Credit Loan Default Detection Report

Deep Patel, Vincent Lee, Wen-Chi Tseng, Yilong Ruan

December 8th, 2023

BANA 212 Data & Programming Analytics

Group 6A

Table of Contents

Table of Contents 2

1. Executive Summary 3

2. Introduction 4

2.1 Business Idea: Credit Loan Default Detection 4

3. Data Summary, Description, Visualization 6

3.1 Dataset Summary 6

3.2 Dataset Features Description 7

3.3 Data Visualization 8

3.3.1 Numerical Columns 8

3.3.2 Categorical Columns 9

3.3.3 Cleaned Dataset Visualization 10

3.3.4 Target Feature Visualization 11

3.3.5 Correlation of Features 12

4. Analysis Method 13

4.1 Models Selection 13

4.1.1 Multinomial Naive Bayes 13

4.1.2 Logistic Regression 13

4.1.3 Random Forest 13

4.1.4 XGBoost 14

4.2 Necessary Data Pre-Processing 14

4.3 Training and Testing Data 15

4.4 Visualization of Benchmark Results 15

5. Improving XGBoost 17

5.1 XGBoost Benchmark Model Evaluation 17

5.2 Improving XGBoost Model Through Pre-Processing 17

5.3 Cost Matrix Evaluation 19

6. Conclusion 21

1. Executive Summary

In today's dynamic financial landscape, the assessment of credit risk stands out as a paramount concern for lenders and financial institutions. Our “Credit Loan Default Detection” analysis report is dedicated to leveraging a comprehensive set of parameters to gauge the likelihood of loan default among applicants. Our dataset encompasses critical factors such as age, income, employment history, credit history, etc., enabling us to make precise decisions. To ensure the data is suitable for advanced machine-learning models, we utilized pre-processing, such as data cleaning and imputation for missing values. Subsequently, we harnessed the power of data visualization to reveal valuable insights about our dataset structure and features. In our pursuit of making an accurate prediction, we employed various models, including Random Forest and Naive Bayes, to deliver actionable results. Our project results in a robust evaluation of model accuracy, providing lenders and financial institutions with a reliable tool to enhance their credit risk assessment processes.

2. Introduction

2.1 Business Idea: Credit Loan Default Detection

The business idea of "detecting loan or credit default individuals" centers around the development and implementation of systems, strategies, and tools to assess the creditworthiness of individuals or businesses applying for loans or credit. This idea is important because it addresses the critical need for lenders and financial institutions to make informed decisions when extending credit. Here's a more detailed explanation of this business concept:

- **Credit Assessment Services:** The core of this business involves offering credit assessment services to lenders. These services may include credit scoring, risk profiling, and predictive modeling to evaluate the likelihood of borrowers defaulting on their loans. Businesses providing such services employ sophisticated data analytics, machine learning, and statistical models to make accurate predictions.
- **Data Analysis and Modeling:** The heart of this business is data-driven. It involves collecting and analyzing vast datasets, which may include information about an applicant's financial history, income, employment status, assets, liabilities, and more. Advanced algorithms are then used to create credit risk models that assess an applicant's creditworthiness.
- **Risk Mitigation:** The primary objective is to help lenders mitigate the risks associated with lending. By identifying potential defaulters or high-risk applicants, lenders can make informed decisions about whether to approve a loan, set interest rates, or impose conditions. This reduces the likelihood of financial losses due to loan defaults.
- **Customized Solutions:** Businesses in this field often tailor their solutions to different types of lending, such as consumer loans, mortgages, small business loans, or corporate credit. Customization ensures that lenders receive insights and risk assessments specific to their target markets.
- **Technological Innovation:** To remain competitive, businesses in this domain continually innovate. They leverage advancements in technology, such as artificial intelligence and

big data analytics, to improve the accuracy of credit assessments and adapt to changing market dynamics.

- **Value Proposition:** The key value proposition is to provide lenders with the ability to make better lending decisions, reduce default rates, improve profitability, and maintain a healthier loan portfolio. This, in turn, fosters trust with borrowers and supports the overall financial stability of the lending institution.

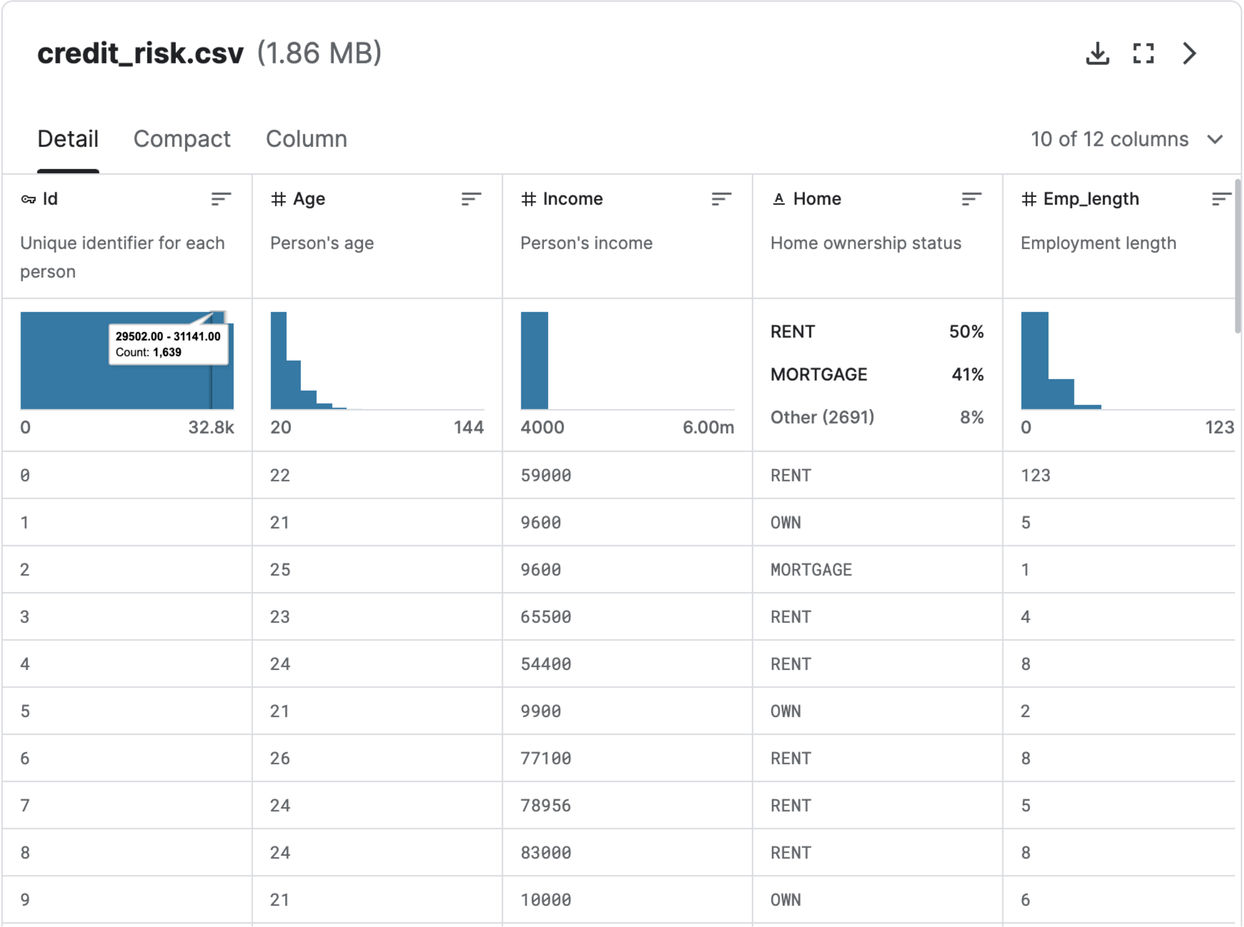
In summary, the business idea of detecting loan or credit default individuals revolves around using data-driven methods to assess and predict credit risk, ultimately helping lenders make informed lending decisions while reducing the likelihood of loan defaults. It is a critical and evolving field that plays a pivotal role in the financial industry's stability and responsible lending practices.

