PREDICTIVE MODELING FOR LOAN APPROVAL

COMPREHENSIVE PROJECT REPORT

PRESENTED BY:

FARAH NAAZ

JAHANAVI GOGINENI

ADEDAMOLA DOSUNMU

DECEMBER 10, 2025

BARNEY SCHOOL OF BUSINESS
UNIVERSITY OF HARTFORD

TABLE OF CONTENTS

1. Introduction

1.1 Description of the Problem

Loan approval is one of the most important decision-making processes available to most financial institutions as it directly affects their financial resolve or profitability, customer satisfaction and stability. Banks should overtly assess the credit worthiness of their loan applicants while balancing the overall goal of minimizing risk and maximizing revenue through loan approval. While digital banking is on the rise combined with an increase in the sum of loan applications, JAF bank faces immense pressure to improve both speed and accuracy of its loan approval process. We can agree that manual review systems lack the resolve to handle large financial data properly and efficiently (Hosmer, Lemeshow, & Sturdivant, 2013). At JAF Bank, borrower risk evaluation and loan approval decisions have traditionally relied on manual assessments and rule-based credit scoring systems. Rule-based systems are time-consuming to maintain, rigid and inflexible, and often fail to capture complex interactions between variables. They are also susceptible to human bias and, importantly, are not predictive, limiting their ability to support accurate, data-driven decision-making.

This study helps address these problems by applying machine learning techniques to develop an automated loa approval prediction system that improves decision accuracy, consistency and efficiency.

1.2 Objectives of the Study

The major objective of the study is to develop a machine learning-based predictive model that pinpoints or accurately classifies loan applications as approved or not approved using the applicant's demographic and financial attributes.

Here is an itemized breakdown of the objectives of the study:

a. Carry out EDA to have an insight into key patterns and relationships amongst applicant variables.

b. Preprocess and balance the dataset.

c. Develop and train many models for loan approval prediction.

d. Compare the performance of the models using appropriate evaluation metrics.

e. Identify the most significant predictors influencing loan approval decisions.

f. Recommend the most suitable model for decision making process.

2. Benchmark Solutions

2.1 Overview of Existing Solutions to the Problem

Loan approval prediction and credit risk assessment has been studied over the past few years by numerous scholars with the aim of trying to develop a better understanding towards designing the perfect model that best suits the scenario at hand, thus maximizing profit while reducing cost implications open to management. Several methods and approaches have been proposed in literature all the way from statistical methods to machine learning techniques. We chose four models to carry out our analysis based on these studies which are discussed below as follows:

Study 1: Machine Learning Models for Credit Risk Assessment

This study which is titled "Enhancing Credit Risk Assessment in Loan Approval: Performance evaluation of Machine Learning Models (Truong, Phan, Truong, & Nguyen, 2025)". The main aim of the study they carried out was to assess and compare the performance of many machine learning algorithms for credit risk classification in loan approval decisions. The researchers used logistic regression, Random Forest and XG boost machine learning models to a well-defined and structured credit risk dataset which contained financial and demographic information.

In this study, XG boost outperformed the other models by achieving an accuracy level of 91.9% and an AUC of 96.5% which made it the most reliable model for forecasting loan approval outcomes. Logistic regression performed the lowest amongst the evaluated models. The conclusion statement of this study showed that boosting techniques provide optimum superior forecasting power for complex financial datasets and should be used for automated credit risk systems.

Study 2: Credit Risk Prediction Using Machine Learning and Deep Learning:

This study tackled credit risk prediction by using traditional machine learning and deep learning methods in their study titled "Credit Risk Prediction Using Machine Learning and Deep learning: A study on Credit Card Customers by (Chang, et al., 2024). Their study focused on forecasting customer credit risk using real world credit card transactions and behavioral data.

The authors measured logistic regression, Adaboost, XG boost, LightGBM and neural networks. Their study showed that XG boost was able to generate the highest forecasting performance with an accuracy of 99.4% which ended up making it the most efficient model in their study. Logistic regression performed the weakest amongst all the models.

Study 3: Machine Learning Framework for Loan Default Prediction:

In this work titled "A proposed framework for loan default prediction using machine learning techniques by (Sharaf Eldin, Idrees, & Ouf, 2025). This study focused on designing a practical and easy to understand forecasting system for financial institutions using well-defined borrower data.

The researchers chose and implemented Decision Tree, Random Forest and gradient boosting algorithms and measured their performance on a loan default dataset. Decision tree model achieved the best performance with an accuracy of 88%. In this study the conclusive statement was that decision trees were highly efficient when forecasting banking environments where transparency and regulatory explainability are vital requirements.

2.2 Justification of the Chosen Methodologies

In this study we have decided to carry out a multi model benchmarking framework using Logistic regression, Decision trees, XG boost, Random Forest to efficiently evaluate predictive performance for loan approval classification. We looked at different related works in the past, compared to our available dataset and selected the mentioned machine learning models based on their results.

Logistic regression was selected as the ***baseline model*** because of its ease of interpretability and simplicity and predominant strong acceptance in the banking industry (Hosmer, Lemeshow, & Sturdivant, 2013). ***Decision tree*** was picked due to its rule-based structure and due to its transparency, which often mimics human self-evaluation in loan decision making. ***Random forest*** was selected due to its ability to handle large and complex data sets, its clear opposition to overfitting and firm generalization performance. It is predominantly regarded by analysts as one of the most widely accepted algorithms for credit risk modeling in real world banking environments (Breiman, Friedman, Olshen, & Stone, 1984). ***XG Boost*** is added because of its high-performance boosting algorithm nature of capturing complex non-linear relationships with superior predictive accuracy (Chen & Guestrin, 2016). Both Truong et al. (2025) and Chang et al. (2024) recorded XG boost as the best performing models which also justified our inclusion of XG boost in this study.

3. Data Collection and Preparation

3.1 Source of the Dataset

https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval/data

We used a dataset from **Kaggle** to leverage historical data and machine learning techniques.

3.2 Data Preprocessing

Based on the comprehensive data preprocessing conducted, the dataset was found to be complete and consistent, with ***no missing values*** or ***duplicate entries***, ensuring a robust foundation for analysis.

To prepare the data for temporal analysis and improve readability, we performed two key preprocessing steps. First, the **'ApplicationDate'** column, initially stored as text, was converted into a proper datetime format. This conversion enabled us to extract meaningful time-based features, such as the application year, which was later used for inflation adjustment. Second, to enhance the clarity of our dataset, we systematically ***renamed all column headers*** from their original camelCase format (e.g., 'AnnualIncome', 'CreditScore') into a more human-readable format with spaces (e.g., 'Annual Income', 'Credit Score').

3.2.1 Inflation Adjustment

To account for economic changes over time, monetary columns such as Annual Income, Loan Amount, and Monthly Income were adjusted for inflation and converted to 2010-dollar values using Consumer Price Index (CPI) data, thereby enabling fair and temporally consistent comparisons across all financial features in the model. Thereafter, we dropped the original columns (Annual Income, Loan Amount, Monthly Income, Savings Account Balance, Checking Account Balance, Total Assets, Total Liabilities, Net Worth) and retained the inflation adjusted columns of these features.

3.2.2 Outlier detection

To ensure the robustness of our predictive models, we conducted a thorough analysis and treatment of outliers within key numerical features. Outliers, which are extreme values that deviate significantly from most of the data, can distort statistical analyses and impair model performance. We employed a multi-strategy approach tailored to the distribution and nature of each variable.

For features like ***Annual Income, Monthly Income, Loan Amount***, and account balances, we applied ***Winsorization***. This technique caps extreme values by replacing the top and bottom 3% of data with the values at the 3rd and 97th percentiles, respectively. This effectively reduces the influence of the most extreme outliers while preserving the overall structure and sample size of the dataset.

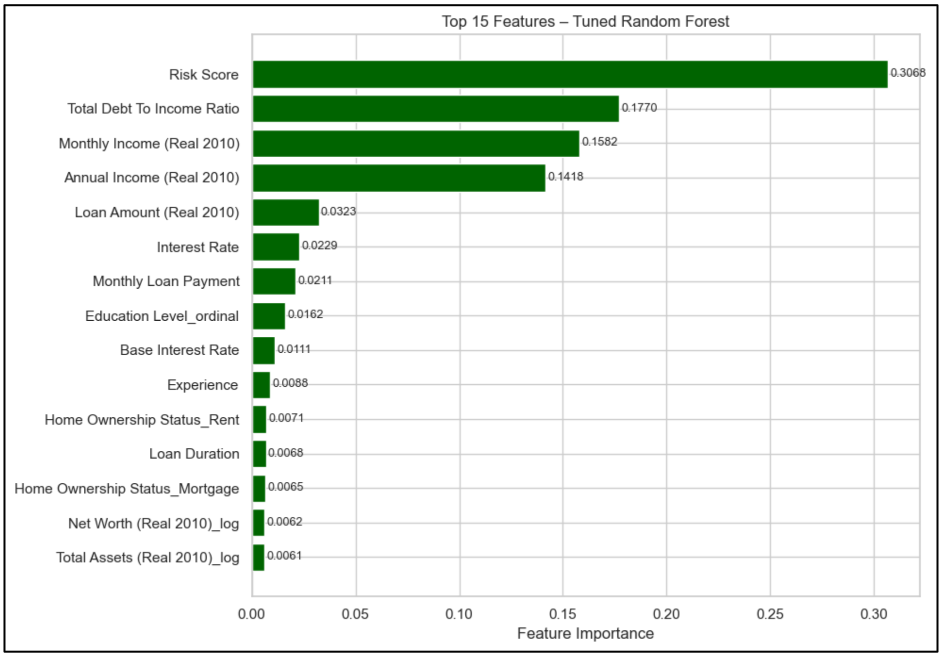
For variables representing regular payments, such as ***Monthly Debt Payments*** and ***Monthly Loan Payment***, we used the ***Interquartile Range (IQR) Capping method***. This method identifies outliers as values falling below Q1 - 1.5*IQR or above Q3 + 1.5*IQR and replaces them with these calculated bounds. This approach is effective for mitigating the skew caused by unusually high or low payment values without removing data points.

Finally, for highly skewed wealth indicators like ***Total Assets**, **Total Liabilities***, and ***Net Worth***, we applied a ***Log Transformation***. By taking the natural logarithm of these values (after adding 1 to handle zeros), we compressed the scale of the data, pulling in extremely long right tails and making the distributions more symmetrical and closer to normal. This transformation is particularly beneficial for linear models, as it helps meet the assumption of normally distributed features and stabilizes variance.

3.2.3 Feature Selection

Based on the correlation matrix, Variance Inflation factor (VIF) and Random Forest, the feature selection was done.

Figure 1: Feature Selection



3.2.4 Normalization / Standardization

During the training of the ***Logistic Regression model with Principal Component Analysis (PCA)***, a ***StandardScaler*** was applied as a critical step within the model pipeline. This scaler standardizes features by removing the mean and scaling to unit variance, transforming the data so that each numerical feature has a mean of zero and a standard deviation of one. This normalization is essential for PCA, as the technique is sensitive to the variances of the initial variables, and it is a best practice for logistic regression to ensure all coefficients are on a comparable scale, preventing features with larger ranges from dominating the model.

3.3 EDA Summary:

The dataset consists of ***20,000 records***, with ***36 features***, comprising both categorical and continuous variables. These include demographic and financial attributes such as age, income, credit score, loan amount, and employment status, as well as economic indicators like debt-to-income ratio and credit utilization. The ***target variable*** indicates whether the ***loan was approved (1) or not approved (0).*** The dataset has been expanded using SMOTENN to simulate new data points, ensuring a balanced distribution of both categorical and continuous features for robust model training.

According to *Figure 1* below, the data is exceptionally imbalanced, with a rate of 23.9% for approved customers.

