Final Project Introduction to Machine Learning

Report by

A N A N D R J

CONTENTS

INTRODUCTION

PROBLEM STATEMENT

BANK DEFAULTER DATA ANALYSIS

This business problem is an example for predicting loan defaults in the German bank.

The goal of the "Loan Default Prediction" project is to develop a replicable and publication-worthy framework. This project will demonstrate the systematic process of building and evaluating machine learning models tailored for predicting loan defaults in the Banking Financial sector.

This is an extremely critical part in any organization that lends money.

INTRODUCTION

PROBLEM STATEMENT

BANK DEFAULTER DATA ANALYSIS

Dataset

The dataset from the German bank provides valuable insights into predicting loan defaults, a major concern for financial institutions. The bank faces the critical task of spotting customers at risk of defaulting on loans, with historical data on customers' loan activities. The primary objective is to build a predictive machine learning model capable of accurately predicting loan defaults based on historical customer information and various associated features.

Objective of this project:

- From the provided dataset which machine learning model is most effective in predicting loan defaults using the German bank dataset?
- How we can improve model performance?
- What role does a customer's credit history play in the probability of loan defaulting or not?

Dataset Explanation

The data set has 17 columns and 1000 rows. Columns are described below and each row is a customer.

- checking_balance - Amount of money available in account of customers
- months_loan_duration - Duration since loan taken
- credit_history - credit history of each customers
- purpose - Purpose why loan has been taken
- amount - Amount of loan taken
- savings_balance - Balance in account
- employment_duration - Duration of employment
- percent_of_income - Percentage of monthly income
- years_at_residence - Duration of current residence
- age - Age of customer
- other_credit - Any other credits taken
- housing- Type of housing, rent or own
- existing_loans_count - Existing count of loans
- job - Job type
- dependents - Any dependents on customer
- phone - Having phone or not
- default - Default status (Target column)

Data Wrangling

Data Cleansing is the crucial part of the Model Building as it will help in improving the accuracy and avoid any anomaly in the ML Models.

Pre-cleaning the data

- The data is about the behavior of the repayment of the credit to the bank.
- The data has 1000 entries of rows and 17 columns (Variables).
- The data has 10 object data type and 7 int data type.
- The data has no entries as Null/Nan values.
- The data has no Duplicate Row.

Cluttered – Before Cleaning

Post-Cleaning the Data

Correcting some typos such as car0 to car in the dataset.

The clean data is ready for the next step to proceed.

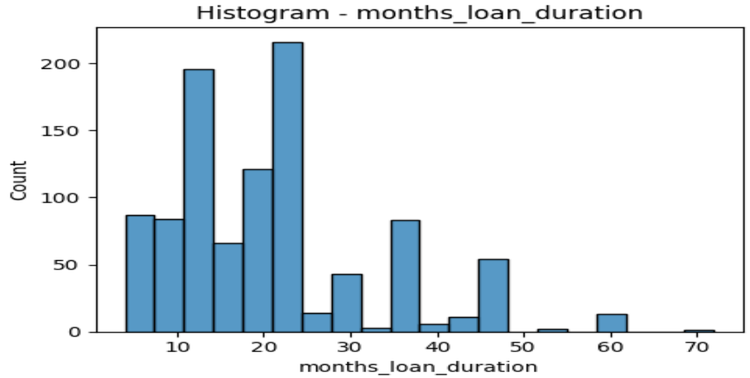
.

