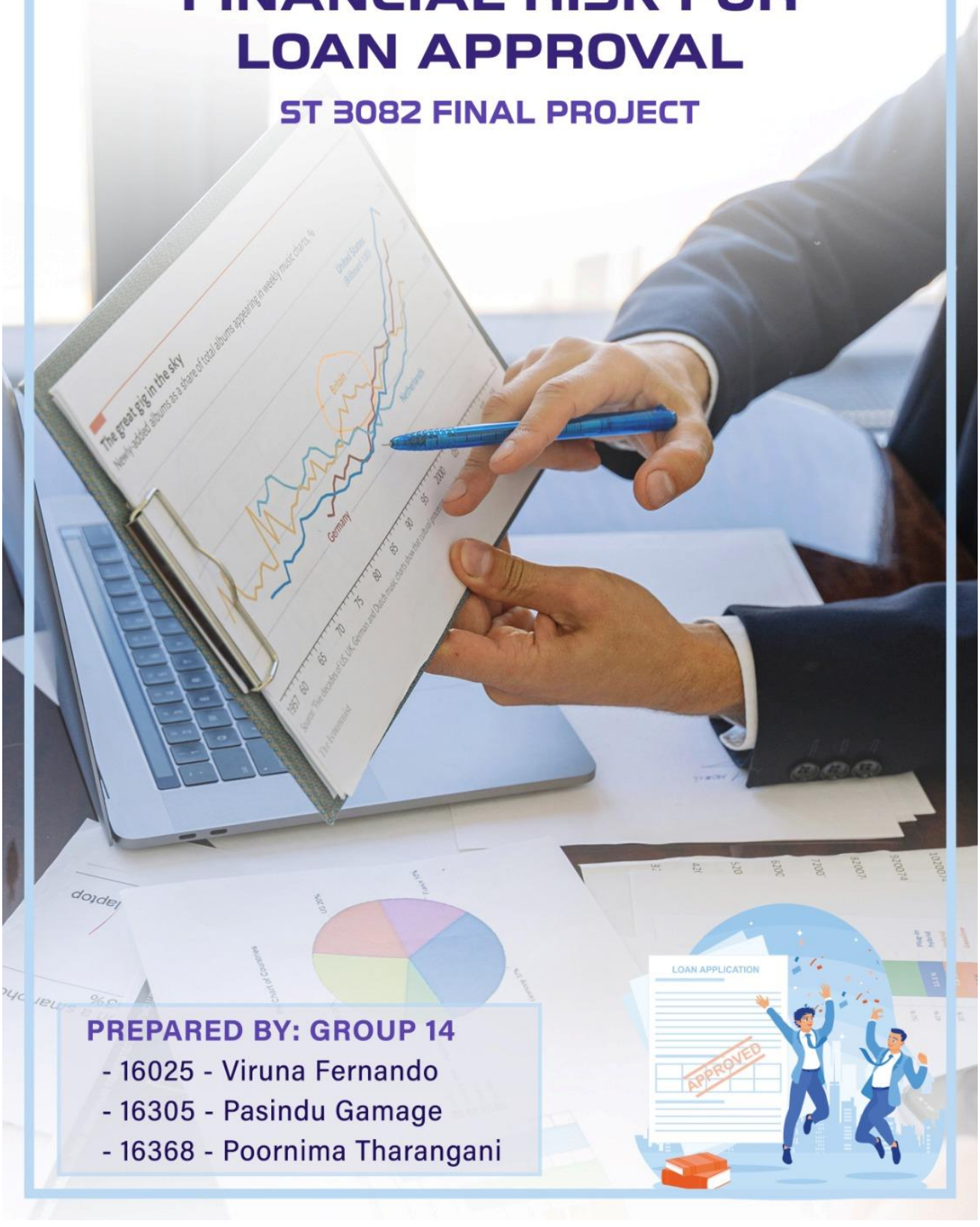
DESCRIPTIVE ANALYSIS OF FINANCIAL RISK FOR LOAN APPROVAL

ST 3082 FINAL PROJECT



PREPARED BY: GROUP 14

- 16025 - Viruna Fernando
- 16305 - Pasindu Gamage
- 16368 - Poornima Tharangani

Contents

List of Tables

List of Figures

Abstract

This report presents a data-driven approach to predicting Credit score, Risk score and Loan approval status based on various financial and demographic factors. The dataset was analyzed using statistical and machine learning techniques. Descriptive analysis was conducted to identify key drivers of predicting these factors. The study also explored the relationship between these, offering valuable insights for both applicants and lenders. By developing predictive models, this research contributes to improve the efficiency of loan approval processes and helps applicants to understand the key criteria for loan eligibility.

Introduction

In the modern financial landscape, loan approval processes have evolved beyond simplistic criteria and now rely on a more complex understanding of an applicant's financial and personal profile. Traditional methods often overlook the complex interplay between various factors such as demographics, income stability, debt obligations, and financial behavior. This has created an opportunity to improve the accuracy of loan approval decisions by leveraging advanced data analytics techniques.

This project focuses on using historical loan application data to develop a framework that can predict the likelihood of loan approval. By examining a variety of key features, the aim is to create a more refined model that enhances the decision-making process. Through this analysis, financial institutions can understand the factors influencing loan outcomes, improve the transparency of the approval process, and offer applicants clearer guidance on how to improve their eligibility.

This project aims to provide a data-driven approach to loan approval prediction, by providing more effective risk management strategies for lenders while enhancing accessibility and fairness for borrowers.

Description of the Problem

In today's financial landscape, traditional loan approval systems often rely on rigid, rule-based criteria that fail to capture the full complexity of an applicant's financial and demographic profile. These systems typically overlook important behavioral and contextual variables such as income consistency, employment status, age, and credit history, leading to limited access to credit for eligible borrowers, while simultaneously increasing the default risk for lenders.

This project addresses these challenges by developing a data-driven predictive framework that improves both the fairness and efficiency of loan evaluation processes. Using a large historical dataset of loan applicants, the project focuses on two major goals:

1. Loan Approval Prediction
   Our Primary goal is Identify key factors influencing Loan Approval status and build a model capable of predicting whether an applicant's loan request will be approved or denied.
2. Credit score and Risk score
   The next goal is to estimate continuous financial risk score and credit score and analyze its relationship with loan approval eligibility.

By incorporating both risk and credit-based insights into the predictive framework, the project delivers a more accurate loan assessment tool. The overall objective is to improve decision-making for financial institutions, based on a comprehensive analysis of their financial behavior and risk profiles.

Description of the Dataset

The dataset used for this analysis, titled *"Financial Risk for Loan Approval,"* from Kaggle comprises 20,000 observations and 36 features, representing a wide array of demographic and financial information related to loan approval decisions. This dataset includes both numerical and categorical variables, capturing important applicant details that are critical for evaluating loan applications.

Variable	Data Type	Description
Age	Numerical	Applicant's age
AnnualIncome	Numerical	Yearly income
CreditScore	Numerical	Creditworthiness score
EmploymentStatus	Categorical	Job situation
EducationLevel	Numerical	Highest education attained
Experience	Numerical	Work experience
LoanAmount	Numerical	Requested loan size
LoanDuration	Numerical	Loan repayment period
MaritalStatus	Categorical	Applicant's marital state
NumberOfDependents	Numerical	Number of dependents
HomeOwnershipStatus	Categorical	Homeownership type
MonthlyDebtPayments	Numerical	Monthly debt obligations
CreditCardUtilizationRate	Numerical	Credit card usage percentage
NumberOfOpenCreditLines	Numerical	Active credit lines
NumberOfCreditInquiries	Numerical	Credit checks count
DebtToIncomeRatio	Numerical	Debt to income proportion
BankruptcyHistory	Categorical	Bankruptcy records
LoanPurpose	Categorical	Reason for loan
PreviousLoanDefaults	Categorical	Prior loan defaults
PaymentHistory	Numerical	Past payment behavior
LengthOfCreditHistory	Numerical	Credit history duration
SavingsAccountBalance	Numerical	Savings account amount
CheckingAccountBalance	Numerical	Checking account funds
TotalAssets	Numerical	Total owned assets
TotalLiabilities	Numerical	Total owed debts
MonthlyIncome	Numerical	Income per month
UtilityBillsPaymentHistory	Numerical	Utility payment record
JobTenure	Numerical	Job duration
NetWorth	Numerical	Total financial worth
BaseInterestRate	Numerical	Starting interest rate
InterestRate	Numerical	Applied interest rate
MonthlyLoanPayment	Numerical	Monthly loan payment
TotalDebtToIncomeRatio	Numerical	Total debt against income
LoanApproved	Categorical	Loan approval status
RiskScore	Numerical	Risk assessment score

Table 1:Variable Description

Feature Engineering

Given the high dimensionality of our dataset, which contains 36 variables, we adopted a combined approach using both statistical analysis and domain knowledge to refine our feature set and improve model performance.

