Consumer Decision Analysis Using HMDA Data [Years 2020-2024]

AUTHOR: SAMIRA JAHANGIRI
ADVISOR: DR. LIQIANG NI

Objectives

1. Consumer Behavior Analysis
2. Predictive Modeling
3. Equity & Disparity Insights
4. Strategic Recommendations

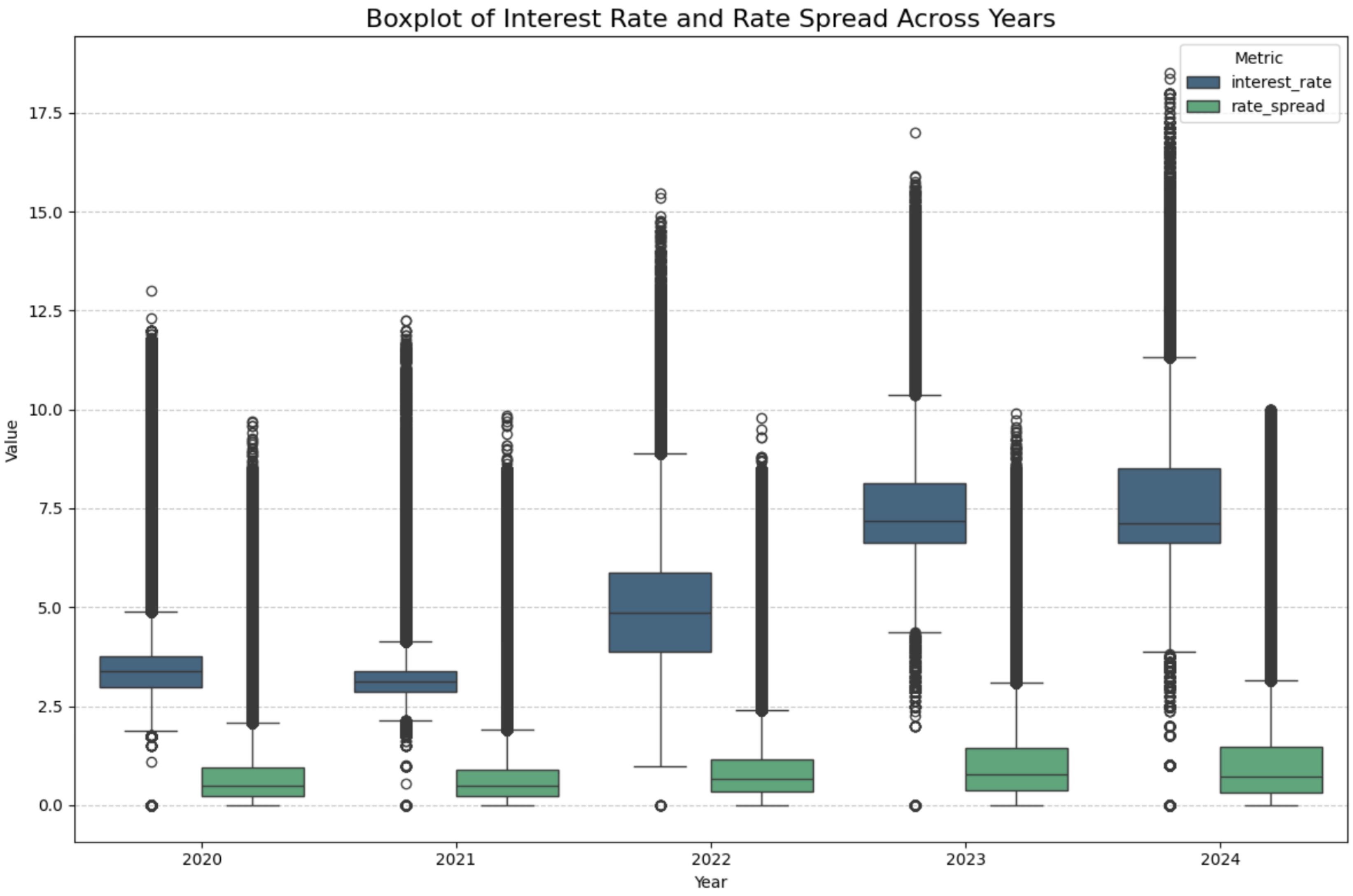
Home Mortgage Disclosure Act

Applicants, loan characteristics, and lender decisions

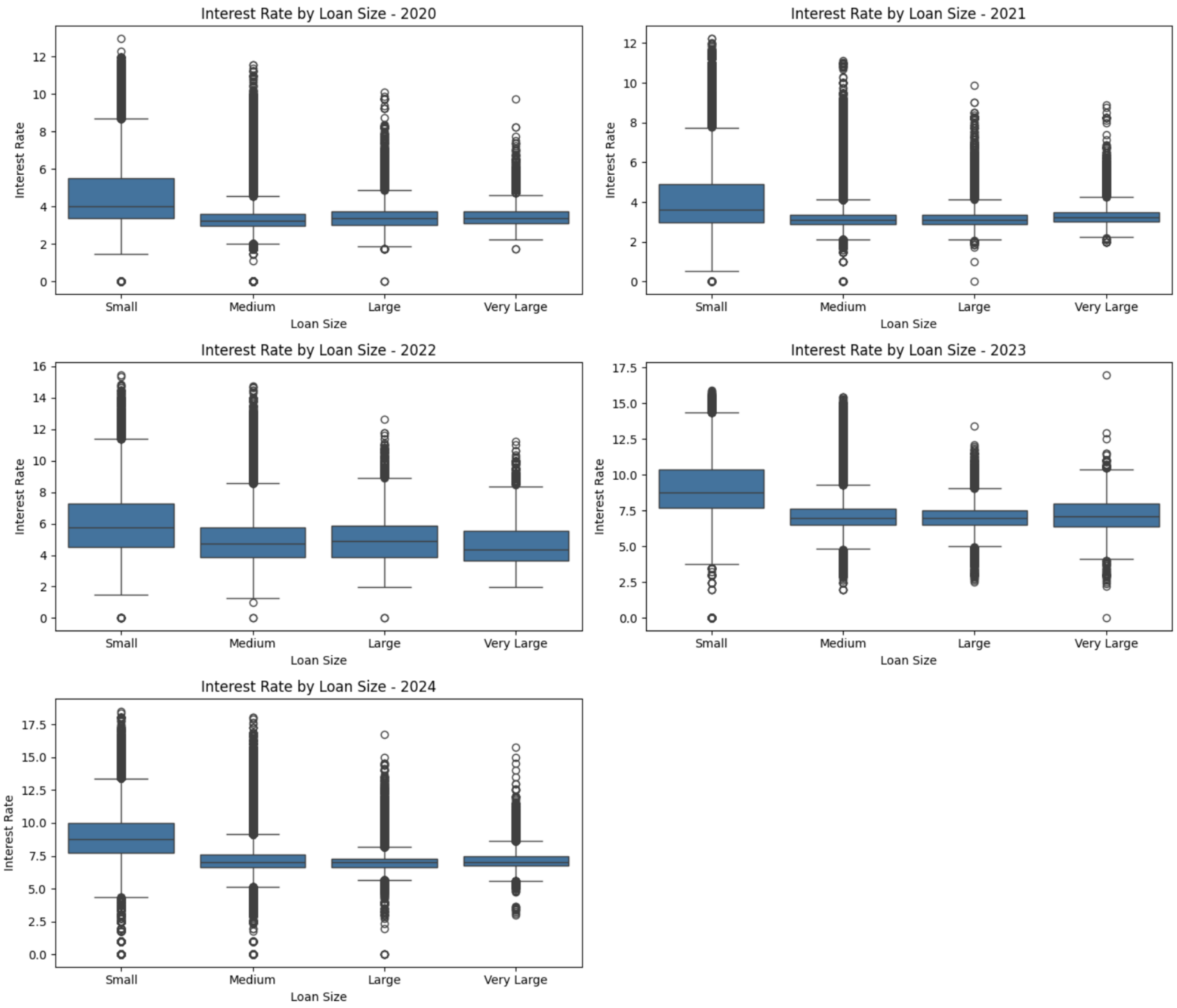
Loan amount, interest rates, applicant demographics (e.g., age, income), and loan purpose

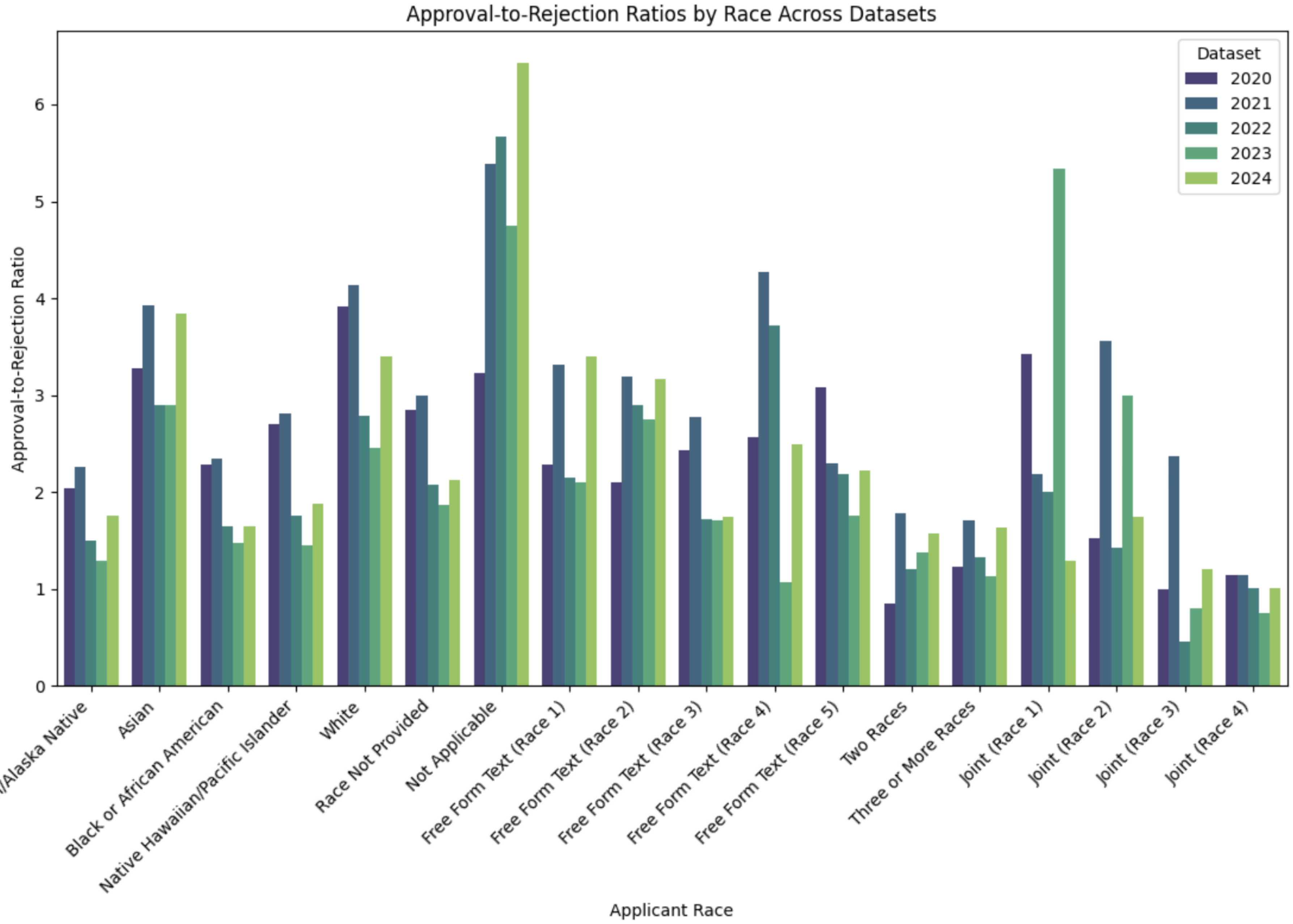
Regulatory compliance, fair lending assessments, and academic research