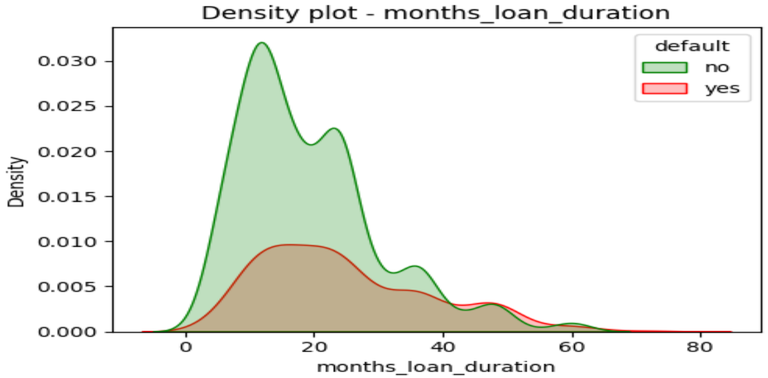
Organised – After cleaning

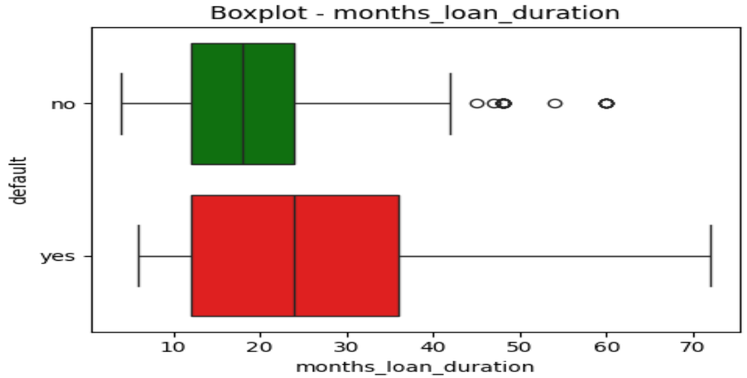
Methods And Material

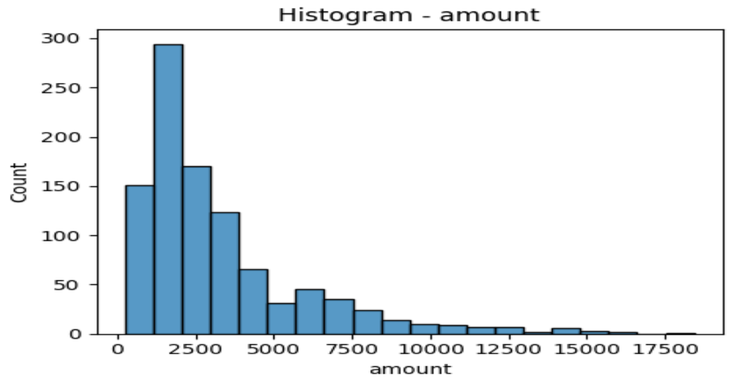
Exploratory Data Analysis

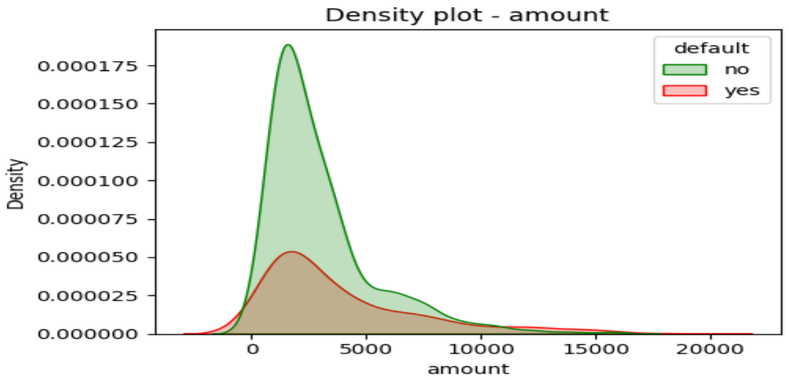
In the Exploratory Data Analysis (EDA) phase, I used visualizations like histograms, density plots, and boxplots to understand the distribution of numerical features and identify outliers. Pair plots and heatmaps were employed to explore correlations between numerical features. The EDA process raised interesting research questions, offering initial insights that may need further study for conclusive findings.

