THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Modeling of Prediction Loan Default with Machine Learning

SEAS 6401

Instructor: Dr. Benjamin Harvey

Group Member:

Rui Zhang

Ran Wei

CONTENTS

INTRODUCTION

- **Project Goal:**
  - Understand the applicant's profile
  - Make a prediction
- **Data Set:**
  - Kaggle – Lending Club 2007-2020(Q3)
  - 2.9million loan records and 142 features

Feature Selection

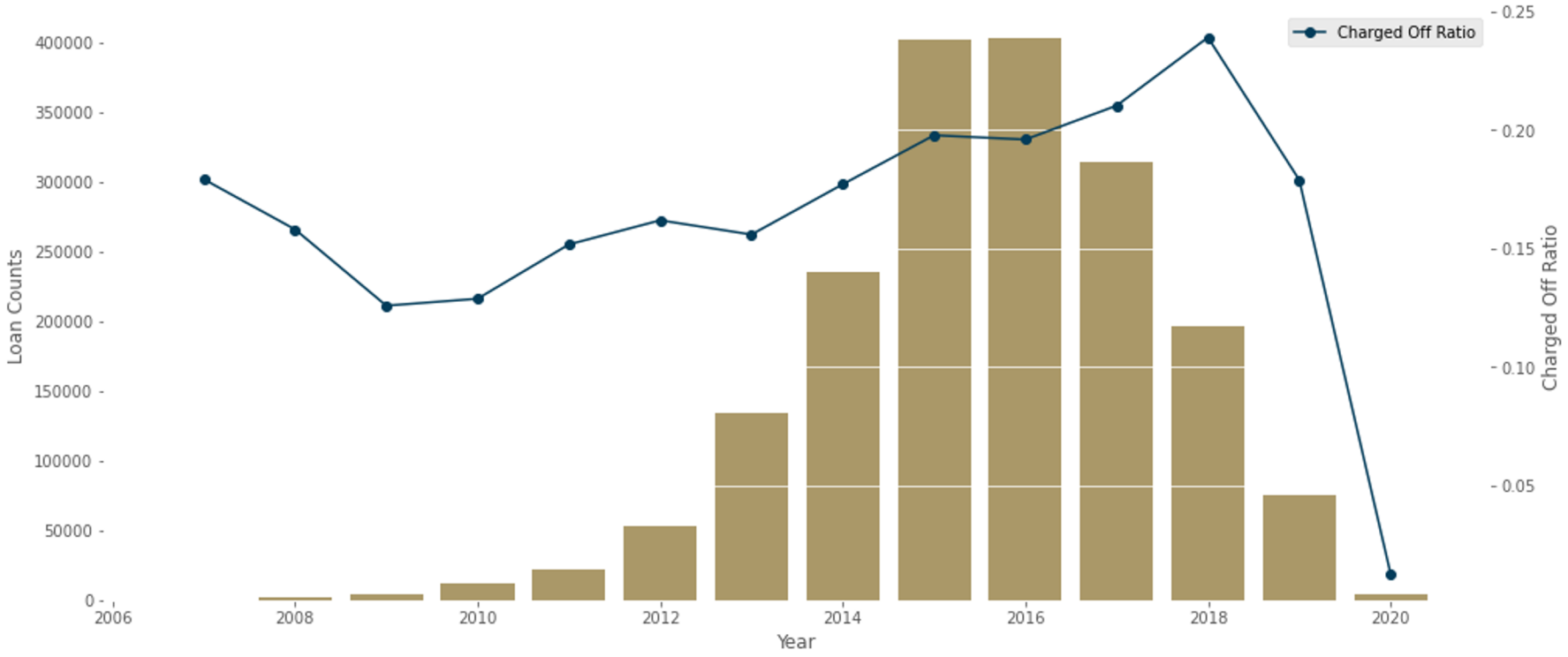
- ❑ Select 10000 rows
- ❑ Set dependent variable
- ❑ Filter loan status
- ❑ Check correlation
- ❑ Features would be available before funding the loan

Dependent variable: 'loan_status'
Features:

1) 'loan_amnt'
2) 'term'
3) 'int_rate'
4) 'sub_grade'
5) 'home_ownership'
6) 'annual_inc'
7) 'verification_status'
8) 'purpose'
9) 'addr_state'
10) 'dti'
11) 'open_acc'
12) 'pub_rec'
13) 'revol_bal'
14) 'revol_util'
15) 'initial_list_status'
16) 'application_type'
17) 'mort_acc'
18) 'pub_rec_bankruptcies'
19) 'loan_status_flag'
20) 'fico'
21) 'earliest_cr_line_y'

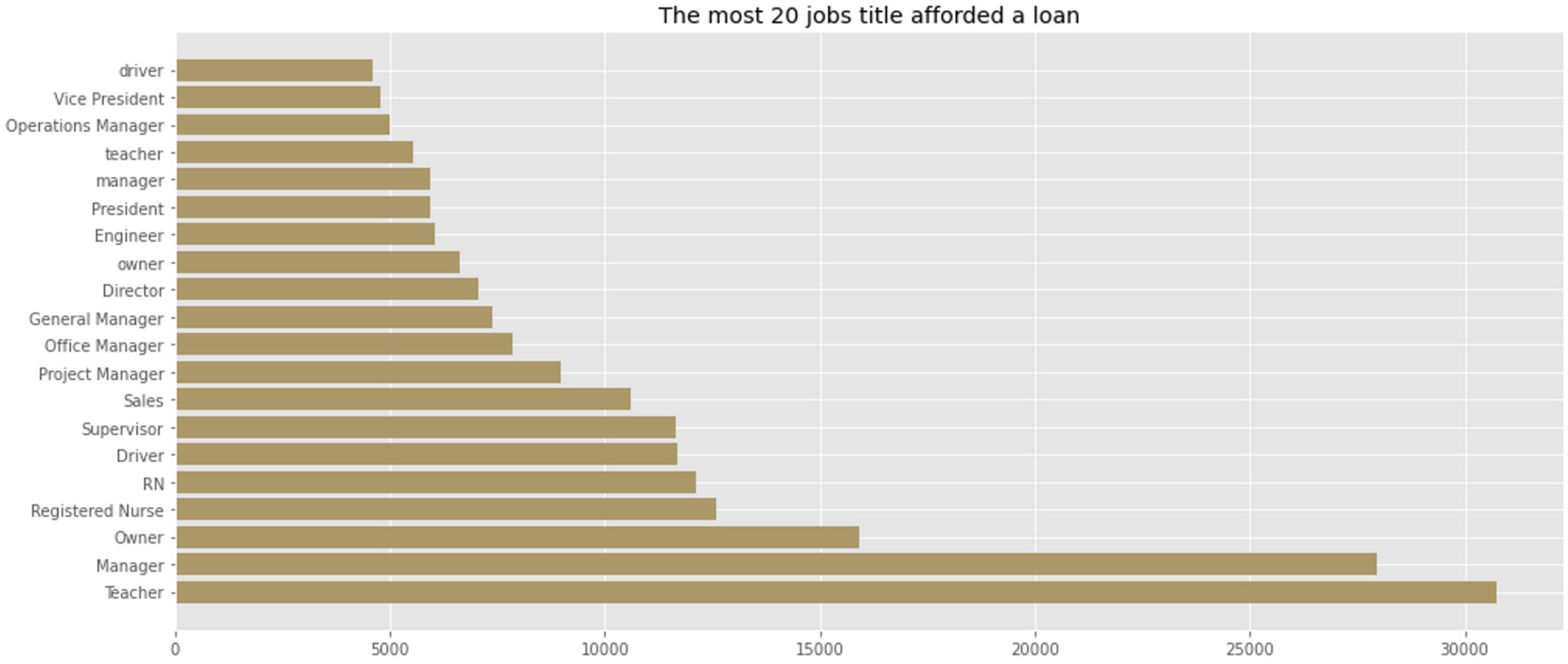
