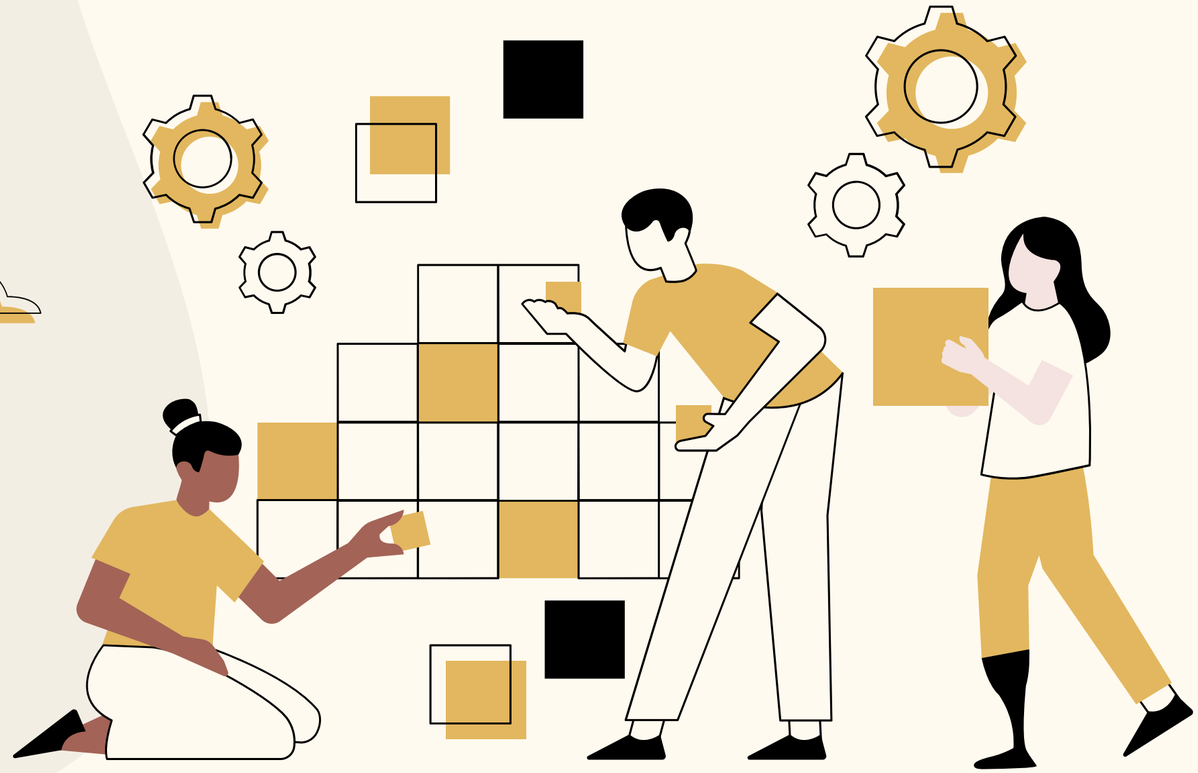
Predictive Modeling for Loan Approval

Created by:
Farah Naaz
Adedamola Dosunmu
Jahnavi Gogineni



Link

Table of content

01 Brief Project Overview

02 EDA Summary

03 Benchmark Solutions

04 Modeling Approach

05 Performance Evaluation

06 Model Interpretation

07 Business Implications

08 Lessons Learned

09 Next Steps

Brief Project Overview

At JAF Bank, borrower risk evaluation and loan approval decisions have traditionally relied on manual assessments and rule-based credit scoring systems.

- Leverage EDA to discover key relationships between variables.
- Identify strongest predictors of loan approval.
- Develop predictive models to determine the likelihood of loan approval.

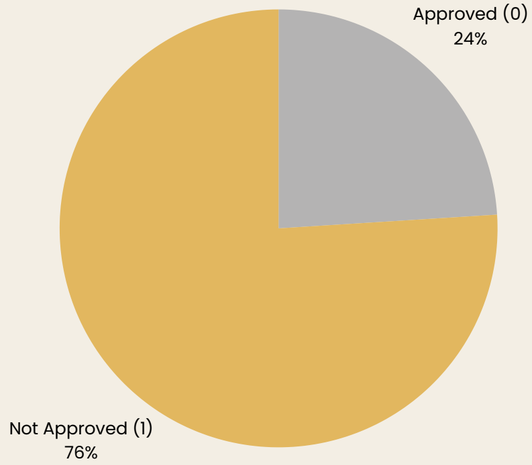
Dataset Overview

20,000 rows

36 columns

Demographic Features: age, education, marital status

Financial metrics: income, debt, credit score



Target Variable: Loan Approved

EDA SUMMARY

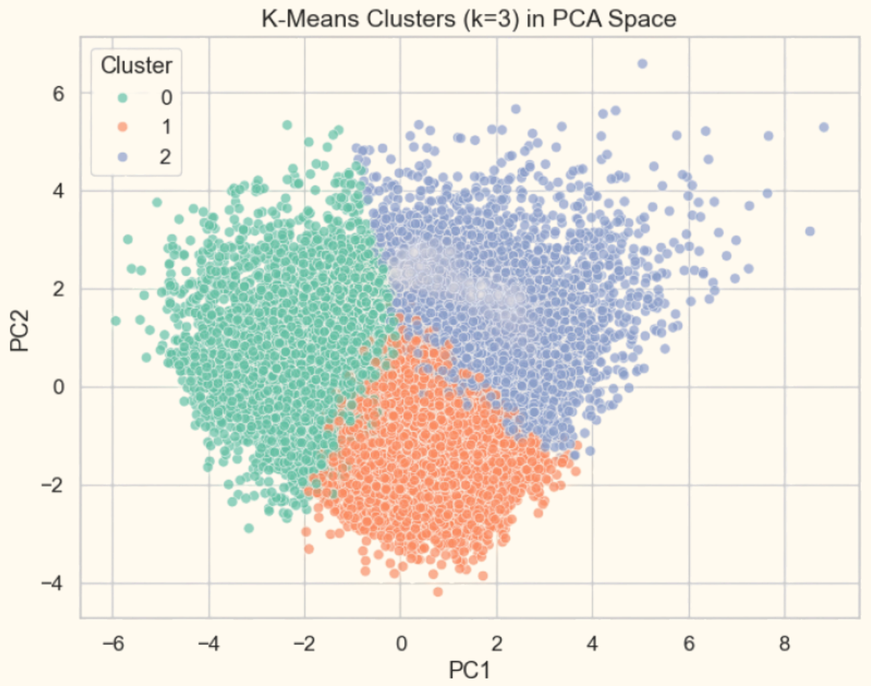
Next Slide

Cluster Analysis

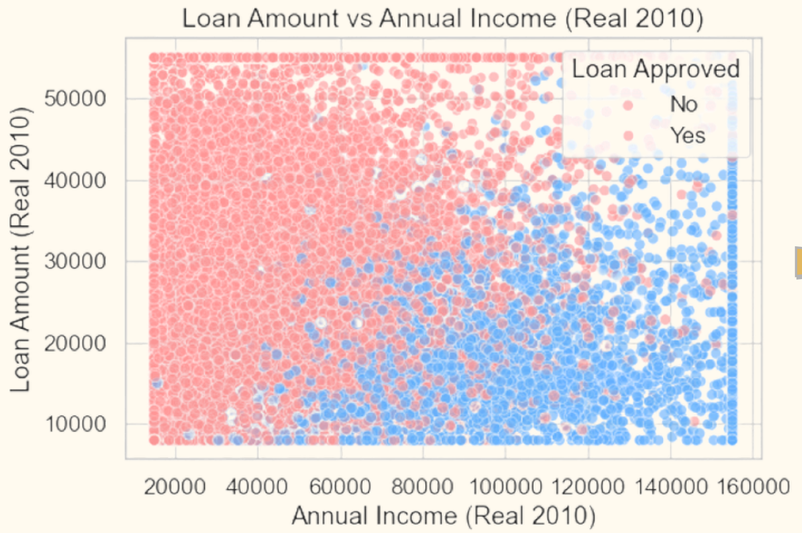
- Cluster 0: “Low-Risk, High-Capacity Borrowers” (Approval Rate: 57%).
- Financially strong applicants with stable profiles, high income, and assets — naturally resulting in the highest loan approval rate.