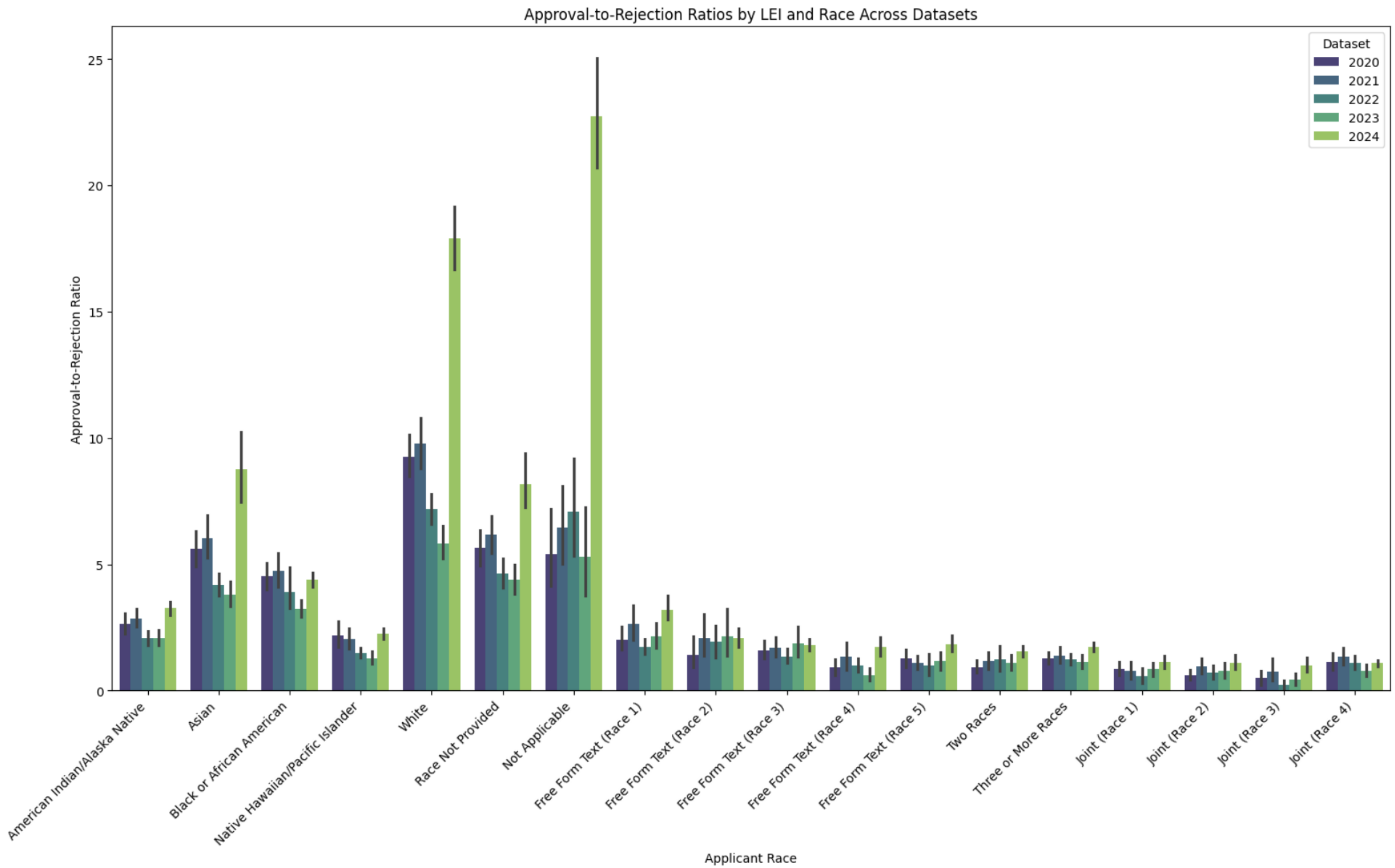
The boxplot shows the distribution of interest rates and rate spreads from 2020 to 2024. Interest rates remained relatively stable and low between 2020 and 2021, then significantly increased starting in 2022, peaking in 2023 and 2024. This trend reflects rising market rates over time. Meanwhile, rate spreads (e difference between offered and benchmark rates) remained consistently low and stable across all years, suggesting that while base rates increased, the relative markup by lenders did not change much. Overall, the data illustrates a clear upward shift in interest rates post-2021 while keeping lending spreads steady



The boxplots show how interest rates vary by loan size (as defined by quartiles of loan amount: Small, Medium, Large, Very Large) across 2020–2024. In each year, smaller loans consistently had higher interest rates than larger ones. This gap became more pronounced from 2022 onward as overall rates rose. Very large loans maintained relatively lower and more stable rates, while small loans showed wider variation and steeper increases. This reflects risk-based pricing, where smaller loans may be seen as riskier or less profitable by lenders.

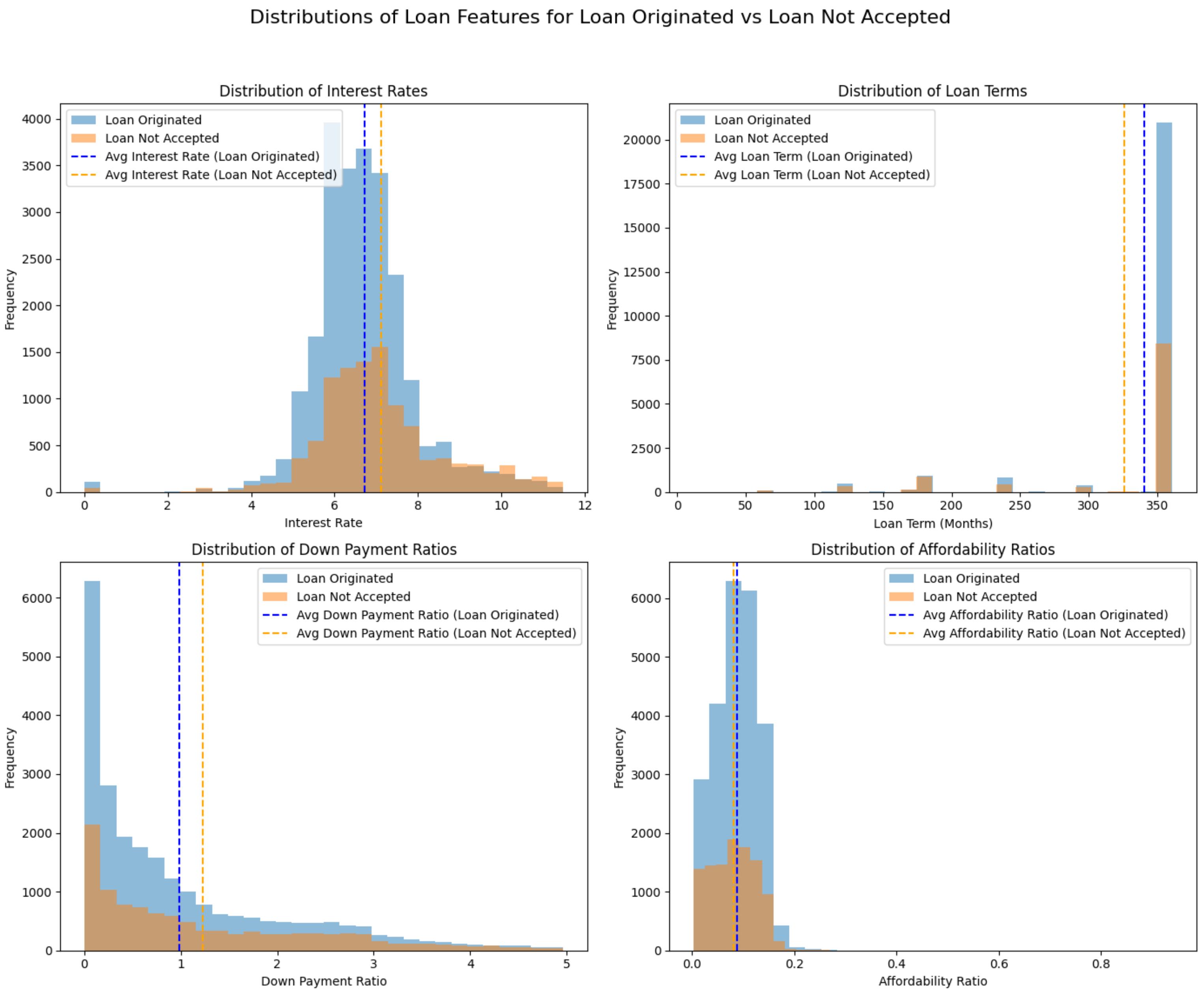



Approval-to-Rejection Ratios by Race Across Datasets
Dataset
2020
2021
2022
2023
2024
Approval-to-Rejection Ratio
0
1
2
3
4
5
6
American Indian/Alaska Native
Asian
Black or African American
Native Hawaiian/Pacific Islander
White
Race Not Provided
Not Applicable
Free Form Text (Race 1)
Free Form Text (Race 2)
Free Form Text (Race 3)
Free Form Text (Race 4)
Free Form Text (Race 5)
Two Races
Three or More Races
Joint (Race 1)
Joint (Race 2)
Joint (Race 3)
Joint (Race 4)
Applicant Race

