MERCER UNIVERSITY

SCHOOL OF BUSINESS

PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE

IN

BUSINESS ANALYTICS

OFFERED BY

STETSON-HATCHER SCHOOL OF BUSINESS

AT

MERCER UNIVERSITY

Project Title: Prediction of Loan Default by First Time Borrowers

Advisor: Nikanor I. Volkov

Prepared by: Sowndarya Saini & Goutham Yallapu

Table of Contents

1. Introduction:

Using the banking and financial services industry as a case study, this investigation explores the combination of exploratory data analysis (EDA) and data mining techniques in risk analytics.

The risk involved in lending, where financial institutions must weigh the necessity to accept loans for expansion against the possibility of suffering financial loss from defaults, is the foundation of this issue. When applicants have no credit history, it might be challenging to appropriately determine their risk using traditional approaches.

By using EDA and data mining approaches to analyse loan application data, the case study seeks to solve this by identifying trends and insights that may be useful in forecasting loan defaults. It aims to accomplish these goals by improving the loan approval decision-making process, lowering the rate of bad debt, and guaranteeing credit is accessible to those who meet the requirements.

This enhances the overall business outcomes of the company while lowering the chance of financial loss.

Because it shows how data can be used in a real-world, business-focused environment, this study is crucial because it goes beyond the theoretical uses of EDA taught in an academic setting. Through the case study, which highlights the critical role that data plays in reducing loan risks and bolstering the financial stability of lending institutions, a basic grasp of risk analytics in the banking industry is introduced.

Applying analytical methods to thoroughly examine loan application data is the goal of this study, as analysing loan applications can be difficult, particularly for those without credit histories. To lower the possibility of financial loss and guarantee that loans are awarded to qualified applicants, the goal is to identify patterns and trends that may guide risk management decisions. By weighing opportunity and risk, the ultimate objective is to maximize commercial outcomes for the financial institution.

The case study illustrates how EDA can be used practically in a real-world corporate setting. It provides insights into risk analytics in the banking and financial sector and goes beyond applying EDA techniques that are acquired in an academic setting. Understanding how data can help reduce the financial risks associated with lending and how data mining techniques can support safer lending practices are the goals of this investigation.

By describing the issue, its significance to the financial sector, and the used data analysis approach, this introduction establishes the framework for the case study. The case study's instructional purpose is also established, demonstrating how EDA and data mining might be applied to risk analytics.

The main risks that financial organizations consider when deciding whether to approve a loan:

Financial Loss: On the other hand, in the event that a loan is authorized for a borrower who is untrustworthy or unable of fulfilling their financial commitments, the lending organization runs the danger of suffering a loss. Either the default itself or the expenses incurred in trying to recoup the money that was due can cause this loss. Losses can also include lost opportunities resulting from choosing to lend money to a high-risk borrower rather than a more trustworthy applicant. Inadequate due diligence done during the loan approval process or aggressive lending practices are frequently associated with this kind of risk.

An organization risks unsustainable losses if it is too permissive and misses out on opportunities for growth and profitable ventures. Accurately evaluating applicants' creditworthiness and making well-informed judgments that will benefit the organization without exposing it to undue risk are key components of effective risk management. In order to identify credit risks based on historical data and predictive analytics, EDA and data mining approaches are helpful in this situation.

2. Literature review:

Financial services are a broad range of offerings made available by organizations to manage and facilitate many elements of monetary transactions, wealth management, and financial well-being. These services are essential to the operation of modern economies, providing individuals, businesses, and governments with tools for effective financial management. Financial services include banking, insurance, investment, and advisory services. Whereas Banking services represent a subset of financial services that primarily focus on the management, safeguarding, and utilization of funds. Banks serve as crucial financial intermediaries, offering a myriad of services to individuals, businesses, and governments. Core banking services include deposit-taking, lending, and payment facilitation.

In our paper, we have included a wide range of features and attributes related to loan default payments in the dataset, which seeks to ease the study of credit risk. According to the study "Risk Management 4.0: The Role of Big Data Analytics in the Bank Sector" by Grazia

Dicuonzo et al., highlights the importance of advanced technological infrastructures in small banking institutions for effective risk management within the constraints of internal regulations and supervisory limits. It also emphasizes the rapid identification and quantification of new risks, transparency in reporting activities, and the integration of traditional information sources with unstructured data using advanced technological tools.

Moreover, The Everest Group's "Analytics in Banking" report highlights the transformative power of analytics in the banking industry, particularly predictive and prescriptive models, which are useful in understanding customer behaviour, optimizing products and portfolios, and mitigating fraud and money laundering risks. Despite the promise, banks have hurdles in implementing analytics due to functional barriers, a personnel shortage, obsolete data systems, and conflicting agendas. It also underlines the importance of constant model refining and leveraging global sources to gain expertise and preserve model integrity. The banking sector's use of analytics is still in its early stages, implying a substantial unmet opportunity for value creation.

Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA): The work focuses on creating a strong loan default prediction model in order to solve the urgent problem of loan defaults in the financial sector, which present serious dangers to economies in both developed and developing countries. A comprehensive assessment of the literature was done, looking at earlier research that used data mining methods to forecast loan default. Information Discovery in Databases (KDD), Sample, Explore, Modify, Model, and Assess (SEMMA), and the Cross-Industry Standard Process for Data Mining (CRISP-DM) are among the approaches examined. A movement away from traditional credit scoring models and toward more dynamic and predictive frameworks is evident in the examined literature, which favors the use of machine learning algorithms and sophisticated statistical models. Comparative Analysis: Because of SEMMA's methodical and thorough approach, the research compares several data mining approaches used in loan default prediction. The goal was to identify the best model for financial risk prediction by evaluating the performance of these approaches according to predetermined criteria. By analysing several predictive models' efficacy in relation to loan defaults, the study adds to the corpus of knowledge. In order to improve financial lending decision-making, it offers a thorough comparison of approaches. By determining the most important variables causing loan defaults and utilizing appropriate predictive models, the research intends to help financial institutions reduce bad loans while preserving the stability of the economy.

3. Data, data sources and data characteristics:

We have selected the dataset from Kaggle Risk Analytics In Banking & Financial Services 2 (kaggle.com) and data

It contains 307505 observations and 122 variables. The dataset appears to be extensive, encompassing several facets of a customer's profile and loan particulars. Indicators of the socioeconomic class, such as income, credit limit, and loan annuity, are included in addition to demographic data, such as gender, vehicle and real estate ownership, and family status. In-depth information is also included, such as the customer's occupation type, educational background, residential area, contact information availability, and a number of characteristics pertaining to the customer's living circumstances, including apartment size and building age.

The dataset also has a significant focus on the customer's social circle, as seen from variables detailing social defaults and inquiries made to credit bureaus. The goals of the case study, which include applying EDA and data mining techniques in risk analytics, are in line with the importance of such data in risk analytics for forecasting loan defaults. In order to conduct additional data analysis, it would be necessary to investigate the distribution of each variable, any missing values, and any potential impact on the target variable—which represents payment difficulties. It looks that the dataset came from a Kaggle competition, which is popular for machine learning and risk assessment. Predictive modeling appears to have been the intended application for the data based on its characteristics, including the format and kind of variables (continuous, categorical, etc.).

4. Prepare data for analysis: collect, clean, and process data:

4.1 Data dictionary:

Refer to Appendix

4.2 Clean and preprocess the data:

To address the missing values in the dataset, we used a range of strategies, according to the requirements of each variable. Based on the quantity of missing values and the knowledge domain, variables are chosen. Initially, we eliminated 93 factors and only looked at 29 variables for more examination.

To determine the optimal imputation technique for numerical variables, we computed mean, median, and skewness. Since the median is less impacted by outliers, it is typically employed for skewed distributions. For amount annuity, for instance, the skewness of the data was calculated and then the median was used to fill in the missing values. For amount good price, for instance, the skewness of the data was calculated and then the mean was used to fill in the missing values. Count of family members, observed 30 days count social circle, default 30 days count social circle, observed 60 days count social circle, default 30 days count social circle, no of credit enquires in a year, days since last phone change are replaced with median

They substituted the mode—the value that occurs the most frequently—for missing values in categorical variables like occupation type. With lacking data, this method assumes that the most prevalent category is the most plausible classification.

Additionally, a new variable occupation type has been developed based on the customer's occupation. Employees working in management, accounting, high-skilled tech, medicine, or IT are classified as highly skilled; employees working in sales, core services, private services, driving, cooking, secretarial work, real estate, or human resources are classified as medium skilled; and employees working in labour, security, cleaning, low-skill labour, waiters/barmen are classified as low skilled. We replaced the numbers based on income levels after seeing that several of the observations lacked information about occupation. As low skilled, a consumer's income falls below the first quartile of the income variable; as high skilled, it falls above the third quartile of the income variable; otherwise, it is classed as medium skilled.

