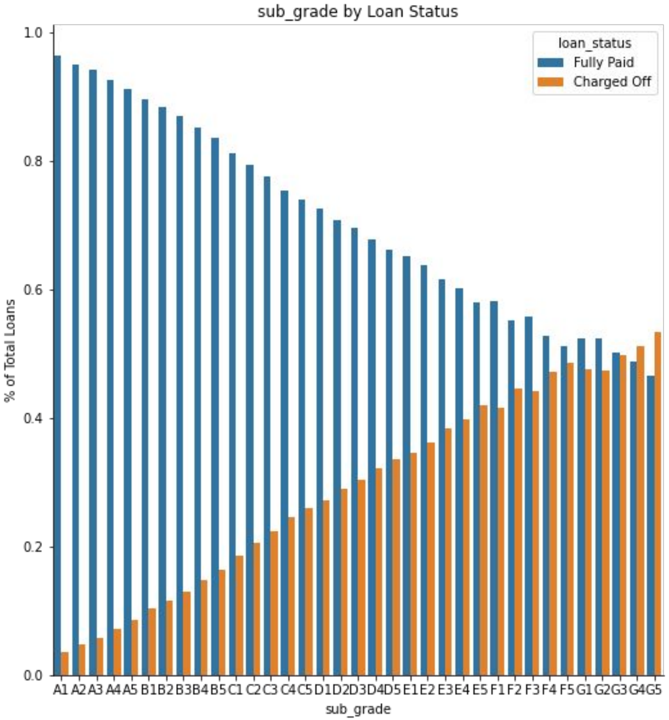


Methods - EDA: Sub Grades

- Clear relationship with decreasing subgrade and increase in charge offs



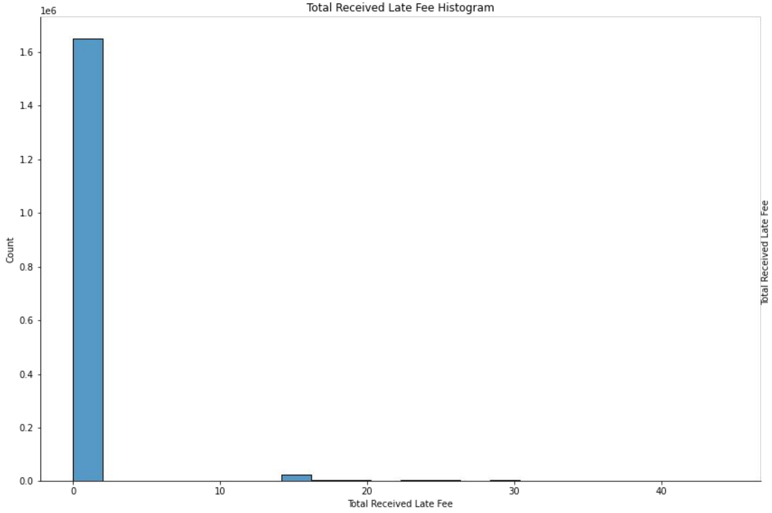
sub_grade Histplot



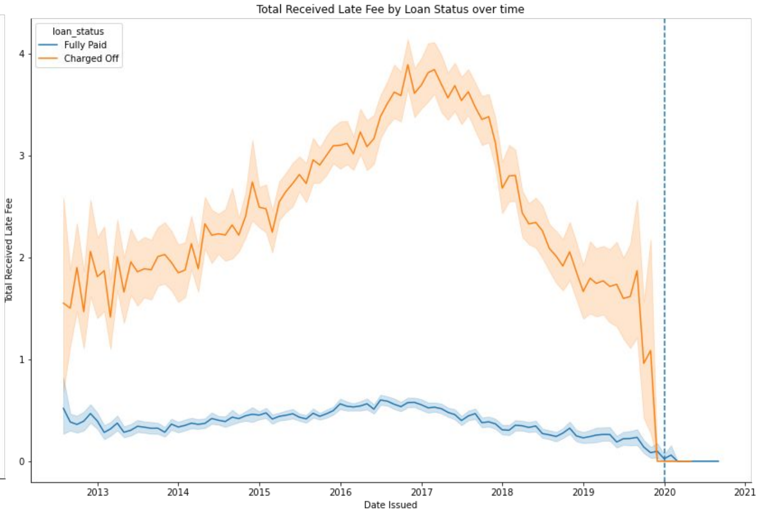
sub_grade by Loan Status

Methods - EDA: Late Fee

- If a client is late for any payment they are top priority for monitoring as most all borrowers who fully pay are never late