- Cluster 1: “Middle-Income Borrowers with Moderate Risk” (Approval Rate: 13%)
- Applicants have average income and modest assets, making them neither high-risk nor very strong applicants.

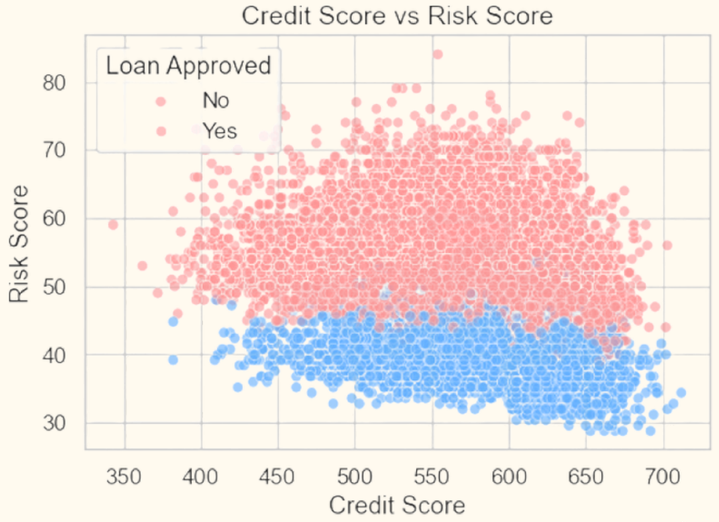
- Cluster 2 — “High-Request, Low-Capacity Borrowers” (Approval Rate: 1.5%)
- This group exhibits mismatch between low financial capacity and high borrowing needs, leading to the lowest approval rate among all clusters.



Essential Relationships

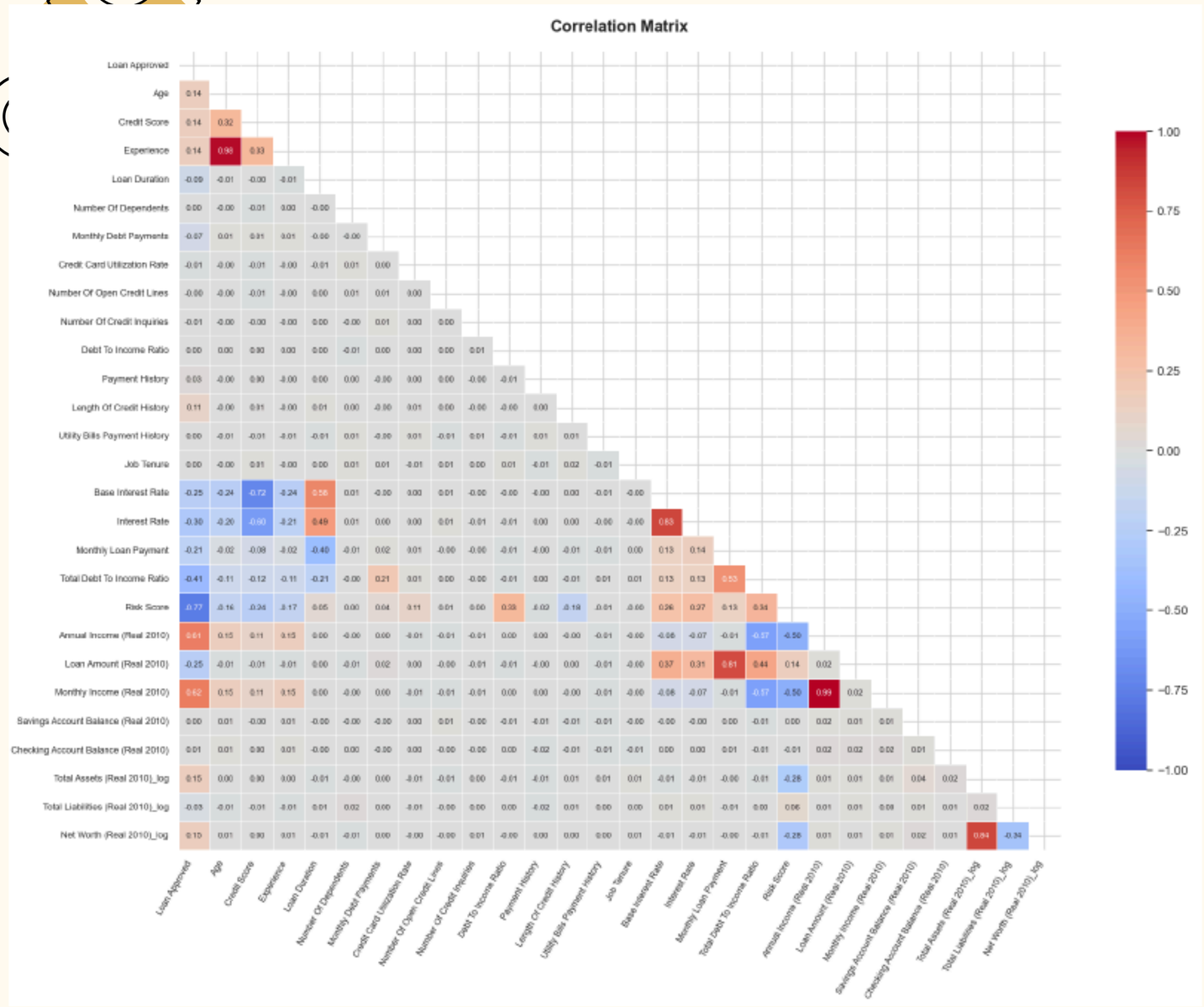


Approvals are much more common at higher annual incomes and moderate loan amounts, while applicants with lower incomes across the loan amount range are predominantly rejected, indicating that income level is a key driver of approval decisions.



Approved applicants (blue) cluster at higher credit scores (500–700) but in a lower risk-score band (30–45), while rejected applicants (pink) occupy the higher risk-score region (50–75) across all credit scores, indicating that lower Risk score values combined with stronger credit scores are associated with higher approval likelihood.

Correlation Matrix



- Monthly Income(0.62) & Annual Income(0.61) → Strongest Positive Predictors; Higher-income applicants are significantly more likely to be approved
- Total Assets(0.15) & Net Worth(0.15) → Positive Correlations; applicants with stronger financial positions tend to receive approvals more frequently
- Risk Score(-0.77) → Strongest Negative Correlation; Higher risk scores are strongly associated with rejection
- Loan Amount(-0.21) & DTI (-0.14) → Moderate Negative Correlations; Higher debt obligations or larger requested loan amounts reduce the likelihood of approval

BENCHMARK SOLUTIONS

Next Slide

Benchmark Solutions

Study: “Enhancing Credit risk Assessment in loan approval: performance evaluation of machine learning Models” by Gia-Huy Truong, Thao-nhi phan phan, Thi-Thuong Truong and To-Hoang Huu Nguyen, 2025 *(https://www.researchgate.net/publication/394367085_Enhancing_Credit_Risk_Assessment_in_Loan_Approval_Performance_Evaluation_of_Machine_Learning_Models)*

Models used: Logistic Regression, Random Forest, XG Boost

Results: XG boost was the best performing model with an accuracy score of 91.9% and AUC of 96.5%.