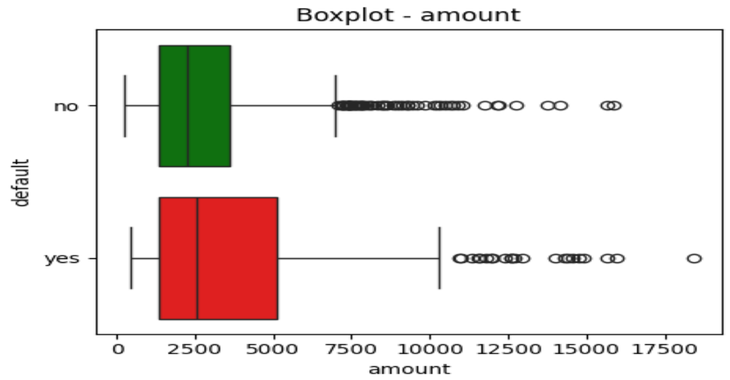


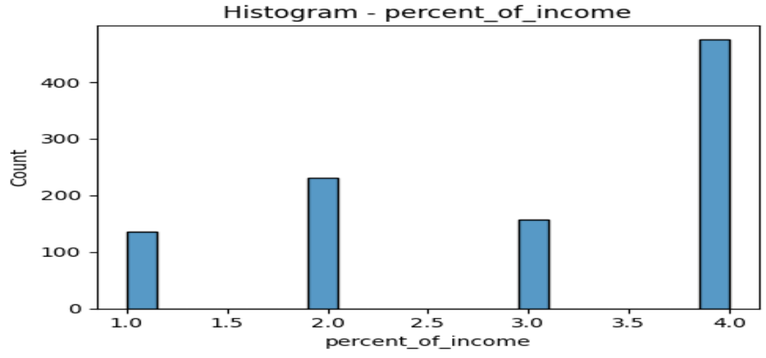


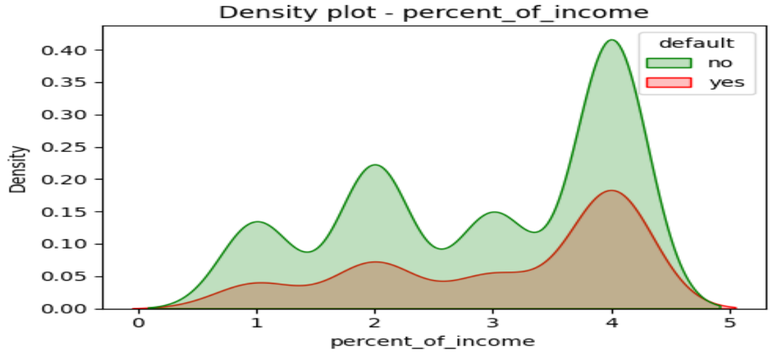


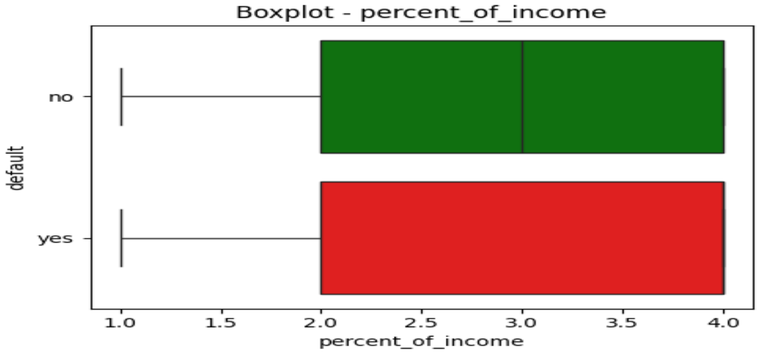








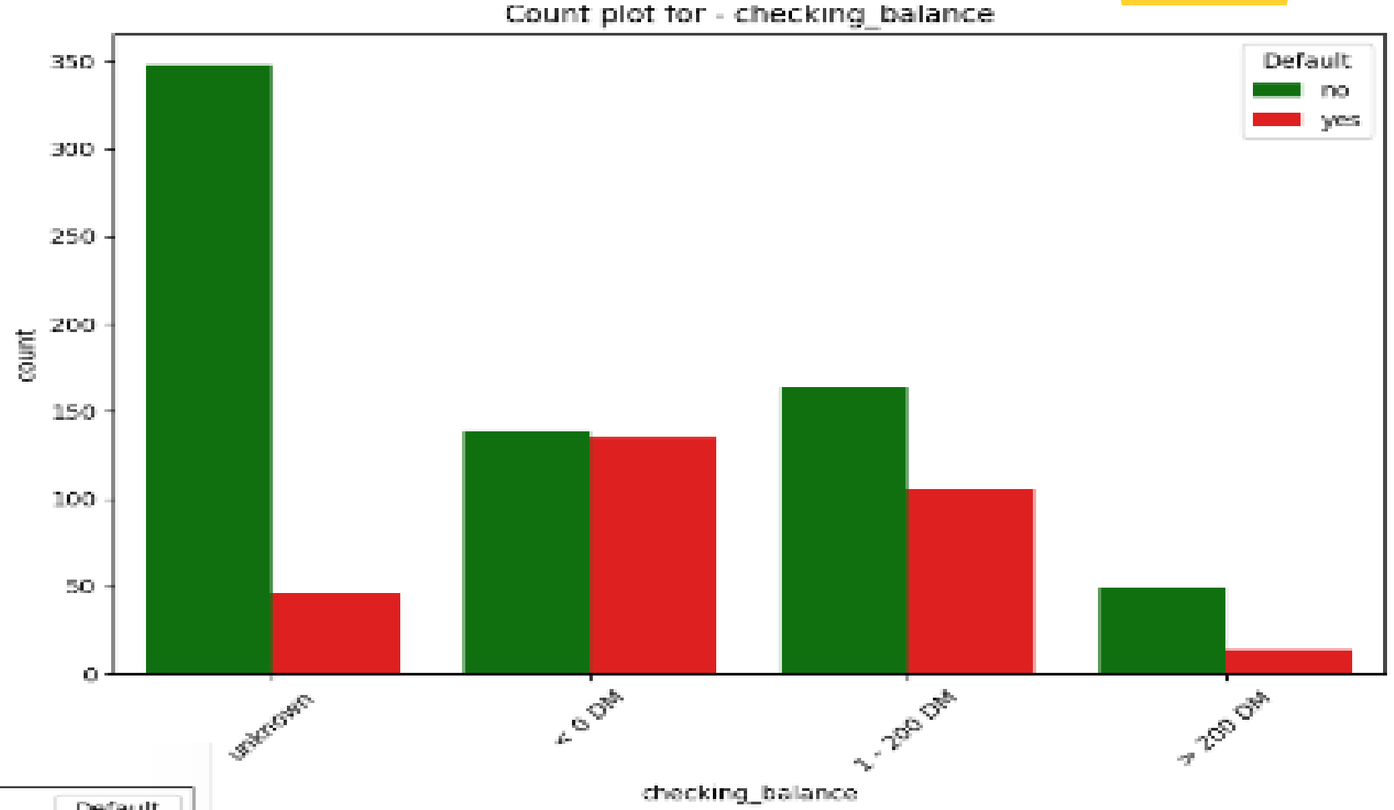


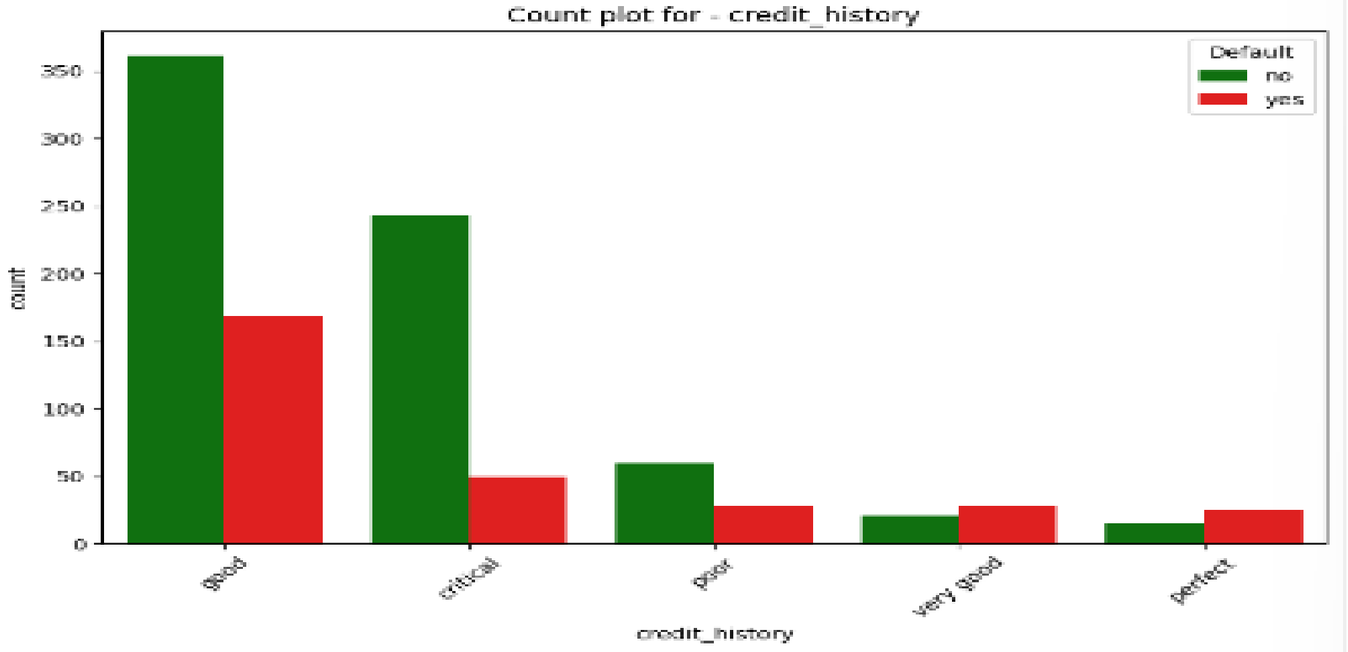


Exploratory Data Analysis

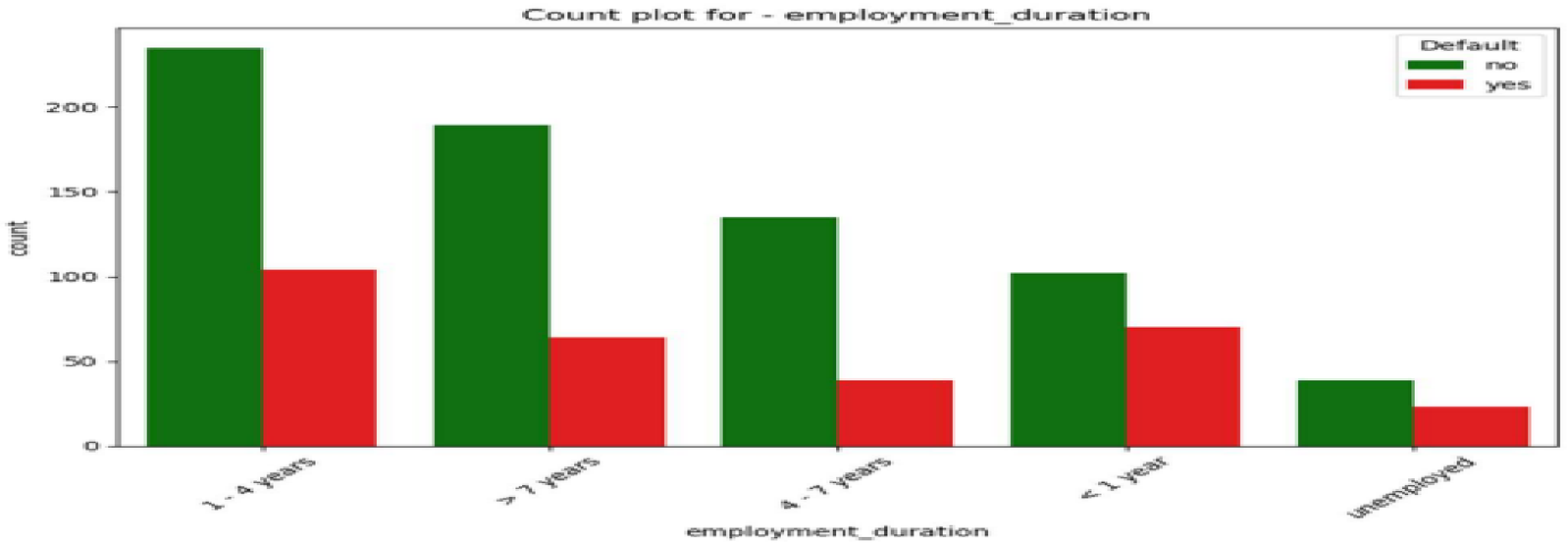
The 'checking balance' indicates the funds available in a customer's checking account for daily financial activities, representing a highly liquid part of their finances. Data visualization shows that default rates rise as the balance in this account decreases.

A significant number of customers with 'very good' and 'perfect' credit history status default on loans, hinting at a possible necessity to review the credit rating system's customer categorization. Although the actual count of such cases is small, the proportion of defaults is notably high, calling for further examination. It's important to consider that the dataset comprises only 1000 observations, and a larger dataset could offer more clarity on this matter.