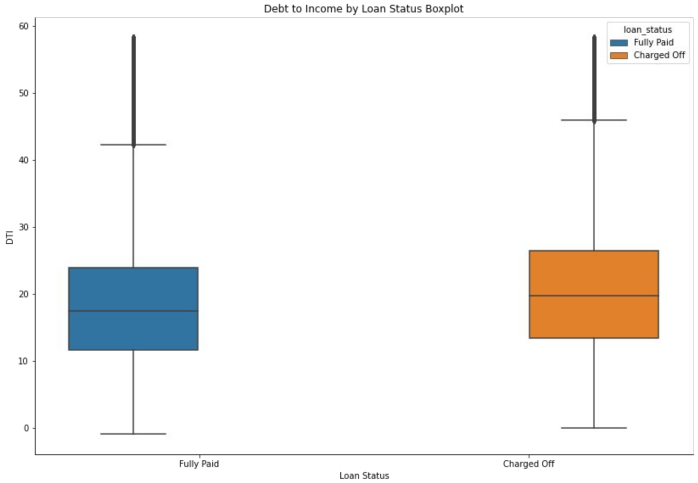
Total Received Late Fee Histogram

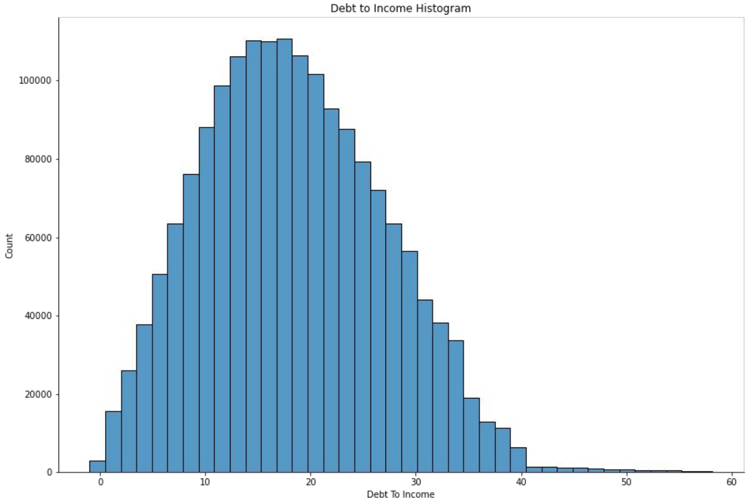


Total Received Late Fee by Loan Status over time

Methods - EDA: DTI

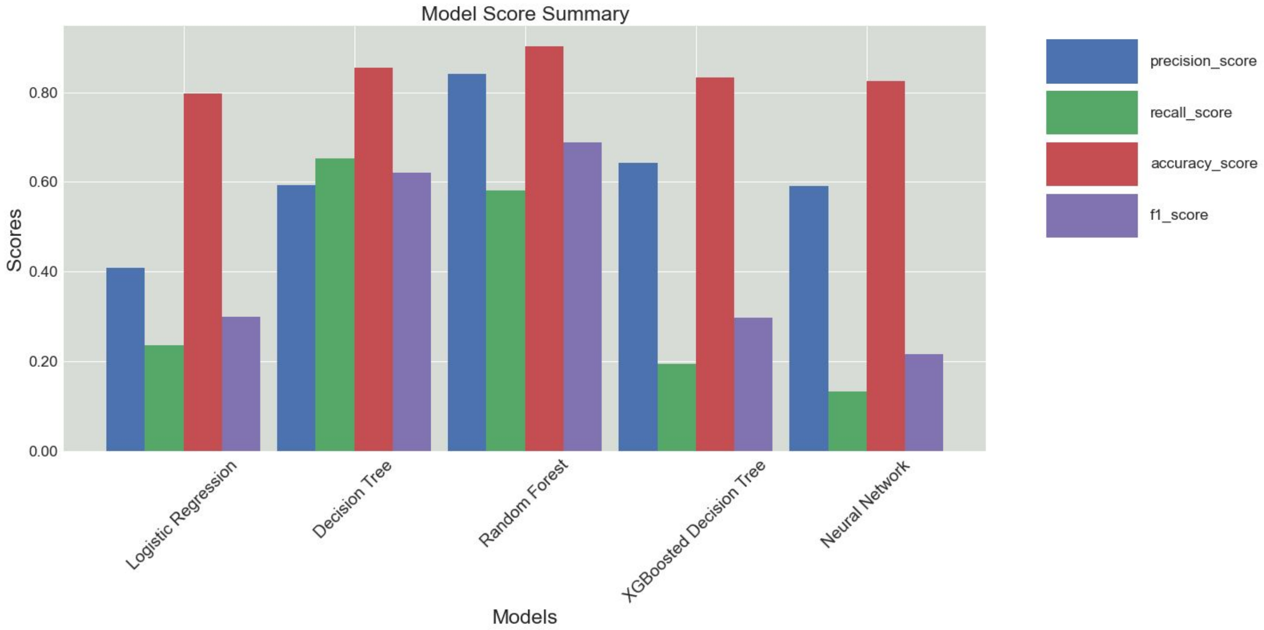
- Higher DTI's lead to higher charge offs