Figure 2: Distribution of Target Variable

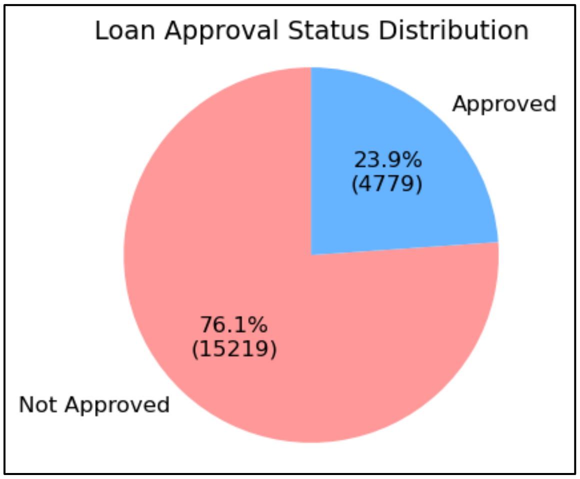
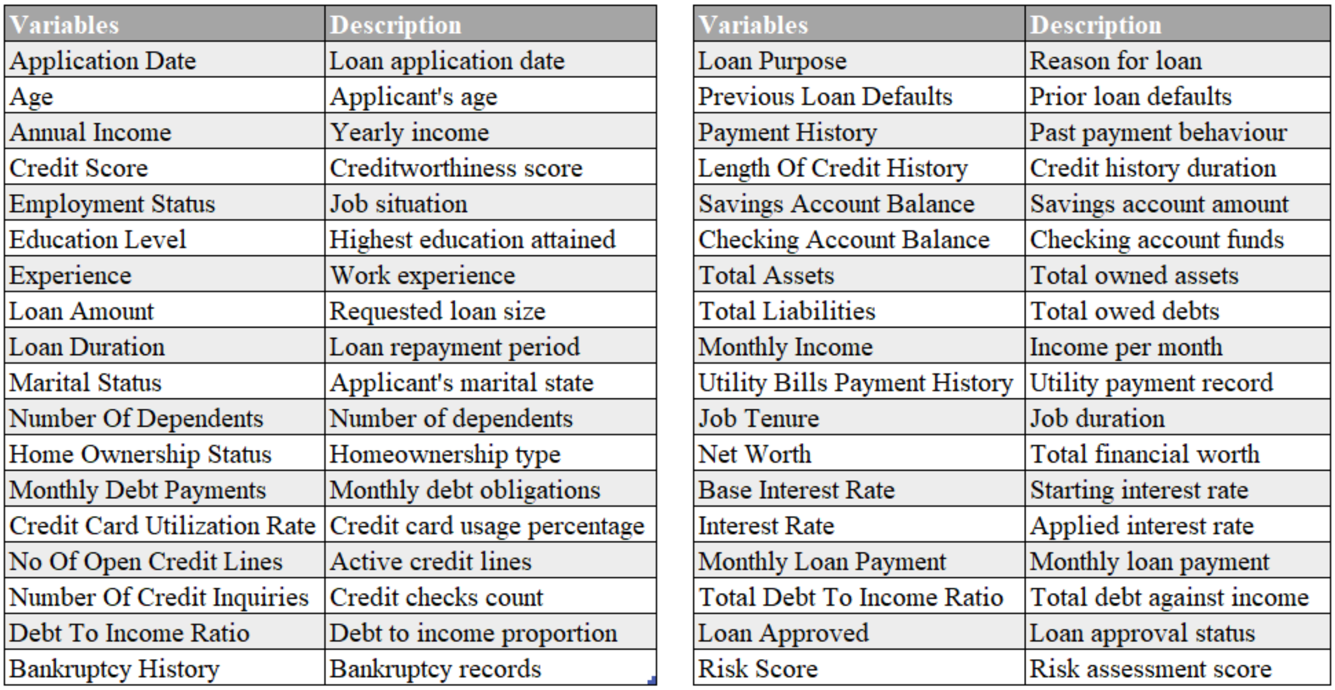
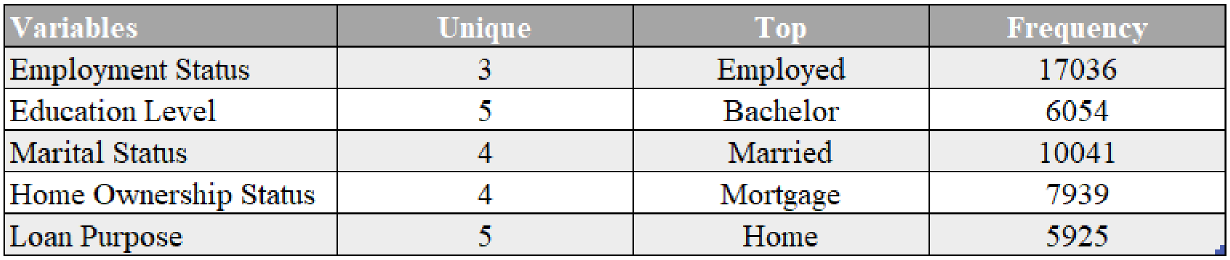


Table 1: Loan Approval Dataset Variables



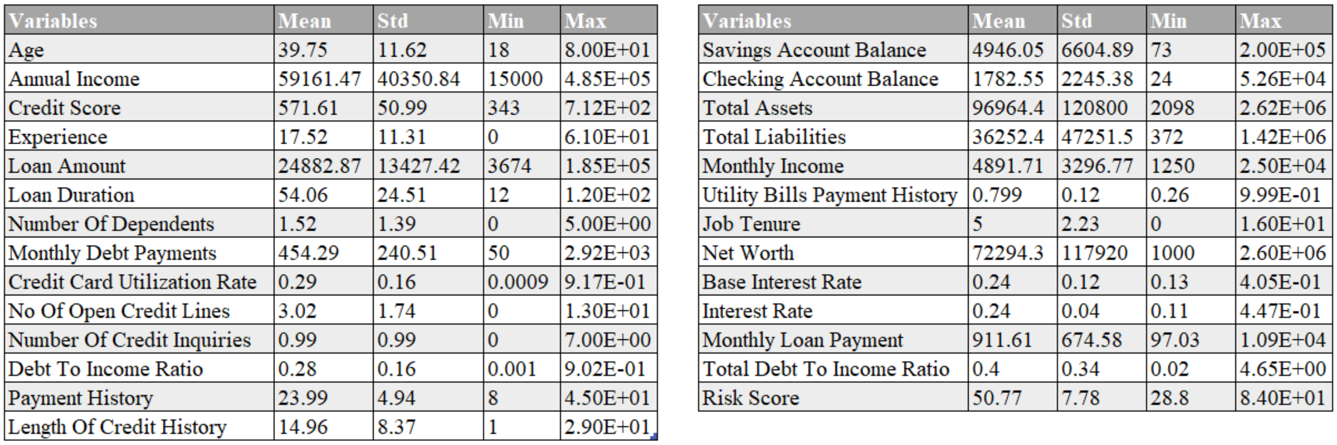
Variables	Description
Application Date	Loan application date
Age	Applicant's age
Annual Income	Yearly income
Credit Score	Creditworthiness score
Employment Status	Job situation
Education Level	Highest education attained
Experience	Work experience
Loan Amount	Requested loan size
Loan Duration	Loan repayment period
Marital Status	Applicant's marital state
Number Of Dependents	Number of dependents
Home Ownership Status	Homeownership type
Monthly Debt Payments	Monthly debt obligations
Credit Card Utilization Rate	Credit card usage percentage
No Of Open Credit Lines	Active credit lines
Number Of Credit Inquiries	Credit checks count
Debt To Income Ratio	Debt to income proportion
Bankruptcy History	Bankruptcy records
Loan Purpose	Reason for loan
Previous Loan Defaults	Prior loan defaults
Payment History	Past payment behaviour
Length Of Credit History	Credit history duration
Savings Account Balance	Savings account amount
Checking Account Balance	Checking account funds
Total Assets	Total owned assets
Total Liabilities	Total owed debts
Monthly Income	Income per month
Utility Bills Payment History	Utility payment record
Job Tenure	Job duration
Net Worth	Total financial worth
Base Interest Rate	Starting interest rate
Interest Rate	Applied interest rate
Monthly Loan Payment	Monthly loan payment
Total Debt To Income Ratio	Total debt against income
Loan Approved	Loan approval status
Risk Score	Risk assessment score

Table 2: Summary Statistics of Categorical Features



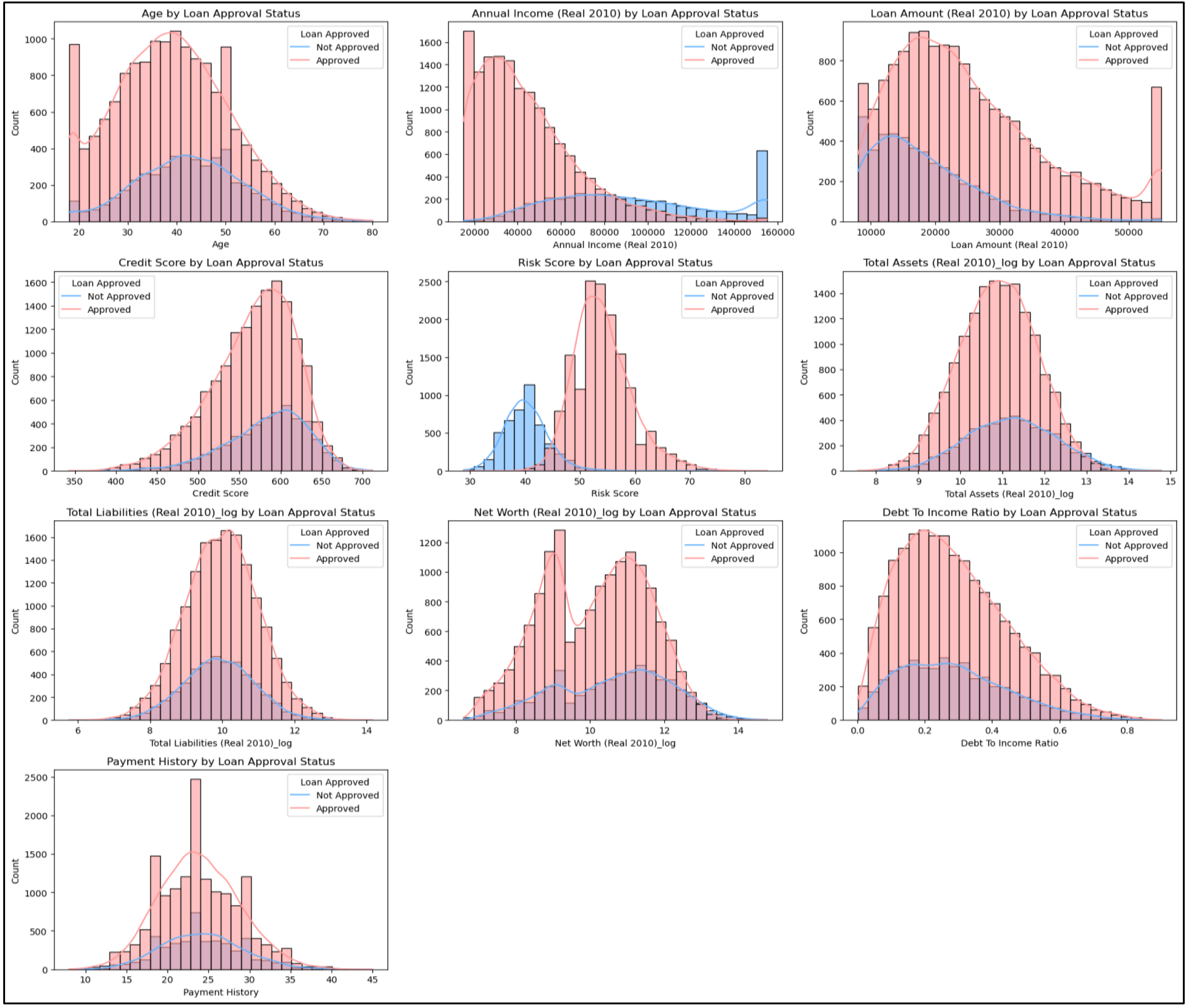
Variables	Unique	Top	Frequency
Employment Status	3	Employed	17036
Education Level	5	Bachelor	6054
Marital Status	4	Married	10041
Home Ownership Status	4	Mortgage	7939
Loan Purpose	5	Home	5925

