Predicting Loan Approval

By Tangyue Gong, Haotian Kang,
Rongsheng Zhang, Aditya Agarwal

Research Motivation

Primary Research Goal - To build a predictive model that will effectively predict loan approval status for new datasets.

Secondary Research Goal - To build a model that can yield interpretative insights about the nature of the relationship between the variables in our dataset

Motivation - Debt is a vital for both our personal finances and the economy in general. As we use debt progress in life, we wanted to analyse how different personal metrics impact an individual's ability to obtain loans.

Example Use Case - Any bank or financial institution that wants to decide whether or not to provide a loan to a potential borrower

Dataset Discussion

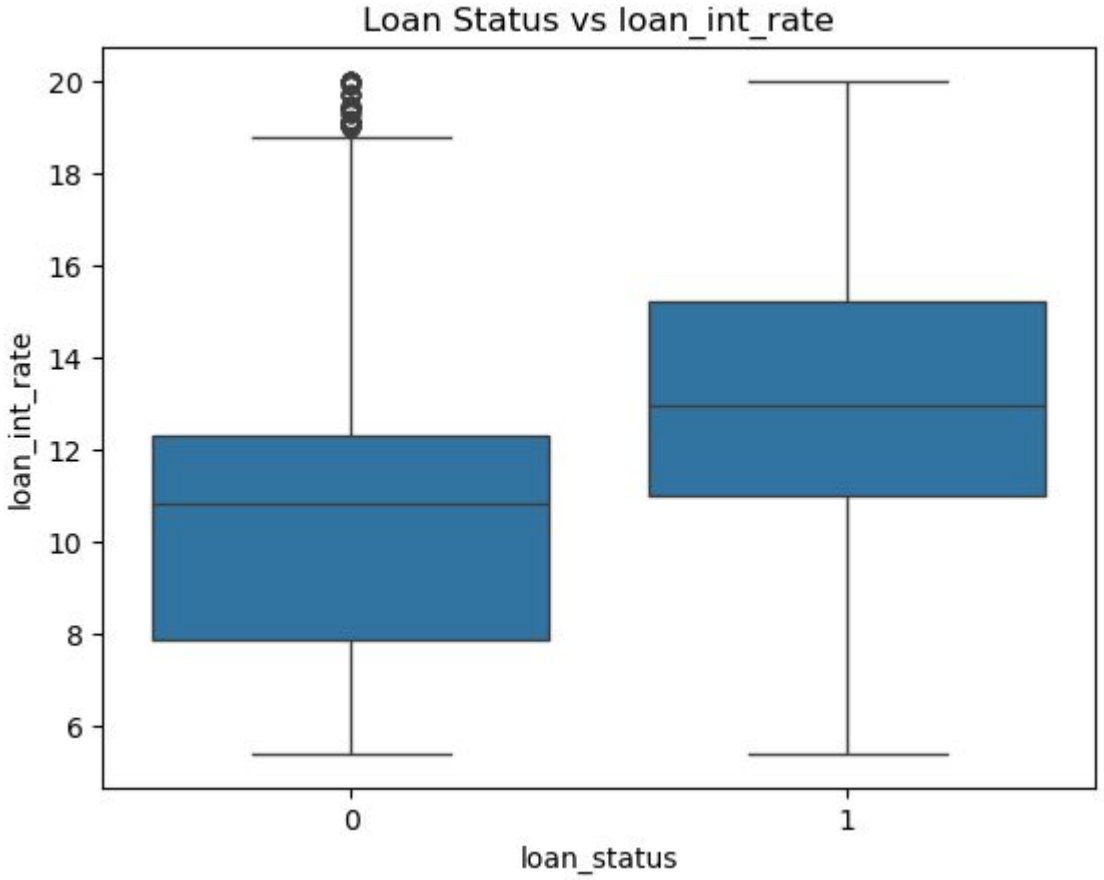
Dataset Origin - Combination of Credit Risk and Financial Risk for Loan Approval Dataset. Contains synthetic data