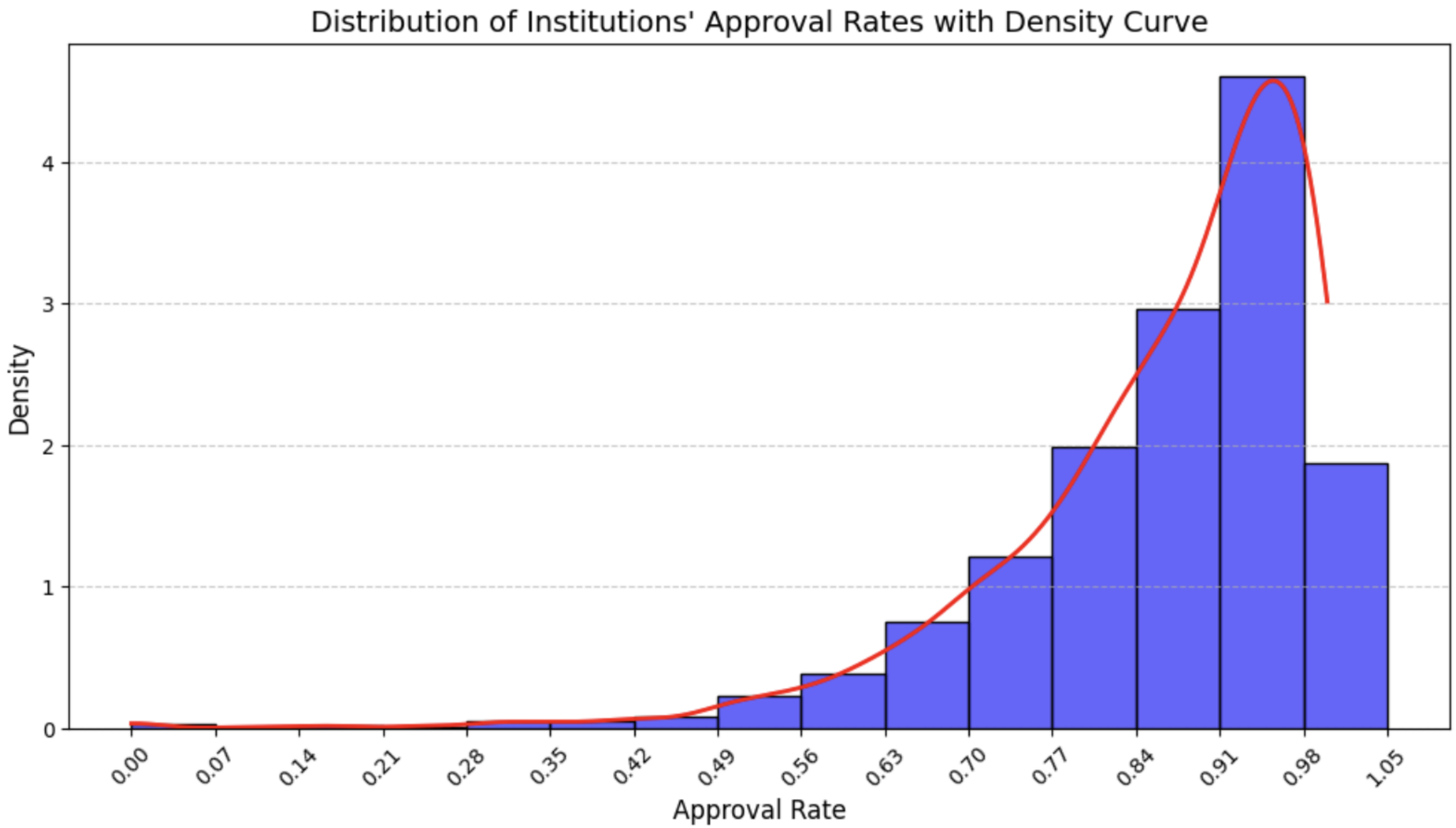


Approval-to-Rejection Ratios by LEI and Race Across Datasets
Approval-to-Rejection Ratio
0
5
10
15
20
25
Dataset
2020
2021
2022
2023
2024
American Indian/Alaska Native
Asian
Black or African American
Native Hawaiian/Pacific Islander
White
Race Not Provided
Not Applicable
Free Form Text (Race 1)
Free Form Text (Race 2)
Free Form Text (Race 3)
Free Form Text (Race 4)
Free Form Text (Race 5)
Two Races
Three or More Races
Joint (Race 1)
Joint (Race 2)
Joint (Race 3)
Joint (Race 4)
Applicant Race

HMDA loans were grouped into three categories: (1) Loan Originated (32,394), (2) Loan Approved but Not Accepted (8,223), and (3) Loan Denied (7,405) (after removing outliers or rows with missing interest rate and rate spread (I could regenerate the missing income, debt-to-income, loan-to-value, and there were not much missing on loan amount or property value)). The plots show feature distributions for Groups 1 and 2, which nearly overlap across all dimensions—interest rate, loan term, down payment ratio, and affordability ratio—indicating minimal difference in loan characteristics. Differences in acceptance behavior may thus stem from borrower-side factors. Group 3 (denied loans) is analyzed separately to assess denial-to-approval patterns across institutions (next slide).