Table 3: Summary Statistics of Numerical Features



Variables	Mean	Std	Min	Max
Age	39.75	11.62	18	8.00E+01
Annual Income	59161.47	40350.84	15000	4.85E+05
Credit Score	571.61	50.99	343	7.12E+02
Experience	17.52	11.31	0	6.10E+01
Loan Amount	24882.87	13427.42	3674	1.85E+05
Loan Duration	54.06	24.51	12	1.20E+02
Number Of Dependents	1.52	1.39	0	5.00E+00
Monthly Debt Payments	454.29	240.51	50	2.92E+03
Credit Card Utilization Rate	0.29	0.16	0.0009	9.17E-01
No Of Open Credit Lines	3.02	1.74	0	1.30E+01
Number Of Credit Inquiries	0.99	0.99	0	7.00E+00
Debt To Income Ratio	0.28	0.16	0.001	9.02E-01
Payment History	23.99	4.94	8	4.50E+01
Length Of Credit History	14.96	8.37	1	2.90E+01
Savings Account Balance	4946.05	6604.89	73	2.00E+05
Checking Account Balance	1782.55	2245.38	24	5.26E+04
Total Assets	96964.4	120800	2098	2.62E+06
Total Liabilities	36252.4	47251.5	372	1.42E+06
Monthly Income	4891.71	3296.77	1250	2.50E+04
Utility Bills Payment History	0.799	0.12	0.26	9.99E-01
Job Tenure	5	2.23	0	1.60E+01
Net Worth	72294.3	117920	1000	2.60E+06
Base Interest Rate	0.24	0.12	0.13	4.05E-01
Interest Rate	0.24	0.04	0.11	4.47E-01
Monthly Loan Payment	911.61	674.58	97.03	1.09E+04
Total Debt To Income Ratio	0.4	0.34	0.02	4.65E+00
Risk Score	50.77	7.78	28.8	8.40E+01

Figure 3: Distribution of Numerical Features

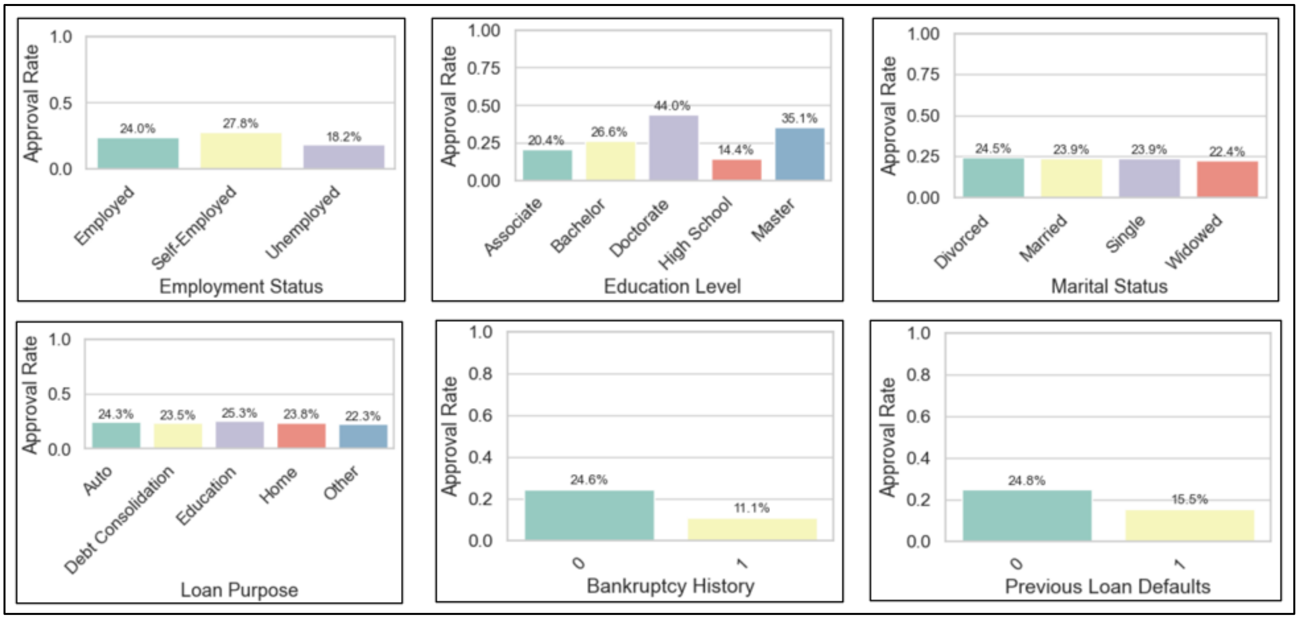


3.3.1 Distribution of Numerical columns (Figure 3 above)

- *Demographic Profile:* Approved borrowers tend to be in the 30–45 age range, whereas rejected applicants are more dispersed and relatively more common at younger (<25) and older (>55) ages, reflecting a preference for borrowers in their prime earning years.
- *Income and Loan Size:* Approved loans are most frequent for moderate annual incomes (~20,000– 60,000 Real 2010) and loan amounts around 15,000–30,000, while very high incomes and very high loan amounts show proportionally more rejections, suggesting JAF Bank favors “typical” retail borrowers and may be more conservative at the extremes.

- *Credit & Risk Measures:* Approved applicants consistently display higher credit scores (520–650) and higher risk scores (48–65) than rejected ones (credit often 400–600, risk 35–50), confirming that the current approval process aligns closely with traditional creditworthiness indicators.
- *Wealth & Balance Sheet Strength:* After log-transform and inflation adjustment, approved customers exhibit higher total assets and higher net worth (log values typically in the 10.5–11.5 and 9–12+ bands, respectively), while rejected customers are more concentrated in the lower asset and net-worth ranges, indicating that stronger balance sheets materially improve approval odds.
- *Liabilities & Leverage:* Total liabilities (log) for approved and rejected borrowers both centre around 9.5–10.5, but rejected cases appear relatively more often at the extremes, suggesting that both very low liabilities (possibly thin-file or low-credit histories) and very high liabilities (over-leveraged clients) are riskier from the bank's perspective.
- *Affordability & Debt Burden:* Debt-to-Income Ratio is one of the clearest separators: approved loans cluster in the 0.10–0.35 band, while rejected loans more frequently fall above 0.40, demonstrating that JAF Bank implicitly enforces a DTI affordability threshold.

Figure 4: Categorical Predictors vs Loan Approved (by rate)



3.3.2 Categorical Predictors vs Loan Approved (Figure 4 above)

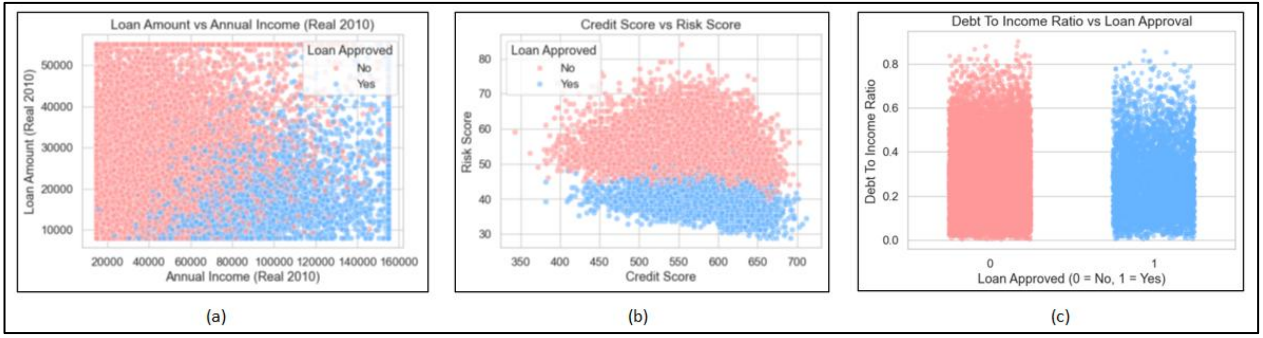
How do approval rates change by education level, employment type, home ownership, bankruptcy history, etc.? Analysis of categorical predictors reveals several meaningful patterns in loan approval outcomes:

- *Employment Type:* Self-Employed applicants exhibit the highest approval rates, suggesting that JAF Bank tends to approve entrepreneurial individuals who show strong financial independence or business income stability.
- *Education Level:* Applicants with Master's and Doctorate degrees have the strongest approval likelihood, indicating that higher educational attainment is viewed as a proxy for income potential, job stability, and lower long-term credit risk.
- *Marital Status:* Approval rates appear uniform across all marital status categories, showing that marital status is not a determining factor in the lending decision.
- *Home Ownership:* Whether an applicant rents, owns, or lives with family does not significantly influence approval outcomes, implying that home ownership is not heavily weighted in the bank's credit evaluation for this product.

- *Loan Purpose:* The reason for seeking a loan shows no meaningful impact on approval decisions, suggesting that JAF Bank evaluates applicants primarily on financial capacity and risk, not on loan intent.
- *Bankruptcy History:* Applicants without a bankruptcy history have noticeably higher approval rates, reflecting the bank’s caution toward individuals with past insolvency events and reinforcing bankruptcy as a key risk indicator.

3.3.3 Essential Relationships

Figure 5: Key Relationships between Features



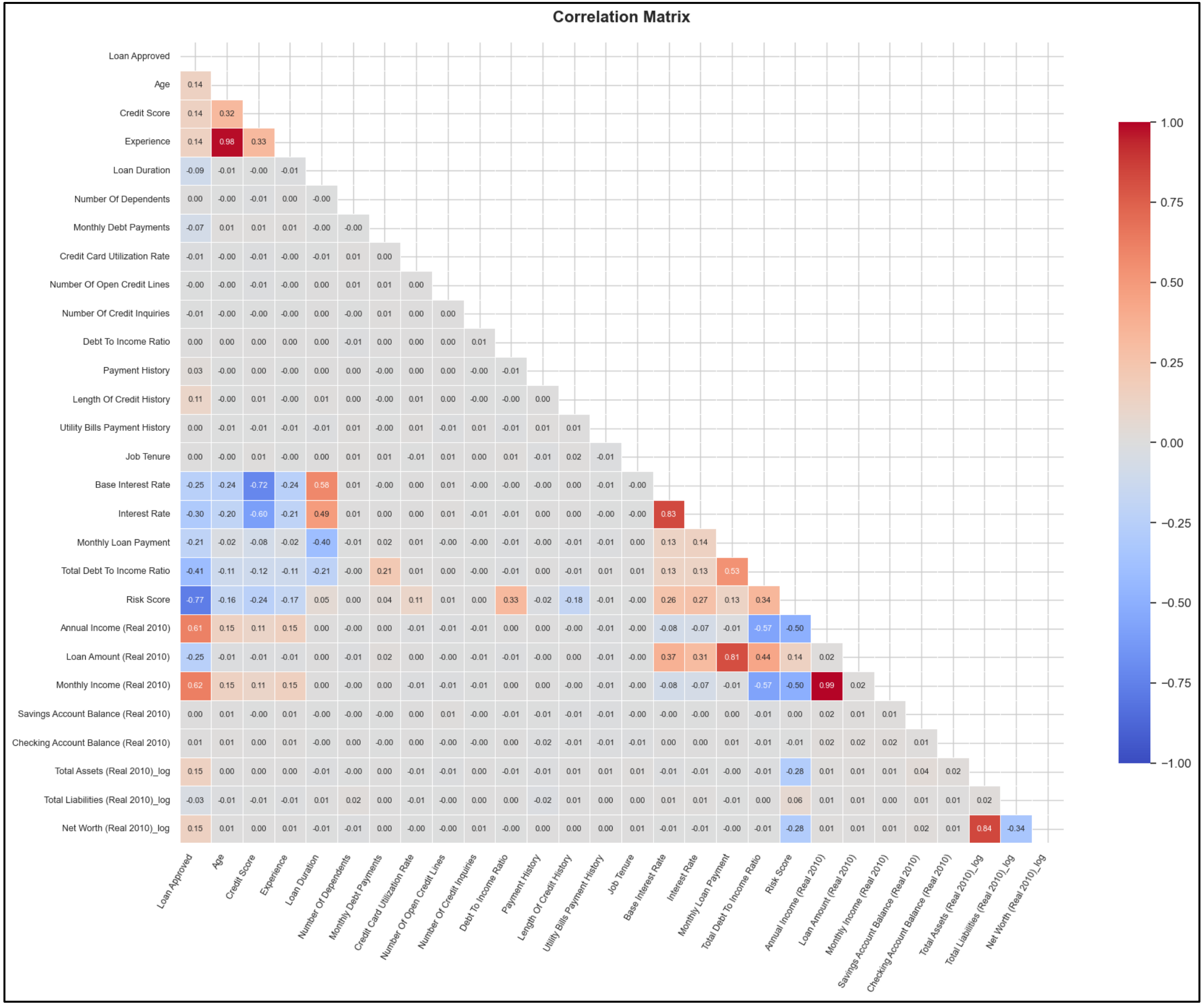
Loan Amount Vs Annual Income (Figure 5-a) - The scatterplot shows that loan approvals are much more common at higher annual incomes and moderate loan amounts, while applicants with lower incomes across the loan amount range are predominantly rejected, indicating that income level is a key driver of approval decisions.