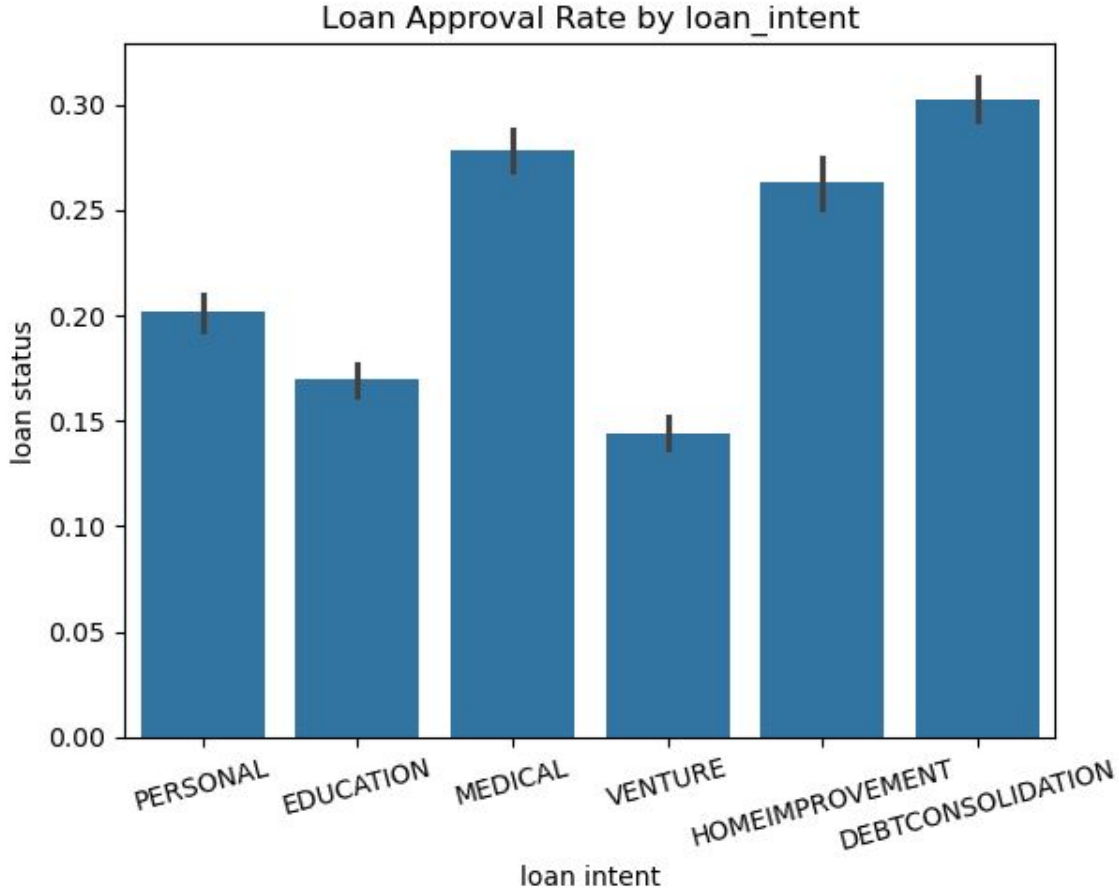
Cleaning - No implicit or explicit missing variables. Eliminated data with person_age > 100

Classifier Preference - Minimizing false positive outcomes.

	person_age	person_gender	person_education	person_income	person_emp_exp	person_home_ownership	loan_amnt	loan_intent	loan_int_rate	loan_percent_income	cb_person_cred_hist_length	credit_score	previous_loan_defaults_on_file	loan_status
0	22.0	female	Master	71948.0	0	RENT	35000.0	PERSONAL	16.02	0.49	3.0	561	No	1
1	21.0	female	High School	12282.0	0	OWN	1000.0	EDUCATION	11.14	0.08	2.0	504	Yes	0
2	25.0	female	High School	12438.0	3	MORTGAGE	5500.0	MEDICAL	12.87	0.44	3.0	635	No	1
3	23.0	female	Bachelor	79753.0	0	RENT	35000.0	MEDICAL	15.23	0.44	2.0	675	No	1
4	24.0	male	Master	66135.0	1	RENT	35000.0	MEDICAL	14.27	0.53	4.0	586	No	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
44995	27.0	male	Associate	47971.0	6	RENT	15000.0	MEDICAL	15.66	0.31	3.0	645	No	1
44996	37.0	female	Associate	65800.0	17	RENT	9000.0	HOMEIMPROVEMENT	14.07	0.14	11.0	621	No	1
44997	33.0	male	Associate	56942.0	7	RENT	2771.0	DEBTCONSOLIDATION	10.02	0.05	10.0	668	No	1
44998	29.0	male	Bachelor	33164.0	4	RENT	12000.0	EDUCATION	13.23	0.36	6.0	604	No	1
44999	24.0	male	High School	51609.0	1	RENT	6665.0	DEBTCONSOLIDATION	17.05	0.13	3.0	628	No	1