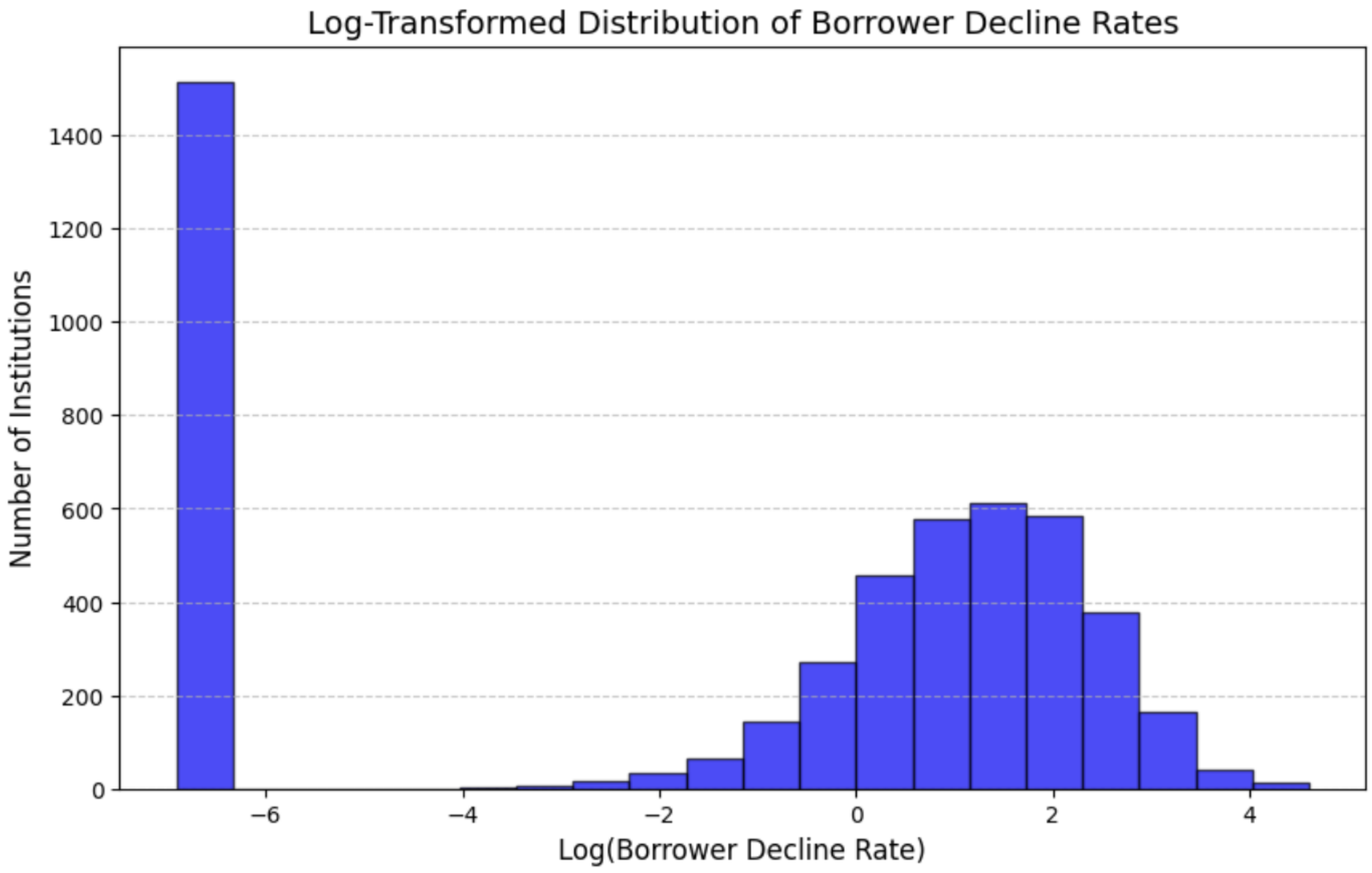


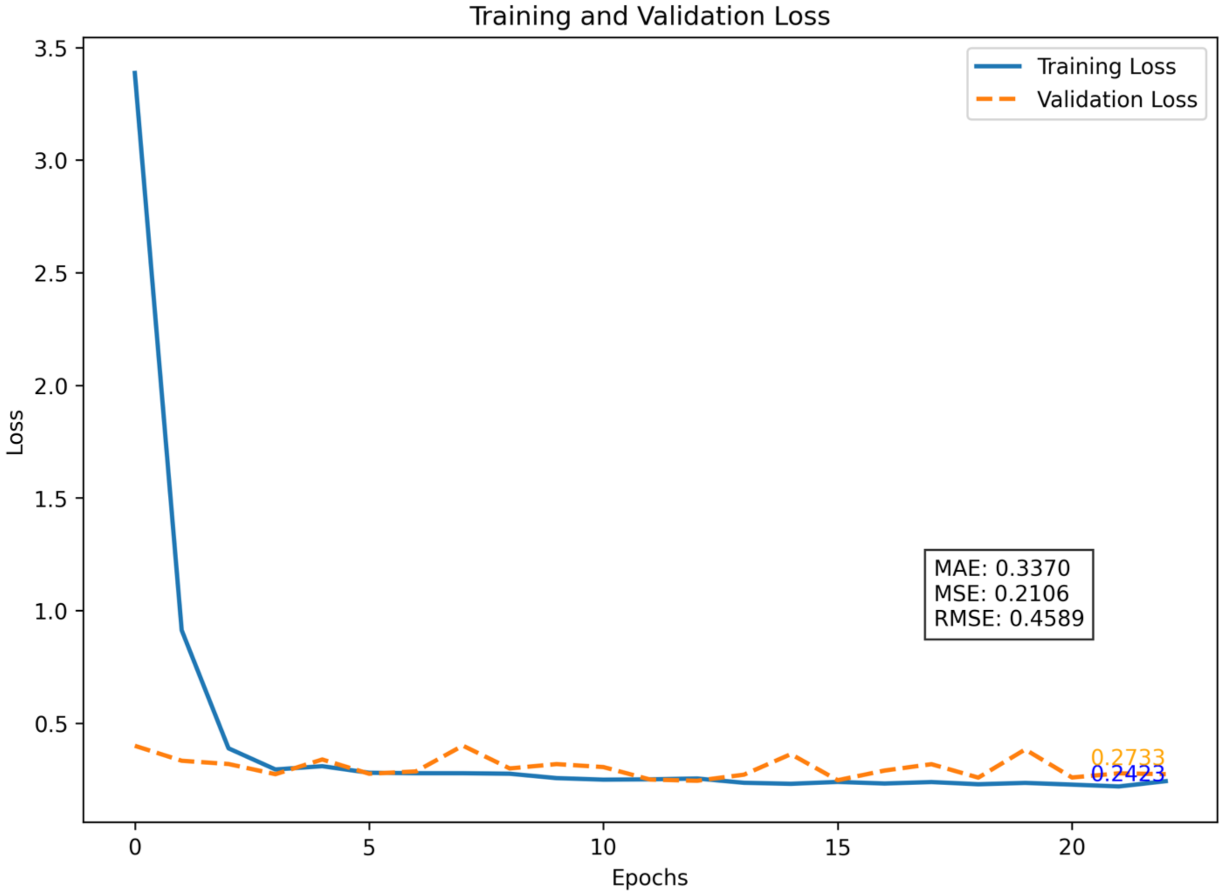
The plot shows the distribution of institutional approval rates for 2024, computed as the ratio of originated loans to total applications with known outcomes (originated + denied), excluding "approved but not accepted" cases to reflect lender decisions only. Out of 4,908 institutions analyzed, **63.63%** approved more than **85%** of applications, indicating a strong tendency toward approval. An additional **31.36%** approved between **60%–85%**, reflecting moderate selectiveness. The KDE curve confirms that most institutions cluster at high approval rates. Only **1.73%** (85 institutions) had over 25% of borrowers decline post-approval, suggesting limited issues with loan attractiveness across the market.

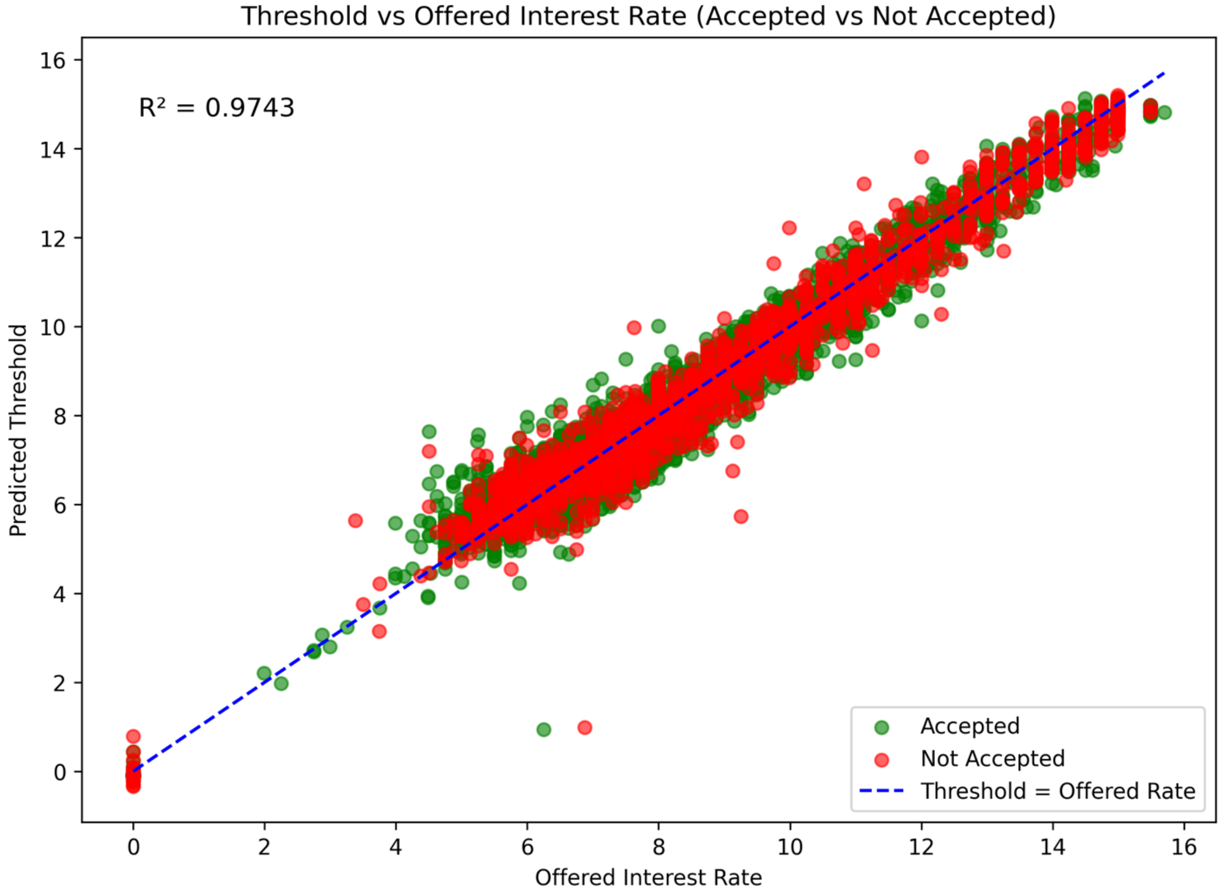


Here's the updated explanation based on your 2024 results:
We calculated each institution's borrower decline rate—the percentage of loans that were approved but not accepted—and applied a **log transformation** to highlight relative differences. The spike on the far left represents institutions with **zero or negligible decline rates**, where nearly all approved applicants accepted the offer. The right-skewed distribution captures institutions with increasing drop-off behavior.

Out of **4,908 total institutions**, only **1.73% (85 institutions)** had **25% or more of applicants decline** their approved loans, suggesting a small subset of lenders are consistently offering uncompetitive products. This analysis helps identify lenders with potential issues in loan attractiveness or consumer trust.







A neural network was trained to predict the interest rate threshold at which an applicant would accept an approved loan. The threshold was modeled as a function of loan and applicant features. If the predicted threshold was below the offered rate, the applicant likely accepted the loan (green); otherwise, they likely declined (red). The model achieved strong performance (MAE = 0.3370, RMSE = 0.4589, R^2 = 0.9743), and the separation between accepted and not accepted offers validates the effectiveness of this interest threshold framework.

Once we have
1- predicted threshold from the neural network
2- offered interest rate (actual from the lender)
3- observed client behavior (accepted or not accepted)
then we apply the decision rule:

$$\text{Predicted Acceptance} = \begin{cases} \text{Accepted,} & \text{if predicted threshold} < \text{offered rate} \\ \text{Declined,} & \text{otherwise} \end{cases}$$

Then compare it to the actual label to compute:

1- accuracy = (Correct predictions) / (Total samples)
2- precision for predicted accepted
3- recall for predicted declined

Nex slides show this classification results (using the predicted interest rate as threshold)

Interest Rate Prediction – Top 20 Observation

derived_loan_product_type	loan_type	loan_purpose	age	term	property_value	loan_amount	income	rate_spread	predicted_IR	IR	action	dp_ratio	affordability_ratio	down	risk_factor	normalized_rf	approval_rate	not_accepted_rate
FHA:First Lien	FHA-insured	Home purchase	50	360	325000	315000	69000	-0.18	5.63	5.63	1	0.03	6.57	10000	150	0.90	0.89	0.15
Conventional:First Lien	Conventional	Home purchase	70	360	315000	235000	107000	1.04	7.37	7.38	0	0.25	13.66	80000	119	0.70	0.92	0.16
Conventional:First Lien	Conventional	Home purchase	40	360	555000	445000	1000	1.78	8.12	8.13	1	0.20	0.07	110000	135	0.81	0.70	0.24
FHA:First Lien	FHA-insured	Home purchase	50	360	195000	185000	68000	1.92	7.50	7.50	0	0.05	11.03	10000	152	0.91	0.88	0.24
Conventional:First Lien	Conventional	Home purchase	70	360	235000	195000	69000	2.25	8.13	8.13	1	0.17	10.62	40000	132	0.79	0.89	0.15
Conventional:First Lien	Conventional	Home purchase	40	360	355000	245000	1000	2.11	8.75	8.75	1	0.31	0.12	110000	80	0.45	0.96	0.42
Conventional:Subordinate Lien	Conventional	Home improvement	60	180	695000	105000	155000	6.96	14.74	14.74	1	0.85	22.14	590000	112	0.66	0.98	0.57
Conventional:First Lien	Conventional	Home purchase	50	360	335000	245000	1000	1.35	7.88	7.88	1	0.27	0.12	90000	83	0.47	0.89	0.15
Conventional:First Lien	Conventional	Home purchase	40	360	385000	245000	1000	1.87	8.50	8.50	0	0.36	0.12	140000	75	0.42	0.96	0.42
FHA:First Lien	FHA-insured	Home purchase	30	360	385000	375000	118000	0.16	6.12	6.13	1	0.03	9.44	10000	144	0.86	0.89	0.15
Conventional:First Lien	Conventional	Home purchase	60	360	605000	485000	133000	0.06	6.56	6.56	1	0.20	8.23	120000	117	0.69	0.92	0.16
VA:First Lien	VA-guaranteed	Cash-out refinancing	75	360	535000	525000	85000	-1.48	5.00	5.00	1	0.02	4.86	10000	153	0.92	0.70	0.53
Conventional:First Lien	Conventional	Home purchase	40	360	295000	205000	251000	2.68	9.00	9.00	1	0.31	36.73	90000	80	0.45	0.70	0.24
Conventional:First Lien	Conventional	Cash-out refinancing	50	360	185000	115000	93100	2.33	8.88	8.88	1	0.38	24.29	70000	117	0.69	0.96	0.42
FHA:First Lien	FHA-insured	Home purchase	50	360	305000	305000	160000	-1.65	5.50	5.50	0	0.00	15.74	0	122	0.72	0.88	0.20
FHA:First Lien	FHA-insured	Home purchase	40	360	425000	415000	132000	1.16	7.12	7.13	1	0.02	9.54	10000	152	0.91	0.92	0.16
Conventional:First Lien	Conventional	Home purchase	75	300	95000	85000	64000	2.63	8.50	8.50	1	0.11	18.82	10000	132	0.79	0.43	0.61
Conventional:First Lien	Conventional	Home purchase	60	360	285000	255000	96000	0.65	6.88	6.88	1	0.11	11.29	30000	139	0.83	0.88	0.20
Conventional:First Lien	Conventional	Home purchase	40	360	295000	215000	89000	-0.03	6.50	6.50	1	0.27	12.42	80000	114	0.67	0.92	0.16
Conventional:Subordinate Lien	Conventional	Other purpose	50	120	465000	65000	170000	5.48	11.60	11.60	0	0.86	26.15	400000	109	0.64	0.63	0.27

