

PREDICTIVE ANALYSIS
MACHINE LEARNING
Project

Kavya Sharawat

Submitted to:

Dr. Kriti Priya Gupta

Introduction

Lending loans is the cornerstone of banking, facilitating economic growth by providing individuals and businesses with the capital they need to pursue opportunities and achieve their goals. However, this practice is inherently risky as borrowers may default on their loans, leading to financial losses for the lending institution. Defaulting occurs when a borrower fails to repay the loan according to the terms agreed upon in the loan contract, which can be due to various reasons such as financial instability, unforeseen circumstances, or deliberate non-payment. The consequences of loan defaulting ripple throughout the financial system, impacting not only the lender's bottom line but also potentially destabilizing the broader economy.

To mitigate this risk associated with loan defaulting, banks are increasingly turning to machine learning (ML) techniques to develop predictive models that can identify potential defaulters before extending credit. By leveraging consumer information and compiling a analytical dataset containing required descriptive information banks aim to build robust ML algorithms capable of accurately classifying whether a new borrower is likely to default on their loan.

Steps involved in prediction of Loan Defaulters:

1. **Data Collection:** Gather historical data on loan borrowers and compiling them into one dataset. The data collected for further predictions include the following variables:
 - **Loan Borrower ID:** Loan Borrower ID serves as a unique identifier for each borrower which helps in facilitating data management and analysis after predicting the customers loan defaulting probability.
 - **Gender:** Gender may impact factors such as income, employment status, and credit history, which are important predictors of loan default risk. ML models can use gender as a feature to identify patterns and correlations with default behaviour.
 - **Approved in Advance (non pre or pre):** Whether a loan is approved in advance or not may indicate the level of scrutiny applied during the approval process. Loans approved in advance may have lower default rates due to stricter eligibility criteria.
 - **Loan Type (Type 1 or Type 2 or Type 3):** Different loan types carry varying levels of risk. ML models can analyse the characteristics of each loan type and their historical default rates to predict future default probability.

- Purpose of Loan (p1 / p2 / p3 / p4): The purpose of the loan provides insights into the borrower's intentions and financial stability. Models can assess the risk associated with different loan purposes and their likelihood of default.
- Type of Credit worthiness (I1 or I2): Creditworthiness reflects a borrower's ability and willingness to repay debts. ML models can use creditworthiness as a feature to assess default risk and make predictions.
- Open Credit or not (opc / nopc): Whether a borrower has open credit lines may indicate their credit utilization and financial obligations. ML models can use this information to assess the borrower's capacity to take on additional debt and predict default probability.
- Business loan or commercial loan: The type of loan (business or commercial) has different risk profiles and default probabilities. ML models can differentiate between these loan types and tailor predictions accordingly.
- Loan Amount: The amount of the loan directly influences the borrower's repayment capacity and default risk.
- Interest Rate: Higher interest rates may indicate higher risk loans and potentially higher default rates.
- Property Value: Property value serves as collateral for secured loans and may influence default risk.
- Borrower's Income: Income levels are one of the strongest indicators of a borrower's ability to repay loans and predict loan default probability.
- Credit Type: Different credit types (e.g., revolving credit, instalment credit) may have different default probabilities. This variable can be used to analyse the characteristics of each credit type and their historical default rates to make predictions.
- Credit Score: Credit score reflects a borrower's creditworthiness and past credit behaviour, making it a strong predictor of loan default probability.
- Borrower's Age: Age can be indicative of financial maturity and stability.

- **Loan to Value Ratio:** Loan-to-value ratio measures the ratio of the loan amount to the appraised value of the collateral.
- **Borrower's Region:** Regional economic conditions may influence default rates. ML models can analyse regional data to identify patterns and correlations with default behaviour.
- **Security Type:** The type of security (e.g., real estate, stocks) may influence default risk.
- **Loan Status (0: Not approved; 1: Approved):** While loan status indicates whether a loan was approved or not, it is also served as a target variable for our models, allowing them to learn patterns and make predictions based on historical loan approval decisions and their outcomes.

Techniques Employed in Data Cleaning

Data cleaning steps and techniques contribute to enhancing the quality of the dataset by reducing outliers, imputing missing values, controlling skewness and kurtosis, and ensuring that the data is suitable for subsequent analyses such as exploratory data analysis and machine learning modelling. The important variables used in this process include loan amount, interest rate, borrower's income, loan to value ratio, property value, credit score, and loan approval status, among others. These variables provide valuable insights into borrower characteristics and loan attributes, helping to identify patterns and relationships that can inform loan default prediction models.

