id/x partners x Rakamin Academy

LendingClub

ANALYSIS

Make numbers tell

Statistics

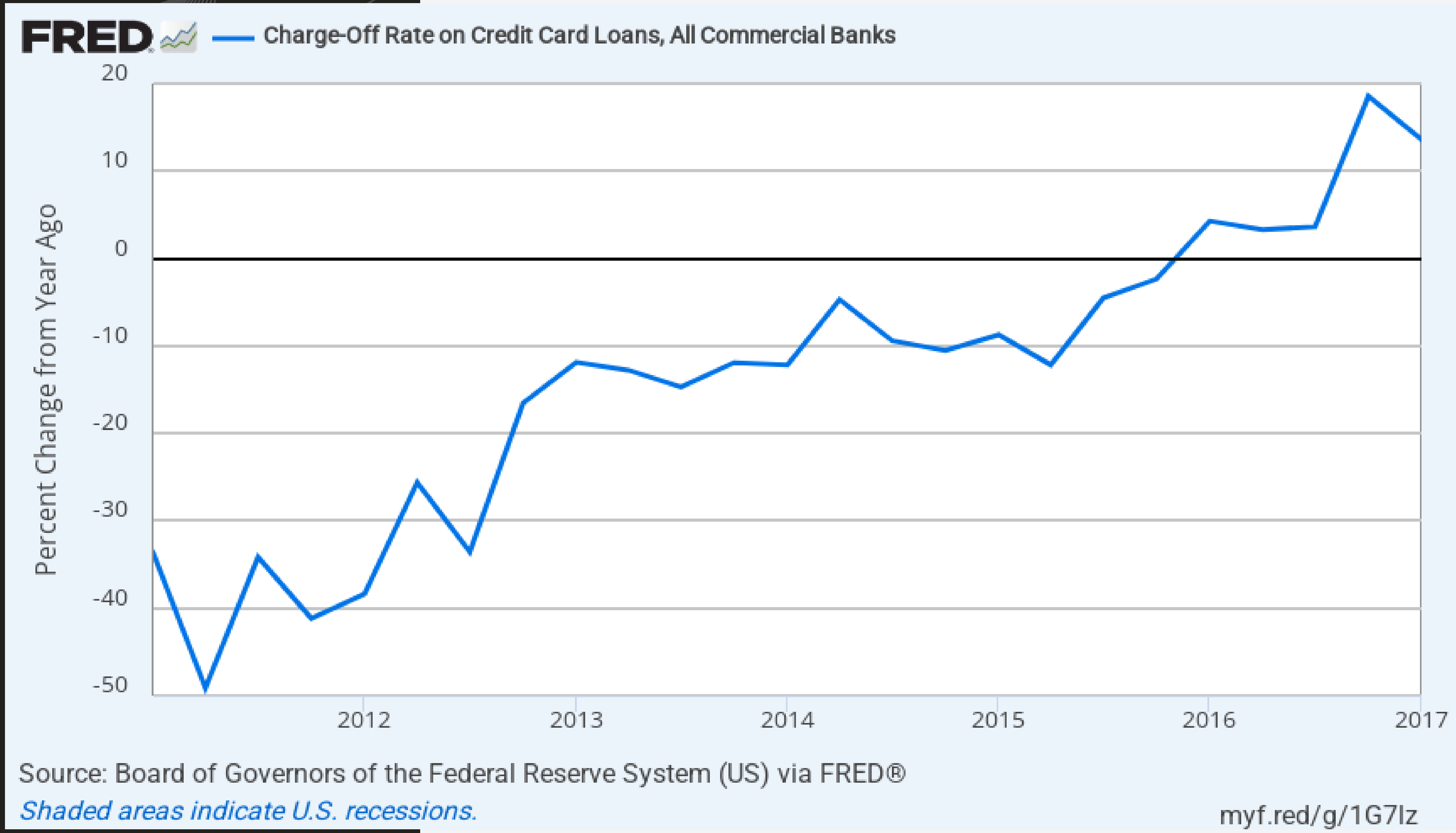
3.7%

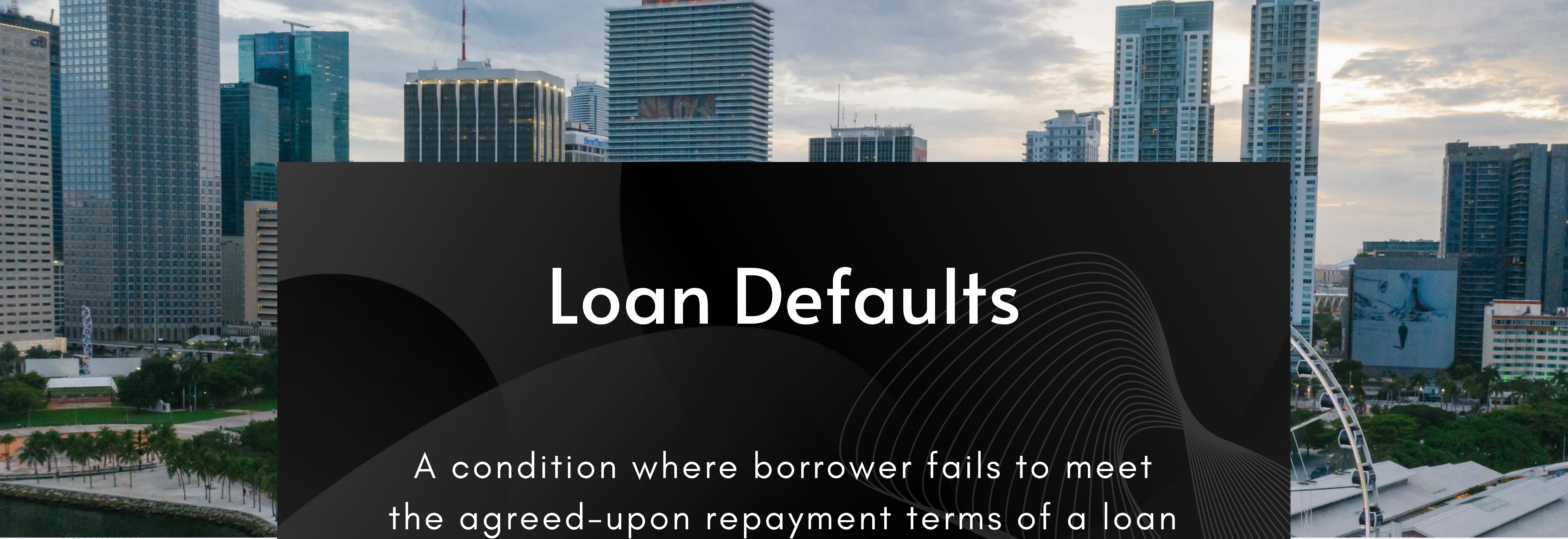
Avg Charge-off

Up to

10 B$

Annually





Loan Defaults

A condition where borrower fails to meet the agreed-upon repayment terms of a loan

Lending Club, one of the pioneers of online peer-to-peer lending, has facilitated billions in loans. Though viewed as a pioneer in the fintech industry and one of the largest such firms, LendingClub experienced problems in early 2016, with difficulties in attracting investors.

