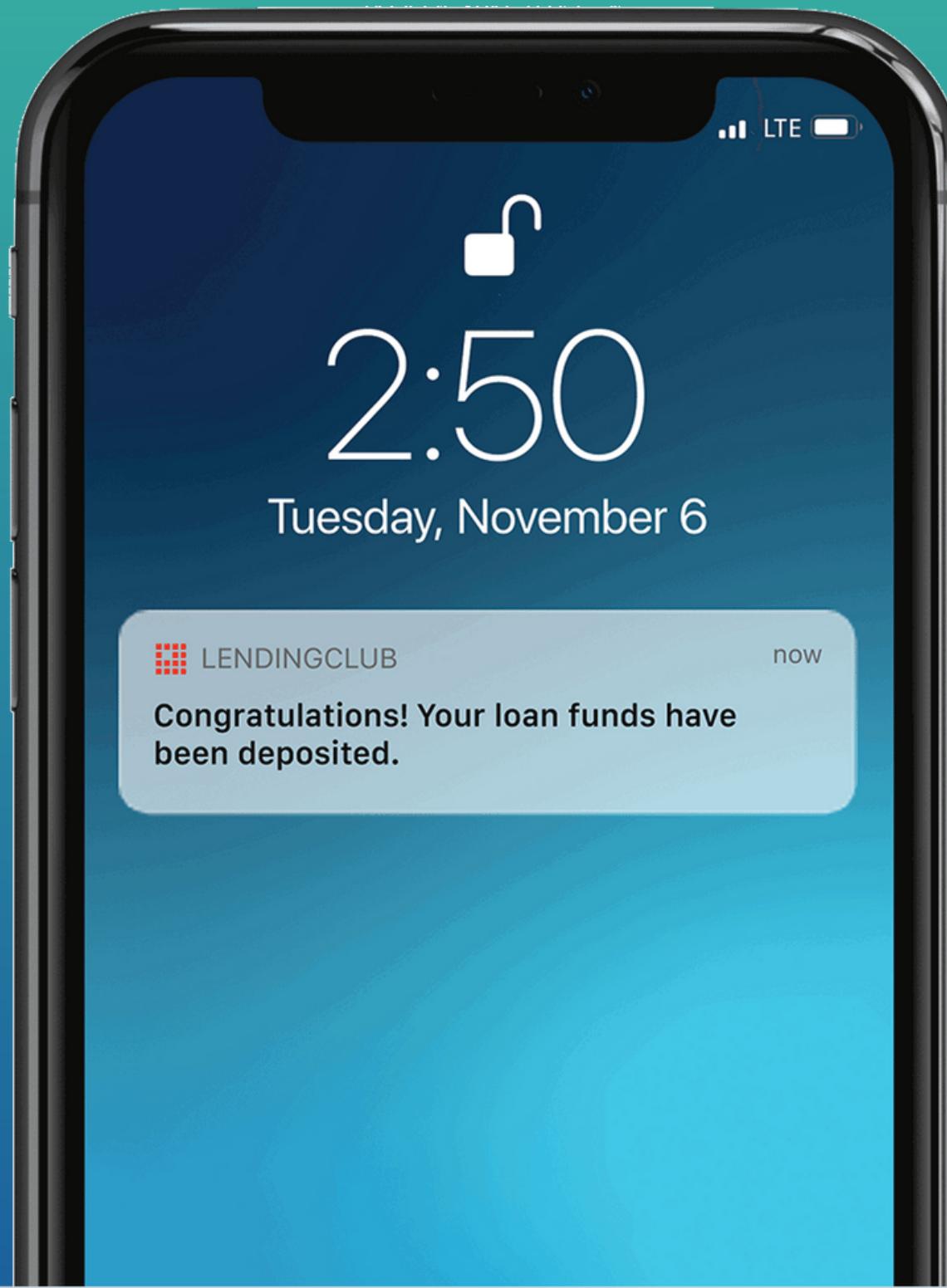


Predicting loan status for risk management in P2P lending market

Thu Ha MAI
Aouatef HAOUFADI
Maria Yamina TEHAR



Table of contents



01

Introduction

02

Material and Methods

- Dataset and Feature understanding
 - Data Preprocessing
 - Methodology
-

03

Results

- Exploratory Data Analysis
- Information Reduction models
- Prediction models

P1. Introduction

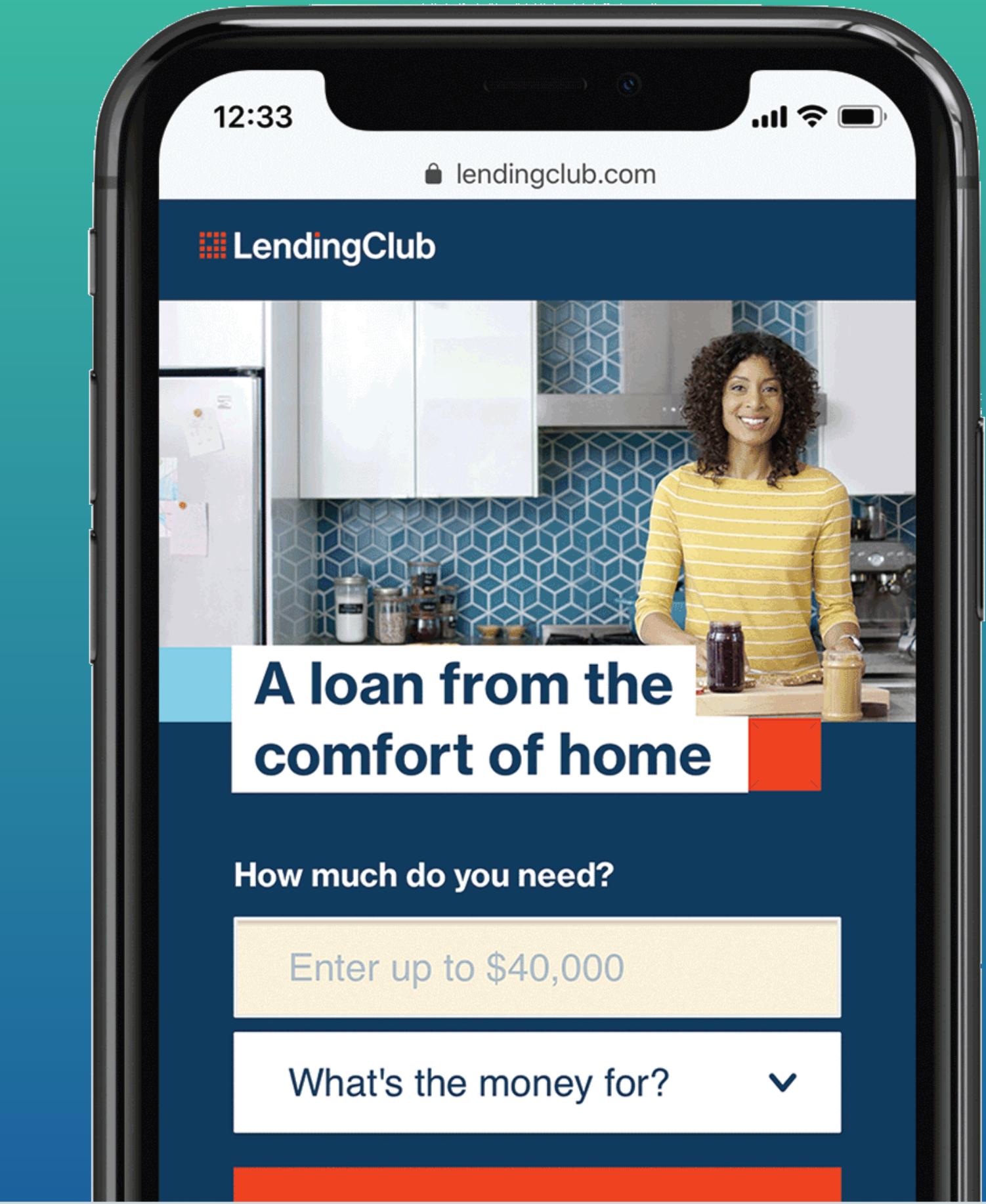
Introduction

01 Context

- Lending Club is a peer-to-peer lending company
- Headquartered in San Francisco, California

02 Objective

- Predict the loan status based on the applicants' profiles
- Understand the driving factors behind loan status



Peer-to-peer Lending

Economics Context



Peer-to-peer lending is pushing the sector of finance toward greater credit accessibility



Making credit available for small businesses means economic growth, support for local entrepreneurs, and potentially more job opportunities

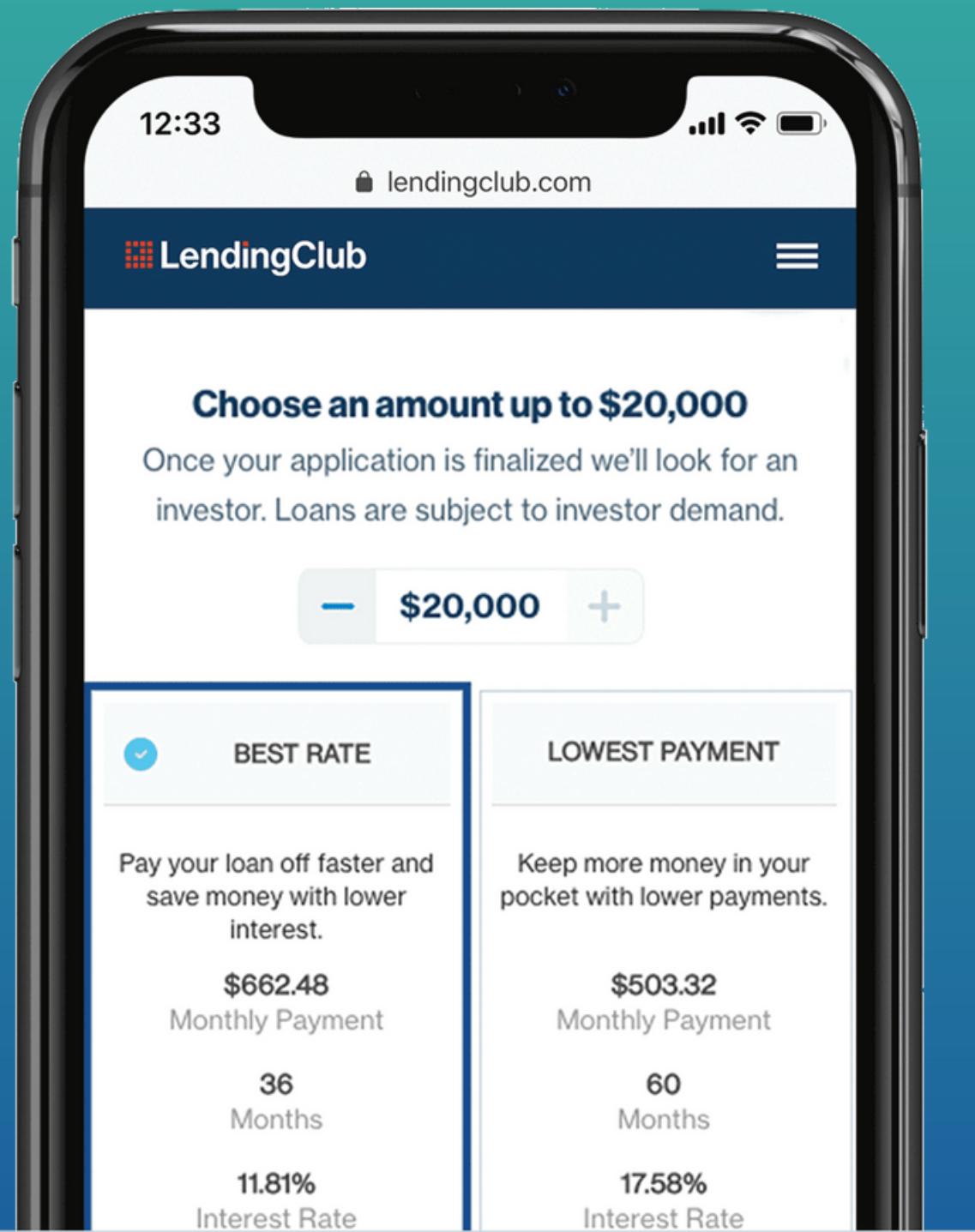


Peer-to-peer lending is quite risky for investors since there is credit risk, liquidity risk and even platform collapse risk