EXPLORATORY DATA ANALYSIS

- **Dependent Variable – Loan Status**



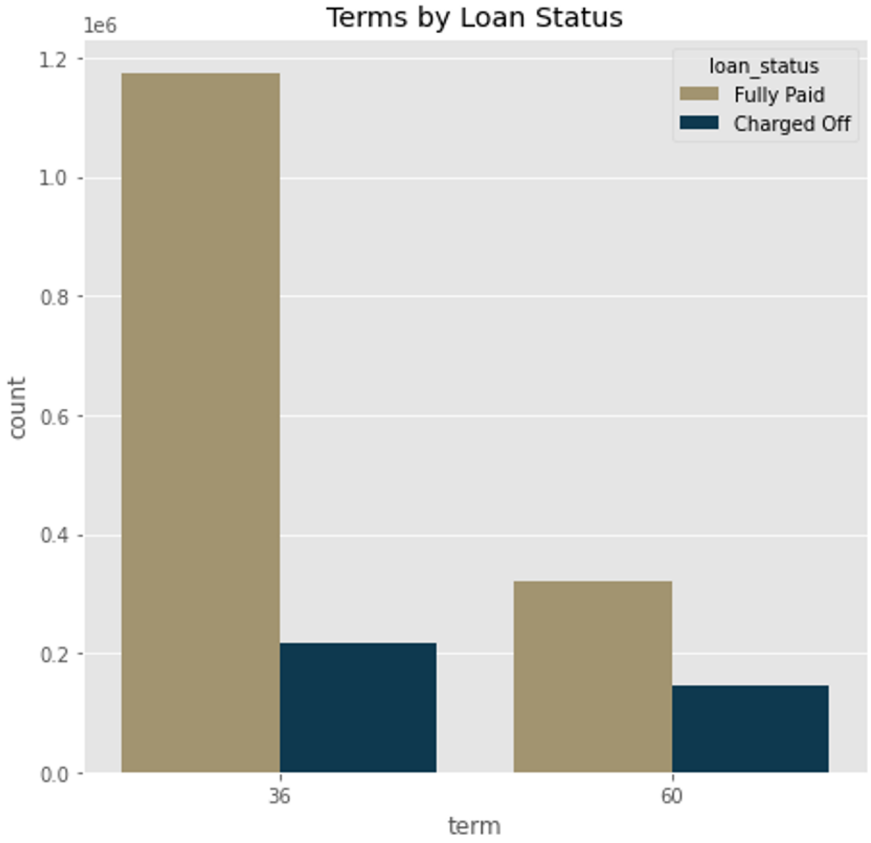
EXPLORATORY DATA ANALYSIS

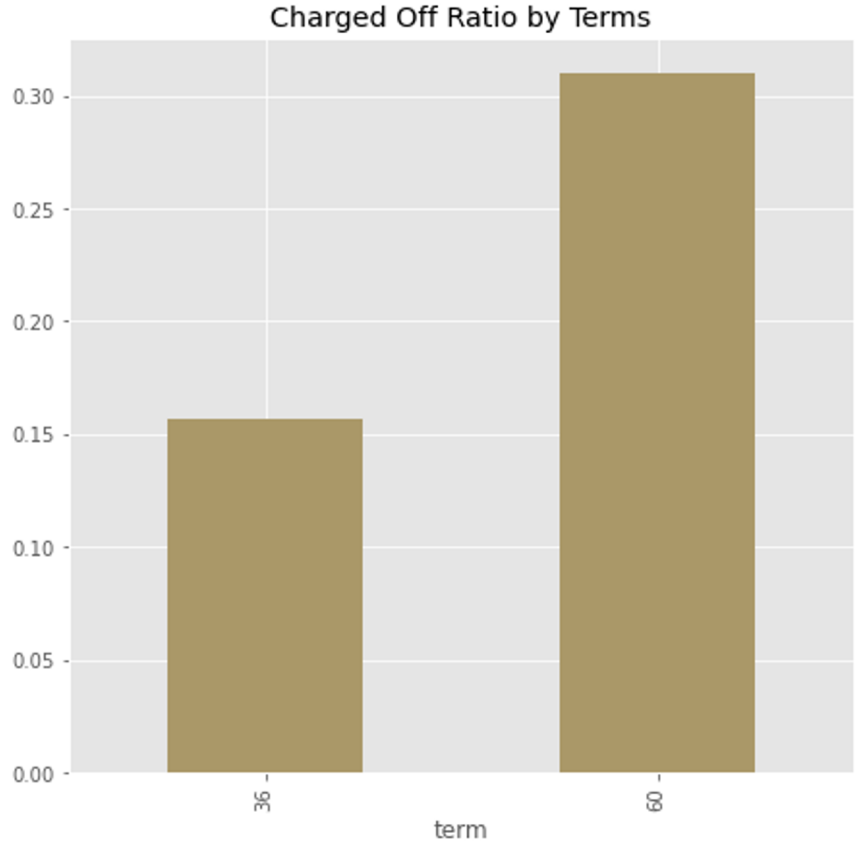
❑ Predictors – Job Title



EXPLORATORY DATA ANALYSIS

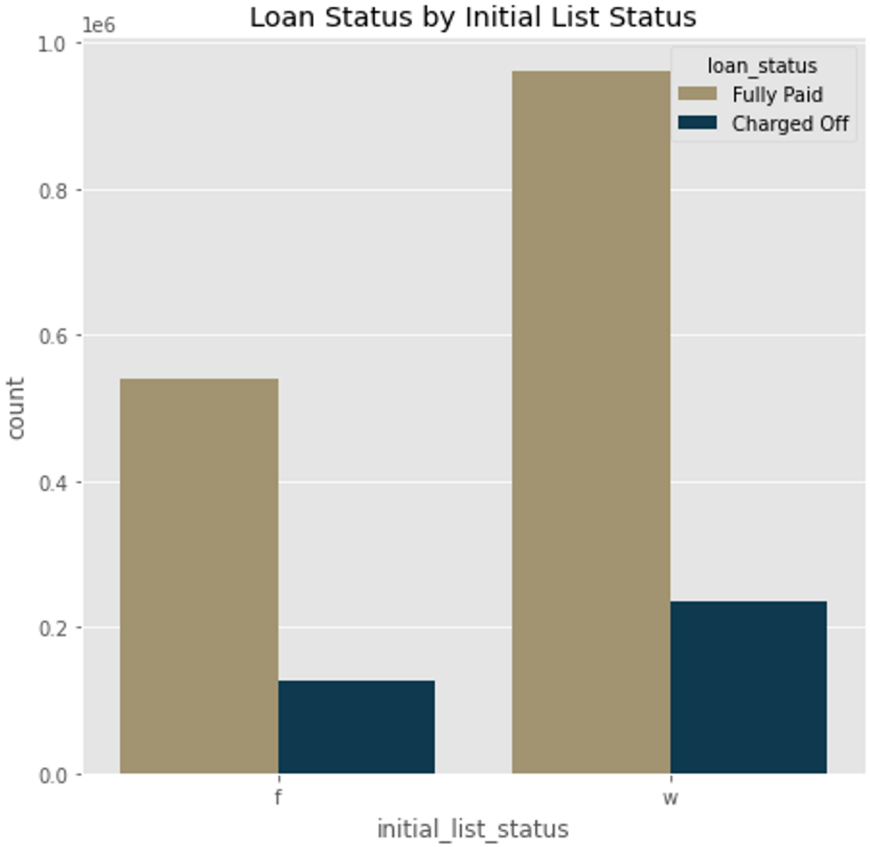
❑ Predictors – Loan Term

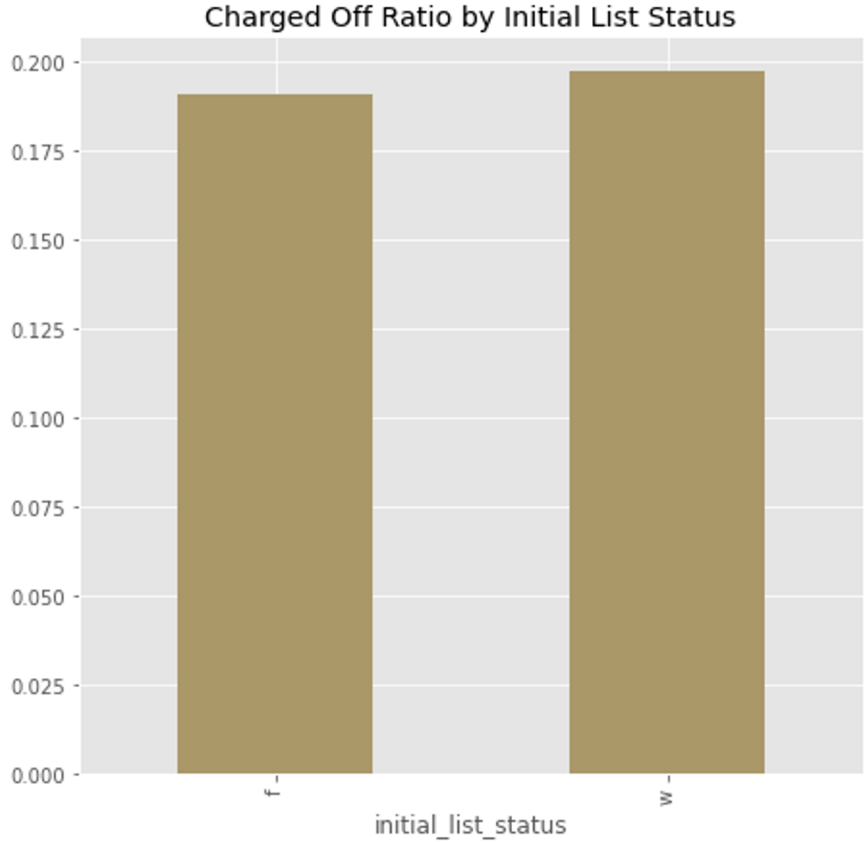




EXPLORATORY DATA ANALYSIS

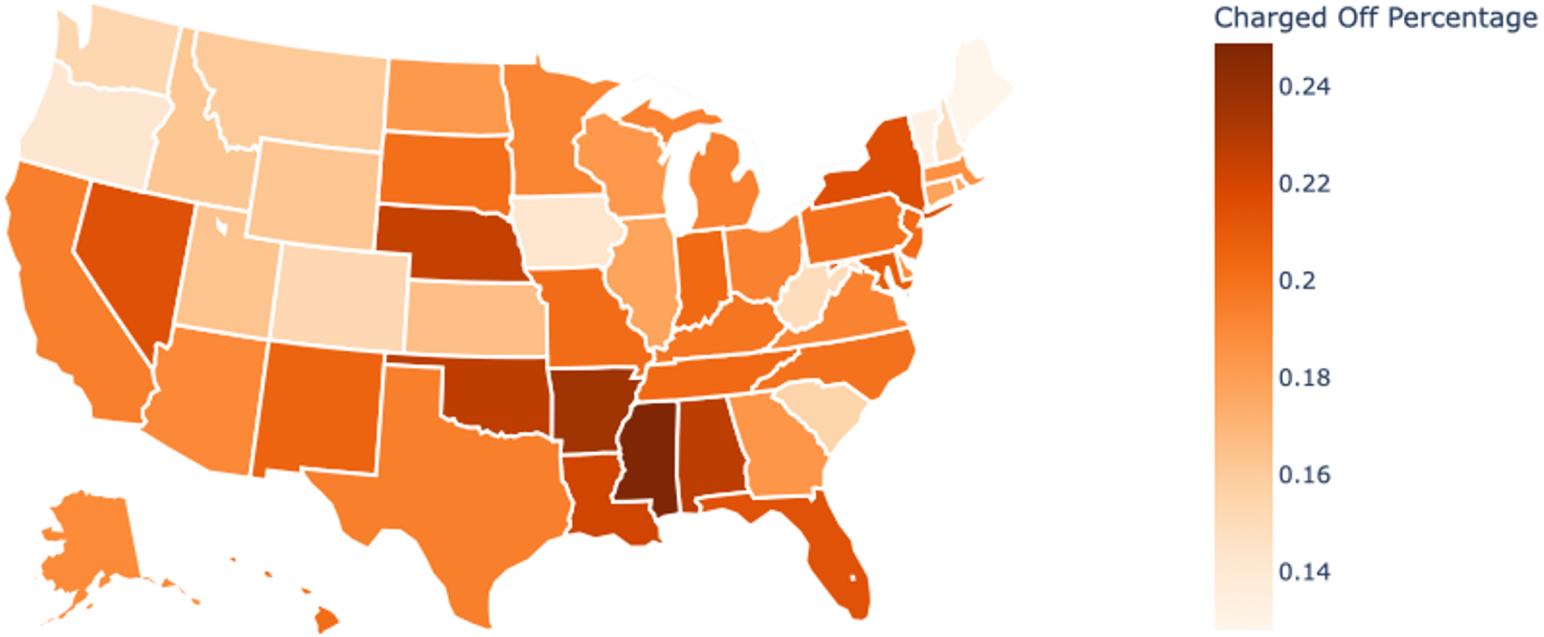