Understanding the factors that contribute to loan defaults on such platforms is vital for investors, borrowers, and the platform's sustainability.

My Journey

2025
March

Kimia Farma X Rakamin Academy
As Big Data Analyst

2024
Januari

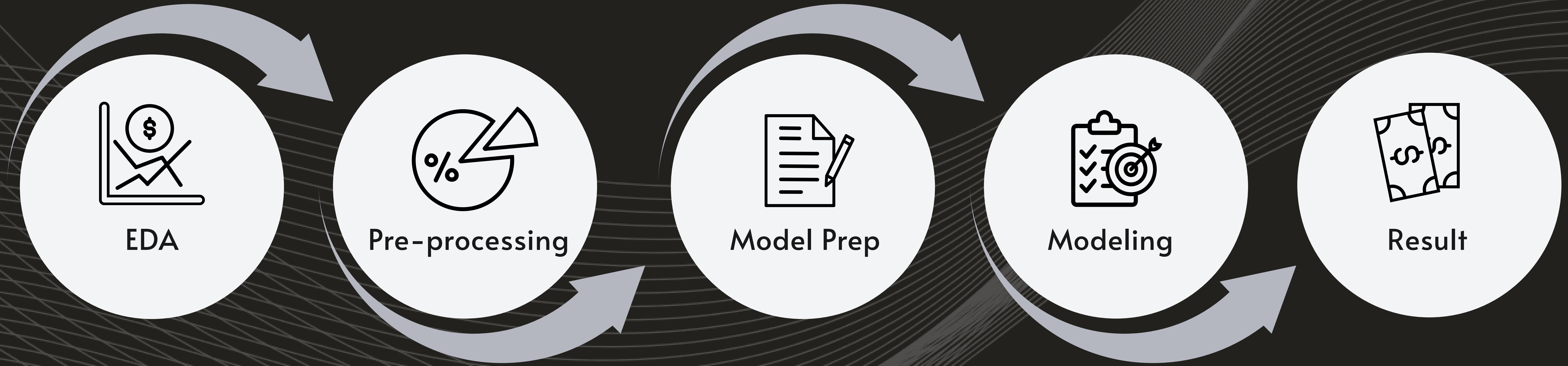
Bank Muamalat X Rakamin Academy
As Business Intelligence Analyst

2024
December

Rakamin Academy Bootcamp Final Project
As Lead Data Scientist

Bramantyo Raka

Highlights



The "Main" Characters

- Loan Amount (The Principal)
- Interest Rate (The Cost of Borrowing)
- Borrower's Income (Their Ability)
- Loan Purpose (The Reason)

[Sources]