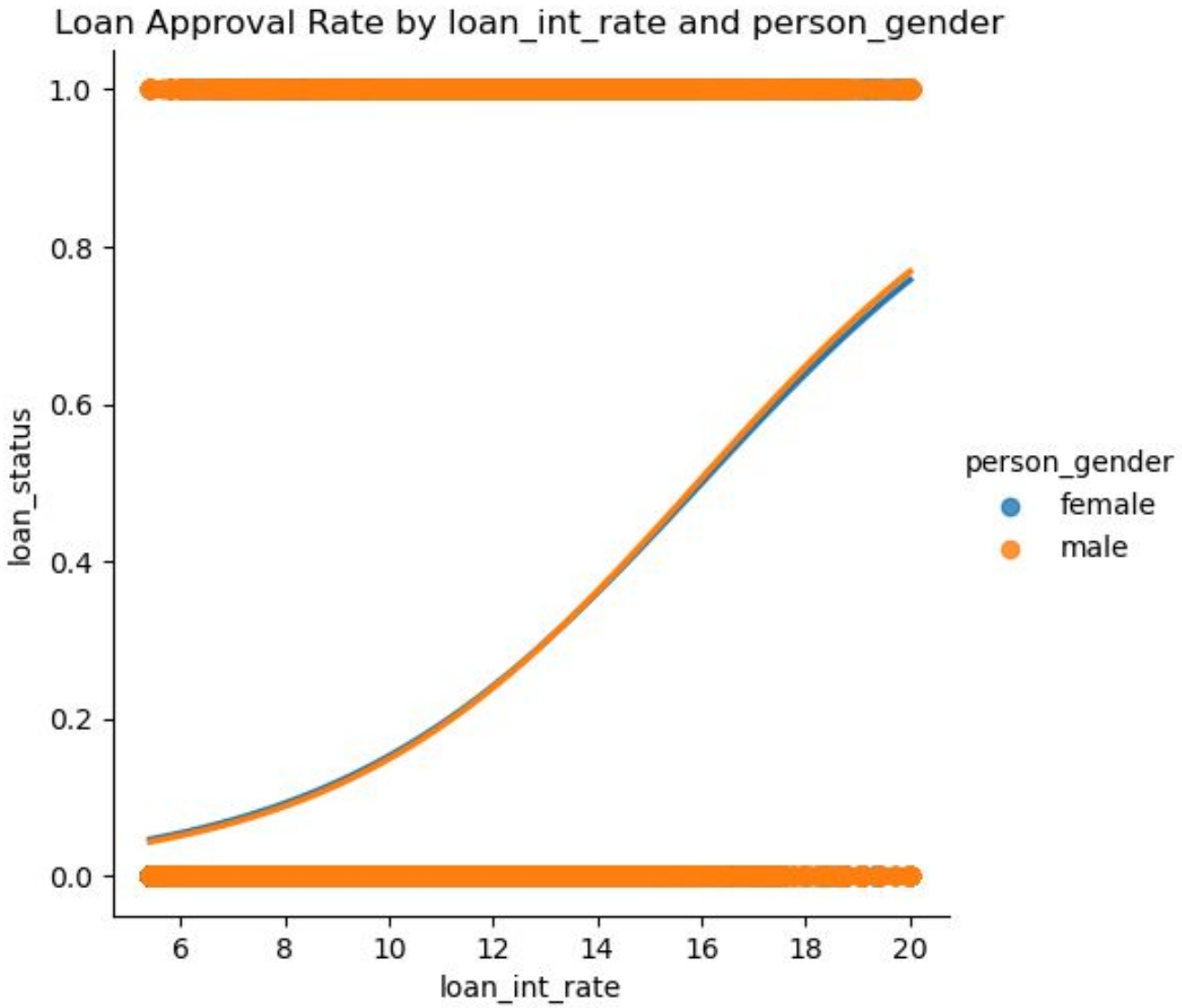
Preliminary Analysis - Single variable





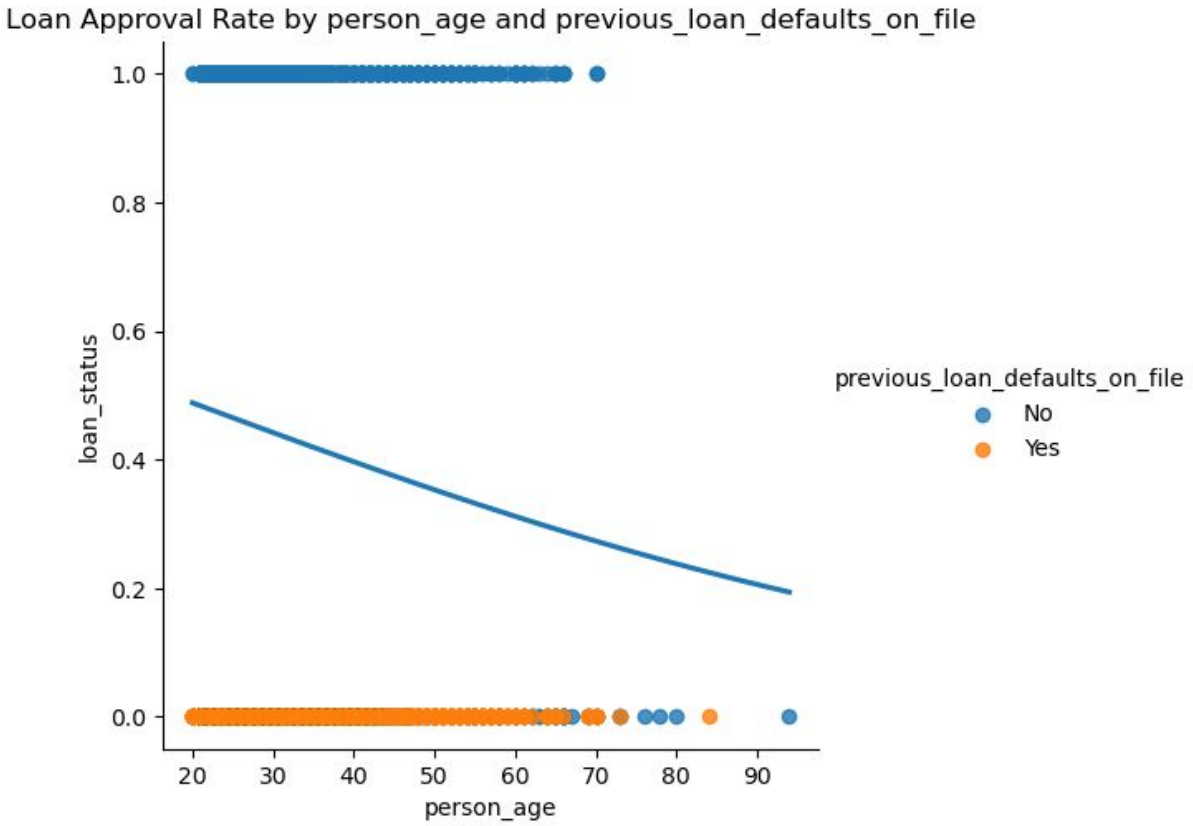
- *Slightly higher interest rates for approval loans*
- *Approval rates vary by purpose*