To ensure that our model focuses on the most relevant and actionable predictors, we first consulted financial domain sources to understand key criteria typically used in real-world loan approval decisions. These sources highlighted the following core factors as most influential:

- **Credit Score**
- **Loan Approval History**
- **Risk Score**
- **Payment History**
- **Current Debt**
- **Income**
- **Utility Expenses**

Based on these financial considerations, we systematically mapped and categorized the 36 original variables under these broader factors. We then conducted a detailed assessment of Statistical relationships on given dataset ,Redundancy and overlap among predictors and their domain relevance ensuring variables align with real-world loan evaluation criteria.

As a result of this evaluation we identified and removed several redundant, overlapping and low importance variables that were unlikely to contribute to loan evaluation criteria as follows ,

ApplicationDate, AnnualIncome, DebtToIncomeRatio, EducationLevel, Experience, PaymentHistory, UtilityBillsPaymentHistory ,JobTenure, SavingsAccountBalance, CheckingAccountBalance ,TotalAssets

Methodology for Calculating Loan Approval Probability

To analyze how loan approval probability varies across different variables, we encountered challenges related to the distribution of the data. Specifically, many variables did not follow a distribution suitable for direct probability estimation. To address this, we adopted the following methodology:

1. Continuous Variables:

- For each continuous variable, we first identified the full range of observed values.
- This range was then divided into a series of class intervals (bins).
- Within each class interval, we calculated the mean value of the "Loan Approved" variable.
- Since the "Loan Approved" variable is binary (0 = Not Approved, 1 = Approved), the mean value within each interval directly represents the loan approval probability for that interval.

For example, if in a given interval the mean is 0.75, it indicates that 75% of applicants in that range were approved.

Note: In some cases, to better understand how important features influence the response variable, we applied this approach more broadly by calculating the mean response within each class interval of the feature. This helped capture non-linear relationships between variables and loan approval likelihood.

2. Categorical Variables:

- For each categorical variable, we counted the number of applicants within each category.
- We then calculated the percentage of approved loans in each category.
- This percentage was interpreted as the loan approval probability for applicants belonging to that specific category.

Data Preprocessing

- **Missing Values and Duplicates**: The dataset was clean : no missing values or duplicate entries were found.
- **Outlier Detection**: Outliers were examined and may explained the natural variability in financial data. They were retained in the dataset.
- **Feature Scaling**: Numerical variables were standardized to ensure consistent contribution in model training.
- **Categorical Encoding**: Variables were encoded using One-Hot Encoding, enabling their integration into trained models.
- **Data Splitting**: The data was divided into training (16,000 observations) and testing (4,000 observations) sets to enable generalization and validation of model performance.

Important Results of descriptive analysis

Distribution of Loan Approval Status

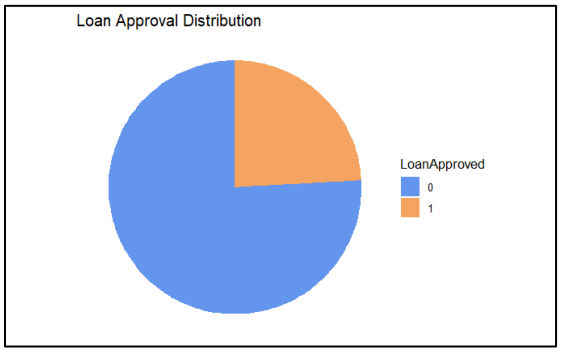


Figure 1:Pie chart for Distribution of LoanApproved

In the data set, Most of the observations are belong to loan rejected group (LoanApproved=0)

Only about 1/4 th of the all observations belong to loan accepted group (LoanApproved=1)

So the data set was imbalanced. We have to consider the imbalance of the LoanApproved when we predicting loan approval status using classification models.

Distribution of Credit Score

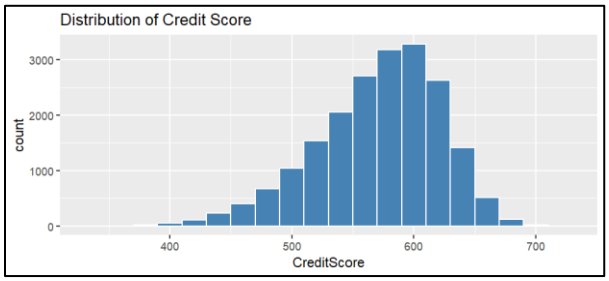


Figure 2: Histogram of CreditScore

The distribution of the credit score variable is left-skewed, with more observations clustered around higher credit score values. This indicates that most applicants tend to have higher credit scores. The mean credit score was 571.61, while the median credit score is about 578. The skewness value was 0.60. This may imply that the majority of applicants have relatively higher credit scores.

Distribution of Risk Score

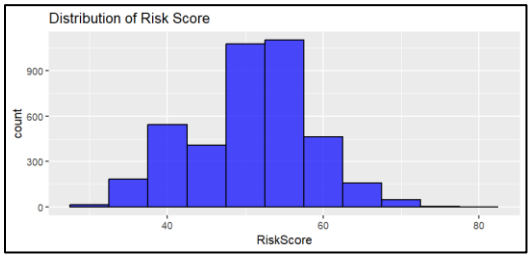
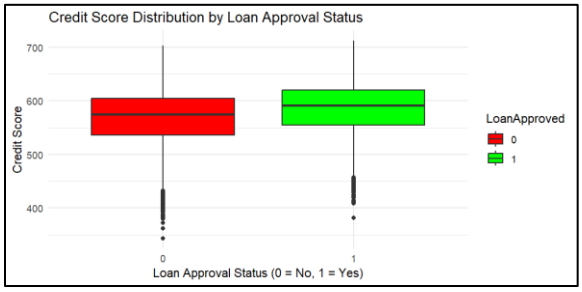


Figure 3:Distribution of RiskScore

The distribution of the Risk Score is approximately symmetric. The mean and median values are 50.76 and 52 respectively, indicating that the scores are fairly balanced around the center. This suggests that there is no strong skewness present, and the Risk Score variable is evenly distributed.

Distribution of Credit Scores by Loan Approval Status



Approved applicants tend to have higher credit scores, as shown by the higher median values. The IQRs are similar, but the distribution for approved loans is shifted upward, supporting the conclusion that higher credit scores are associated with increased loan approval probability.

Figure 4:boxplot for CreditScore vs LoanApproved

Distribution of Credit Score vs Age

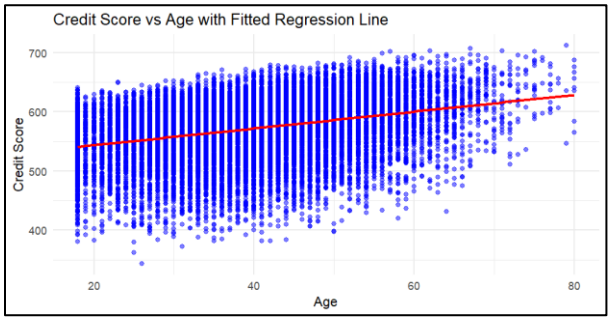


Figure 5:Scatterplot for Credit score vs Age

It indicates a weak relationship between Age and Credit Score.(R square 0.142) This suggests that Age alone may not a strong predictor of Credit Score, and implies the presence of other variables or non-linear relationships that a simple linear model cannot capture.

Credit Score vs. Loan Approval Probability Curve

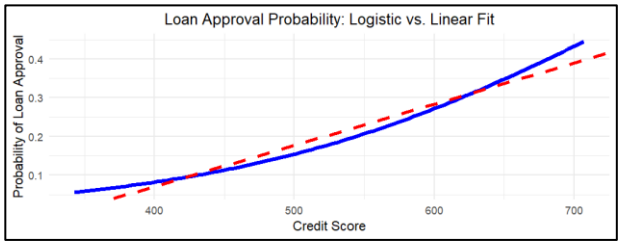


Figure 6:CreditScore vs LoanApproval probability

The Credit Score vs. Loan Approval Probability Curve demonstrates a linear relationship between credit score and the probability of loan approval, with the fitted regression equation: Loan Approval Probability = -0.3548 + 0.0011 * Credit Score**.** The R-squared value (0.9633) signifies a strong correlation. As credit score increases, the probability of loan approval rises steadily, reflecting the general practice of granting loans to individuals with higher credit scores, as they have low financial risk to lenders.