Benchmark Solutions

Study: “Credit risk prediction using machine learning and deep learning: A study on credit card customers” by Victor Chang, Sharuga Sivakulasingam, Hai Wang, Siu Tung Wong, Meghana Ashok Ganatra and Jiabin Luo *(https://www.mdpi.com/2227-9091/12/11/174)*

Models used: Logistic Regression, AdaBoost, XG Boost, LightGBM, Neural networks

Results: XG boost resulted with best model achieving 99.4% accuracy. Logistic regression was the weakest.

Benchmark Solutions

Study: “A proposed framework for loan default prediction using machine learning techniques” by Mona Aly Sharaf Eldin, Amira M. Idrees, Shimaa Ouf, 2025 *(https://thesai.org/Publications/ViewPaper Volume=16&Issue=6&Code=IJACSA&SerialNo=40)*

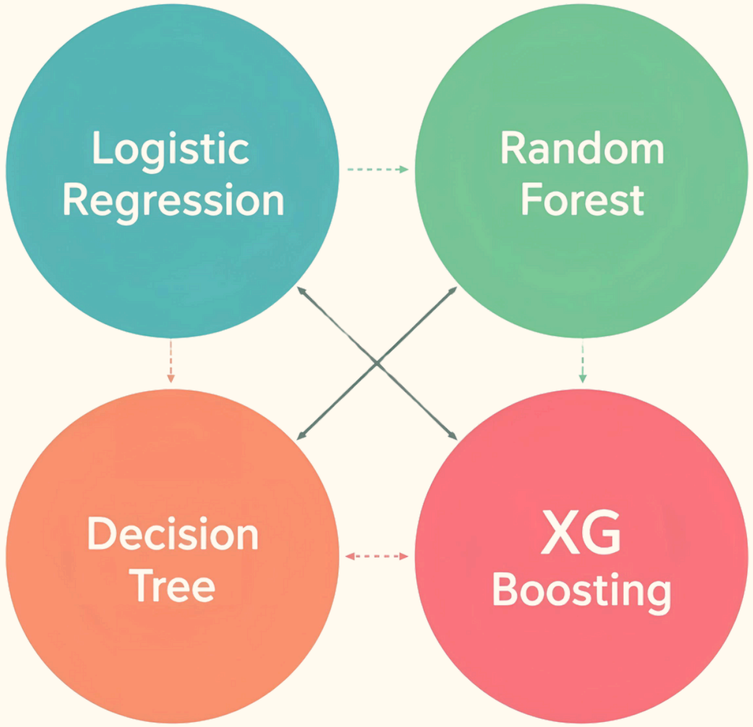
Models used: Decision tree, Random forest, Gradient boosting.

Results: Decision tree achieved the best interpretability and accuracy at 88%

MODELING APPROACH

Next Slide

Model Selection



Handling Imbalanced Data

Random Undersampling → Data Shape (9560, 14); created a perfect 1:1 class balance (50% majority & 50% minority) but lost over 50% of original data potentially removing valuable patterns from the majority class.

SMOTE + ENN → Data Shape (23693, 14); successfully created an optimal 1.15:1 class balance (53.5% majority & 46.5% minority) by combining synthetic minority class generation while retaining original information

Train-Test Split

We used 80:20 for our dataset

Accepted standard in machine learning, backed by substantial study and real-world application.

This split guarantees reasonable training times without compromising model performance with its 36 features and sophisticated algorithms (XGBoost, Random Forest).

Feature Selection

Preserved High-Impact Predictive Variables

• Risk Score (-0.77 correlation) - Primary risk indicator
• Monthly Income (0.62 correlation)—Key affordability measure
• Total Debt-to-Income Ratio (-0.41 correlation) - Critical capacity metric
• Credit Score (0.14 correlation) - Traditional credit assessment

Strategic Elimination of Non-Predictive Features

•Removed 9 low-correlation variables (correlation < |0.02|) that added noise and provided no predictive value. E.g. job tenure, number of dependents, savings A/C balance, etc.
• Eliminated redundant features in case of high redundancy pairs. E.g. dropped Experience - kept Age; droped interest rate - kept base interest rate, etc.

Evaluation Metrics

Accuracy

Precision

F1 - Score

Recall

Balanced Accuracy

Type I Error

Type II Error

Power

Confusion Matrix

ROC- AUC

Cross Validation

Stratified K-Fold Cross-Validation (K=5) → splits the data into five folds while preserving the original class proportions in each fold, ensuring that every subset is representative of the overall dataset.

Meaningful choice for our imbalanced loan approval data, because it prevented biased splits, allowed every model to train and validate on balanced class distributions, and produced more reliable and stable performance estimates.

Hyperparameter Tuning

Randomized Search CV - most suitable for our data. Instead of testing every combination (like Grid Search), Randomized Search explores a smarter subset of parameters, making tuning much faster for models with many hyperparameters like XGBoost and Random Forest.

It reduces computation time significantly while maintaining strong performance, which is ideal for gradient boosting models.

PERFORMANCE EVALUATION

Next Slide

Logistic Regression

Metric	Test Set	CV Mean ± Std
Accuracy	0.9648	0.9705 ± 0.0047
Precision	0.9636	0.9688 ± 0.0038
Recall	0.9662	0.9722 ± 0.0062
F1-Score	0.9649	0.9705 ± 0.0048
Balanced Acc.	0.9648	0.9705 ± 0.0047
ROC-AUC	0.9943	0.9954 ± 0.0011
Type I Error	0.0365	–
Type II Error	0.0338	–
Power (1-Type II)	0.9662	–

- Interpretable baseline
- Reliable but clearly outperformed by tree-based models.

Decision Tree

Metric	Test Set	CV Mean ± Std
Accuracy	0.9900	0.9887 ± 0.0009
Precision	0.9895	0.9884 ± 0.0020
Recall	0.9905	0.9892 ± 0.0028
F1-Score	0.9900	0.9888 ± 0.0009
Balanced Acc.	0.9900	0.9887 ± 0.0009
ROC-AUC	0.9900	0.9887 ± 0.0009
Type I Error	0.0105	–
Type II Error	0.0095	–
Power (1 - Type II)	0.9905	–

- Good performance
- More prone to variance

Random Forest

Metric	Test Set	CV Mean ± Std
Accuracy	0.9926	0.9922 ± 0.0009
Precision	0.9928	0.9923 ± 0.0011
Recall	0.9924	0.9920 ± 0.0010
F1-Score	0.9926	0.9922 ± 0.0009
Balanced Acc.	0.9926	0.9922 ± 0.0009
ROC-AUC	0.9998	0.9998 ± 0.0000
Type I Error	0.0072	–
Type II Error	0.0076	–
Power (1 - Type II)	0.9924	–