3. Data Summary, Description, Visualization

3.1 Dataset Summary

This is a credit risk analysis dataset we found on Kaggle. This dataset provides essential information about loan applicants and their characteristics, including their loan rate, income, age, credit length, etc. This dataset provides a simplified view of the factors contributing to credit risk, presenting an excellent opportunity for us to apply our machine learning analysis in determining whether a loan applicant is likely to default.

Data Source: https://www.kaggle.com/datasets/nanditapore/credit-risk-analysis/data



credit_risk.csv (1.86 MB)

Id	# Age	# Income	Home	# Emp_length
Unique identifier for each person	Person's age	Person's income	Home ownership status	Employment length
29502.00 - 31141.00 Count: 1,639 0 32.8k	20 144	4000 6.00m	RENT 50% MORTGAGE 41% Other (2691) 8%	0 123
0	22	59000	RENT	123
1	21	9600	OWN	5
2	25	9600	MORTGAGE	1
3	23	65500	RENT	4
4	24	54400	RENT	8
5	21	9900	OWN	2
6	26	77100	RENT	8
7	24	78956	RENT	5
8	24	83000	RENT	8
9	21	10000	OWN	6

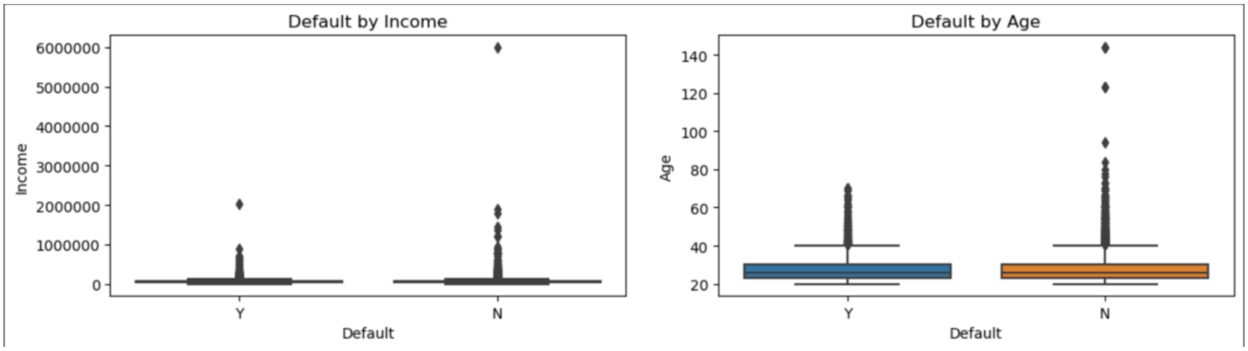
3.2 Dataset Features Description

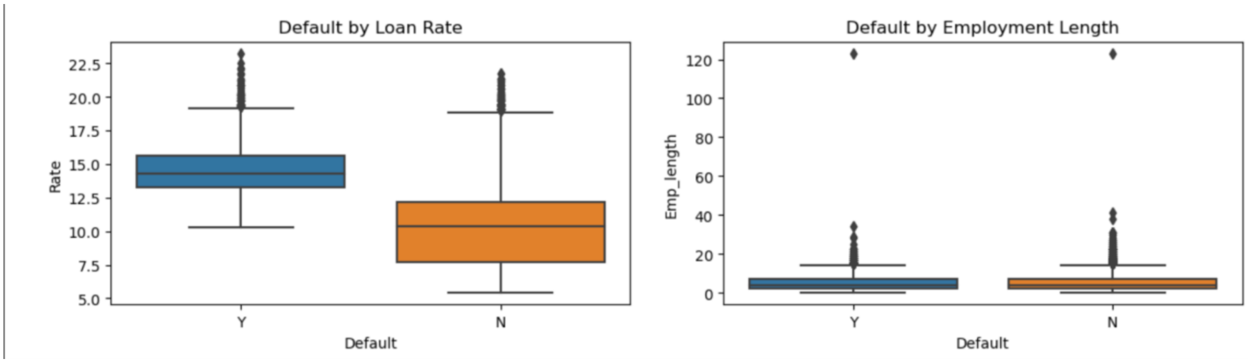
- **ID**: Unique identifier for each loan applicant.
- **Age**: Age of the loan applicant.
- **Income**: Income of the loan applicant.
- **Home**: Home ownership status (Own, Mortgage, Rent).
- **Emp_Length**: Employment length in years.
- **Intent**: Purpose of the loan (e.g., education, home improvement).
- **Amount**: Loan amount applied for.
- **Rate**: Interest rate on the loan.
- **Status**: Loan approval status (Fully Paid, Charged Off, Current).
- **Percent_Income**: Loan amount as a percentage of income.
- **Default**: Whether the applicant has defaulted on a loan previously (Yes, No).
- **Cred_Length**: Length of the applicant's credit history.