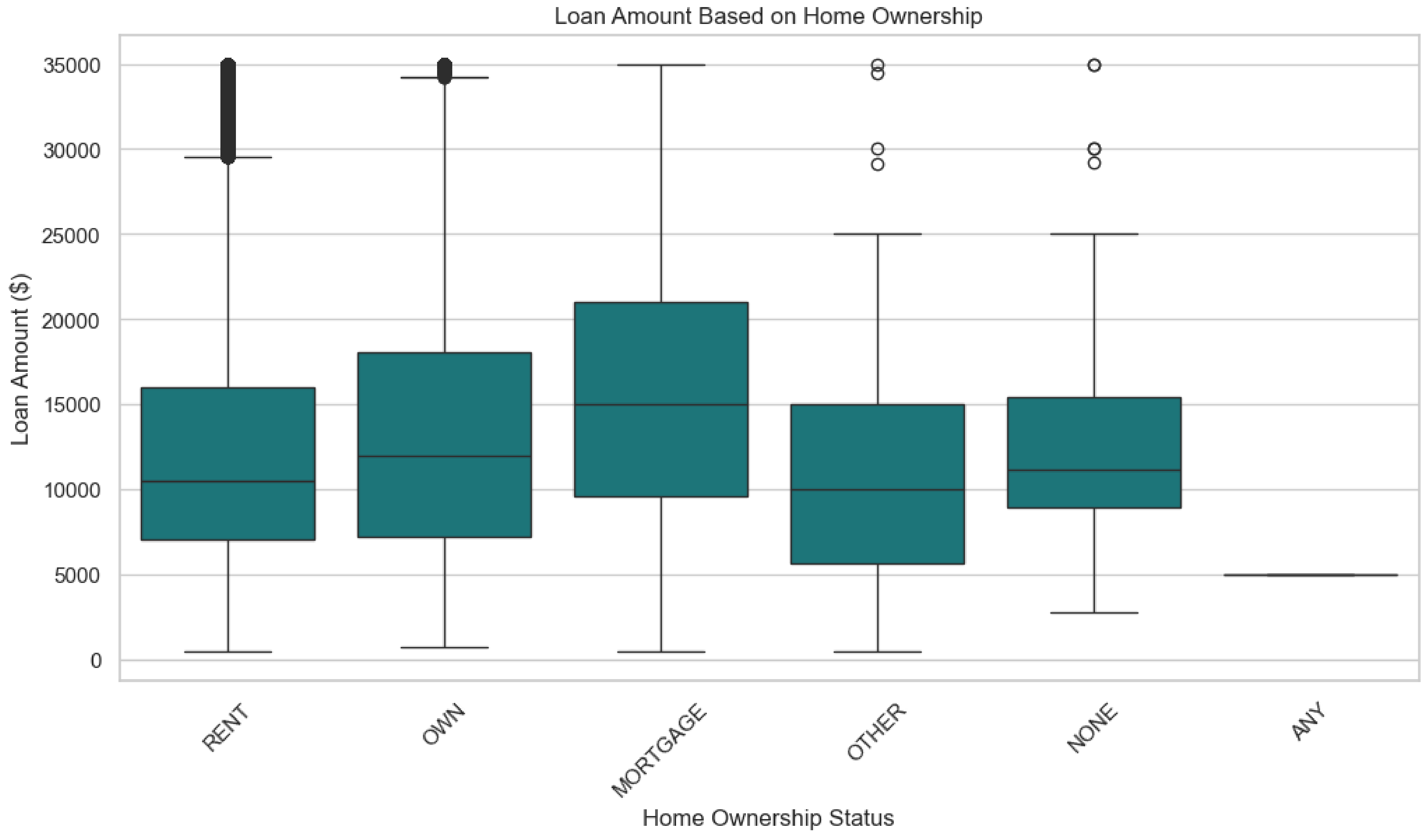
Loan Amt Distribution

Moderate Fund

For Most Loan

Other Purposes

Tail Ends



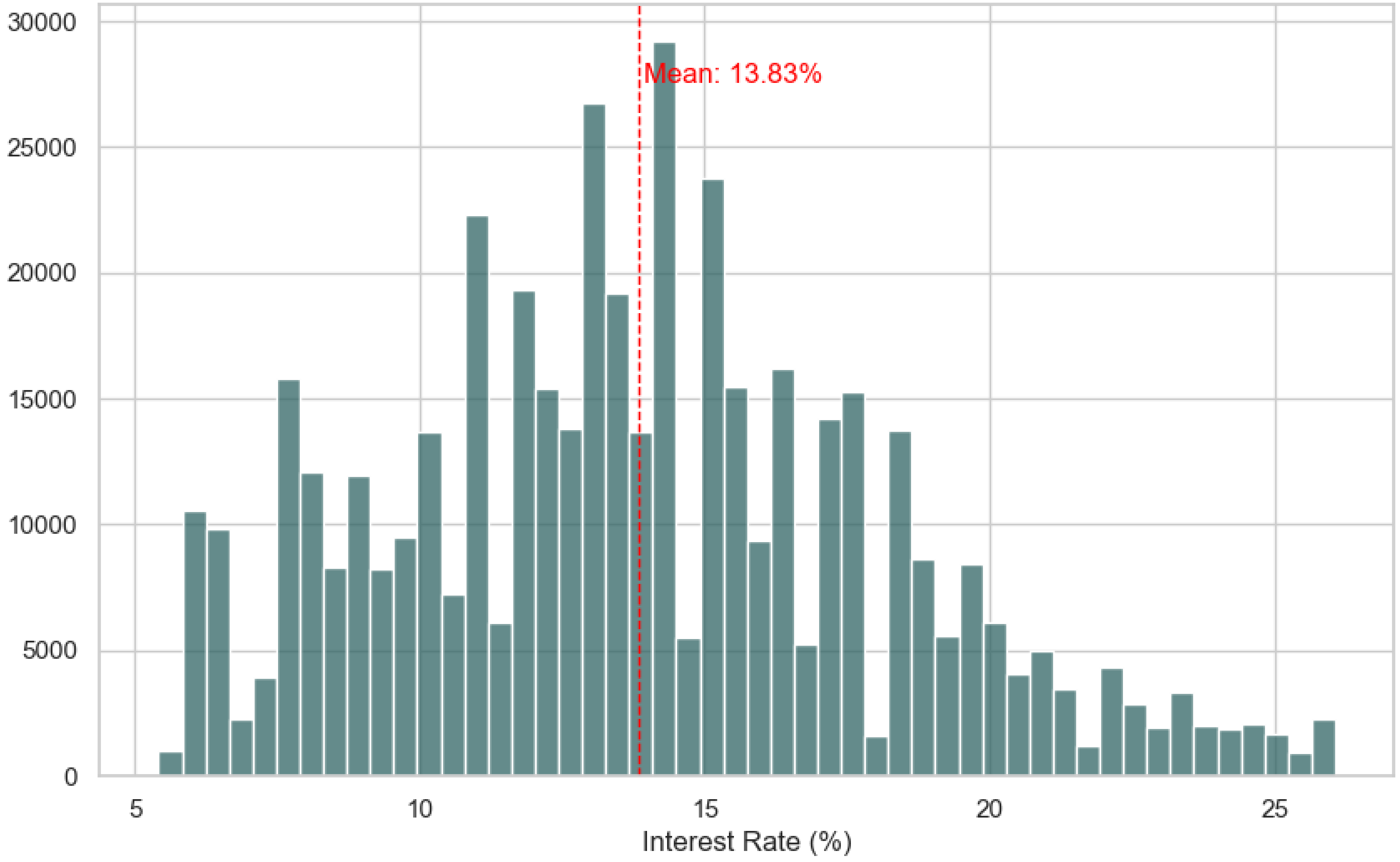
Interest Rate Distribution