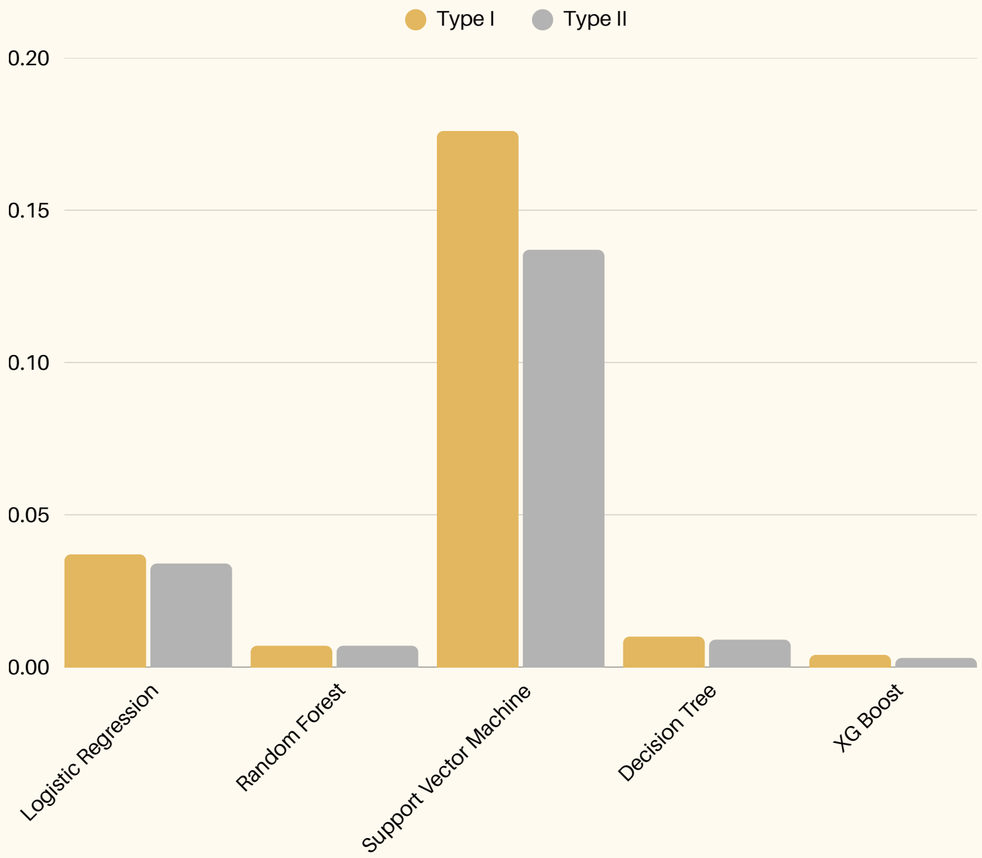
- Strong runner-up
- Consistent generalization

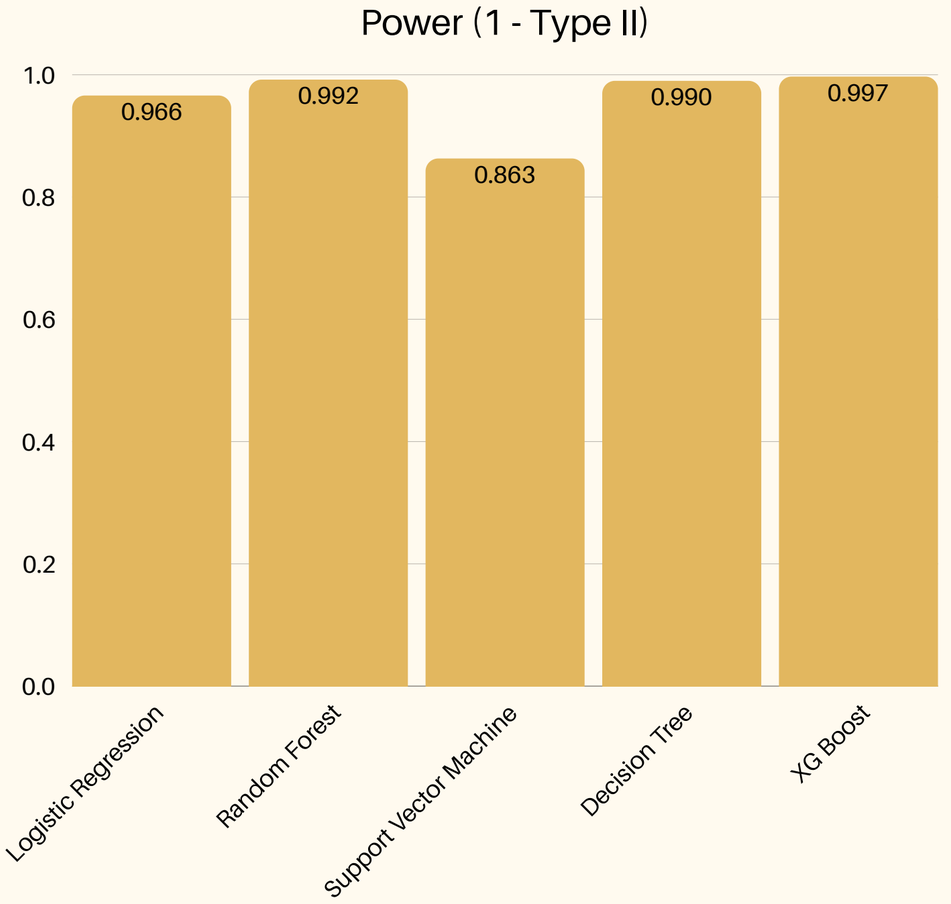
XG Boost

Metric	Test Set	CV Mean ± Std
Accuracy	0.9962	0.9952 ± 0.0015
Precision	0.9957	0.9955 ± 0.0025
Recall	0.9967	0.9950 ± 0.0008
F1-Score	0.9962	0.9952 ± 0.0015
Balanced Acc.	0.9962	0.9952 ± 0.0015
ROC-AUC	0.9999	0.9999 ± 0.0000
Type I Error	0.0043	–
Type II Error	0.0033	–
Power (1 - Type II)	0.9967	–

- Overall winner
- Almost all good and bad loans correctly classified
- Highly stable and not overfitting

Error Analysis





XGBoost and Random Forest show the strongest performance with the lowest Type I & Type II errors and the highest statistical power, while SVM performs noticeably worse with higher error rates and substantially lower power.

MODEL INTERPRETATIONS

Next Slide

Logistic Regression

Intercept: –8.658

When all predictors are at zero, the log-odds of loan approval are strongly negative.

Top Negative Predictors (Reduce Approval Chances)

Feature	Coefficient	Odds Ratio	Interpretation
Risk Score	–8.59	0.000186	- Strongest predictor. Higher risk score drastically reduces approval.
Total Debt To Income Ratio	–2.69	0.068006	- Higher debt burden → lower approval probability.
Loan Amount	–1.77	0.170419	- Larger requested loans reduce approval chances.
Loan Duration	–0.88	0.414986	- Long-term loans are seen as riskier.
Home Ownership	–0.66	0.519131	- Less stable housing situations (other/rent) reduce approval likelihood.