***Credit Score vs Risk Score (Figure 5*-b)** - 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 in this dataset lower Risk Score values combined with stronger credit scores are associated with higher approval likelihood.

***Debt to Income Ratio vs Loan Approval (Figure 5*-c)** - The scatterplot clearly shows that approved applicants consistently have lower Debt to-Income ratios, while higher DTI values (especially above ~0.35–0.40) are concentrated among rejected borrowers, confirming DTI as a strong negative predictor of loan approval.

3.3.4 Correlation Analysis

Figure 6: Correlation Matrix



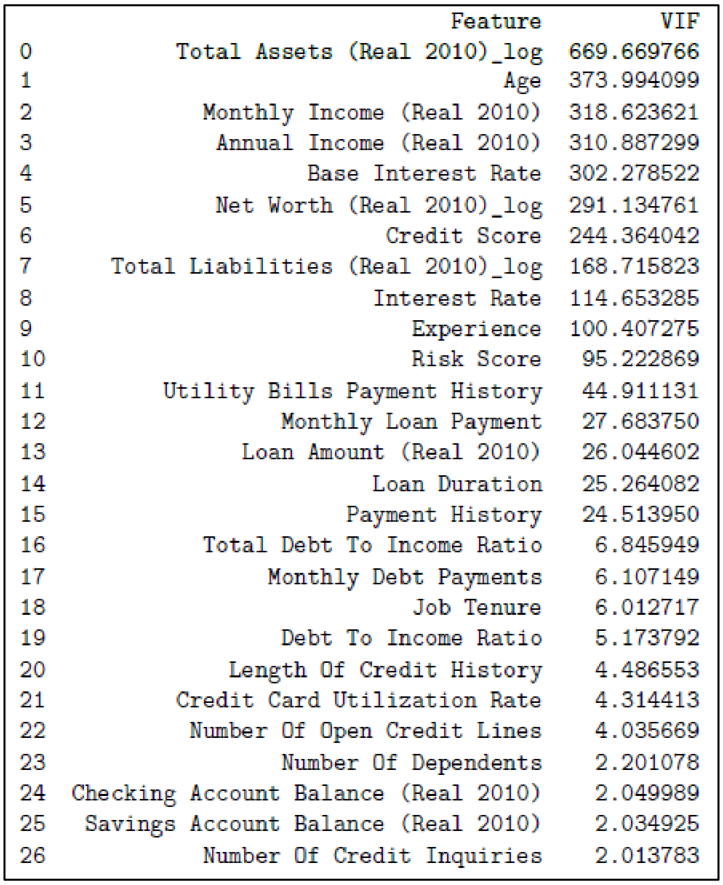
	Loan Approved	Age	Credit Score	Experience	Loan Duration	Number Of Dependents	Monthly Debt Payments	Credit Card Utilization Rate	Number Of Open Credit Lines	Number Of Credit Inquiries	Debt To Income Ratio	Payment History	Length Of Credit History	Utility Bills Payment History	Job Tenure	Base Interest Rate	Interest Rate	Monthly Loan Payment	Total Debt To Income Ratio	Risk Score	Annual Income (Real 2010)	Loan Amount (Real 2010)	Monthly Income (Real 2010)	Savings Account Balance (Real 2010)	Checking Account Balance (Real 2010)	Total Assets (Real 2010)_log	Total Liabilities (Real 2010)_log	Net Worth (Real 2010)_log
Loan Approved																												
Age	0.14																											
Credit Score	0.14	0.32																										
Experience	0.14	0.98	0.33																									
Loan Duration	-0.09	-0.01	-0.00	-0.01																								
Number Of Dependents	0.00	-0.00	-0.01	0.00	-0.00																							
Monthly Debt Payments	-0.07	0.01	0.01	0.01	-0.00	-0.00																						
Credit Card Utilization Rate	-0.01	-0.00	-0.01	-0.00	-0.01	0.01	0.00																					
Number Of Open Credit Lines	-0.00	-0.00	-0.01	-0.00	0.00	0.01	0.01	0.00																				
Number Of Credit Inquiries	-0.01	-0.00	-0.00	-0.00	0.00	-0.00	0.01	0.00	0.00																			
Debt To Income Ratio	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.01																		
Payment History	0.03	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	-0.01																	
Length Of Credit History	0.11	-0.00	0.01	-0.00	0.01	0.00	-0.00	0.01	0.00	-0.00	-0.00	0.00																
Utility Bills Payment History	0.00	-0.01	-0.01	-0.01	-0.01	0.01	-0.00	0.01	-0.01	0.01	-0.01	0.01	0.01															
Job Tenure	0.00	-0.00	0.01	-0.00	0.00	0.01	0.01	-0.01	0.01	0.00	0.01	-0.01	0.02	-0.01														
Base Interest Rate	-0.25	-0.24	-0.72	-0.24	0.58	0.01	-0.00	0.00	0.01	-0.00	-0.00	-0.00	0.00	-0.01	-0.00													
Interest Rate	-0.30	-0.20	-0.60	-0.21	0.49	0.01	0.00	0.00	0.01	-0.01	-0.01	0.00	0.00	-0.00	-0.00	0.83												
Monthly Loan Payment	-0.21	-0.02	-0.08	-0.02	-0.40	-0.01	0.02	0.01	-0.00	-0.00	-0.01	-0.00	-0.01	-0.01	0.00	0.13	0.14											
Total Debt To Income Ratio	-0.41	-0.11	-0.12	-0.11	-0.21	-0.00	0.21	0.01	0.00	-0.00	-0.01	0.00	-0.01	0.01	0.01	0.13	0.13	0.53										
Risk Score	-0.77	-0.16	-0.24	-0.17	0.05	0.00	0.04	0.11	0.01	0.00	0.33	-0.02	-0.18	-0.01	-0.00	0.26	0.27	0.13	0.34									
Annual Income (Real 2010)	0.61	0.15	0.11	0.15	0.00	-0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	-0.00	-0.01	-0.00	-0.08	-0.07	-0.01	-0.57	-0.50								
Loan Amount (Real 2010)	-0.25	-0.01	-0.01	-0.01	0.00	-0.01	0.02	0.00	-0.00	-0.01	-0.01	-0.00	0.00	-0.01	-0.00	0.37	0.31	0.81	0.44	0.14	0.02							
Monthly Income (Real 2010)	0.62	0.15	0.11	0.15	0.00	-0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	-0.00	-0.01	-0.00	-0.08	-0.07	-0.01	-0.57	-0.50	0.99	0.02						
Savings Account Balance (Real 2010)	0.00	0.01	-0.00	0.01	-0.00	-0.00	-0.00	0.00	0.01	-0.00	-0.01	-0.01	-0.01	-0.01	-0.00	-0.00	-0.00	0.00	-0.01	0.00	0.02	0.01	0.01					
Checking Account Balance (Real 2010)	0.01	0.01	0.00	0.01	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.02	-0.01	-0.01	-0.01	0.00	0.00	0.01	-0.01	-0.01	0.02	0.02	0.02	0.01				
Total Assets (Real 2010)_log	0.15	0.00	0.00	0.00	-0.01	-0.00	0.00	-0.01	-0.01	0.00	-0.01	-0.01	0.01	0.01	0.01	-0.01	-0.01	-0.00	-0.01	-0.28	0.01	0.01	0.01	0.04	0.02			
Total Liabilities (Real 2010)_log	-0.03	-0.01	-0.01	-0.01	0.01	0.02	0.00	-0.01	-0.00	0.00	0.00	-0.02	0.01	0.00	0.00	0.01	0.01	-0.01	0.00	0.06	0.01	0.01	0.00	0.01	0.01	0.02		
Net Worth (Real 2010)_log	0.15	0.01	0.00	0.01	-0.01	-0.01	0.00	-0.00	-0.00	0.01	-0.00	0.00	0.00	0.00	0.01	-0.01	-0.01	-0.00	-0.01	-0.28	0.01	0.01	0.01	0.02	0.01	0.84	-0.34	

- The correlation values with Loan Approved indicate that Monthly Income and Annual Income are the strongest positive predictors of loan approval, with correlations of 0.62 and 0.61, respectively, suggesting that higher-income applicants are significantly more likely to be approved by the bank.
- Wealth indicators such as Total Assets (0.1467) and Net Worth (0.1459) also show positive correlations, reflecting that applicants with stronger financial positions tend to receive approvals more frequently.

- Among non-financial attributes, variables like Credit Score (0.142), Age (0.141), Experience (0.1407), and Length of Credit History (0.1059) have mild positive correlations, indicating that stable credit backgrounds and longer financial histories slightly increase approval likelihood, though the strength of these relationships is modest.
- Payment behavior shows only a weak positive association: Payment History (0.038) and Utility Bills Payment History (0.0038) have very low correlations, suggesting these variables alone do not strongly influence approval decisions.
- On the negative side, Risk Score (-0.766) has the strongest negative correlation with loan approval, indicating that higher risk scores (in this dataset) are strongly associated with rejection, confirming its role as a highly predictive risk indicator.
- Similarly, Base Interest Rate (-0.247) and Interest Rate (-0.301) have moderately negative correlations, suggesting that loans offered at higher interest rates tend to be associated with riskier applicants who are more frequently rejected.
- Variables like Loan Amount (Real 2010) (-0.25), Monthly Loan Payment (-0.206), and Total Debt to Income Ratio (-0.41) also show moderate negative correlations, indicating that higher debt obligations or larger requested loan amounts reduce the likelihood of approval.

3.3.5 Variance Inflation Factor (VIF)

Table 4: VIF Results



```
                                   Feature         VIF
0            Total Assets (Real 2010)_log  669.669766
1                                      Age  373.994099
2               Monthly Income (Real 2010)  318.623621
3                Annual Income (Real 2010)  310.887299
4                       Base Interest Rate  302.278522
5              Net Worth (Real 2010)_log  291.134761
6                             Credit Score  244.364042
7       Total Liabilities (Real 2010)_log  168.715823
8                            Interest Rate  114.653285
9                               Experience  100.407275
10                              Risk Score   95.222869
11            Utility Bills Payment History   44.911131
12                    Monthly Loan Payment   27.683750
13                 Loan Amount (Real 2010)   26.044602
14                           Loan Duration   25.264082
15                         Payment History   24.513950
16              Total Debt To Income Ratio    6.845949
17                   Monthly Debt Payments    6.107149
18                              Job Tenure    6.012717
19                    Debt To Income Ratio    5.173792
20                 Length Of Credit History    4.486553
21            Credit Card Utilization Rate    4.314413
22             Number Of Open Credit Lines    4.035669
23                    Number Of Dependents    2.201078
24  Checking Account Balance (Real 2010)    2.049989
25   Savings Account Balance (Real 2010)    2.034925
26              Number Of Credit Inquiries    2.013783
```

The VIF table shows extremely high multicollinearity among several financial variables, especially ***Total Assets, Age, Monthly Income, Annual Income, Base Interest Rate, and Net Worth***, all with VIF values far above the acceptable threshold (typically VIF > 10 indicates concern). These features are highly correlated with each other and may distort coefficient estimates in linear models like Logistic Regression.

Variables such as ***Credit Score, Total Liabilities, Interest Rate, and Experience*** also exhibit strong multicollinearity, while features with VIF < 10 show acceptable levels of redundancy.

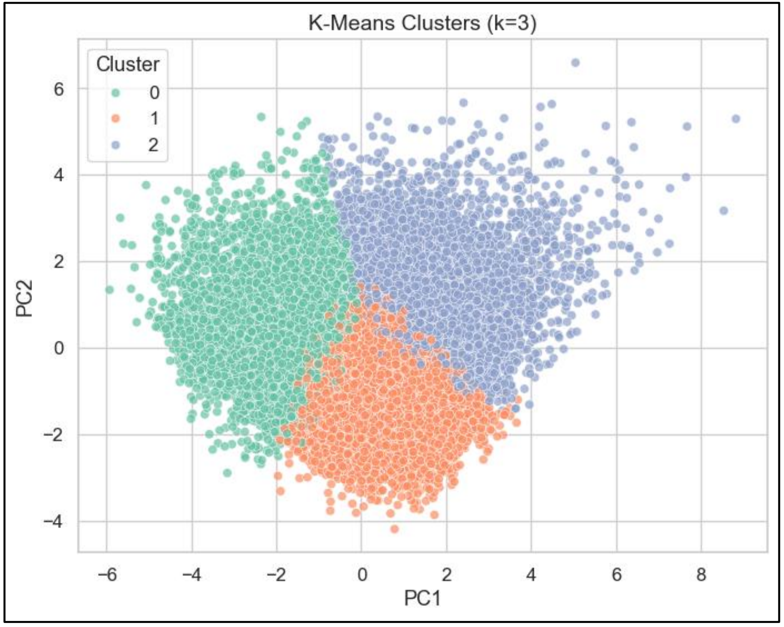
Recommendations for Model Training

- Remove or combine extremely high VIF variables (e.g., Monthly Income vs Annual Income, Total Assets vs Net Worth), as they convey overlapping information.
- Apply dimensionality reduction techniques such as PCA if retaining variance is important.
- Use tree-based models (RF, XGBoost) where multicollinearity is less problematic, but still monitor redundancy for interpretability.