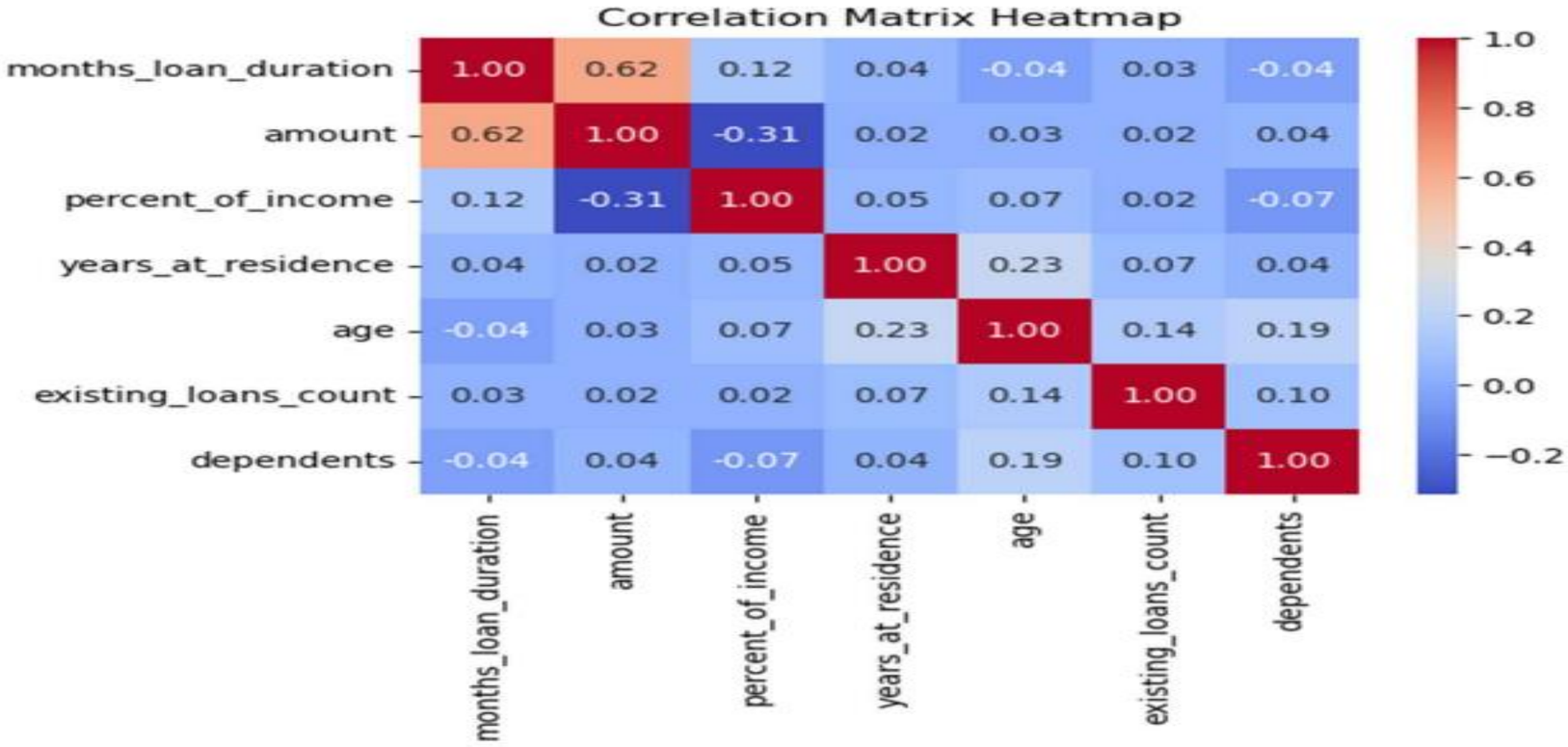




Exploratory Data Analysis



A quick look at the graph shows that individuals with longer employment histories tend to have fewer defaulters. This observation suggests that longer-tenured employees are more financially stable and responsible, leading to better loan repayment compliance. However, a detailed analysis is necessary to confirm and draw conclusive insights from this observation.



Overall, the correlation matrix, heatmap, and pair plot indicate a weak relationship among the numerical predictors, with some exceptions. The pair plot, which includes scatterplots (not shown here), visually confirms these weak correlations among the numerical features.

To summarize, the EDA phase yielded important insights into the dataset's characteristics and behavior, revealing potential relationships between features and default likelihood. It also highlighted key variables essential for predicting loan defaults, laying a solid groundwork for the following stages of machine learning model development and assessment.

Results

Model Building

During data preprocessing, nominal categorical features were encoded using one-hot encoding, ordinal features were ranked using scikit-learn, and standardization ensured uniform scaling. The dataset was then split into training and testing sets in a 75-25 ratio, with stratification to preserve class distribution in the imbalanced dataset.

Model Training and Hyperparameter Tuning:

I examined different supervised machine learning models like logistic regression, k-nearest neighbors, random forest, gradient boosting and AdaBoost,for predicting defaults. Hyperparameter tuning was done using GridSearchCV and cross-validation on the training set, focusing on optimizing 'recall' to reduce false negatives and identify potential defaulters accurately.

Model Fitting and Evaluation:

After hyperparameter tuning, I utilized the optimal settings for all models on the full training set and evaluated their performance on both training and testing data. The classification report provided insights into metrics such as precision, recall, and F1-score per class, while visualizing the confusion matrix helped interpret the model's predictions in terms of true positives, true negatives, false positives, and false negatives.

Results

Model Building(Confusion Matrix and Hyper tuning)

Outcomes : i. True Positive : Predicted values are default, here outcome is really default
ii. True Negative : Predicted values are not default, here outcome is not default
iii. False Positive : Predicted values are default, but outcome is not default
iv. False Negative : Predicted values are not default, but outcome is default

Accuracy : to measure how well a classification model correctly predicts the class labels of a dataset

Accuracy: (TP + TN) / (TP + TN + FP + FN)

Sensitivity : Sensitivity, also known as Recall, is a performance metric that measures the ability of a classification model to correctly identify positive instances from the total actual positive instances in a dataset.

Sensitivity = TP / (TP + FN)

Specificity : Specificity, also known as Precision, is a performance metric that measures the ability of a classification model to correctly identify negative instances from the total actual negative instances in a dataset.

Specificity = TN / (TN + FP)

F-measure : The F-measure, also known as the F1 score, is a performance metric that combines both precision and recall into a single value.

F-measure = 2 * (Precision * Recall) / (Precision + Recall)

AUC (Area Under the Curve) : It is a performance metric to evaluate the quality of a binary classification model based on its receiver operating characteristic (ROC) curve.
The AUC value ranges from 0 to 1, where:
For example, when AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.

Model Hyper Tuning:

Grid search is a technique used in machine learning to find the optimal hyperparameters for a given model. Hyperparameters are parameters that are not learned directly from the data, but rather set by the user before training the model.

Grid Search CV helps to avoid overfitting to a specific validation set and provides a more reliable estimation of a model's performance. By considering multiple combinations of hyperparameters and evaluating them on different subsets of data, it can help identify the best hyperparameter configuration that generalizes well to unseen data.