Slightly Skewed Higher End

Risk-Based Pricing

Balance

Risk = Return



Sources

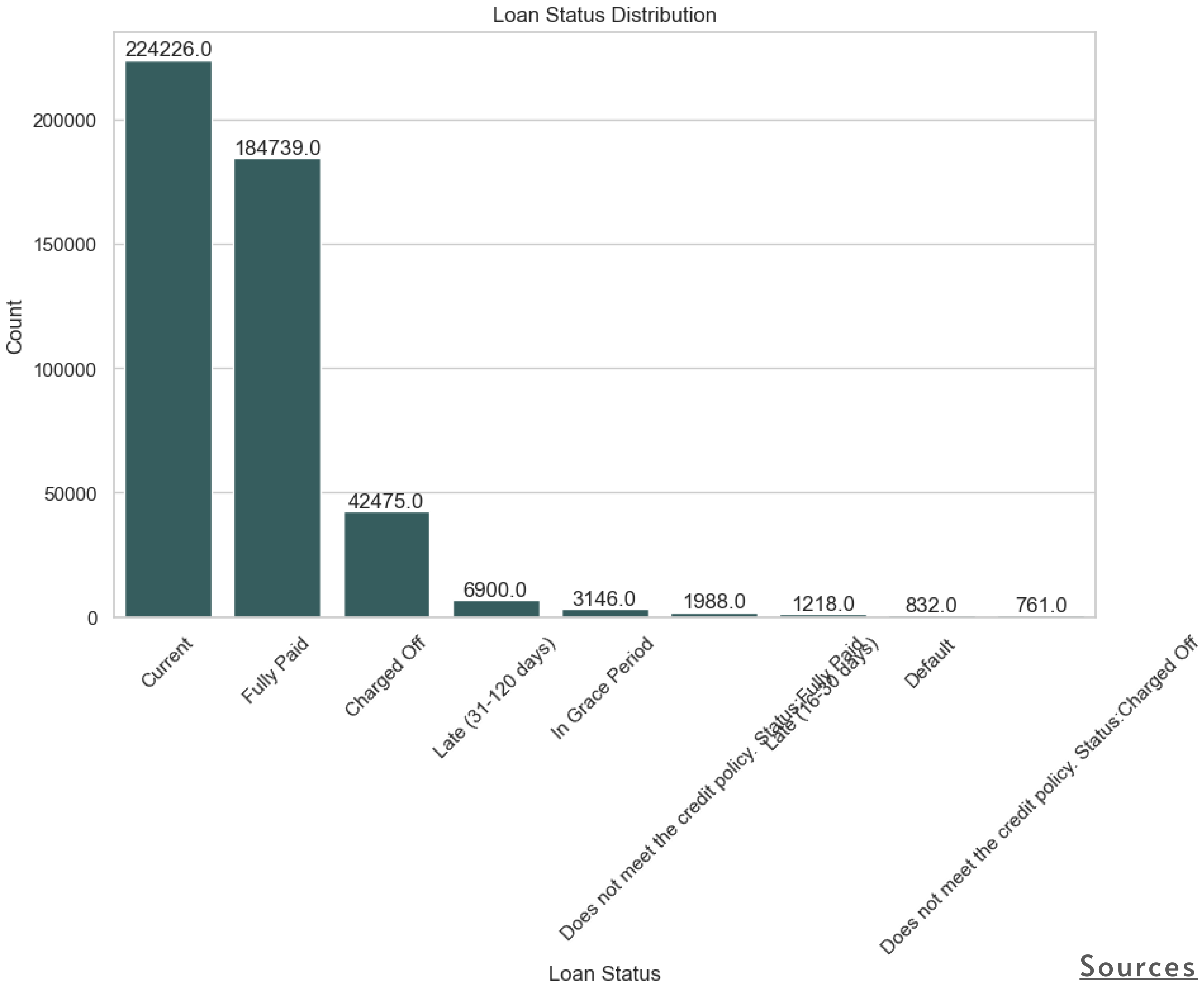
Loan Status Distribution

Fully Paid

Our “Happy Ending”