P2. Material & Methods

Material & Methods

Dataset and Feature Understanding



01

The dataset

- Official website of Lending Club
- Source of the dataset: Kaggle
- 396,030 observations and 27 features

02

The loan status is the target variable

- Fully-paid
- Charged-off

03

The variables can be divided into two segments

- Features related to the borrowers
- Features related to the loan characteristics

Material & Methods

Data Preprocessing

Detection and Treatment of duplicates and missing values

- No duplicate rows in our data
- Existence of missing values in 06 columns
 - Remove 'emp_title', 'emp_length', 'title'
 - Remove missing values in 'revol_util', 'pub_rec_bankruptcies'
 - Special treatment to 'mort_acc'

Feature Engineering

- Drop the 'grade' and 'issue_d' columns
- Extract the 'zip_code' column from the 'address' column
- Convert our categorical variables to numerical ones

Material & Methods

Data Preprocessing

Detection and treatment of outliers

- Only on the training set to avoid leaking information from the test set
- **Inter Quartile Range (IQR)** approach
- **Trimming method** since the outliers account for only 0.5 to 1 % of the training set

Dealing with imbalanced data

- Fully paid: 80%
- Charged-off: 20%
- Synthetic Minority Over-sampling Technique (SMOTE)

Material & Methods

Methodology

Information Reduction

Elastic Net
Autometrics

Prediction

Logistic Regression
Random Forest

Empirical Strategy

Model Selection
Cross Validation
GridSearch CV
SMOTE-NC

Performance Evaluation

Accuracy
Precision
Recall
F1 score
AUC
Specificity

P3. Results

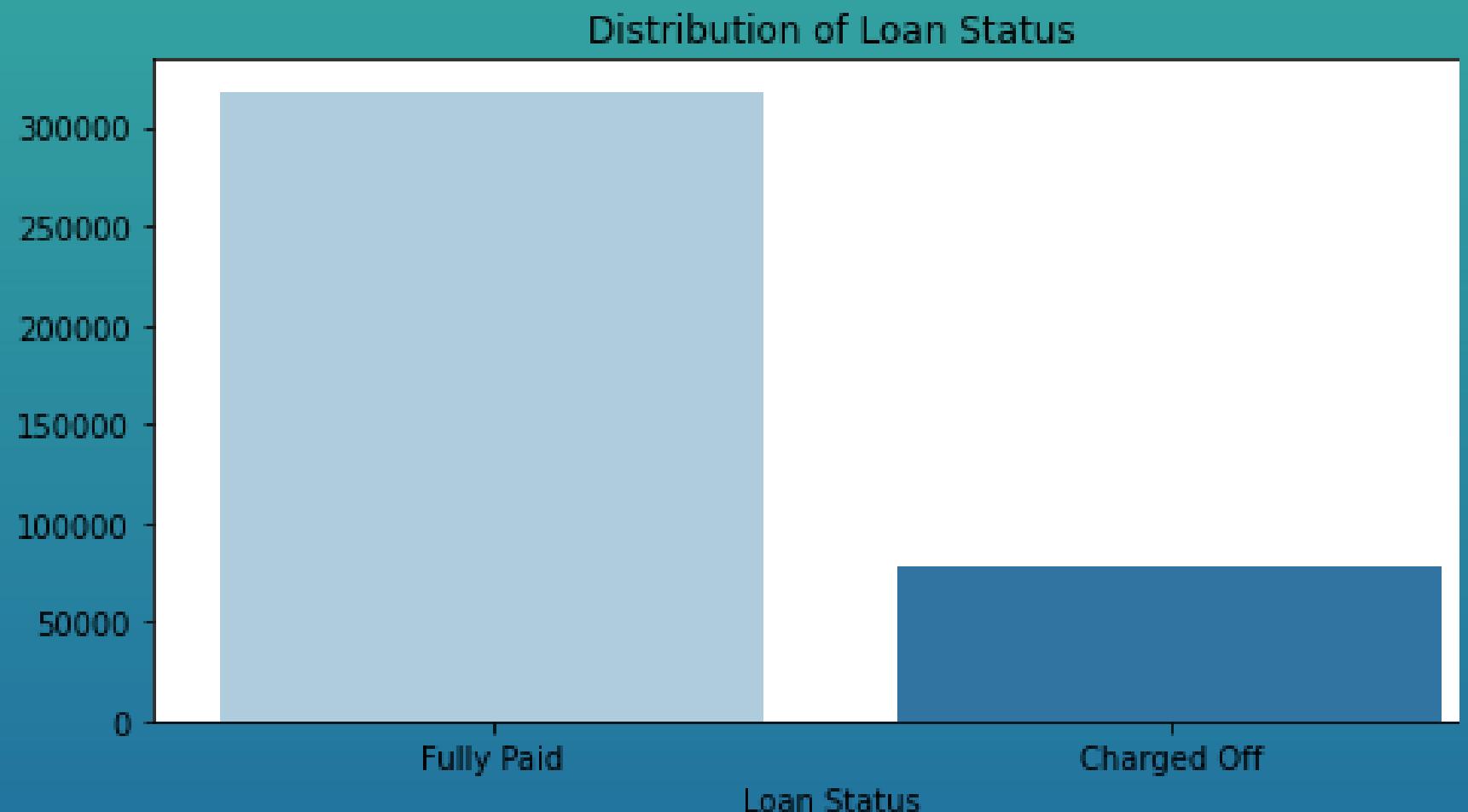
P3.1. Exploratory Data Analysis

Exploratory Data Analysis

Univariate analysis

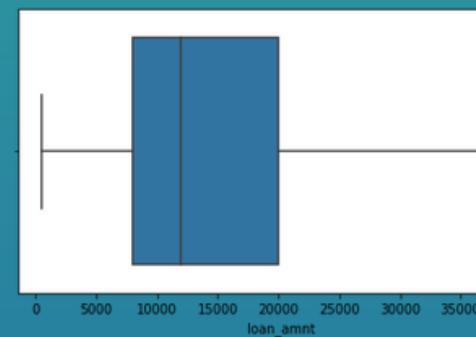
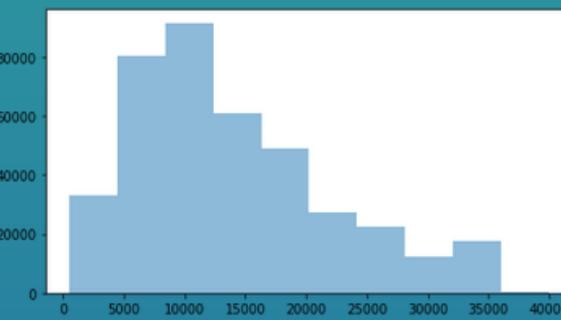
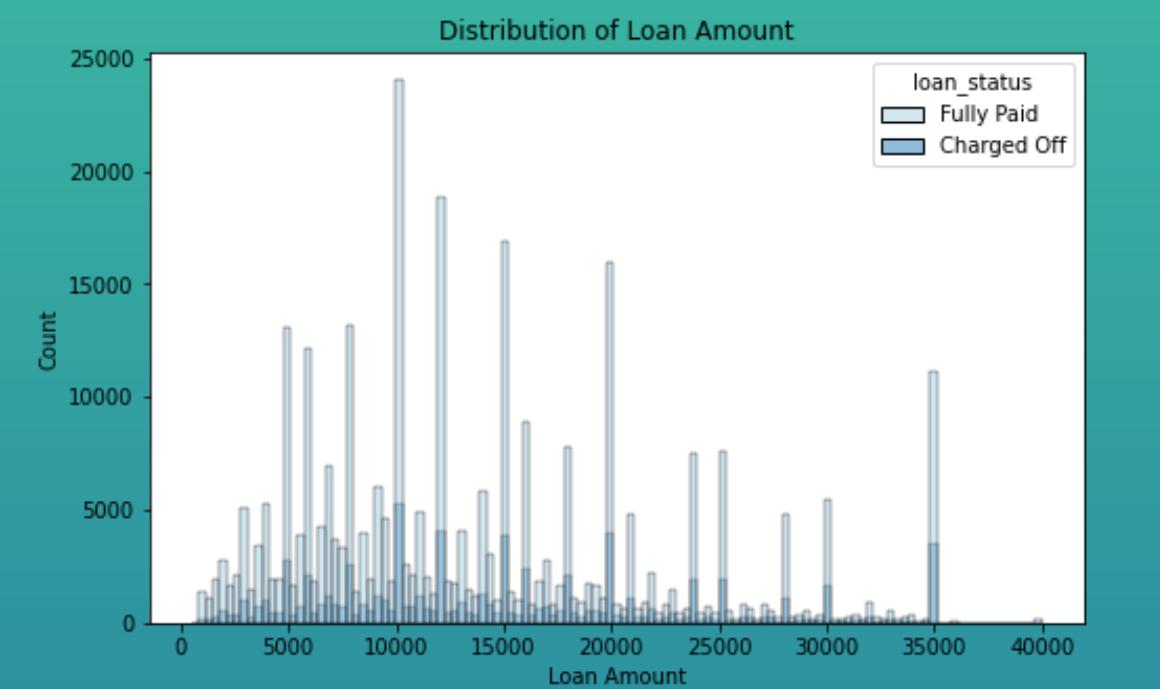
Loan status

- The dependent variable for our analysis is a categorical variable.
- Takes only two values : " Fully paid" or "Charged Off"
- Imbalanced data
 - Fully paid: 80% # observation
 - Charged off: 20% # observation