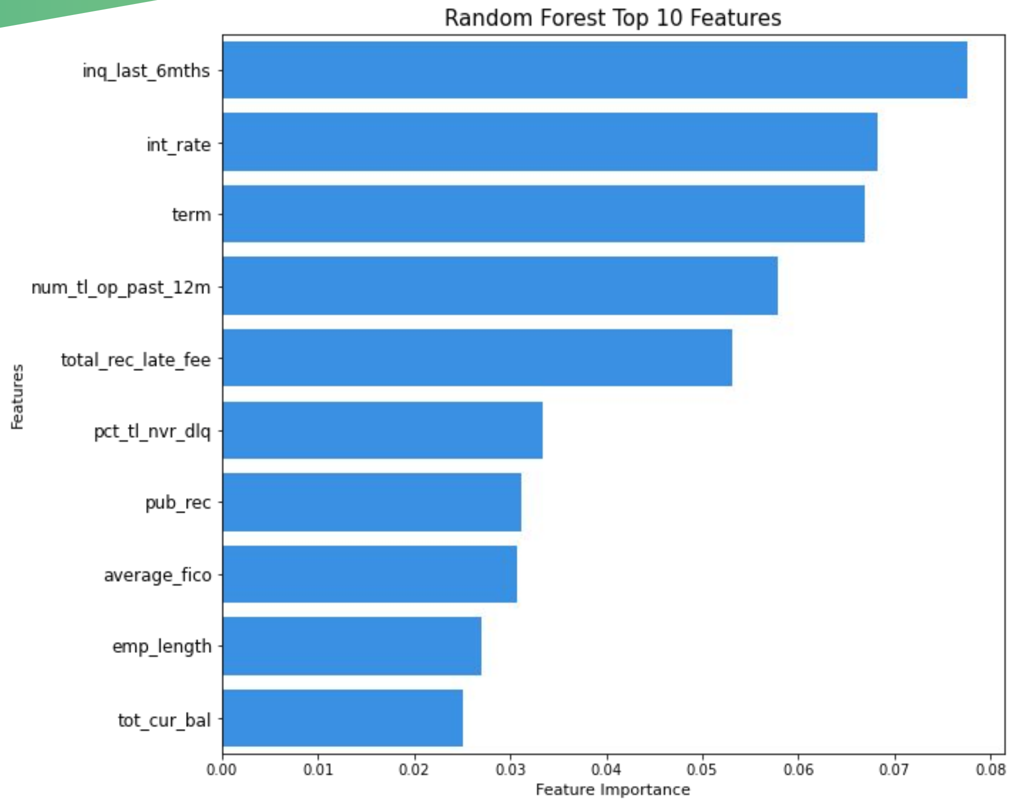
Results

- Across all models, Random Forest had highest accuracy at .90



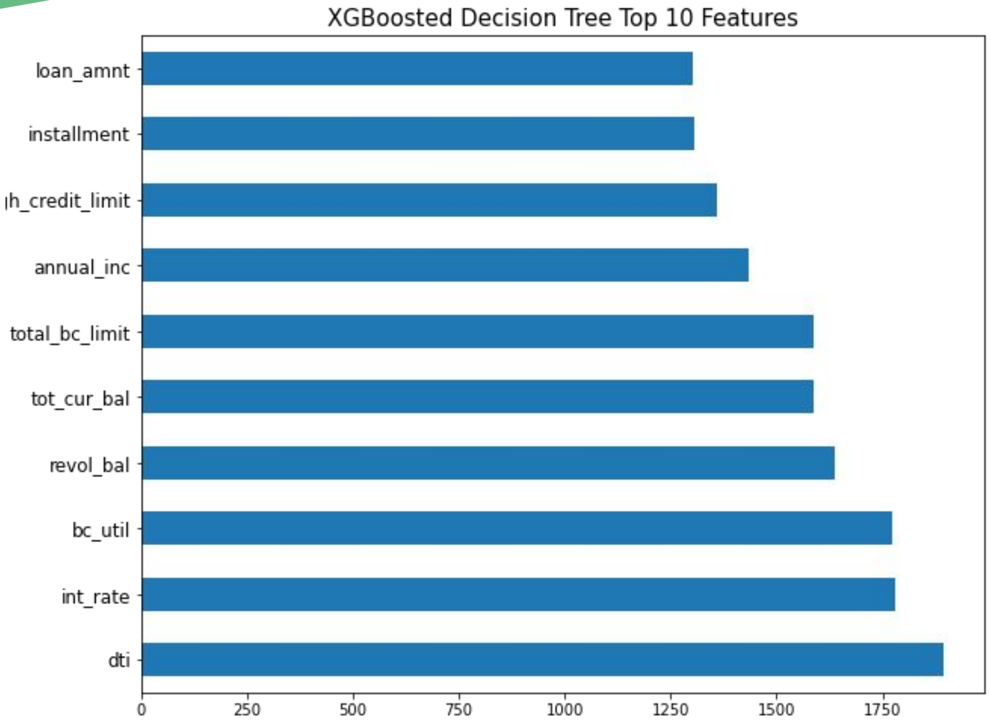
Results - Random Forest

- Top Features from our Random Forest model:
  - Inquiries in the last 6 months
  - Interest Rate
  - Term
  - % Loans never delinquent