Preliminary Analysis - Weak Interaction

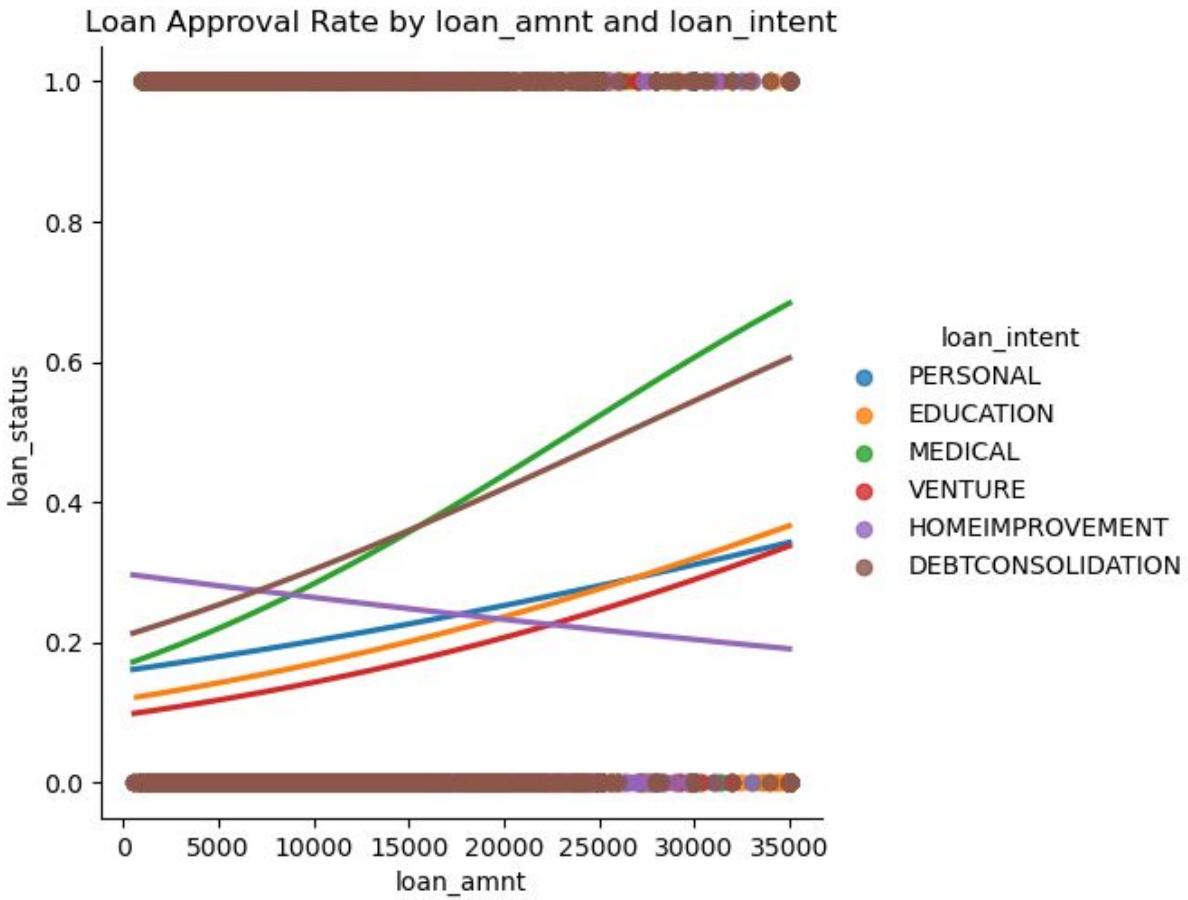


Lines overlap → no clear interaction between gender & interest rate.