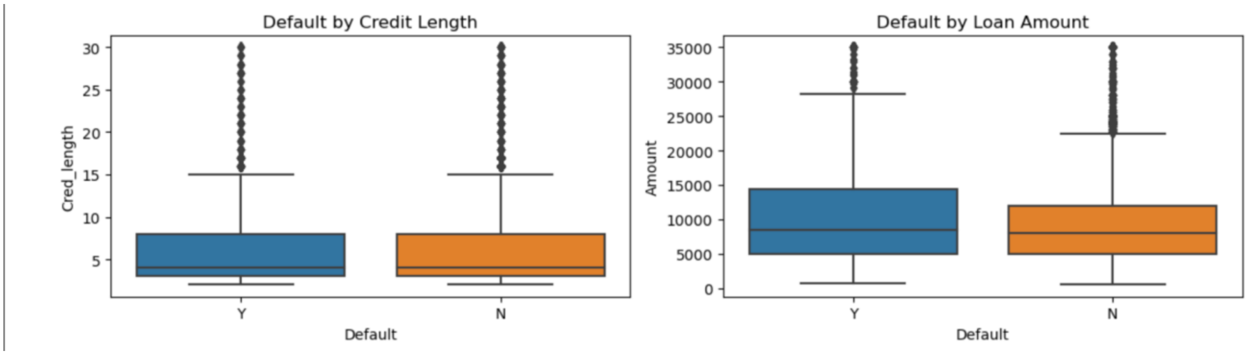
3.3 Data Visualization

3.3.1 Numerical Columns

We visualized our numerical columns using boxplots to identify any patterns or significant details in the data. As shown, the income, age, and employment length columns have an extreme case of outliers. We also noticed that the number of people who default or don’t default have the same distributions except for the loan rate, which has more people defaulting at an average rate of around 14%. The overall pattern of our numerical columns is that there is a tendency for right-skewness or a lot of outliers in the higher value ranges.

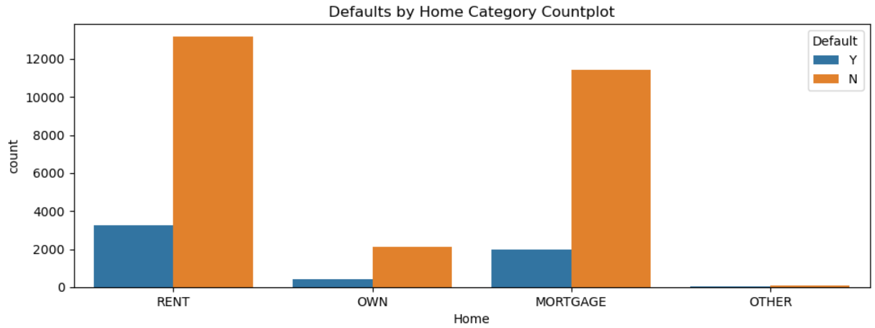


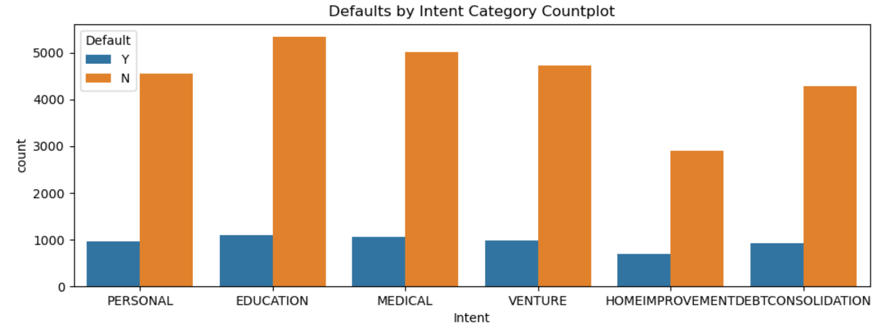


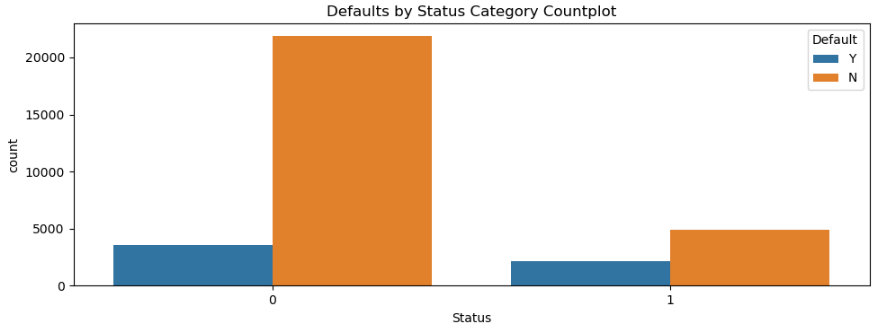


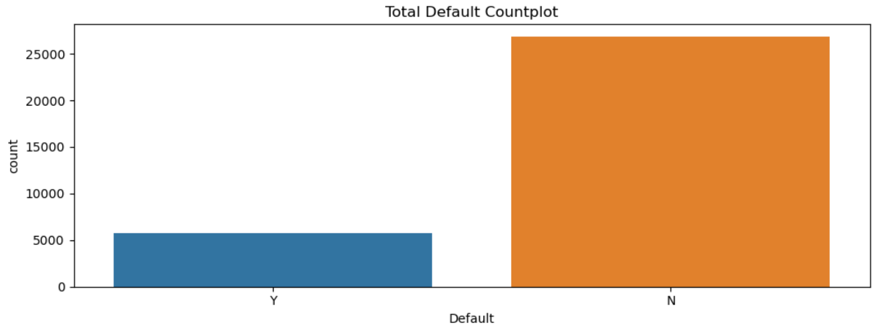
3.3.2 Categorical Columns

We used count plots to visualize our categorical columns and how they affect the number of defaults. Since our dataset contains substantially more amounts of people who didn’t default compared to people who did, the graphs depict an accurate representation. However, it is interesting to note that people who default tend to not own a home compared to those who didn’t default.