Relationship Between Credit Score vs Loan Duration

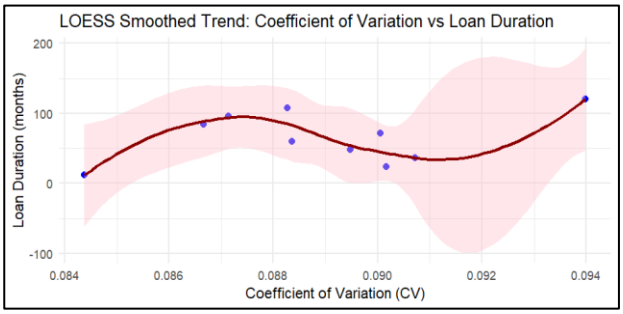
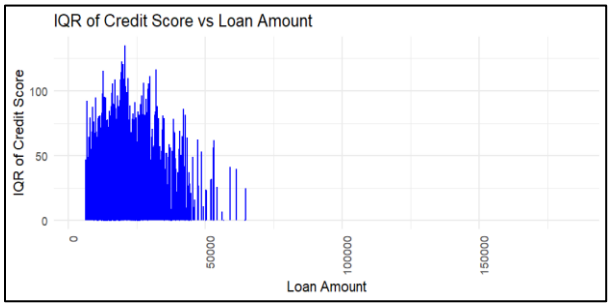


Figure 7:CreditScore vs LoanApproval probability

A Pearson correlation coefficient of approximately 0.34 indicates a weak to moderate positive linear relationship between loan duration and the coefficient of variation of credit score. This suggests that, on average, as the loan duration increases, the relative variability of credit scores tends to increase slightly**.**

While the relationship is not strong, the positive direction implies a general upward trend. However, due to the moderate strength, this trend is not consistent to conclude a strong linear association. So overall, we can say There appears to be a mild positive relationship between loan duration and the coefficient of variation of credit score, but the association is not strong or strictly linear.



The analysis reveals a significant negative relationship between LoanAmount and the Interquartile Range (IQR) of the CreditScore. Both linear regression and Spearman's rank correlation tests show that as the loan amount increases, the variability (IQR) in the credit score decreases, with a statistically significant negative correlation (Spearman's rho = -0.1593, p-value < 2.2e-16), indicating that higher loan amounts are associated with lower variability in credit scores.

Mean Credit Score vs Interest rate

Figure 8:Mean CreditScore vs InteresRate

The analysis of the relationship between interest rate and credit score reveals a clear downward trend. As the interest rate increases, the average credit score of applicants tends to decrease across the defined class intervals. Applicants receiving loans with the lowest interest rates generally have the highest average credit scores, indicating stronger creditworthiness, suggesting that applicants in higher interest rate brackets tend to have weaker credit profiles. This pattern highlights a negative association between interest rate levels and borrower credit quality.

Probability of Loan Acceptance vs Risk Score Curve

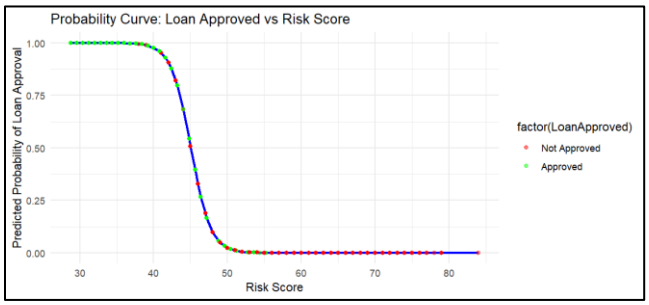
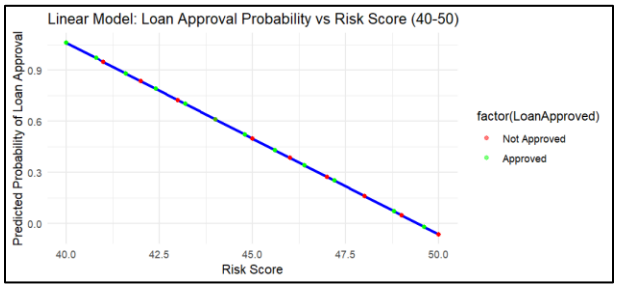


Figure 9:prob curve for LoanApproved vs RiskScore

The probability of loan approval as a function of risk score follows a Z-shaped curve. In the range of risk scores from 0 to 40, the probability of approval remains consistently high, close to 1, suggesting that applicants with lower risk scores are more likely to be approved. Between risk scores of 40 and 50, the probability of approval decreases in a linear fashion, indicating a transition phase where the likelihood of approval drops as the risk increases. Beyond a risk score of 50, the probability of approval approaches zero, indicating a negative relationship between higher risk scores and loan approval. This Z-shaped curve effectively illustrates the non-linear relationship between risk score and loan approval probability. The below plot shows the distribution in between 40-50 more clear.



There is a clear linear relationship between RiskScore and Loan Approval Probability within the 40-50 range. As the RiskScore increases, the Approval probability decreases in a near-linear fashion. This implies that higher risk scores are associated with a lower likelihood of loan approval, and the relationship between risk and approval is relatively strong within this range.

Loan Approval Rate by Education Level

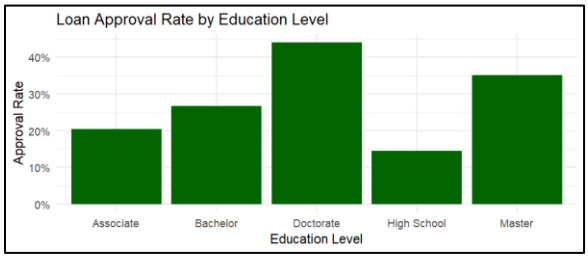


Figure 10:Bar chart of LoanApproved Rate and Edu Level

The results indicate a upward trend in the loan approval rate with increasing education level. Applicants with a High School education have the lowest approval rate at 14.4%, followed by those with an Associate degree at 20.4%. The approval rates rise for applicants holding a Bachelor's degree at 26.6% and a Master's degree at 35.1%, with those possessing a Doctorate achieving the highest approval rate at 44%. This pattern suggests that higher education is associated with higher loan approvalRates, likely reflecting lenders' perception that better-educated individuals pose lower financial risk.

Loan Approval Rate by Employment Status

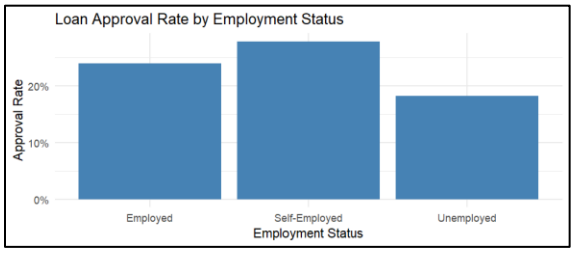


Figure 11:Bar chart for LoanApproved Rate and Emp Status

Employed applicants exhibit an approval rate of 24%, while self-employed individuals experience a slightly higher rate of 27.8%. In contrast, unemployed applicants have the lowest approval rate at 18.2%. This pattern suggests that both traditional and self-employment are viewed more favorably by lenders compared to unemployment, likely due to the perception of greater financial stability and earning potential in these groups.

Loan Approval Rate by Home Ownership Status

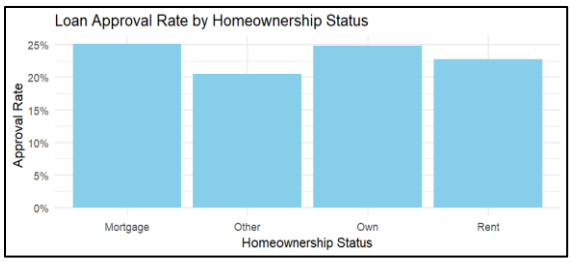


Figure 12:Bar chart for LoanApproved rate by HomeOwnershipStatus

The analysis of loan approval rates based on homeownership status reveals some interesting patterns. Applicants with a mortgage have the highest approval rate at 25.2%, followed by those who own their homes at 24.9%. Renters have a slightly lower approval rate of 22.8%, while applicants in the "Other" category have the lowest approval rate at 20.5%. This suggests that financial institutions may view individuals with mortgages or homeownership as more financially stable and less risky, leading to higher approval rates. Renters, on the other hand, appear to have a slightly higher perceived risk, possibly due to concerns about long-term financial security. In conclusion, homeownership status plays a significant role in loan approval decisions.