Preliminary Analysis - Strong Interaction



Age × Previous Defaults:
Strong difference: approval drops with age for no-default group, stays near zero for defaults.

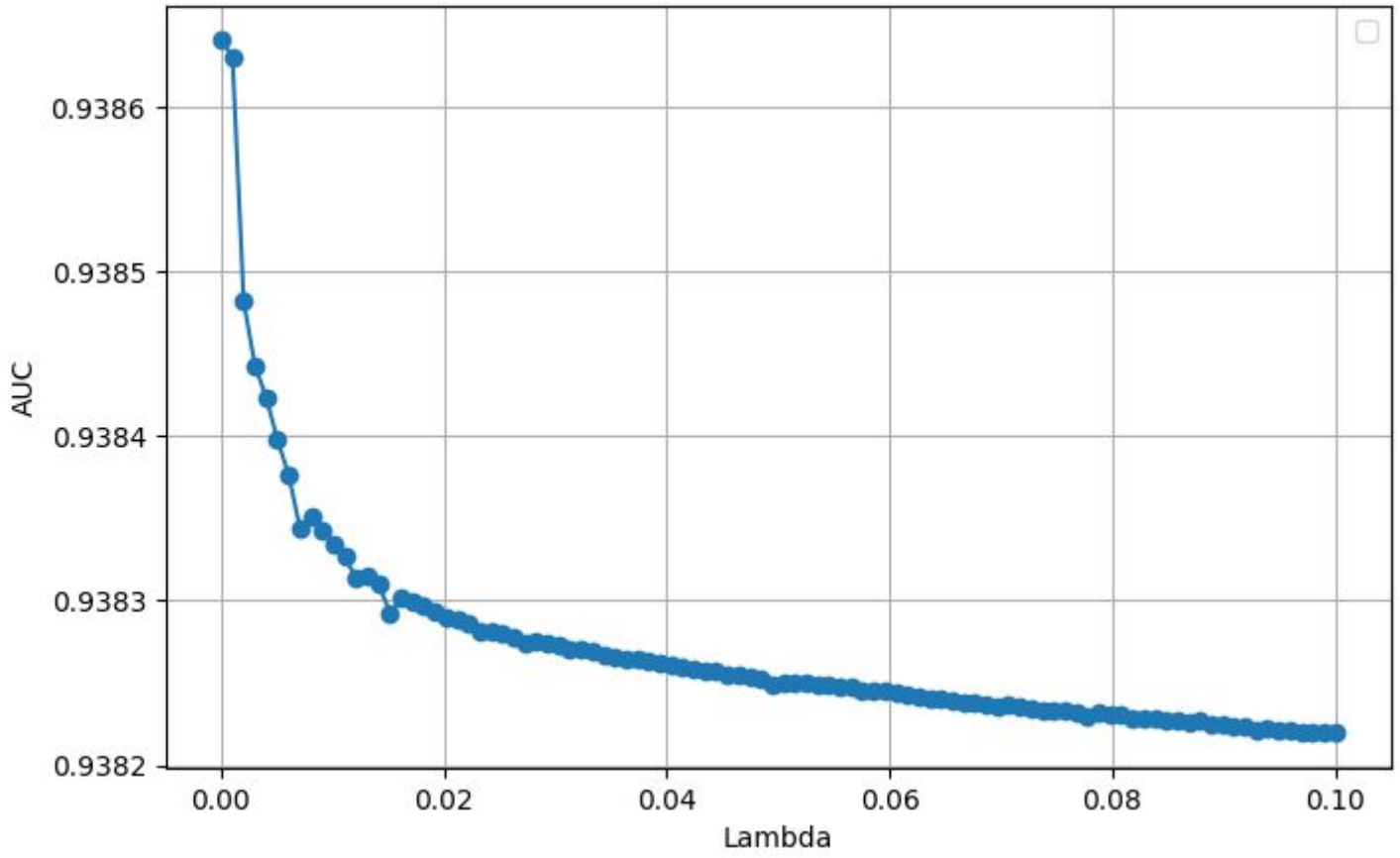


Loan Amount × Loan Intent:
Slopes vary by intent → moderate interaction between amount & purpose.