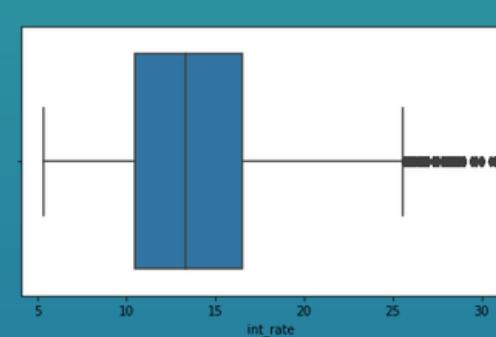
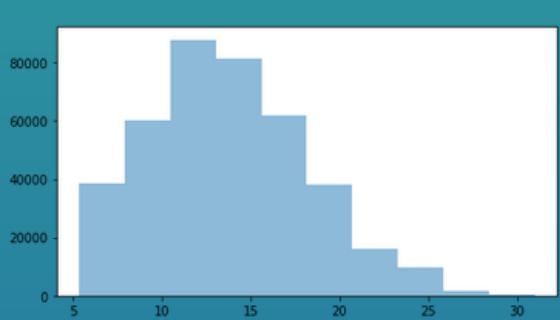
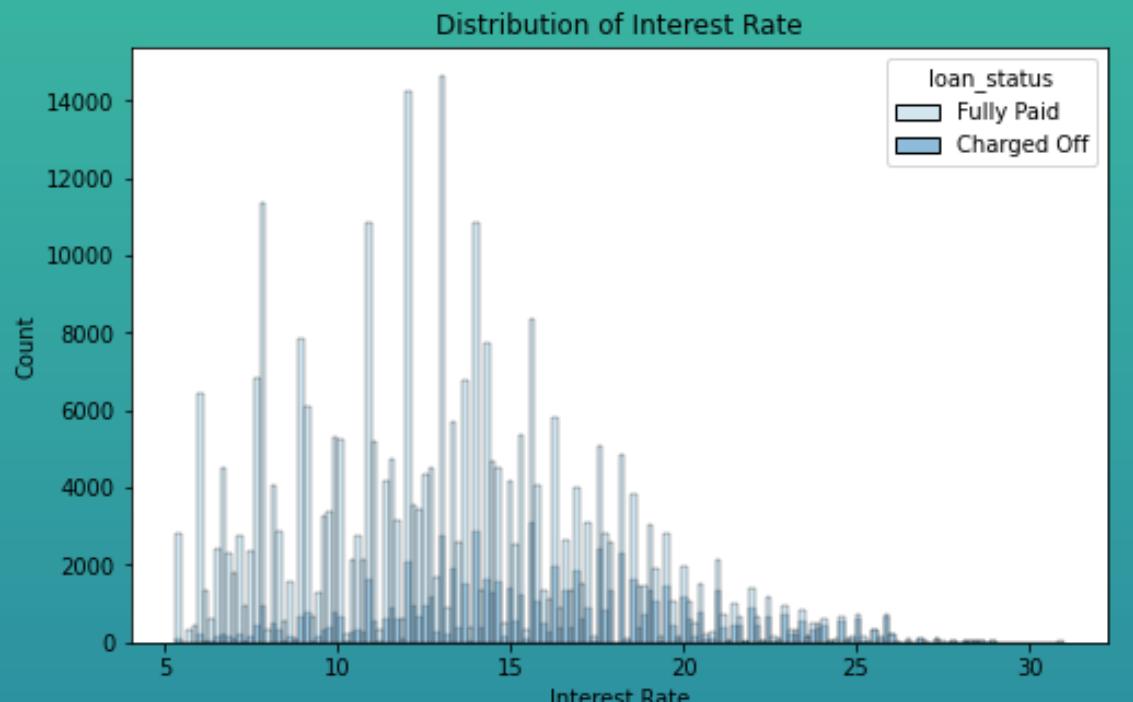
Exploratory Data Analysis

Multivariate analysis : Loan characteristics feature analysis



Loan amount

- Loan amounts is ranging from \$5000 to \$20000
- Normally distributed, slightly right skewed
- No outliers

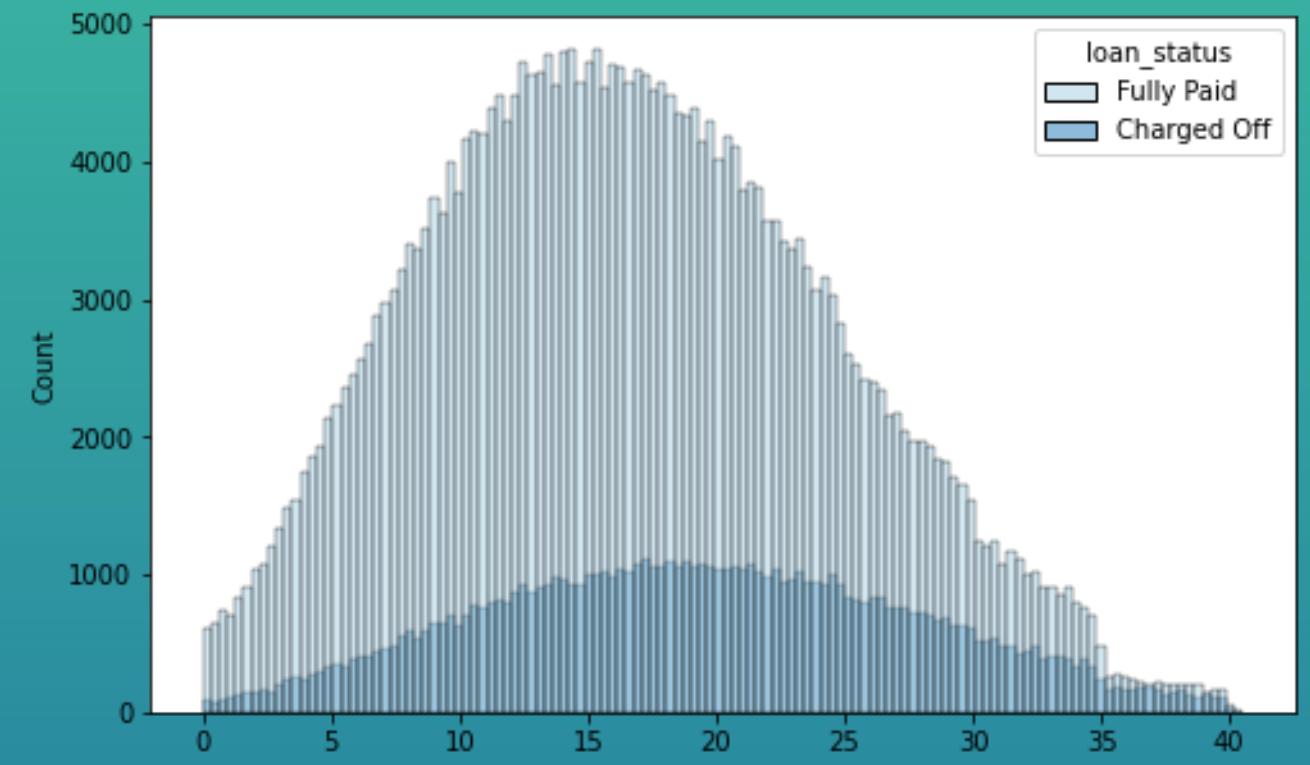


Interest Rate

- Interest rate is ranging from 10% to 15%
- Normally distributed, slightly skewed
- Few outliers => poor credit ratings

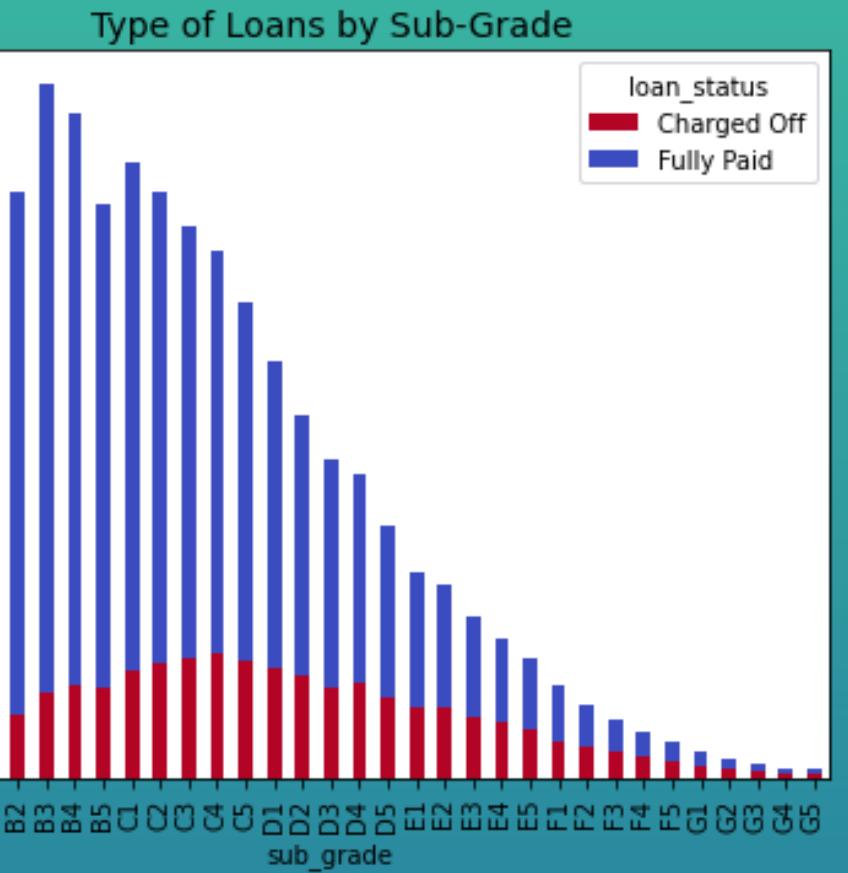
Exploratory Data Analysis

Multivariate analysis : Loan characteristics feature analysis



DTI - Debt to Income ratio

- Normally distributed
- Ranging from 0 to 40
- (bound is zero = no debt)

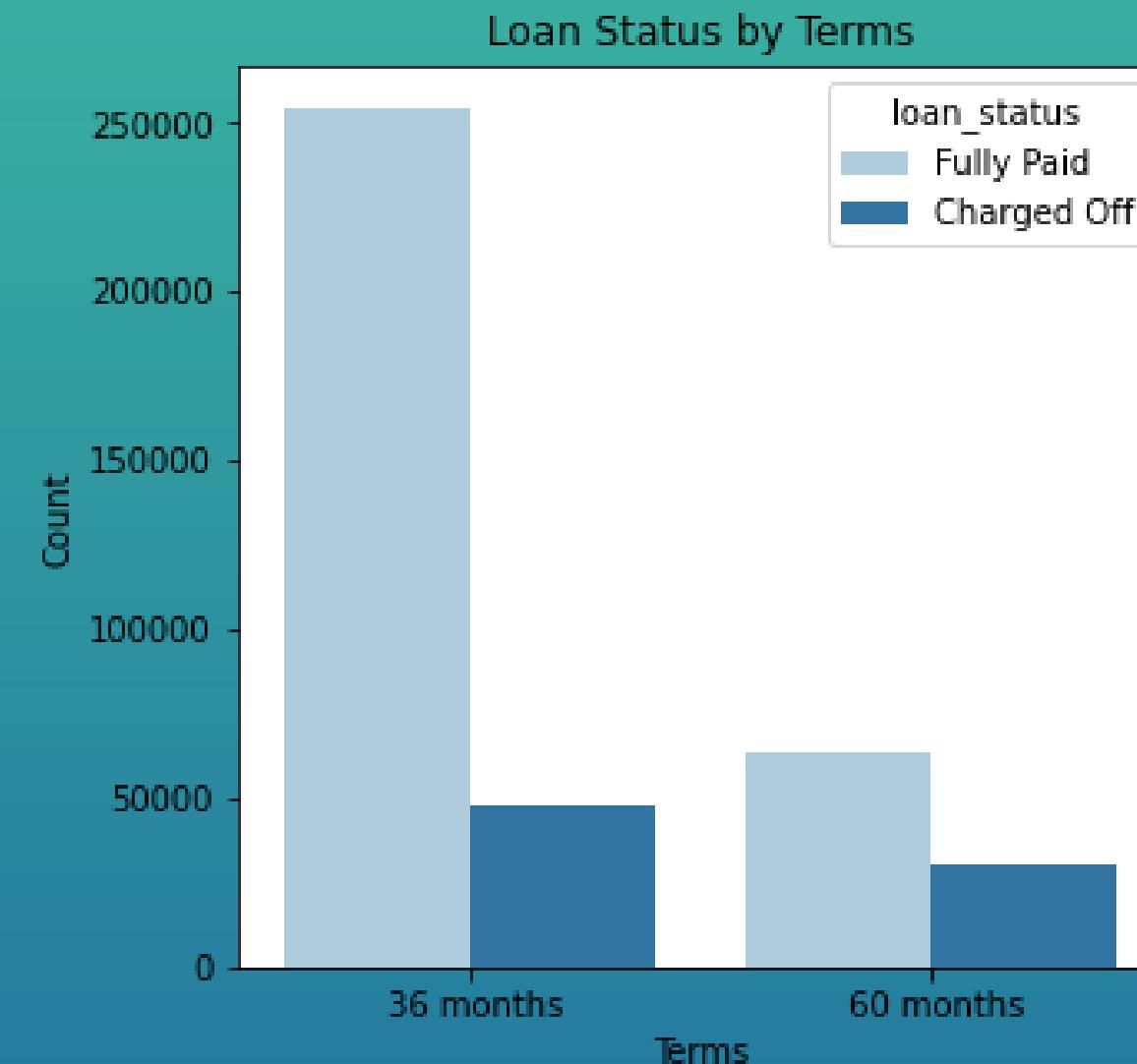


Sub-grade

- A is the highest grade => lowest interest rate
- Normally distributed with a heavy right tail
- The trend of higher fraction of loans charged-off