4.3 Reduce the data dimension:

Correlation analysis was used to evaluate the association between different numerical features in the dataset. The highly linked variables, which frequently suggest redundancy, were found by using the correlation matrix. For example, they discovered that the correlation coefficient between default 30 days count social circle and default 60 days count social circle was roughly 0.86, indicating a very significant positive link. Count family members and count children also showed a strong correlation, with a coefficient of about 0.89.

They decided to remove some variables in order to reduce multicollinearity, which can skew the outcomes of statistical models like logistic regression, in response to observations of high correlation. They eliminated several aspects, such as amount good price as it is correlated with amount of credit, region rating client with city is correlated with region client. This method is widely used in data analysis to guarantee that the predictors offer distinct information about the response variable and the models are strong. Hence, we omitted some of the variable based on correlation with other variables.

For additional modelling and analysis, we then employed the updated dataset with reduced multicollinearity. In order to increase model performance and interpretability, this is an essential stage in the preparation of data for machine learning algorithms.

4.4 Determine methods and techniques appropriate for data and problem:

Predictive modeling convention dictates that the dataset be split into training and testing sets. 70% of the data was set aside for training, which is how the models were built and adjusted. To assess the models' efficiency and verify their applicability to new, untested data, the remaining thirty percent were used as a testing set. This split keeps a balance between the amount of data needed to test the models' predictions and the amount needed to adequately train the models.

Logistic regression since it is a suitable tool for binary classification problems, such as forecasting loan defaults (where Target variable is 0 or 1 where 1 represents loan default and otherwise zero). This helped define the best approaches and strategies for assessing the loan application data. For risk assessment in financial contexts, logistic regression is a perfect tool since it can accurately calculate the likelihood of a binary outcome based on predictor factors.

They also considered other data pre-treatment methods that are essential for enhancing model interpretability and sensitivity, like managing missing values and lowering multicollinearity across predictor variables. In order to make educated lending decisions, it was necessary to comprehend how different predictors affect the probability of default and make sure the model could be comprehended in a commercial setting. This led to the decision to use logistic regression.

Random Forest: This is a decision tree classifier-based ensemble learning technique. It provides robustness by averaging several deep decision trees that were trained on various portions of the same training set. This helps to lower the variance of the model. When managing non-linear data with a large dimensionality, Random Forest is very helpful as it may reveal the significance of individual features.

Using the concept of predictor independence, the Naive Bayes algorithm is a probabilistic classifier that relies on the Bayes theorem. Naive Bayes works well in situations where computing efficiency is critical and is especially effective when the dimensionality of the input is high in comparison to the amount of data points.

By offering different angles on the data and supporting the validation or improvement of your model's predicted accuracy and robustness, these techniques could be used in addition to logistic regression analysis.

Exploratory data analysis

1. **Age:** According to Graph 1, most borrowers across all age categories fall into the Non-Default category, showing a general trend of dependable repayment among borrowers. The percentage of defaults seems to be comparatively larger in younger individuals.

   There are multiple possible reasons for the greater default rates among younger borrowers, as observed:

   - Financial Stability: Due to their recent entry into the workforce and entry-level pay, younger people may have less secure financial circumstances.
   - Credit History: Their credit histories tend to be shorter or less established, which may not accurately reflect their financial habits or trustworthiness.
   - Financial obligations: The ability of younger persons to fulfil all of their financial obligations may be impacted by their greater variable costs, such as school debt.
   - Lack of experience managing credit may also cause them to make poor judgments that negatively impact their credit standing.

Comprehending these variables can facilitate the creation of customized financial advice and risk evaluation tactics for younger debtors.

2. **Type of Loans:** According to Graph 2, the percentage of defaults in the Cash Loans category is marginally greater than in the Revolving Loans category.
   - Cash loans are a type of instalment loan that may entail more risk because of their larger sums and longer payback terms. This category's somewhat higher default rate may be an indication of the potential financial burden these terms may place on borrowers.
   - Revolving Loans: Typically, they give borrowers additional freedom, like the option to borrow back funds up to a predetermined amount. Because of its flexibility, which enables borrowers to better manage payments based on their financial circumstances, it may have a reduced default rate in this instance.

3. **Skill Levels**: This Graph 3 can be used to quickly visualize the possible variations in loan default rates among various skill-level-defined groupings. Compared to people with medium and high skill levels, the percentage of defaults appears to be higher among low skill persons, indicating that skill level may be a significant predictor of loan default.
   - Economic Factors: Jobs requiring less expertise typically pay less and may have less job security, which raises the possibility of financial difficulties culminating in loan defaults. Their capacity to regularly fulfil debt obligations may be impacted by this financial volatility.
   - Financial Literacy and Access: People's debt management and financial planning strategies may be impacted by skill levels that are correlated with financial literacy. Reduced financial literacy may be correlated with lower skill levels, which could result in worse financial judgments and a rise in default rates. Furthermore, it may be more difficult for less skilled people to acquire financial resources like emergency credit or advantageous loan conditions, which makes up for any deficits in income.
4. **Region Rating Client:** As per Graph 4**,** The percentage of defaults is observably greater in "Low" rated regions than in "Medium" and "High" rated regions. This points to a pattern where default rates are higher in areas with lower ratings. The graph suggests that a good indicator of loan default risk may be the region rating. Lower rated regions may be linked

to variables like less financial literacy, lower average wages, or unstable economies, all of which could raise the risk of loan defaults.

5. **Credit Income Ratio:** As per Graph 5, It is a financial metric used by the lenders to assess the borrowers ability to repay the loan. It is capped at 5 times the annual income of the borrower, with observations exceeding this threshold removed from the analysis to focus on a financially reasonable range. Histogram (graph 5) shows the distribution of credit income ratios, which exhibits a right-skewed pattern, showing that the majority of borrowers have relatively low credit income ratios with a modal peak in the lower range. This shows that most borrowers practice good financial management by keeping their credit levels modest in relation to their income. However, the distribution shows multiple spikes in higher ratio values, indicating the occurrence of outliers where borrowers have much higher ratios. These higher values show a subset of the borrower population that may be overleveraged, implying financial weakness and an increased chance of default. Understanding the distribution and extremes of credit income ratios allows lenders to enhance risk assessment models and adapt lending policies to reduce potential risks associated with high debt levels.
6. **Gender:** Graph 6 indicates that males (M) and females (F) exhibit similar patterns in loan repayment behaviour. The tiny red segments at the bottom of each bar indicate the percentage of people who have defaulted. Notably, males had a little larger share of default than females, indicating a difference in default rates between the two genders. This graphic is critical for understanding gender-specific financial activities and can help financial institutions modify risk management and client engagement strategies.
7. **Education level:** Graph 7 clearly illustrates how educational achievement connects with loan repayment patterns. Loan defaults are relatively low across all education levels, but are significantly higher in lower secondary education and incomplete higher education. This pattern may imply that higher educational attainment is associated with better financial management or greater financial stability, resulting in lower default rates. The graphic emphasizes the need of including educational background in risk assessment models employed by financial institutions, as it appears to be a substantial predictor of possible default risk.
8. **Population**: This Graph 8 shows various peaks, indicating that both groups are present in a variety of differentially inhabited areas. However, the density of defaulters increases slightly at certain places, notably at lower population levels, implying that default is slightly more likely in less populous areas. Overall, while regional population size does show

significant difference in density between defaulters and non-defaulters, the influence appears to be minor, implying that other indicators may be more predictive of loan default risk. This analysis can help financial institutions understand how area demographics relate to loan repayment practices, but it should be combined with other socioeconomic indicators for a more complete risk assessment.

5. <u>Methodology</u>:

This chapter includes a detailed description of three chosen machine learning algorithms namely Logistic Regression Model, Random forest model and Naïve Bayes Model including but not limited to evaluation matrices, namely, specificity, sensitivity, accuracy and
Area under Curve(ROC)

Logistic Regression:

Logistic Regression model is a statistical technique for simulating the likelihood of a binary result depending on one or more predictor variables. It is utilized only when the dependent variable is binary (or dichotomous). For instance, in our project, it may forecast whether loan would default or not. The formula for Logistic regression can be expressed as follows:

$$P(y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}$$

Here, P(y=1|X) is the probability that dependent variable equals a case (often referred as 1), given the predictors X. $\beta 0$, $\beta 1$,…,βn are referred as coefficients, and e is the base for natural logarithm.

In logistic Regression, for every unit change in the relevant independent variable, coefficient estimates the change in the dependent variable's log odds. This model is extensively used in financial services field which helps in predicting the chances of default on credit payments, bankruptcy etc.

