Machine Learning-Based Prediction of Loan Default

Barjinder Singh

University of Oklahoma

DSA/ISE 5103

Dr. Nicholson

12/12/2024

Executive summary

Problem Statement: This study aims to predict loan defaults in a dataset with a binary target variable (0 for non-default and 1 for default). More specifically, the ultimate goal is to build a model that would predict accurately whether a borrower would default on the loan based on three categories – demographic information, financial history, and loan-related variables. This study majorly uses supervised learning methods like regressions, neural networks, XGBoost, random forest, and support vector machines.

Major Concerns/Assumptions:

It is pertinent to point out the crucial features of the dataset that led to major decision-making in this study.

1. **Imbalanced Data**: The fact that the target variable is a binary categorical variable (loan default being affirmative or negative), and in reality, the number of loan defaults is less than the number of normal loans (without default) which our dataset nicely captures, could lead to biased predictions. In the dataset, 88% of the Loan IDs were non-defaults and 12% were default cases.
2. **Feature Selection**: The selection of the most relevant predictors of loan default is crucial to avoiding issues in modeling like overfitting. One major concern was whether to keep the interaction terms. Although it made logical sense to have them included in our modeling process, it diminished model performance.
3. **Evaluation Metrics**: Traditional accuracy might not be sufficient due to class imbalance, so metrics such as Accuracy, Sensitivity, Specificity, and ROC-AUC will be used for evaluation. This will provide the necessary information for model selection.
4. **Hygiene Issues**: These are common issues that need to be properly dealt with. Some of them are missing values, outliers, and duplicated values.

Summary of Findings:

The findings of the study incorporate results from key steps.

1. **Data Preprocessing**: After handling missing and duplicated values and encoding categorical variables, the data was ready for modeling. The imbalanced nature of the dataset necessitated the usage of techniques that dealt with class imbalance issues. Although oversampling the minority class gave better results than SMOTE in terms of balancing the dataset, SMOTE increased the AUC thereby leading to SMOTE being selected as the most ideal technique.
2. **Feature Importance**: Certain features, such as credit score, loan amount, and income, were found to be more significant predictors of loan default. Interaction terms were created but did not contribute to the overall model performance.
3. **Data Split**: The data was split into 80% for training and 20% for testing the model.
4. **Model Performance**: In this study, the selection of model was not of paramount difficulty because the random forest outperformed the other models (logistic regression, Neural Networks, Support Vector Machine) with an Area Under Curve (AUC) of 0.8085 followed by the XGBoost model with an AUC of 0.7911.

Recommendation:

Based on the findings of the analysis, it is hereby recommended to prioritize income verification and setting minimum income thresholds, as income is the strongest predictor of loan default. Credit scores should be used to screen borrowers, with stricter terms for those with lower scores. Loan amounts should be capped relative to income and credit scores to ensure affordability. For the high-risk age group (21-30), financial literacy programs and alternative underwriting strategies, such as requiring guarantors, or certain guarantor waiver programs (see the company Leap) can reduce default risks. Additionally, implementing data-driven predictive models (for instance, random forest model) to flag high-risk applicants and updating these models regularly will enhance risk management. Preemptive measures like flexible repayment options and financial counseling can further mitigate potential defaults.

Problem background

Problem description, context, and background:

The prediction of loan defaults poses a critical challenge for financial institutions, as it has a direct impact on profitability and financial stability. Appropriate estimates of loan default made by financial institutions help to hedge the occurrence of such potential risks in a bid to avoid an adverse financial situation. New data points to the delinquency rate on consumer loans at all commercial banks posting at 2.7% during the third quarter of 2024 (Federal Reserve Bank, 2024). However, this number jumps to as high as 10.7% in the first quarter of 2024 the highest level in 12 years among younger borrowers, those aged between 21 and 30 (Associated Press News, 2024).

Key among the significant determinants are income, loan amount, and credit score which would likely determine if the person will default. For instance, loan default rates are nearly three times higher for the bottom versus the top income quartile. Besides, credit card balances went up to over $6,000 on average, with interest rates skyrocketing to 22% (Wall Street Journal, 2024). These statistics indeed confirm that the emergence of potent predictive models, in light of borrowers' demographics and financial history, can very well predict and manage default risks.

The influx of lending activities due to a rise in global demand for commodities and services, in general, makes the identification of the most at-risk borrowers one important measure to mitigate losses and lend optimally. However, these conventional loan approval processes have been dominated by static rules, which do not at most times capture the intricate risk pattern in borrower behavior. Recently, advances in data analytics and machine learning have allowed for a more nuanced, data-driven approach using borrower demographics, financial history, and loan-specific features to predict defaults better. Taking forward these relationships, this study develops a predictive model that will assist the lender in determining potential defaults to deal with other related challenges, including data quality, class imbalance, and interpretability.

Dataset Description:

1. **Dataset Overview:**

Dataset URL: - https://www.kaggle.com/datasets/nikhil1e9/loan-default

The dataset consists of **255,347 rows**, each representing a loan application. The attributes provide details about borrowers and loan characteristics.

Variable description:

The variable description is mentioned in the Appendix.

2. **Data Attributes:**

The dataset contains a mix of variables. Some of the numeric variables are Income, Loan amount, Credit score, and DTI ratio. The binary variables capture important attributes - whether a customer has a mortgage (yes/no), has dependents (yes/no), and has a cosigner (yes/no). Among the categorical variables are

Education level (Bachelor's, High School, Master's, PhD), Employment type (Full-time, Part-time, Self-employed, unemployed), marital status (Single, Married, Divorced), and loan purpose (Auto, Business, Home, Education, Other). The target variable is the default indicator which is a binary variable (0/1). The loan default is 0 when a customer has not defaulted and 1 when the customer has defaulted.