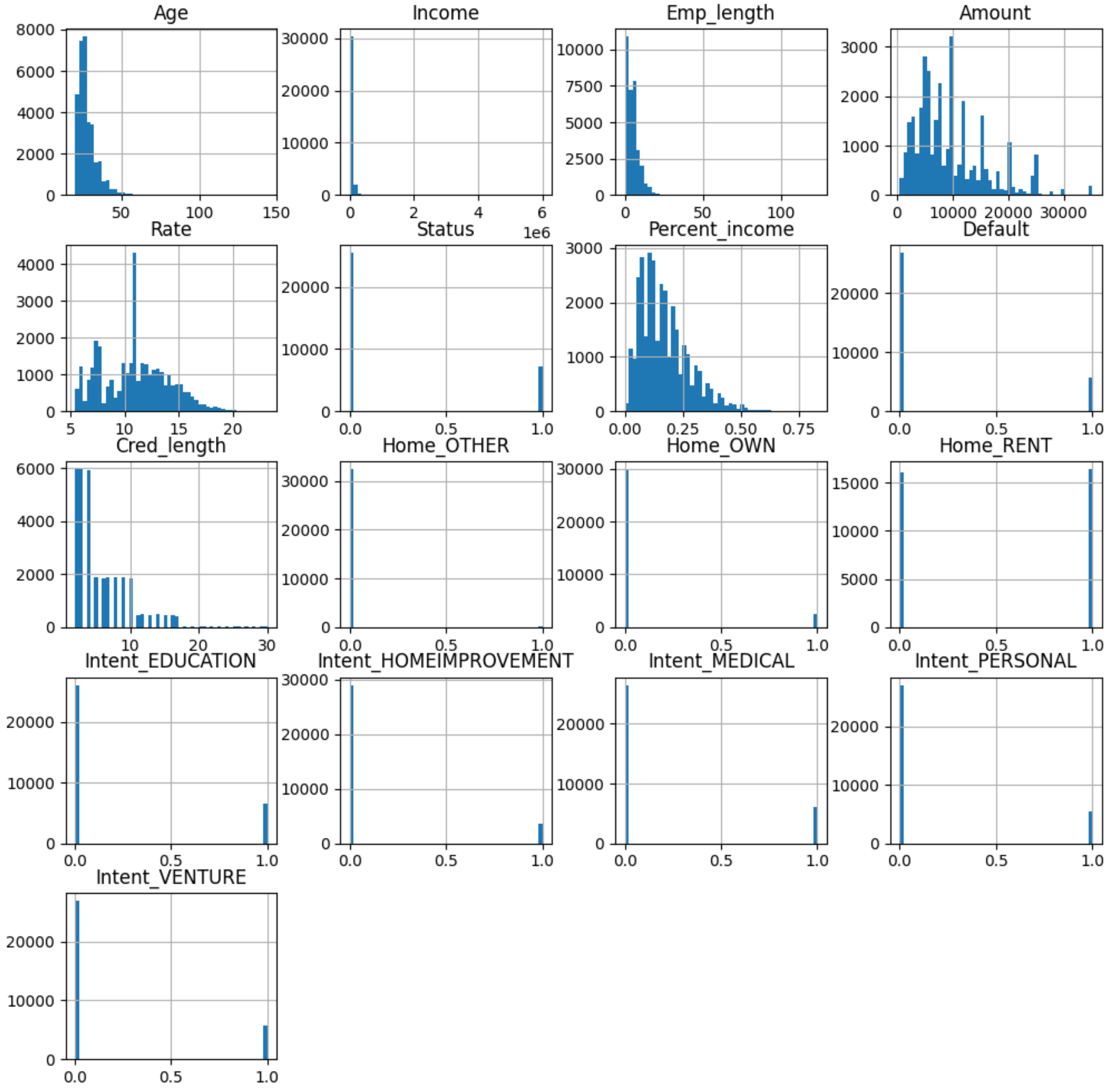






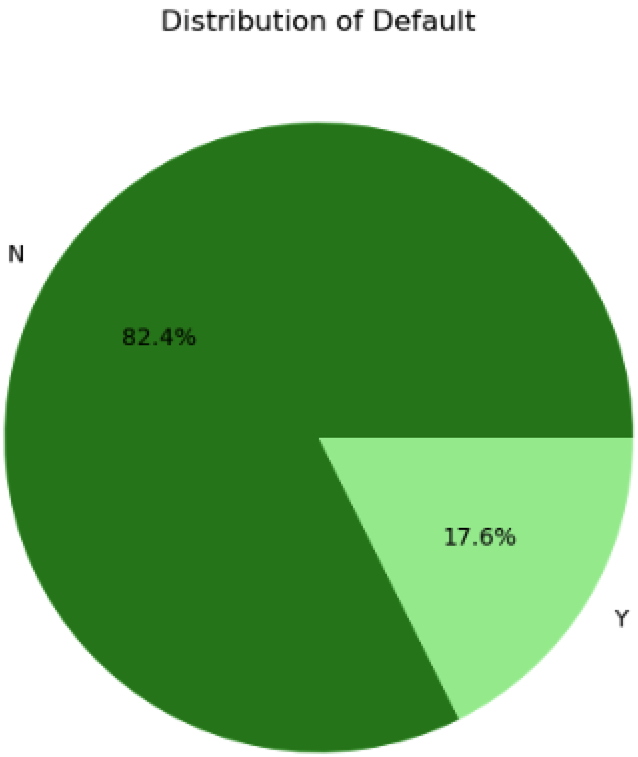
3.3.3 Cleaned Dataset Visualization

The charts represent our final cleaned data after handling missing values and transforming the categorical columns by using dummy variables.