Steps involved in data checking and cleaning:

1. Conversion to Factors: The following categorical variables are converted to factors:

- Gender
- Approved in Advance (non pre or pre)
- Loan Type (Type 1 or Type 2 or Type 3)
- Purpose of Loan (p1 / p2 / p3 / p4)
- Type of Credit worthiness (I1 or I2)
- Open Credit or not (opc / nopc)
- Business loan or commercial loan
- Credit Type
- Borrower's Age
- Borrower's Region
- Security Type
- Loan Status (0: Not approved; 1: Approved)

This ensures that R recognizes them as categorical variables rather than numerical ones, facilitating categorical data analysis.

2. Checking for Missing Values: The `sum(is.na(.))` function is used to identify missing values in each column of the dataset. This step is crucial as missing values can impact the accuracy and reliability of subsequent analyses. Number of missing values in the following variables:

Approval in advance	19
Rate of Interest	718
Property Value	301
Age	4
Loan to Value Ratio (LTV)	301
Income	168

3. Imputation of Missing Values: Missing values are imputed using different techniques:

For variables within Approval in Advance and Credit Score **mode** imputation is used. It replaces the missing values with the most frequent value.

For Loan Amount, Interest Rate, Borrower's Income, and Loan to Value Ratio, **median** imputation (replacing missing values with the median) is used.

4. Skewness and Kurtosis Analysis: Skewness and kurtosis are analysed for numerical variables:

- Loan Amount
- Interest Rate
- Property Value
- Borrower's Income
- Credit Score
- Loan to Value Ratio

Skewness measures the asymmetry of the distribution, while kurtosis measures the peakiness or flatness of the distribution. Identifying and addressing skewness and kurtosis is important for ensuring that the data follows normal distribution assumptions. The skewness and kurtosis of variables before transformation:

Variables	Skewness	Kurtosis
Loan Amount	1.532701796	9.290111
Rate of Interest	0.353708200	3.918559
Property Value	4.169592057	41.252515
Income	12.255668646	265.651109
Credit Score	-0.005807775	1.829448
Loan to Value Ratio	-0.827689483	3.978403

5. Boxplot and Outlier Detection: Boxplots are used to visualize the distribution of numerical variables and identify potential outliers. Outliers, which are extreme values that deviate significantly from the rest of the data, can distort statistical analyses and machine learning models.

6. Outlier Removal: Outliers are removed from the dataset based on boxplot analysis. This is done for columns loan amount, property value, rate of interest, LTV, and income. Outliers are detected using Boxplot which uses the interquartile range (IQR) method, and observations containing outliers are removed from the dataset.

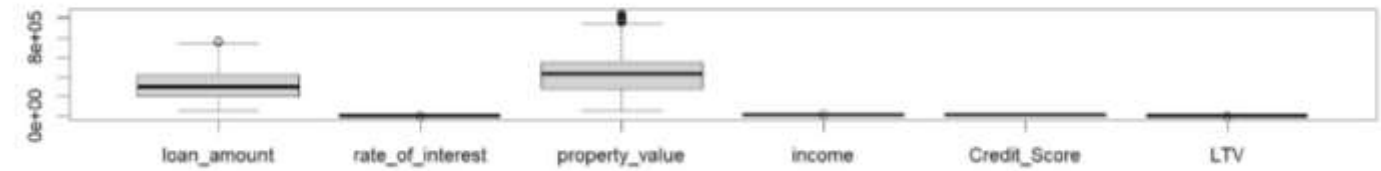
7. Further Imputation of Outliers: After removing outliers, missing values are imputed again for the cleaned dataset (L1). Outliers are detected using boxplot analysis for

columns loan amount, property value, rate of interest, LTV, and income. Missing values in these columns are imputed using median imputation.

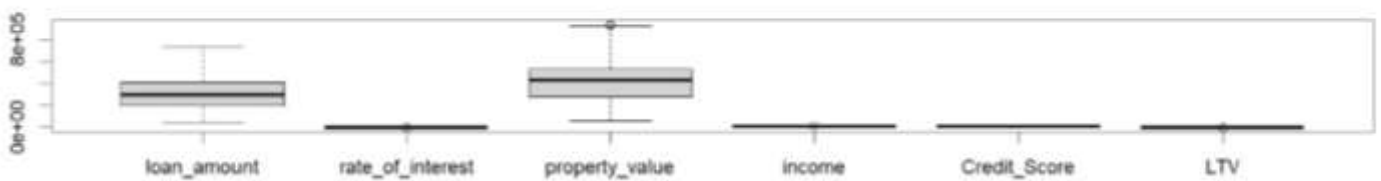
8. Transformation: Various transformations are applied to address skewness and kurtosis and achieve more symmetrical distributions for the numeric variables. Square root transformation, logarithmic transformation, and log10 transformation are explored for columns: loan amount, rate of interest, property value, income, and LTV.

Exploratory Data Analysis

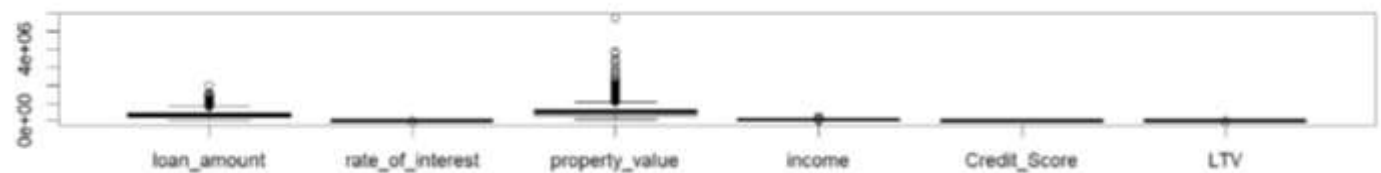
1. Before treating outliers



2. After using method 1. Imputing outliers.



3. After removing outliers and then inputting



Machine Learning Models

1. Logistic Regression (Model1):

Accuracy: 87.02%

Sensitivity: 52.31%

Specificity: 99.45%

Pos Pred Value: 97.14%

Neg Pred Value: 85.34%

Kappa: 60.76%

Balanced Accuracy: 75.88%

Summary: While achieving high specificity, the sensitivity of this model is relatively low, indicating that it might struggle with correctly identifying positive cases.

2. Naive Bayes (Model2):

Accuracy: 86.41%

Sensitivity: 67.69%

Specificity: 93.11%

Pos Pred Value: 77.88%

Neg Pred Value: 88.95%

Kappa: 63.47%

Balanced Accuracy: 80.40%

Summary: This model shows improved sensitivity compared to logistic regression while maintaining a relatively high specificity, making it better at correctly identifying positive cases.

3. Random Forest (Model3):

Accuracy: 88.44%

Sensitivity: 71.54%

Specificity: 94.49%

Pos Pred Value: 82.30%

Neg Pred Value: 90.26%

Kappa: 68.92%

Balanced Accuracy: 83.01%

Summary: Random forest performs well with higher sensitivity and specificity compared to the previous models, indicating its effectiveness in correctly identifying both positive and negative cases.

4. Decision Tree (Model4):

Accuracy: 88.44%

Sensitivity: 71.54%

Specificity: 94.49%

Pos Pred Value: 82.30%

Neg Pred Value: 90.26%

Kappa: 68.92%

Balanced Accuracy: 83.01%

Summary: The decision tree model performs similarly to the random forest.

Why did we not use linear regression method?

We are not using linear regression because the variables which are using is categorical in nature (defaulters or non-defaulters) and linear regression is used in continuous variables.

Evaluation

- **Accuracy Metrics:** Overall accuracy of each model to understand how well they perform in predicting the target variable (Status).

Model No.	Model	Accuracy Rate (%)
1.	Logistic Regression	87.02%
2.	Naive Bayes	86.41%
3.	Random Forest	88.44%
4.	Decision Tree	88.44%

- **Sensitivity and Specificity:** Evaluate the sensitivity (true positive rate) and specificity (true negative rate) of each model to assess their ability to correctly identify positive and negative instances, respectively.

Model No.	Model	Sensitivity Rate (%)	Specificity Rate (%)
1.	Logistic Regression	52.31%	99.45%
2.	Naive Bayes	67.69%	93.11%
3.	Random Forest	71.54%	94.49%
4.	Decision Tree	71.54%	94.49%

- **Balanced Accuracy:** It is taken out as the average of the true positive and true negative rates i.e. $(\text{sensitivity} + \text{specificity})/2$

Considering the balanced accuracy helps to ensure that the model performs well across all classes.

Model No.	Model	Balanced Accuracy
1.	Logistic Regression	75.88%
2.	Naive Bayes	80.40%
3.	Random Forest	83.01%
4.	Decision Tree	83.01%

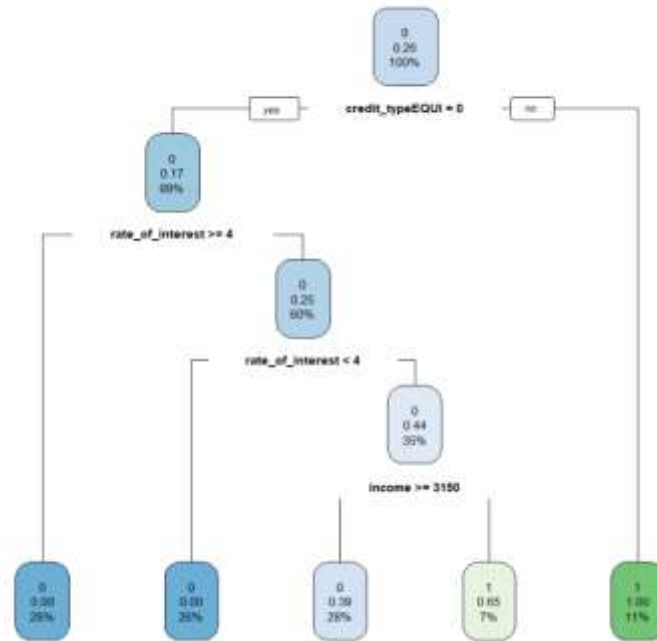
- **Comparison of Models:**

Comparing the performance of the different models across all metrics helps in identifying which model consistently performs better across various evaluation criteria.

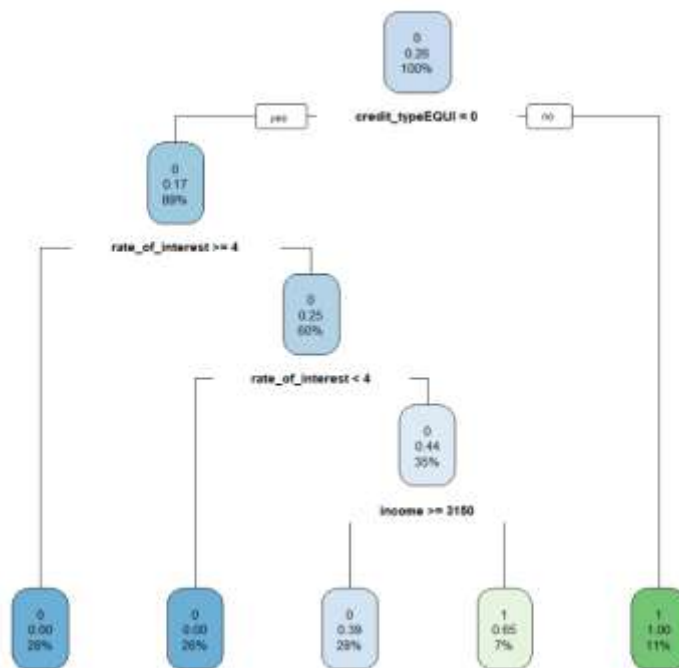
1. **Logistic Regression (Model1):** While achieving high specificity, the sensitivity of this model is relatively low, indicating that it might struggle with correctly identifying positive cases. The balanced accuracy is also lower compared to other models.
2. **Naive Bayes (Model2):** This model shows improved sensitivity compared to logistic regression while maintaining a relatively high specificity, making it better at correctly

identifying positive cases. The balanced accuracy is higher compared to logistic regression but slightly lower than the random forest.

3. Random Forest (Model3): Random Forest performs well with higher sensitivity and specificity compared to logistic regression and naive Bayes. It also has the highest balanced accuracy among the models evaluated, indicating its effectiveness in correctly identifying both positive and negative cases.