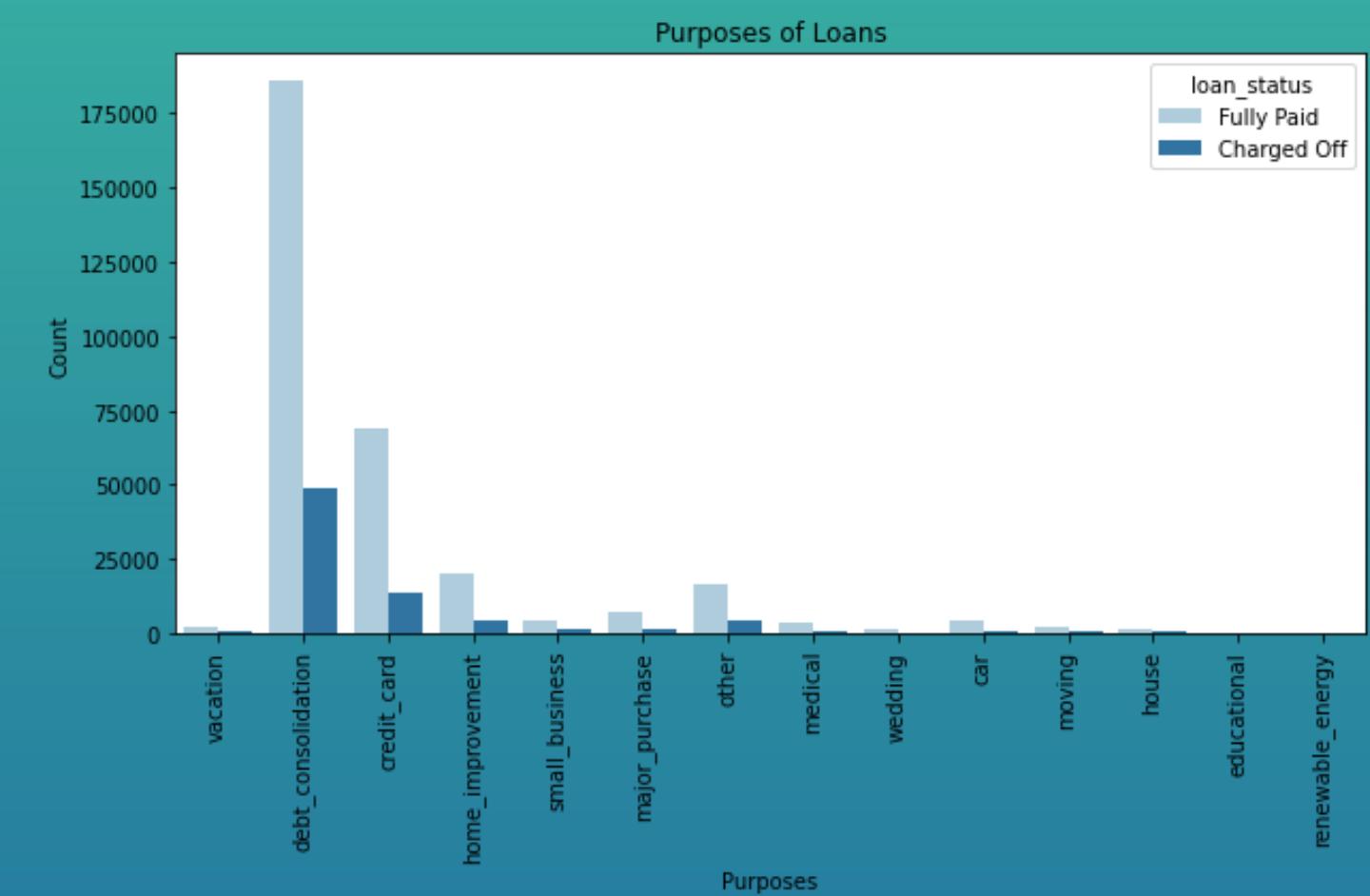
Exploratory Data Analysis

Multivariate analysis : Loan characteristics feature analysis



Loan status by terms

- 02 Categories: 36 months and 60 months
- The lower fraction of loans are fully paid within 60 months

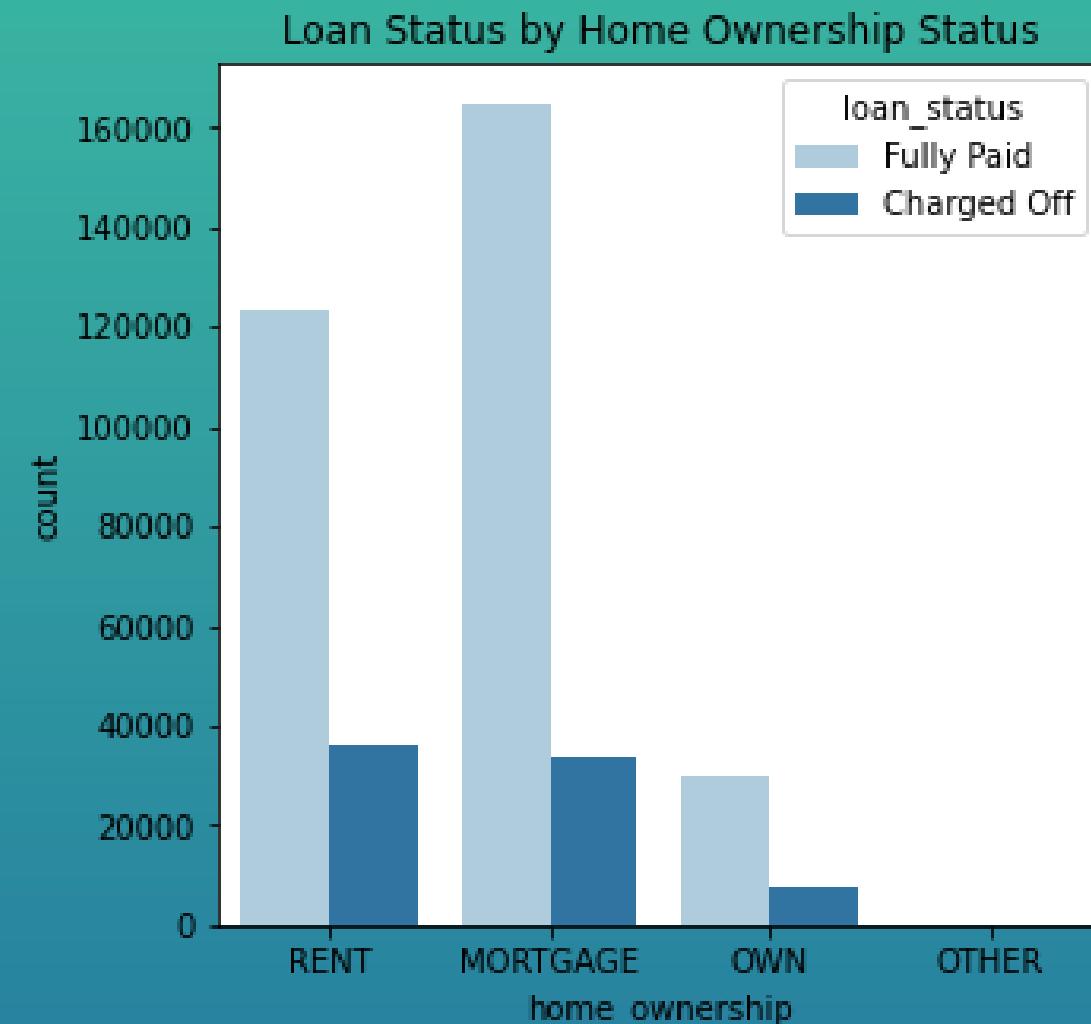


Loans purposes

- The highest category: Debt consolidation
- Debt consolidation has less probability of being charged-off