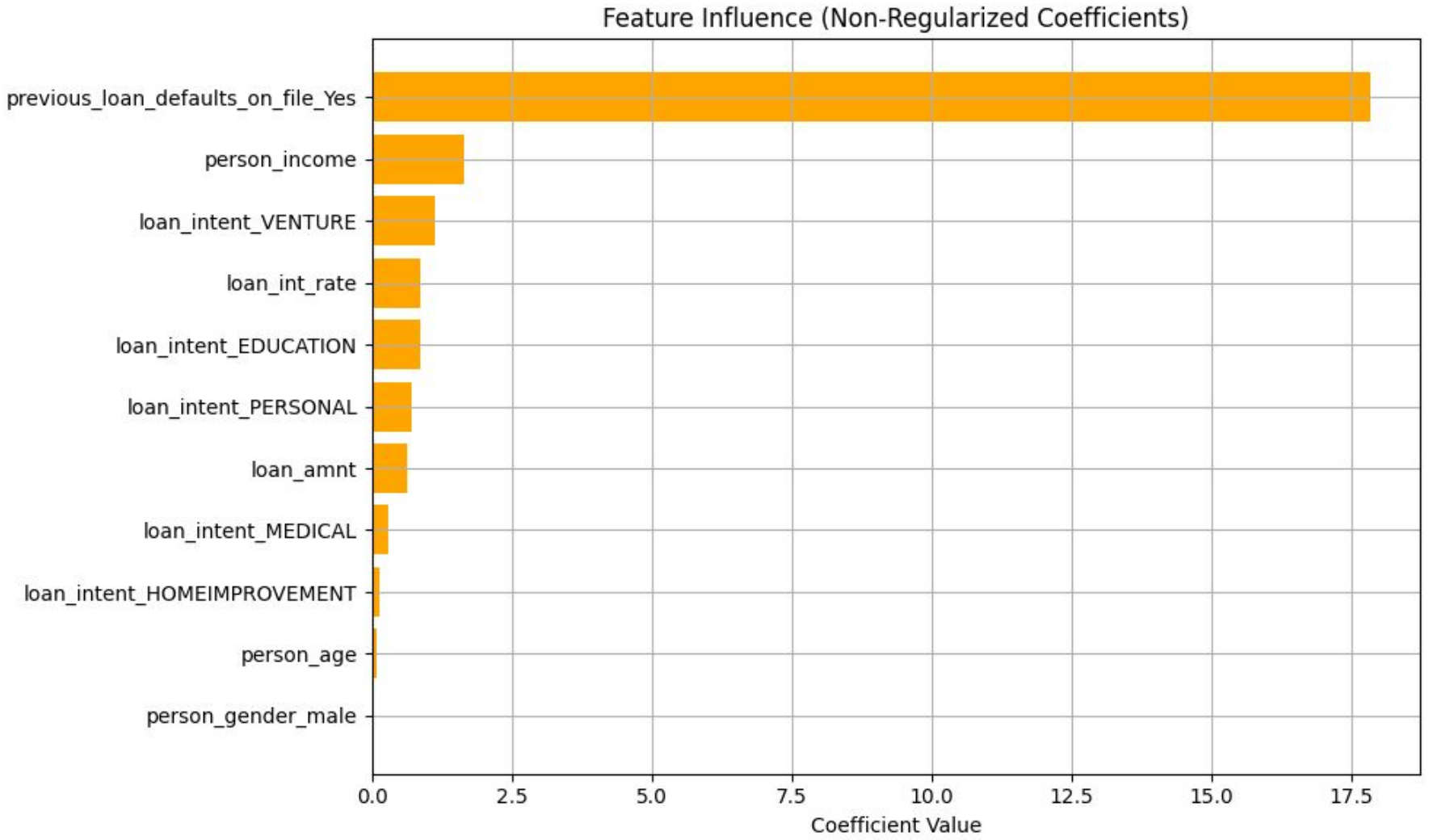
Feature Selection

Experiment *- lambda ranging from 0.00001 to 0.1*