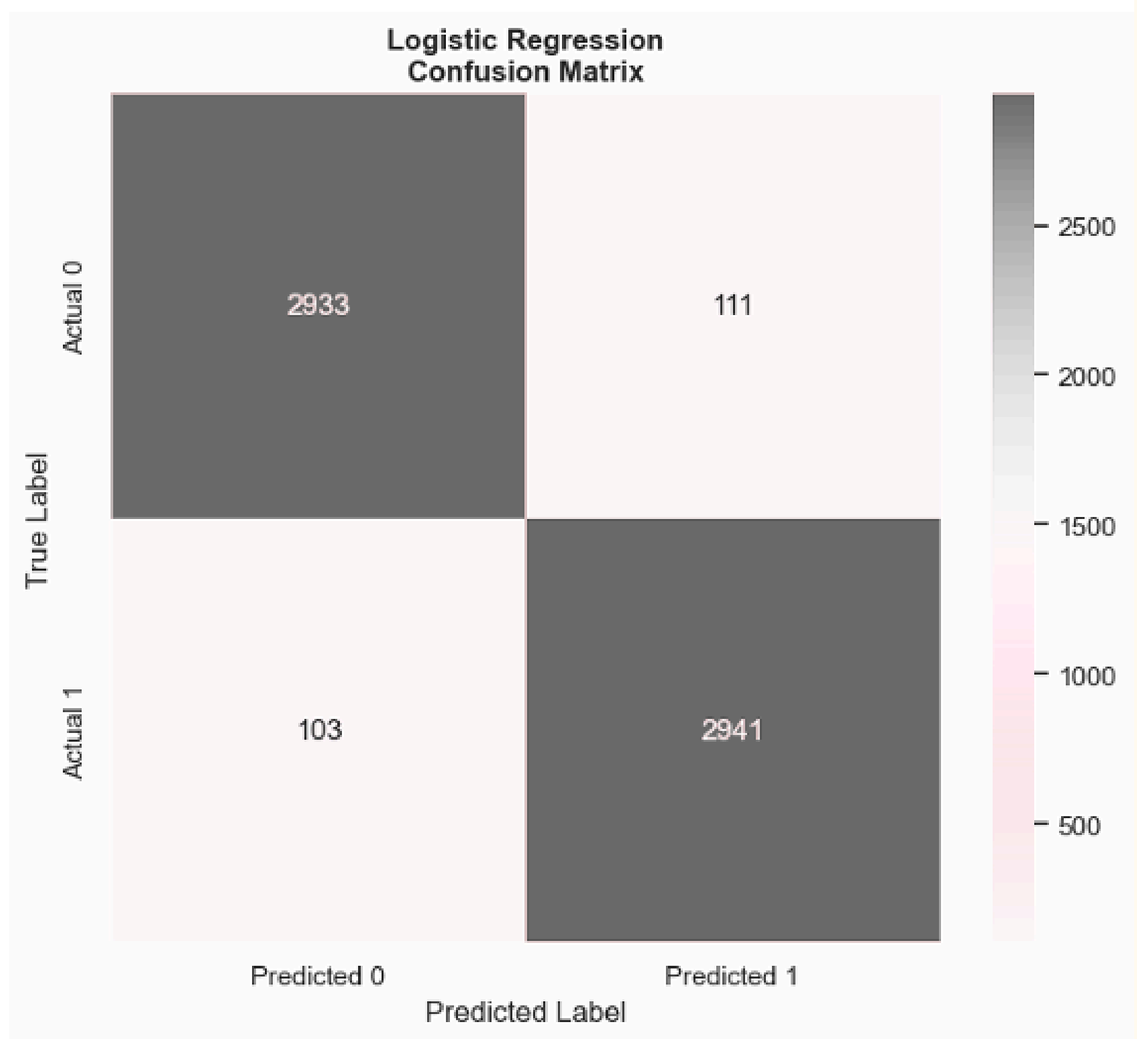
The model heavily penalizes risky financial behavior, high borrowing, and weak stability indicators.

Top Positive Predictors (Increase Approval Chances)

Feature	Coefficient	Odds Ratio	Interpretation
Bankruptcy History	7.19	1319.74	- Since “1” appears to mean No bankruptcy, not having a bankruptcy hugely increases approval odds.
Previous Loan Defaults	3.69	40.17	- Encoded as 1 = No defaults → increases approval.
Employment Status: Self-Employed	2.90	18.27	- Self-employed borrowers are 18x more likely to be approved
Monthly Income	1.45	4.27	- Higher income increases approval probability.
Education Level	0.49	1.63	- More educated applicants are viewed as lower risk.

Good financial stability & higher income, strongly boost approval odds.

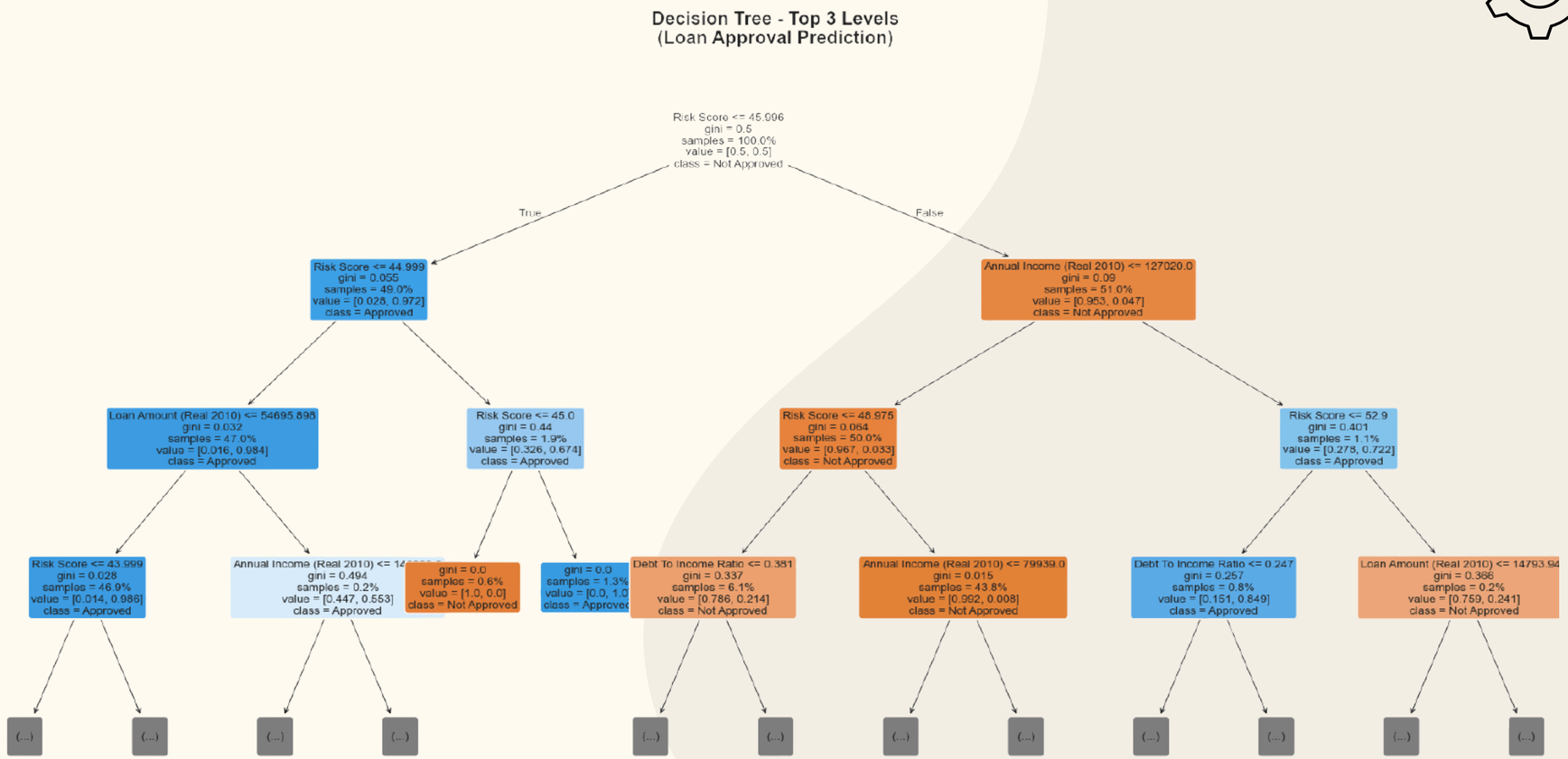
LR - Confusion Matrix



- Correctly classifies most observations, showing solid overall accuracy.
- 111 false negatives (not approved but actually approved).
- 103 false positives (approved but actually not approved).
- Higher error counts than Random Forest and XGBoost.
- Precision (96.36%) and recall (96.62%) are good but not exceptional.
- Works well as a baseline model but is outperformed by tree-based models.