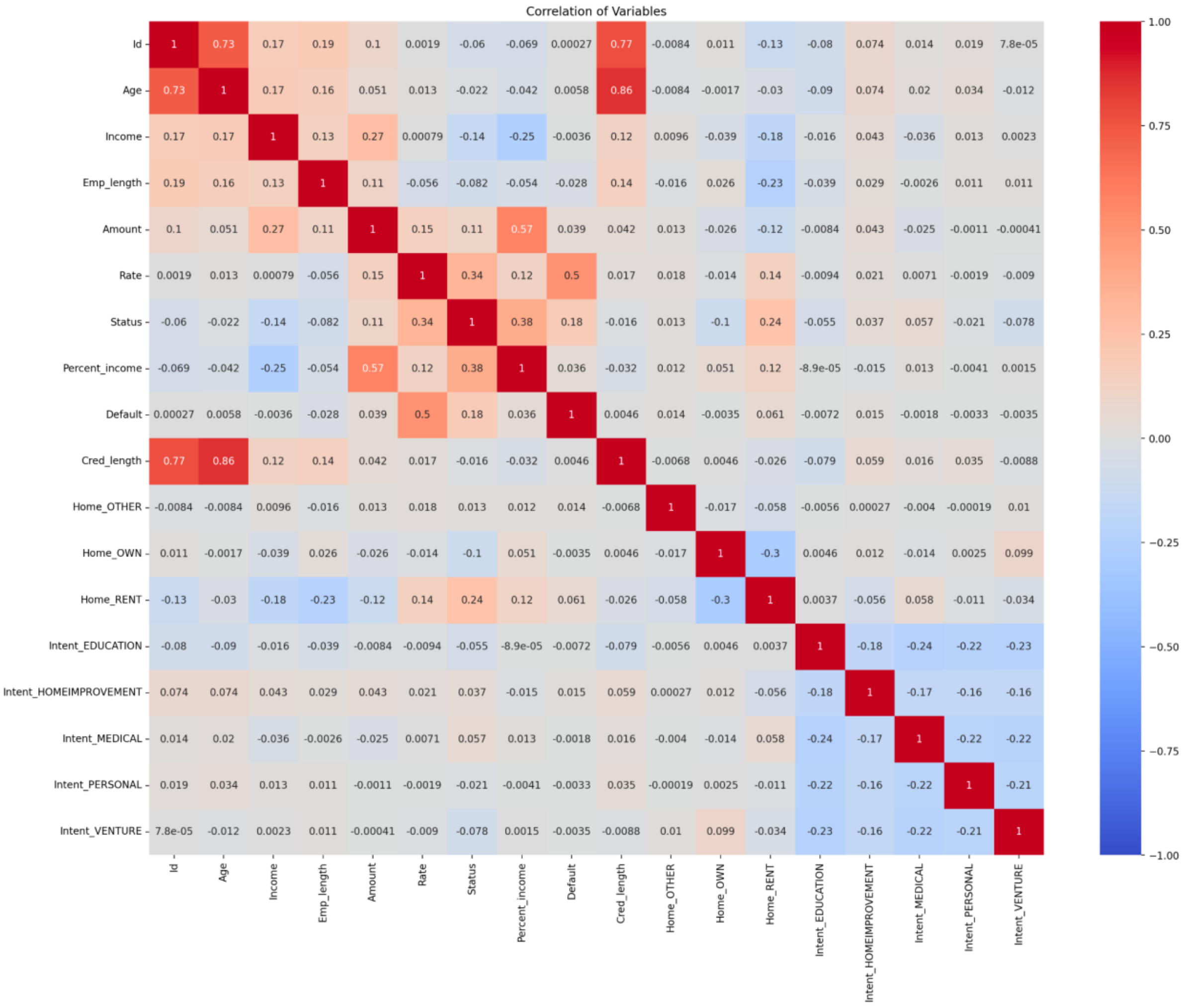
3.3.4 Target Feature Visualization

Our target feature “Default” is visualized in a simple pie graph below. This chart highlights that there is a significant class imbalance in our target feature column because there are significantly more “not default” values than “yes default” values. This will be an important factor to take into consideration as we build our prediction model.



3.3.5 Correlation of Features

The final graph visualizes the correlations of all our features. Using this graph, we can identify if any features are too closely correlated and may be unnecessary to include both in a model. The only two variables with a considerable amount of correlation are age and credit length. However, since we are interested in predicting the “Default” feature, it is important to note that the loan rate feature is the most correlated with default at a correlation coefficient of 0.5.



4. Analysis Method

The type of analysis we chose to implement was using machine learning models to predict the probability of credit default on our dataset. This analysis will aid significantly in the context of our dataset, as machine learning is an essential technology in predicting credit defaults in the real world. Out of the various models available, we decided to test our dataset on the models Naive Bayes, Logistic Regression, Random Forest, and XGBoost. We provide our reasonings for choosing these models below.

4.1 Models Selection

4.1.1 Multinomial Naive Bayes

This classifier incorporates the concept of conditional probability, which refers to the probability of event A occurring under the condition that event B has already occurred. We chose this classifier for its straightforward interpretation and computational efficiency in training our data. Given the presence of three categorical features in our dataset, Multinomial Naive Bayes was preferred over alternatives such as Gaussian Naive Bayes for our training purposes.

4.1.2 Logistic Regression

The logistic regression model is a fundamental component of our project's machine-learning framework and plays a pivotal role in predicting binary outcomes. This model is well-known for its simplicity and fast processing time, which makes it compelling to experiment with for our dataset. Logistic regression is also very suitable for binary classification predictions, which is true for our dataset as we are only predicting “0” (not default”) or “1” (yes default) class values.

4.1.3 Random Forest

Random Forest's strength comes from its unique approach of creating a 'forest' of decision trees, each made from a random selection of data points and features. This variety allows each tree to be different, making the group decision stronger and more reliable than any single tree's guess. It's like getting several opinions before making a big decision — the more advice you get, the better your decision is likely to be.

In the context of predicting loan defaults, this means that the Random Forest considers a multitude of scenarios and patterns before concluding. Each tree is trained on a random sample of the data, ensuring that different aspects of the data are highlighted. This randomness helps the model remain unbiased and identify complex interactions between variables that might indicate a risk of default. Additionally, the model can handle a large variety of data types with ease. It doesn't get overwhelmed by the volume or complexity of the data, making it particularly suitable for the multifaceted nature of financial information. Also offers an effective tool that balances sophistication with practical insight, thereby helping to uphold the financial integrity and stability of the lending process.

4.1.4 XGBoost

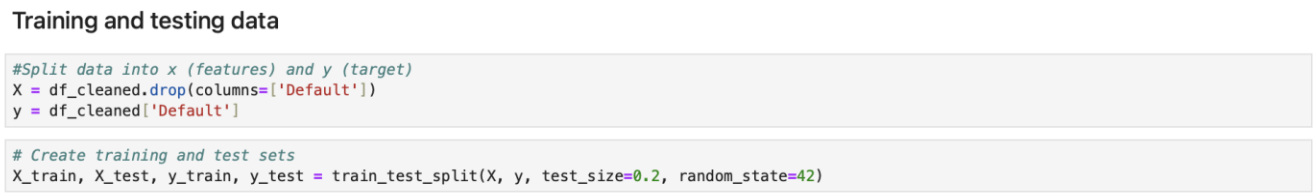
The XGBoost model is well known for its fast processing time and versatility for various datasets. It is an ensemble machine-learning method released in 2014 as an improved version of the classic gradient-boosting model. Its powerful performance was what attracted many data scientists to use this model, and especially in the 2010s many data science competitions were won by implementing an XGBoost model. This model uses decision trees as its base learner and will generally provide accurate predictions due to its ensemble method. In the context of our dataset, we have implemented XGBoost due to its popularity in data science competitions and its overall high-rated performance.