Loan Approval Rate by Number of Credit Inquiries

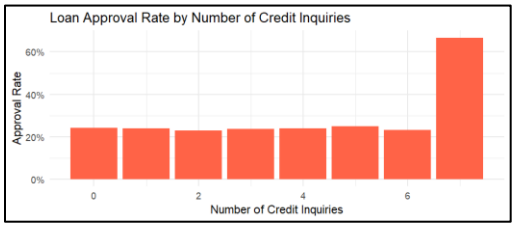


Figure 13:LoanApproved Rate by Number of Credit Inquiries

The analysis of loan approval rate based on the number of credit inquiries reveals an interesting pattern. Initially, the approval rate remains relatively stable, around 0.24 (24%) for the first few categories of credit inquiries (0 to 3). However, as the number of inquiries increases, the approval rate fluctuates slightly, with the highest approval rate of 66.7% observed for applicants with 7 credit inquiries. This indicates that while there may be a general tendency for higher approval rates for individuals with fewer inquiries, the small number of applications in higher inquiry categories might skew the data. Overall, the approval rate does not show a consistent downward trend with increasing credit inquiries, suggesting that the relationship between credit inquiries and loan approval might not be as straightforward.

Loan Approval Rate by Bankrupt History

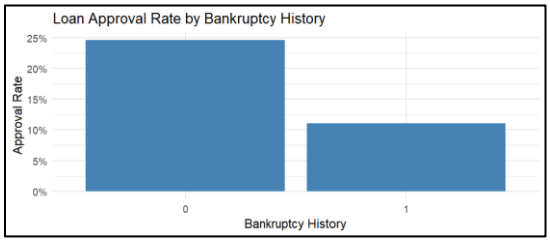


Figure 14:Bar chart for LoanApproved vs Bankruptcy History

The loan approval rate is notably lower for applicants with a bankruptcy history. Applicants without a bankruptcy history have an approval rate of 24.6%, while those with a bankruptcy history have an approval rate of only 11.1%.. bankruptcy history significantly reduces the likelihood of loan approval, indicating that lenders associate bankruptcy with higher financial risk.

Loan Approval Probability by Monthly Income

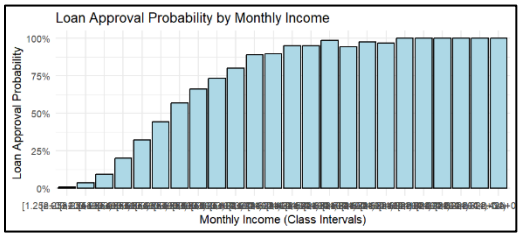


Figure 15:Prob of LoanApproved by MonthlyIncome plot

The plot displays the relationship between monthly income class intervals and the loan approval probability. As we observe from the results, there is a clear positive correlation between monthly income and loan approval probability. As the income increases, the approval probability rises.This indicates that individuals with higher monthly incomes are more likely to be approved for loans.

Loan Approval Probability by Total Debt-to-IncomRatio

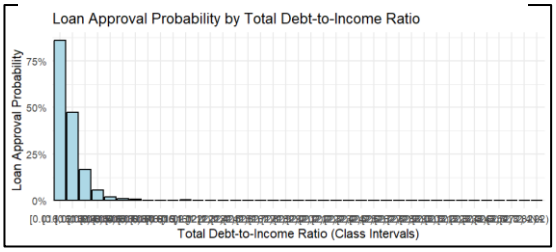


Figure 16:prob of LoanApproved by Total DTI plot

The analysis of loan approval probability based on the Total Debt-to-Income Ratio (TDI) reveals an inverse relationship between the TDI ratio and the likelihood of loan approval. As the TDI ratio increases, the approval rate declines sharply. In the higher TDI values, the approval rate falls to as low as suggesting that the Total Debt-to-Income Ratio is a critical determinant in loan approval decisions. Therefore, it can be suggested that maintaining a low TDI ratio is crucial for improving the likelihood of loan approval.

Loan Approval Probability by Interest Rate

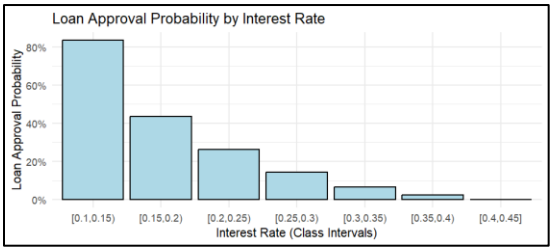


Figure 17:Prob of LoanApproved vs Interest rate

The analysis of loan approval probability based on interest rates reveals a clear inverse relationship. As the interest rate increases, the probability of loan approval tends to decrease. In lower interest rate ranges, the approval probability is relatively high, but as the interest rate moves into higher ranges, the likelihood of approval drops significantly. This trend suggests that higher interest rates may indicate higher perceived risk for lenders and a lower number of approved loans. Ultimately, the data reflects that interest rates play an important role in influencing loan approval decisions.

Loan Approval Probability by loan Duration

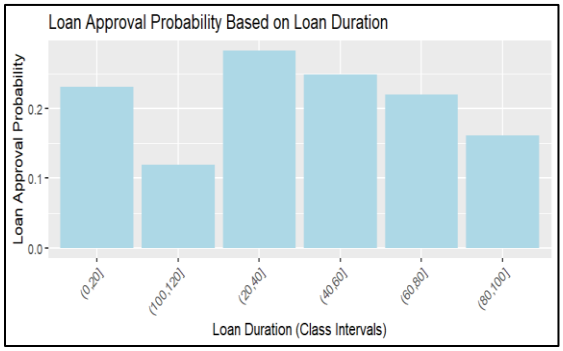


Figure 18:Loan Approval Prob by Loan Duration

The analysis of loan approval probability based on loan duration reveals varying approval rates across different loan duration classes. In the shorter loan duration intervals, the probability of loan approval tends to be moderate, with an increasing trend as the loan duration increases. However, as the loan duration extends further, the approval probability gradually decreases. The longest loan duration class shows the lowest probability of approval. This indicates that while loans with shorter durations may have a higher likelihood of approval, longer durations are associated with a declining approval rate, suggesting that lending institutions might be more cautious with longer-term loans.

Loan Approval Probability by Loan Amount

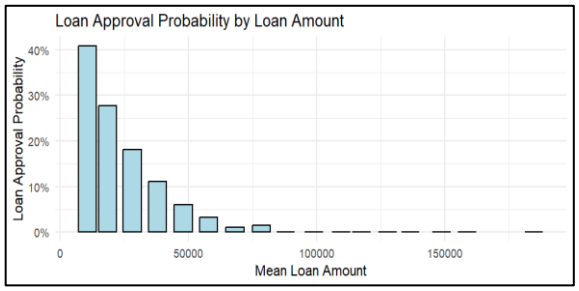


Figure 19:Prob of LoanApproved vs LoanAmount

The analysis of loan approval probabilities based on loan amounts reveals a negative trend between the size of the loan and the probability of loan approval. As the loan amount increases, the approval probability generally decreases. For smaller loan amounts, the approval rate is relatively higher, indicating that smaller loans are more likely to be approved. However, once the loan amount surpasses a certain threshold, the number of approved loans sharply declines, and in some of the higher loan amount categories, no loans were approved at all. This suggests that the likelihood of approval is inversely related to the loan amount, with larger loans facing more stringent approval criteria. The conclusion from this analysis is that smaller loan amounts are favored for approval, while larger loans may face more significant challenges in obtaining approval, reflecting possible risk management or policy constraints.

Risk Score vs Total Debt-to-Income Ratio

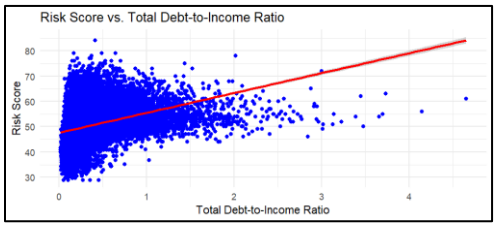


Figure 20:Scatterplot for RiskScore vs Total DTI

The linear regression model between Risk Score and Total Debt-to-Income Ratio (TDI) reveals a positive relationship, with a coefficient of 7.86, indicating that for each unit increase in TDI, the Risk Score increases by approximately 7.86 units. The p-value for the TDI coefficient is less than 0.001, confirming its statistical significance. However, the Multiple R-squared value of 0.1174 suggests that TDI explains only about 11.74% of the variation in Risk Scores, indicating that other factors likely contribute to the variability in Risk Score. The residual standard error of 7.308 highlights some level of spread around the fitted values, which is typical in real-world data. While TDI has an impact on Risk Scores, its explanatory power is limited, suggesting that other variables or more complex models are necessary to better capture the full range of influences on Risk Scores.