Decision Tree

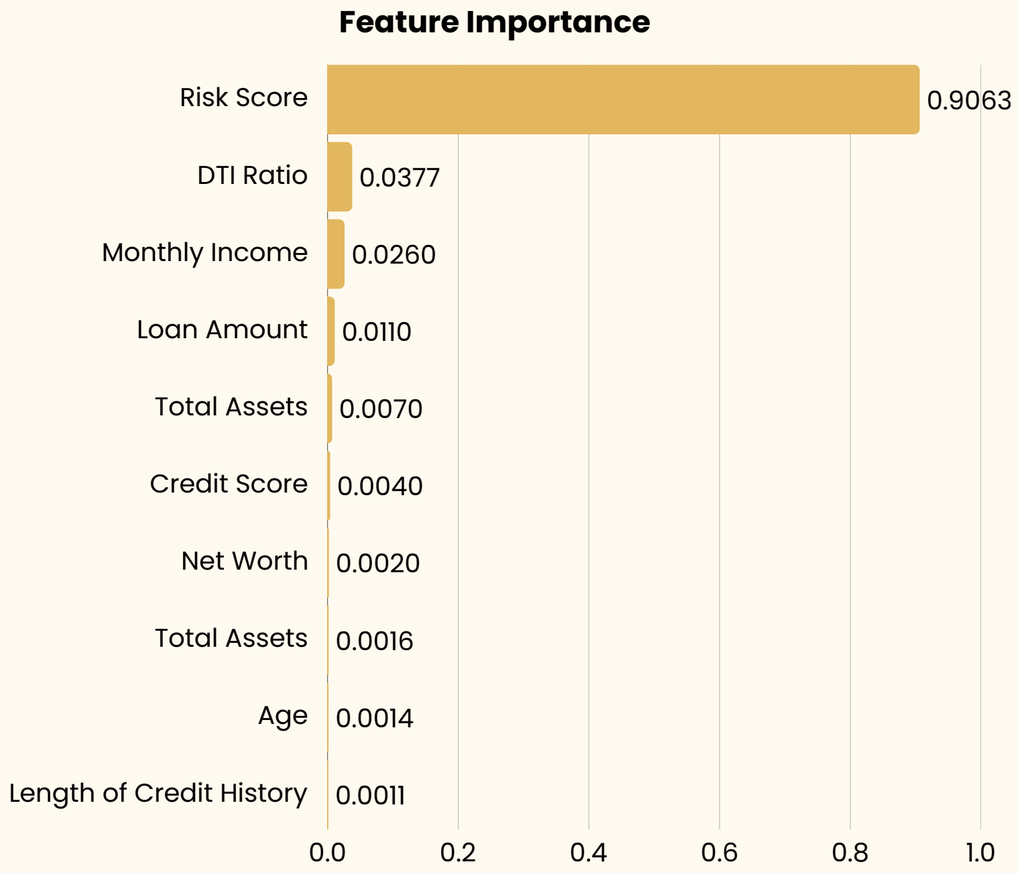
- Total nodes: 369
- Leaf nodes: 185
- Max depth: 16
- No. of features used: 14



If Risk Score ≤ 45.99 & Risk Score ≤ 44.99 & Loan Amount ≤ 54695.89 & Risk Score ≤ 43.99 & Annual Income ≤ 21364.4463 & Debt To Income Ratio ≤ 0.1942 then:

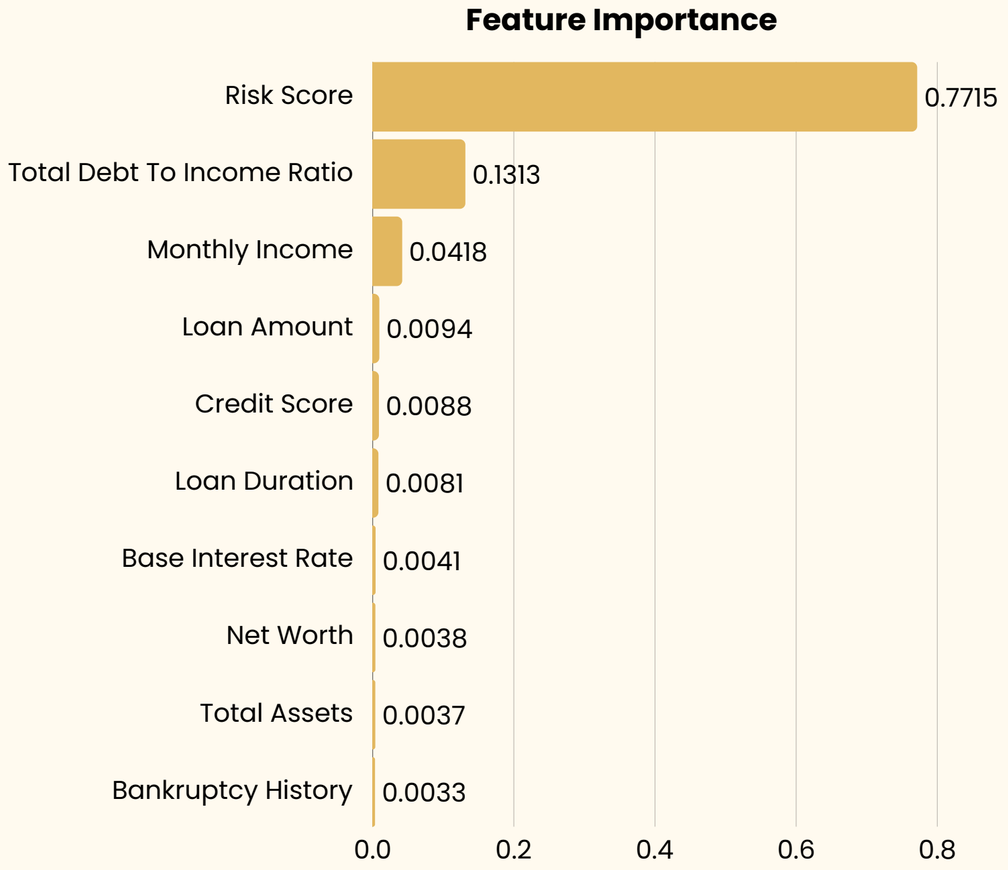
Predicted: REJECT (Not Approved = 1)