Model Performance Summary – Classification Based on Predicted Threshold

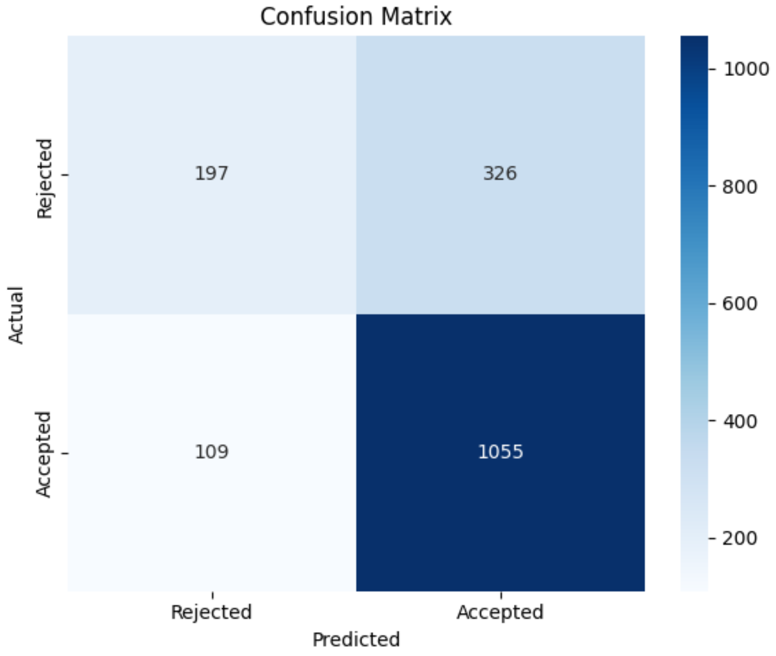
•Confusion Matrix Breakdown:

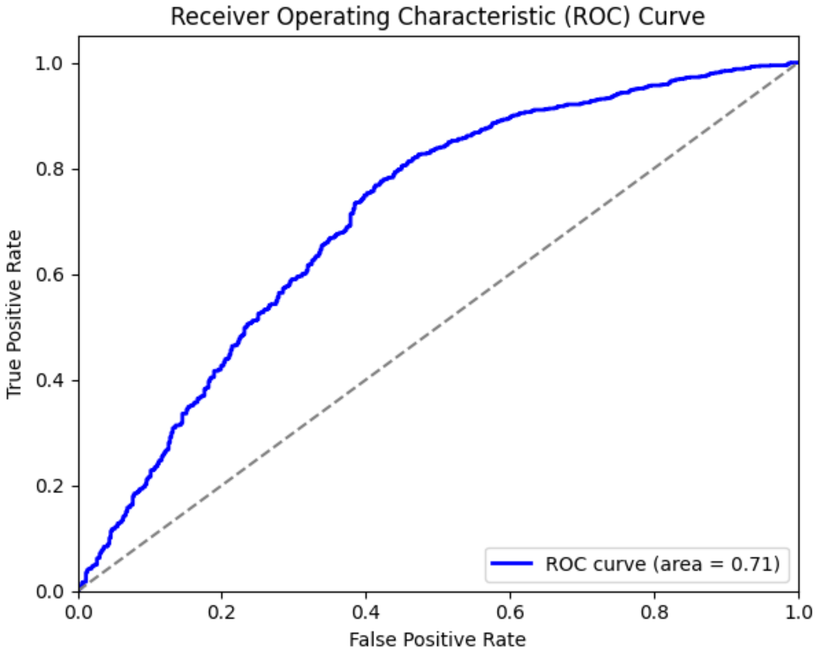
- True Positives (Accepted correctly predicted): 1,055
- True Negatives (Rejected correctly predicted): 197
- False Positives (Rejected predicted as accepted): 326
- False Negatives (Accepted predicted as rejected): 109

•Key Observations:

- The model tends to overpredict acceptance, as shown by the high number of false positives.
- False negatives are lower, meaning fewer eligible applicants were wrongly rejected.
- The ROC AUC = 0.71 indicates moderate discrimination ability between accepted and rejected applications.

While the model captures client decision behavior reasonably well, adjusting the decision threshold could improve accuracy and reduce risk for lenders.





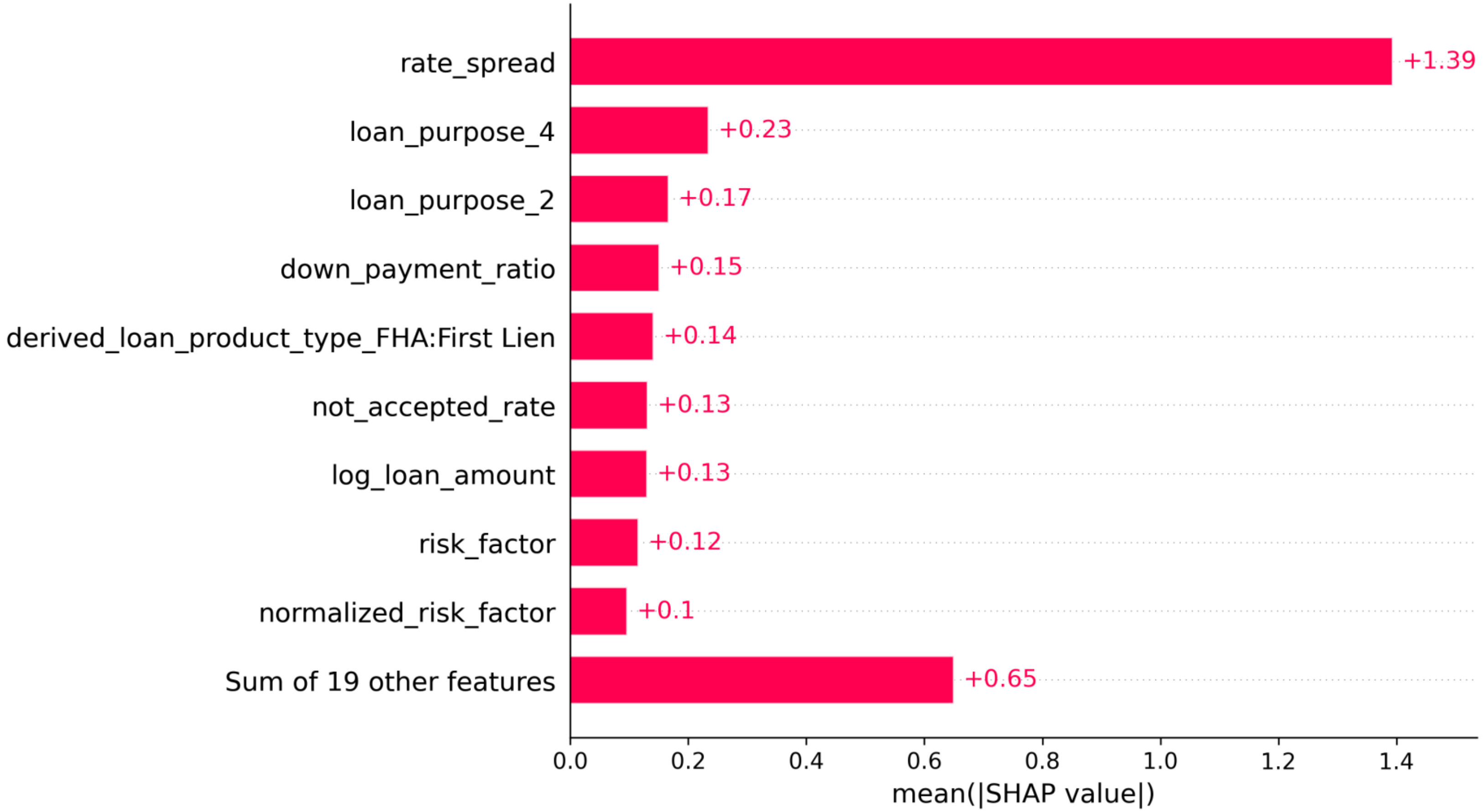
The SHAP analysis shows that

1- Rate_spread is by far the most influential feature in predicting loan acceptance, with a SHAP value exceeding +1.3, indicating it heavily drives the model's output.

2- Loan purpose variables (loan_purpose_4, loan_purpose_2) and down_payment_ratio also contribute meaningfully to the prediction.

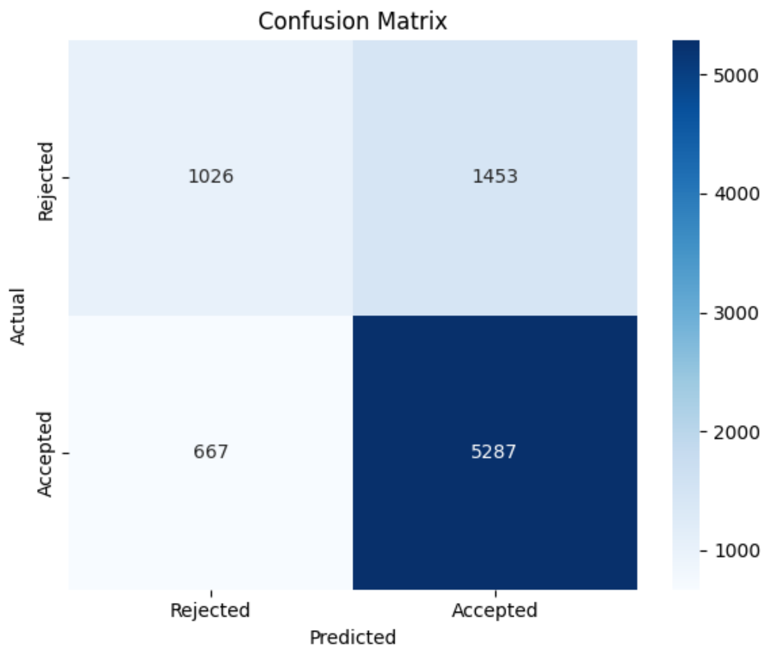
3- Features such as derived_loan_product_type_FHA:First Lien, not_accepted_rate, and log_loan_amount show moderate influence, shaping the model's confidence in whether a loan will be accepted.