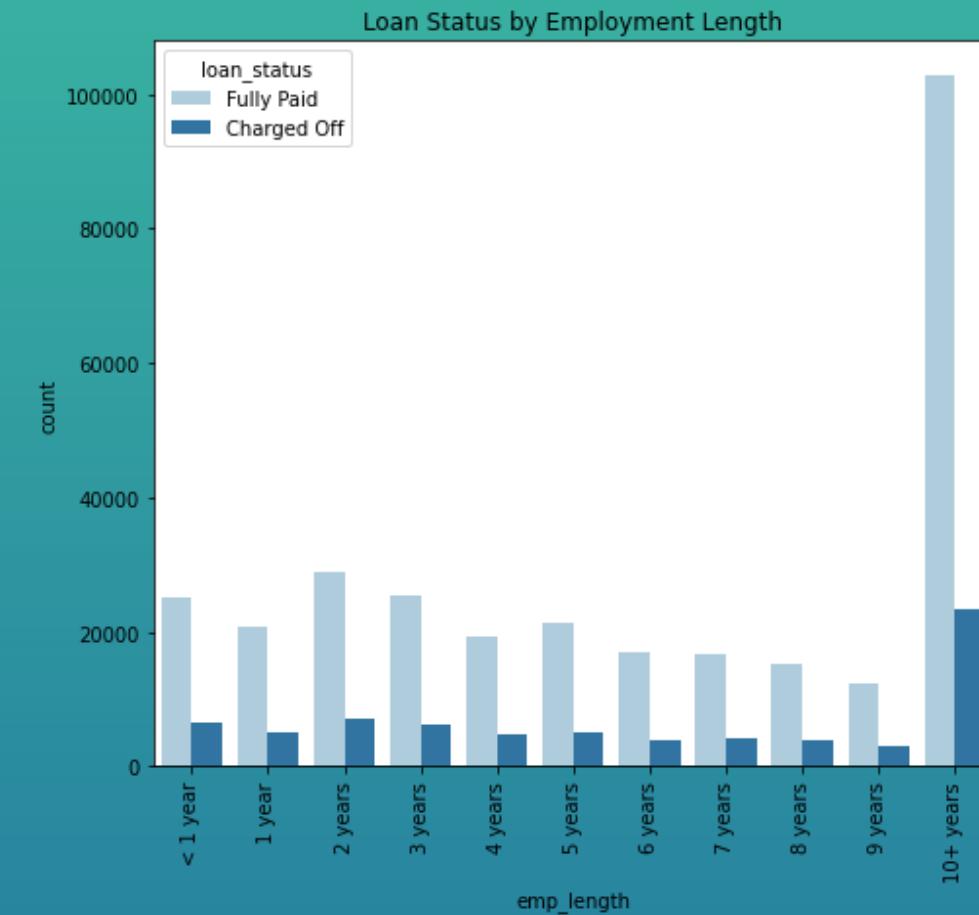
Exploratory Data Analysis

Multivariate analysis : Applicants' demographic feature analysis



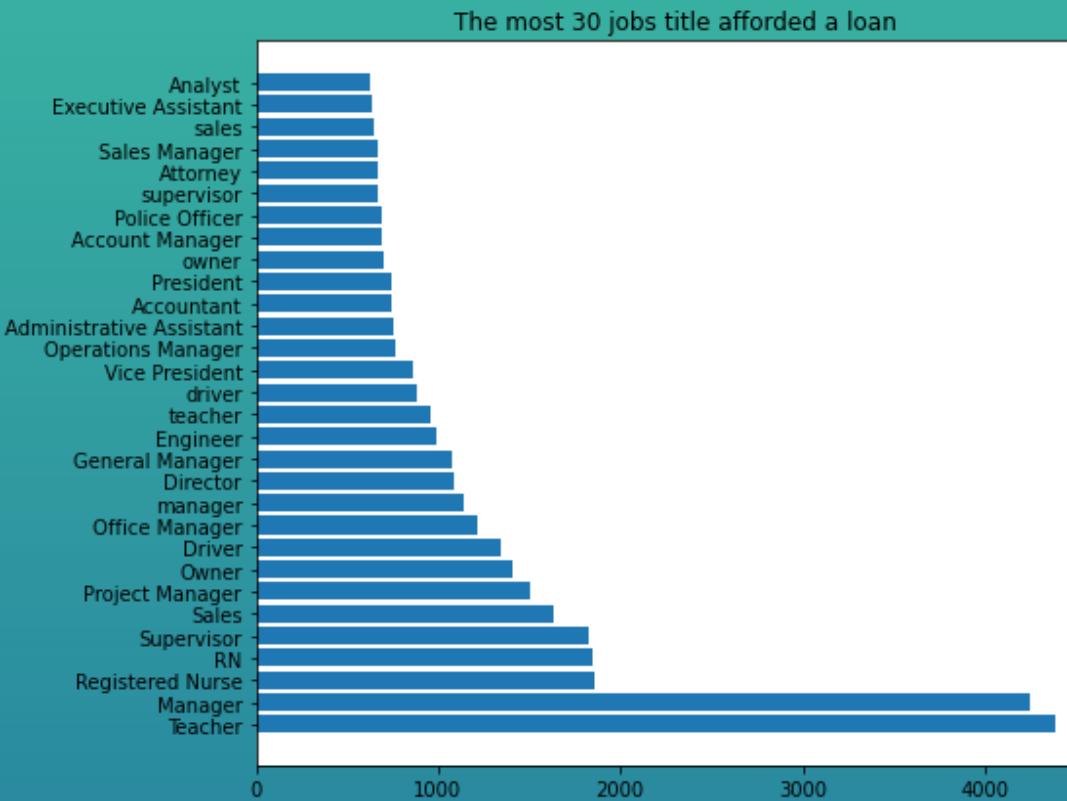
Home ownership

- Small difference in charged-off rates
- Mortgage homes:
 - The most popular category
 - Less probability of being charged-off



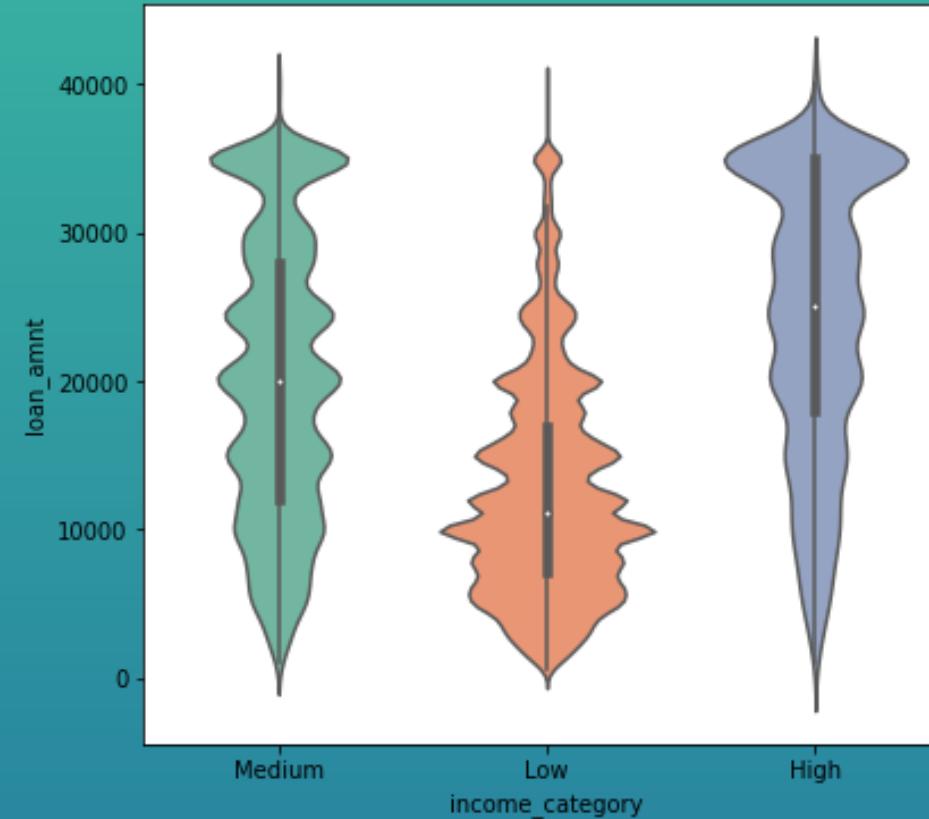
Employment title & Employment length

- More than 10 years of job experience:
 - Highest category
 - Higher probability of being fully-paid



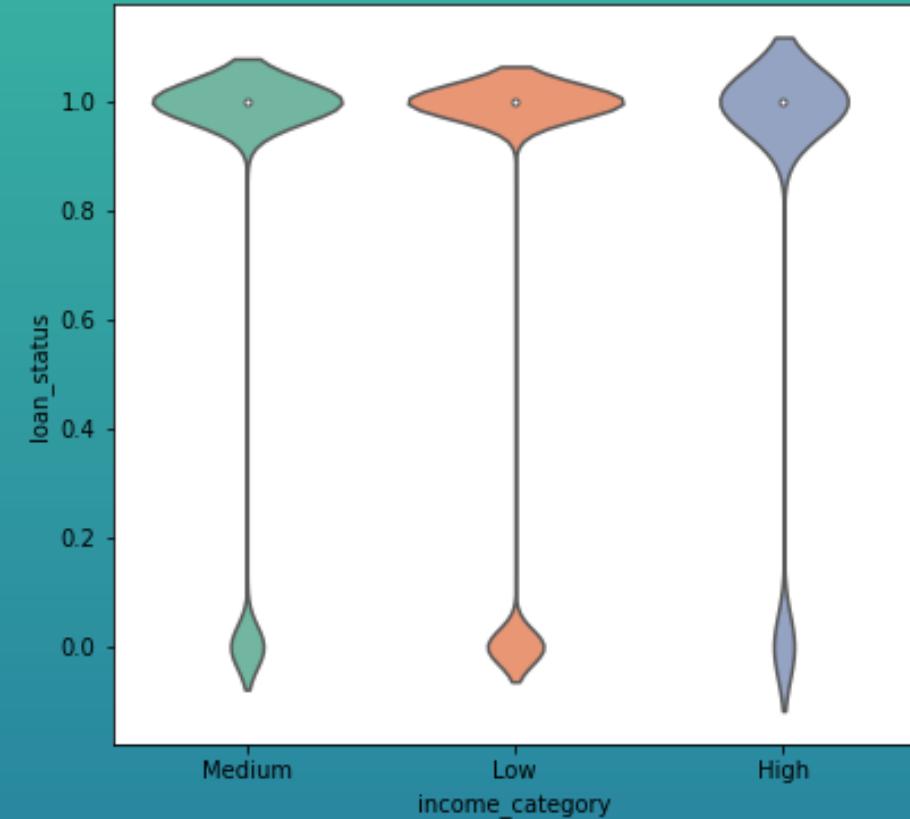
Exploratory Data Analysis

Multivariate analysis: Applicants' demographic feature analysis



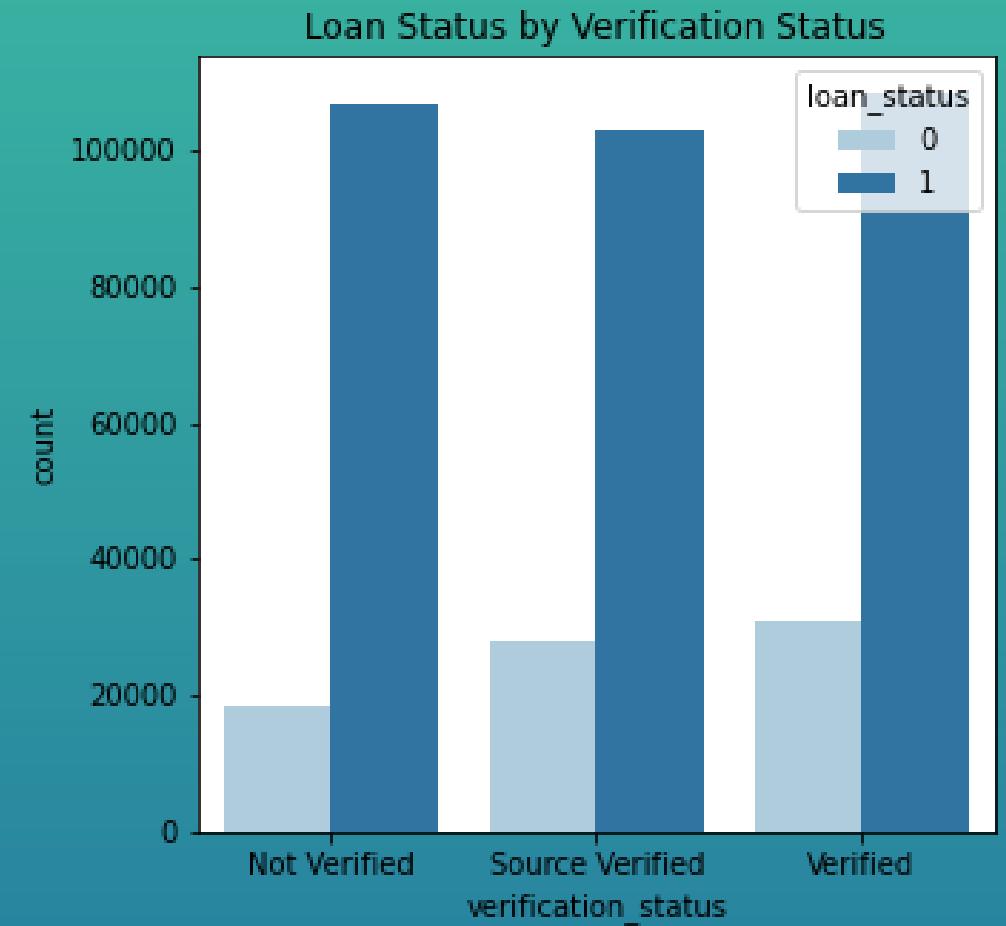
Income category

- Higher-income category => Higher loan amount
- Low-income category => Higher chance of being charged-off



Verification status

- 3 categories
- Verified loans => higher chance of charged-off (would guess the opposite)



