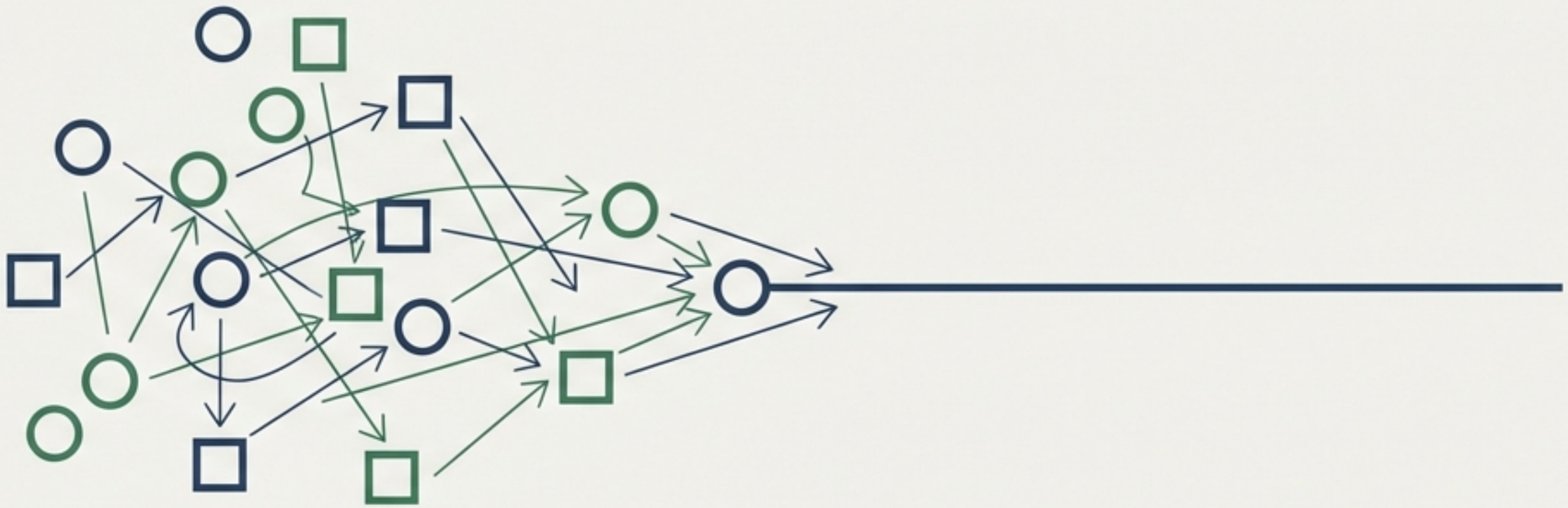
Loan Default Prediction Using Machine Learning

A Comparative Analysis of Classification Models for Credit Risk Assessment



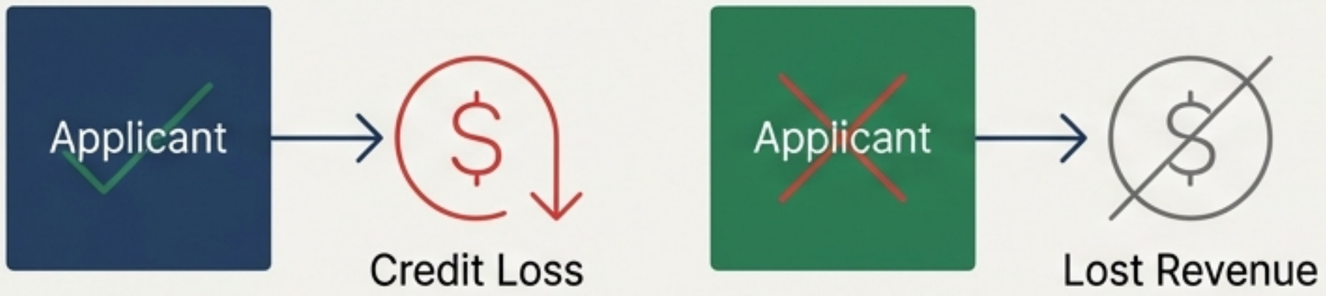
- An academic project based on the analysis of borrower demographic, financial, and credit-related data.
- The study follows a structured pipeline of exploratory data analysis, feature engineering, and comparative model evaluation to identify the most suitable model for loan default prediction.

Group 1: Pragya Gupta RBA55, Sahil Jain RBA39, Aman Dwivedi RBA24, Krishna Agrawal RBA23
Date: January 3, 2026

The Business Imperative: Mitigating Loan Default Risk

The Business Problem

- **Financial Risk:** Loan defaults represent a primary source of credit loss for financial institutions, directly impacting profitability and regulatory capital.
- **Cost of Misclassification:**
  - **Approving a 'bad' loan (False Negative)** leads to direct financial loss.
  - **Rejecting a 'good' loan (False Positive)** results in lost revenue and customer opportunity.
- **Challenge:** Traditional linear models may fail to capture complex, non-linear borrower behaviors, limiting predictive accuracy.



Project Objectives

1. **Predict**: Develop robust machine learning models to accurately predict loan default status (a binary classification task).
2. **Compare**: Systematically evaluate the performance of linear, regularized, and tree-based ensemble models using business-critical metrics.
3. **Identify**: Uncover the key risk drivers that most significantly influence loan default outcomes.