4.2 Necessary Data Pre-Processing

Before we can begin running our benchmark models, several essential data pre-processing steps must be taken for the model to run at its bare minimum. We dropped the “Id” column as it provided no valuable information to our models. Second, we filled in the missing values in our dataset with the mean of its column using the Simple Imputer package. Lastly, we transformed our categorical variables using the “pd.get_dummies” function to generate dummy variable columns. These are all the necessary steps required for us to proceed with testing our benchmark models.

4.3 Training and Testing Data

Before running the models, we first split the data into 20% testing and 80% training. We then set the random state to 42 to make sure all of our models are run in the same environment.



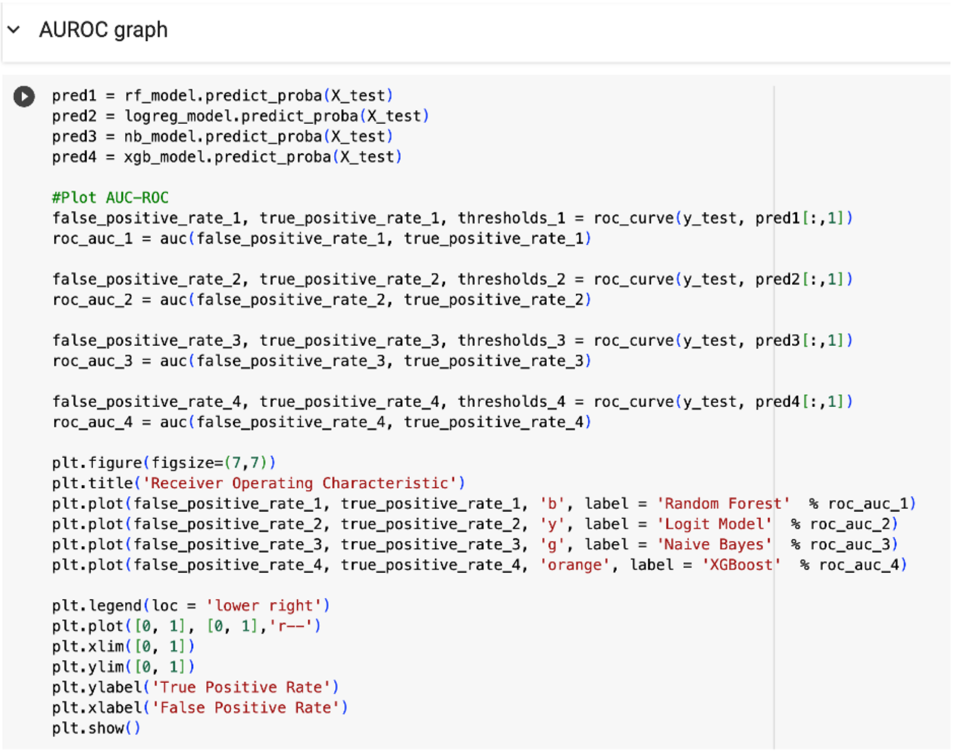
Training and testing data

```
#Split data into x (features) and y (target)
X = df_cleaned.drop(columns=['Default'])
y = df_cleaned['Default']
```

```
# Create training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4.4 Visualization of Benchmark Results

In this section, we present a visual assessment of the benchmark performance of our models. We evaluate the models using the Area Under the Receiver Operating Characteristic (AUC-ROC) curve, a metric commonly used to gauge the quality of binary classification models.



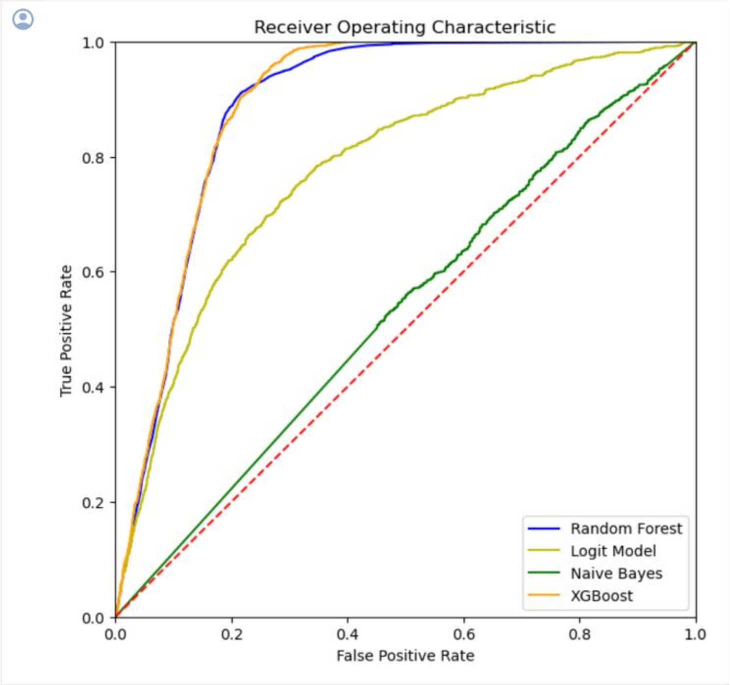
AUROC graph

```
pred1 = rf_model.predict_proba(X_test)
pred2 = logreg_model.predict_proba(X_test)
pred3 = nb_model.predict_proba(X_test)
pred4 = xgb_model.predict_proba(X_test)

#Plot AUC-ROC
false_positive_rate_1, true_positive_rate_1, thresholds_1 = roc_curve(y_test, pred1[:,1])
roc_auc_1 = auc(false_positive_rate_1, true_positive_rate_1)

false_positive_rate_2, true_positive_rate_2, thresholds_2 = roc_curve(y_test, pred2[:,1])
roc_auc_2 = auc(false_positive_rate_2, true_positive_rate_2)

false_positive_rate_3, true_positive_rate_3, thresholds_3 = roc_curve(y_test, pred3[:,1])
roc_auc_3 = auc(false_positive_rate_3, true_positive_rate_3)

false_positive_rate_4, true_positive_rate_4, thresholds_4 = roc_curve(y_test, pred4[:,1])
roc_auc_4 = auc(false_positive_rate_4, true_positive_rate_4)

plt.figure(figsize=(7,7))
plt.title('Receiver Operating Characteristic')
plt.plot(false_positive_rate_1, true_positive_rate_1, 'b', label = 'Random Forest'  % roc_auc_1)
plt.plot(false_positive_rate_2, true_positive_rate_2, 'y', label = 'Logit Model'  % roc_auc_2)
plt.plot(false_positive_rate_3, true_positive_rate_3, 'g', label = 'Naive Bayes'  % roc_auc_3)
plt.plot(false_positive_rate_4, true_positive_rate_4, 'orange', label = 'XGBoost'  % roc_auc_4)

plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

Code Description:

- We begin by predicting class probabilities for each model (**rf_model**, **logreg_model**, **nb_model**, and **xgb_model**) on the test dataset.