3.3.6 Cluster Analysis:

The K-Means clustering (k=3) in PCA space reveals three clearly separated applicant segments, showing distinct underlying financial behavior patterns even without using the loan approval label. Cluster characteristics confirm meaningful differences across income, credit quality, loan size, and asset strength.

Figure 7: Cluster Analysis



Cluster 0: "Low-Risk, High-Capacity Borrowers" (Approval Rate: 57%). This cluster represents 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%). These applicants have average income and modest assets, making them neither high-risk nor very strong applicants. Their approval rate is significantly lower than Cluster 0, reflecting borderline creditworthiness.

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.

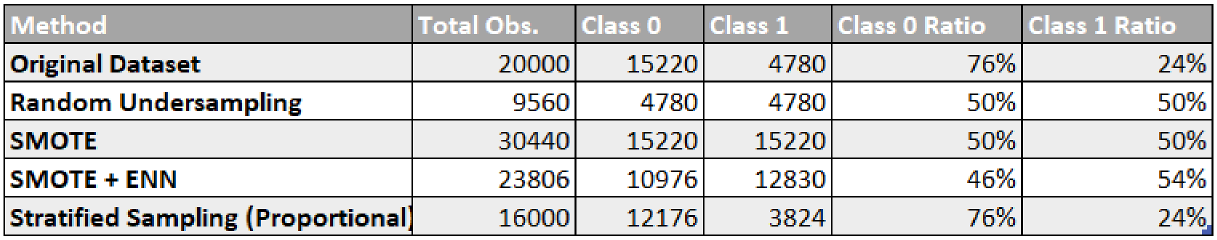
4. Modeling Approach

4.1 Model Selection

This study takes a supervised machine learning framework to make forecasts on loan approval outcomes. It consists of data partitioning, model training and model testing. The models selected were - ***Logistic Regression, Decision tree, Random Forest and XG Boost*** to give room for comparisons between traditional statistical techniques and advanced ensemble learning methods. These structured approaches made sure that models were measured under consistent conditions and that performance results were not biased by data leakage and overfitting (Hastie, Tibshirani, & Friedman, 2017).

4.2 Handling Data Imbalance

Table 5: Class distribution under different imbalance-handling strategies



Method	Total Obs.	Class 0	Class 1	Class 0 Ratio	Class 1 Ratio
Original Dataset	20000	15220	4780	76%	24%
Random Undersampling	9560	4780	4780	50%	50%
SMOTE	30440	15220	15220	50%	50%
SMOTE + ENN	23806	10976	12830	46%	54%
Stratified Sampling (Proportional)	16000	12176	3824	76%	24%

We tested four imbalance-handling techniques - Random Undersampling, SMOTE, SMOTE+ENN, and Stratified Sampling on the original dataset (76% Class 0 vs. 24% Class 1). Among these, ***SMOTE+ENN*** produced the most meaningful balance (46% Class 0, 54% Class 1) while also improving data quality by removing noisy synthetic points. Because this method offered the best trade-off between class balance and dataset integrity, SMOTE+ENN was selected as the final resampling strategy for training all predictive models.

4.3 Cross Validation

We used ***Stratified K-Fold Cross-Validation (K=5)*** to ensure that each fold preserved the original class proportions, making every subset representative of the full dataset. This approach was especially important given the imbalance in our loan approval data, as it prevented biased or skewed splits that could distort model learning. By training and validating on folds that consistently maintained class distribution, the models produced ***more reliable, fair, and stable performance estimates***, ultimately improving the robustness of our evaluation.

4.4 Hyperparameter Tuning

We used ***Randomized Search CV*** as the hyperparameter tuning method because it is the most efficient and practical choice for models with large parameter spaces, such as XGBoost and Random Forest. Unlike Grid Search, which tests every possible combination, Randomized Search evaluates a strategically selected subset of parameters, making the process significantly faster without sacrificing performance quality. This approach greatly reduced computational time while still identifying strong, well-optimized configurations—making it ideal for our gradient boosting–based modelling pipeline.

4.5 Training Process

i) Logistic Regression

To train the Logistic Regression model, we built a pipeline consisting of three steps:

- ***StandardScaler*** – This step standardizes all numeric features, so they have a similar scale.
- ***PCA (Principal Component Analysis)*** – We reduced the dimensionality of the data while keeping 95% of the variance, helping remove noise and multicollinearity and improving model stability.

- ***Logistic Regression Classifier*** – The final step fits the actual model using the optimized LBFGS solver and a higher iteration limit (1000) to ensure proper convergence.

ii) Random Forest

To optimize the Random Forest model, we defined a parameter search space that allowed the algorithm to explore different model configurations during Randomized Search CV:

- ***Number of Trees (n_estimators):*** We tried forests with 150, 200, and 300 trees to see how many trees are needed for strong and stable predictions.
- ***Maximum Tree Depth (max_depth):*** We tested tree depths of 5, 7, 10, and unlimited to compare simpler trees versus more complex ones.
- ***Minimum Samples to Split a Node (min_samples_split):*** Values of 2, 5, and 10 were used to control how detailed each tree becomes when creating new branches.
- ***Minimum Samples in a Leaf (min_samples_leaf):*** We used 1, 2, and 4 to prevent overly small leaf nodes that can lead to overfitting.
- ***Number of Features to Consider at Each Split (max_features):*** We tested "sqrt" and "log2", which randomly limit the number of features evaluated at each split and help improve generalization.

iii) Decision Tree

To tune the Decision Tree model, we tested several configuration options to understand how different tree shapes and splitting rules affect performance:

- ***Maximum Tree Depth (max_depth):*** We tried depths of 3, 4, 5, 6, and unlimited. This allowed us to compare shallow trees (simpler, less overfitting) versus deeper trees (more complex, more detailed splits).
- ***Minimum Samples to Split a Node (min_samples_split):*** We tested 2, 20, 50, and 100 as the minimum number of samples required before a node can be split. Higher values make the tree more conservative and help prevent overfitting.

- ***Minimum Samples in a Leaf (min_samples_leaf):*** Values of 1, 5, 20, and 50 were used to control how many samples must remain in the final leaf nodes. Larger leaf sizes lead to smoother and more generalizable trees.
- ***Splitting Criterion (criterion):*** We compared "gini" and "entropy", two different ways of measuring impurity to decide how the tree chooses the best split. Testing both ensured we captured the most effective splitting strategy.

By exploring these settings through Randomized Search CV, the model was able to determine the best structure and splitting rules, leading to an optimized Decision Tree with strong predictive performance and reduced risk of overfitting.

iv) XG Boost

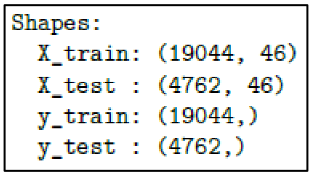
To optimize the XGBoost model, we tested a range of important configuration options that control how trees are built and how the model learns:

- ***Maximum Tree Depth (max_depth):*** We used depths of 3, 5, and 7 to compare shallow versus deeper tree structures. Deeper trees can capture more complex patterns but may risk overfitting.
- ***Learning Rate (learning_rate):*** Values of 0.02, 0.05, and 0.1 were tested. The learning rate controls how quickly the model updates during training, smaller values make learning slower but more precise, while larger values speed learning but may overshoot.
- ***Number of Trees (n_estimators):*** We tried 200, 300, and 500 boosting rounds to determine how many sequential trees the model needs for optimal performance.
- ***Subsample Ratio (subsample):*** The model was tested with 70%, 80%, and 100% of rows sampled for each tree. Lower values help prevent overfitting by adding randomness.
- ***Column Sampling per Tree (colsample_bytree):*** We tried 0.6, 0.8, and 1.0, which specify what fraction of features are randomly selected for each tree. This improves diversity among trees.
- ***Penalty Term for Split Complexity (gamma):*** We used values of 0, 1, and 5. Higher gamma makes the model more conservative by requiring stronger justification for new splits.

Testing these configurations with Randomized Search CV helped identify the best-performing XGBoost model and ensured a strong balance between accuracy, generalization, and computational efficiency.

4.6 Train-Test Split (Testing Process)

We divided the data into an **80:20 split** on training and testing sublets which can be considered widely recommended for datasets of this capacity to guarantee strong generalization performance (Kuhn & Johnson, 2013).



```
Shapes:
  X_train: (19044, 46)
  X_test : (4762, 46)
  y_train: (19044,)
  y_test : (4762,)
```

This shows 19,044 observations were allocated to the train set and 4762 observations were reserved for the test set and each of the predictors contains 46 predictor variables.

5. Results of the Study

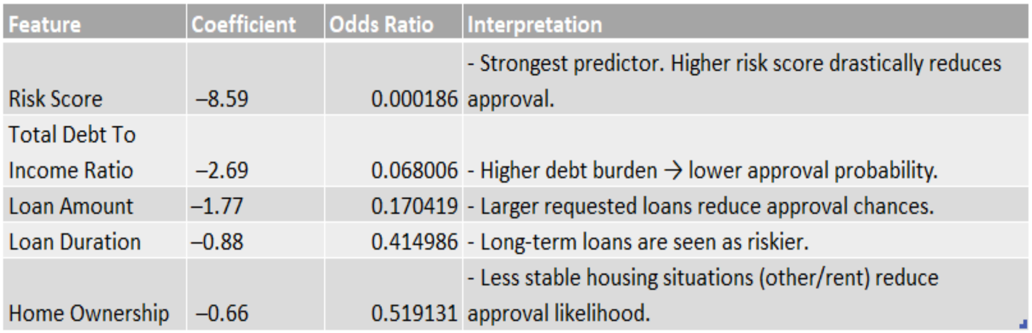
5.1 Model Outcomes

i) **Logistic Regression** – Some of the important results from training our baseline model (logistic regression) are discussed below.

The ***intercept value*** of our model is ***-8.658***, when all predictors are at zero, the log odds of loan approval are strongly negative.

Top Negative Predictors (Reduce Approval Chances) – From table 6 below we notice that the model heavily penalizes risky financial behavior, high borrowing, and weak stability indicators.

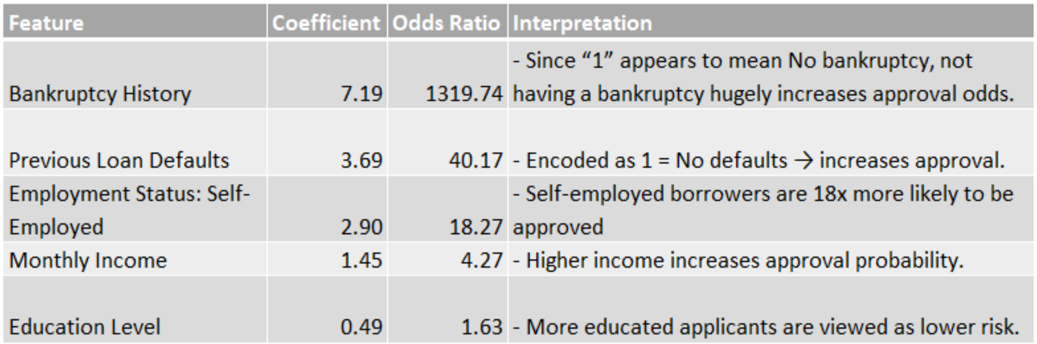
Table 6: Top Negative Predictors



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.

Top Positive Predictors (Increase Approval Chances) – From table 7 below we notice that good financial stability & higher income, strongly boost approval odds.