Random Forest Model:

Random Forest model is a powerful machine learning technique which is widely used for both classification and regression tasks. It operates by constructing a large number of decision trees during training phase and outputs the class i.e., the mode of the classes (for classification tasks) or mean prediction (for regression tasks) of each individual trees.

How Random Forest Model works:

Random Forest Model works by building the technique bootstrapping which means randomly selecting a subset of the training dataset to build each tree. Every tree in the forest was constructed using a replacement sample that was taken from the training set. Rather than selecting the best split among all features, the best split from a random subset of the features is used to split a node throughout the tree-building process. This increases the model's variety, which lowers the correlation between the trees and strengthens the predictions taken as a whole.

$$\text{Random Forest Prediction} = \text{mode}(\text{predictions of all trees})$$

Then each tree is grown to its full potential without pruning, resulting in trees that are deep and complex. In classification tasks, Random Forest employs the Gini impurity or entropy, while in regression, it uses mean-squared error reduction to find points (or "splits") that best segregate the data in each node.

After that, Random Forest uses the majority vote from all the trees in case if it is for classification. Or else for a regression, it averages the results from all trees. This is known as aggregation which helps to reduce model variance by mitigating the overfitting problem of individual decision trees to training data.

Advantages of using a Random Forest Model is that it can handle large datasets and it is known for its robustness which is effective for datasets with missing values and maintains accuracy. It is widely used in the field of Medicine, E-commerce and Banking industry which is used for credit scoring and predicting loan default.

Naïve Bayes Model:

Naive Bayes model is a popular probabilistic classification technique based on Baye’s Theorem, with a strong (naive) assumption of independence among predictors. It is particularly known for its efficiency, effectiveness, and simplicity, especially in large datasets. Naïve Bayes classifiers are basic "probabilistic classifiers" that apply Bayes' theorem with the assumption of conditional independence between each pair of features given the value of the class variable.

How Naïve Bayes Model works:

Naïve Bayes classifiers heavily rely on Baye’s Theorem. It is expressed as follows:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Here, P(A|B) is the probability of class (A, target) given predictor (B, attributes). P(B|A) is the likelihood which is the probability of predictor given class. P(A) is the prior probability of class. P(B) is the prior probability of predictor.

In this model, for each class the algorithms learn the prediction distribution. It calculates the probabilities of predictor values for each class based on the training data's frequencies. When presented with several predictors, the method simply multiplies the probabilities to calculate the likelihood. This multiplication is correct on the naive assumption that the qualities are conditionally independent in terms of the target class.

In order to make a prediction, Naive Bayes takes the likelihoods computed for each class, multiplies them by the prior probabilities of each class, then normalizes them to provide probabilities ranging from zero to one. The prediction will result in the class with the highest posterior probability.

Naive Bayes model has several strong advantages that make it extremely valuable in the field of machine learning, particularly for classification tasks. To begin, its efficiency and simplicity stand out; the model can swiftly provide predictions even when working with big datasets, making it a good choice for applications that require rapid processing, such as real-time prediction systems. Furthermore, its resistance to irrelevant features aids in instances when the dataset contains a huge number of features, not all of which are valuable for prediction. These properties make Naive Bayes a useful and adaptable tool in the armoury of data science tools.

Performance evaluation metrics:

Accuracy is the ratio of correct predictions to total observations. Specificity (Precision) referred as the proportion of true negatives correctly identified. Sensitivity (Recall) defined as the ratio of true negatives correctly identified. Accuracy, precision, and recall are calculated using the following formulas:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Area Under Curve - Receiver Operating Characteristic (AUC-ROC):

The AUC-ROC curve is an important evaluation statistic for classification models because it provides a comprehensive measure of model performance across many threshold settings. The Receiver Operating Characteristic (ROC) illustrates the relationship between True Positive Rate (TPR) and the False Positive Rate (FPR) at various levels, demonstrating a binary classifier system's diagnostic capacity. The Area Under the Curve (AUC) measures the model's ability to differentiate across classes. An AUC of 1 represents perfect prediction, but an AUC closer to 0.5 suggests performance comparable to random chance. The AUC is especially useful since it gives a single metric to compare multiple models and is independent of the classification threshold, making it perfect for evaluating models in a variety of scenarios and requirements.

5.2 Detailed Empirical Results:

Dependent variable: Target (Binary outcome)

Independent variable: Credit to Income Ratio, Population Density, Client Region Rating, No. of credit enquiries, 60 days Default count, Education level, Gender, Type of loan, Years Employed, Age, and Occupation.

i) Interpretation of Logistic Regression Model:

Logistic regression analysis identified several key predictors significantly influencing the likelihood of loan default.

Coefficients Interpretation:

Credit to Income Ratio: For each unit increase in the credit to income ratio, the odds of a loan default increase by approximately 4.1%. This finding is statistically significant, indicating a relationship between higher credit utilization and increased default risk. This aligns with conventional financial wisdom that higher debt burdens relative to income can strain borrowers' financial stability.

Population: A negative coefficient for population (-1.7646) suggests that higher population metrics are associated with a lower probability of default. This outcome is significant, potentially reflecting stronger socio-economic activities or better financial support systems in more populous areas.

Regional Rating: The coefficients for medium (0.4023) and low (0.7463) regional ratings indicate that regions with lower ratings significantly increase the likelihood of default compared to high-rated regions. This effect is more pronounced for regions rated as low, which aligns with potential socio-economic challenges, crime rates or lesser creditworthiness in these areas.

Gender: The analysis shows that males are more likely to default than females, with 22.5% higher odds of defaulting. This difference is statistically significant and may reflect varying financial behaviours or obligations between genders.

Type of Loan: Revolving loans are associated with a 39.9% lower odds of default compared to cash loan types, suggesting that these loans, typically with more flexible repayment terms, might be managed more effectively by borrowers.

Years Employed: Longer employment duration correlates strongly with reduced default risk, with each additional year of employment decreasing the odds of default by 20.9%. This result is highly significant and underscores the stability that sustained employment provides in meeting financial commitments.

Age: Each additional year in age increases the odds of default decreases by 1.8%, a finding that is highly significant. Older individuals may have more savings, reducing their likelihood of default.

Occupation Type: The coefficients for low and medium-skilled occupations suggest that individuals in these categories have higher odds of default compared to those in high-skilled occupations, emphasizing the impact of job nature and economic stability on financial reliability.

Performance Evaluation:

Confusion matrix for Logistic Regression shows that it is highly effective at identifying the negative class but at the cost of a large number of false negatives with 91.19% of Specificity. Sensitivity which correctly identified actual positives stands at 22.1% and Accuracy which corrected both negatives and positives at 28.33%

The ROC curve starts at the bottom left of the plot (0,0) and rises towards the top left corner and then moves horizontally towards the top right corner. This ROC curve suggests the model does better than random guessing. The AUC value of the curve stands at 65.1% (Refer Figure 1)

ii) Interpretation of Random Forest Model:

The Out-of-Bag(OOB) error rate is a mean prediction error on each training sample, using only the trees that did not have the sample in their bootstrap sample stands at 11.73%. In other words, it suggests that the model has a high accuracy rate classifying approximately 88.27% which is good for many practical applications.

The model correctly predicted '0' (non-default) for 46,301 instances. The model incorrectly predicted the non-default class as default for 9,298 instances. The model incorrectly predicted the default class as 'non-default' for 3,699 instances. The model correctly predicted '1' (default class) for 51,472 instances.

Class Error for '0' stands at 16.72% indicates the proportion of actual non-target instances that were incorrectly classified by the model. Class Error for '1' stands at 6.70% indicates the proportion of actual target instances that were incorrectly classified by the model.

Overall, The Random Forest model shows strong performance with a relatively low OOB error rate, suggesting it is effectively capturing the relationships within the data to distinguish between the target and non-target classes.

Performance Evaluation:

Confusion matrix for Random Forest Model shows that it is highly effective at identifying the negative class but at the cost of a large number of false negatives with 85.77% of Specificity. Sensitivity which correctly identified actual positives stands at 28.1% and Accuracy which corrected both negatives and positives at 33.3%.

Feature Importance Chart:

Random Forest Model offers metric gauges the impact of each feature on the model's predictive accuracy by calculating the decrease in model accuracy when the data for that feature is permuted while keeping all other features constant. (Refer Figure 2)

Years Employed: This feature shows the highest mean decrease in accuracy, indicating that it is the most important predictor in the model. This might reflect the stability or reliability of an individual's income, which is crucial in many predictive scenarios like loan default prediction.

Education level: Following closely is the education level, which has a significant impact on model accuracy. This suggests that an individual’s degree of education has a considerable impact on the model's predictions, potentially functioning as a proxy for their earning potential or financial literacy.