  - Average FICO
  - Number of Public Records (derogatory)



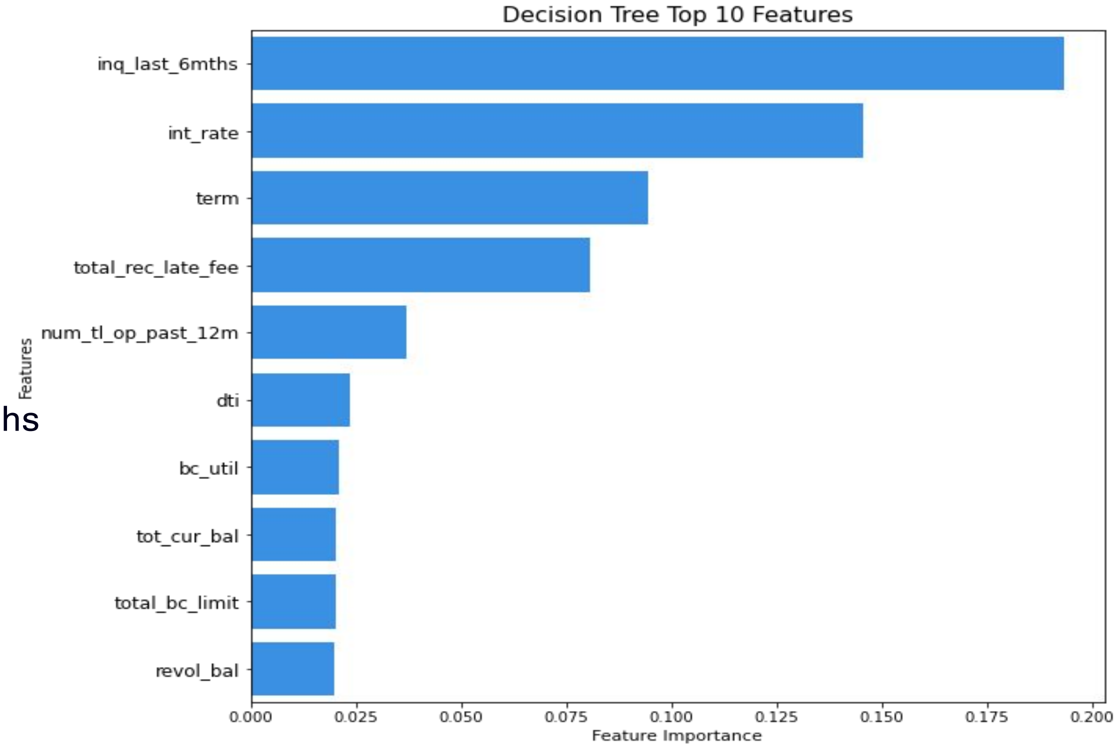
Results - XGBoosted Decision Tree

- Top Features from our XG Decision Tree:
  - Debt to Income
  - Interest Rate
  - Bankcard Utilization
  - Revolving Loan Balance
  - Total Current Balance