Correlation analysis 1/2

Correlation between numerical features

Heat map of Pearson correlation

A strong correlation between:

- loan amount and instalment with 95%
- pub_rec and pub_rec_bankruptcies with 87%
- total-acc and open-acc with 68%

Variance Inflation Factor

- High VIF for loan amount, installment, year
- Existence of multicollinearity

	loan_amnt	term	int_rate	installment	annual_inc	loan_status	dti	open_acc	pub_rec	revol_bal	revol_util	total_acc	mort_acc	in bankruptcies	year
loan_amnt	1	0.39	0.17	0.95	0.34	-0.06	0.016	0.2	-0.1	0.33	0.1	0.22	0.21	-0.11	-0.15
term	0.39	1	0.43	0.15	0.06	-0.17	0.036	0.079	-0.02	0.085	0.055	0.1	0.11	-0.021	-0.029
int_rate	0.17	0.43	1	0.16	-0.057	-0.25	0.079	0.012	0.069	-0.011	0.29	-0.037	-0.09	0.058	0.11
installment	0.95	0.15	0.16	1	0.33	-0.041	0.016	0.19	-0.092	0.32	0.12	0.2	0.18	-0.1	-0.13
annual_inc	0.34	0.06	-0.057	0.33	1	0.054	-0.082	0.14	-0.033	0.3	0.028	0.19	0.2	-0.053	-0.14
loan_status	-0.06	-0.17	-0.25	-0.041	0.054	1	-0.062	-0.028	-0.018	0.011	-0.082	0.018	0.069	-0.0083	-0.039
dti	-0.016	0.036	0.079	0.016	-0.082	-0.062	1	0.14	-0.018	0.063	0.088	0.1	-0.0087	-0.014	-0.0097
open_acc	0.2	0.079	0.012	0.19	0.14	-0.028	0.14	1	-0.022	0.22	-0.13	0.68	0.14	-0.029	-0.12
pub_rec	-0.1	-0.02	0.069	-0.092	-0.033	-0.018	-0.018	-0.022	1	-0.13	-0.09	0.034	0.042	0.87	-0.064
revol_bal	0.33	0.085	-0.011	0.32	0.3	0.011	0.063	0.22	-0.13	1	0.23	0.19	0.17	-0.13	-0.2
revol_util	0.1	0.055	0.29	0.12	0.028	-0.082	0.088	-0.13	-0.09	0.23	1	-0.1	-0.0017	-0.086	-0.008
total_acc	0.22	0.1	-0.037	0.2	0.19	0.018	0.1	0.68	0.034	0.19	-0.1	1	0.34	0.042	-0.28
mort_acc	0.21	0.11	-0.09	0.18	0.2	0.069	-0.0087	0.14	0.042	0.17	-0.0017	0.34	1	0.048	-0.27
in bankruptcies	-0.11	-0.021	0.058	-0.1	-0.053	-0.0083	-0.014	-0.029	0.87	-0.13	-0.086	0.042	0.048	1	-0.061
year	-0.15	-0.029	0.11	-0.13	-0.14	-0.039	-0.0097	-0.12	-0.064	-0.2	-0.008	-0.28	-0.27	-0.061	1

Correlation analysis 2/2

Correlation between categorical features

Heat map of the p-values for the Chi-Square test

- High correlation between most of the categorical features because most of the p-values < 0.05

		Chi-Square Test Results		
		A2	A3	A4
sub_grade_A2	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_A4	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_B1	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_B3	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_B5	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_C2	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_C4	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_D1	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_D3	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_D5	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_E2	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_E4	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_F1	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_F3	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_F5	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_G2	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
sub_grade_G4	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
home_ownership_OTHER	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
home_ownership_RENT	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
verification_status_Verified	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
purpose_debt_consolidation	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
purpose_home_improvement	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
purpose_major_purchase	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
purpose_moving	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
purpose_renewable_energy	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
purpose_vacation	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
initial_list_status_w	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
application_type_JOINT	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
income_category_Medium	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
zip_code_11650	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
zip_code_29597	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
zip_code_48052	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000
zip_code_86630	-0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000

P3.2. Information Reduction Models

Information Reduction Models 1/2

Before SMOTE-NC

Both ElasticNet and Autometrics models:

- Perform pretty well in classifying the positive classes (high Recall, F1 and Precision scores)
- But poorly in detecting actual negatives (low Specificity score)

=> Mainly due to the imbalanced dataset

Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Default ElasticNet Classification before SMOTE-NC	0.887771	0.989005	0.934104	0.884978	0.572985	0.730378
Tuned ElasticNet Classification before SMOTE-NC	0.887759	0.989036	0.934099	0.884943	0.572836	0.730297
Autometrics before SMOTE-NC	0.887506	0.988706	0.93394	0.884922	0.572823	0.730164

Information Reduction Models 2/2

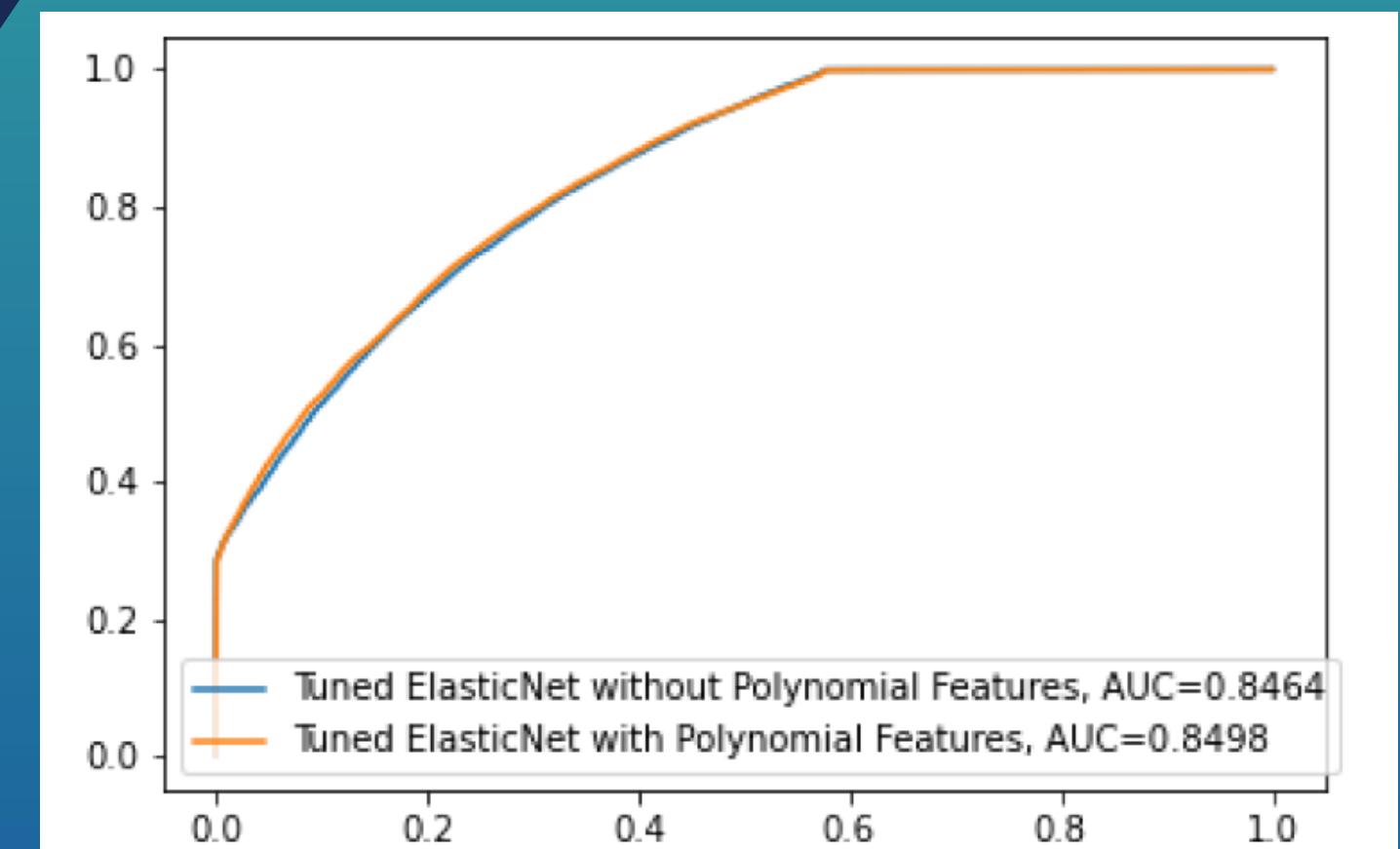
After SMOTE-NC

ElasticNet: Default & Tuned Hyperparameters

Classifiers	Accuracy	Recall	F1	Precision	Specificity
Tuned ElasticNet Classification before SMOTE-NC	0.887759	0.989036	0.934099	0.884943	0.572836
Default ElasticNet Classification after SMOTE-NC	0.783171	0.808428	0.85709	0.911987	0.704636
Tuned ElasticNet Classification after SMOTE-NC	0.783133	0.808381	0.857062	0.911982	0.704627

- The two models perform quite similarly
- But noticeably the Specificity scores for these 02 models are better than the before SMOTE-NC ones
- However, as a trade-off, the Recall and F1 scores decrease quite significantly