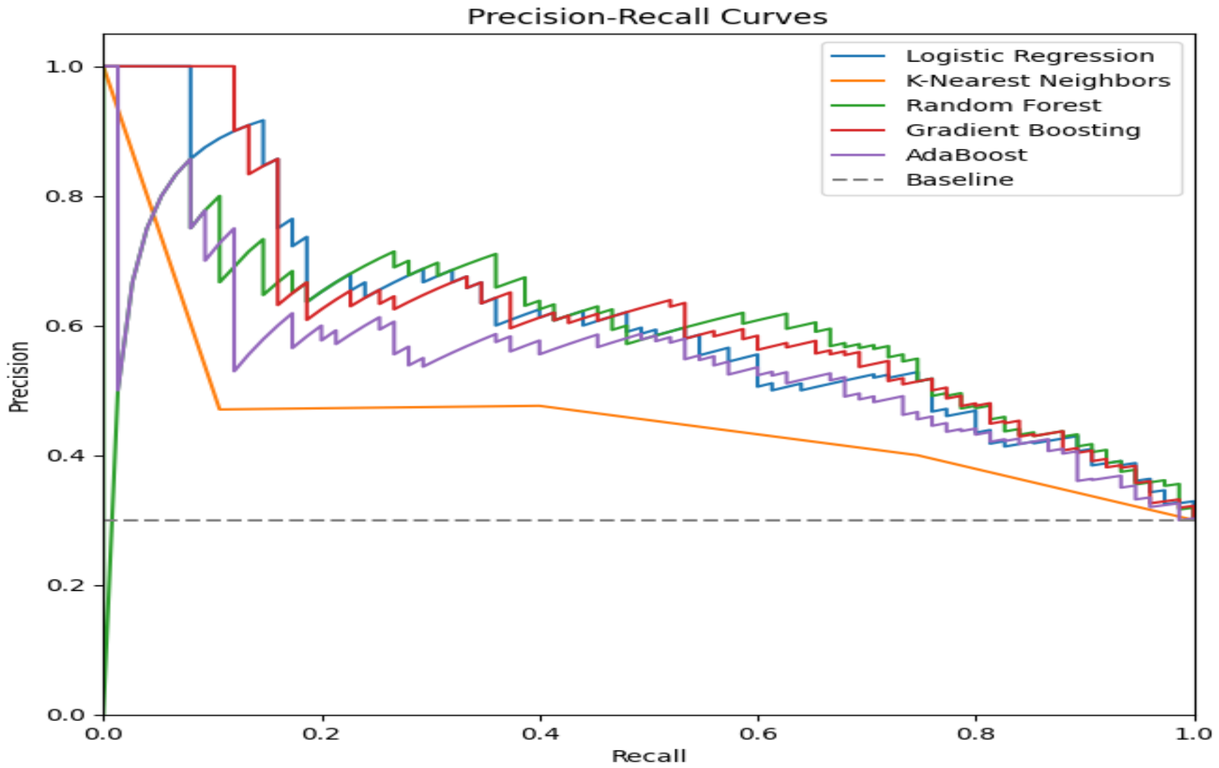
It's important to note that Grid Search CV can still be computationally expensive, especially if the hyperparameter space is large or the dataset is large. However, it provides a more robust and accurate way of tuning hyperparameters compared to a simple validation set approach.

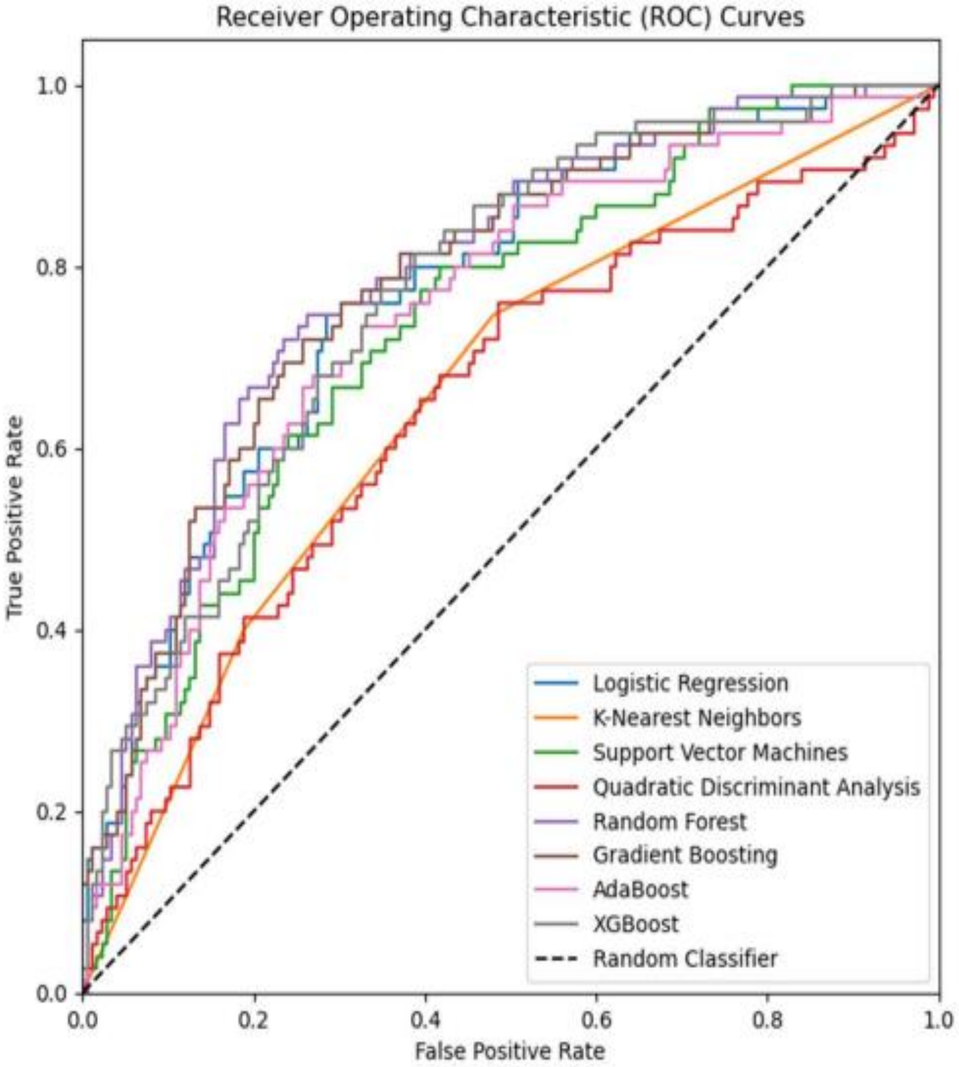
Results

Model Building

Model Comparison and Selection:

I used precision-recall (PR) curves and their AUC values to compare models and select the final one. Due to the dataset's imbalance, the main performance metric was AUC-PR, where higher values indicated better performance. The chosen model was adjusted with a custom threshold to improve recall. These evaluations highlighted the balance between precision and recall in loan default prediction.





Discussion

Insight of the various Models used

The following are the ML Models in Order of Performance (Based on PR-AUC) from lowest to highest, obtained for training on the German bank loan dataset:

1. K-Nearest Neighbors
2. AdaBoost
3 Random Forest
4. Logistic Regression
5. Gradient Boosting

Let's review the project results using PR-AUC scores, crucial for imbalanced data like loan defaults, indicating better identification of default cases with higher PR-AUC values.