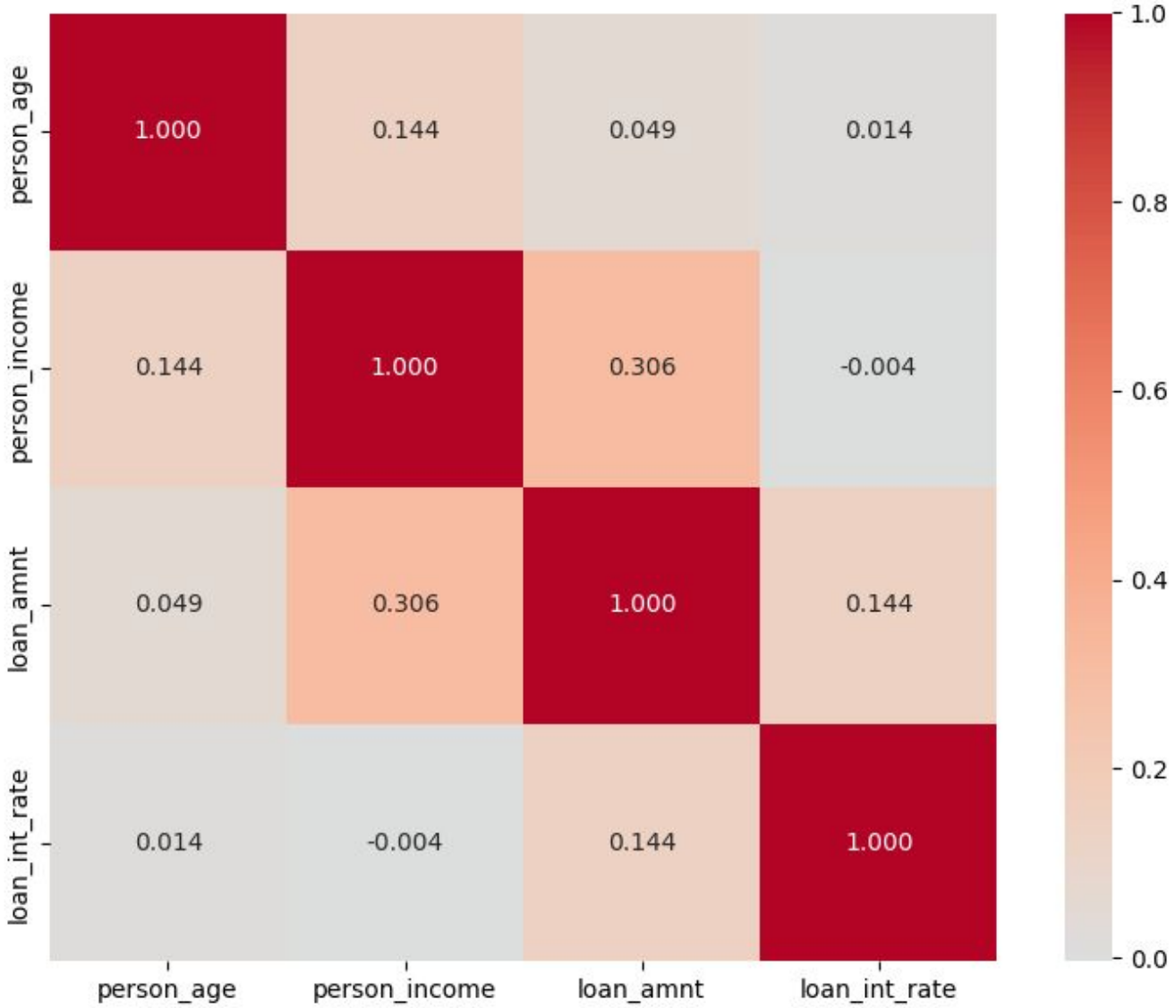
Analysis *- Highest AUC when lambda is small. All features bring enough predictive power.*