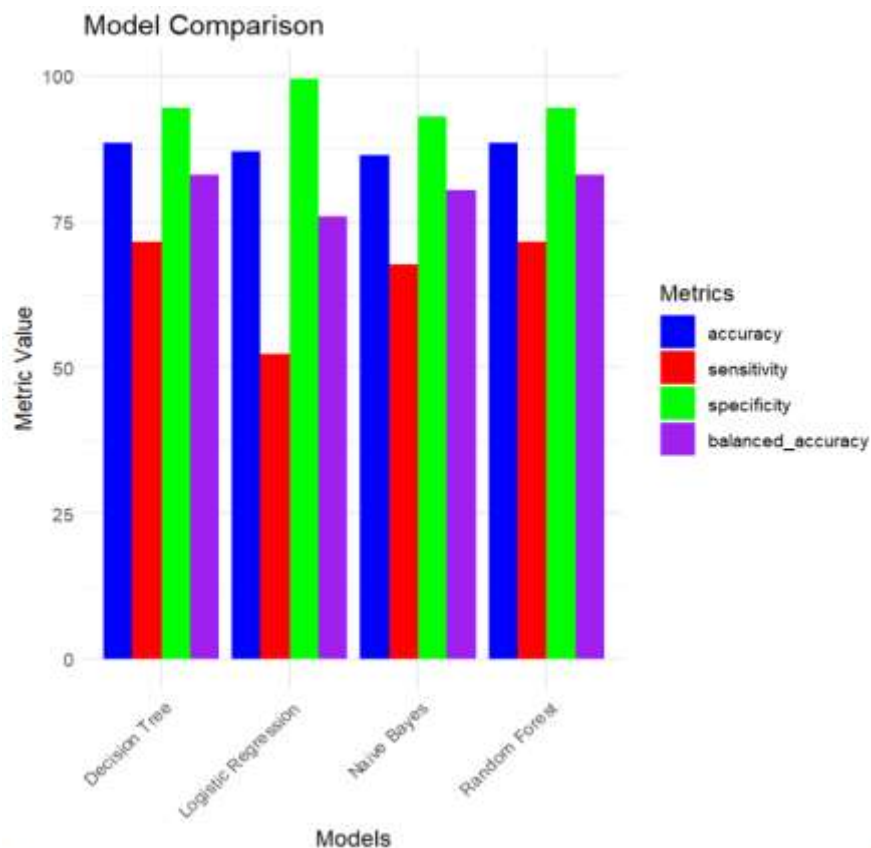


4. Decision Tree (Model4): The decision tree model performs similarly to the random forest. Its performance metrics are identical to the random forest.



Considering these evaluations, random forest model and decision tree model appears to be the best choice. Both the models showcase the same results. They achieve the highest accuracy among the models, indicating better overall performance in correctly identifying both positive and negative cases.

- Visualizations:



The visualisation showcases a bar plot comparing different evaluation metrics for each model. We can visually inspect which model performs best across the metrics that we are interested in.

- Model Interpretability:

Considering from the business contexts it is important to understand the major factors driving predictions for better decision-making.

In the context of predicting loan defaulters, the most important metric to consider when making decisions is specificity.

Reasons:

1. **Minimizing False Positives:** Specificity measures the proportion of actual negative cases (non-defaulters) that are correctly identified as negative by the model. In the context of loan default prediction, high specificity means the model can accurately identify borrowers who are not likely to default on their loans. Minimizing false positives is critical because falsely identifying a non-defaulter as a defaulter could lead to denying loans or imposing unfavorable terms on creditworthy borrowers, potentially resulting in lost revenue for the bank and negative customer experiences.

2. **Balancing Risk and Opportunity:** Banks aim to maximize their revenue while minimizing risk. High specificity helps achieve this balance by accurately identifying low-risk borrowers who are likely to repay their loans as agreed. By focusing on specificity, banks can confidently extend loans to creditworthy borrowers, thereby maximizing revenue opportunities, while simultaneously reducing the risk of default and financial losses.
3. **Enhancing Trust and Reputation:** Building a reputation for responsible lending practices is essential for banks to attract customers and maintain trust in the financial market. High specificity ensures that loan decisions are based on reliable predictions, enhancing the bank's reputation for sound risk management practices and ethical lending standards.

While sensitivity and precision are also important metrics, prioritizing specificity ensures that banks can make informed lending decisions that optimize revenue generation while effectively managing the risk of loan defaults.

- **Recommendations:**

Random forest models are known for their robustness to overfitting and ability to handle large datasets with high dimensionality, making them suitable for our dataset about 2900 data variables and a wide range of classification tasks.

It is also a simpler model which is as effective as the decision tree.

Contribution

1. Krrish Manchanda

Data Cleaning and major description regarding modelling.

2. Manya Agarwal

Research regarding which model to use and data cleaning.

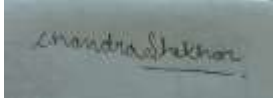



3. Kavya Sharawat

Documentation of the word document and applying various MS Models.

4. Chandra Shekhar Shukla

Documentation of the word document and applying various MS Models.

Signature:

			
Chandra Shekhar	Krrish Manchanda	Kavya Sharawat	Manya Agarwal