The Dataset has “default” = 1000 rows in which it has 700 rows as “no” (0) and rest 300 rows as “yes”(1).

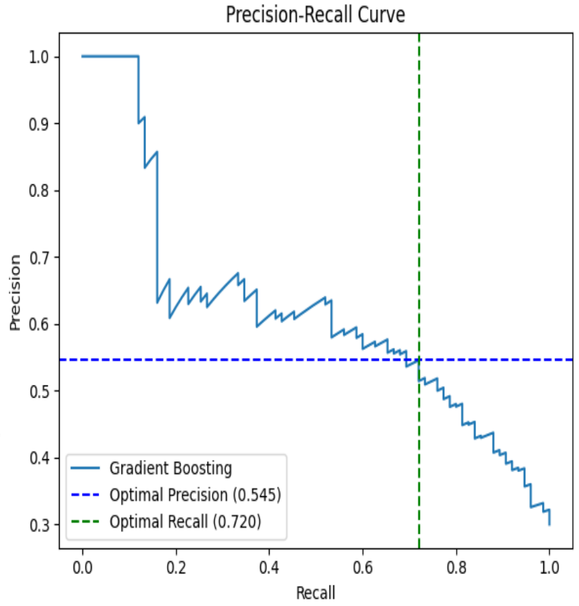
Model	AUC-PR
Gradient Boosting	0.617257
Logistic Regression	0.607566
Random Forest	0.587175
AdaBoost	0.539437
K-Nearest Neighbors	0.457832

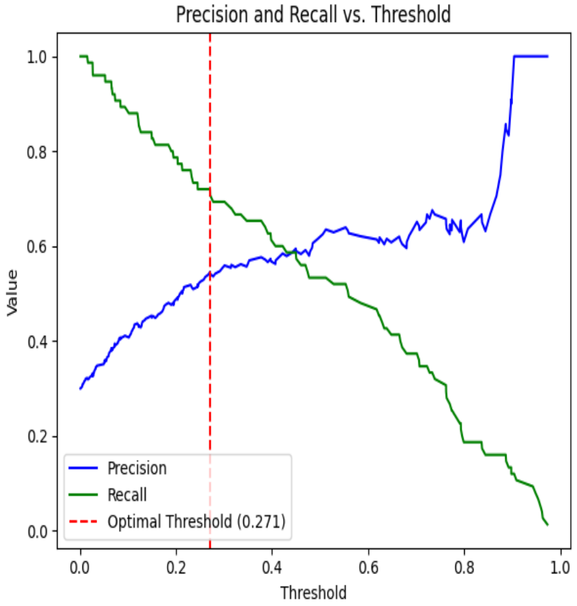
Discussion

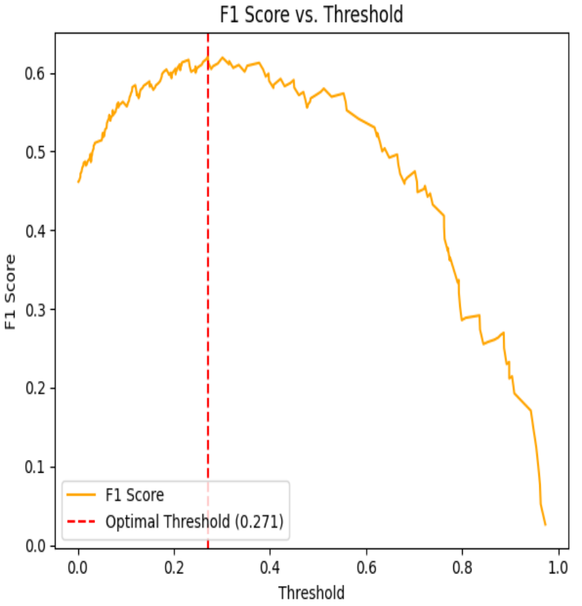
Insight of the various Models used

Low to Moderate performing model:

K-Nearest Neighbors with AUC-PR score of 0.4578
Adaboost with AUC-PR score of 0.539437
Random Forest with AUC-PR Score of 0.587175





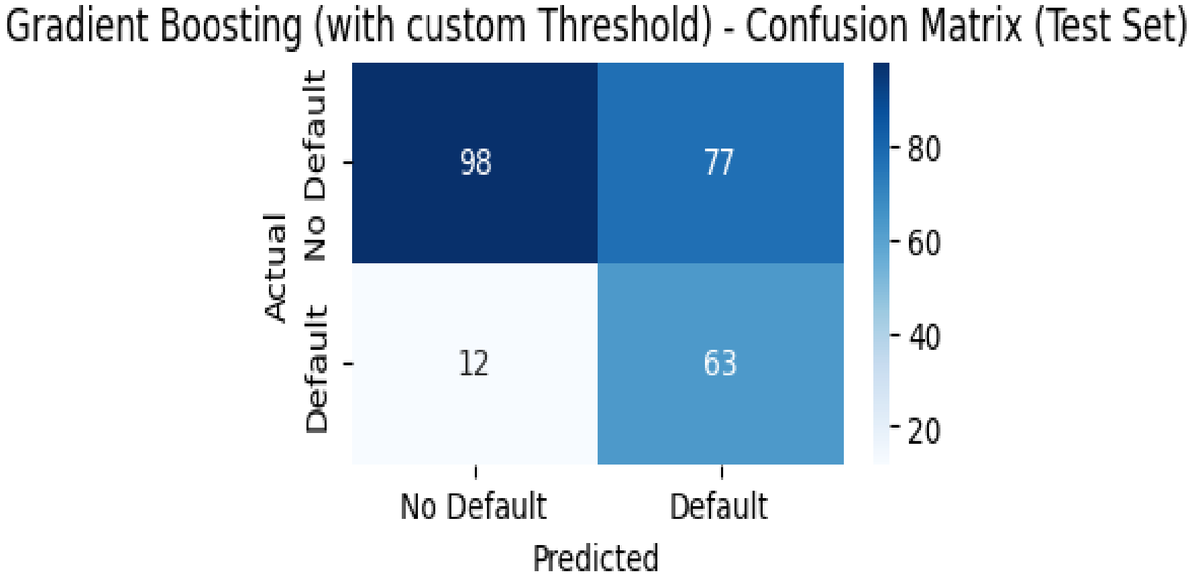


Best performing model:

Both Gradient Boosting models, with and without a custom threshold, demonstrated outstanding performance with a PR-AUC score of 0.620, excelling in high recall while maintaining good precision. This makes Gradient Boosting a robust option for identifying default cases, crucial for minimizing false negatives and managing the bank's risk effectively.