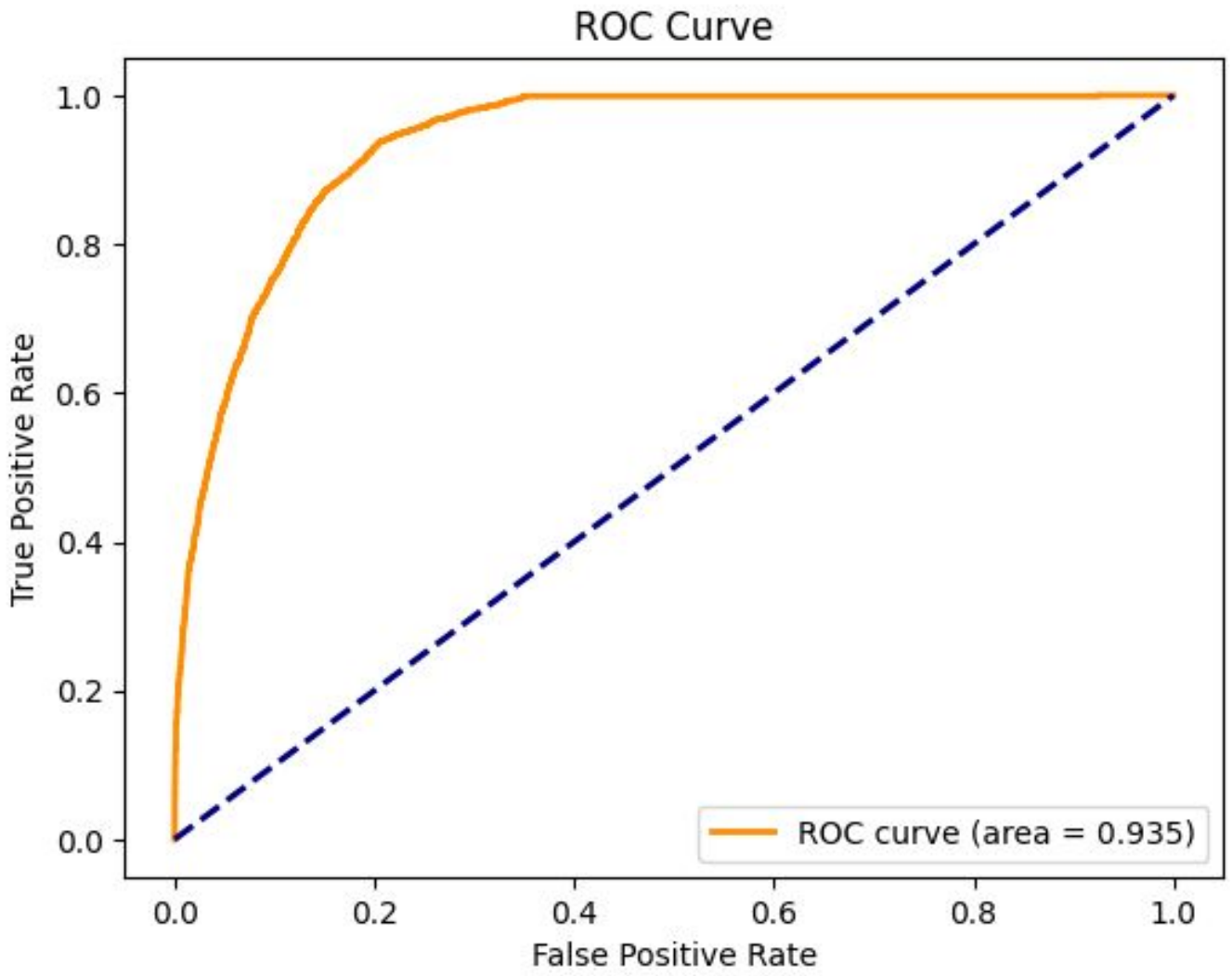
Feature Selection - Result



Best Model Discussion



Best Model Discussion



Best Model *- Low FPR*

FPR *- 0.0530*

TPR *- 0.600*

Best Model Discussion - Equation

$$\hat{p} = \frac{1}{1 + \exp\left(-\left[\begin{array}{l} -1.18539112 \\ -0.022838 \cdot \text{person\_age} \\ -0.000026 \cdot \text{person\_income} \\ +0.000098 \cdot \text{loan\_amnt} \\ +0.260390 \cdot \text{loan\_int\_rate} \\ +0.010904 \cdot \text{person\_gender\_male} \\ -14.967081 \cdot \text{previous\_loan\_defaults\_on\_file\_Yes} \\ -0.927160 \cdot \text{loan\_intent\_EDUCATION} \\ -0.193122 \cdot \text{loan\_intent\_HOMEIMPROVEMENT} \\ -0.341704 \cdot \text{loan\_intent\_MEDICAL} \\ -0.782793 \cdot \text{loan\_intent\_PERSONAL} \\ -1.179984 \cdot \text{loan\_intent\_VENTURE} \end{array}\right]\right)}$$

Additional Analysis - K-Means

cluster	person_age	person_income	loan_amnt	loan_int_rate	loan_status
0	28.787065	73870.252341	7449.259528	8.205212	0.103022
1	29.549471	133890.468783	19319.571794	12.275271	0.274154
2	26.042338	61273.942107	7135.830852	12.893825	0.303528

Three clusters:

low-risk, medium-risk and high-risk

Compare the differences between each cluster:

By comparing cluster 0 and cluster 2, under the condition that person_income is similar, the higher the loan_int_rate is, the higher the probability of loan approval is

Project Summary and Recommendations

Research Motivation - Develop an accurate and interpretable model to assist banks in making informed loan decisions.

Dataset Discussion - Use a synthetic dataset from kaggle and clean up the abnormal data of age.

Preliminary Analysis - Find loan_int_rate and loan_intent have strong impact; identify key interaction effects, especially with previous_loan_defaults.

Feature Selection - Use K-fold cross-validation on LASSO logistic regression model for robust feature selection.

Model Discussion - Fit a LASSO logistic regression model, achieving a high test AUC score of 0.935, indicating accurate prediction.

Shortcomings

A synthetic dataset
lacks extreme values

Limit to LASSO logistic regression

Limit explanatory variables

Future Work

Use more realistic datasets

Try more models and feature selection methods

Incorporate more financial ratios (debt-to-income, loan-to-value)

Reference

- (1) Mian, Atif, and Amir Sufi. House of Debt: How They (and You) Caused the Great Recession, and How We Can Prevent It from Happening Again. University of Chicago Press, 2014. https://press.uchicago.edu/ucp/books/book/chicago/H/bo20832545.html
- (2) Stein, Tristan, and Emily Wielk. "Student Loan Default: How Policy and Politics Are Failing Borrowers." Bipartisan Policy Center, 16 Apr. 2025, https://bipartisanpolicy.org/blog/student-loan-default-how-policy-and-politics-are-failing-borrowers/

Thank you for listening