Decision Tree



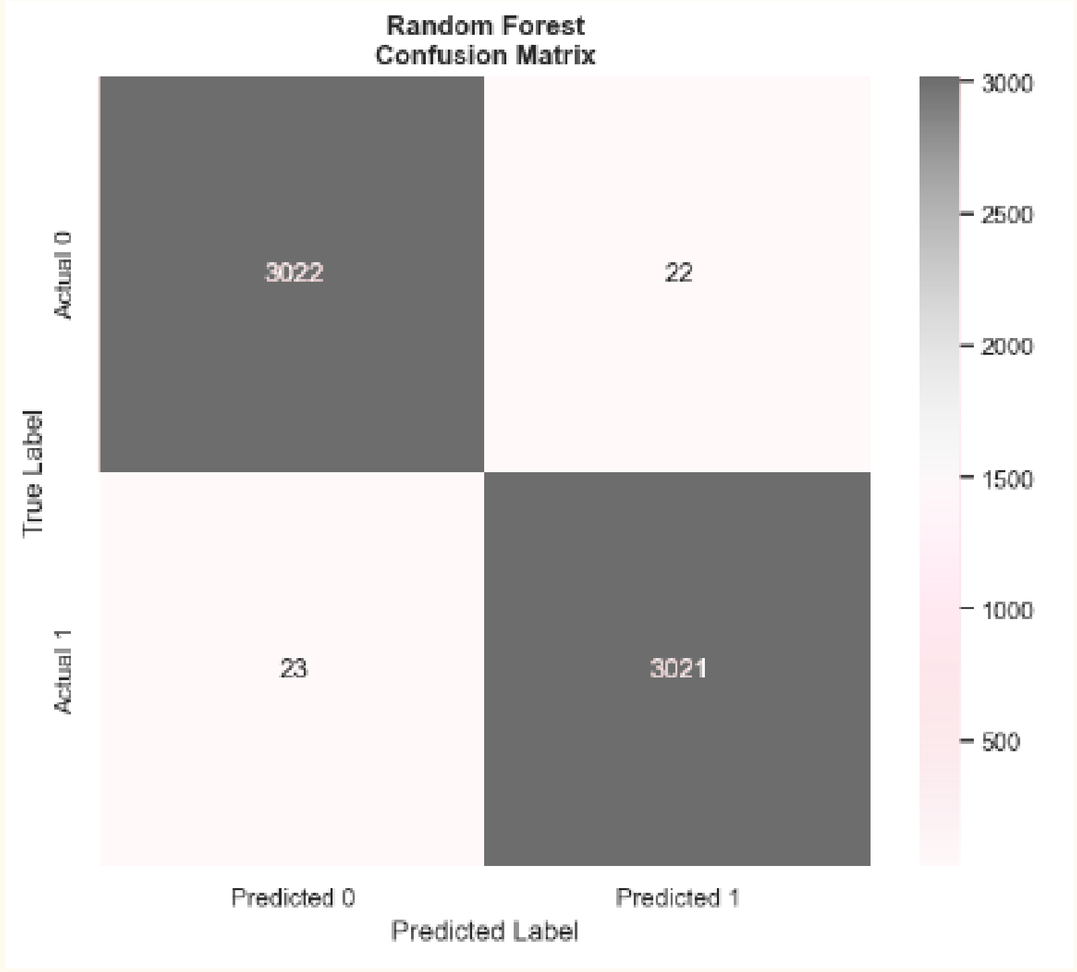
- Risk Score: Root split, meaning it is the strongest determinant. Lower → Loan approved, higher → Loan rejected.
- Debt-to-Income Ratio: Higher DTI indicates financial strain → Increases chance of rejection.
- Monthly Income: Higher income improves repayment capability → Increases likelihood of approval.
- Secondary splits involve Annual Income, Debt-to-Income Ratio, and Loan Amount, showing that income stability and repayment capacity strongly influence the decision.
- The tree shows clear separation of approved vs. not-approved clients early in the structure, indicating high model confidence and low impurity at top splits.

Random Forest



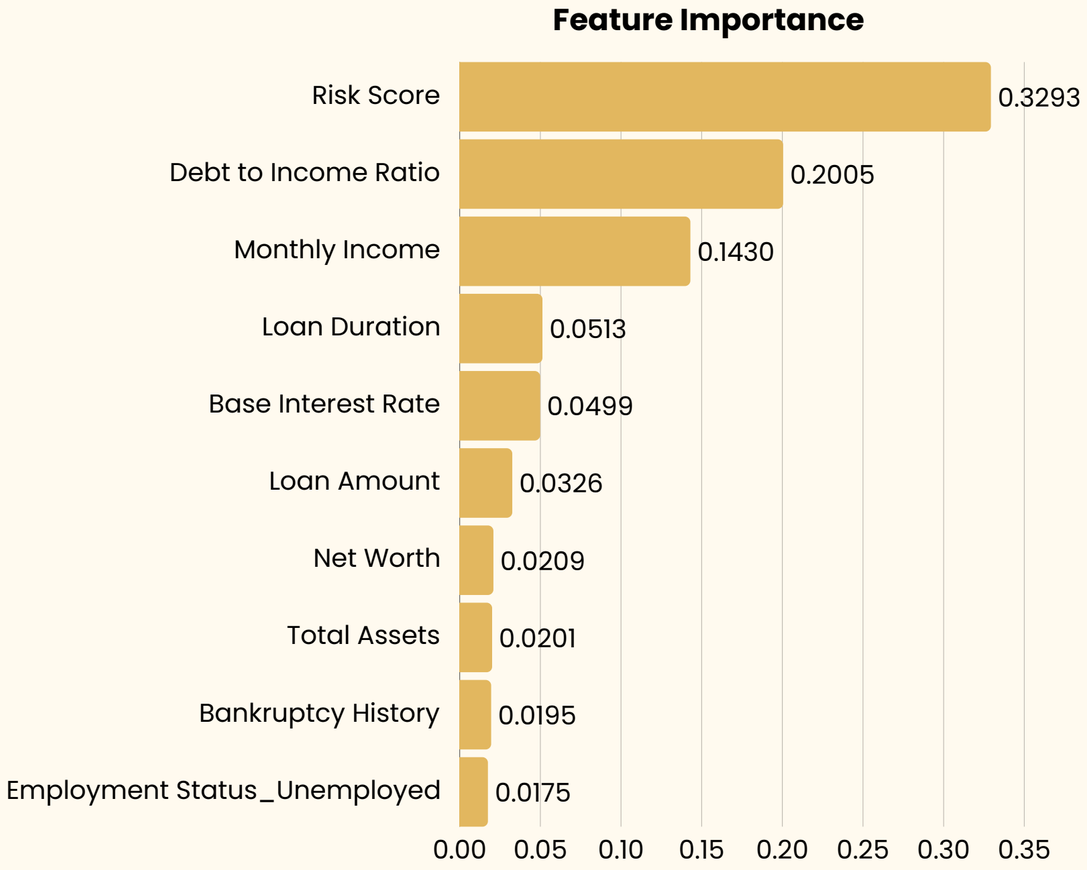
- Top Most Important Predictors - Risk Score, Total Debt to Income Ratio, Monthly Income, Loan Amount, Credit Score. These variables consistently drive loan approval decisions.
- Moderately Important Predictors - Loan Duration, Base Interest Rate, Net Worth, Total Assets, Bankruptcy History
- Low-Importance Predictors - Length of Credit History, Monthly Debt Payments, Previous Loan Defaults
- Delivered excellent predictive performance, with an accuracy of 99.1% and ROC-AUC of 99.9% after tuning.
- Cross-validation confirmed model stability.
- Findings indicate that borrower risk profile, debt burden, and income capacity are the primary drivers of loan approval decisions.