Table 7: Top Positive Predictors

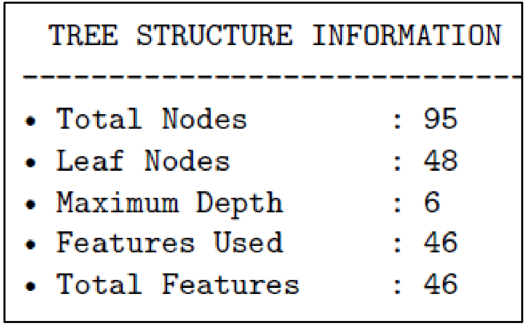


Feature	Coefficient	Odds Ratio	Interpretation
Bankruptcy History	7.19	1319.74	- Since "1" appears to mean No bankruptcy, not having a bankruptcy hugely increases approval odds.
Previous Loan Defaults	3.69	40.17	- Encoded as 1 = No defaults → increases approval.
Employment Status: Self-Employed	2.90	18.27	- 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.

ii) Random Forest – The findings from this model indicate that borrower risk profile, debt burden, and income capacity are the primary drivers of loan approval decisions. These findings are discussed in detail in ***section 5.2*** below.

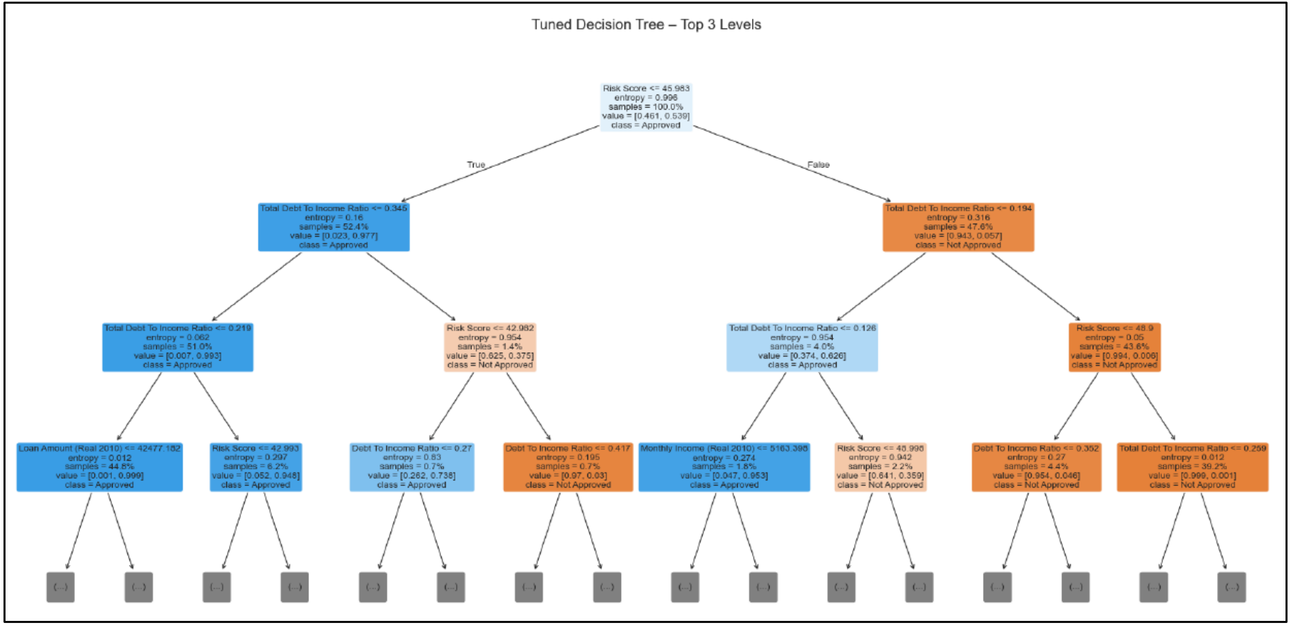
iii) Decision Tree – The tree structure of the model is presented in the figures below.

Figure 8: Tree Structure



```
  TREE STRUCTURE INFORMATION
-----------------------------
• Total Nodes          : 95
• Leaf Nodes           : 48
• Maximum Depth        : 6
• Features Used        : 46
• Total Features       : 46
```

Figure 9: Decision Tree



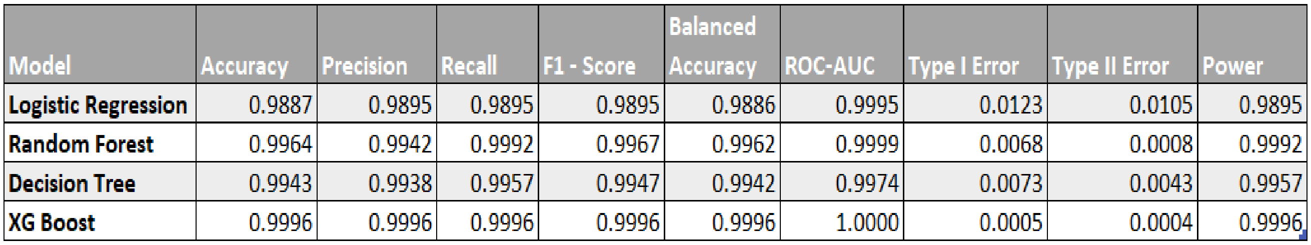
Risk Score is the root split, meaning it is the strongest determinant. A lower risk score improves chances of loan approval while a higher risk score decreases chances of loan approval. Other primary splits include Total Debt to Income Ratio. Secondary splits involve Loan Amount, Debt to Income Ratio and Monthly income, 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.

iv) **XG Boost** - The findings from this model indicate that risk, debt burden, income, and ability to repay dominate the decision logic. Wealth and employment status play secondary but meaningful roles. These findings are discussed in detail in ***section 5.2*** below.

5.2 Model Comparison and Interpretation

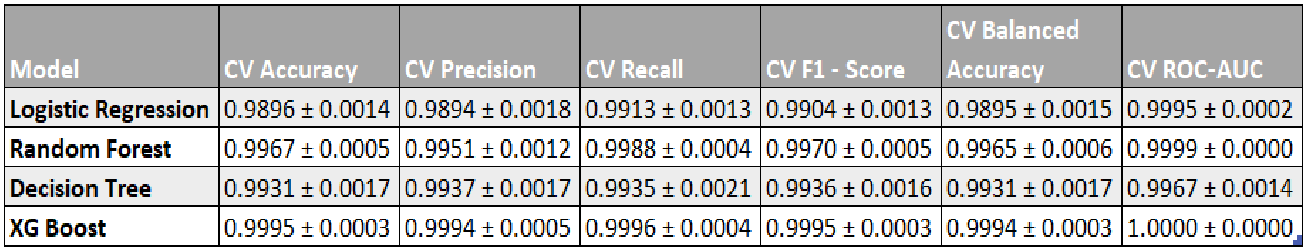
In this section we have included a comprehensive comparison and interpretations of our model performances.

Table 8: Test Set Performance Metrics for Tuned/Final Models



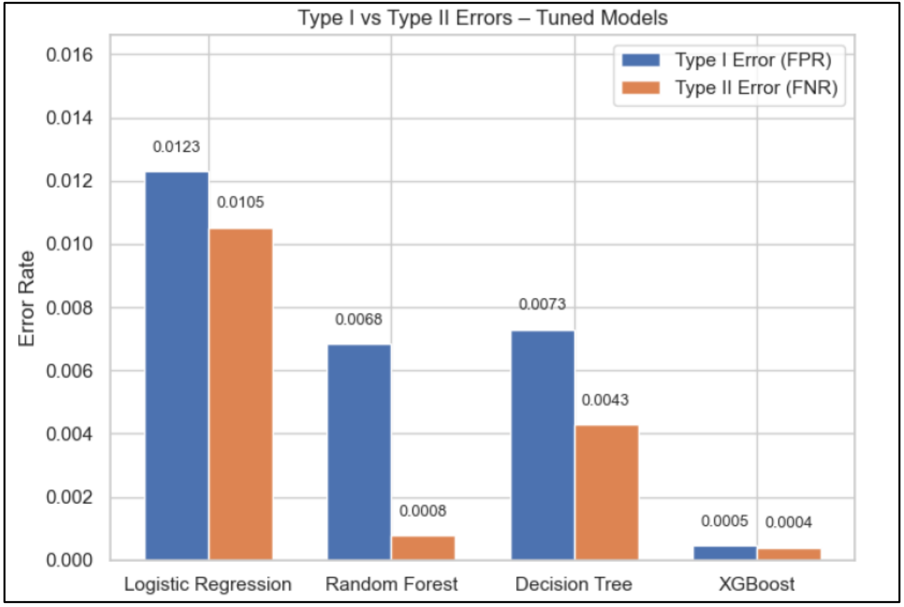
Model	Accuracy	Precision	Recall	F1 - Score	Balanced Accuracy	ROC-AUC	Type I Error	Type II Error	Power
Logistic Regression	0.9887	0.9895	0.9895	0.9895	0.9886	0.9995	0.0123	0.0105	0.9895
Random Forest	0.9964	0.9942	0.9992	0.9967	0.9962	0.9999	0.0068	0.0008	0.9992
Decision Tree	0.9943	0.9938	0.9957	0.9947	0.9942	0.9974	0.0073	0.0043	0.9957
XG Boost	0.9996	0.9996	0.9996	0.9996	0.9996	1.0000	0.0005	0.0004	0.9996

Table 9: 5-Fold Cross Validation Results for Tuned/Final Models (Mean ± Std)



Model	CV Accuracy	CV Precision	CV Recall	CV F1 - Score	CV Balanced Accuracy	CV ROC-AUC
Logistic Regression	0.9896 ± 0.0014	0.9894 ± 0.0018	0.9913 ± 0.0013	0.9904 ± 0.0013	0.9895 ± 0.0015	0.9995 ± 0.0002
Random Forest	0.9967 ± 0.0005	0.9951 ± 0.0012	0.9988 ± 0.0004	0.9970 ± 0.0005	0.9965 ± 0.0006	0.9999 ± 0.0000
Decision Tree	0.9931 ± 0.0017	0.9937 ± 0.0017	0.9935 ± 0.0021	0.9936 ± 0.0016	0.9931 ± 0.0017	0.9967 ± 0.0014
XG Boost	0.9995 ± 0.0003	0.9994 ± 0.0005	0.9996 ± 0.0004	0.9995 ± 0.0003	0.9994 ± 0.0003	1.0000 ± 0.0000

Figure 10: Type I vs Type II Errors

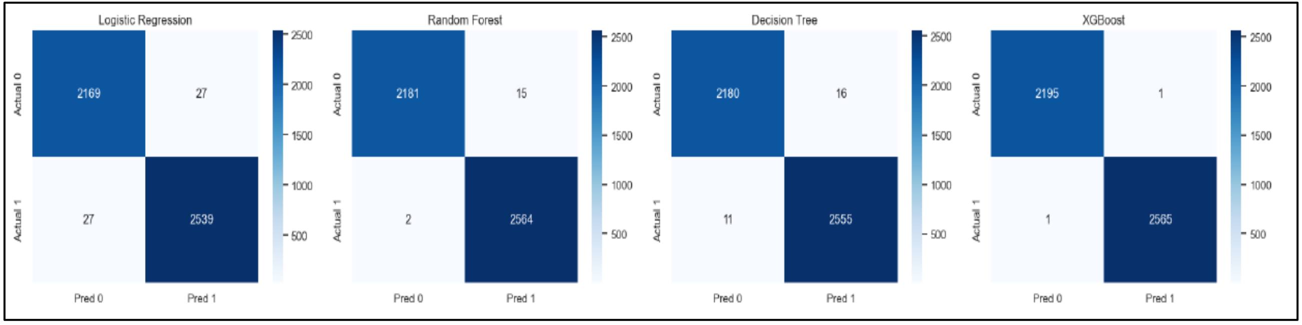


The Test Set Performance Metrics and the 5-Fold Cross-Validation (CV) results provide a comprehensive comparison of the four tuned classification models: Logistic Regression, Random Forest, Decision Tree, and XG Boost. Across both evaluation tables (table 8 and 9) and from figure 10, all models demonstrate strong predictive ability, with consistently high accuracy, precision, recall, F1-score, balanced accuracy, and ROC-AUC values. Error rates (Type I and Type II) are also extremely low, indicating reliable classification with minimal false positives and false negatives.

Among the models, XG Boost is the top performer. It achieves the highest values across almost all test metrics. Accuracy (0.9996), Precision (0.9996), Recall (0.9996), F1-score (0.9996), Balanced Accuracy (0.9996), and ROC-AUC (1.0000), while maintaining the lowest Type I and Type II Errors. The CV results further confirm its stability, with extremely tight standard deviations, indicating that XG Boost performs consistently across different data folds.