Credit income ratio: This feature, representing the ratio of amount credit to their income, also ranks highly in importance. Its significant role underscores the direct relationship between an individual's financial commitments and their financial health or risk profile.

Credit enquiries: This predictor, which likely measures the frequency of credit checks or applications for new credit, also impacts model accuracy considerably. Frequent credit inquiries might indicate financial stress or higher credit activity, factors relevant in assessing credit risk.

Population: The size or density of the population in an individual's area has a moderate impact on the model's accuracy, suggesting that demographic and geographic factors play a notable role in the prediction.

Age and Type of loan: Both features have a relatively lower but notable influence on the accuracy. Age might relate to financial stability or experience, while the type of loan could reflect different levels of risk associated with different credit products.

Region Rating, Occupation Type, days60_Default_count and Gender: These features demonstrate the least impact on model accuracy in the current analysis. While still important, their lower positions suggest that other factors might be more predictive of the outcome in the specific context of this model.

iii) Interpretation of Naïve Bayes Model

The Naive Bayes classification model demonstrates a balanced a-priori probability between Class 0 and Class 1, each approximating 50%, indicating an almost evenly distributed dataset. The conditional probabilities for each predictor highlight key insights into how various attributes differentiate between the classes. For instance, the credit_income_ratio is marginally higher for Class 1, suggesting a slight tendency towards higher credit utilization in this group. Similarly, the population attribute shows a lower mean in Class 1, which could reflect demographic trends relevant to the classification task. The education level data reveals a higher proportion of individuals with secondary education in Class 1, hinting at potential socioeconomic factors influencing class distinctions. Notably, the Region_Rating distribution indicates a significantly higher proportion of individuals in low-rated regions within Class 1, aligning with the observed trends in socio-economic status and geographical influence. Furthermore, employment-related attributes such as Years_Employed and Age show that individuals in Class 0 tend to be older and have longer employment histories, suggesting that stability through age and employment might correlate with the outcomes associated with Class 0. The detailed conditional probabilities for attributes like Typeofloan and Occupation_Type further enrich our understanding, showing a higher prevalence of cash loans in Class 1 and a greater proportion of low-skilled workers compared to Class 0. These nuances captured by the Naive Bayes model are invaluable for making informed predictions and provide a robust foundation for risk assessment, policy making, or resource allocation decisions based on the probabilistic understanding of influential factors.

Performance Evaluation:

Confusion matrix for Naïve Bayes Model shows that it is highly effective at identifying the negative class but at the cost of a large number of false negatives with 94.3% of Specificity. Sensitivity which correctly identified actual positives stands at 14.51% and Accuracy which corrected both negatives and positives at 21.71%.

The ROC curve appears to rise moderately steeply from the lower left towards the upper left of the graph before flattening out towards the upper right corner. This shape suggests a reasonable ability of the Naive Bayes model to distinguish between the two classes over a range of thresholds. (Figure 3). The AUC value stands at 64.3% which offers a reasonable classification performance for this problem.

<u>Selection of Model:</u>

For our business objective, reducing the number of false positives is crucial. Incorrectly identifying a defaulting customer as a non-defaulter can lead to financial loss and could damage bank reputation. Although sensitivity is notably lower than that of the Random Forest model and low accuracy, Naive Bayes model with its highest specificity, aligns well with the need to maintain customer trust and satisfaction by ensuring that customers are not turned away.

6. Interpretation in a business context, conclusions, and recommendations

Conclusions:

- **Identification of Risk characteristics:** The analysis concluded that key predictors of loan default include credit income ratio, education level, occupation type, regional rating, age, population density, the frequency of defaults within 60 days, credit inquiries, and years employed.

- **Demographic Insights**: Age and occupation significantly affect loan repayment, with older individuals and specific job sectors showing lower default rates, highlighting the impact of experience and stable employment on financial behavior.
- **Financial Indicators**: A higher credit-to-income ratio was identified as a critical risk marker, suggesting that borrowers with higher debt relative to their income are more likely to default, emphasizing the importance of this metric in risk assessment.
- **Geographic Variation**: The analysis also noted that default rates vary by region, which could be influenced by economic conditions and regional stability, pointing to the necessity of considering geographic factors in risk evaluation.

Recommendations:

Detailed Explanation of Recommendations:

- **Credit Policy Adjustment**: Financial institutions are advised to refine their credit scoring models by placing greater emphasis on key risk factors like the credit-income ratio, occupation, and age. This strategic adjustment will enhance the evaluation of risk associated with new loan applications, leading to more informed lending decisions. Implementing cautionary profiles, especially for high-risk occupations, can secure loans with full collateral.
- **Targeted Financial Products**: Based on the analysis identifying higher-risk groups, it is recommended to develop financial products specifically tailored for these segments. This could include options such as lower credit limits or secured credit products, which mitigate the lender's risk while accommodating the financial needs of higher-risk borrowers.
- **Enhanced Data Collection**: Improving the accuracy of predictive models can be achieved by collecting additional data on applicants’ financial obligations, past credit history, and the economic conditions of the areas they live in. This broader data collection will provide a more comprehensive basis for assessing loan applications.
- **Preventive Measures for High-Risk Regions**: For regions identified with a higher risk of defaults, implementing rigorous follow-up and recovery processes is crucial. This proactive approach will help in minimizing potential financial losses from defaults.
- **Continuous Monitoring and Updating of Models:** Regular updates to predictive models are essential, incorporating new data and feedback from actual loan outcomes.

This ongoing refinement helps capture shifts in economic conditions and borrower behavior, keeping the risk assessment models current and effective.

- **Educational Programs:** Offering financial education programs to applicants in high-risk categories can empower them to better manage their finances and understand the implications of credit. This not only aids the borrower in making informed financial decisions but also reduces the risk of defaults.

Appendix:

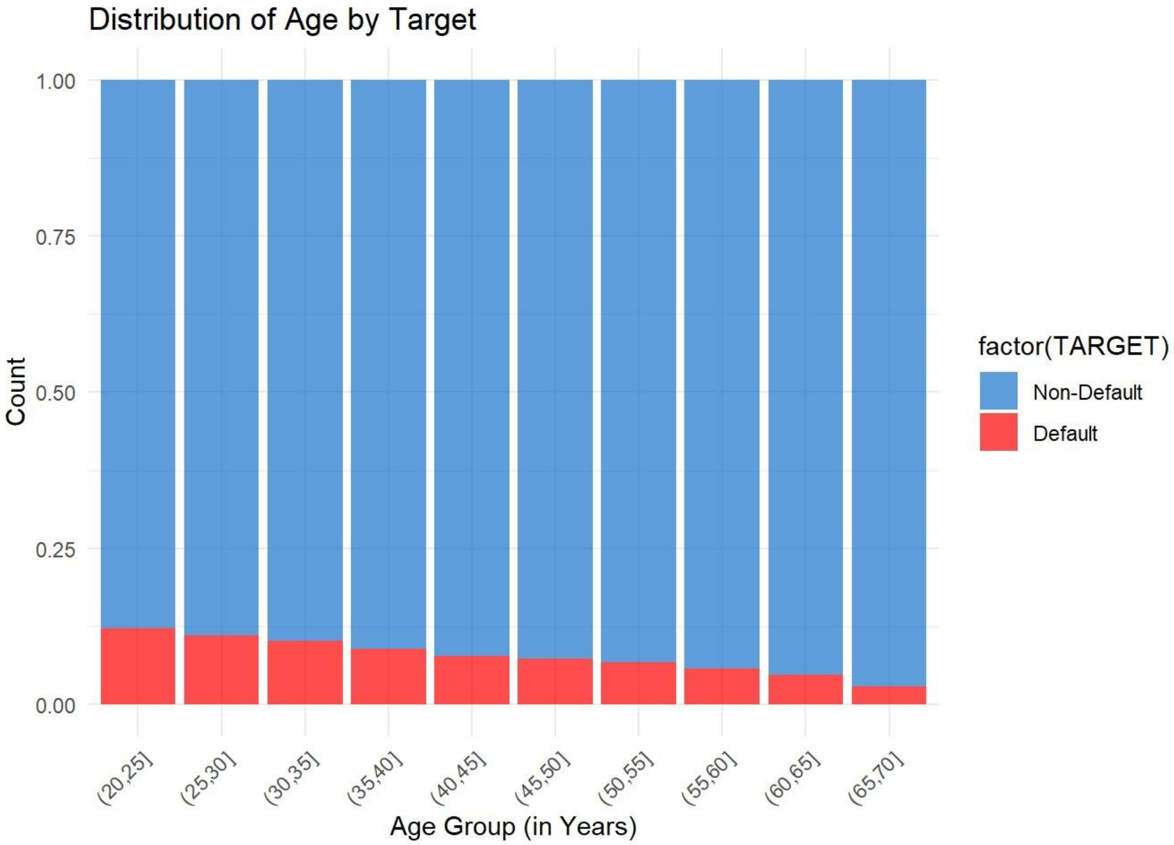
Sk_Id_Curr	Customer ID of loan
Target	Target variable (1 - customer with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan, 0 - all other cases)
Name_Contract_Type	Identification if loan is cash or revolving
Code_Gender	Gender of the customer
Flag_Own_Car	Customer owns a car (1 for Yes, 0 for No)
Flag_Own_Realty	Customer owns a house or flat (1 for Yes, 0 for No)
Cnt_Children	Number of children the customer has
Amt_Income_Total	Income of the customer
Amt_Credit	Credit amount of the loan
Amt_Annuity	Loan annuity
Amt_Goods_Price	For consumer loans it is the price of the goods for which the loan is given
Name_Type_Suite	Who was accompanying customer when he was applying for the loan
Name_Income_Type	Customers income type (businessman, working, maternity leave,…)
Name_Education_Type	Level of highest education the customer achieved
Name_Family_Status	Family status of the customer
Name_Housing_Type	What is the housing situation of the customer (renting, living with parents, ...)
Region_Population_Relative	Normalized population of region where customer lives (higher number means the customer lives in more populated region)
Days_Birth	Customer's age in days at the time of application