Results - Decision Tree

- Top Features from our Decision Tree:
  - # of Inquiries in last 6 months
  - Interest Rate
  - Loan Term
  - # of Late Fees Received
  - # of accounts opened in last 12 months



Conclusion

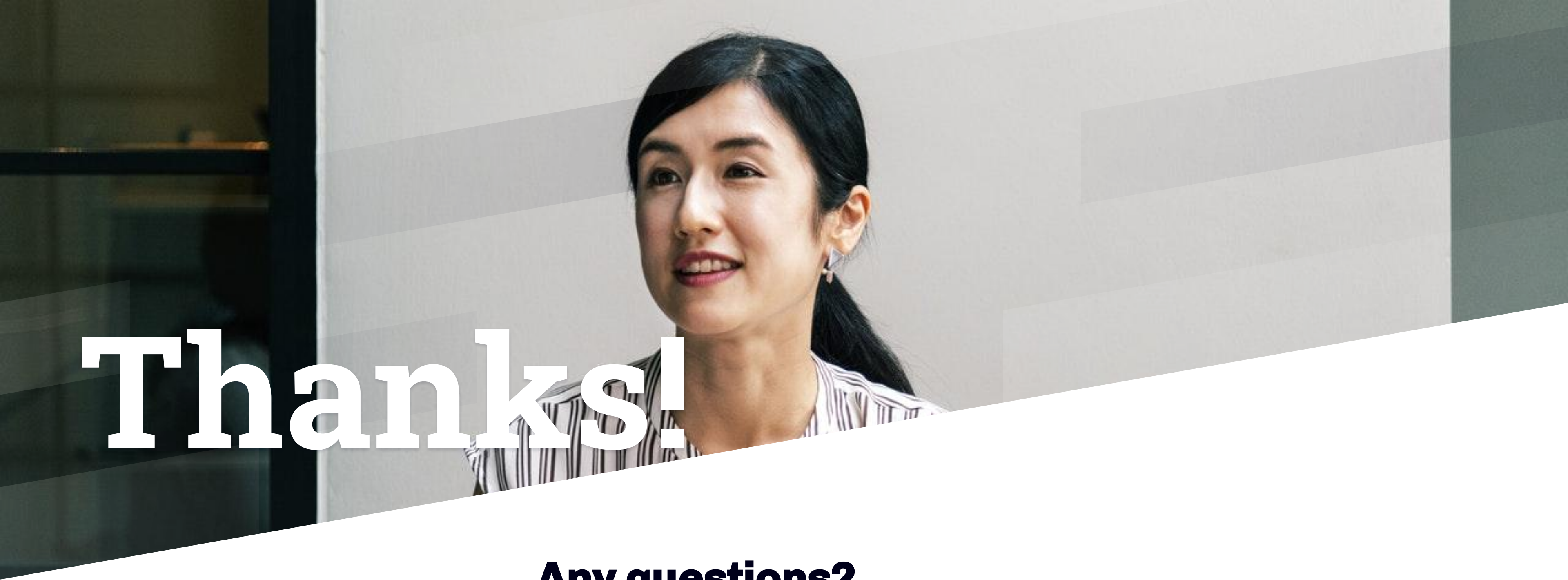
In conclusion, we would advise Lending Club to use a random forest model to predict loan losses of potential borrowers and perhaps. Criteria that Lending Club should be most aware in making the decision to extend a line of credit are the following:

- Interest Rate
- Debt to Income Ratio
- Number of inquiries in the last 6 months
- Term of the loan
- Average FICO
- What Sub Grade they would be classified in
- Bankcard Utilization

A combination of predictive modeling and criteria awareness by Lending Club should allow them to reduce loan losses for prospective borrowers.

Future Work

- It would be relevant to analyze how this model performed throughout COVID as it is a prime example of an unexpected shock that would test any risk model.
- Additionally, to be able to create an even more complex neural network would likely increase our NN models score, but I digress.
- Other considerations would be including additional economic indicators or consumer sentiment/outlook numbers
- Investigate the potential pricing of the underlying loans to the extent that they are sold to investors on the secondary market.
- Lastly, it would be interesting to have a model that considers active loans and probability of default for those loans in order to better prepare future windfall to Lending Club and its investors.



Thanks!

Any questions?

You can find me at:

- Email: srodriguez2742@gmail.com
- GitHub: @srodriguez2742