Random Forest also performs exceptionally well, ranking closely behind XGBoost, while Logistic Regression and Decision Tree follow with slightly lower (but still strong) performance metrics. Overall, the results clearly ***indicate that XG Boost is the most accurate, robust, and reliable model*** for predicting loan approval outcomes in this dataset.

Figure 11: Confusion Matrices

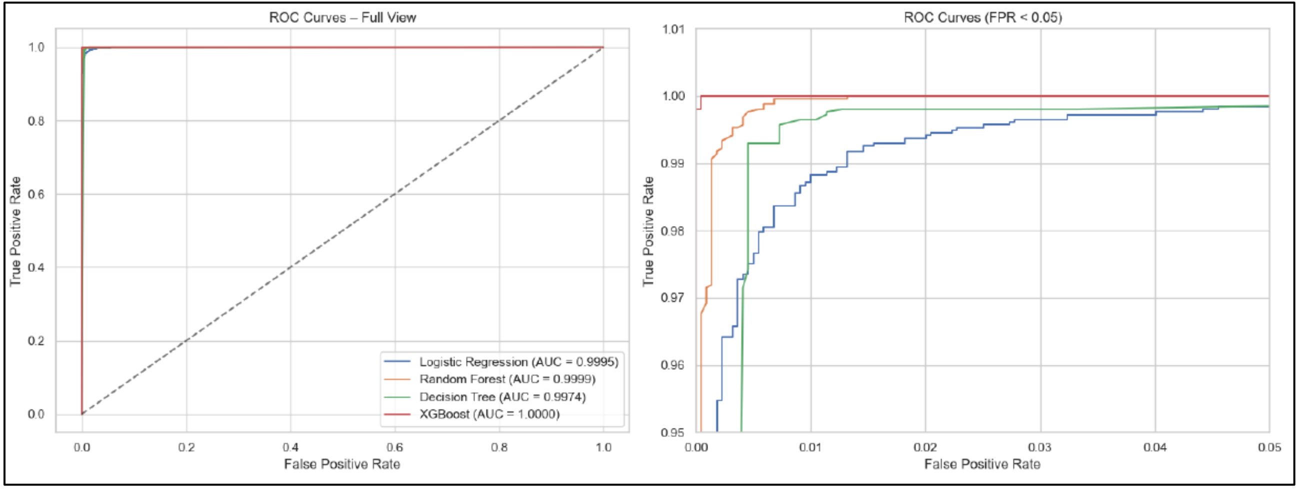


The confusion matrices in Figure 11 show that all four models classify both loan not approved (0) and loan approved (1) cases with very high accuracy. ***Logistic Regression*** performs strongly, correctly identifying most cases, but it shows ***27 false negatives*** (approved loans predicted as not approved) and ***27 false positives***, slightly ***higher than the tree-based models***.

Random Forest improves on this by reducing misclassifications to ***15 false positives*** and only ***2 false negatives***, showing better sensitivity toward approved loans. ***Decision Tree*** performs similarly well with ***16 false positives*** and ***11 false negatives***, maintaining strong balance across both classes.

XG Boost clearly outperforms all models, with ***only 1 false positive and 1 false negative***, demonstrating ***exceptional ability to correctly classify both non-approved and approved loans***. This minimal error indicates the strongest discriminatory power and the most reliable predictions among all models.

Figure 12: ROC Curves



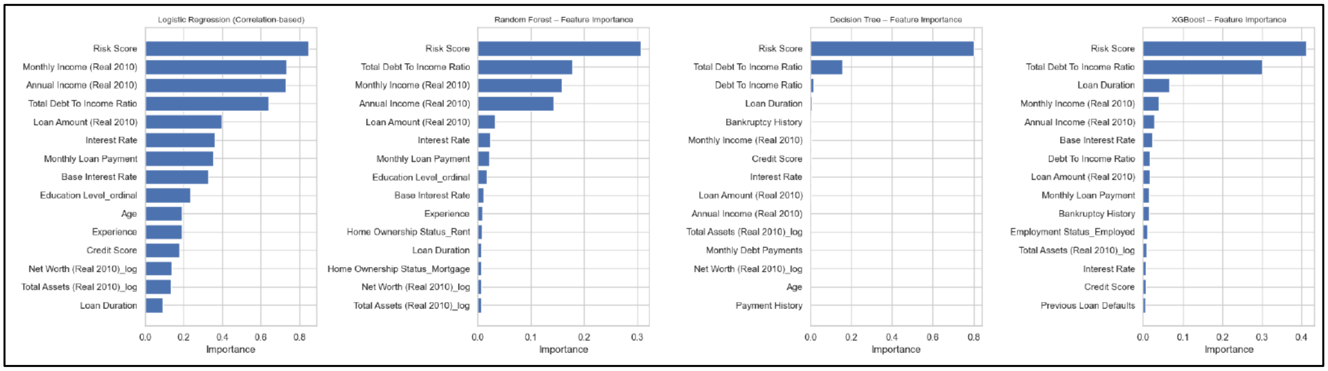
The ROC curves in Figure 12 illustrate the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for all four models. In the full view, all models show excellent performance with ROC curves that hug the top-left corner of the plot, consistent with their high AUC values.

When zooming into the low-FPR region (FPR < 0.05), which is especially important for loan approval applications to minimize misclassification of denied loans, the differences between models become clearer. ***Logistic Regression performs well but shows a slightly lower TPR in the very low FPR region compared to the tree-based models. Decision Tree and Random Forest display stronger early lift, maintaining higher TPRs at small FPR values.***

XG Boost consistently delivers the strongest curve shape, reaching near-perfect TPR even when FPR is extremely low, which is reflected in its perfect AUC = 1.0000. This indicates the model's superior ability to distinguish between approved (1) and not-approved (0) loans across all threshold levels, especially in high-precision decision regions.

Overall, while all models show excellent discriminatory power, XG Boost provides the most dominant ROC performance, particularly where minimizing false positives is critical.

Figure 13: Feature Importance



Across all four models, ***Risk Score consistently emerges as the most influential predictor of loan approval,*** showing the strongest relationship with the target variable in both correlation-based methods (Logistic Regression) and tree-based feature importance rankings (Random Forest, Decision Tree, XGBoost). This indicates that creditworthiness, as captured by the Risk Score, is the dominant factor driving loan approval decisions.

Income-related variables, ***Monthly Income, Annual Income, and Debt-to-Income Ratio*** also ***appear prominently across models.*** These features reflect an applicant's financial stability and repayment capacity, making them key determinants in the approval process. Logistic Regression highlights strong correlations with income and debt variables, while Random Forest and XG Boost further reinforce their importance through high contribution scores.

XG Boost provides the most refined importance distribution, again emphasizing ***Risk Score and Total Debt to Income Ratio*** as ***primary drivers***, followed by loan-specific variables such as ***Loan Duration, Loan Amount, and interest-related factors.***

Overall, while the magnitude and scaling of importance differ by model type, ***the patterns are highly consistent: Risk Score, income level, and debt burden*** collectively form the ***core predictors*** of loan approval. This agreement across statistical and machine learning models reinforces the reliability of these variables in shaping credit decisions.

Figure 14: Partial Dependence Plot – Logistic Regression

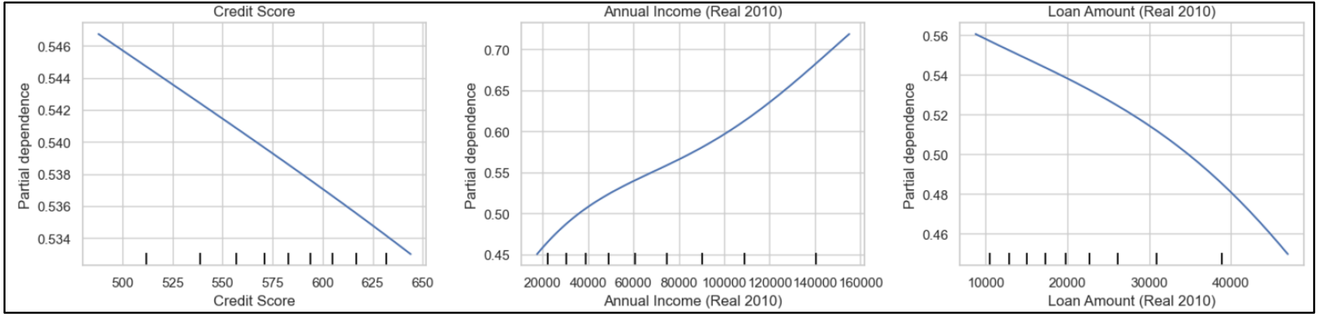


Figure 15: Partial Dependence Plot – Random Forest

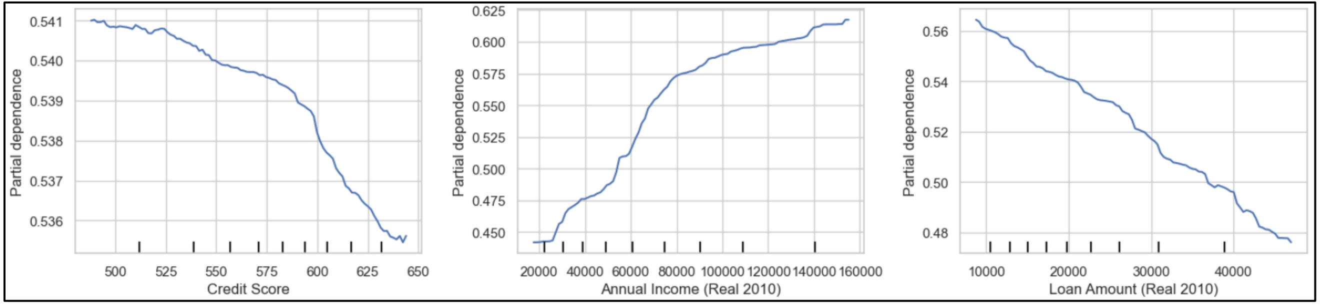


Figure 16: Partial Dependence Plot – Decision Tree

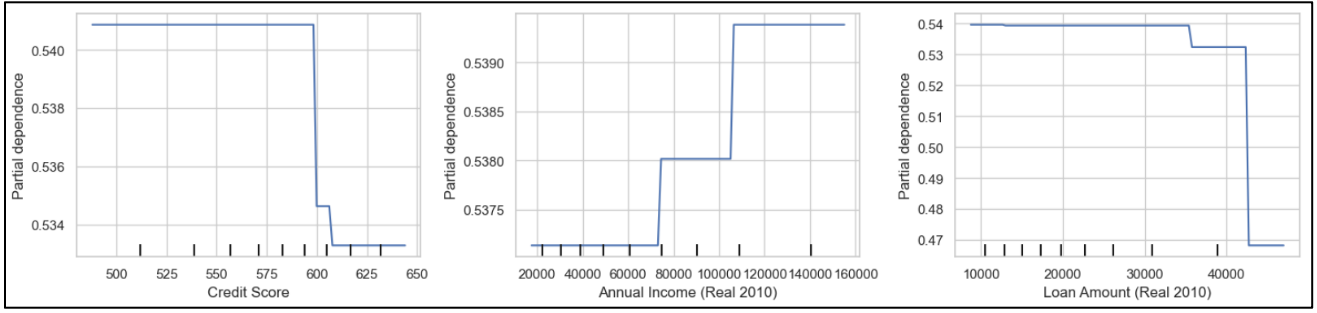
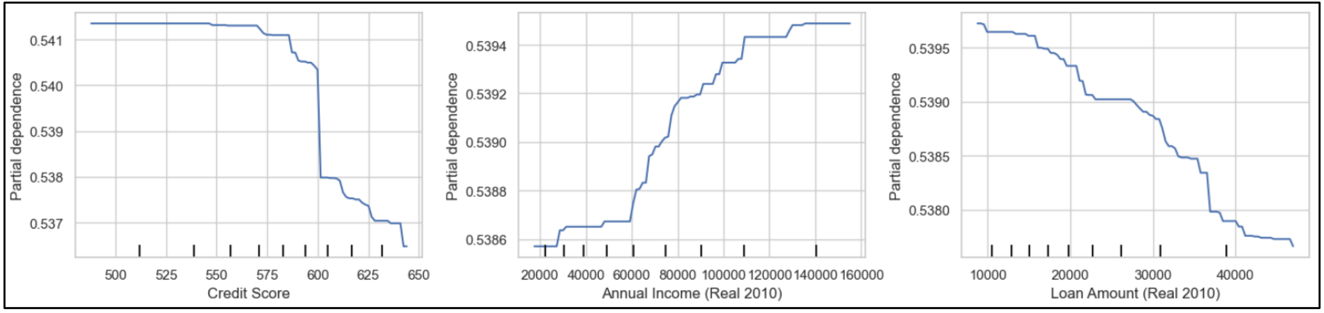


Figure 17: Partial Dependence Plot – XG Boost



Partial Dependence Plots (PDPs) were generated for three key predictors - ***Credit Score, Annual Income, and Loan Amount*** across all four tuned models to understand how each feature influences the predicted probability of loan approval (class 1).