Days_Employed	How many days before the application the person started current employment
Days_Registration	How many days before the application did customer change his registration
Days_Id_Publish	How many days before the application did customer change the identity document with which he applied for the loan
Own_Car_Age	Age of customer's car
Flag_Mobil	Did customer provide mobile phone (1 for Yes, 0 for No)
Flag_Emp_Phone	Did customer provide work phone (1 for Yes, 0 for No)
Flag_Work_Phone	Did customer provide home phone (1 for Yes, 0 for No)
Flag_Cont_Mobile	Was mobile phone reachable (1 for Yes, 0 for No)
Flag_Phone	Did customer provide home phone (1 for Yes, 0 for No)
Flag_Email	Did customer provide email (1=YES, 0=NO)
Occupation_Type	What kind of occupation does the customer have
Cnt_Fam_Members	How many family members does customer have
Region_Rating_Client	Our rating of the region where customer lives (1,2,3)
Region_Rating_Client_W_City	Our rating of the region where customer lives with taking city into account (1,2,3)
Weekday_Appr_Process_Start	On which day of the week did the customer apply for the loan
Hour_Appr_Process_Start	Approximately at what hour did the customer apply for the loan
Reg_Region_Not_Live_Region	Flag if customer's permanent address does not match contact address (1 for different, 0 for same, at region level)
Reg_Region_Not_Work_Region	Flag if customer's permanent address does not match work address (1 for different, 0 for same, at region level)
Live_Region_Not_Work_Region	Flag if customer's contact address does not match work address (1 for different, 0 for same, at region level)
Reg_City_Not_Live_City	Flag if customer's permanent address does not match contact address (1 for different, 0 for same, at city level)
Reg_City_Not_Work_City	Flag if customer's permanent address does not match work address (1 for different, 0 for same, at city level)
Live_City_Not_Work_City	Flag if customer's contact address does not match work address (1 for different, 0 for same, at city level)
Organization_Type	Type of organization where customer works
Ext_Source	Normalized score from external data source
Apartments_Avg	Average Apartment size

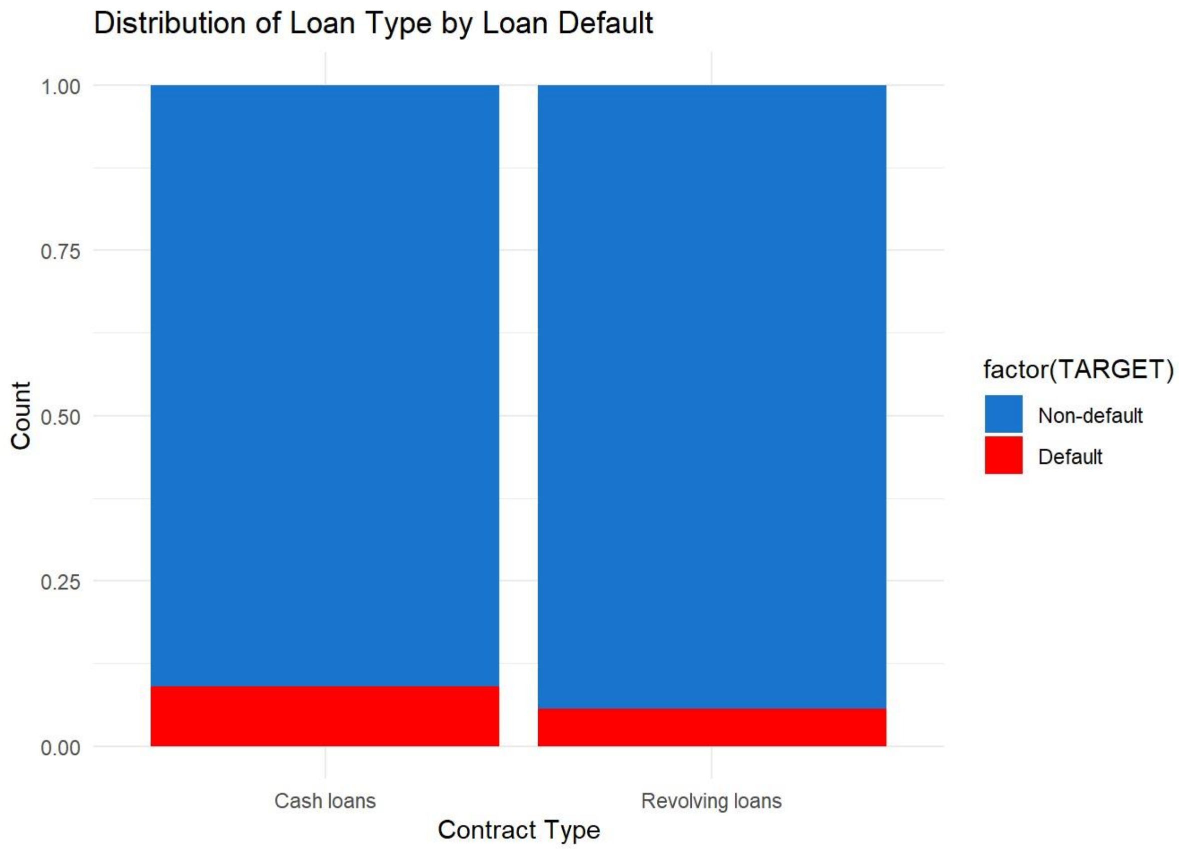
Basementarea_Avg	Average Basement area
Years_Beginexpluatation_Avg	Average Years Since Exploitation Began
Years_Build_Avg	Average Years Since Building Was Built
Commonarea_Avg	Average Common Area
Elevators_Avg	Average Number of Elevators
Entrances_Avg	Average Number of Entrances
Floorsmax_Avg	Average Maximum Floors
Floorsmin_Avg	Average Minimum Floors
Landarea_Avg	Average Land Area
Livingapartments_Avg	Average Living Apartments
Livingarea_Avg	Average Living Area
Nonlivingapartments_Avg	Average Non-Living Apartments
Nonlivingarea_Avg	Average Non-Living Area
Apartments_Mode	Mode Apartment Size
Basementarea_Mode	Mode Basement Area
Years_Beginexpluatation_Mode	Mode Years Since Exploitation Began
Years_Build_Mode	Mode Years Since Building Was Built
Commonarea_Mode	Mode Common Area
Elevators_Mode	Mode Number of Elevators
Entrances_Mode	Mode Number of Entrances
Floorsmax_Mode	Mode Maximum Floors
Floorsmin_Mode	Mode Minimum Floors
Landarea_Mode	Mode Land Area
Livingapartments_Mode	Mode Living Apartments
Livingarea_Mode	Mode Living Area
Nonlivingapartments_Mode	Mode Non-Living Apartments
Nonlivingarea_Mode	Mode Non-Living Area
Apartments_Medi	Median Apartment Size
Basementarea_Medi	Median Basement Area
Years_Beginexpluatation_Medi	Median Years Since Exploitation Began
Years_Build_Medi	Median Years Since Building Was Built
Commonarea_Medi	Median Common Area
Elevators_Medi	Median Number of Elevators
Entrances_Medi	Median Number of Entrances

Floorsmax_Medi	Median Maximum Floors
Floorsmin_Medi	Median Minimum Floors
Landarea_Medi	Median Land Area
Livingapartments_Medi	Median Living Apartments
Livingarea_Medi	Median Living Area
Nonlivingapartments_Medi	Median Non-Living Apartments
Nonlivingarea_Medi	Median Non-Living Area
Fondkapremont_Mode	Mode Fondkapremont
Housetype_Mode	Mode House Type
Totalarea_Mode	Mode Total Area
Wallsmaterial_Mode	Mode Walls Material
Emergencystate_Mode	Mode Emergency State
Obs_30_Cnt_Social_Circle	How many observation of customer's social surroundings with observable 30 DPD (days past due) default
Def_30_Cnt_Social_Circle	How many observation of customer's social surroundings defaulted on 30 DPD (days past due)
Obs_60_Cnt_Social_Circle	How many observation of customer's social surroundings with observable 60 DPD (days past due) default
Def_60_Cnt_Social_Circle	How many observation of customer's social surroundings defaulted on 60 (days past due) DPD
Days_Last_Phone_Change	How many days before application did customer change phone
Flag_Document	Did customer provide documents
Amt_Req_Credit_Bureau_Hour	Number of enquiries to Credit Bureau about the customer one hour before application
Amt_Req_Credit_Bureau_Day	Number of enquiries to Credit Bureau about the customer one day before application (excluding one hour before application)
Amt_Req_Credit_Bureau_Week	Number of enquiries to Credit Bureau about the customer one week before application (excluding one day before application)
Amt_Req_Credit_Bureau_Mon	Number of enquiries to Credit Bureau about the customer one month before application (excluding one week before application)
Amt_Req_Credit_Bureau_Qrt	Number of enquiries to Credit Bureau about the customer 3 month before application (excluding one month before application)
Amt_Req_Credit_Bureau_Year	Number of enquiries to Credit Bureau about the customer one day year (excluding last 3 months before application)