RF - Confusion Matrix



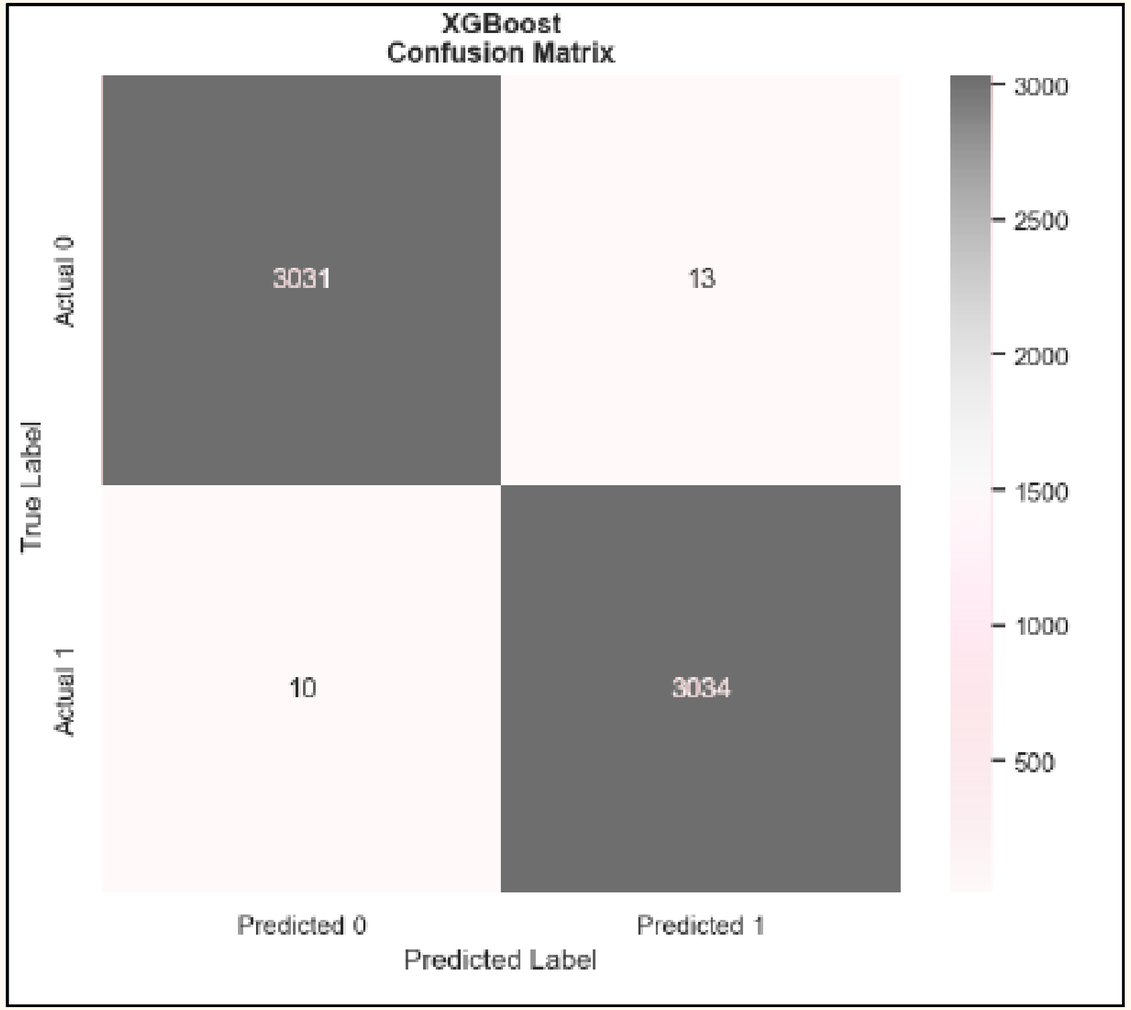
- Correctly classifies the vast majority of observations, indicating strong overall predictive accuracy.
- Only 22 false negatives (predicted not approved but actually approved).
- Only 23 false positives (predicted approved but actually not approved).
- Very high precision (99.28%) and recall (99.24%), showing that the model rarely misclassifies either approved or denied loans.
- Confirms Random Forest as a highly reliable model for predicting loan approval outcomes, with balanced error rates and excellent classification performance.

XG Boost



- Top Most Important Predictors - Risk Score, Total Debt to Income Ratio, Monthly Income, Loan Duration, Base Interest Rate
- Risk, debt burden, income, and ability to repay dominate the decision logic.
- Wealth and employment status play secondary but meaningful roles.
- Delivered the best performance among all four models: highest overall accuracy, perfect ROC-AUC, excellent recall of the minority class
- Stable validation metrics (low variance across folds)

XGB - Confusion Matrix



- Perfectly strong performance overall, with 3,031 true negatives and 3,034 true positives dominating the matrix.
- The model makes very few errors, only 13 false negatives and 10 false positives out of 6,088 total predictions.
- Recall ≈ 99.7% → the model almost never misses actual approved cases, meaning it is excellent at correctly identifying “Approved” applicants.
- Precision ≈ 99.6% → when the model predicts “Approved,” it is almost always correct.
- Overall, the confusion matrix reflects near-perfect classification, proving XGBoost is highly reliable for loan approval prediction in this dataset.

Best Performing Model

XG Boost

	Model	F1-Score	Rank
0	XGBoost	0.996224	1
1	Random Forest	0.992607	2
2	Decision Tree	0.989985	3
3	Logistic Regression	0.964895	4

XGBoost delivered the best performance among all four models (LR, RF, DT, XGB) with:

- Highest overall accuracy (99.62%)
- Perfect ROC-AUC (99.99%)
- Excellent recall (99.7%)of the minority class
- Stable validation metrics (low variance across folds)
- Captures both linear and nonlinear relationships effectively.
- Robust to multicollinearity, interactions, and skewed distributions.

Managerial Implications

Automate approvals for low-risk applicants.

Improve speed and consistency of decisions.

Reduced losses from risky approvals.

Use key predictors (Risk Score, DTI, Income & Loan Duration) to refine lending policies and applicant eligibility guidelines.

Lessons Learnt

Throughout the project, we gained important insights into how data quality, model selection, and feature behavior influence predictive performance. These lessons will guide future development and validation efforts.

1.

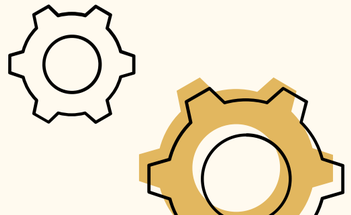
High-quality data preprocessing is essential for strong model performance (Outlier treatment, log transformations, and feature engineering significantly improved accuracy.)

2.

Risk and Debt Burden are the #1 Decision Factors. This mirrors real-world lending where credit risk and affordability dominate underwriting.

3.

XGBoost and Random Forest captured non-linear patterns that Logistic Regression could not



Next Steps

Even though we successfully built and evaluated five predictive models and identified the best-performing one, all results are based on a synthetic dataset. Now that we clearly understand the key drivers influencing loan approval, the next phase is to validate and apply the solution in real-world settings.

1. Partner with small financial institutions to collect real loan application and repayment datasets.

2. Retrain and validate the best-performing model on real-world data to ensure reliability, fairness, and operational accuracy.

3. Deploy the refined model as a decision-support tool to assist lenders in improving approval consistency, reducing risk, and enhancing overall credit assessment efficiency.

References

1. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier.

2. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R and Python (2nd ed.). Springer.

Thank You

We hope our findings help drive smarter, fairer, and more efficient loan approval decisions at JAF Bank.