Charged Off / Default

Not so “Happy Ending” ☹

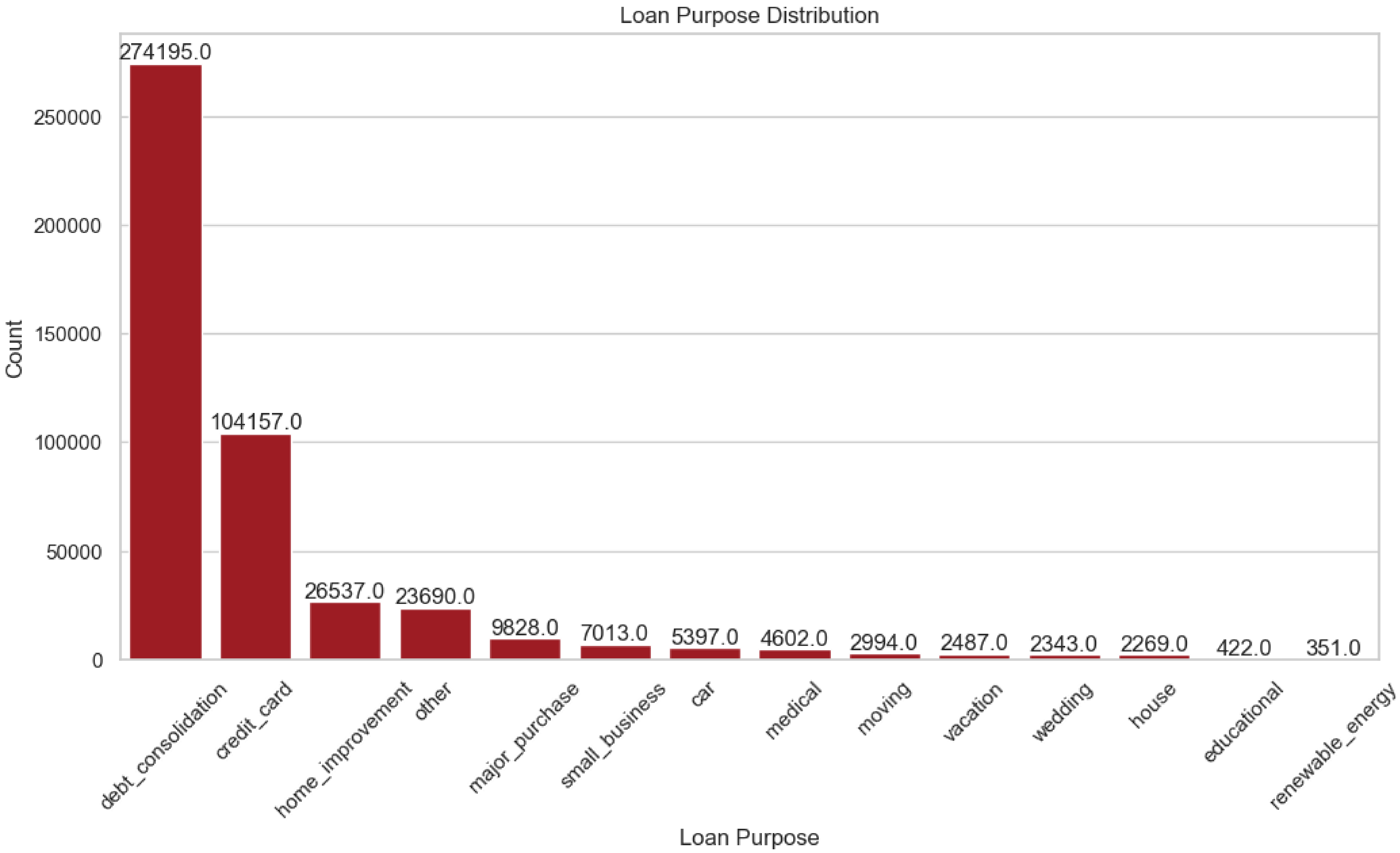


Sources

Purpose and Default

Begs a Question

“ Are loans taken for 'small businesses' more likely to default than those for 'debt consolidation’? ”



Sources

Economic Conditions

Plays a major role, for instance take the 2008 financial crisis where default rate surges across the board

Sources

Pre-processing

Handling Missing Values

Dropped Features w/ > 50% Missing Values
Imputed Missing Numerical Values w/ Median
And Missing Categorical Values w/ Mode

Encoding

Created binary ‘target’ from ‘loan_status’
One-hot encoded categorical variables

Datetime Conversion

Converted ‘date’ columns to datetime format

Model Prep

Feature Selection / Engineering

Created new features such as ‘issue_year’, ‘issue_month’
‘credit_history_length’
Removed redundant features

Scaling

Scaled numerical features using StandardScaler

Handling Class Imbalance

Used SMOTE to oversample the minority class

The Storytellers

Logistic Regression

Random Forest

XGBoost

Sources

Evaluation Metrics

Classification Reports

AUC-ROC

Confusion Matrices

Crossval

Results

Metric	Logistic Regression	Random Forest	XGBoost
Accuracy	0.95	0.97	0.98
AUC-ROC	0.9398	0.9569	0.9872
Mean Cross-Validation AUC-ROC	0.9615	0.9988	0.9985
Precision (Class 0)	0.98	0.97	0.98
Recall (Class 0)	0.96	1	1
F1-score (Class 0)	0.97	0.98	0.99
Precision (Class 1)	0.73	0.98	0.98
Recall (Class 1)	0.82	0.73	0.84
F1-score (Class 1)	0.77	0.84	0.9
Weighted Avg F1-score	0.95	0.97	0.98