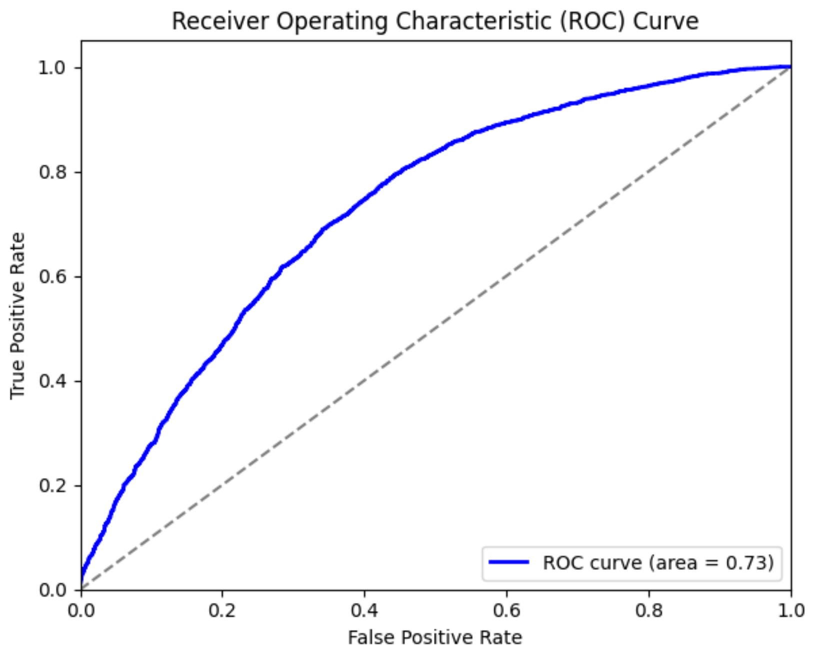
4- Additionally, risk-related factors like risk_factor and normalized_risk_factor play smaller but still important roles. Overall, the SHAP results highlight that both financial terms and institutional patterns influence the model's loan acceptance predictions.



This neural network was trained to directly classify whether a client would accept or reject an approved loan, based solely on application and loan data, without estimating any threshold.
The confusion matrix shows strong model performance with:
•**True Positives (Accepted predicted as Accepted):** 5,287
•**True Negatives (Rejected predicted as Rejected):** 1,026
•**False Positives:** 1,453 (rejected predicted as accepted)
•**False Negatives:** 667 (accepted predicted as rejected)

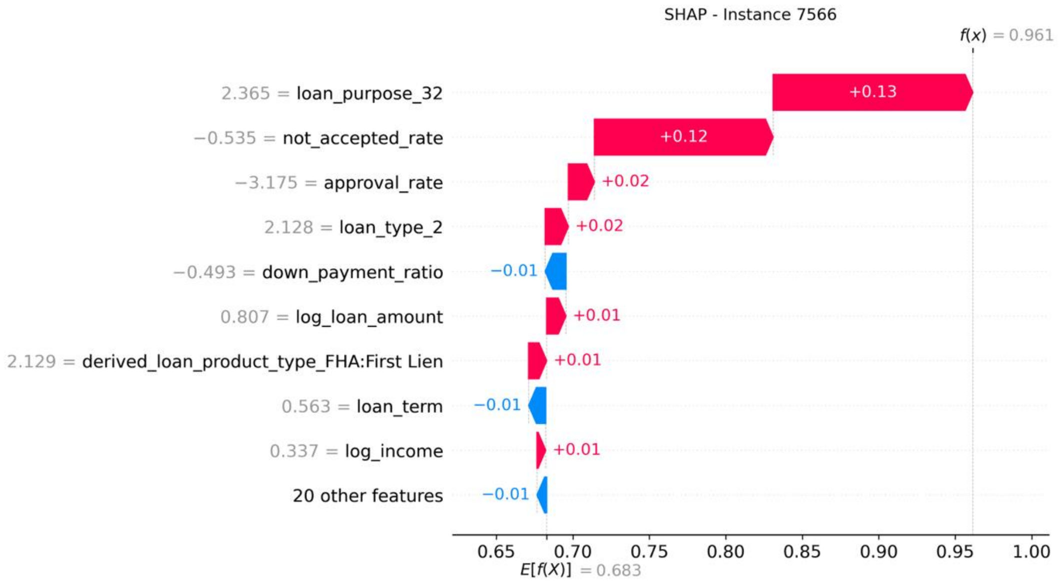
Despite the imbalance, the model demonstrates good separation power, with an ROC AUC of **0.73**, indicating improved discrimination ability over the threshold-based approach. The results show that the model can learn behavioral patterns directly from input data to predict loan acceptance decisions.

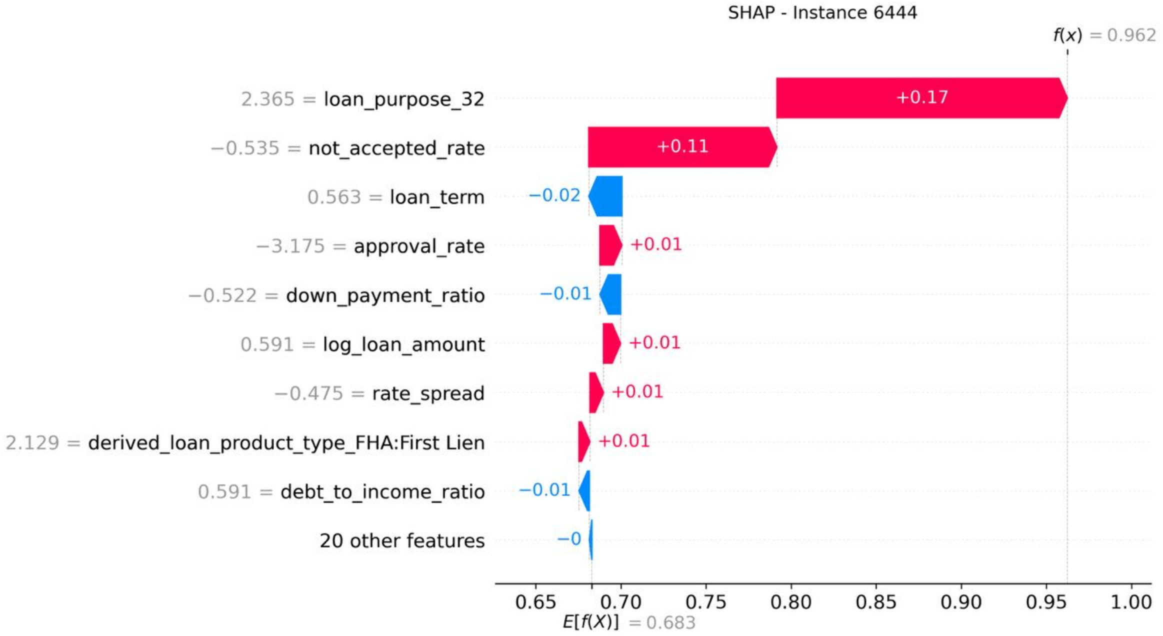




Decision Prediction – Observations With Highest Probability of Accepting the Approved Loan – SHAP Analysis

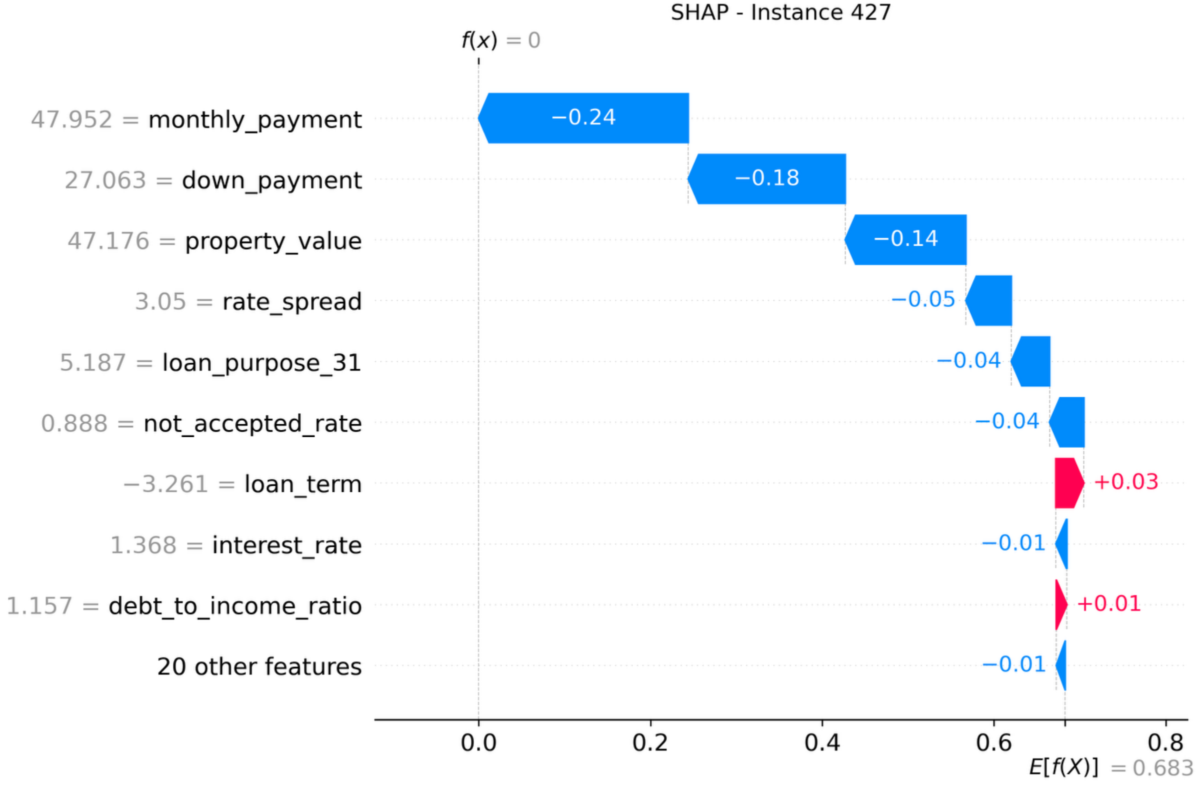
loan_product_type	loan_type	loan_purpose	rate_spread	age	term	property_value	loan_amount	interest_rate	probability	action	income	down	affordability_ratio	risk_factor	normalized_rf	approval_rate	not_accepted_rate	down_payment_ratio
FHA:First Lien	FHA-insured	Cash-out refinancing	0.11	40	360	505000	385000	6.75	0.9614931	1	131000	120000	10.21	122.80	0.73	0.24	0.21	0.24
FHA:First Lien	FHA-insured	Cash-out refinancing	-0.42	40	360	355000	285000	6.375	0.9619256	1	100000	70000	10.53	135.00	0.81	0.24	0.21	0.20