- Next, we compute the Receiver Operating Characteristic (ROC) curve and calculate the AUC (Area Under the Curve) for each model. The ROC curve visualizes the trade-off between true positive rate and false positive rate for different classification thresholds.
- The AUC value serves as a summary metric, indicating the overall performance of each model. Higher AUC values are indicative of better discrimination between positive and negative classes.
- Finally, we create a comprehensive ROC curve plot to visually compare the performance of all four models. Each curve is labeled with the model's name and its corresponding AUC score.

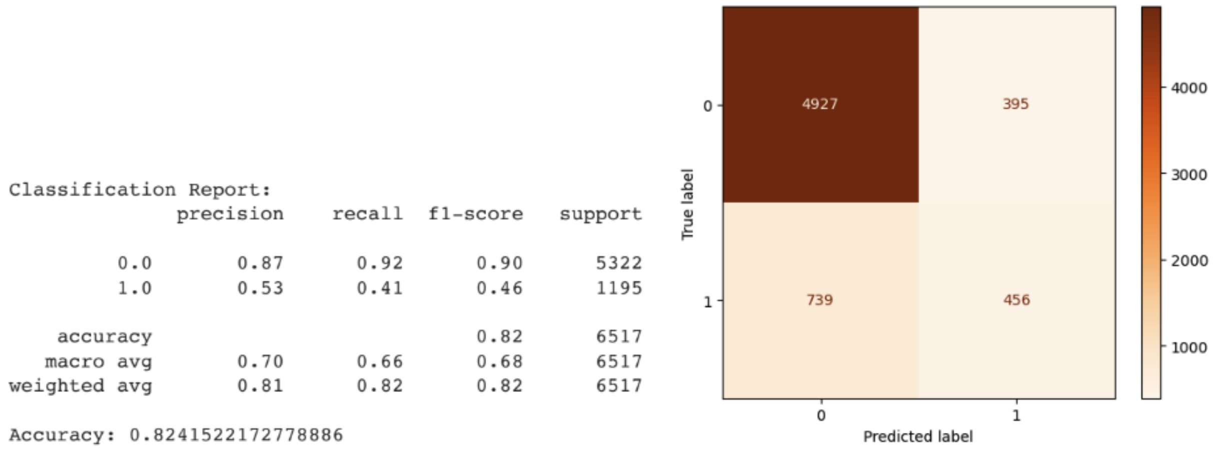


This visualization allows for a direct and intuitive comparison of the models' abilities to distinguish between classes, providing valuable insights into their relative performance. As shown from our graph, XGBoost performed the best out of the four models we tested. Therefore, we will proceed with focusing on fine-tuning our XGBoost model to create our final credit default risk prediction model.

5. Improving XGBoost

5.1 XGBoost Benchmark Model Evaluation

Our benchmark model produced an accuracy of 0.824, making it the most overall accurate benchmark model in our analysis.

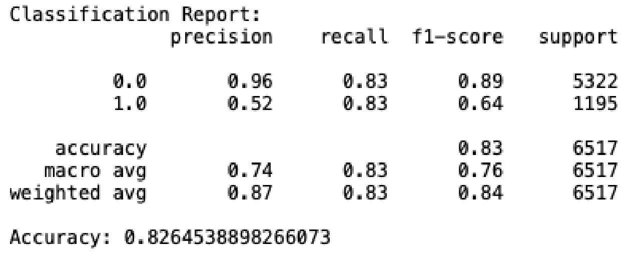


```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.87      0.92      0.90      5322
         1.0       0.53      0.41      0.46      1195

    accuracy                           0.82      6517
   macro avg       0.70      0.66      0.68      6517
weighted avg       0.81      0.82      0.82      6517

Accuracy: 0.8241522172778886
```

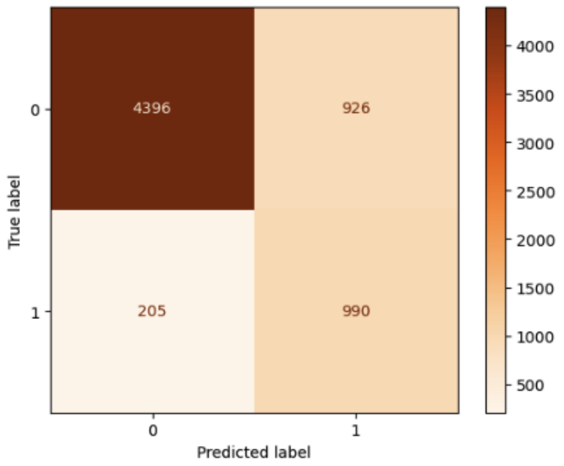
Our benchmark model provided an accuracy of 0.82, which proved to be our best benchmark model accuracy. The class variable proportions are distributed with the “0” class, or “Not Default”, having 5322 predictions, and the “1” class, or “Default”, having 1195 predictions.

5.2 Improving XGBoost Model Through Pre-Processing

SMOTE, Standardizing, Parameter Tuning: Several pre-processing steps were taken in an attempt to improve our model and particularly increase the accuracy of predicting the “1” class. First, the training data was resampled using SMOTE. After resampling the data, we used StandardScaler() to standardize the training data to decrease the impact of outliers. We also tuned the “scale_pos_weight” parameter in the model code to adjust for the class imbalance as there are significantly more “0” class values than “1” class values. After fine-tuning the parameters, the most optimal scale_pos_weight value was 3. In total, 3 pre-processing steps were taken to enhance our XGBoost model: resampling the data through SMOTE, standardizing the resampled data, and fine-tuning the “scale_pos_weight” parameter. Our results below only slightly increased our overall accuracy, but decreased our precision in predicting the “1” class. More importantly, the amount of false negative predictions was lowered.



```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.96      0.83      0.89      5322
         1.0       0.52      0.83      0.64      1195

    accuracy                           0.83      6517
   macro avg       0.74      0.83      0.76      6517
weighted avg       0.87      0.83      0.84      6517

Accuracy: 0.8264538898266073
```