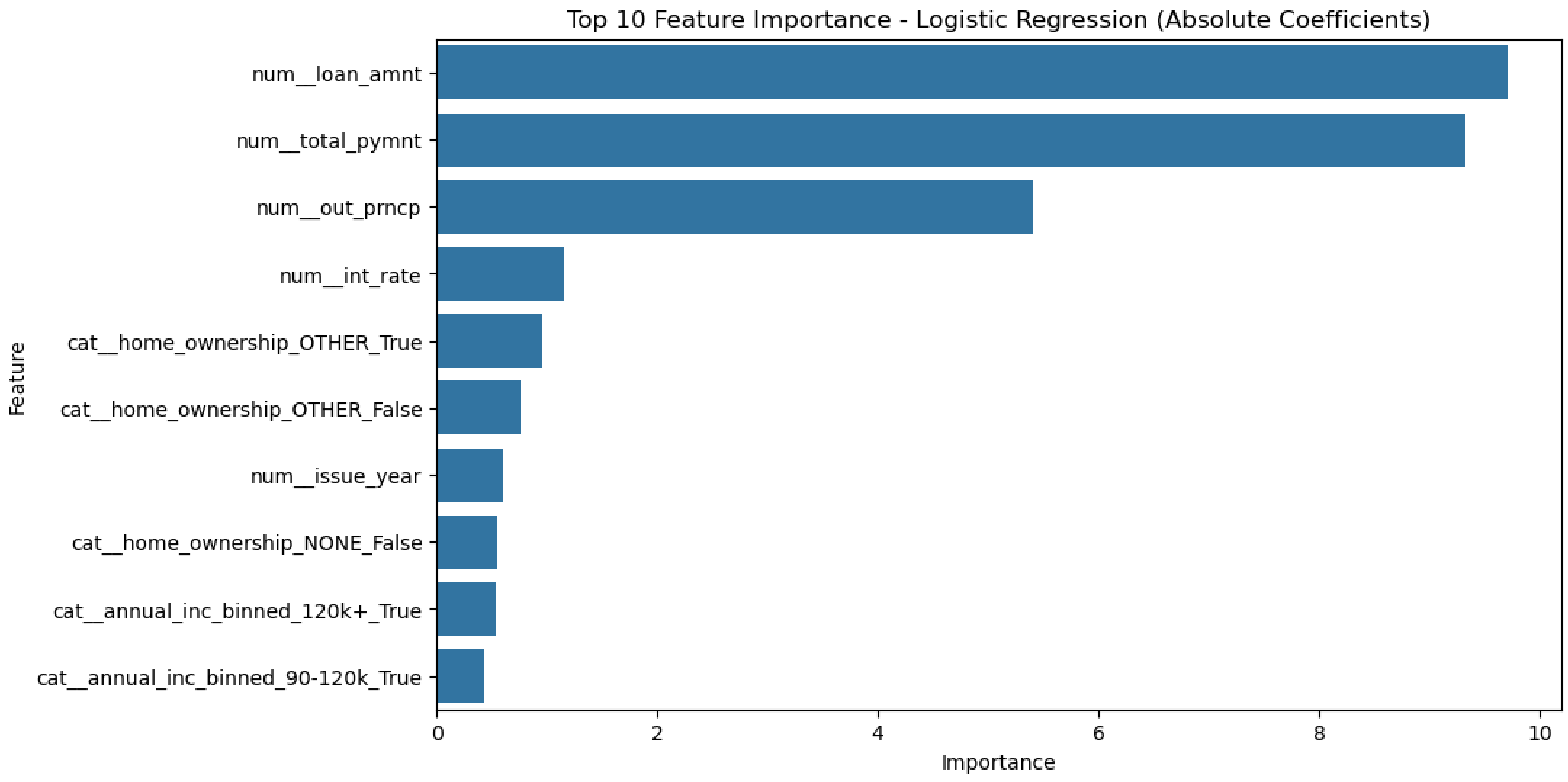
Logistic Regression, The Chosen One

It offers a strong combination of good predictive performance, interpretability, and resistance to overfitting.

Feature Importance

Influential Characters

“ Loan amount, total payment, outstanding principal, and interest rate are the most influential characters. ”