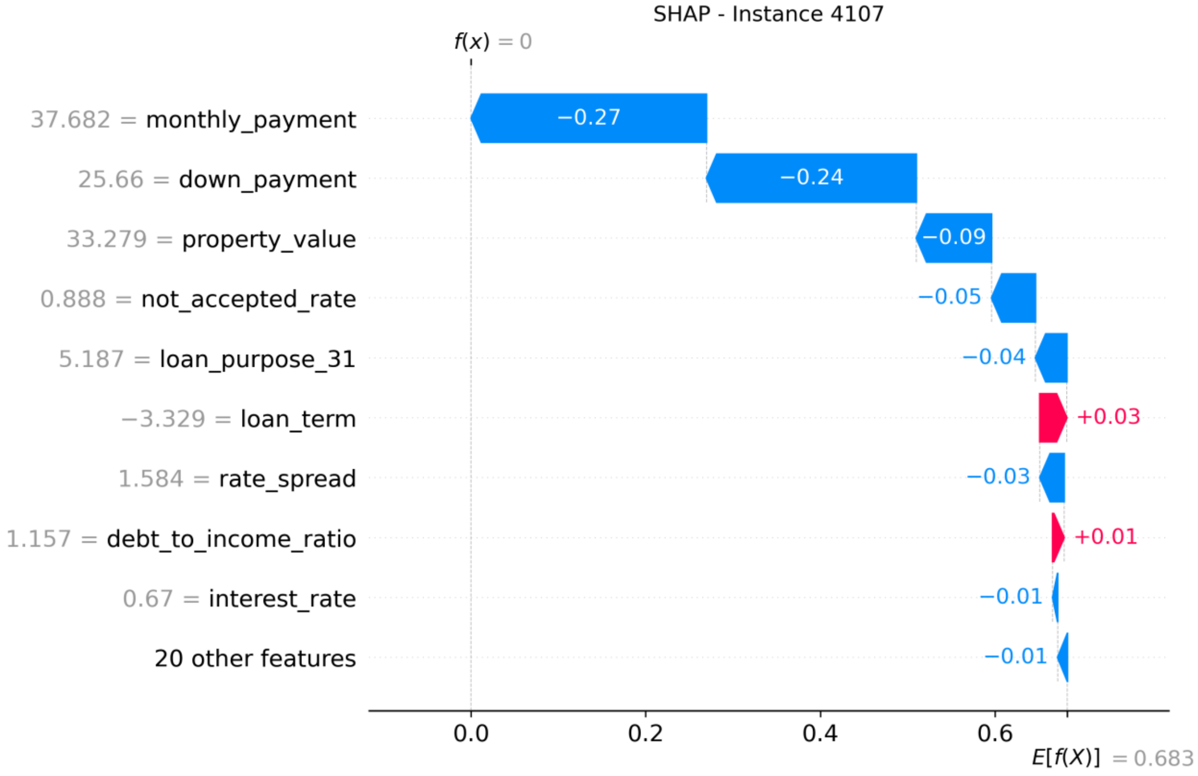




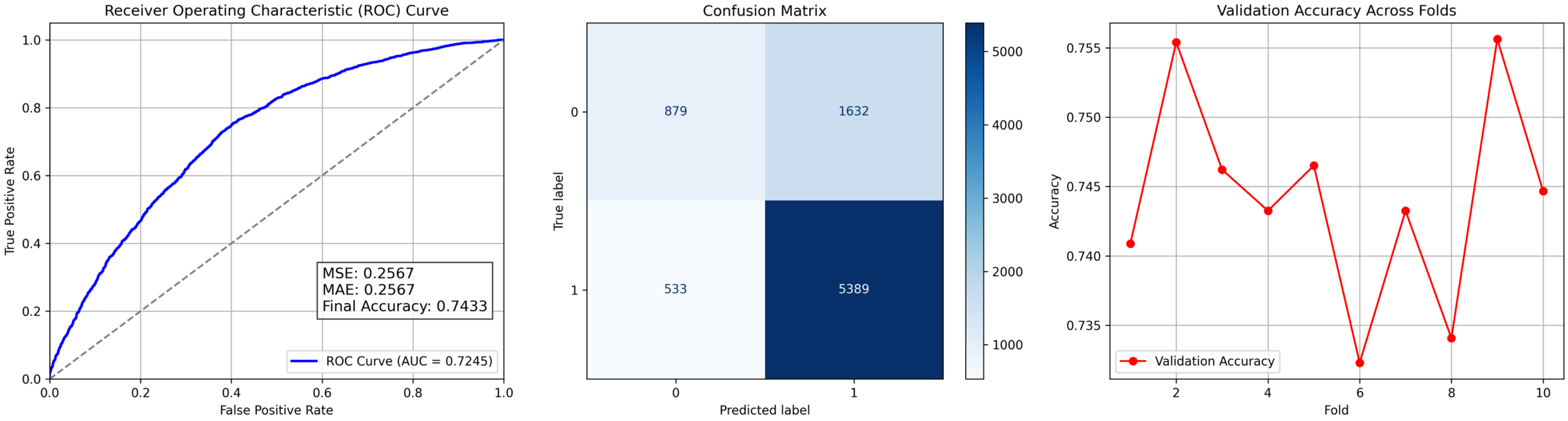
Decision Prediction – Observations With Lowest Probability of Accepting the Approved Loan – SHAP Analysis

loan_product_type	loan_type	loan_purpose	rate_spread	age	term	property_value	loan_amount	interest_rate	probability	action	income	down	affordability_ratio	risk_factor	normalized_rf	approval_rate	not_accepted_rate	down_payment_ratio
Conventional:First Lien	Conventional	Refinancing	9.61	50	24	80715000	54255000	12	2.19E-05	0	93100	26460000	0.00	122.22	0.72	0.94	0.43	0.33
Conventional:First Lien	Conventional	Refinancing	5.75	50	18	57085000	31985000	9.99	0.00018268	0	85900	25100000	0.00	111.03	0.65	0.94	0.43	0.44





Decision Prediction – Random Forest



This model uses a Random Forest classifier trained directly on application and loan data to predict whether a client would accept or decline an approved loan. Cross-validation was applied to assess generalization, with validation accuracy ranging from 0.735 to 0.755 across folds. The final test accuracy was 0.7433, and the ROC AUC reached 0.7245, indicating solid discriminative power. The confusion matrix shows 5,389 true positives and 879 true negatives, with 1,632 false positives and 533 false negatives. Overall, the model captures client behavior reasonably well and performs comparably to the neural network classifier while offering better interpretability through ensemble-based feature splits.

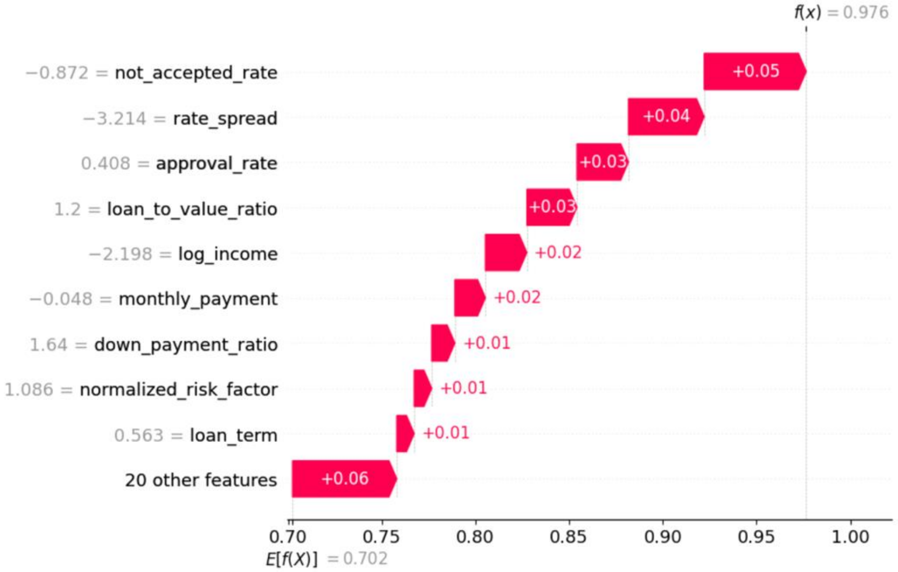
Applicant Loan Acceptance Prediction Using Random Forest