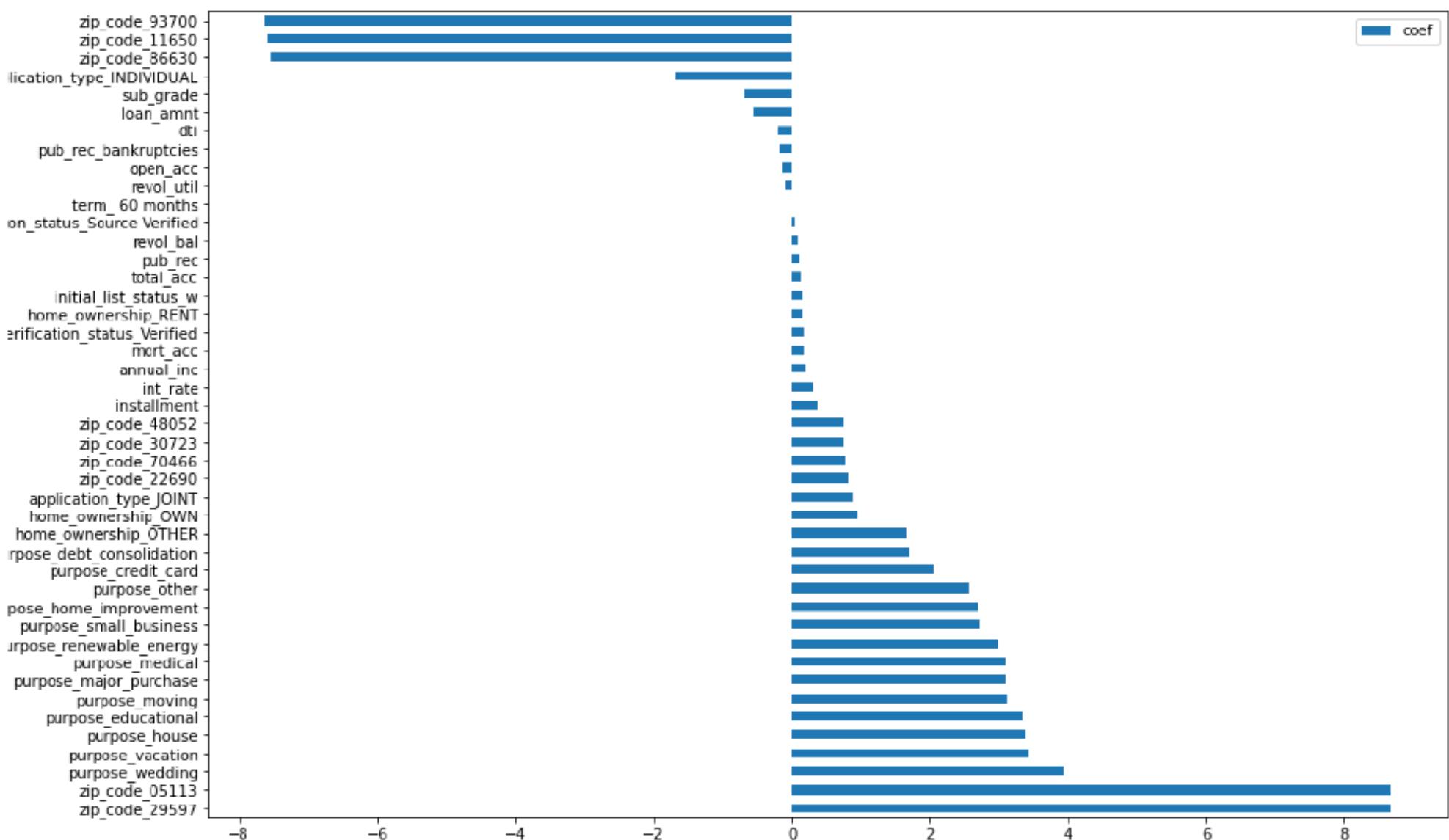
ElasticNet:
With Polynomial Features



Feature Importance

Interpretation of ElasticNet Classification model

- The **zip_code** feature plays a vital role in our classifier's predictive power
- Individual loans: charge-off indicator >< Joint loans: fully-paid
- The higher the **sub-grade**, the lower the likelihood that the applicants will fully pay the loans
- DTI feature performs as expected

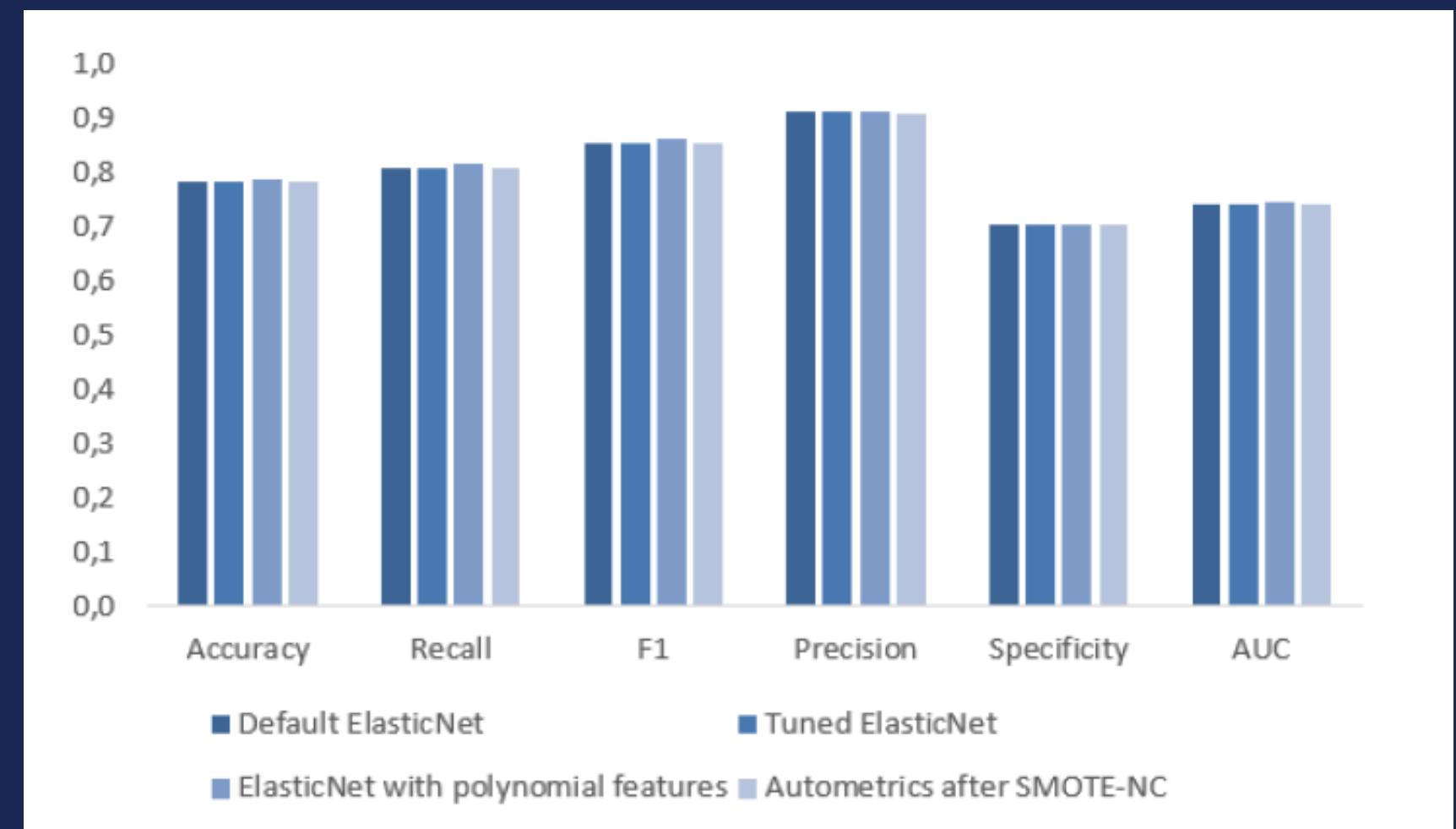


Information Reduction Models 2/2

After SMOTE-NC

Autometrics Models

- The regressor significance level is chosen as $\alpha=0,1$
- Terminal model
 - Before SMOTE-NC contains 28 variables (not including the constant)
 - After SMOTE-NC contains 42 out of a total of 44 features
- Regarding the power of information reduction, there is no superior technique in the case of our study