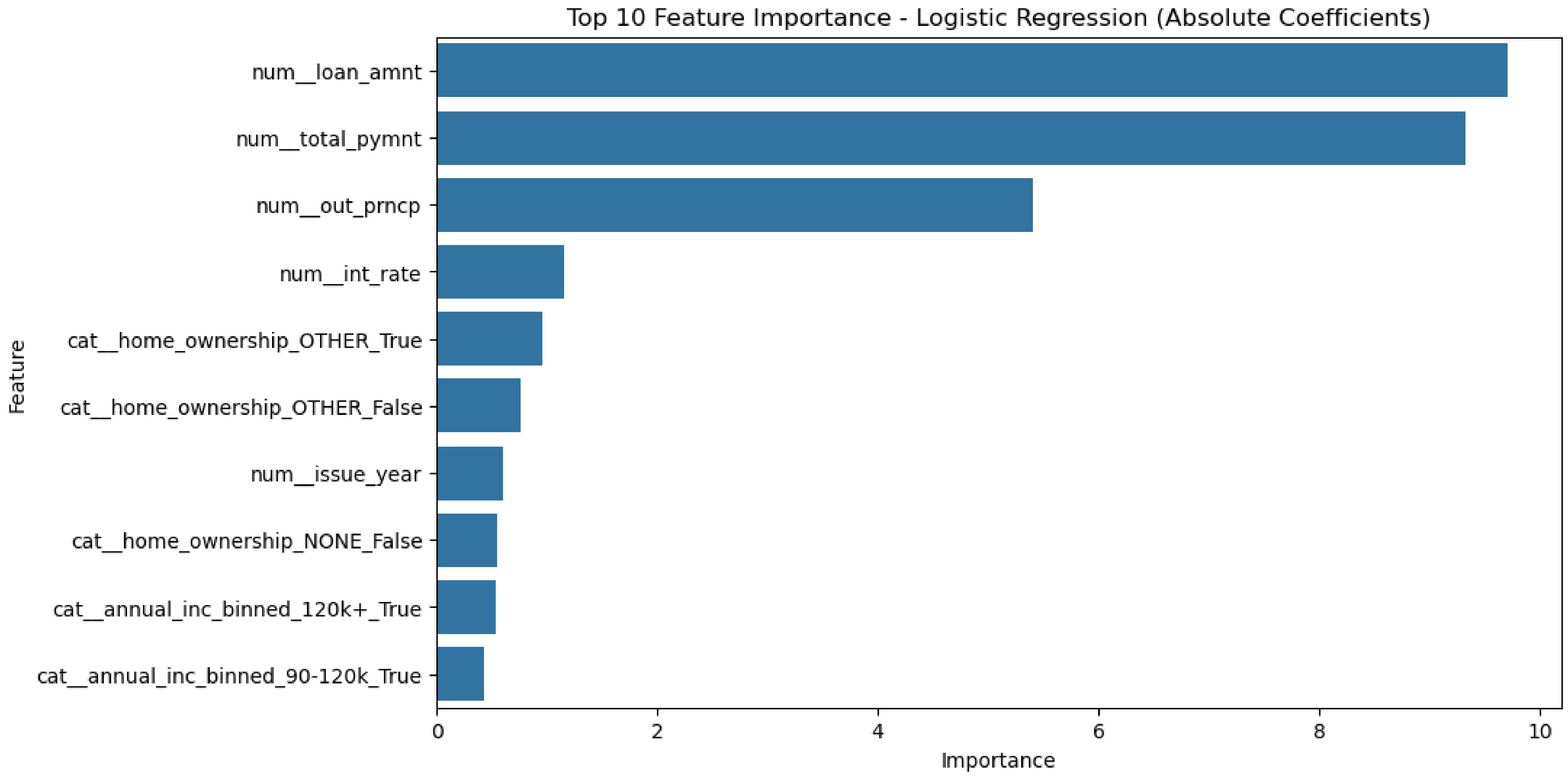


Sources

Feature Importance

Borrower's Characters

" Homeownership and income level also contribute to the prediction. "

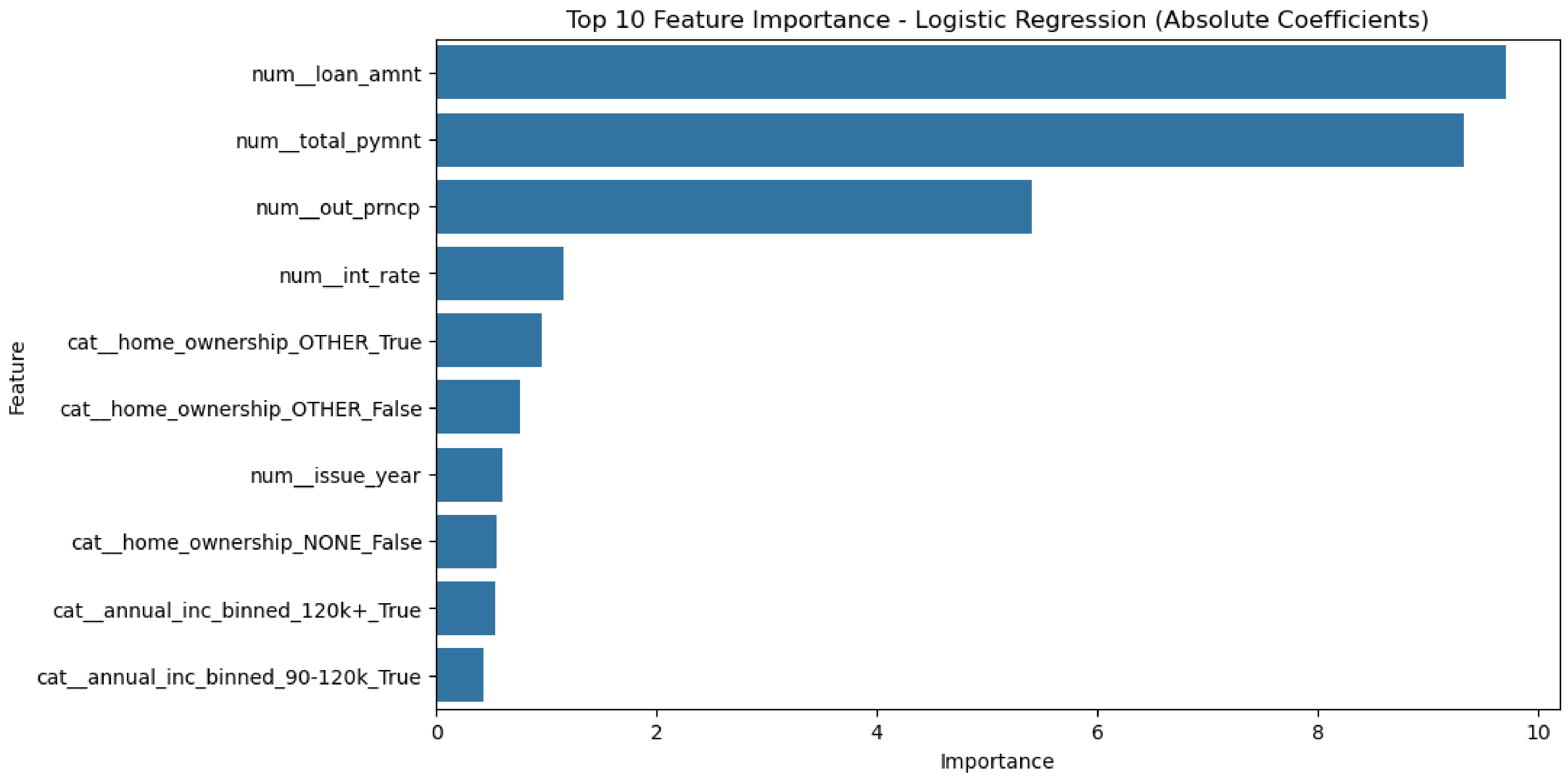


Sources

Feature Importance

Temporal Context

" The issue year suggests that the time the loan was issued is relevant. "



Conclusion

Logistic Regression offers interpretable insights on Loan Default Factors including :