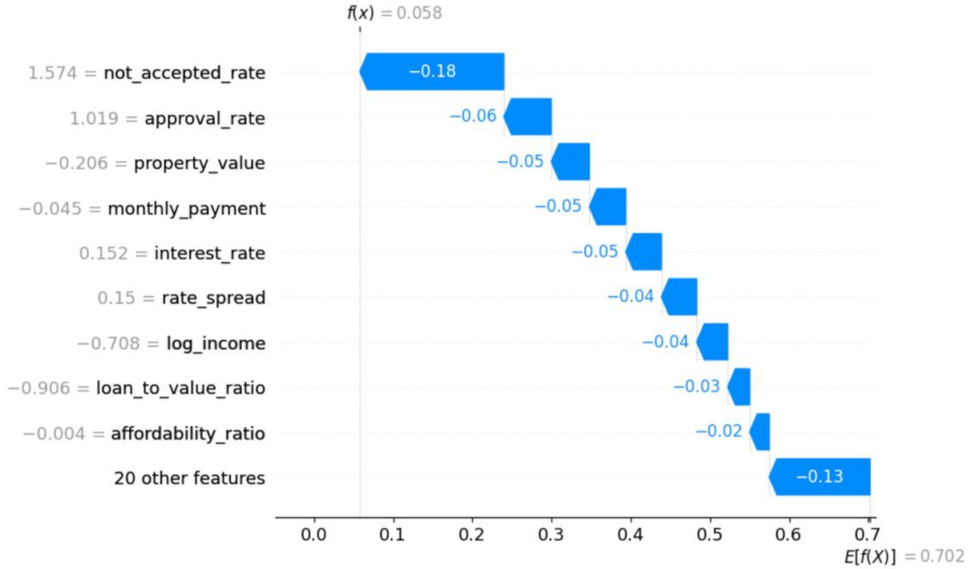
5 highest and lowest probabilities

Rank	derived_loan_product_type	loan_type	loan_purpose	rate_spread	applicant_age	loan_term	Property value	Interest rate	down	Risk factor	normalized_risk_factor	approval_rate	Not accepted rate	down_payment_ratio	affordability_ratio	predicted_probability	action_taken	loan_amount	income
1	Conventional:First Lien	Conventional	Refinancing	2.40	50	9	365000	9.49	150000	113	0.66	0.94	0.43	0.41	0.23	0.98	1	215000	65800
2	Conventional:First Lien	Conventional	Refinancing	2.41	50	9	365000	9.49	150000	115	0.67	0.94	0.43	0.41	0.23	0.98	1	215000	65800
3	Conventional:Subordinate Lien	Conventional	Home purchase	-6.91	50	360	355000	0.00	340000	144	0.86	0.89	0.15	0.96	150.00	0.98	1	15000	75000
4	Conventional:Subordinate Lien	Conventional	Home purchase	-6.96	50	360	285000	0.00	270000	147	0.88	0.89	0.15	0.95	148.00	0.98	1	15000	74000
5	Conventional:Subordinate Lien	Conventional	Home purchase	-7.18	40	360	345000	0.00	330000	148	0.89	0.89	0.15	0.96	164.00	0.98	1	15000	82000
6	Conventional:First Lien	Conventional	Cash-out refinancing	1.96	50	360	145000	8.50	70000	110	0.64	1.00	0.54	0.48	35.84	0.06	0	75000	89600
7	Conventional:First Lien	Conventional	Cash-out refinancing	1.97	50	360	145000	8.50	70000	111	0.65	1.00	0.54	0.48	35.84	0.06	0	75000	89600
8	Conventional:First Lien	Conventional	Cash-out refinancing	1.97	50	360	145000	8.50	70000	111	0.65	1.00	0.54	0.48	35.84	0.06	0	75000	89600
9	Conventional:First Lien	Conventional	Cash-out refinancing	1.97	50	360	145000	8.50	70000	111	0.65	1.00	0.54	0.48	35.84	0.06	0	75000	89600
10	Conventional:First Lien	Conventional	Cash-out refinancing	1.97	50	360	145000	8.50	70000	111	0.65	1.00	0.54	0.48	35.84	0.06	0	75000	89600

Observations with highest and lowest probabilities of acceptance (Rank 1 and 6 in the table above)

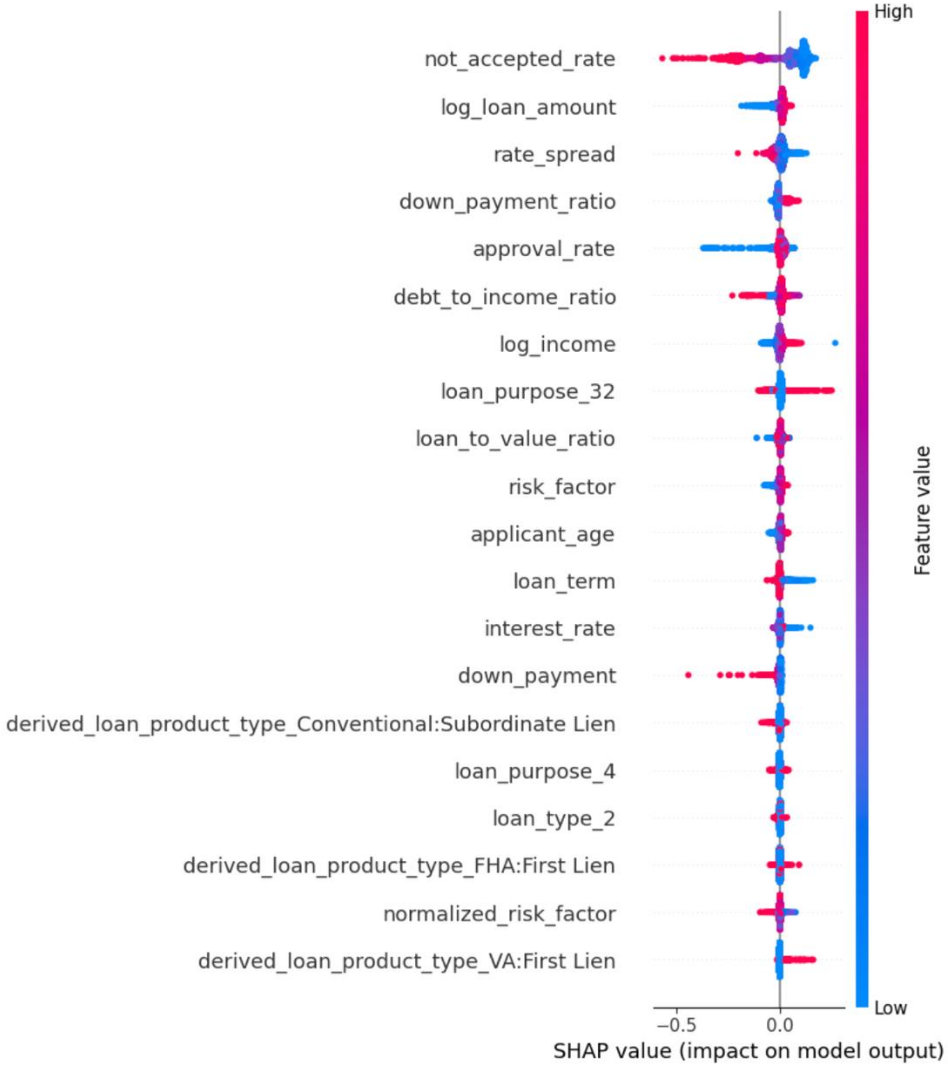
The first applicant (high probability f(x)=0.976) was **strongly favored** for accepting the loan offer, mainly by terms offered the institution might have offered (even though the institution has a high rate of applicants not accepting the loan from them) (not_accepted_rate **(+0.03)**) and **loan-to-value ratio (+0.03)**, institution **approval rate (94 %) (+0.03)**, and **rate spread (2.40) (+0.04)**, which increased the predicted probability.
However, **not_accepted_rate (-0.87)** and **rate_spread (-3.21)** had negative effects but were outweighed by positive contributions.
The second applicant (low probability f(x)=0.058f(x) = 0.058) was **strongly disfavored**, with **not_accepted_rate (-0.18)** (54% of applicant did not accept the loan from this corresponding institution, and this negatively impacted the decision), **log_income (-0.71)**, and **loan-to-value ratio (-0.91)** reducing the probability of accepting the loan. Even though **approval_rate** of the institution (100%) **(+0.06)** had a positive influence, it wasn't enough to offset the negative factors.
Overall, **not_accepted_rate** played a crucial role in both cases, but the high-probability applicant had mitigating positive contributions, while the low-probability applicant had overwhelmingly negative influences.



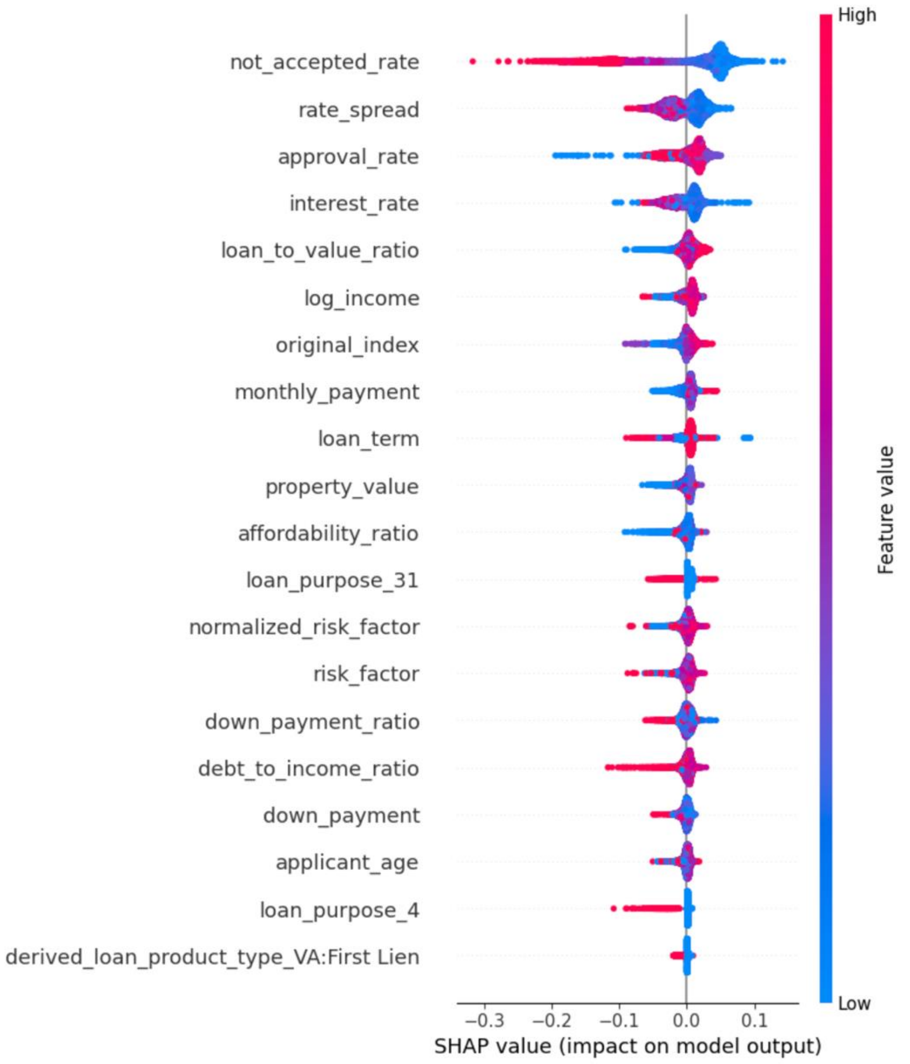


SHAP Summary Plots for Both Neural Network and Random Forest Models

The SHAP summary plot for the **neural network (NN)** shows a smoother distribution of feature impacts, with **not_accepted_rate, log_loan_amount, and rate_spread** having strong influences, indicating a more continuous learning process. The **random forest** (RF) plot, while similar in key feature importance, shows sharper boundaries in SHAP values, reflecting the model's reliance on discrete tree-based decisions. Both models highlight **not_accepted_rate** as the most critical factor, but RF appears to assign more extreme negative impacts, while NN spreads its effects more gradually across features.



Neural Network



Random Forest

References

- FFIEC – Home Mortgage Disclosure Act:
https://ffiec.cfpb.gov/data-browser/data/2023?category=states&items=FL

- SHAP Python Package:
https://shap.readthedocs.io/en/latest/index.html

- TensorFlow Python Library:
https://www.tensorflow.org/

- Scikit-learn (sklearn) Python Library:
https://scikit-learn.org/stable/

Thank You!