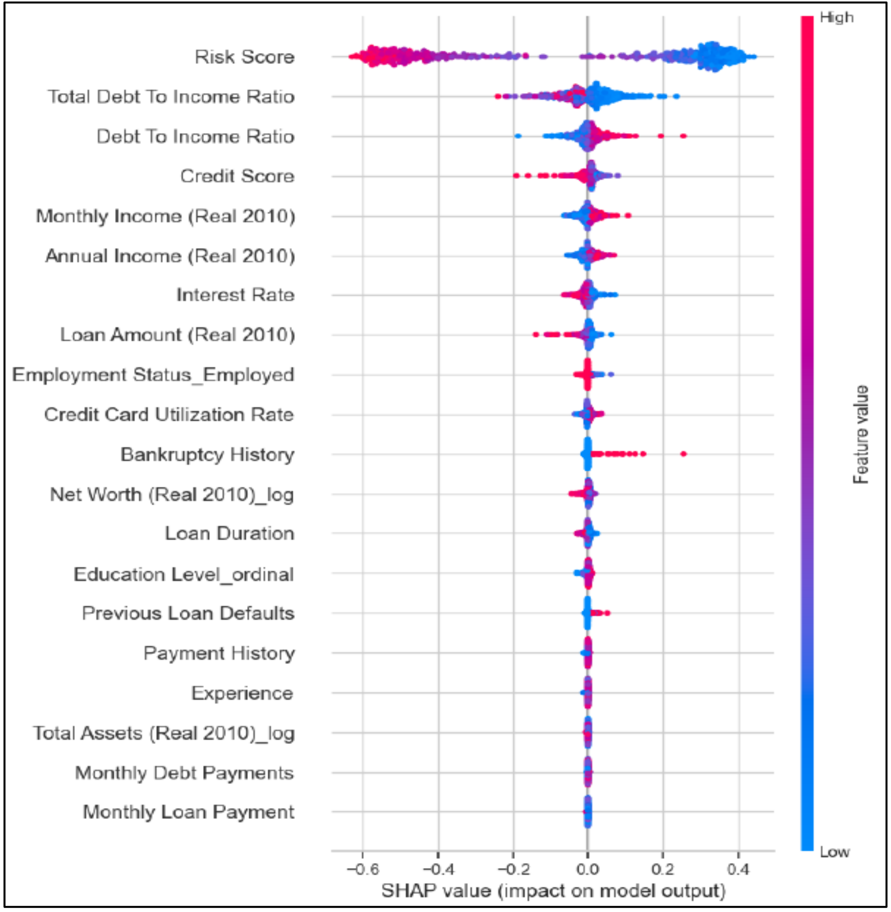
Credit Score - All models show a ***slight negative relationship*** between credit score and approval probability. Although counterintuitive, this trend is consistent across models and reflects the underlying dataset, where higher credit-score applicants may also have characteristics that increase perceived lending risk (e.g., high debt, large loan amounts).

Annual Income - Annual Income displays a ***strong and consistent positive effect*** across all models. Higher income significantly increases the predicted probability of approval, reflecting its importance as a financial stability indicator. Logistic Regression and Random Forest show smooth increases, while Decision Tree and XG Boost show threshold-based jumps.

Loan Amount - Loan Amount has a ***clear negative effect*** on approval probability for all models. As loan size increases, predicted approval decreases, indicating higher perceived repayment risk. Tree-based models again show stepwise declines, whereas linear models show smooth downward trends.

The PDP analysis confirms that the models have learned economically logical patterns, with income increasing approval likelihood and high loan amounts reducing it. The unexpected negative influence of credit score appears to be a ***dataset-specific phenomenon***, not a modeling issue. Overall, the PDPs complement the feature importance results and help validate how the models arrive at their predictions.

Figure 18: SHAP Values for Best Performing Model (XG Boost)



The SHAP summary plot (figure 18) highlights how each feature influences XG Boost's loan approval predictions, revealing both the **direction** and **magnitude** of impact at the individual observation level.

Most Influential Features

- ***Risk Score*** is the single most impactful feature. Higher Risk Scores (blue points) push the prediction towards approval, while lower scores (pink points) push it towards rejection, showing a strong monotonic relationship.
- ***Total Debt-to-Income Ratio and Debt-to-Income Ratio*** have substantial negative effects. Higher values consistently reduce the predicted likelihood of approval, aligning with lending risk guidelines.
- ***Credit Score*** shows a mixed but generally positive effect, higher values tend to increase approval probability, though the influence is smaller compared to Risk Score.
- ***Income variables (Monthly and Annual Income)*** positively contribute to approval predictions. Applicants with higher incomes generally push the SHAP values upward, indicating lower perceived repayment risk.

Moderate Influence Features

- ***Interest Rate and Loan Amount*** mostly show negative SHAP contributions, meaning higher interest or larger loan sizes decrease approval probability**.**
- ***Employment Status (Employed)*** shifts predictions toward approval, as expected.
- ***Credit Card Utilization Rate and Bankruptcy History*** strongly push predictions downward when values are high, highlighting risk concerns.

Lower Impact Features

Features such as ***Loan Duration, Education Level, Payment History, Experience, and Assets/Net Worth*** contribute minimally, suggesting the model relies more on debt, income, and risk measures when distinguishing approvals.

Overall, XG Boost’s SHAP values indicate that ***approval decisions are primarily driven by Risk Score, Debt-to-Income Ratios, Income, and Credit Score***, with financial stress indicators (high utilization, bankruptcy, large loan amounts) reducing approval likelihood. The model's SHAP behavior aligns well with standard credit risk assessment practices.

Best Performing Model – XG Boost delivered the best performance among all four models with:

- Highest overall accuracy (99.96%)
- Perfect ROC-AUC (100%)
- Excellent Recall (100%) of the minority class
- Stable validation metrics (low variance across folds)
- Captures both linear and non-linear relationships effectively
- Robust to multicollinearity, interactions, and skewed distributions

6. Conclusion

6.1 Summary of Findings

The objectives of this project were to leverage EDA to discover key relationships between variables, identify strongest predictors of loan approval, and develop predictive models to determine the likelihood of loan approval for applicants. All three objectives were successfully achieved through a structured analytical workflow integrating statistical exploration, feature engineering, and multiple machine-learning models.

EDA Findings:

Initial exploratory analysis highlighted strong relationships between financial stability indicators and approval outcomes. Higher Risk Scores, Income, and Credit Scores were associated with increased approval likelihood, while higher Debt-to-Income Ratios and larger Loan Amounts reduced approval chances.

Key Predictors:

Across correlation analysis, model-based feature importance, SHAP values, and partial dependence plots, a consistent set of dominant predictors emerged. Risk Score was the most influential feature, followed by Debt-to-Income Ratios, Income, Loan Amount, Interest Rate, and Credit Score. These variables played a central role in shaping approval decisions.

Model Performance:

Four models - Logistic Regression, Random Forest, Decision Tree, and XG Boost were trained and tuned using stratified cross-validation. All performed strongly, but XG Boost achieved the best overall results with near-perfect accuracy, precision, recall, F1-score, and ROC-AUC, along with the lowest misclassification rates. SHAP analysis further enhanced interpretability, confirming the model's alignment with observed financial patterns.

6.2 Business Implications

The insights and predictive models developed in this project offer several high-value implications for JAF Bank's lending operations:

- ***Automated Approvals for Low-Risk Applicants:*** With highly accurate models, especially XG Boost, JAF Bank can confidently automate approvals for applicants exhibiting strong financial stability (high Risk Score, low DTI, high Income). This reduces manual workload and accelerates throughput without compromising decision quality.
- ***Improved Speed, Consistency & Efficiency:*** Data-driven predictions eliminate subjective variations in decision-making. Standardized model-based evaluations ensure consistent treatment across applicants and significantly shorten processing times.
- ***Reduced Losses from Risky Approvals:*** By accurately identifying high-risk profiles (e.g., high Debt-to-Income Ratios, large loan amounts relative to income, poor credit behavior), the bank can

minimize default risk and improve portfolio quality. Models can flag borderline or high-risk cases for enhanced manual review rather than outright rejection.

- ***Refined Lending Policies & Eligibility Guidelines:*** Key predictors i.e. Risk Score, DTI, Income, Credit Score, and Loan Duration, should be incorporated into updated lending frameworks. JAF Bank can strengthen eligibility thresholds, adjust acceptable parameter ranges, or introduce tier-based risk pricing based on these insights.
- ***Better Customer Segmentation & Targeting:*** Predictive insights allow the bank to segment applicants into risk tiers and tailor communication, pre-approved offers, and credit products to each group. This supports smarter marketing and customer acquisition strategies.
- ***Enhanced Monitoring & Early Warning Systems:*** Predictive features can be leveraged for post-approval monitoring. Applicants whose risk profiles begin deteriorating (e.g., rising debt levels, declining income stability) can be flagged early for intervention.

6.3 Limitations Of the Study

While the project delivered strong predictive performance and valuable insights, several limitations should be acknowledged:

- ***Use of Synthetic Data:*** The dataset is artificially generated and may not fully capture the complexity, irregularities, and noise present in real-world banking data. As a result, the model's performance in production may differ from the high accuracy observed during testing.
- ***Potential Overfitting to Clean Data:*** Synthetic datasets tend to lack missing values, inconsistencies, or errors. Real applicant data often contains incomplete records, reporting inaccuracies, and behavioral patterns that are harder to model, which could impact predictive accuracy.

- ***Static Modeling Approach:*** The models were trained on a snapshot of data rather than time-evolving financial behavior. Economic conditions, interest rate changes, and borrower risk profiles may shift over time, requiring periodic model retraining.

6.4 Recommendations for Future Work

Although the project successfully developed and evaluated four predictive models, identifying XG Boost as the strongest performer, the results are based on a synthetic dataset. To ensure real-world reliability and operational value, several strategic next steps are recommended:

- ***Collect Real-World Data***: Partner with small financial institutions to obtain actual loan application and repayment datasets.
- ***Retrain and Validate the Model:*** Re-train the best-performing model (XGBoost) on real data to ensure accuracy, fairness, and operational readiness.
- ***Deploy as a Decision-Support Tool:*** Integrate the refined model into the loan approval workflow to improve consistency, reduce manual workload, and minimize risk.
- ***Monitor and Govern the Model:*** Implement ongoing monitoring to detect performance drift, bias, and changes in applicant behavior.
- ***Enhance Features Over Time:*** Consider adding behavioral, transactional, and alternative data sources for improved predictive power.
- ***Provide Staff Training:*** Educate loan officers on interpreting model outputs to support informed decision-making.

7. References

a. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 1263-1284. Retrieved from https://ieeexplore.ieee.org/document/5128907

b. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

c. Batista, G. E., Bazzan, A. L., & Monard, M. C. (2004). *A study of the behavior of several methods for balancing machine learning training data.* ACM SIGKDD Explorations Newsletter. Retrieved from https://doi.org/10.1145/1007730.1007735

d. Breiman, L. (2001). *Random forests Machine Learning*. Retrieved from https://link.springer.com/article/10.1023/A:1010933404324

e. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Chapman & Hall.

f. Chen, T., & Guestrin, C. (2016). GBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785-794). Retrieved from https://doi.org/10.1145/2939672.2939785

g. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

h. Batista, G. E., Bazzan, A. L., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 20-29. Retrieved from https://doi.org/10.1145/1007730.1007735

i. Breiman, L. (2001). *Random forests. Machine Learning*, 45(1), 5–32. Retrieved from https://doi.org/10.1023/A:1010933404324

j. Chen, T., & Guestrin, C. (n.d.). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785-794). Retrieved from https://doi.org/10.1145/2939672.2939785

k. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

l. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

m. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer.

n. Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.