Exploratory Data Analysis:

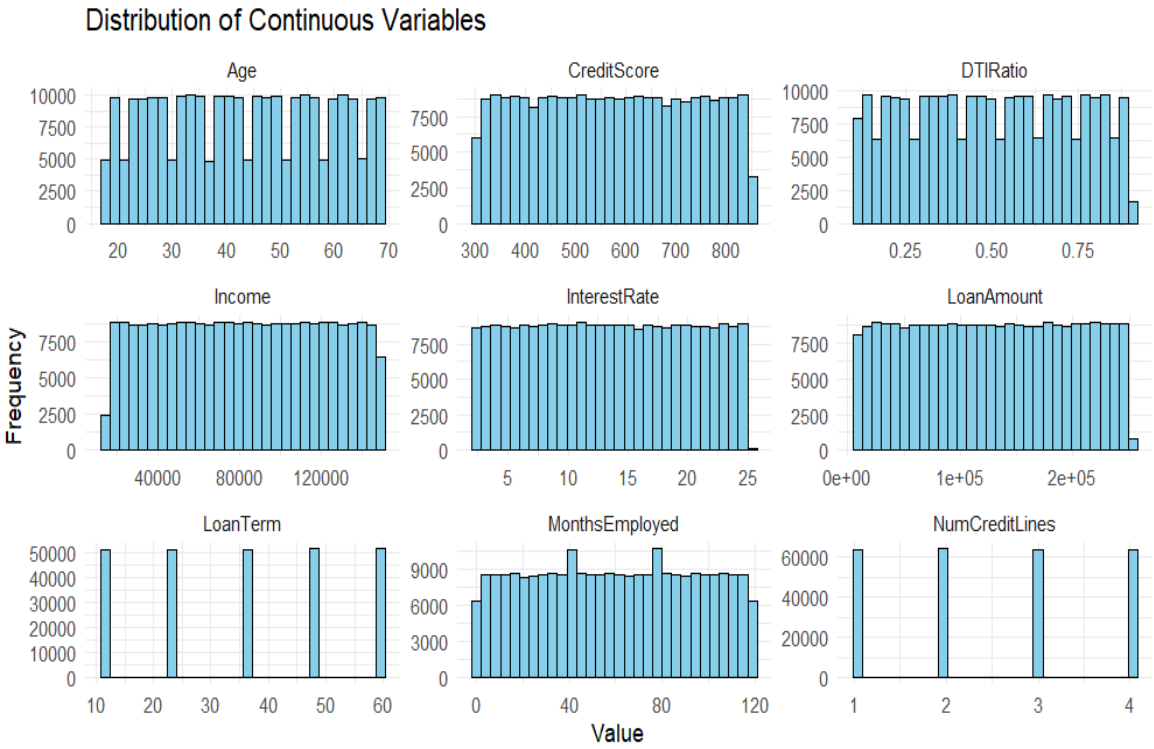
1. **Visualizations**

The visualizations are done to see initial trends and reveal patterns in the dataset.

a. Distribution of all Numeric Variables

First, the distribution of all numeric variables is checked in this visualization. This is a crucial step for data processing in statistics. From Figure 1, It is noticed that several variables such as Age, credit score, DTI Ratio, Income, Interest rate and loan amount appear to have a uniform distribution with very similar frequencies across bins.

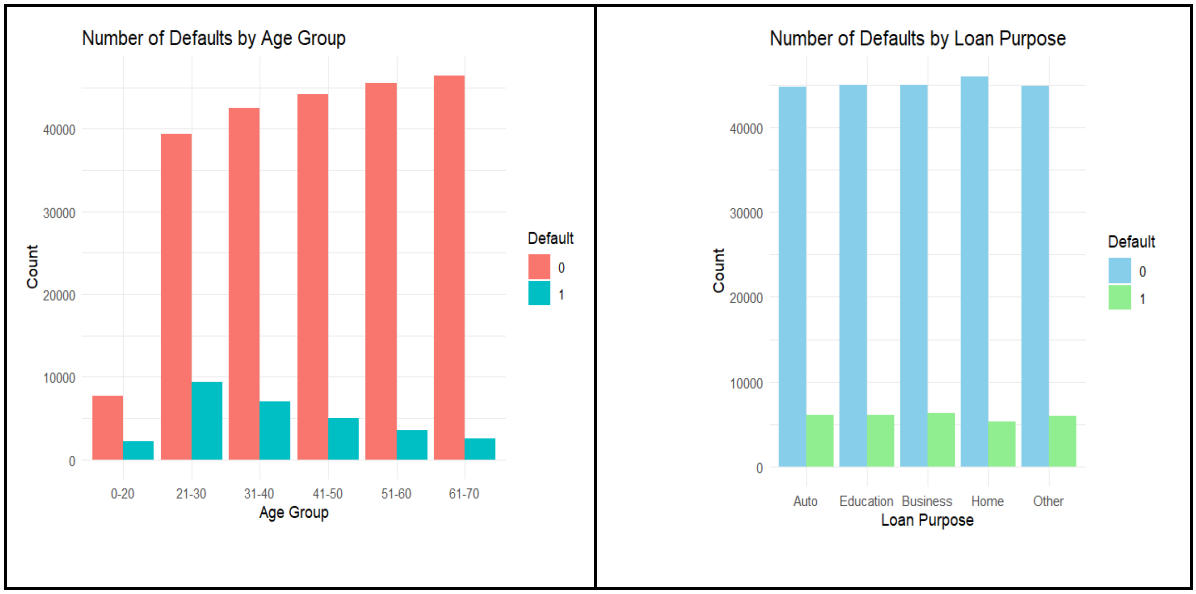
Figure 1: Distribution of Continuous Variables



b. Key Demographics

In many datasets, Age is usually treated as a continuous variable as it can be a decimal figure, for instance, 25.5 years, 30.75 years, etc. However, age is treated as a discrete variable in many large datasets as it adds value to interpretation if it is treated as a classifier. As a part of demographic analysis, group the Number of Defaults by age. The initial trends show that the 21-30 age group is most likely to default. The 0-20 age group can be treated as redundant as they form a very small part of the dataset. In the second panel, Home loans default less than their other counterparts.

Figure 2: No. of defaults by Age Group and Loan Purpose



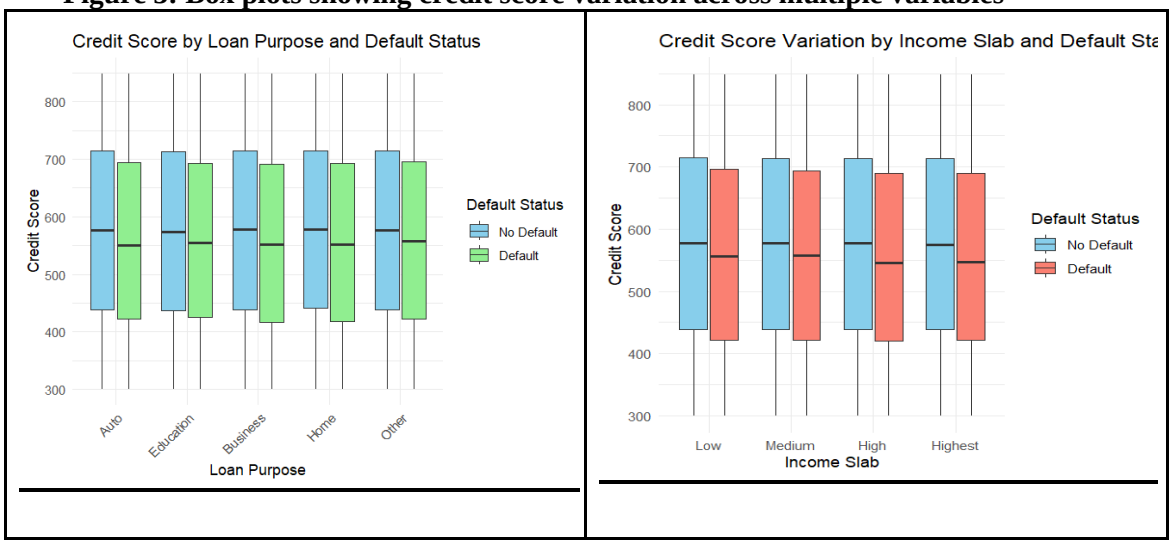
Note: In this figure, the no. of defaults in each age group were shown. The default variable is a binary variable where 0 is for "no defaults" and 1 indicates "default".

c. Important Insights:

i. Figure 3:

In Figure 3, boxplots are plotted showing the relationship between credit score and Loan purpose. It is observed that the loans that are in the default category have lower credit scores than the non-default category. When seen from the Loan Purpose, the differences are almost inconsequential except for some variation in the Home and Business categories. In the second panel, income slabs are again used (divide income into equally spaced categories into Low, Medium, High, and Highest) and see the credit score variation among default and non-default categories. It is observed that high-income individuals have a higher credit score variation in the default vs non-default category.

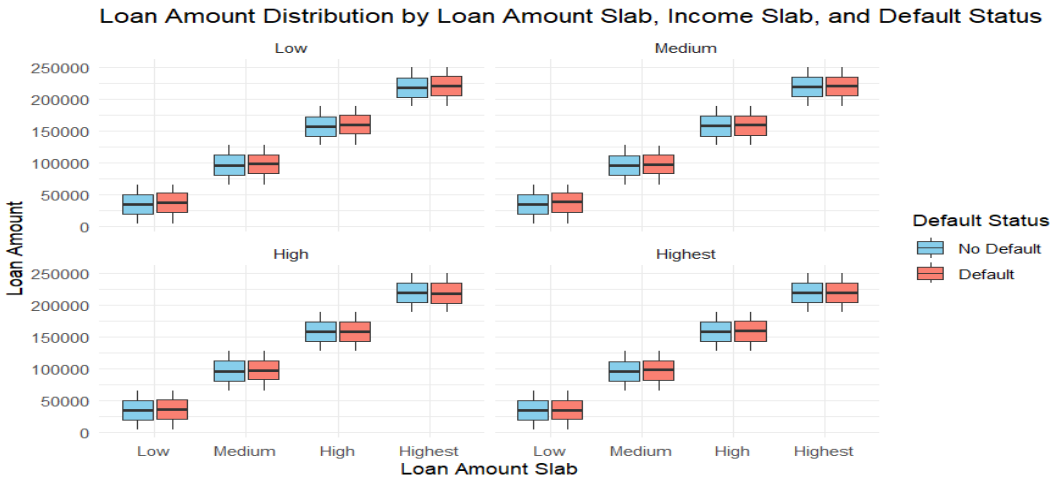
Figure 3: Box plots showing credit score variation across multiple variables



ii. **Figure 4:**

In Figure 4, the distribution of Loan Amount within Loan Amount Slabs (Low, Medium, High, Highest) across different Income Slabs (Low, Medium, High, Highest) is plotted and separated by Default Status (No Default vs. Default). Within each income slab, the loan amount increases thus showing a positive trend in the relationship. The propensity to take loans increases with the increase in income. The default and non-default categories do not vary much. There is no drastic variation between income slabs thus showing that these variables might not be informative when taken alone. We need to think about more meaningful interaction variables.

Figure 4: Loan amount distribution according to Loan amount slab, Income slab, and default status



Methodology

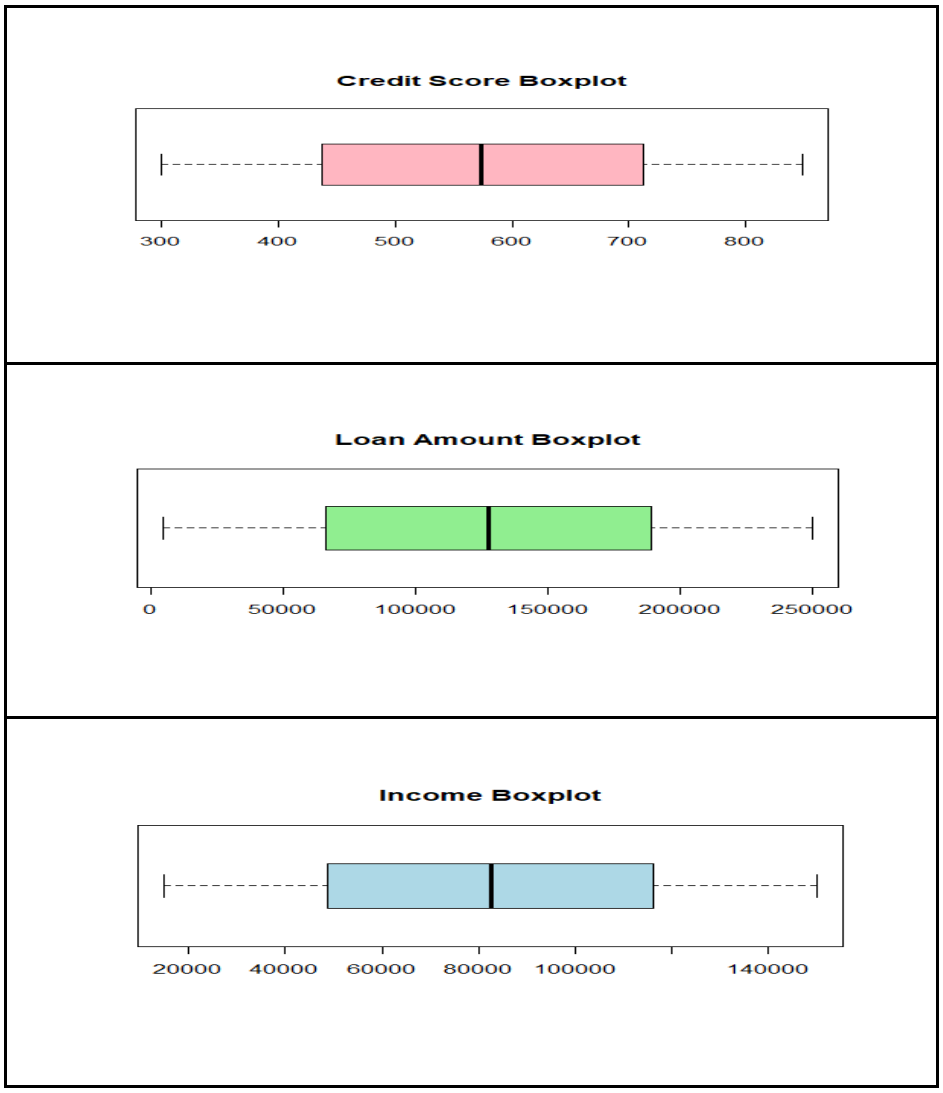
The methodology contains the following subsections :

1. **Data pre-processing - This section contains feature selection, outlier processing, missing value detection, and dealing with class imbalance.**
2. **Modeling Choices - This section involves the models that were considered to predict loan default.**
3. **Data Handling and Training - This section explains how data was handled and trained**
4. **Model Validation approach - This section explains the validation approach that was chosen for the model**

Data Pre-processing- Initial Insights:

The study focuses on predicting loan default among loan customers of a financial institution. After loading the dataset, the basic data analysis is done- for instance, treatment of missing values and outlier detection. Thankfully, the dataset does not have any missing values and outliers (see boxplots in Figure 5).

Figure 5: Boxplots of Credit Score, Loan Amount, and Income Indicating No Outliers



Therefore, there is no need to perform imputation techniques. Certain numeric variables need to be scaled for modeling purposes. This standardizes the variables and aids comparative analysis (independent of units). This is simply because credit score is a score and the loan amount is measured in currency. One cannot compare between the two, however, scaling enforces the range of the variables from 0 to 1 thereby leading to an empirically correct way to study relationships.

As a next step, the binary categorical (character) variables (yes/no) like whether the customer Has a Mortgage, has dependents, and has cosigners are transformed into numerical dummy variables (1/0). Additionally, One-hot encoding certain variables like Education, Employment Type, Marital status, and loan purpose was considered. Then, this was combined with the main dataset. At this stage, any cases of duplication of Loan IDs were removed.

The next step was to check for class imbalance. Since this dataset deals with a binary predictor, it might so happen that the non-defaulters (0) are more than defaulters (1). The abundance of the number of non-defaulters might make it challenging for the models to interject the minority class effectively.

In our dataset, there is a class imbalance as quite naturally the number of non-defaulters is more than the number of defaulters. There are 88% non-defaults and 12% defaults. There are various methods to handle class imbalance:

a. Oversampling the Minority class (ideal for small datasets)

b. Under sampling the majority class (for large datasets)

c. SMOTE (best for no information loss)

d. Class weighting (adjust weights during model training without modifying data)

Among the above methods, oversampling the minority class and SMOTE (Synthetic Minority Oversampling method) were chosen for treating the class imbalance. From the first method, it was noticed that the dataset was balanced perfectly- the ratio of defaults to non-defaults was 1:1. There were 225694 defaults and 225694 non-defaults. The SMOTE method gave us a subpar result – the balance ratio was approximately 4:1. However, SMOTE was preferred because of overfitting issues as it gave a better AUC in further models. As SMOTE creates synthetic samples for the minority class by interpolation, the data generation process is more realistic for the minority class. This helps in better generalization.

Further, interaction terms are investigated as well. We use a logistic regression to assess the significance of the potential interaction terms. An example would be income x loan amount. From the results, we find that all predictors (Income and loan amount and their interaction) are highly significant. This means that the effects on the probability of default are highly significant. In practical terms, for higher-income borrowers, the risk of default is reduced even if they take larger loans. Conversely, lower-income borrowers taking larger loans face a higher risk of default. The AIC of this model increases indicating that the additional terms are adding to the model performance. Income x Loan Amount, DTIRatio x CreditScore, MonthsEmployed x InterestRate, and LoanAmount x LoanTerm are statistically significant and improve the model's explanatory power. The interaction income x DTI ratio is not significant and hence will be removed.

Interpretation of Interactions:

To capture relationships between variables, interaction terms were explored. A key interaction analyzed was **Income × LoanAmount**, as these two features are critical indicators of default risk.

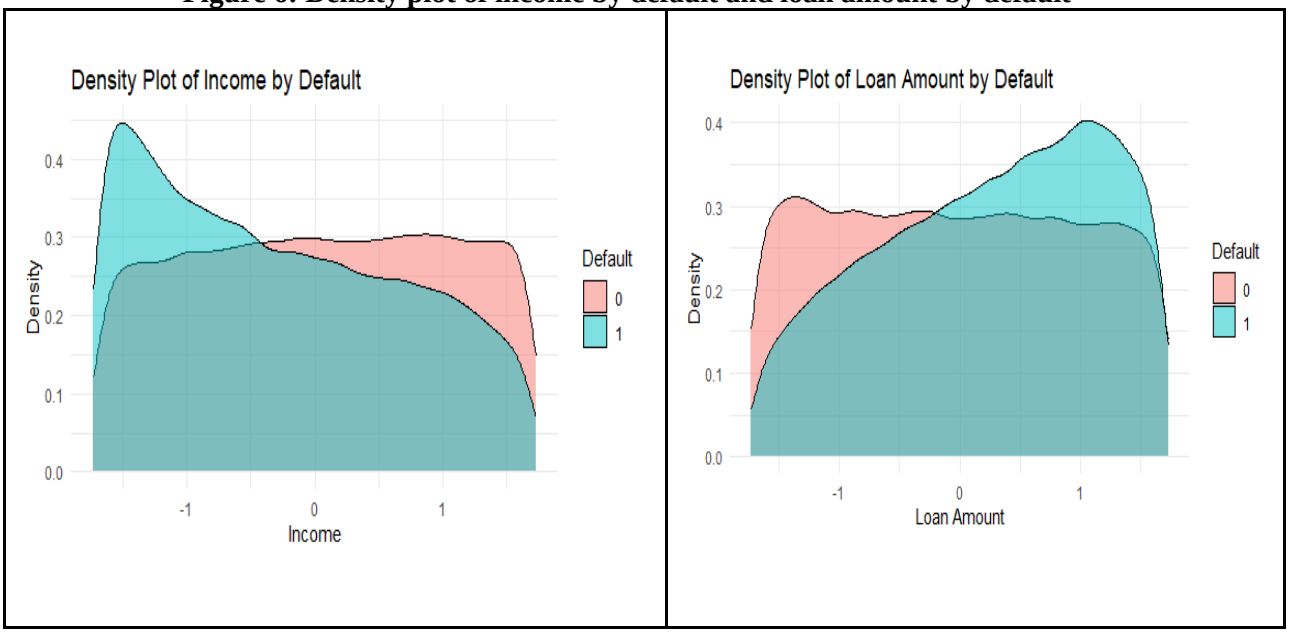
Instead of a scatter plot, **density plots** were used to provide a clearer view of the distribution of Income and Loan Amount based on Default status:

- **Income Distribution by Default** (Figure 3): This plot highlights that borrowers with lower incomes are more likely to default. The income density for defaulters shows a shift toward lower values compared to non-defaulters.

- **Loan Amount Distribution by Default** (Figure 4): Borrowers who default typically request higher loan amounts, as indicated by the density curve for defaulters, which is skewed toward higher loan amounts.
- **DTIRatio × CreditScore**:
    - High debt-to-income ratios offset the benefits of good credit scores.
- **MonthsEmployed × InterestRate**:
    - Employment stability protects against the impact of high interest rates.

These plots in Figure 6 demonstrate that both **Income** and **Loan Amount** play significant roles in predicting default, and their interaction was further examined in the modeling stage.

Figure 6: Density plot of income by default and loan amount by default



A crucial issue is multicollinearity in large datasets. Therefore, the Variance inflation factor (VIF) was calculated for each variable and the cross-relations and VIFs above 5 were observed depicting high multicollinearity. Hence, penalized logistic models, for instance, the LASSO and Ridge regressions which handle multicollinearity and overfitting problems more efficiently.

Despite the added feature engineering, the **Random Forest AUC dropped to 0.7274.** This may have been caused by noise introduced through interaction terms or over-reduction in feature selection. Since interaction terms added complexities, the modeling part of the study was approached without considering the interaction terms.

Summary of findings from Data pre-processing:

- Data is checked for **missing values, outliers, and any cases of duplication.**
- A thorough examination revealed no missing values across all features. Hence, no imputation techniques were required.
- Visual analysis through boxplots indicated the absence of outliers, eliminating the need for outlier treatment.

- The **numeric variables are scaled** for better interpretability.
- The **binary variables** were transformed into **dummy variables** (0/1). To avoid the dummy variable trap, **one-hot encoding** of certain **categorical variables** was used, for instance, Education, Employment Type, Marital status, and loan purpose.
- The dataset showed significant **class imbalance**, with 88% of loans being non-defaults (Default = 0) and only 12% being defaults (Default = 1).
- Two methods of treatment of class imbalance were considered. O**versampling the Minority Class** and SMOTE method. While the former was effective, it risked overfitting the model. Although SMOTE resulted in a slightly imperfect balance ratio (~4:1), it provided better generalization and improved AUC during modeling.
- Key insights from the interactions: High income protects against default, but the benefit decreases with larger loans, high credit scores generally reduce default risk, but this benefit is offset by high debt-to-income ratios, Employment stability offers slight protection against the impact of higher interest rates, and Large loans with long terms may be less risky than expected, possibly due to better loan structuring.
- Since interaction terms added complexities, the modeling part of the study was approached without considering the interaction terms.

Modeling Choices:

1. **Penalized Logistic Regression**:
   - Applied LASSO regularization to handle multicollinearity and improve generalization.
   - Served as a baseline but struggled with non-linear patterns in the data.
2. **Random Forest**:
   - Used an ensemble approach to capture non-linear relationships.
   - Decision trees usually suffer from overfitting; to mitigate this issue, a random forest model is used as it prevents overfitting due to averaging across trees.
   - Also provides feature importance scores for interpretation.
3. **XGBoost**:
   - Combines weak learners (typically decision trees) sequentially. Each new model focuses on reducing the errors made by the previous models.
   - Iteratively corrected prediction errors through boosting by optimizing a loss function (log-loss for binary classification).
   - Also includes parameters to handle a class imbalance in the data, although the class imbalance in our data was addressed with the help of SMOTE.
   - The evaluation metric used was “auc” which is commonly used for binary classification
   - Max depth was set to 6, which is a good starting point, as more depth can lead to overfitting.
4. **Neural Networks**:
   - Captured complex patterns with a single-layer neural network.
   - Used a combination of 5 and 10 neurons for the single-layered network and decay of 0.1 and 0.5 for regularization. The regularization parameter prevents overfitting by penalizing large weights in the neural network.
   - The hidden layer size controls the complexity of the neural network. More neurons can capture complex relationships, but this can also lead to overfitting.

5. **Support Vector Machine (SVM)**:
    - Used a radial kernel which produces smooth, continuous decision boundaries making it effective for our classification task.
    - Utilized stratified sampling to reduce the computational cost for training and hyperparameter tuning by creating a smaller yet representative subset of data.
6. **Gradient Boosting Machine (GBM):**
    - Combines weak learners typically decision trees, to create a strong classifications model. This technique makes trees sequentially, where successive trees aim to correct the errors made by the previous trees.
    - Tuned the number of trees, interaction depth, and learning rate.
    - Excels in handling both regression and classification tasks. Although here it's being utilized only to perform classification.
    - Is computationally expensive and requires careful tuning for optimal performance.
7. **Decision Tree:**
    - A model which uses a tree-like structure to split data into different branches and make predictions accordingly. The splits are done mostly based on criteria such as entropy and Gini index. Only nodes with high impurity or non-leaf nodes are considered for further splits and stops if the threshold i.e. the depth is reached or there are no more nodes to split.
    - Serves as a baseline for classification and is comparable to other more complex tree-like models.
    - Are intuitive, and easy to visualize and interpret.
    - Prone to overfitting with deep trees unless they are not regularized using techniques such as pruning or setting constraints.
8. **Elastic Net Regression:**
    - A variant of penalized logistic regression combining LASSO and Ridge regression penalties.
    - Selected lambda via cross-validation to ensure the balance between predictive performance and model simplicity.
    - Requires careful tuning because of both the L1 and L2 penalty parameters.
    - Highly effective in dealing with multicollinearity which is common in real-world datasets.
9. **Naive Bayes (NB):**
    - Naive Bayes assumes independence among predictors, making it computationally efficient and fast to train.
    - Probabilistic predictions were generated using class conditional probabilities and Bayes' theorem.
    - Despite its simplicity and independence assumption, Naive Bayes performed well, especially given the preprocessed data's reduced multicollinearity.
10. **K-Nearest Neighbors (KNN):**
    - The KNN algorithm relies on the proximity of data points for classification.
    - Data was scaled to ensure that features with larger magnitudes (e.g., income) did not dominate the distance calculations.
    - The number of neighbors (k) was set to 5 after experimenting with other values to balance bias and variance.

- While computationally expensive for larger datasets, KNN was effective for identifying simple patterns in the data.

Data Handling and Training:

- **Class Imbalance Handling:** SMOTE (Synthetic Minority Oversampling Technique) was used to balance the dataset by generating synthetic samples for the minority class.
- **Data Splitting:** The dataset was divided into 80% training and 20% testing after applying SMOTE to ensure that both classes were well-represented in the splits.
- **Evaluation Metric:** AUC-ROC was chosen as the evaluation metric to compare the discriminatory power of the models.

Model Validation approach:

The model validation approach used in this project was 5-fold cross-validation to mitigate overfitting and to ensure robust evaluation. This method has divided the dataset into 5 equal parts and then it uses 4 parts for training and 1 part is used for testing iteratively and this ensures that all data points were used for both validation and training. This approach is chosen as it provides a more comprehensive assessment of the performance of the model when compared with the single handout method, which might introduce variance or bias due to the randomness of the split.

Results

Overall Model Performance:

The table below summarizes the performance of all evaluated models based on AUC-ROC:

Table 1: Model Performance Summary with AUC-ROC and Key Observations

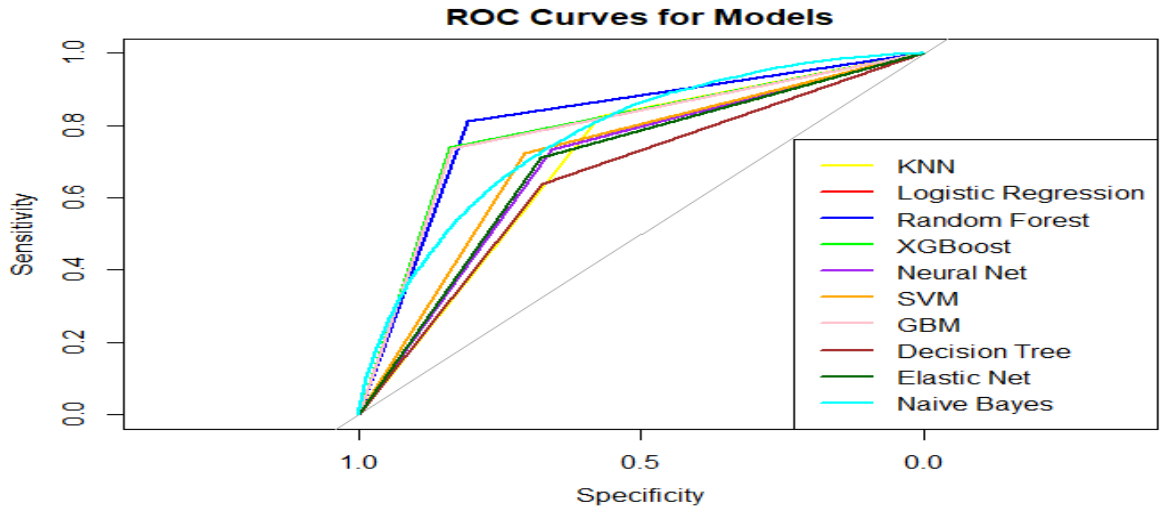
Model	AUC-ROC	Key Observations
Penalized Logistic Regression	0.6948	Performs moderately; acts as a baseline.
Random Forest	0.8081	Best overall performance.
XGBoost	0.7911	Competitive with Random Forest; robust model.
Neural Network	0.6944	Moderate performance; computationally heavy.
Support Vector Machine (SVM)	0.7144	Moderate performance with simple features.

Gradient Boosting Machine (GBM)	0.7857	Competitive performance; good generalization.
Decision Tree	0.6562	Simple and interpretable but with lower accuracy.
Elastic Net Regression	0.6948	Similar to Logistic Regression.
Naive Bayes	0.7697	Efficient and competitive for simpler problems.
K-Nearest Neighbors (KNN)	0.6986	Moderate performance; effective for simple patterns.

Combined ROC Curves:

The ROC curves in Figure 7 illustrate the trade-offs between sensitivity and specificity for each model. Random Forest consistently outperformed other models, followed by XGBoost and GBM. Decision Tree and Logistic Regression exhibited weaker discriminatory power, while Naive Bayes provided competitive results with minimal computational overhead.

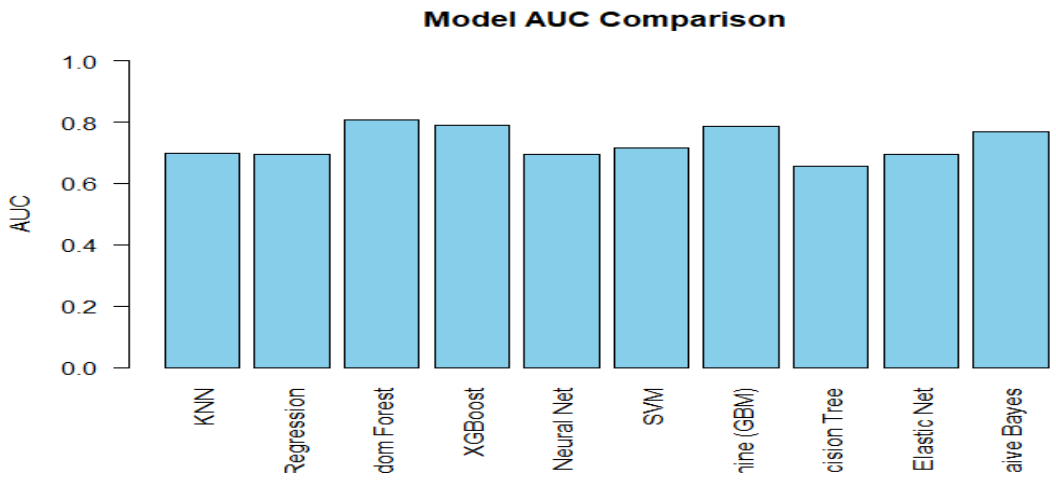
Figure 7: ROC Curves for Models



Bar Plot:

A bar plot in Figure 8 was created to visualize the AUC-ROC values for all models. Ensemble methods (Random Forest and XGBoost) stood out as top performers, while simpler models (Decision Tree, Logistic Regression) were less effective.

Figure 8: Model AUC Comparison



Key Observations:

- **Top Performer:** Random Forest achieved the highest AUC-ROC (0.8087), demonstrating its robustness and capability to handle complex interactions.
- **Boosting Models:** XGBoost and GBM offered competitive results, highlighting the strength of boosting techniques for imbalanced datasets.
- **Naive Bayes:** Despite its simplicity and strong independence assumptions, Naive Bayes performed surprisingly well, with an AUC of 0.7697.
- **Limitations of Simpler Models:** Decision Tree and Logistic Regression underperformed, reflecting the dataset's complexity and non-linear patterns.