Mean Risk Score vs MonthlyIncome

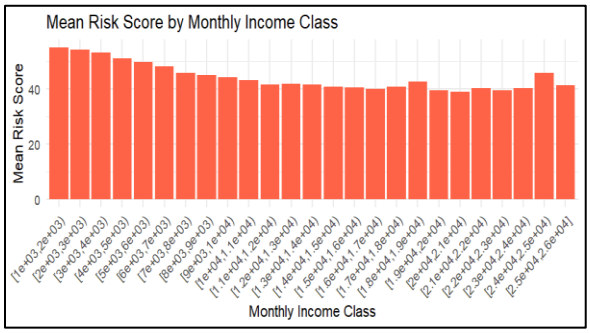


Figure 21:Mean RiskScore by Income class

The analysis reveals a clear decreasing trend in the mean risk score as monthly income increases. This suggests that individuals with higher monthly incomes tend to have lower risk scores on average, indicating a potential inverse relationship between income and financial risk.

Mean Risk Score by Total DTI

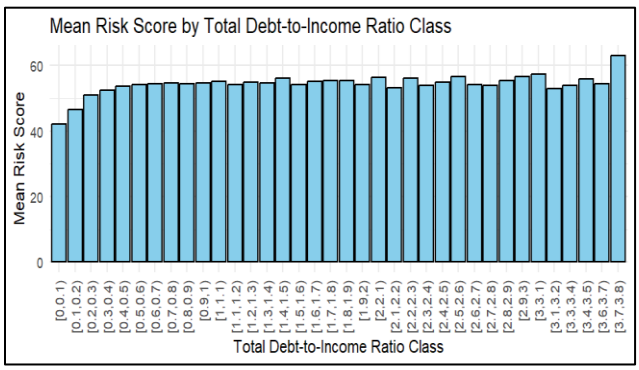


Figure 22:Mean RiskScore by Total DTI

The data shows the relationship between the Total Debt-to-Income (TDTI) ratio and the Mean Risk Score across different TDTI class intervals. As the TDTI ratio increases, the Mean Risk Score also tends to rise, suggesting a positive relationship between the two variables. This indicates that individuals with higher TDTI ratios tend to have higher risk scores, which may reflect the increased financial strain associated with higher debt levels relative to income.

Bankrupt history vs RiskScore

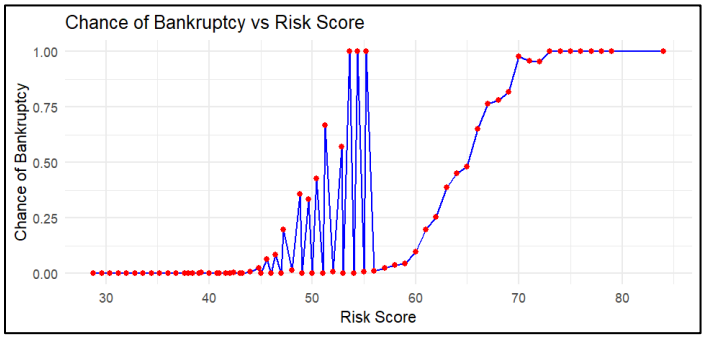


Figure 23:Chance of Bankrupty vs RiskScore

The analysis of the relationship between Risk Score and Chance of Bankruptcy reveals a clear trend: as the Risk Score increases the likelihood of bankruptcy is increased. For lower Risk Scores (below 50), the Chance of Bankruptcy is near zero, with most individuals having no bankruptcy history. However, as the Risk Score rises, particularly above 50, the Chance of Bankruptcy increases significantly, reaching nearly 1 for scores above 70, where almost all individuals have a bankruptcy history. This suggests a positive relationship between Risk Score and the probability of bankruptcy, with higher-risk individuals being much more likely to experience financial distress, which can be critical for risk management and decision-making.

Mean RiskScore vs NetWorth

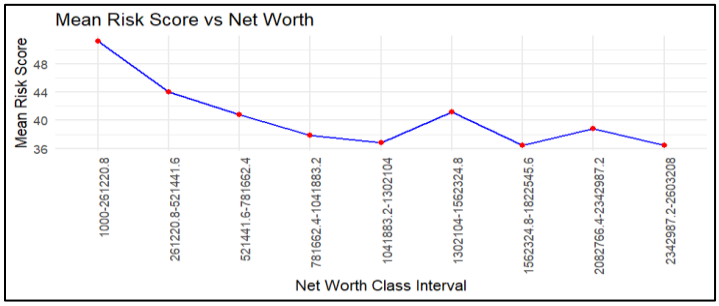


Figure 24:Mean RiskScore vs NetWorth

The analysis reveals an inverse relationship between net worth and average risk score, where individuals with lower net worth tend to exhibit higher average risk scores. As net worth increases, the average risk score generally declines, indicating lower credit risk among wealthier individuals. However, due to the smaller number of observations in higher net worth categories, there is some variability in the trend at the upper end of the distribution. This pattern suggests that net worth can be a meaningful indicator when assessing an individual's financial risk profile.

Mean RiskScore vs InterestRate

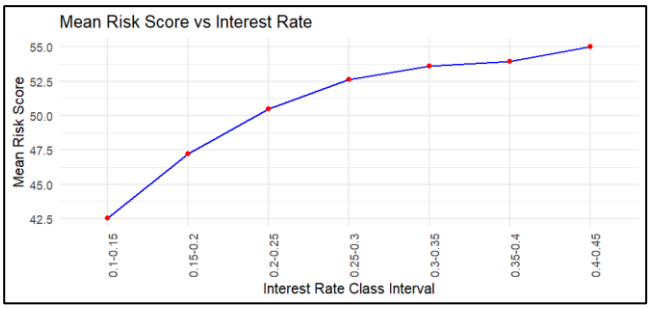


Figure 25:Mean RiskScore vs InterestRate

The analysis of interest rate intervals in relation to average risk scores reveals an upward trend, where higher interest rate ranges are generally associated with increased average risk scores. This pattern suggests that as the interest rises, the risk profile of the borrowers also tends to increase. This distribution supports the interpretation that lenders might be pricing higher-risk borrowers with elevated interest rates, aligning with typical credit risk-based pricing strategies.

Previous LoanDefaults vs RiskScore

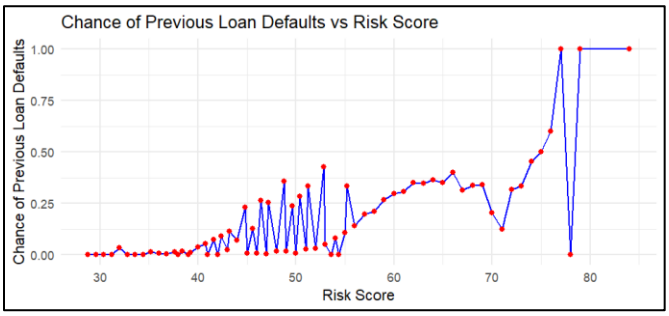
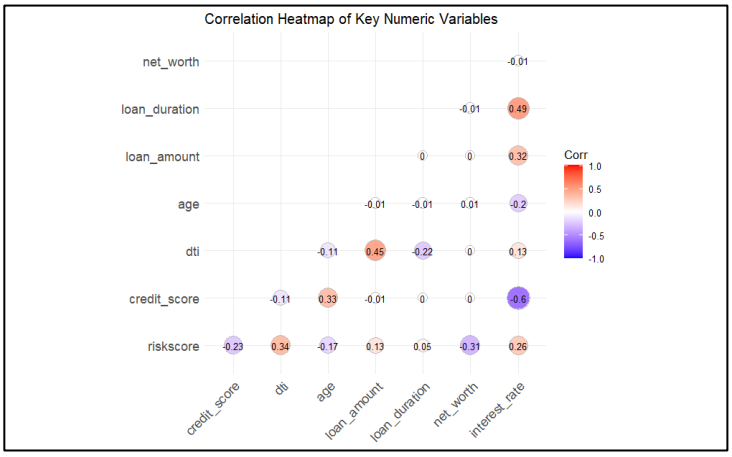


Figure 26:Previous LoanDefaults vs Riskscore

The analysis reveals an upward trend in the likelihood of previous loan defaults as the risk score increases. Borrowers with lower risk scores exhibit minimal to negligible default histories, while those with higher risk scores show a progressively higher probability of having defaulted on past loans. This pattern underscores a positive association between the assigned risk score and historical credit behavior, suggesting that the risk scoring mechanism effectively captures the default propensity based on prior borrowing conduct.