Using a custom threshold of '0.14' significantly boosts Gradient Boosting's ability to detect defaults, aligning with the bank's goal. It achieves impressive PR-AUC scores, crucial for minimizing false negatives and accurately identifying defaulters, making it the best choice for maximizing default case identification while managing precision.

there are several opportunities for improvement and expansion in this project. These possibilities include exploring feature engineering techniques, employing ensemble methods tailored for large datasets, and adopting advanced techniques for handling imbalanced datasets.



Discussion

Insight of the various Models used

Limitations and Future Directions:

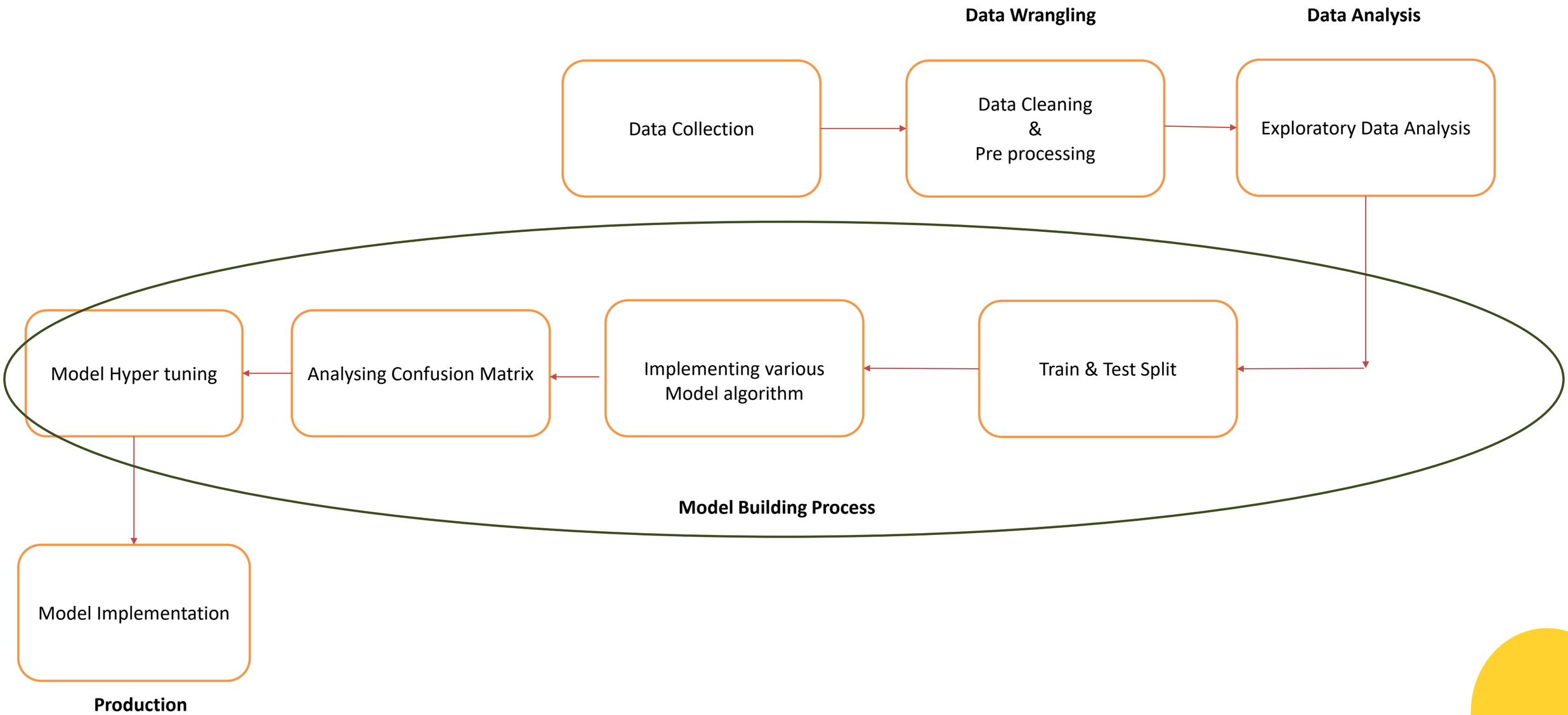
The dataset's small size of 1000 samples may limit model generalization, suggesting a need for a more extensive dataset covering diverse customer profiles and economic conditions. Addressing class imbalance is crucial to avoid bias and improve model performance, possibly through oversampling or undersampling techniques and custom thresholds. Enhancing feature engineering by deriving new features or integrating external data could reveal hidden patterns and improve predictions, like debt-to-income ratios or economic indicators. Further hyperparameter tuning and experimenting with advanced techniques like neural networks could enhance predictive capabilities. While complex models like Gradient Boosting offer high performance, simpler models like Logistic Regression provide interpretability, aiding in understanding features' contributions. Integrating external data like economic indicators or customer behavior data can enrich model predictions, such as incorporating unemployment rates for contextual default patterns..

Conclusion:

The project aimed to build a predictive model for loan defaults using data from a German bank. The systematical approach of exploring and visualization of the data and then proceed to apply various machine learning algorithms. From which we prioritized Gradient Boosting as the most promising model for accurate loan default prediction. As it consistently demonstrated efficient performance in terms of AUC-PR and Recall.

However there are few limitations like a small dataset and class imbalance. Despite these limitations, the above insights provides a significant value of advanced machine learning in financial risk assessment.

Machine Learning Modelling Steps Followed :



THANK YOU