❑ Predictors – Application Type





EXPLORATORY DATA ANALYSIS

- **Predictors – State**



DATA PREPROCESSING

- **Remove unnecessary features**
- **Missing Value**

mort_acc	0.0254
revol_util	0.0008
dti	0.0006
pub_rec_bankruptcies	0.0004

Impute

Drop

- **One-Hot Encoding**

MODEL

❑ Train Test Split

Train Data	80%
Test Data	20%

❑ Downsampling

Fully Paid	80%
Charged Off	20%

Fully Paid	50%
Charged Off	50%

MODEL

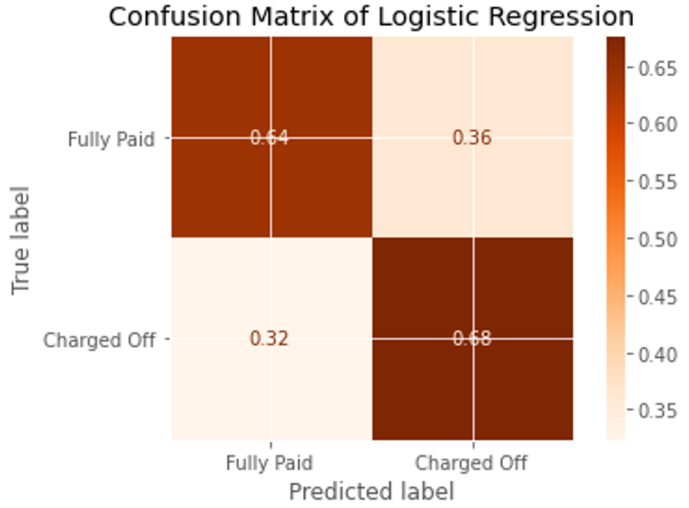
- Model Selection
  - Logistic Regression
  - Random Forest
  - XGBoost
- Hyperparameter Tuning
  - Randomized Grid Search

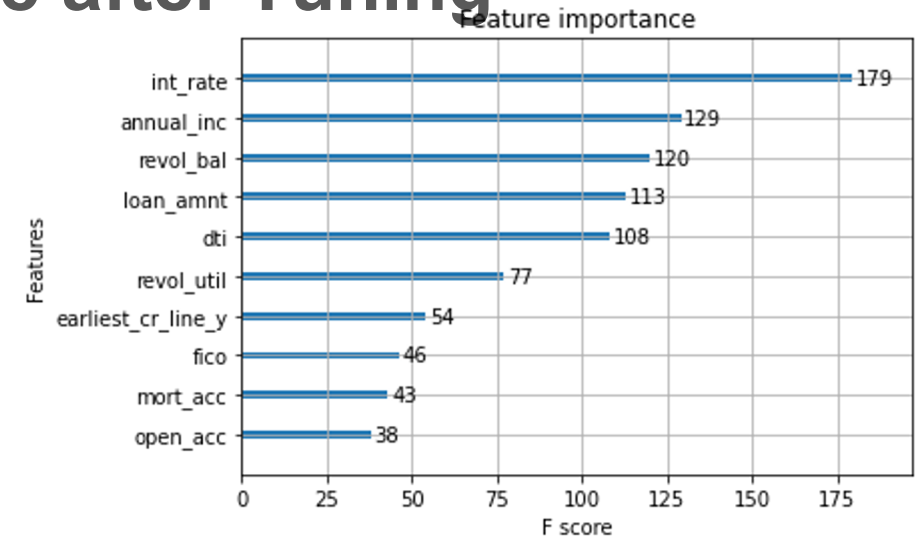
Summary of model performance on the test set

	Logistic Regression	Random Forest	XGBoost
Precision	0.65	0.65	0.65
Recall	0.53	0.66	0.68
F1	0.58	0.65	0.66
AUC_ROC	0.68	0.71	0.72
Accuracy	0.62	0.65	0.66

RESULT

❑ Model Performance after Tuning





❑ Web Application

❑ Streamlit Python Library

CONCLUSION

- We built a predictive model to predict the loan status at the closing time.
- The performance of the model is not very good.
- We can choose a high threshold and use this model as a preventive tool for the investors to avoid risky loans.

Thank You
&
Questions?

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC