INFORMATION SCIENCE: MACHINE LEARNING

BANKING DEFAULT PREDICTION WITH MACHINE LEARNING

AUTHOR: MOHAMMED SAMEER SYED(ID: 23915373)

AUGUST 26, 2024

Abstract

This project aims to develop a fully reproducible machine learning workflow to predict loan default risks for a German bank using historical customer data. The dataset includes various customer attributes such as credit history, loan amount, employment duration, and savings balance. The project involves systematically exploring and cleaning the data, performing feature engineering, and training multiple machine learning models to identify the most accurate predictor of loan defaults. The report will detail the methods, results, and interpretations, ensuring that the entire workflow is transparent and can be replicated by others.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Project Specification

In the banking industry, effectively managing the risk of loan defaults is essential to maintaining financial stability. In light of this, a German financial institution has compiled a dataset consisting of historical information from 1,000 customers who have taken loans. The goal is to build a machine learning model that can accurately predict loan defaults, providing the bank with a proactive tool for risk management. The dataset, named **Germanbank.csv**, includes 17 features that represent various financial and demographic aspects of the customers, such as account balance, loan characteristics, credit history, employment duration, and other pertinent attributes.

The target variable, 'default,' signifies whether a customer has defaulted on their loan. An initial visual exploration of this variable shows a notable imbalance in the data: roughly 700 customers did not default, while around 300 did. This significant class imbalance suggests the necessity of employing specialized techniques, such as class weighting or resampling, to ensure that the predictive model accurately reflects both default and non-default scenarios but with a risk of data leakage. Addressing this imbalance is crucial for the model to achieve balanced performance metrics and make fair predictions across all classes and it has been showcased in the report further.

The core objective of this project is to develop a machine learning model capable

of predicting potential loan defaulters using the provided historical data. By leveraging these predictive insights, the bank aims to bolster its risk assessment processes and make more informed lending decisions. Successfully implementing this model could contribute to minimizing financial losses due to defaults and improving the overall quality of the bank's loan portfolio.

1.2 Objective

The banking sector is inherently susceptible to risks associated with loan defaults, which can have a profound impact on financial stability and operations. Given this challenge, banks must leverage data-driven approaches to identify potential loan defaulters early and mitigate their risks. In this project, we aim to explore and address the problem of loan defaults using machine learning techniques. By analyzing the historical loan data of 1,000 customers provided by a German bank, our objective is to build a predictive model that can accurately classify whether a customer will default on their loan.

Several interesting questions will be investigated throughout this project.

- Firstly, which machine learning models are most effective at predicting loan defaults, considering the imbalanced nature of the dataset? We will compare models such as Random Forest, Gradient Boosting, Logistic Regression, and others to evaluate their performance.
- Secondly, how do various features such as employment duration, savings balance, and credit history contribute to predicting loan defaults? Feature importance analysis and visualization will help identify the most significant predictors of loan default.
- Finally, what are the potential biases or limitations within the dataset that could impact the predictive performance of the models, and how can they be mitigated?

By addressing these questions, we aim to develop a robust and reproducible workflow that leverages machine learning to enhance the bank's ability to assess loan risk. The outcomes of this project could contribute to more informed decision-making in the banking sector, potentially reducing the occurrence of loan defaults and improving financial sustainability. .

2 Implementation

2.1 EDA

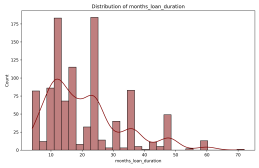
The first step of our analysis involved thorough data cleaning and preprocessing to ensure that the dataset was prepared for modeling. During the Exploratory Data Analysis (EDA), we identified and corrected inconsistencies in the data. For instance, a redundant category, 'car0' in the purpose attribute was replaced with 'car' as it was already present. Additionally, we addressed the inconsistency in the employment duration feature by ensuring that all instances where employment duration was marked as 'unemployed' were consistently labeled in the job column. Following this, ordinal encoding was applied to ordinal attributes, and one-hot encoding was used for nominal attributes. The target variable 'default' was separated from the dataset, and the remaining numerical variables were scaled to ensure that all features were on a comparable scale.

2.2 Visualization

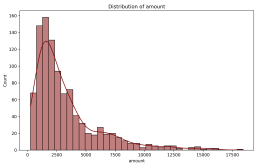
To gain insights into the distribution and relationships of the features, we utilized a variety of visualizations.

2.2.1 KDE + Histogram for each numerical feature

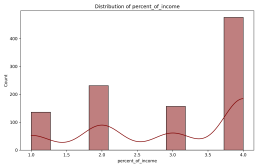
Kernel Density Estimate (KDE) plots combined with histograms were generated for each numerical feature to visualize their distributions.



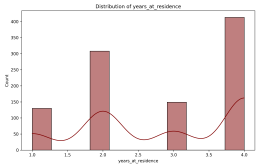
(a) Month loan duration



(b) Amount

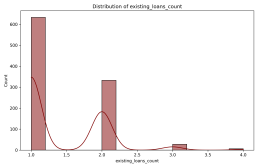


(c) Percent of Income

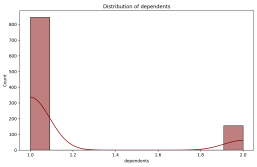


(d) Residence years

(e) Age



(f) Exisiting loans

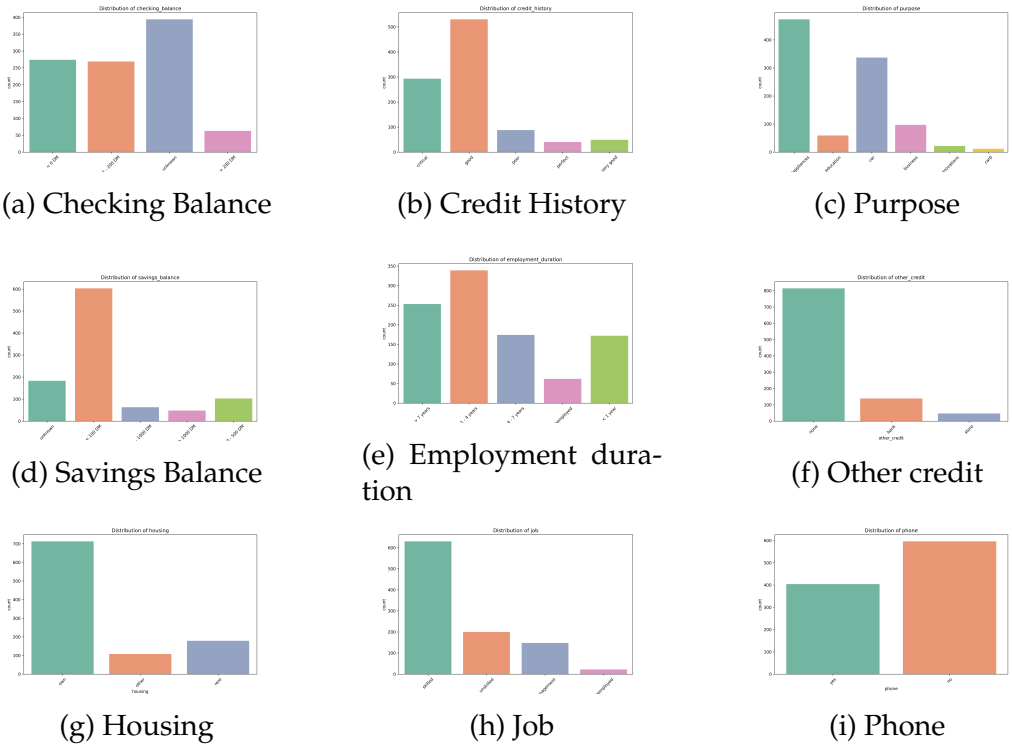


(g) Dependents

Figure 2.1: KDE + Histogram for each numerical feature

2.2.2 Bar plots for each categorical feature

Bar plots were created to analyze the distribution of categorical features, providing an overview of their frequency across different categories.



(a) Checking Balance (b) Credit History (c) Purpose

(d) Savings Balance (e) Employment duration (f) Other credit

(g) Housing (h) Job (i) Phone

Figure 2.2: Bar plots for each categorical feature

2.2.3 Pair plot between numerical variables

A pair plot was constructed to explore relationships between numerical variables and identify potential correlations or patterns.

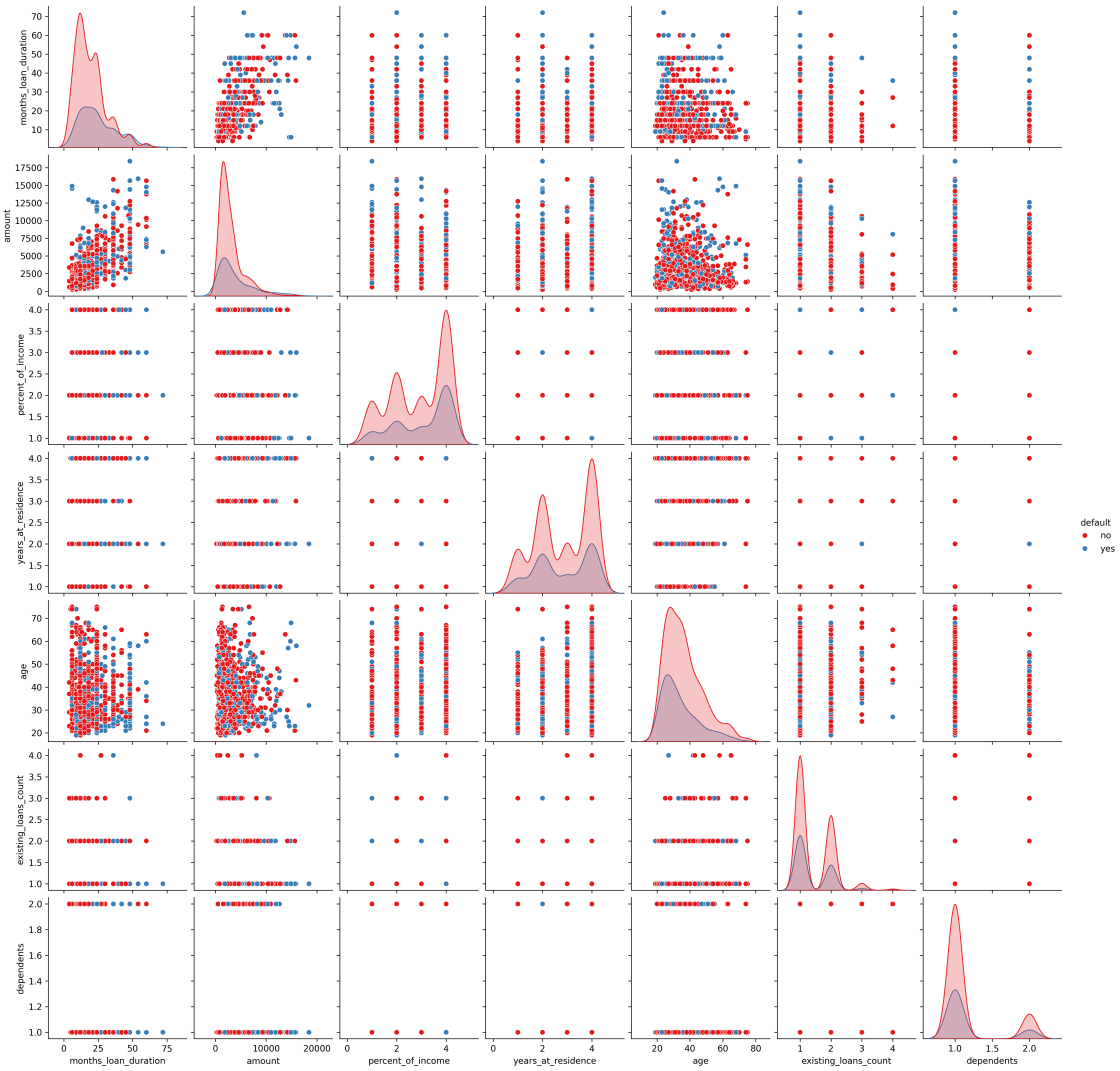


Figure 2.3: Pair Plot

2.2.4 Heatmap of correlation

Lastly, we generated a correlation heatmap to quantify the strength of relationships between features and identify highly correlated variables that could impact model performance.

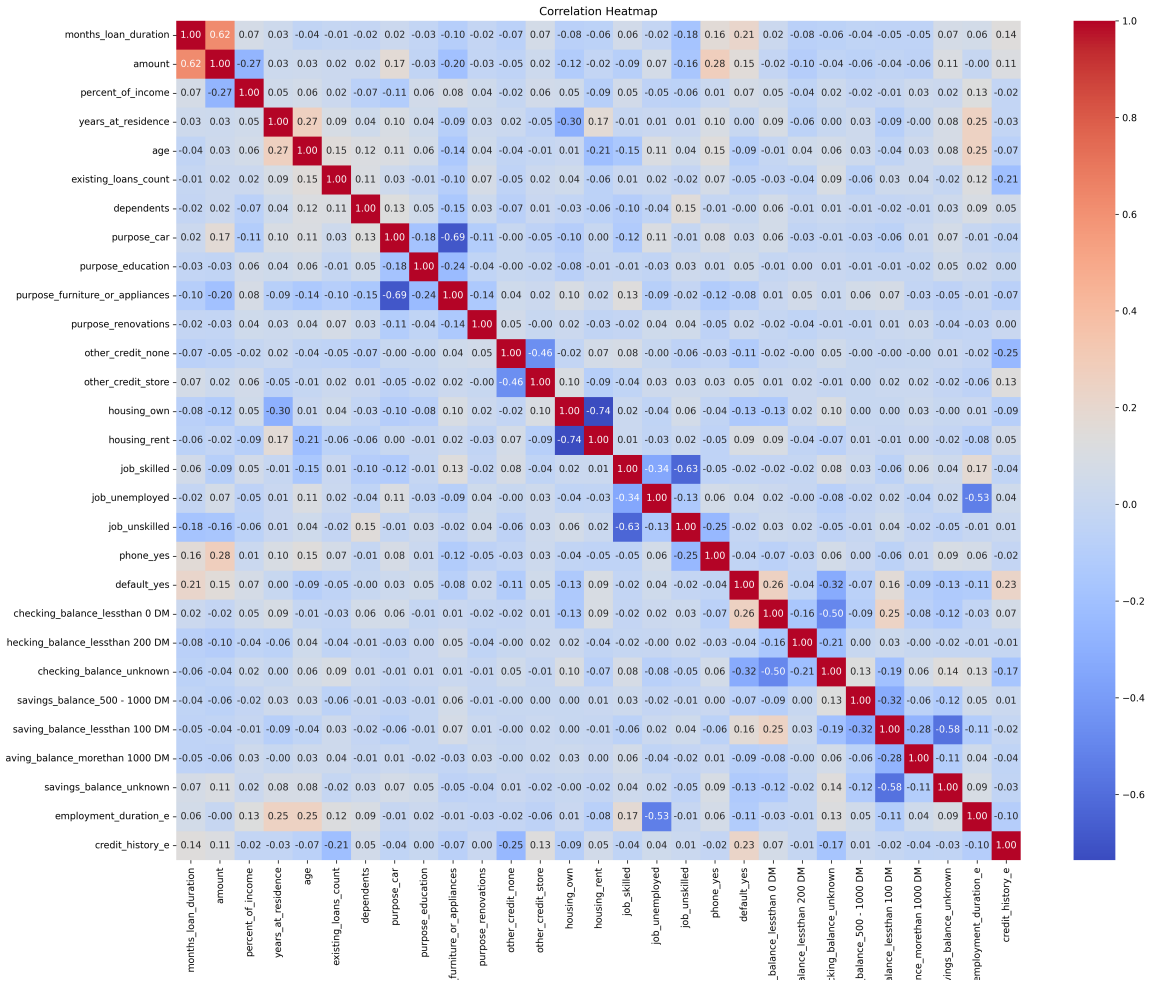


Figure 2.4: Heatmap of correlation

2.3 Feature Selection

Feature selection is a crucial step in machine learning to identify the most relevant features and reduce the dimensionality of the dataset. This helps improve model performance and interpretability. In our project, we employed the Recursive Feature Elimination with Cross-Validation (RFECV) technique to select the optimal subset of features. RFECV iteratively removes features that have the least impact on the model's performance, ensuring that the selected features contribute significantly to the prediction task.

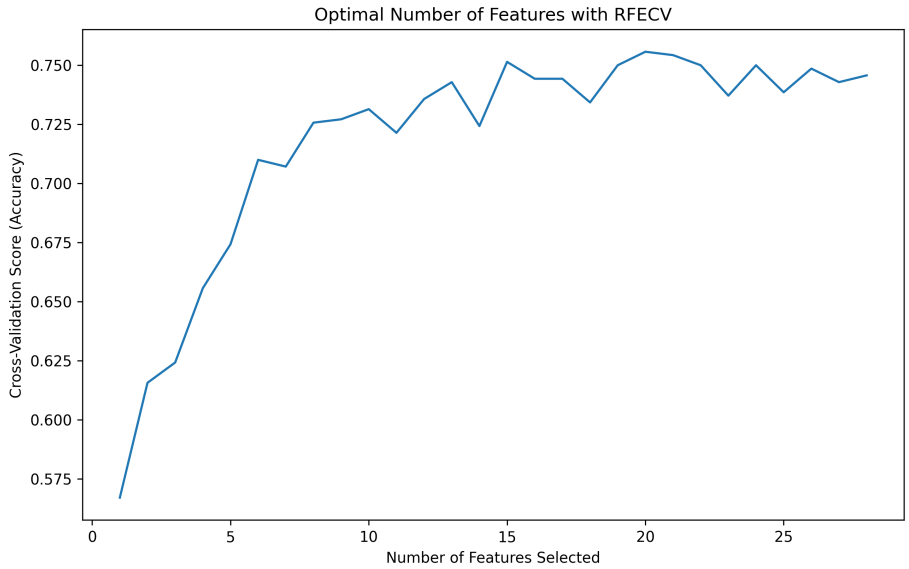


Figure 2.5: Feature Selection plot

3 Results And Performance

3.1 Model development

Our analysis began with the implementation of base models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boost, XGBoost, KNN Classifier, Naive Bayes Classifier, and LightGBM. These models were trained on the dataset without any special adjustments for class imbalance. Initial evaluation metrics, such as accuracy, precision, recall, and ROC AUC scores, were recorded and analyzed across all models. The base models provided a broad understanding of the dataset's characteristics and the performance potential of different algorithms.

Following this, we introduced penalty-based adjustments to address the class imbalance issue, which disproportionately favored non-default cases. Logistic Regression, Decision Tree, and Random Forest models were re-trained with class penalties to balance the prediction accuracy across both default and non-default classes. Additionally, we explored anomaly detection techniques using Isolation Forest and One-Class SVM to identify potential outliers or anomalies that could signify default cases. This approach aimed to enhance the model's ability to detect rare instances of default.

Finally, hyperparameter tuning was performed on the Random Forest and XGBoost models to optimize the performance. Tuning process involved adjusting parameters such as the number of estimators, maximum depth, and learning rate to improve predictive accuracy and overall model performance.

Model	Accuracy	Precision	Recall	ROC AUC
Gradient Boosting	0.7567	0.6197	0.4889	0.6801
Random Forest	0.75	0.6315	0.4	0.65
KNN Classifier	0.75	0.6190	0.4333	0.6595
XGBoost	0.7433	0.5802	0.5222	0.6801
LightGBM	0.7267	0.5444	0.5444	0.6746
Decision Tree	0.7133	0.5212	0.5444	0.6650
Naive Bias	0.7067	0.5102	0.5556	0.6634
Logistic Regression	0.67	0.4705	0.8	0.7071

Table 3.1: Model Performance

3.2 Performance Evaluation

In addition to model comparisons, ROC AUC curves and Precision-Recall curves were plotted for all models to visualize their performance. These curves provided insights into the trade-off between precision and recall, and the overall discriminative ability of each model. A summary table of results, including all key metrics, was generated to consolidate our findings.

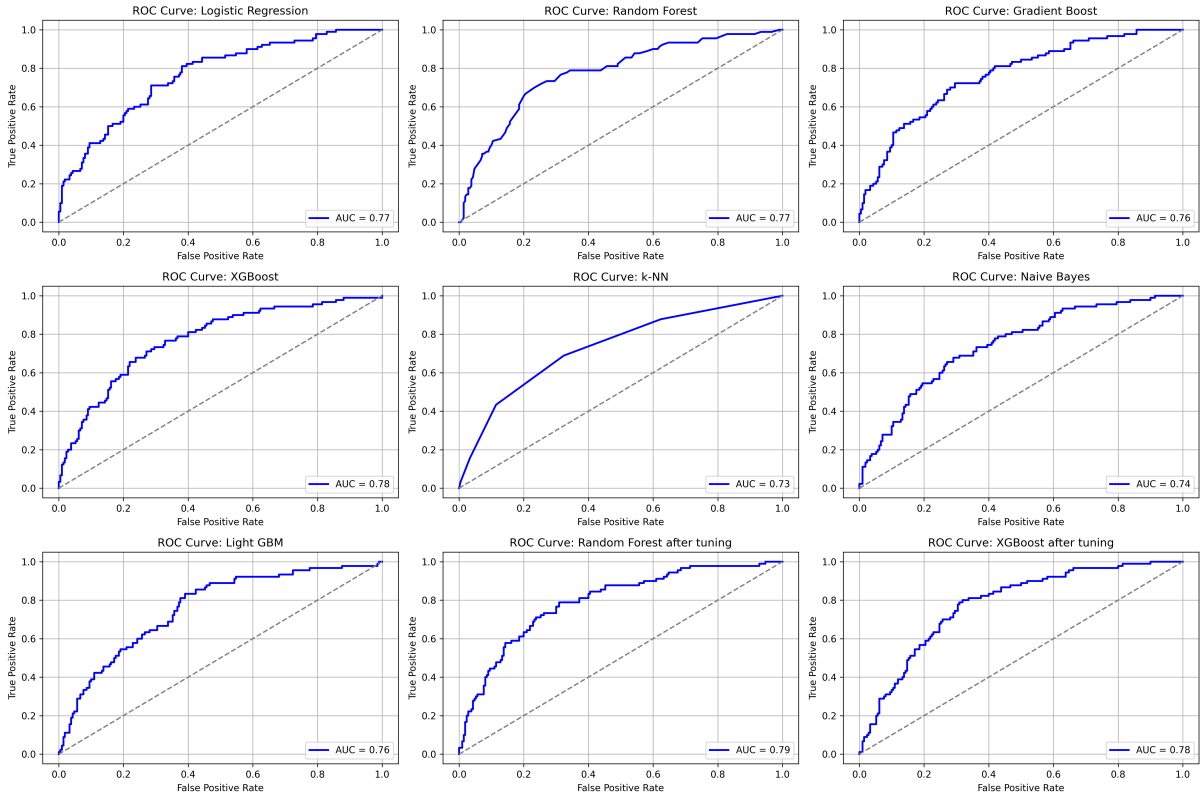


Figure 3.1: ROC AUC plot

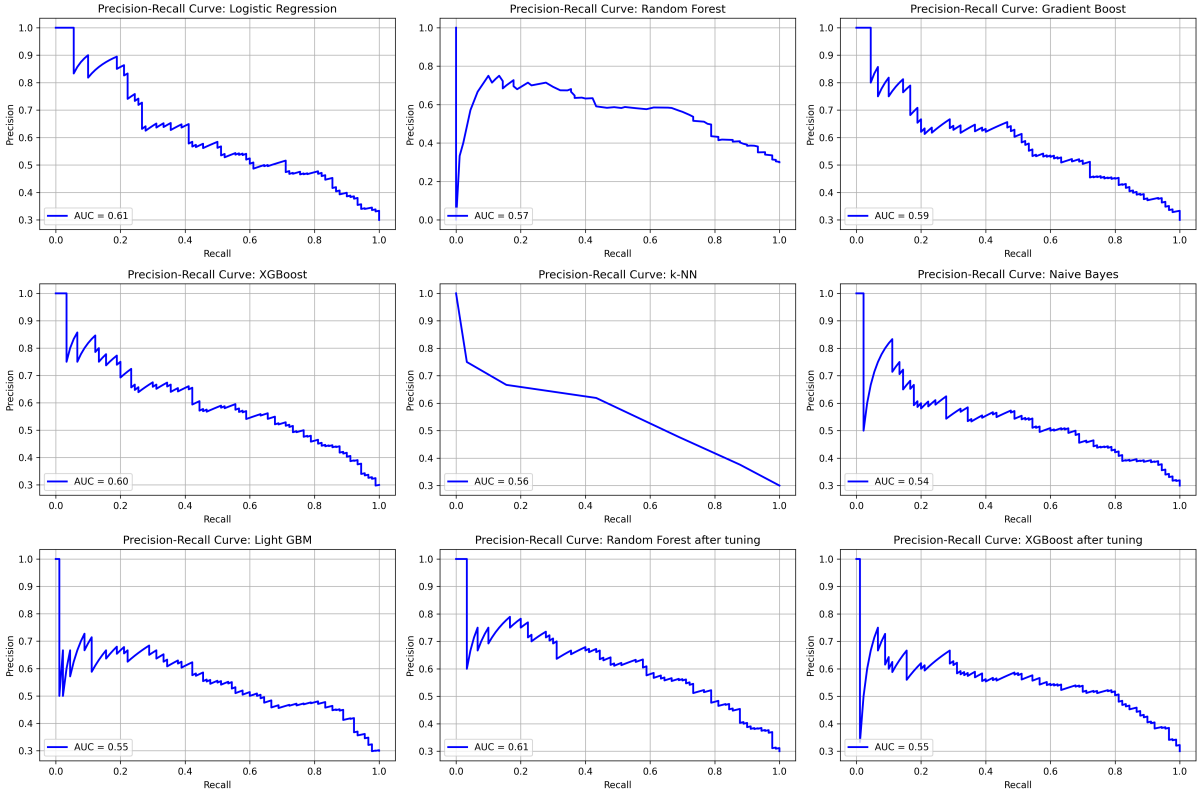


Figure 3.2: Precision Recall plot

4 Discussions

4.1 Summary

The major findings from this study indicate that Gradient Boosting achieved the highest overall accuracy (0.7567), while Random Forest and KNN Classifier performed similarly with accuracies of 0.75. Gradient Boosting and XGBoost demonstrated strong ROC AUC scores (0.6802 and 0.6802, respectively), indicating good discriminative ability between loan defaulters and non-defaulters. Logistic Regression, despite having the lowest accuracy (0.67), exhibited the highest recall (0.80), which highlights its strength in identifying default cases at the cost of precision. Hyperparameter tuning of Random Forest and XGBoost led to a significant increase in performance, with the tuned Random Forest model achieving the highest accuracy (0.9) and balanced precision and recall metrics, making it the most robust model in this analysis.

4.2 Interpretations

The findings suggest that while base models offer a good starting point for predictions, addressing class imbalance through penalty methods and hyperparameter tuning can significantly enhance model performance. The improvement in Random Forest after tuning underscores the importance of fine-tuning model parameters to achieve better predictive accuracy. However, models such as KNN Classifier, which initially showed competitive performance,

were surpassed by more sophisticated algorithms like Random Forest and XGBoost after tuning.

4.3 Limitations

This study's limitations stem primarily from the imbalanced nature of the dataset. Attempts to address the imbalance, such as upsampling, resulted in models that exhibited high accuracy but were at risk of data leakage, potentially compromising the generalizability of the model. Future work could explore alternative techniques for handling imbalance, such as SMOTE (Synthetic Minority Over-sampling Technique) or ensemble methods that balance the dataset more effectively. Additionally, expanding the dataset with more diverse and recent customer data could enhance the model's robustness and applicability to different financial contexts.

4.4 Handling Class Imbalance: Upsampling Study

To further explore the effects of class imbalance, we conducted additional experiments using upsampling techniques. These techniques increased the representation of the minority class (defaulters) in the training set. After upsampling, models such as Random Forest and XGBoost achieved even higher accuracy, with the tuned Random Forest model reaching 0.9 accuracy. However, this improvement came with the risk of data leakage. The following table and ROC curves summarize the performance of the models after upsampling.

Model	Accuracy	Precision	Recall	ROC AUC
Random Forest after tuning	0.9	0.9005	0.9056	0.8998
XGBoost	0.8771	0.8549	0.9167	0.8759
Random Forest	0.8742	0.8469	0.9222	0.8728
XGBoost after training	0.8742	0.8505	0.9167	0.8730
KNN Classifier	0.82	0.7695	0.9278	0.8168
Gradient Boosting	0.8	0.7864	0.8389	0.7988
Naive Bias	0.68	0.6954	0.6722	0.6802

Table 4.1: Model Performance After Upsampling

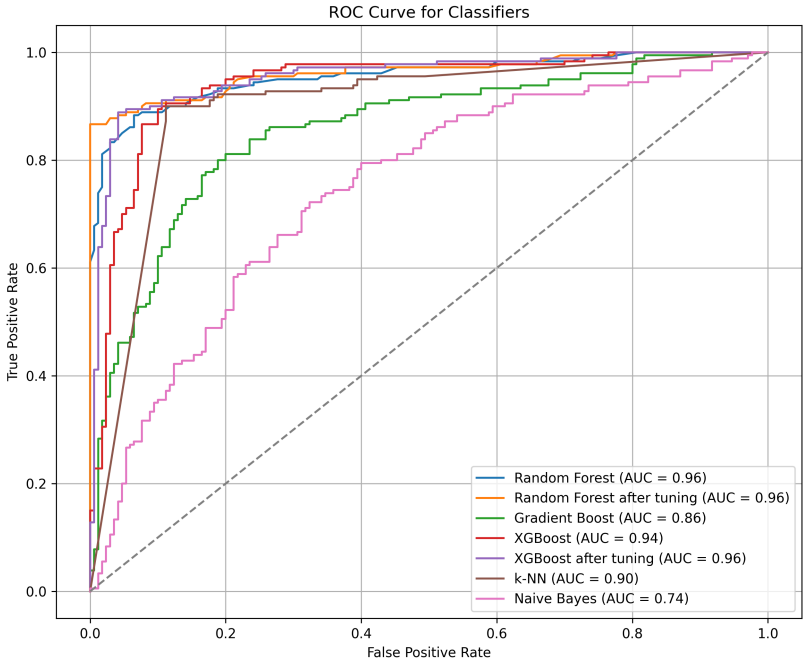


Figure 4.1: ROC Curves for Models After Upsampling

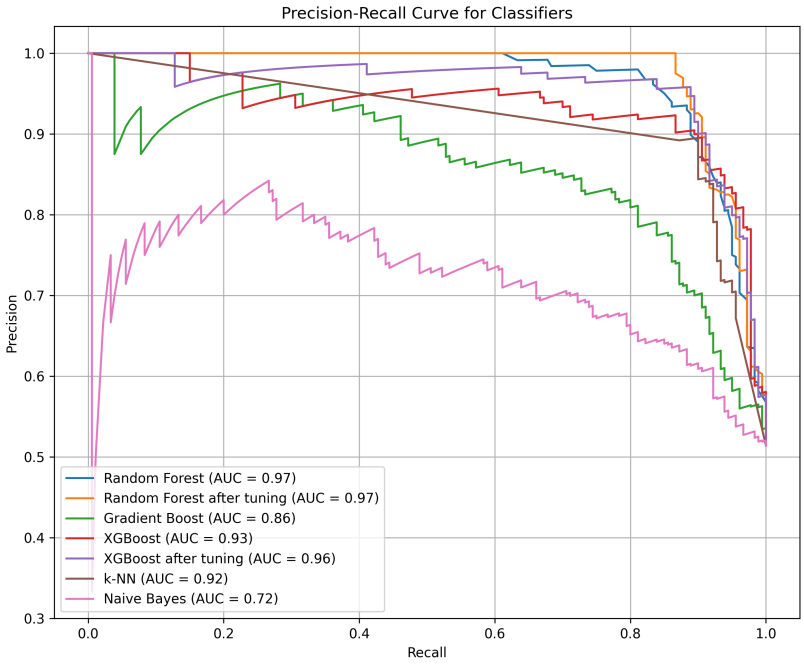


Figure 4.2: Precision-Recall Curves for Models After Upsampling

5 | Conclusion

In conclusion, this study highlights the effectiveness of ensemble models such as Random Forest and XGBoost, particularly after hyperparameter tuning, in predicting loan defaults. While addressing class imbalance through techniques like upsampling can improve accuracy, it also introduces potential risks of overfitting and data leakage. The main takeaway is that careful handling of class imbalance and model tuning is crucial to building robust and generalizable predictive models. Future work should focus on refining these approaches and expanding the dataset to further enhance prediction accuracy and applicability in real-world banking scenarios.