Understanding the Data: Borrower and Loan Attributes

Dataset Snapshot

- **Observations:** 45,000 loan records
- **Original Variables:** 14 features
- **Data Types:** 6 float, 3 integer, 5 object (categorical)
- **Target Variable (loan_status):**
  - 0 = Loan Approved / Non-Default
  - 1 = Loan Rejected / Default

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45000 entries, 0 to 44999
Data columns (total 14 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   person_age                  45000 non-null  float64
 1   person_gender               45000 non-null  object
 2   person_education            45000 non-null  object
 3   person_income               45000 non-null  float64
 4   person_emp_exp              45000 non-null  int64
 5   person_home_ownership       45000 non-null  object
 6   loan_amnt                   45000 non-null  float64
 7   loan_intent                 45000 non-null  object
 8   loan_int_rate               45000 non-null  float64
 9   loan_percent_income         45000 non-null  float64
 10  cb_person_cred_hist_length  45000 non-null  float64
 11  credit_score                45000 non-null  int64
 12  previous_loan_defaults_on_file  45000 non-null  object
 13  loan_status                 45000 non-null  int64
dtypes: float64(6), int64(3), object(5)
memory usage: 4.8+ MB
```

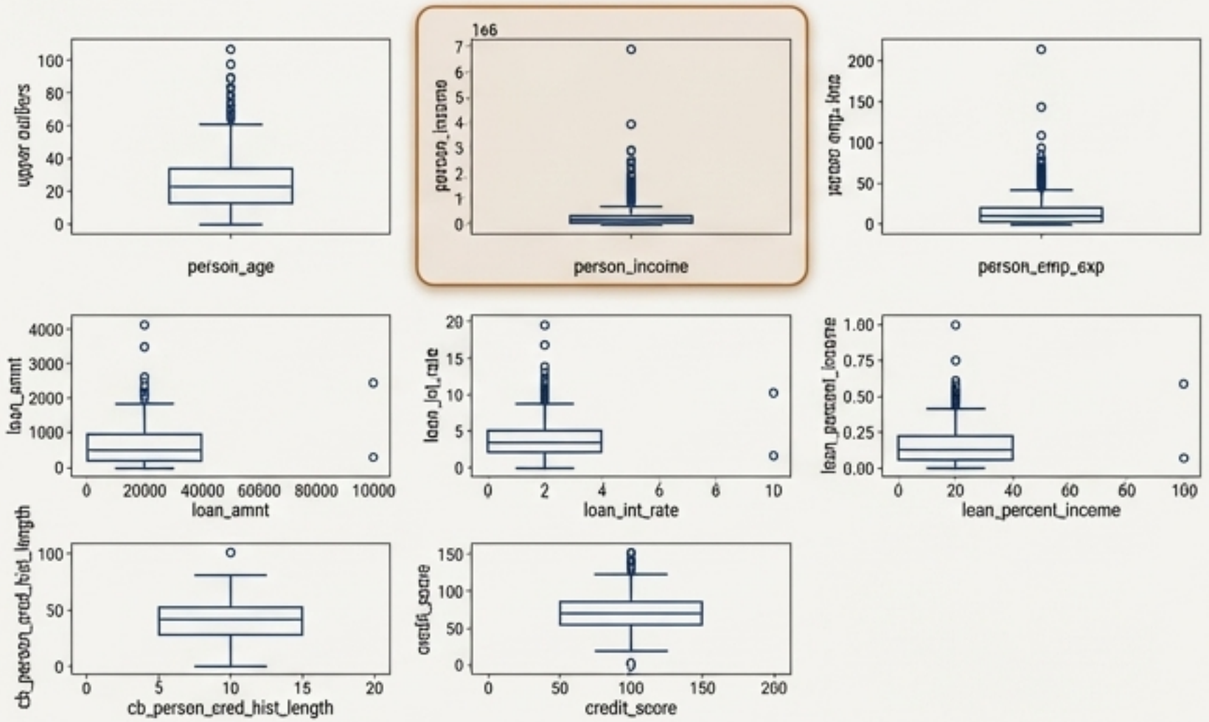
Key Original Variables & Business Interpretation

Variable Name	Business Interpretation
person_income	Core measure of repayment capacity
credit_score	Primary indicator of past credit behavior
loan_percent_income	Measures repayment burden on the borrower
loan_int_rate	Higher rates signal higher perceived credit risk
previous_loan_defaults_on_file	Strong predictor of future default risk
person_home_ownership	Proxy for financial stability and asset backing

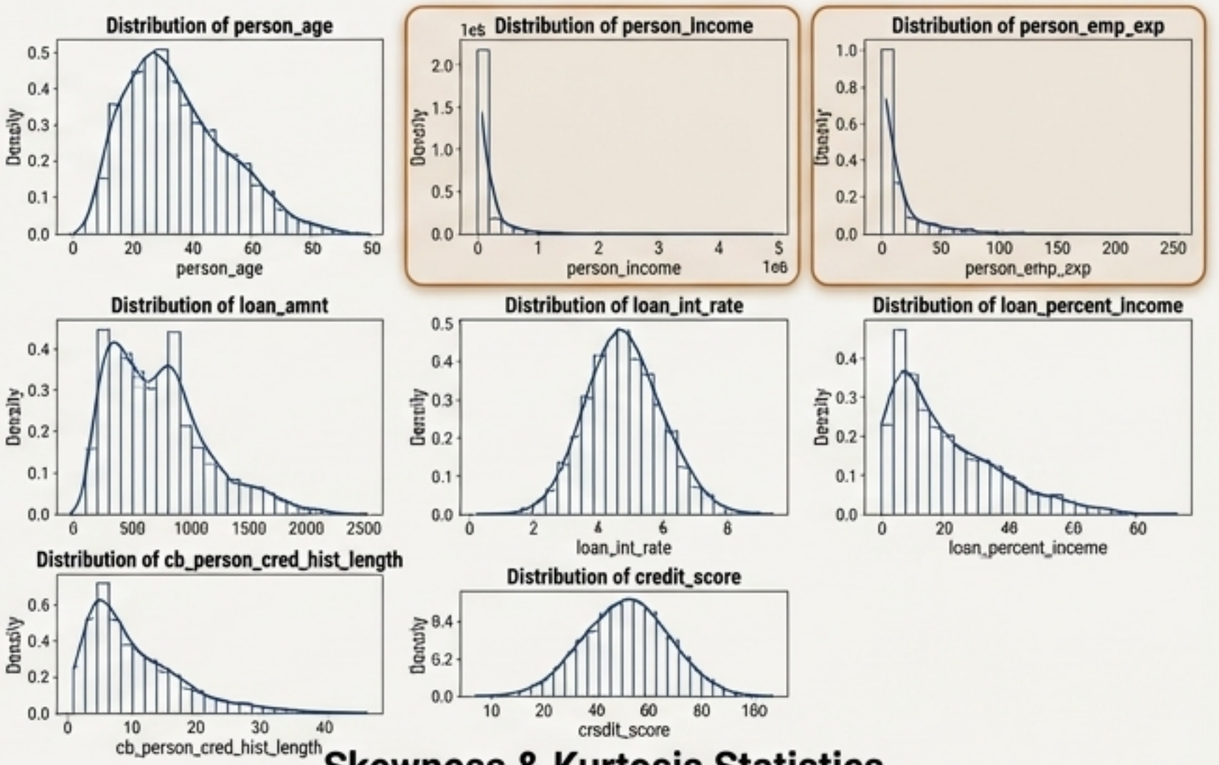
Univariate Analysis Reveals Skewness and Outliers in Financial Data

The dataset is clean with no missing values or duplicates. However, key financial variables exhibit significant right-skewness and heavy-tailed distributions, which can destabilize model training.

Visual Evidence (Boxplots)



Visual Evidence (Histograms)



Skewness & Kurtosis Statistics

Variable	Skewness	Kurtosis	Insight
person_income	34.14	2398.42	Extremely skewed with extreme outliers
person_emp_exp	2.59	19.17	Right-skewed with heavy tails
person_age	2.55	18.65	Highly right-skewed with heavy tails
loan_int_rate	0.21	-0.42	Approximately normal

Skewness & Kurtosis Statistics

Variable	Skewness	Kurtosis	Insight
person_income	34.14	2398.42	Extremely skewed with extreme outliers
person_emp_exp	2.59	19.17	Right-skewed with heavy tails
person_age	2.55	18.65	Highly right-skewed with heavy tails
loan_int_rate	0.21	-0.42	Approximately normal

Strategy: Stabilizing Distributions with Percentile Capping and Log Transformation

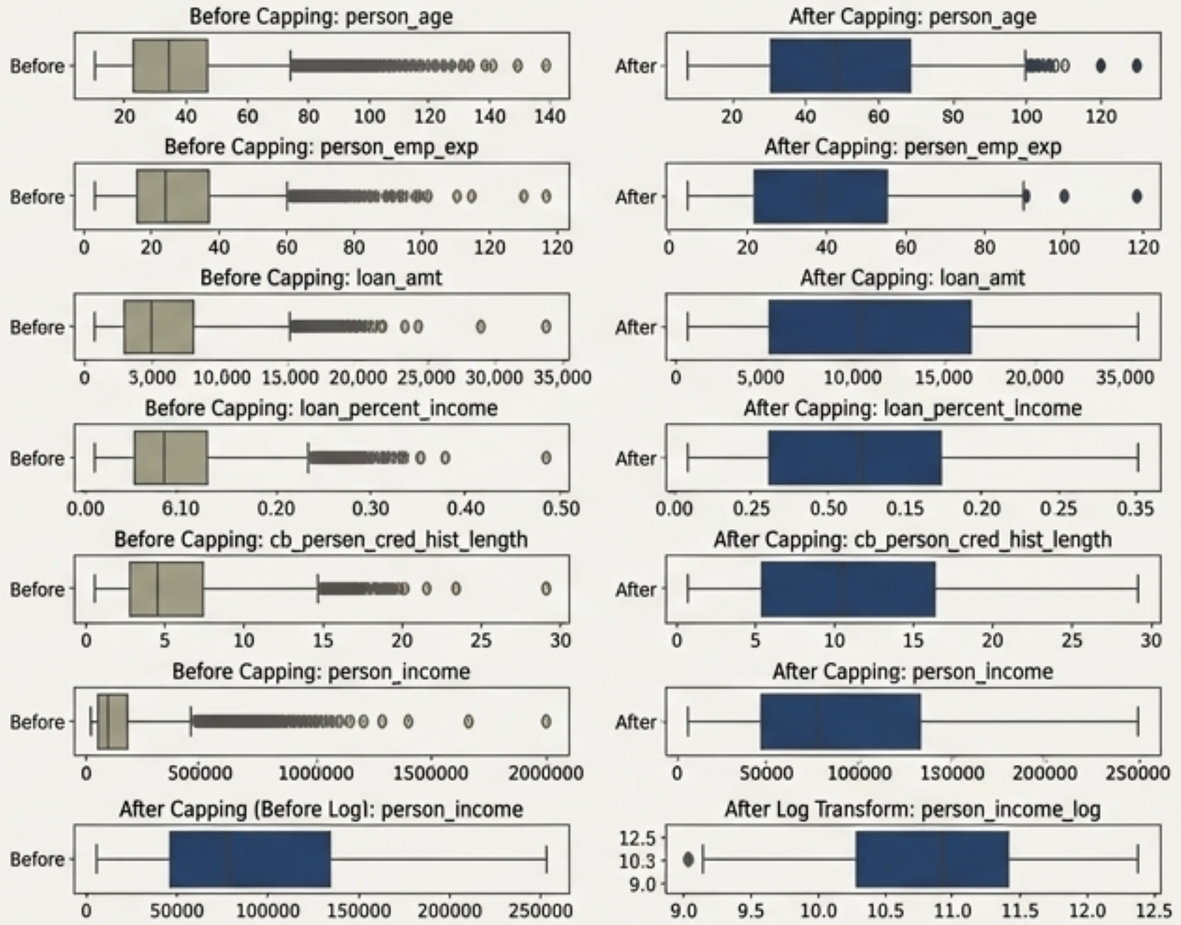
Rationale

- Outright deletion of extreme values is avoided, as they often represent legitimate high-value or high-risk borrowers.
- A **1st-99th percentile capping** approach was applied to limit the influence of extreme observations while retaining all records.
- For the extremely skewed `person_income` variable, a **logarithmic transformation** (`log1p`) was also used to normalize its distribution further.
- Variables with near-normal distributions (`loan_int_rate`, `credit_score`) were intentionally left untreated.

Outcome

This systematic approach improves model robustness and real-world applicability without compromising data integrity.

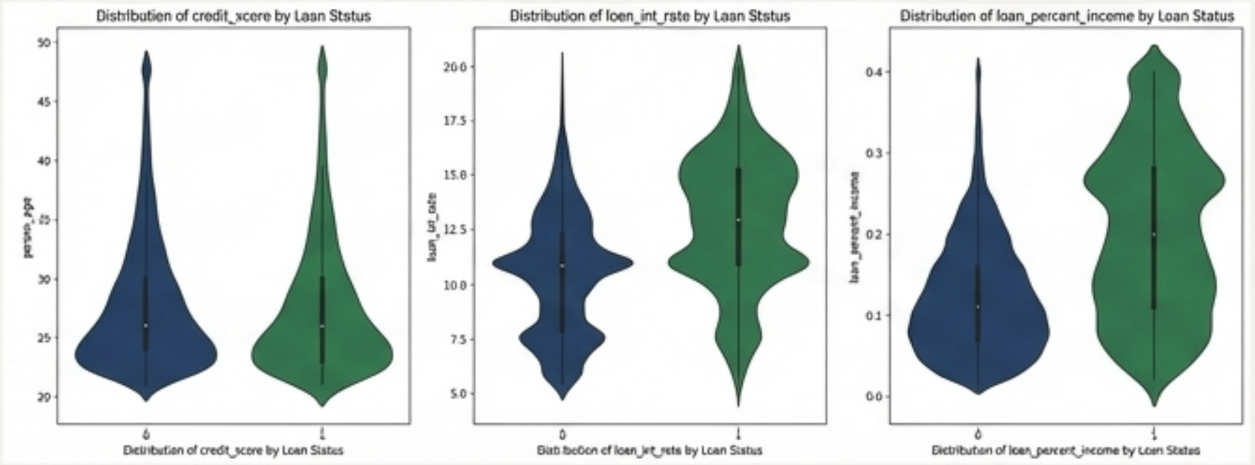
Visual Impact: Before vs. After Treatment



Bivariate Analysis: Identifying Key Drivers of Loan Default

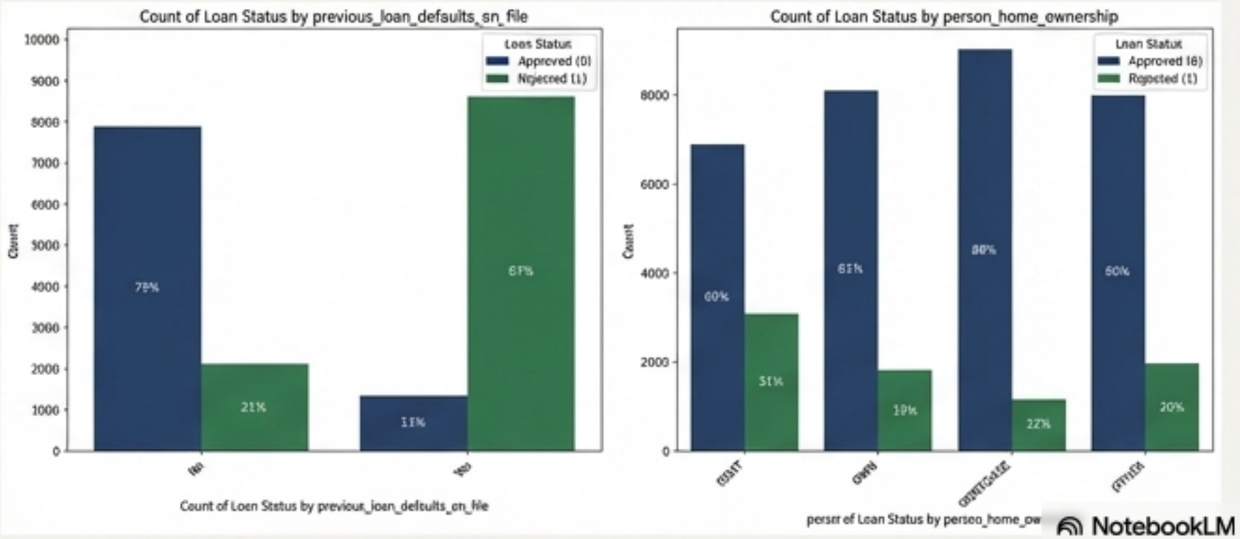
Numerical Risk Indicators

- Violin plots show clear distributional differences for approved (0) and rejected (1) loans.
- Key Separators:
  - **credit_score**: Approved loans have a significantly higher median credit score.
  - **loan_int_rate**: Rejected loans have a distribution shifted towards much higher interest rates.
  - **loan_percent_income**: Rejected loans are associated with a higher repayment burden (loan amount as a percentage of income).



Categorical Risk Indicators

- Count plots reveal strong correlations between categorical features and loan rejection rates.
- Strongest Predictor:
  - A history of previous defaults is a critical risk signal.
  - **87%** of applicants with previous defaults are rejected.
  - Only **21%** of those without previous defaults are rejected.
- Other Factors: Renters show higher rejection rates (~31%) compared to homeowners (~19%).



Feature Preparation: Encoding, Scaling, and Managing Multicollinearity

Feature Engineering & Scaling

- **One-Hot Encoding:** Categorical variables (person_gender, person_education, etc.) were converted into numerical format, expanding the feature set from 14 to 29 columns.
- **Standard Scaling:** All features were scaled using StandardScaler to ensure numerical stability for regression and other distance-based algorithms.

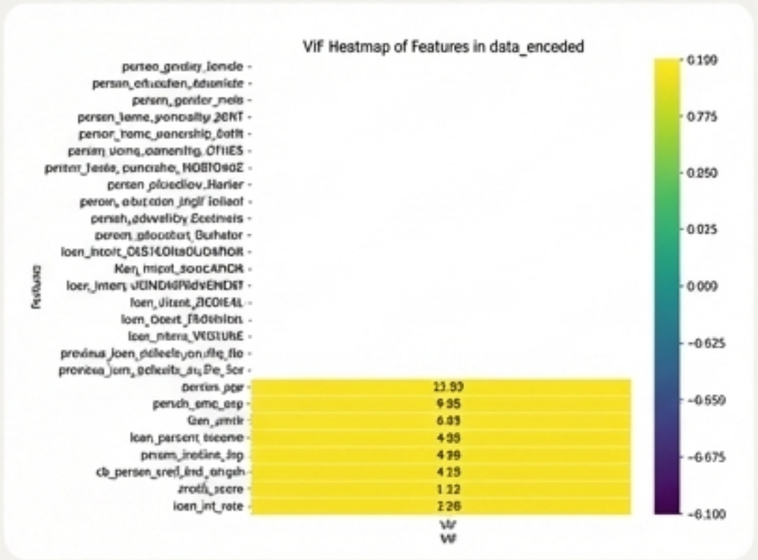
Multicollinearity Analysis & Resolution

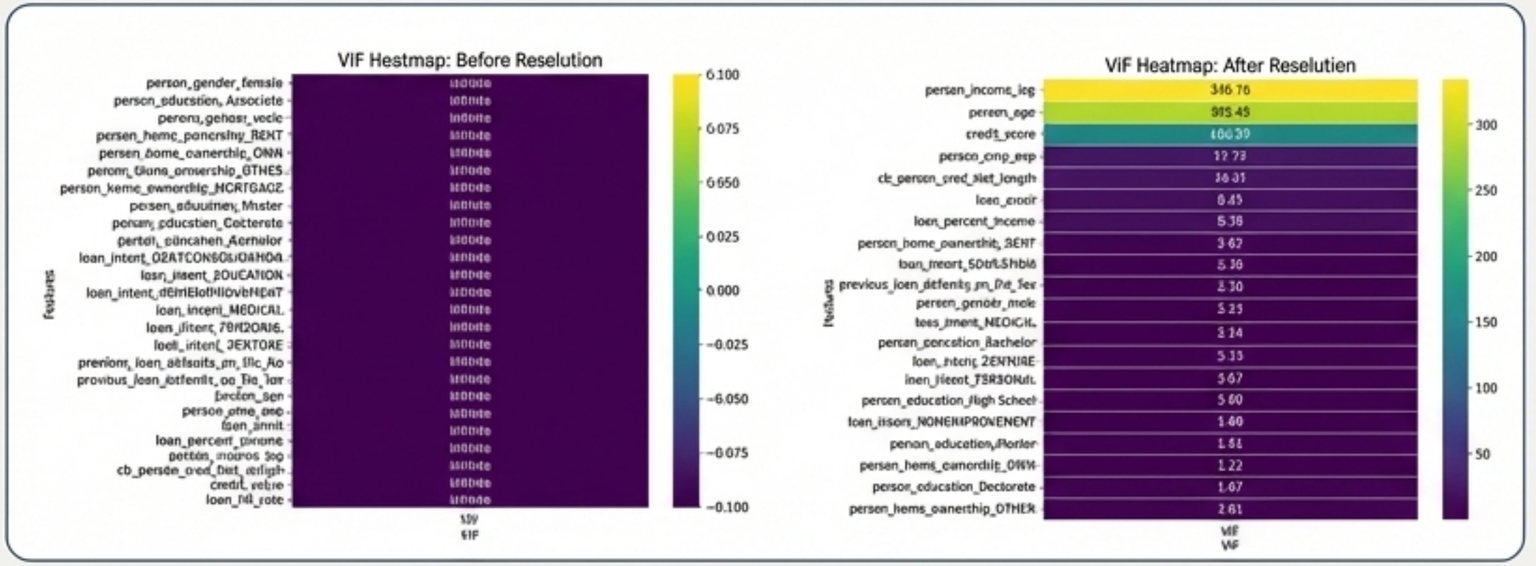
- **Problem:** Initial VIF analysis revealed perfect multicollinearity (infinite VIFs) due to the

Visual Evidence: VIF Heatmaps Before and After

the "dummy variable trap" from one-hot encoding.

- **Solution:** One dummy variable from each original categorical feature was dropped to break the perfect linear dependency (e.g., dropped person_gender_female, person_education_Associate).
- **Result:** This standard practice resolved the infinite VIFs, enabling stable coefficient estimation in estimation in regression models. regression models.





Baseline Model: Establishing an Interpretable Benchmark with Logistic Regression

Serves as a highly interpretable baseline to understand key risk drivers and establish a performance benchmark.

Initial Model Challenges: Quasi-Separation

- The first model using all features failed to converge properly.
- **Reason:** "Possibly complete quasi-separation" was detected. This occurs when a predictor, like `previous_loan_defaults_on_file`, almost perfectly separates the outcome (default vs. non-default).
- **Impact:** This leads to unstable, infinitely large coefficients and non-convergence, making the model unreliable.

```
                                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
                                 0.84         nan        nan        nan        nan        nan
        person_education_Asseciate  1.35317e-83         nan        nan        nan        nan
     person_home_ownership_MORTGAGE  1.29553e-63         nan        nan        nan        nan
      loan_intent_DEBTCOMSOLIDATION  1.39385e-83         nan        nan        nan        nan
      loan_intent_DEBTCONSOLIDATION  1.68792e-63         nan        nan        nan        nan
       loan_intent_DEBTCONSCIDATION  1.28389e-63         nan        nan        nan        nan
     loan_intent_ownership_Assciate  1.39252e-83         nan        nan        nan        nan
               loan_intent_OSRECTON  2.28159e-83         nan        nan        nan        nan
               loan_intent_FREEGTCY  1.42892e-63         nan        nan        nan        nan
                     loan_intent_LAN  1.16789e-63         nan        nan        nan        nan
  previous_loan_defaults_on_file_No     78683e+91         nan        nan        nan        nan
  previous_loan_defaults_on_file_No  1.28979e-93         nan        nan        nan        nan
  previous_loan_defaults_on_file_No  1.35853e-63         nan        nan        nan        nan
  previous_loan_defaults_on_file_No  1.28639e-63         nan        nan        nan        nan
  previous_loan_defaults_on_file_No  1.68395e-81         nan        nan        nan        nan
  previous_loan_defaults_on_file_No  1.40032e-61         nan        nan        nan  0.3525270        nan
 previous_loan_defaults_en_File_Tes -3.95637e-63         nan        nan        nan        nan        nan
Possibly complete quasi-separation: A fraction 6.51 of observatiens can be perfectly predicted.
Thts aight indicate that there is complete quasi-separation. In this case some parameters will not
be identified.
```

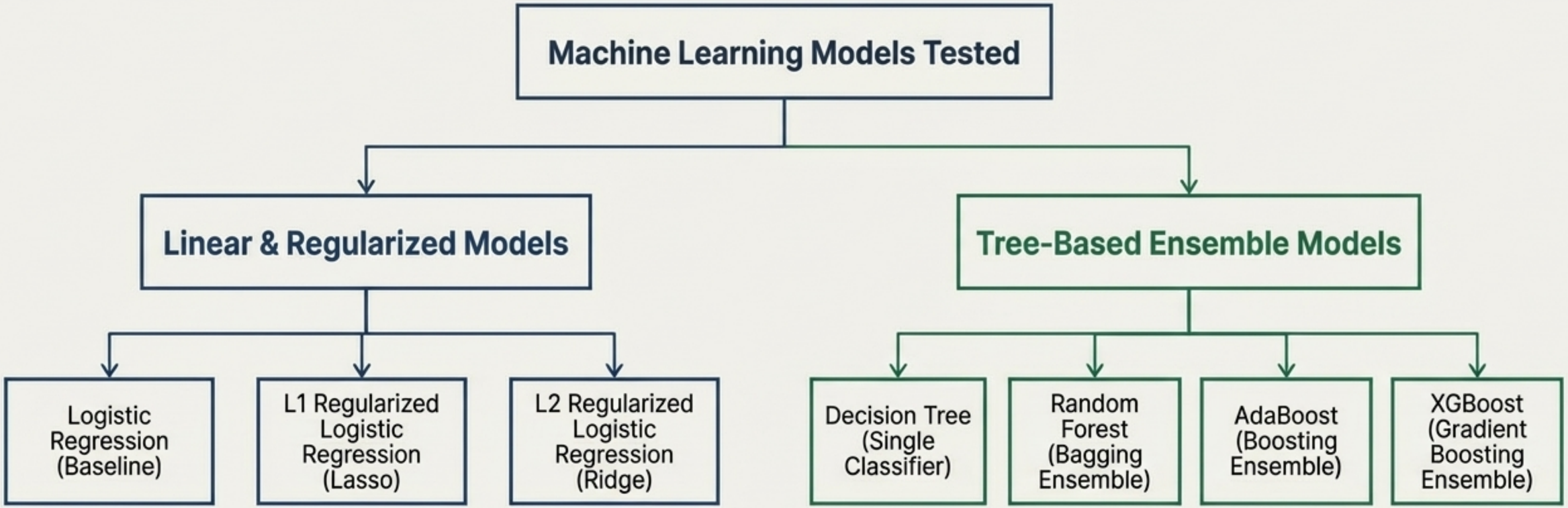
Refined Model & Performance

- **Solution:** To ensure model identifiability and stable inference, `previous_loan_defaults_on_file` was excluded *from the logistic regression model only*.
- The refined model successfully converged, providing stable and meaningful coefficients.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Refined Logit	0.8506	0.7348	0.5125	0.6038	0.8642

Evaluating a Spectrum of Machine Learning Models

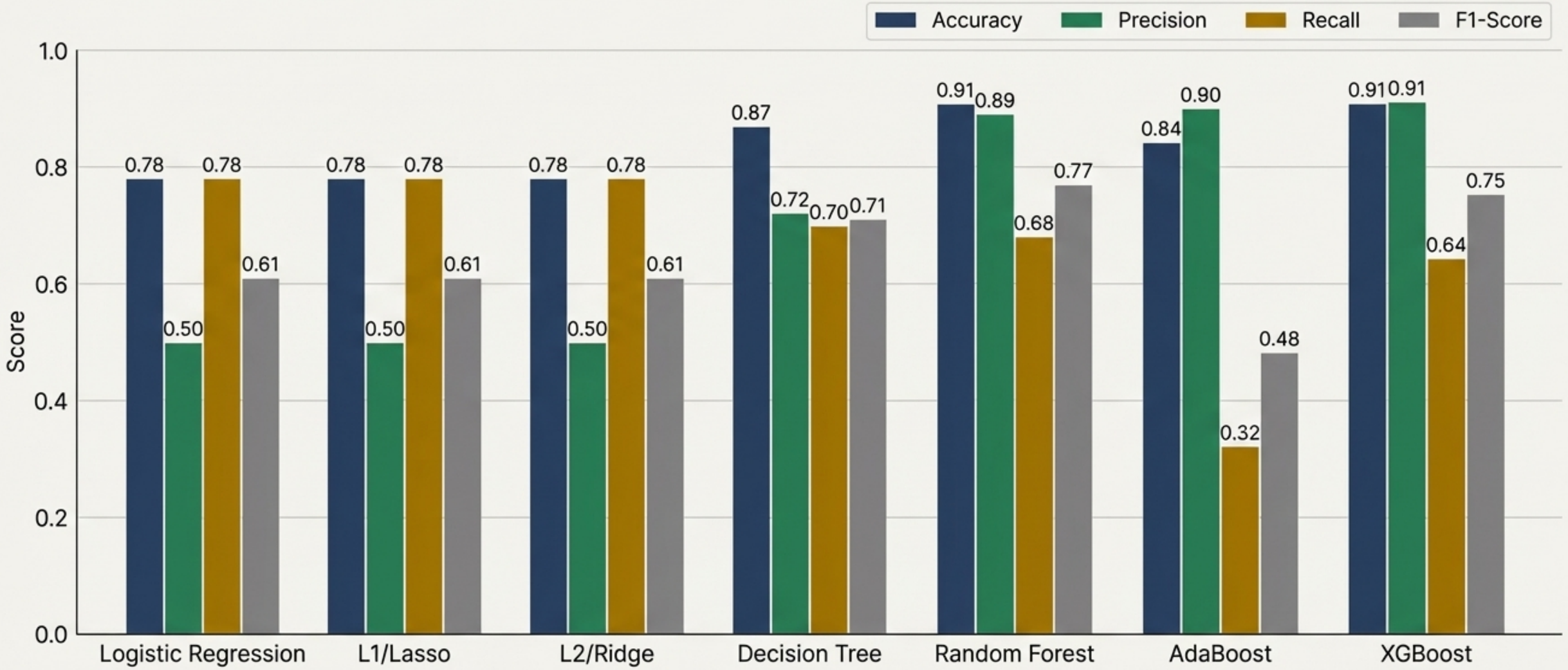
To move beyond the linear baseline, a range of more complex algorithms were developed and tested on the engineered feature set. The goal is to compare their predictive power, especially their ability to capture non-linear relationships and interactions.



Evaluation Focus: While all standard metrics are reported, special emphasis is placed on **Recall** (ability to identify true defaulters) and **ROC-AUC** (overall risk discrimination ability), as these are most critical for credit risk management.

Performance Showdown: Comparing All Models Across Key Metrics

Tree-based ensemble models, particularly **Random Forest** and **XGBoost**, significantly **outperform** the linear models in overall **accuracy**, **precision**, and **F1-Score**. AdaBoost shows high precision but suffers from extremely low recall.



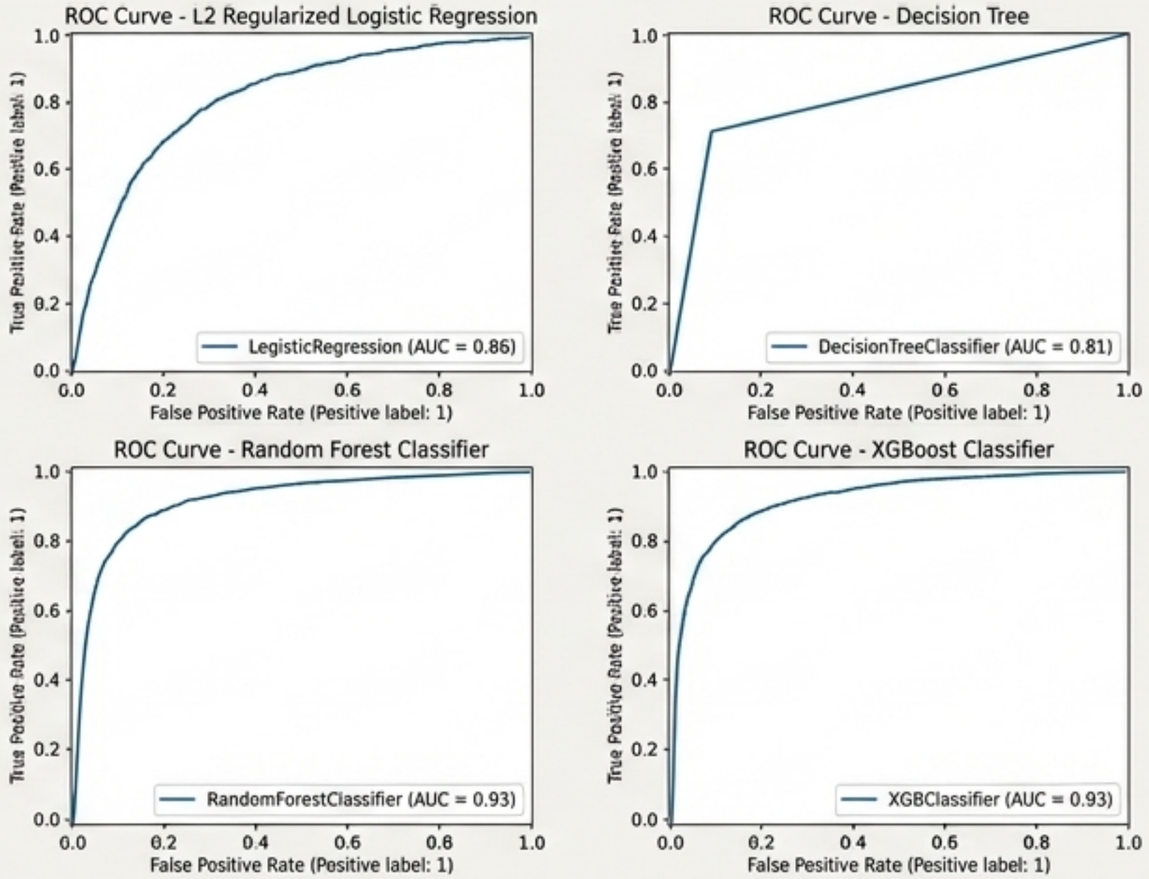
Assessing Risk Discrimination Ability with ROC-AUC

What ROC-AUC Measures: A model's ability to correctly distinguish between positive (default) and negative (non-default) classes. A higher AUC value (closer to 1.0) indicates better performance in ranking borrowers by their predicted risk. This is critical for risk-based pricing and setting approval thresholds.

Performance Comparison

Model	ROC-AUC Score
Random Forest	**0.9304**
XGBoost	**0.9284**
AdaBoost	0.8707
Logistic Regression (all variants)	~0.8630
Decision Tree	0.8109

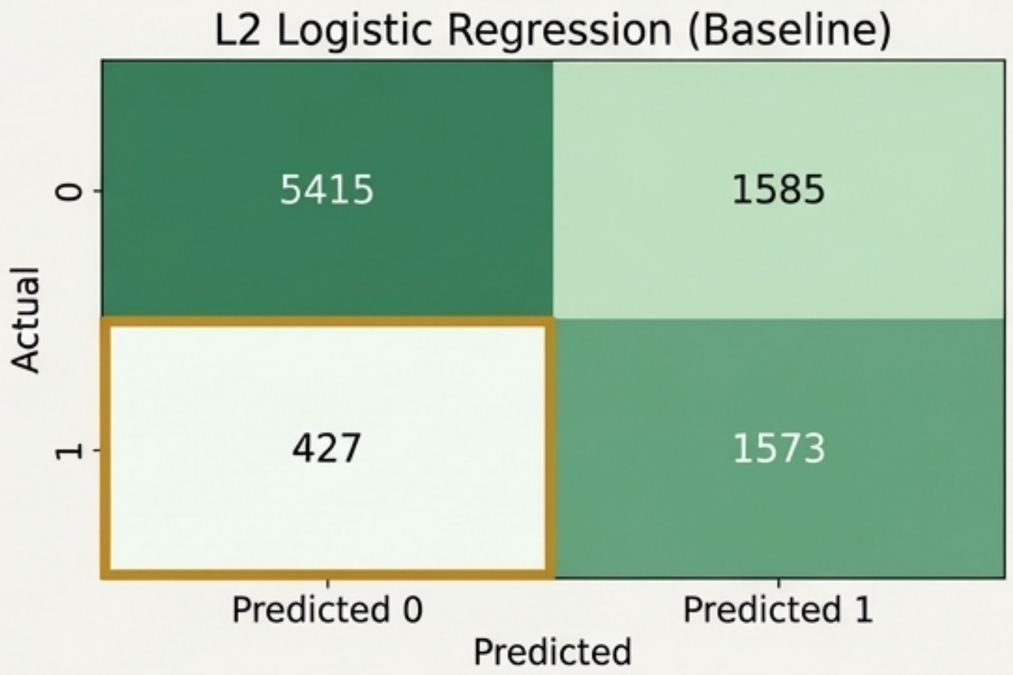
Visual Evidence: ROC Curves

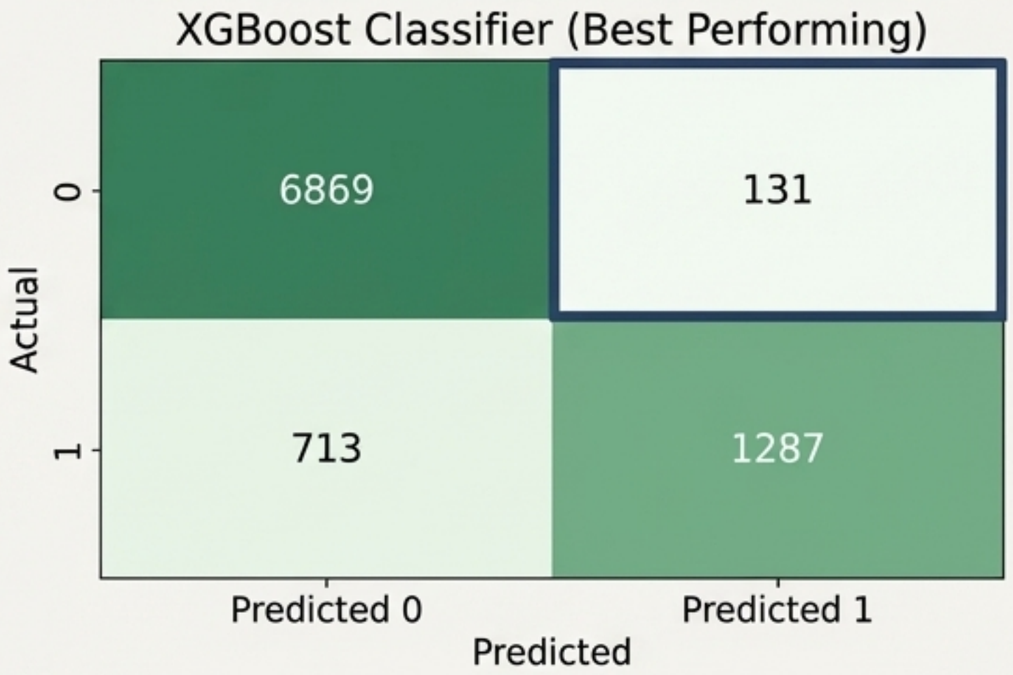


Error Analysis: Comparing the Real-World Impact of Model Decisions

A model's value is determined by the quality of its decisions. Comparing the confusion matrices shows the tangible improvement in classification, especially in reducing costly errors.

Key Metric: False Negatives (bottom-left quadrant) represent high-risk applicants who are incorrectly approved. Minimizing this number is paramount to reducing credit losses.





While Logistic Regression has fewer False Negatives (higher recall), XGBoost is far more precise, drastically reducing False Positives (incorrect rejections) from 1,585 to 131. This highlights the fundamental trade-off. The final model choice depends on the business's tolerance for different types of errors. Based on overall performance (ROC-AUC, Accuracy, F1), XGBoost is superior.

Final Model Selection and Business Recommendation

Overall Model Ranking Based on Performance

1. **Best for Overall Predictive Power (ROC-AUC & F1): XGBoost & Random Forest** consistently delivered the highest performance, excelling at capturing non-linear relationships.
2. **Best for Identifying Defaulters (Recall): Logistic Regression models** (Baseline, L1, L2) showed the highest recall, making them effective at flagging potential defaults, albeit with lower precision.
3. **Best for Interpretability: Logistic Regression** remains unparalleled for its stable, easily explainable coefficients, which are crucial for regulatory compliance and explaining decisions.

Final Selection: XGBoost

XGBoost is selected as the final predictive model due to its superior balance of performance across all key metrics, especially its high ROC-AUC score (0.9284) indicating excellent risk discrimination.

Business Recommendation: A Hybrid Deployment Strategy

Internal Risk Scoring: Use the **XGBoost model** for internal credit scoring, portfolio monitoring, and early warning systems to maximize predictive accuracy and minimize overall credit losses.

Compliance & Explainability: Retain the refined **Logistic Regression model** for regulatory reporting, internal audits, and generating customer-facing decision explanations, satisfying the need for transparency and fairness.

Conclusion: A Data-Driven Framework for Advanced Credit Risk Management

1 End-to-End Analytical Rigor is Essential

This study demonstrated a complete pipeline from data exploration and statistically sound preprocessing to systematic model comparison. This rigorous process builds credibility and ensures the final model is both robust and reliable.

2 Business-Aligned Model Selection is Key

The results confirm a clear trade-off between predictive performance and interpretability. Tree-based ensembles (XGBoost) excel at risk discrimination, while Logistic Regression provides transparency. The optimal strategy is not to pick one "best" model, but to align the right model with the right business function—performance for internal scoring, and interpretability for compliance.

3 The Future is a Hybrid Approach

By combining the predictive power of advanced machine learning with the accountability of interpretable models, financial institutions can create a credit decision framework that is more accurate, transparent, and resilient. This balances the pursuit of business performance with the principles of responsible and ethical lending.