Recommendations:

- Focus on ensemble methods like Random Forest or XGBoost for deployment, as they consistently deliver robust performance.
- Optimize Naive Bayes for simpler tasks or resource-constrained environments.
- Leverage SMOTE preprocessing for imbalanced datasets and explore additional sampling techniques.
- Further, tune hyperparameters and explore feature engineering to improve model performance across the board.

Conclusion

The main aim of the project is to address the challenge of loan default prediction using machine learning models so that it can help in improving risk management for financial institutions. By using a dataset that has diverse financial, loan-related, and demographic attributes, a robust modeling pipeline has been

implemented in this project that includes the preprocessing of data, feature selection, and also evaluation across various supervised learning methods.

Summary of Problem, Approach, and Findings:
The challenge required developing robust predictive models for managing the imbalance between default and non-default cases, as 88 percent of the loans that are present in the dataset were non-defaults. To handle this, SMOTE (Synthetic Minority Oversampling Technique) has been used to balance classes, making sure that the models aren't biased toward the majority class. Preprocessing and feature selection are essential steps for refining the dataset, with important predictors like credit score, loan amount, and income standing out. Interaction terms, while logical in theory, are removed because of their negative impact on the performance of models.

The modeling approach used diverse algorithms to compare their predictive power:

- **Random Forest** with an AUC-ROC of 0.8085 became the top performer which shows its ability to model complex feature interactions.
- **XGBoost** with an AUC-ROC of 0.7911 became the close contender to random forest and also showed robust efficiency and performance.
- **Elastic Net Regression** and **Penalized Logistic Regression** with an AUC-ROC of 0.6948 each, served as the reliable baselines.
- **SVM** and **neural networks** did capture non-linear relationships but could not outperform ensemble methods because of tuning challenges and computational complexity.

Key Insights:

- **Credit Score and Income as Primary Predictors:** "Income" was found to be the strongest predictor of loan default, followed by credit score. Borrowers who have poor credit scores and lower income are much more likely to default.
- **Random Forest as the Best Model:** So, out of all models, random forest has achieved the highest performance with an AUC-ROC of 0.8085 and this makes it the most reliable model for loan default prediction.
- **Effectiveness of SMOTE:** Class imbalance was handled using SMOTE and it has improved model performance significantly and also helped in better detection of defaults.
- **Limited Impact of Interaction Terms:** While interaction terms are logical to include, they have reduced the model's performance and were removed for optimal results.

Critical Assumptions and Limitations:

- SMOTE effectiveness was a critical assumption, as SMOTE will rebalance the dataset for enhancement of model learning. Yet, synthetic examples might not be able to perfectly represent the scenarios of the real world.
- Even though the removal of interaction terms has improved model performance, it might result in overlooking nuanced relationships between the predictors.
- The performance of models depends on high-quality data availability. The real-world deployment will require continual updates ans also validation as economic conditions and borrower behaviors evolve.

Final Recommendations and Impact:

Upon observing results, the Random Forest model is best for deployment as it provides a balanced trade-off between interpretability and accuracy. Random Forest is well-suited for the identification of high-risk borrowers and will also be able to help financial institutions in the implementation of pre-emptive measures like flexible repayment options, stricter lending criteria, and financial counseling for the borrowers that are flagged as high-risk.

While the models performed very well in this controlled environment, it is essential to acknowledge that no model will be able to guarantee a perfect prediction. Regular monitoring, updates, and integration with broader risk management strategies are required to maintain model effectiveness. Also, financial institutions must combine these developed predictive models with domain expertise to enhance decision-making and minimize the risks of loan default effectively.

References

- Federal Reserve Bank of St. Louis. (n.d.). *The delinquency rate on consumer loans, all commercial banks (DRCLACBS)*. Retrieved December 11, 2024, from https://fred.stlouisfed.org/series/DRCLACBS
- Associated Press. (2024, March 29). *Severe delinquencies among young borrowers reach highest levels in over a decade*. AP News. Retrieved December 11, 2024, from https://apnews.com/article/1eb685347b091ff53361e141d80405ad
- Rubin, B. (2024, September 25). *Credit card debt surges as rates hit record highs*. The Wall Street Journal. Retrieved December 11, 2024, from https://www.wsj.com/finance/credit-card-debt-loans-high-interest-rates-8da11e83

Appendix

	Column_name	Column_type	Data_type	Description
0	LoanID	Identifier	string	A unique identifier for each loan.
1	Age	Feature	integer	The age of the borrower.
2	Income	Feature	integer	The annual income of the borrower.
3	LoanAmount	Feature	integer	The amount of money being borrowed.
4	CreditScore	Feature	integer	The credit score of the borrower, indicating their creditworthiness.
5	MonthsEmployed	Feature	integer	The number of months the borrower has been employed.
6	NumCreditLines	Feature	integer	The number of credit lines the borrower has open.
7	InterestRate	Feature	float	The interest rate for the loan.
8	LoanTerm	Feature	integer	The term length of the loan in months.
9	DTIRatio	Feature	float	The Debt-to-Income ratio, indicating the borrower's debt compared to their income.
10	Education	Feature	string	The highest level of education attained by the borrower (PhD, Master's, Bachelor's, High School).
11	EmploymentType	Feature	string	The type of employment status of the borrower (Full-time, Part-time, Self-employed, Unemployed).
12	MaritalStatus	Feature	string	The marital status of the borrower (Single, Married, Divorced).
13	HasMortgage	Feature	string	Whether the borrower has a mortgage (Yes or No).
14	HasDependents	Feature	string	Whether the borrower has dependents (Yes or No).
15	LoanPurpose	Feature	string	The purpose of the loan (Home, Auto, Education, Business, Other).
16	HasCoSigner	Feature	string	Whether the loan has a co-signer (Yes or No).
17	Default	Target	integer	The binary target variable indicating whether the loan defaulted (1) or not (0).