Correlation Heatmap for Some Numerical Variables



To better understand the interrelationships among the key numerical predictors in our dataset, we constructed a correlation heatmap. We focused on some of the variables , chosen based on domain relevance in financial decision-making and statistical importance.

Figure 27:Correlation Heatmap

It suggest that interest rate and risk score has considerable association. and also Age and total DTI show an association with credit score.

For the Risk score , Toal DTI , Credit score , interest rate , net worth show considerable association compared to other.

Cluster analysis

Cluster analysis is performed to identify distinct groups or clusters within the dataset. Initially, we applied Principal Component Analysis (PCA) to identify whether clusters existed or not.

After conducting PCA, we plotted the score plot, which showed that the data points did not form any obvious clusters in the reduced-dimensional space.

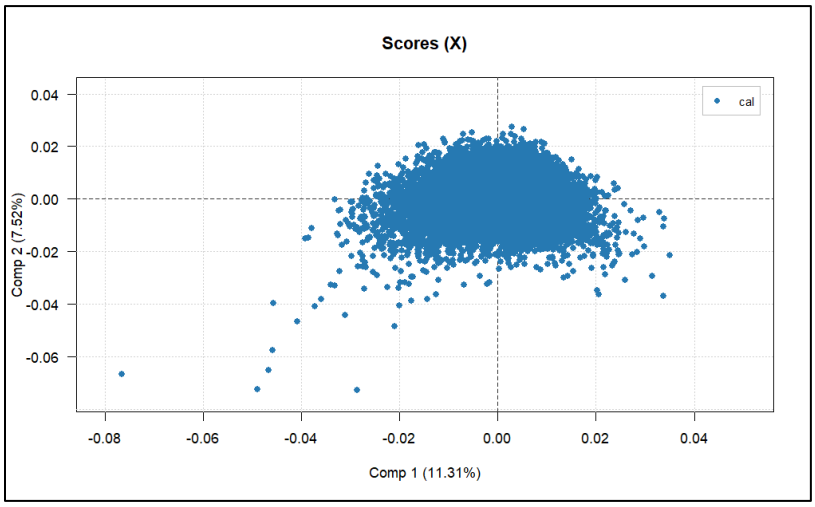


Figure 28:PCA Score Plot

Since we have categorical variables , we next conduct Fisher's Method of Discriminant Analysis (FMDA).but The high dimensionality of the dataset and the presence of multicollinearity between the features hindered the performance of FMDA.

To address this issue, we first performed feature importance analysis to identify the most influential variables (using Random Forest model).This helped us reduce the dimensionality and mitigate the problem of multicollinearity before proceeding to the next step.

To overcome the multicollinearity problem and focus on the most important variables, we applied FMDA with selected variables

K-Means Clustering

We then applied K-Means clustering to identify distinct groups in the data, using k=3 clusters.

The model produced a clustering solution, but the average silhouette score for k=3 was 0.0072, indicating a very weak clustering structure. The silhouette score, which ranges from -1 (poor clustering) to 1 (strong clustering), suggests that the data points were not clearly assigned to their appropriate clusters.

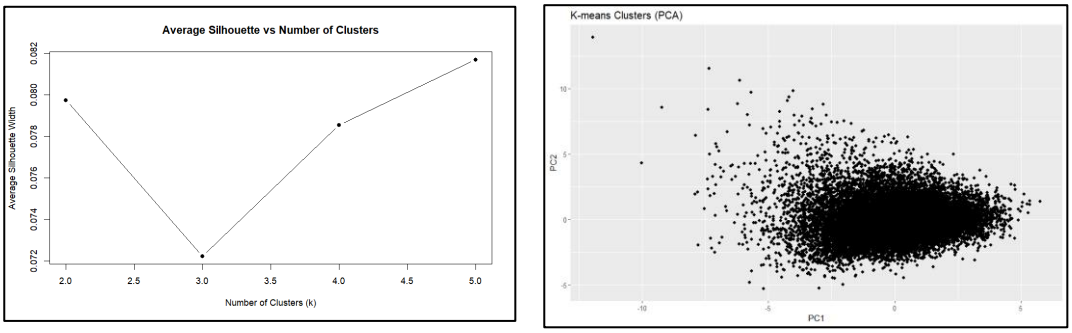


Figure 29:Avg. Sil values vs No of clusters and Cluster plot

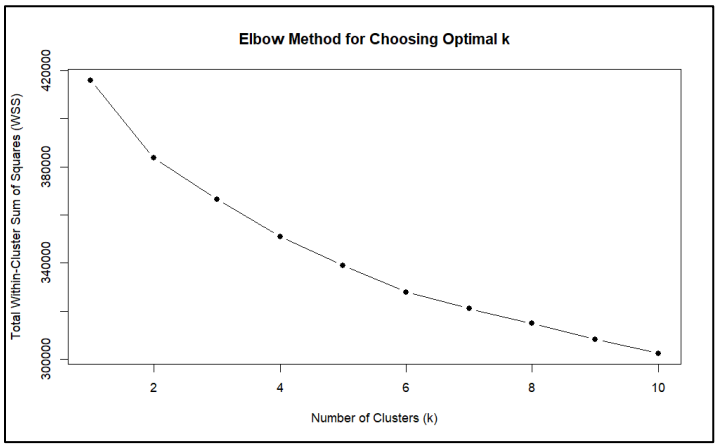


Figure 30:Elbow plot

The elbow plot gives that 3 or 4 clusters is probably optimal for our data. But to further explore the optimal number of clusters, we evaluated the silhouette score across different values of **k** (the number of clusters).

For each **k** value tested, the silhouette score remained low, indicating that the clustering structure was not well-defined.

Based on the results of PCA, FMDA and K-Means clustering, we concluded that the dataset does not contain clear and distinct clusters.

Important Results Of Advanced Analysis

In the advanced analysis phase of this study, the focus was on building predictive models for three key outcomes: **loan approval status**, **risk score**, and **credit score**. These responses were selected to comprehensively capture an applicant's financial profile and support effective decision-making in loan evaluation.

To build robust models for predicting Risk score and Loan Approval status , we followed a two-step modeling strategy:

1. Initial Modeling with All Variables
2. Refined Modeling with Selected Features
   After assessing feature importance using tree-based, we re-trained the models using only the top important variables.

Greater emphasis was placed on the refined models, as our end goal is to integrate the model into a data product that allows for user input. By focusing on a reduced set of key variables, we enhance the model's efficiency, usability, and generalizability , ensuring that it remains both accurate and practical for real-world application.

Building a model to predict Loan Approval Status

In the dataset, the target variable LoanApproved was imbalanced, with significantly more rejected ones (0) than approved loans (1). This imbalance can lead to biased model performance. So we used SMOTE (Synthetic Minority Over-sampling Technique) before model training.

Gradient Boosting classifier

Gradient Boosting Classifier show strong performance and It handles both numerical and categorical variables well.

Random Forest classifier

Random Forest is an ensemble method that builds multiple decision trees and aggregates their results, reducing variance and the risk of overfitting. specially with techniques like SMOTE, Random Forest performs well with class imbalance.

Extreme Gradient Boosting Classifier

XGBoost has ability to handle overfitting and learn from errors efficiently. It Highly optimized for speed ,which is ideal for large or complex datasets.

Important Variables identified by the model

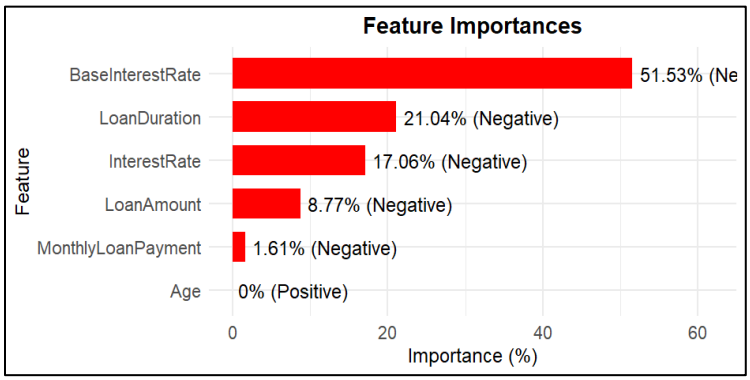


Figure 31:VarImp plot for XG Boost -LoanApproved

For the XG Boost model, BaseInterestRate , Loan Duration , InterestRate , LoanAmount and MonthlyLoanPayment are the key factors in estimating ones loan approved likeability.