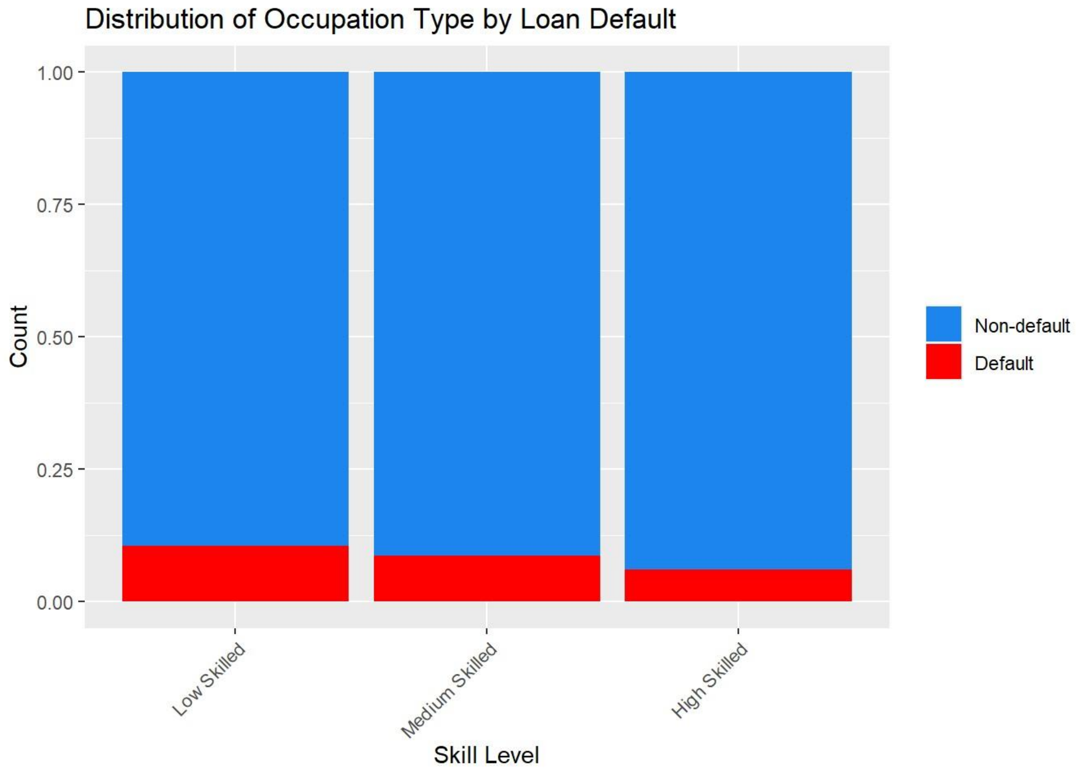
Graph1: Age



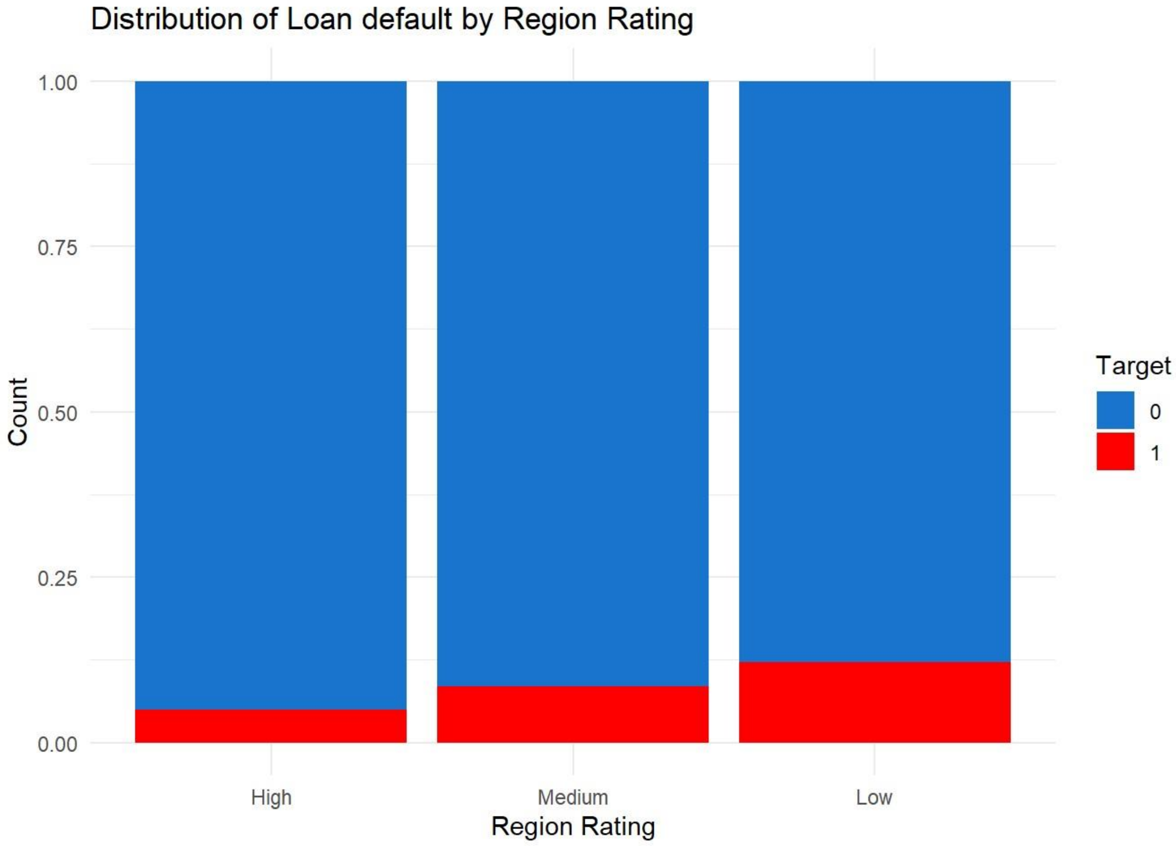
Graph2: Type of Loans



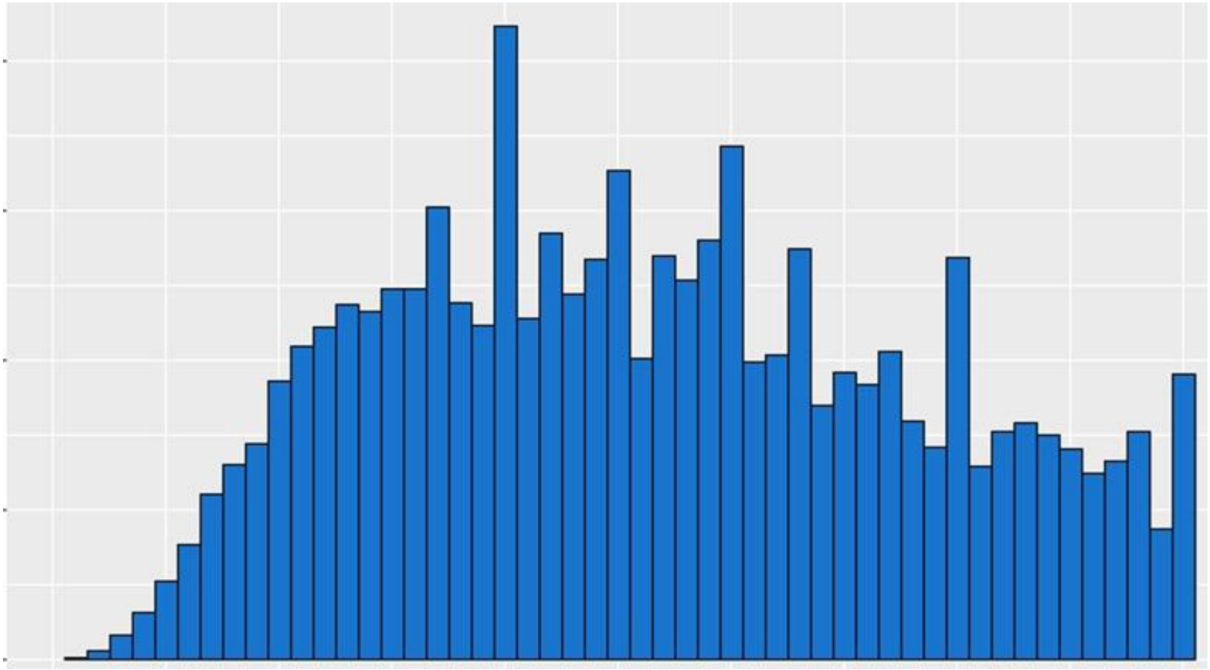
Graph3: Skill Levels



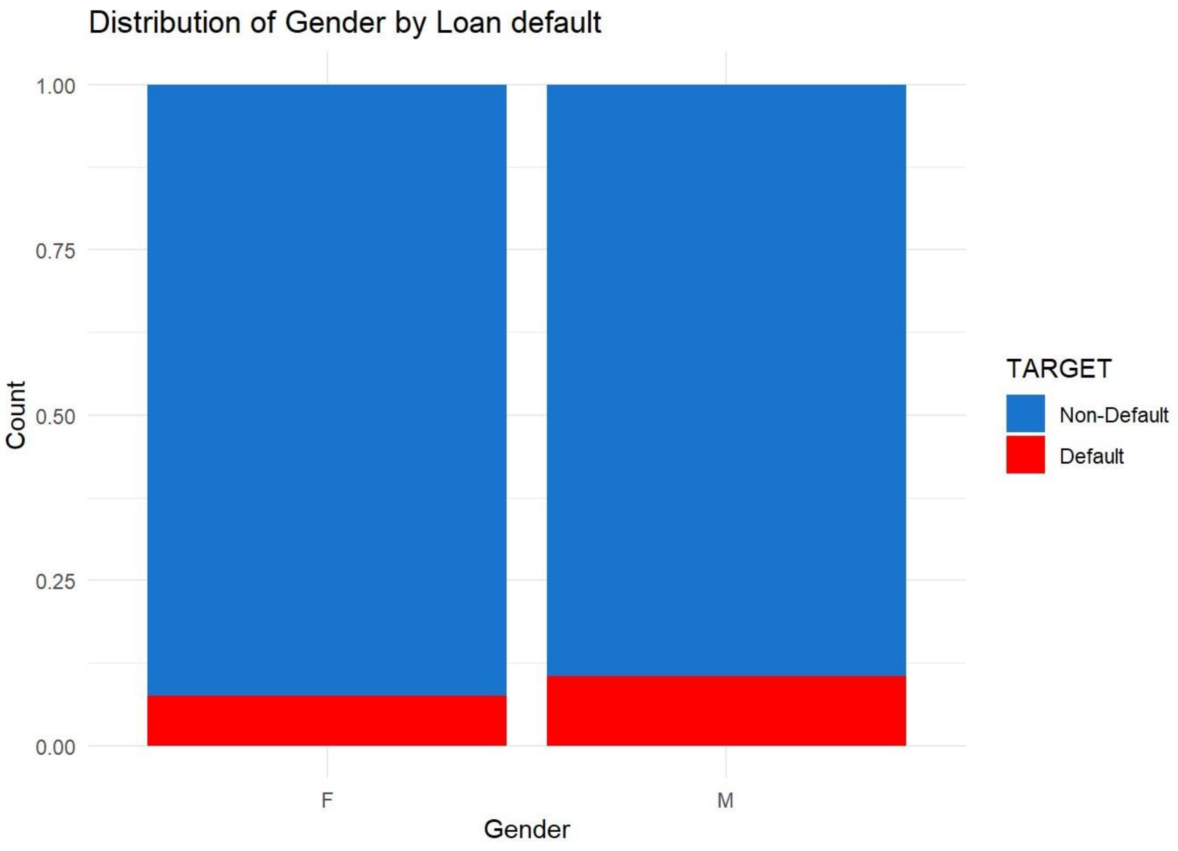
Graph4: Region Rating Client



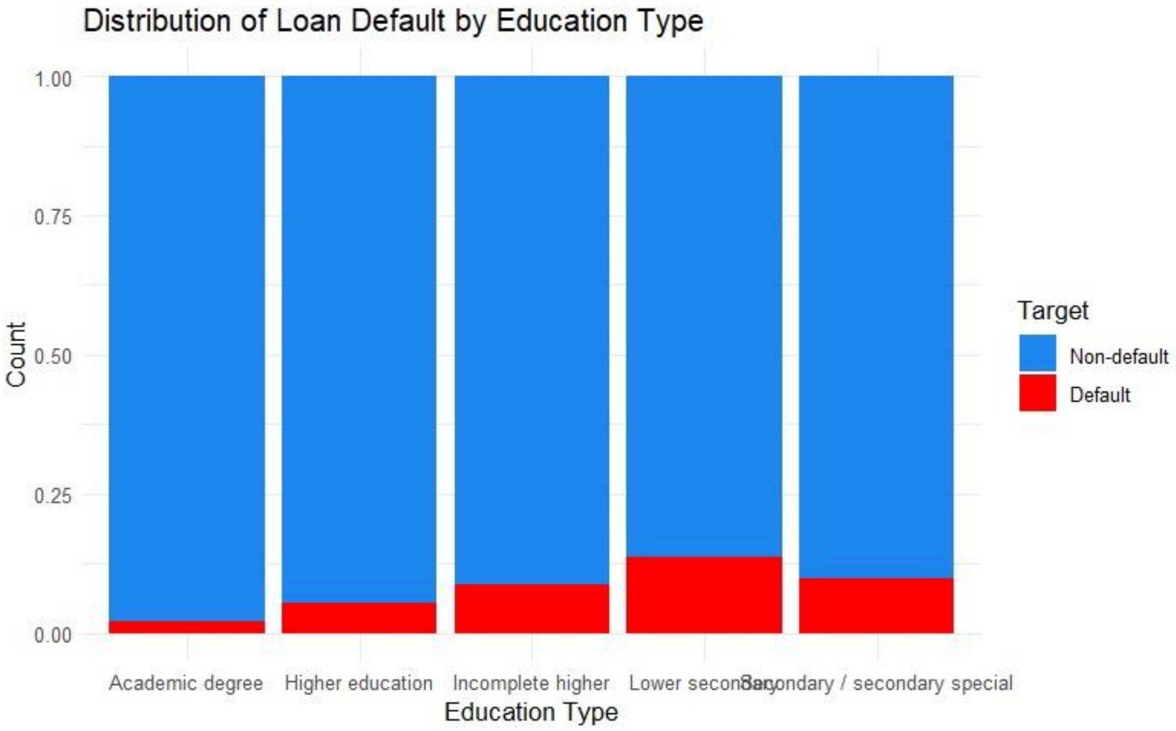
Graph 5: Credit Income Ratio



Graph 6: Gender



Graph 7: Education level



Graph 8: Population

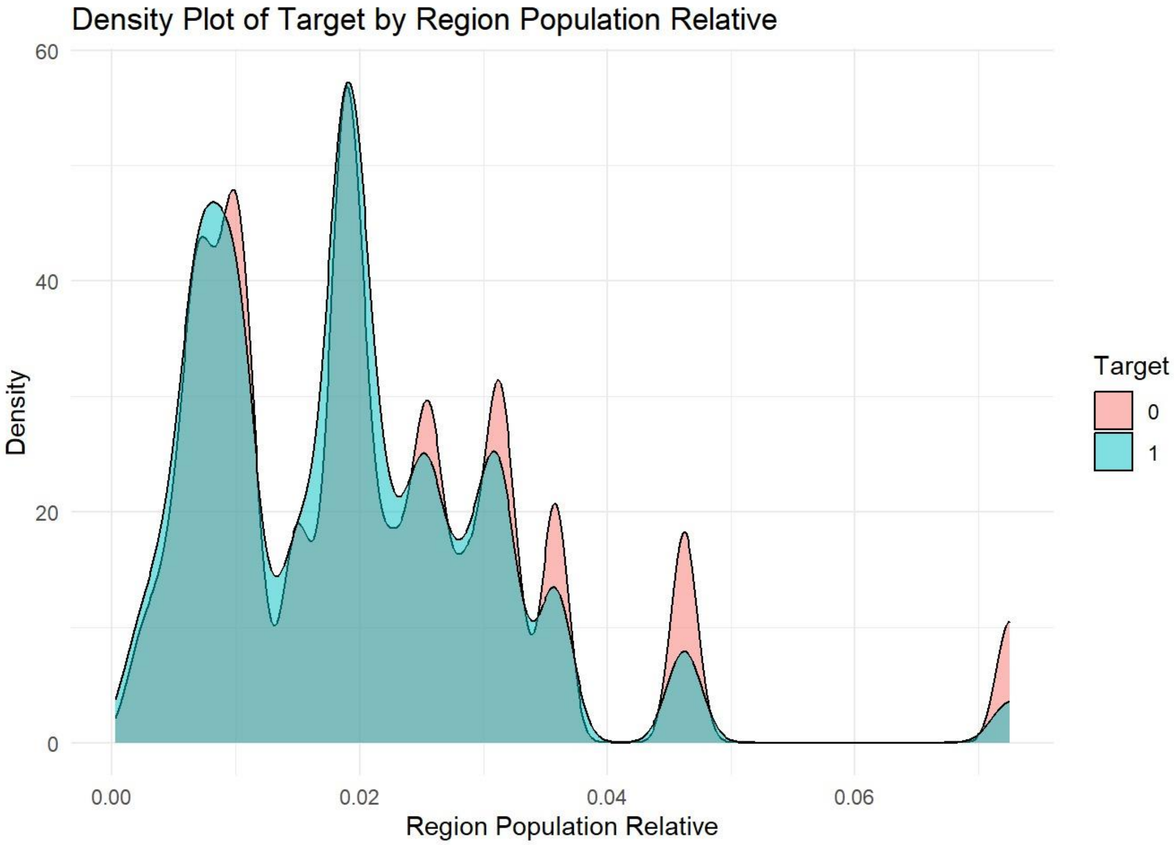


Figure 1:

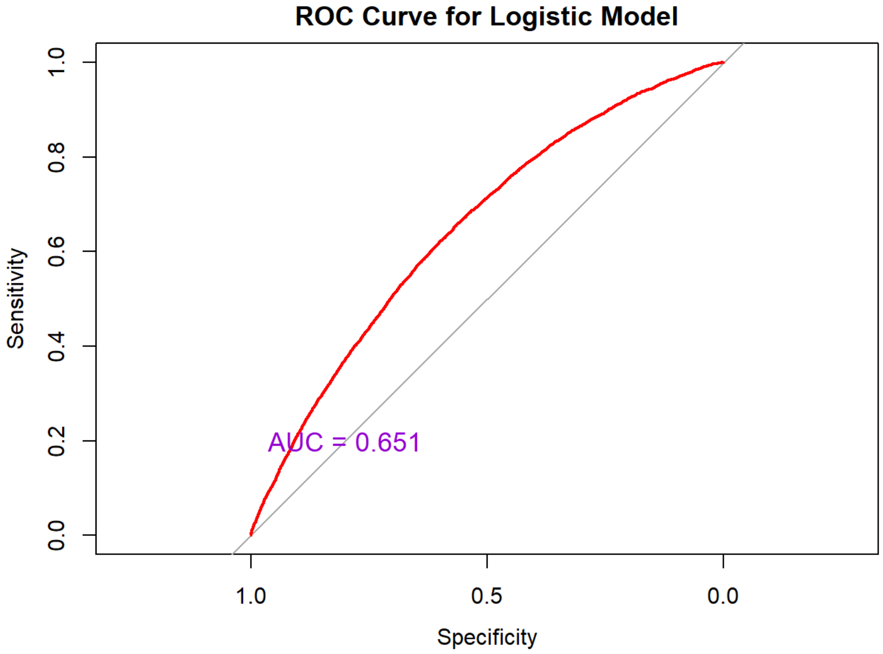


Figure 2:

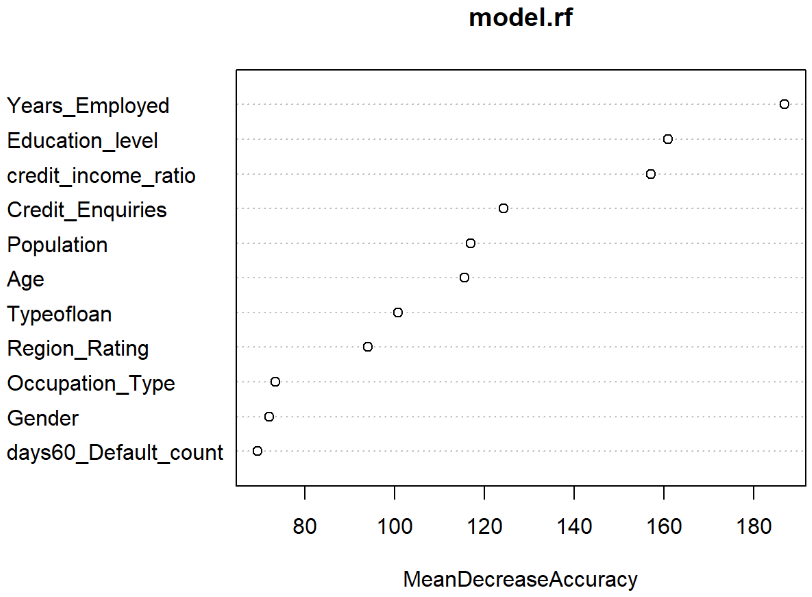
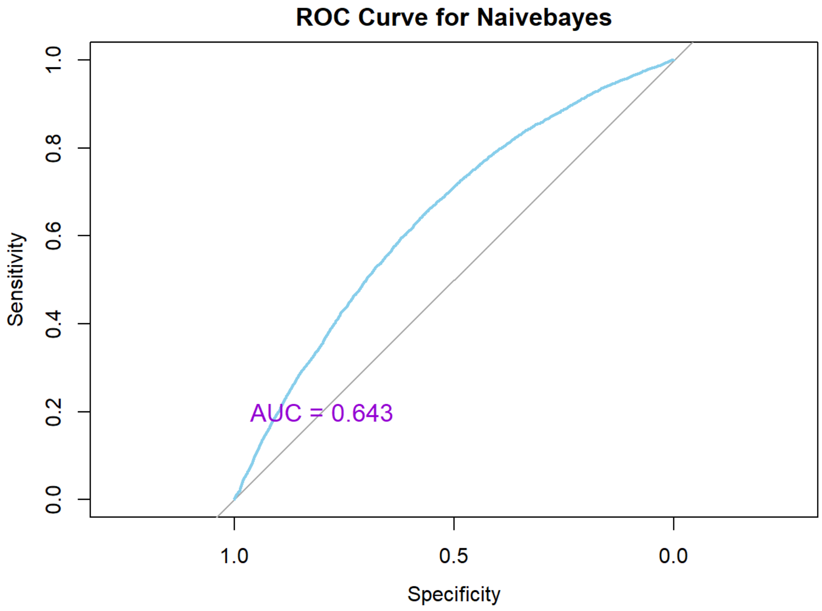

model.rf
Years_Employed
Education_level
credit_income_ratio
Credit_Enquiries
Population
Age
Typeofloan
Region_Rating
Occupation_Type
Gender
days60_Default_count
80
100
120
140
160
180
MeanDecreaseAccuracy

Figure 3:


ROC Curve for Naivebayes
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.643
1.0
0.5
0.0
Specificity

References:

1. Tariq, H. I., Sohail, A., Aslam, U., & Batcha, N. K. (2019). Loan default prediction model using sample, explore, modify, model, and assess (SEMMA). Journal of Computational and Theoretical Nanoscience, 16(8), 3489-3503.

2. Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012042). IOP Publishing.

3. Qiu, Z., Li, Y., Ni, P., & Li, G. (2019, December). Credit risk scoring analysis based on machine learning models. In 2019 6th International Conference on Information Science and Control Engineering (ICISCE) (pp. 220-224). IEEE.

4. Tshivhidzo, R. (2022). Comparative study of Machine learning techniques for loan fraud prediction (Doctoral dissertation, University of the Witwatersrand, Johannesburg).

5. Matthys, C. PREDICTING AN APPLICANT'S CAPABILITY OF REPAYING A BANK LOAN.