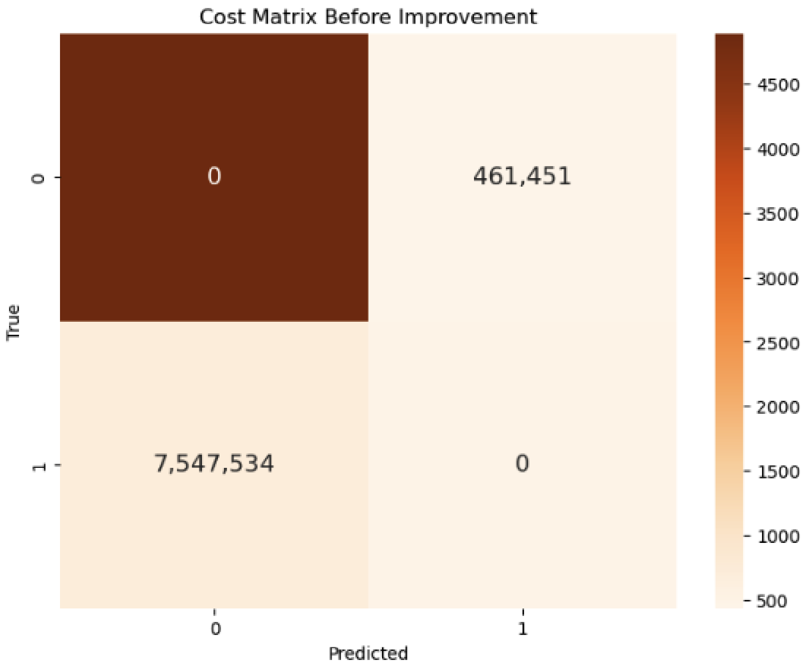


5.3 Cost Matrix Evaluation

In the process of testing and improving our XGBoost model, we have primarily focused on the overall accuracy score from the classification report and the stratified accuracy results from the confusion matrix. However, there are various other metrics we need to consider when judging a model’s performance, especially within the context of our dataset. In a credit default event, lenders are losing a significant amount of money. Therefore, costs need to be taken into consideration in determining which model is more suitable for our credit risk scenario. This can be evaluated using a cost matrix to define the cost implications of a false negative and a false positive prediction. We have simplified each false prediction’s cost below:

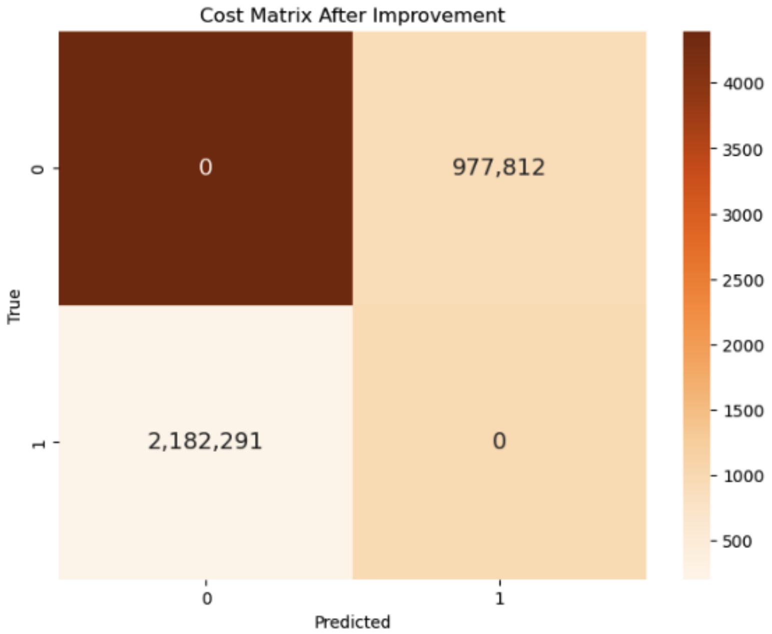
- **Cost of False Negative Prediction** = Average Loan Amount * (1 + Average Interest Rate)
- **Cost of False Positive Prediction** = Average Loan Amount * (Average Interest Rate)

A false negative prediction is more costly because a default will occur in that scenario. Whereas in a false positive prediction, a default will not occur because the lender would not have issued the loan in the first place to a customer with a predicted chance of default. Below is a comparison of the cost matrix of our XGBoost model before and after improvement. The improvement utilizing SMOTE, standardization, and tuning of the “scale_pos_weight” parameter achieved far more cost savings than the benchmark model. This improvement saved about $5 million in potential costs compared to the benchmark model.



Cost Matrix Before Improvement

```
Total Costs: $8,008,985
```



Cost Matrix After Improvement

```
Total Costs: $3,160,103
```

6. Conclusion

Our team's research and analysis in the Credit Loan Default Detection Report have culminated in a comprehensive exploration of the complexities surrounding credit risk in the financial sector. Our approach, utilizing rigorous data analysis and innovative machine learning techniques, has led to insightful findings and practical solutions for detecting loan defaults.

Innovative Application of Machine Learning Models

Throughout our investigation, we employed various machine learning models such as Multinomial Naive Bayes, Logistic Regression, Random Forest, and XGBoost. The decision to focus on XGBoost emerged from its exceptional performance in initial evaluations. This choice underscores our commitment to adopting the most effective tools and techniques in our analysis. The process of refining the XGBoost model, especially through pre-processing steps like SMOTE, data standardization, and parameter tuning, demonstrates our team's adeptness in navigating the intricacies of predictive modeling. We have shown particular attentiveness to the challenge of class imbalances, ensuring the accuracy and reliability of our model's predictions.

Balancing Model Accuracy Measures and Business Relevance

Our methodology in fine-tuning the XGBoost model reflects a thoughtful approach that balances technical precision with business relevance. We looked at more metrics besides the overall accuracy, such as the expected value of cost implications from the false predictions generated. This serves a more relevant purpose as lenders will want to know what the cost of errors are in implementing our model more so than its overall accuracy. This balance is crucial in making our findings not only statistically sound but also practically applicable in real-world scenarios.

Contribution to Financial Risk Management

The insights and methodologies presented in our report are tailored to assist financial institutions and lenders in enhancing their approaches to credit risk assessment. We have highlighted the significance of data-driven decision-making in the financial sector and the transformative role that machine learning can play in modern credit risk management.

A Step Forward in Credit Risk Analysis

In summary, this report is a reflection of our team's dedication, innovation, and thorough understanding of the challenges and opportunities in credit risk assessment. We have produced a thorough data exploration of the dataset and business problem along with choosing an appropriate analysis framework consisting of machine learning prediction models. In the end, we have generated a deployable prediction model as shown in our code file that can take inputs and generate a probability of default.