Models Comparison

Model	Accuracy		F1 Score	
	Train	Test	Train	Test
GBoost (All Variables)	0.9683	0.968	0.9790	0.9682
GBoost (Imp Variables)	0.9692	0.9685	0.9793	0.9690
Random Forest (All Variables)	0.9978	0.9835	0.9970	0.9893
Random Forest (Imp Variables)	0.9980	0.9872	0.9982	0.9917
XGBoost (All Variables)	0.9987	0.9915	0.9987	0.9945
XGBoost (Imp Variables)	0.9953	0.9900	0.9953	0.9935

Table 2:Model comparison plot for LoanApproved

While the full model provides a slight performance boost, the reduced feature model balances predictive power with practical implementation considerations. By focusing on key variables, we ensure that it remains efficient, user-friendly, and effective in real-world scenarios.

Thus, XGBoost with the reduced variable set is the optimal choice for deployment in the loan approval prediction system.

Building a model to predict Risk Score

Multiple Linear Regression (Forward)

	Train Set	Test Set
RMSE	5.4345133	5.2840702
R-Squared	0.5162032	0.5213442
MAE	4.0562564	3.9217683

Table 3:Evaluation metric of MLR-RiskScore

The Multiple Linear Regression (MLR) model performed reasonably well on both the training and test sets, with an R-squared value of approximately 52% in both sets, indicating that it explains around half of the variance in the target variable. RMSE values were 5.43 for the training set and 5.28 for the test set, suggesting that the model's predictions deviate moderately from the actual values. Additionally, MAE was 4.06 for the training set and 3.92 for the test set. Before making final conclusions, it is important to verify the validity of the model assumptions.

1.Linearity

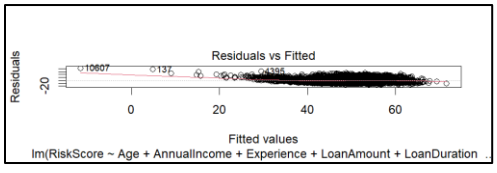


Figure 32:Residual plot of MLR - RiskScore

In examining the linearity assumption through the Residuals vs Fitted plot, the residuals appear to be scattered evenly across the range of predicted values (30-65) without displaying any distinct pattern. This lack of a clear trend suggests that the linearity assumption is likely satisfied.

2.Independence of Residuals

The Durbin-Watson test for independence of residuals yields a test statistic of 2.0044 with a p-value of 0.6105 suggests that there is no significant autocorrelation in the residuals. Since the p-value is much

greater than the typical significance level (e.g., 0.05), we fail to reject the null hypothesis, which indicates that there is no significant autocorrelation present in the residuals. Therefore, we can conclude that the residuals are independent, satisfying the assumption of independence for the model.

3. Homoscedasticity

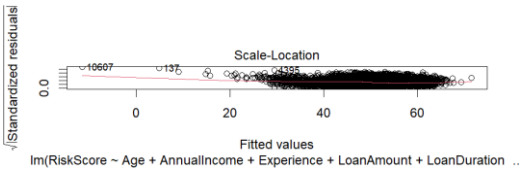


Figure 33:Standized residuals vs fiited values

In the plot of standardized residuals versus fitted values, the distribution appears to be randomly scattered without any clear pattern or trend. This suggests that the variance of the residuals is constant across the range of fitted values. Therefore, we can conclude that the assumption of homoscedasticity holds for this model.

4.Normality of Residuals

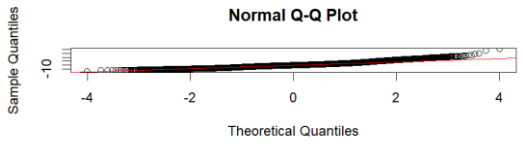


Figure 34:Normal Q-Q plot MLR - RiskSCore

The Q-Q plot of the residuals shows a relatively straight line, with the points closely following the diagonal reference line. This indicates that the residuals are approximately normally distributed, satisfying the normality assumption for linear regression.

5.Multicollinearity

The Variance Inflation Factor (VIF) values for the predictor variables in the model suggest concerns with multicollinearity. Variables such as Age (29.61), AnnualIncome (50.34), Experience (29.73), and NetWorth (41.15) exhibit high VIF values, well above the threshold of 10, indicating strong multicollinearity with other predictors. Although several other variables have VIF values closer to 1, suggesting minimal multicollinearity, the high VIF values for these key variables could lead to instability in the model's coefficient estimates and inflate standard errors. As a result, the multicollinearity assumption of the multiple linear regression model is not fully satisfied. To address this issue, it may be necessary to apply regularization techniques such as Ridge or Lasso regression to reduce the impact of multicollinearity on the model.

Lasso Regression

Lasso is a linear regression technique that applies L1 regularization, which helps to improve model performance by shrinking coefficients of less important features to zero. It effectively handles multicollinearity, improving coefficient stability and interpretability.

For predicting the risk score, we first implemented Lasso regression with a regularization parameter (λ = 0.00611) determined through hyperparameter tuning.

The results are below ,

	RMSE	R squared
Train	4.3492	0.6896
Test	4.3789	0.6824

Table 4:Evaluation matric for Lasso - Risk score

When we evaluate the model , although RMSE for test set is small , the R square is 0.6824. It means

that only about 68% of the total variation of risk score is explained by the model. So we tried other model to increase the model performance

Random Forest and XGBoost

Random Forest and XGBoost are non-parametric models that can capture complex non-linear patterns and interactions between features. They are particularly useful for tasks like risk score prediction, where the relationships between variables which are likely to be non-linear.

	Train		Test	
Model	RMSE	R Squared	RMSE	R Squared
Random Forest (All Variables)	2.242751	0.9472	4.1381	0.7584
Random Forest (Imp Variables)	2.0887	0.9352	3.700	0.7774
XGBoost (All Variables)	2.8393	0.9372	4.138	0.7967
XGBoost (Imp Variables)	2.5722	0.9420	3.6536	0.8163

Table 5:Model comparison - RiskScore

First we fitted the Random Forest model and it give relatively small RMSE , but when we check the R square value also , it might be overfitted. So we fit a reduced model with only variables that are important from variable important as follows ,

Then the RMSE and R square compared for both train and test set , the model performance was increased giving test RMSE as 3.7 and R square 0.7774.

Therefore about 78% of variation is explained by the model.

Next we perform XGBoost and it gave high RMSE than Random forest and test RMSE looks same. But test R square have increased up to 0.7967.

Variable importance was done and refitted the model again with only important variables.

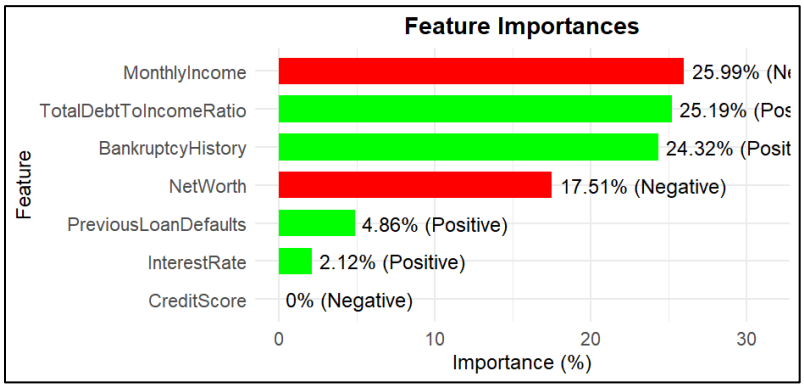


Figure 35:Var Imp plot for XGBoost

For the reduced model test RMSE was low compared to reduced model and test R square has increased as about 81% of the total variation of risk score is explained by the model.

We selected the XGBoost with only important variables in the case of predicting the Risk Score.

Building a model to predict Credit Score

Multiple Linear Regression (Forward)

	Train Set	Test Set
RMSE	0.0055	0.0057
R-Squared	0.9964	0.9961
MAE	0.0037	0.0038

Table 6:Evaluation matrc for MLR-Creditscore

The model achieved a very low RMSE of 0.0055 on the training set and 0.0058 on the test set, indicating that the model's predictions are highly accurate. The R-squared value is also impressive, with 99.64% of the variance explained on the training set and 99.61% on the test set.