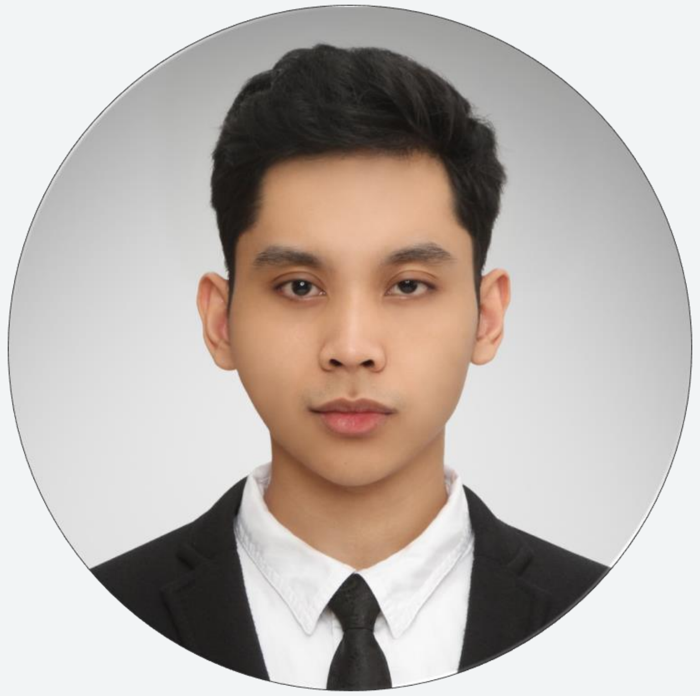
1. Borrower's Financial Stability
2. Loan Terms
3. Repayment Behavior

More on that here

While tree-based models demonstrate better predictive accuracy, it comes with overfitting issue.

Excited to work with us?

Reach Us!



Contact Information

- +62813-3353-2005
- brambsns@gmail.com
- linkedin.com/in/bramraka666/

Gerente General

id/x partners x Rakamin Academy

Gerente General

Sources (2008 Crisis)

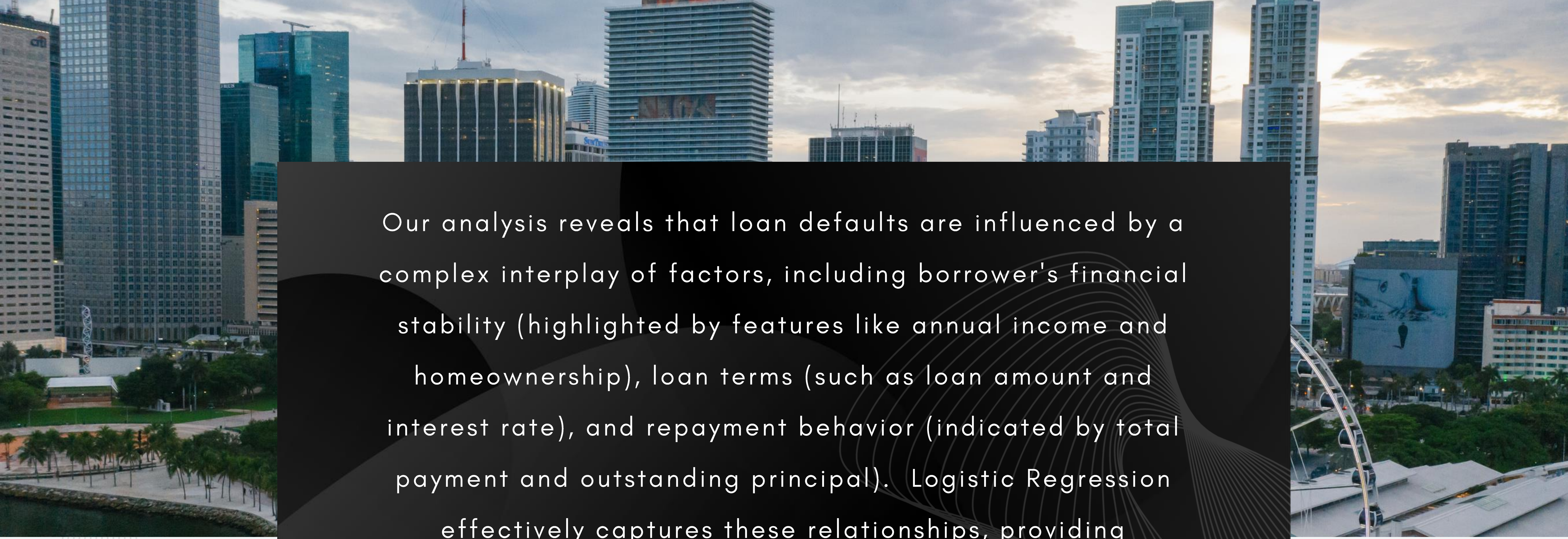
https://unctad.org/system/files/official-document/gdsmdp20101_en.pdf

https://www.govinfo.gov/content/pkg/GPO-FCIC/pdf/GPO-FCIC.pdf

Sources (Models)

"Logistic Regression, randomforest, XGBoost and AdaBoost are used to predict the loan default.These algorithms are common methods for binary classification problems."

https://www.shs-conferences.org/articles/shsconf/pdf/2024/01/shsconf_icdeba2023_02008.pdf



Our analysis reveals that loan defaults are influenced by a complex interplay of factors, including borrower's financial stability (highlighted by features like annual income and homeownership), loan terms (such as loan amount and interest rate), and repayment behavior (indicated by total payment and outstanding principal). Logistic Regression effectively captures these relationships, providing interpretable insights, while tree-based models (Random Forest, XGBoost) demonstrate higher predictive accuracy but with potential overfitting concerns.

Links

Presentation Video:

https://youtu.be/CeZbM6SgSqs

This Project Repository:

https://github.com/Bramasta66/IDX-Partners-VIX