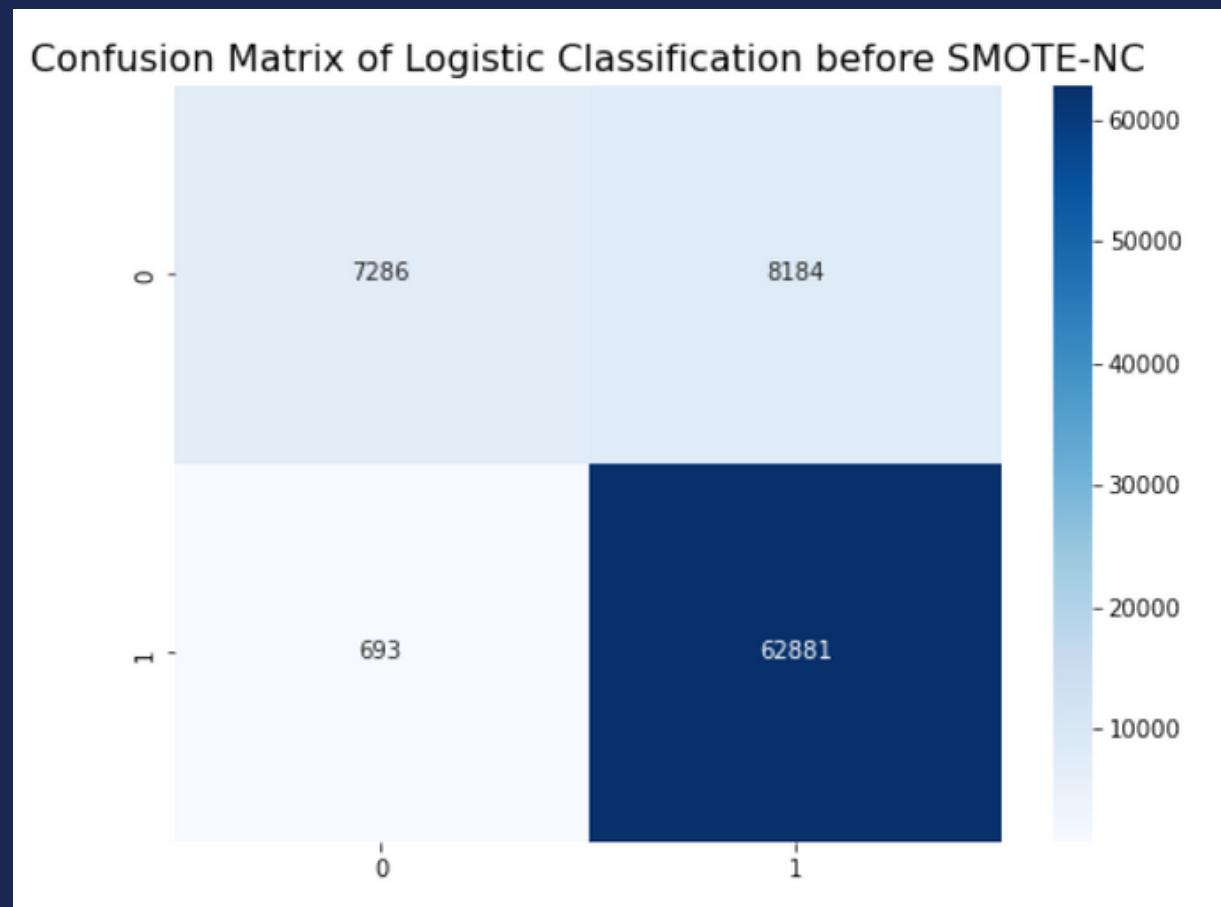


P3.3. Prediction Models

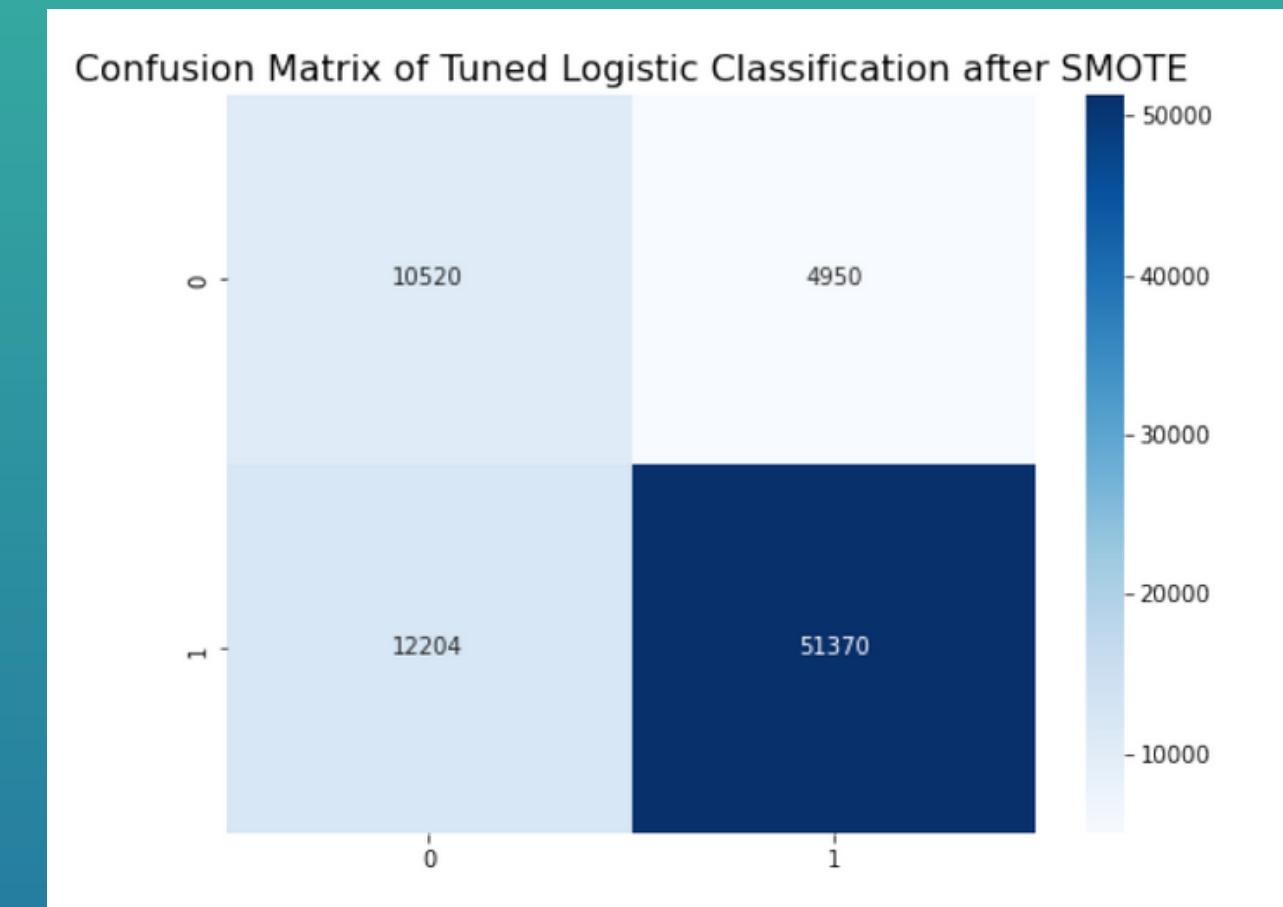
Prediction Models 1/2

Logistic Regression

Before SMOTE-NC



After SMOTE-NC



- The model after SMOTE can detect far more true negatives (i.e., actual charged-off loans) compared to the before SMOTE model
- This is what we expected when we synthetically increased the number of charged-off loans by using SMOTE-NC

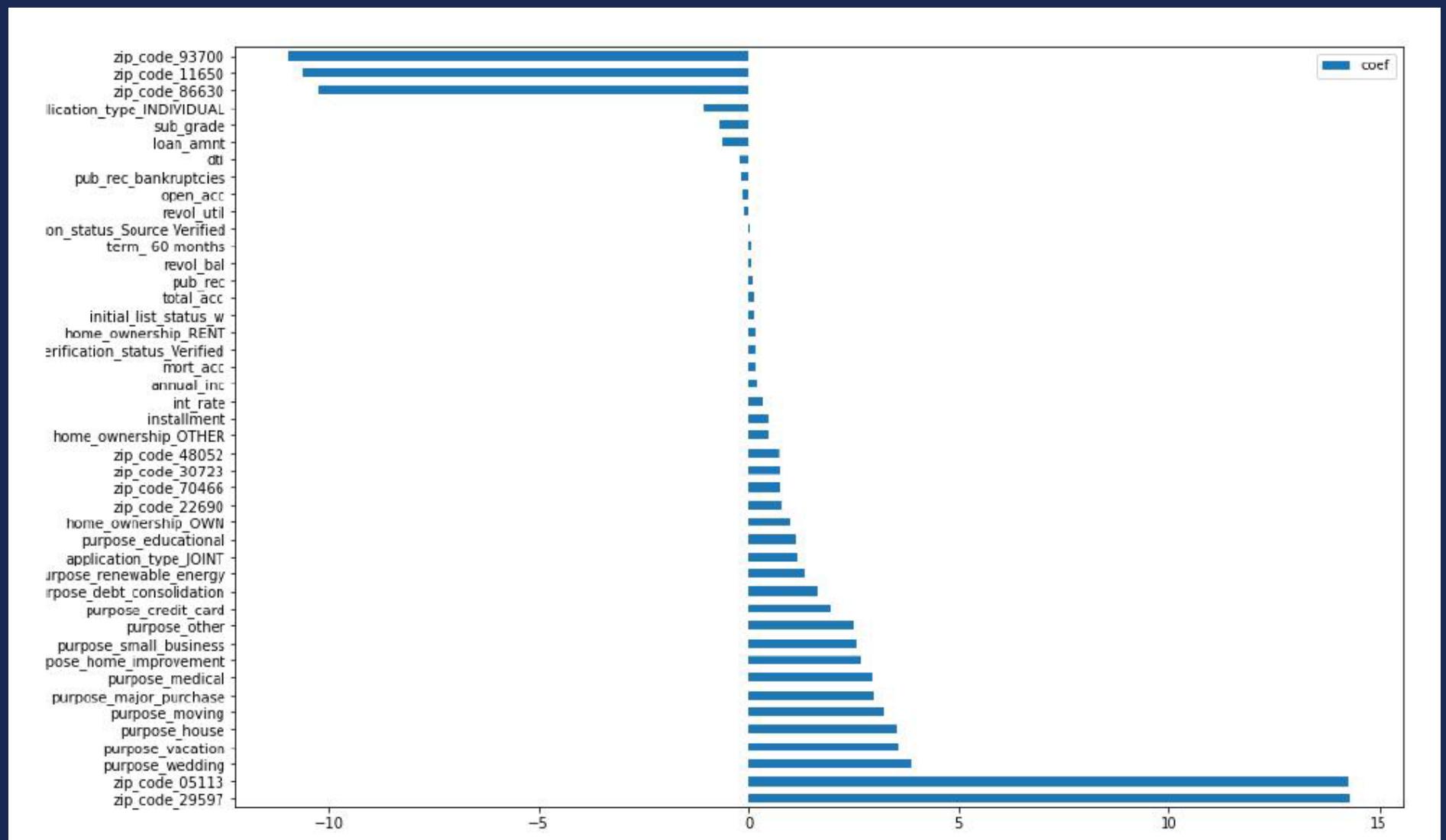
Feature Importance

Interpretation of Logistic Classification model

- Provides quite similar information to that of ElasticNet we interpreted before
- The coefficient magnitudes of Logistic Regression are higher than those of ElasticNet

=> Mainly due to multicollinearity

=> Implies that ElasticNet performs well in cases with highly-correlated data like ours



Prediction Models 1/2

Logistic Classification with Polynomial Features

- After adding polynomial features and interactions degree 02, Recall and F1 scores are improved while the Specificity and AUC scores decrease, both slightly
- Along with the mild results from ElasticNet with polynomial features analyzed in the section above, this result suggests that nonlinearity may not exist in our true Data Generation Process

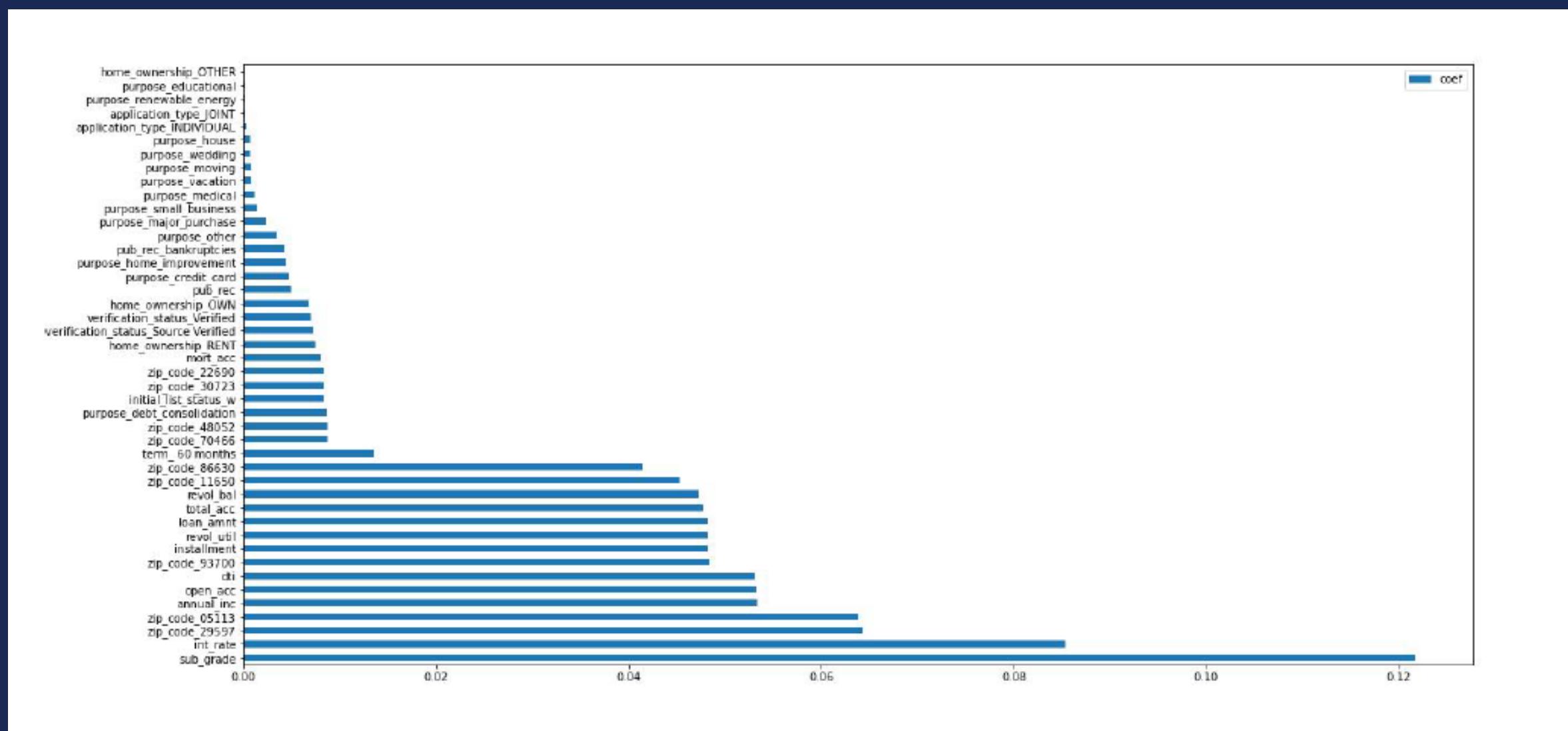
Classifiers	Accuracy	Recall	F1	Precision	Specificity	AUC
Tuned Logistic Classification after SMOTE-NC	0.782982	0.808035	0.856924	0.912109	0.705079	0.74403
Logistic Classification after SMOTE-NC with polynomial features	0.785309	0.812675	0.858936	0.910782	0.700216	0.742763

Prediction Models 2/2

Random Forest

- The evaluation metrics Recall and F1 show improvement compared to the other models
- However, the Specificity score of RF is lower than those of other models
- Sub-grade and interest rate are the two most vital variables
- Other variables which are not considered as important in other models appear in this plot of RF; for example: annual income, open account

Classifiers	Accuracy	Recall	F1	Precision	Specificity
Logistic Classification after SMOTE-NC	0.782918	0.807988	0.856881	0.912072	0.704966
Tuned Logistic Classification after SMOTE-NC	0.782982	0.808035	0.856924	0.912109	0.705079
Logistic Classification after SMOTE-NC with po...	0.785309	0.812675	0.858936	0.910782	0.700216
Random Forest after SMOTE-NC with polynomial f...	0.856181	0.924985	0.911861	0.899104	0.642236



Conclusion

01 Contribution

02 Limitation

03 Further research

Thank you for your attention!