Given these excellent results, the next step is to ensure the validity of the model assumptions. To confirm the reliability of the model, a residual analysis will be conducted.

1.Linearity

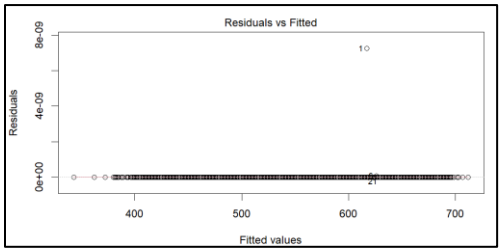


Figure 36:Residual plot for MLR

The residuals vs fitted plot shows that most residuals are randomly scattered around the horizontal line at zero, with no obvious patterns or systematic structure. This suggests that the assumption of linearity.

2.Independence of Residuals

To evaluate the independence of residuals, the Durbin-Watson test was conducted. The test yielded a Durbin-Watson statistic of 0.99877 with a p-value less than 2.2e-16. Therefore, we conclude that the assumption of independence of residuals in the multiple linear regression model is violated**.**

3.Homoscedasticity

To test this assumption, the Breusch-Pagan test was conducted. The test returned a BP statistic of 16.188, with 23 degrees of freedom, and a p-value of 0.8471. Thus, there is no statistical evidence of heteroscedasticity, and it can be concluded that the assumption of homoscedasticity is satisfied.

4.Normality of Residuals

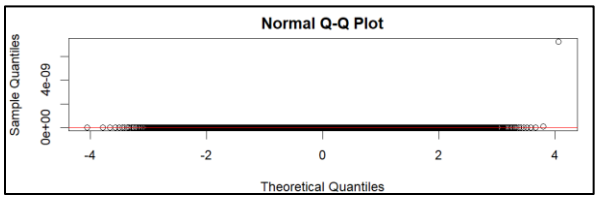


Figure 37:Normal Q-Qplot

In the Q-Q Plot , vast majority of points are hugging the horizontal axis near zero. In a normal Q-Q plot, the points should lie approximately along the diagonal. So it suggest that normality assumption is violated.

Since the assumptions of MLR are violated we used tree based methods Random Forest and XGBoost. They are free of assumptions and avoid overfitting.

Random Forest

	Train Set	Test Set
RMSE	4.872396	6.76518
R-Squared	0.9941715	0.9599265
MAE	3.164816	5.894666

Table 7:Evaluation matric for Random forest - Credit score

when we compared the training and testing measures it gives that the model showed some what better fit with train data that test set ,suggesting that the model's predictive accuracy decreases when applied to new, unseen data. This discrepancy points to potential overfitting on the training data, where the model performs well on known data but struggles to generalize.

XGBoost

	Train Set	Test Set
RMSE	2.28755	4.888673
R-Squared	0.9979948	0.9806792
MAE	1.722231	2.979503

Table 8:Evaluation matrics for XGBoost-LoanApproved

The XGBoost showed good performance for both train and test set than the random forest. For the test set the RMSE gives 4.88673 and R square given as 0.98 suggesting that 98% of total variation of credit score is explained by the model. Overall, the model shows strong performance.

Important variables identified by the Model

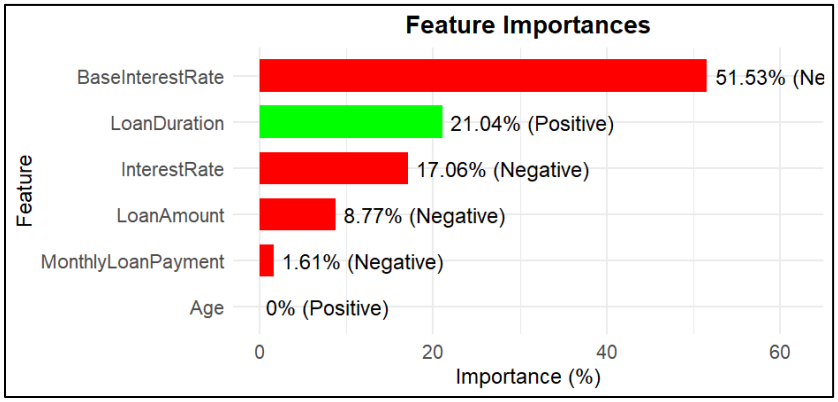


Figure 38:Var Imp for LoanApproved

Base InterestRate , LoanDuration , Interest Rate , LoanAmount and Monthly loan Payement are the important variables identified by the model.

Model Comparison

	Train set		Test set	
Model	RMSE	R squared	RMSE	R squared
MLR	0.0055	0.9964	0.0057	0.9961
Random Forest	4.872396	0.9942	6.7652	0.9599
XG Boost	2.2875	0.9979	4.8886	0.9806

Table 9:Model comparison - Credit score

After comparing all models, XGBoost is recommended as the best model for predicting credit scores. It provides strong predictive performance with minimal overfitting, capturing most of the variance in the data and generalizing well to new observations.

Conclusion

Based on the advanced analysis conducted, the optimal predictive model for each target variable—Loan Approval Status, Risk Score, and Credit Score—was identified through careful model building, comparison, and validation.

For Loan Approval Status, although all three models (Gradient Boosting, Random Forest, and XGBoost) performed well, the XGBoost model using a reduced set of important features emerged as the most suitable for deployment. It balanced high accuracy (99.00%) and strong F1 score (99.35%) on the test set while maintaining computational efficiency and interpretability. The reduced feature set enhances practical usability, especially in data products that require quick, user-input-based predictions. Variables such as Risk Score, Credit Score, Debt-to-Income (DTI) Ratio, Income, and Interest Rate played significant roles in determining loan approval likelihood.

In the case of Risk Score prediction, the XGBoost model with selected important features again proved to be the best option. It delivered a low Root Mean Squared Error (RMSE) of 3.65 and an R-squared value of 0.8163 on the test set, indicating that over 81% of the variation in Risk Score is explained by the model. This model outperformed both Lasso regression and Random Forest, particularly in its ability to generalize to unseen data without significant overfitting. Key predictors included Credit Score, DTI, Income, and Payment History.

For Credit Score estimation, while the Multiple Linear Regression (MLR) model showed exceptionally high R-squared values (~99.6%) and very low RMSE, it violated critical regression assumptions such as independence and normality of residuals. Therefore, tree-based models were also evaluated. Between Random Forest and XGBoost, the XGBoost model once again outperformed the others, achieving an R-squared of 0.9807 on the test set and demonstrating minimal overfitting. This model successfully explained 98% of the variance in Credit Score while maintaining robustness across train and test datasets.

In summary, XGBoost with reduced, important features is the recommended model for all three target variables due to its high predictive power, generalizability, and suitability for real world applications. It stands out as the most effective and practical choice for deployment in a loan evaluation system.

References

1. Investopedia. (n.d.). *Credit Scoring*. Retrieved from https://www.investopedia.com/terms/c/credit_scoring.asp

2. Intaver Institute. (n.d.). *Risk Scores in Project Risk Analysis*. Retrieved from https://intaver.com/blog-project-management-project-risk-analysis/risk-scores-2/

3. FasterCapital. (n.d.). *Loan Risk Modeling: How to Measure and Mitigate Your Loan Risks*. Retrieved from https://fastercapital.com/content/Loan-Risk-Modeling--How-to Measure-and-Mitigate-Your-Loan-Risks.html

4. Money Help Center. (n.d.). *What Is My Credit Score?*. Retrieved from https://www.moneyhelpcenter.com/calculators/what-is-my-credit-score/

5. FasterCapital. (n.d.). Loan Risk Assessment: How to Use *Credit Scoring and Rating Models to Evaluate Your Loan Applicants*. Retrieved from https://fastercapital.com/content/Loan-Risk-Assessment--How-to-Use-Credit-Scoring and-Rating-Models-to-Evaluate-Your-Loan-Applicants.html

6. Paisabazaar. (n.d.). *How to Track the Status of a Personal Loan Application*. Retrieved from https://www.paisabazaar.com/personal-loan/how-track-status-personal-loan application/

Appendix

1. **Dataset Source**

   The dataset used in this analysis is obtained from Kaggle. It contains detailed financial and demographic information useful for evaluating the risk involved in loan approvals.

   - Title: *Financial Risk for Loan Approval*
   - Link: https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for loan-approval

2. **R Code Repository**

   All R scripts used for data preprocessing, exploratory data analysis, model fitting, and evaluation are available in the following Google Drive folder:

   - Google Drive Link: Loan Approval Probability Analysis – Supporting Materials