

Predicting Loan Defaults: A Data-Driven Approach to Credit Risk Analysis

BEE2041 - Data Science in Economics

Student Number - 720017170

Table of contents

| | |
|--|----|
| 1. Introduction | 2 |
| 2. Data | 2 |
| 2.1 Descriptive Statistics | 2 |
| 2.2 Distribution Analysis | 3 |
| 2.3 Correlation Analysis | 5 |
| 3. Results and Discussion | 6 |
| 3.1 Random Forest | 6 |
| 3.2 Naive Bayes | 8 |
| 3.3 XGBoost | 10 |
| 3.4 Model Evaluation and Comparisons | 11 |
| 4. Conclusion | 12 |

1. Introduction

2. Data

Table 1: Variable Information

| Variable | Data Type | Definition |
|----------------|-----------|---|
| Age | int64 | Age of the borrower |
| Income | int64 | Income of the borrower |
| LoanAmount | int64 | Loan amount requested by the borrower |
| CreditScore | int64 | Credit score of the borrower |
| MonthsEmployed | int64 | Number of months the borrower has been employed |
| NumCreditLines | category | Number of credit lines the borrower has |
| InterestRate | float64 | Interest rate of the loan |
| LoanTerm | category | Term of the loan in months |
| DTIRatio | float64 | Debt-to-Income ratio of the borrower |
| Education | object | Education level of the borrower |
| EmploymentType | object | Employment type of the borrower |
| MaritalStatus | object | Marital status of the borrower |
| HasMortgage | object | Whether the borrower has a mortgage |
| HasDependents | object | Whether the borrower has dependents |
| LoanPurpose | object | Purpose of the loan |
| HasCoSigner | object | Whether the borrower has a co-signer |
| Default | category | Whether the borrower defaulted on the loan |

2.1 Descriptive Statistics

Table 2: Summary Statistics of Numeric Variables

| Variable | N | Mean | Median | SD | Min | Max |
|----------------|--------|----------|----------|---------|---------|----------|
| Age | 5930.0 | 40.5 | 39.0 | 14.9 | 18.0 | 69.0 |
| Income | 5930.0 | 78831.0 | 76708.0 | 40289.1 | 15014.0 | 149944.0 |
| LoanAmount | 5930.0 | 134944.3 | 137330.5 | 70970.4 | 5000.0 | 249929.0 |
| CreditScore | 5930.0 | 568.9 | 568.0 | 158.7 | 300.0 | 849.0 |
| MonthsEmployed | 5930.0 | 55.6 | 54.0 | 34.8 | 0.0 | 119.0 |
| InterestRate | 5930.0 | 14.7 | 15.3 | 6.6 | 2.0 | 25.0 |
| DTIRatio | 5930.0 | 0.5 | 0.5 | 0.2 | 0.1 | 0.9 |

2.2 Distribution Analysis

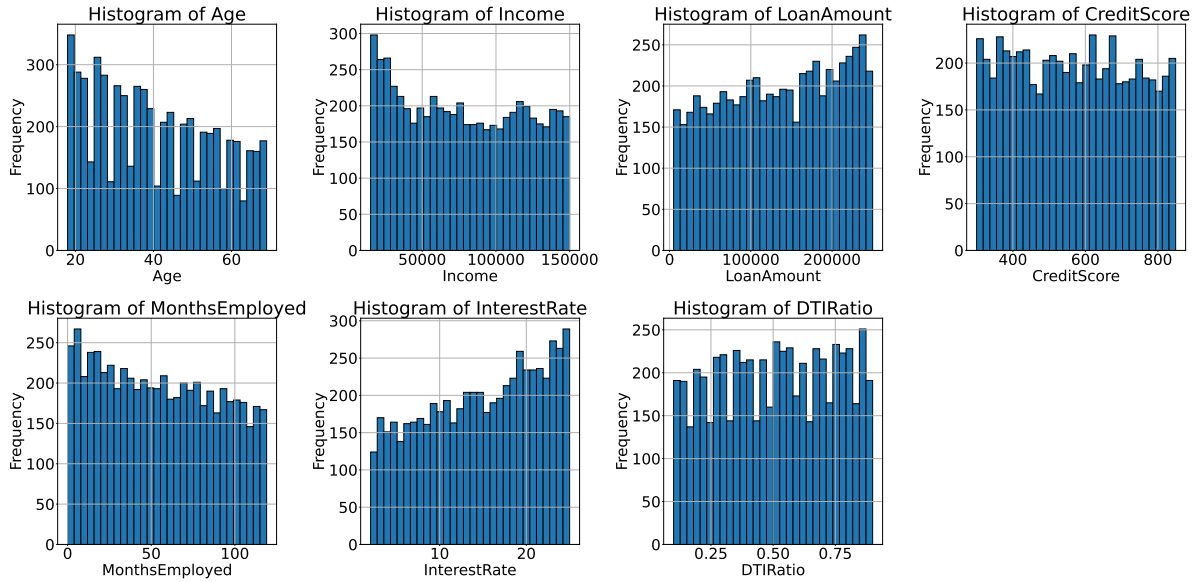


Figure 7: Histograms of all Numeric Variables

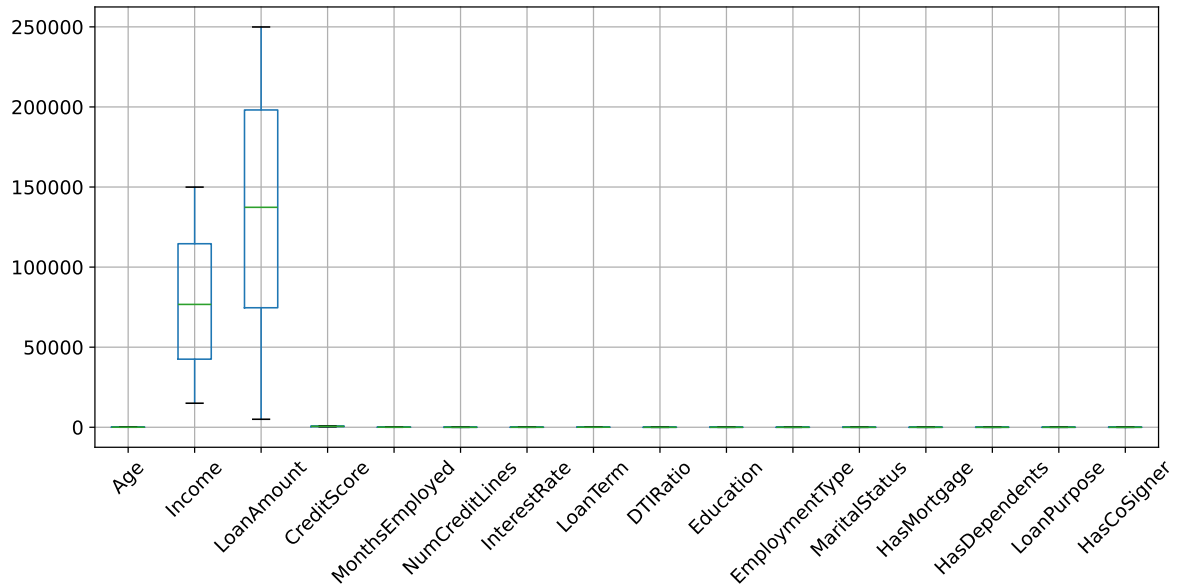


Figure 8: Box Plots of All Variables Before Normalisation

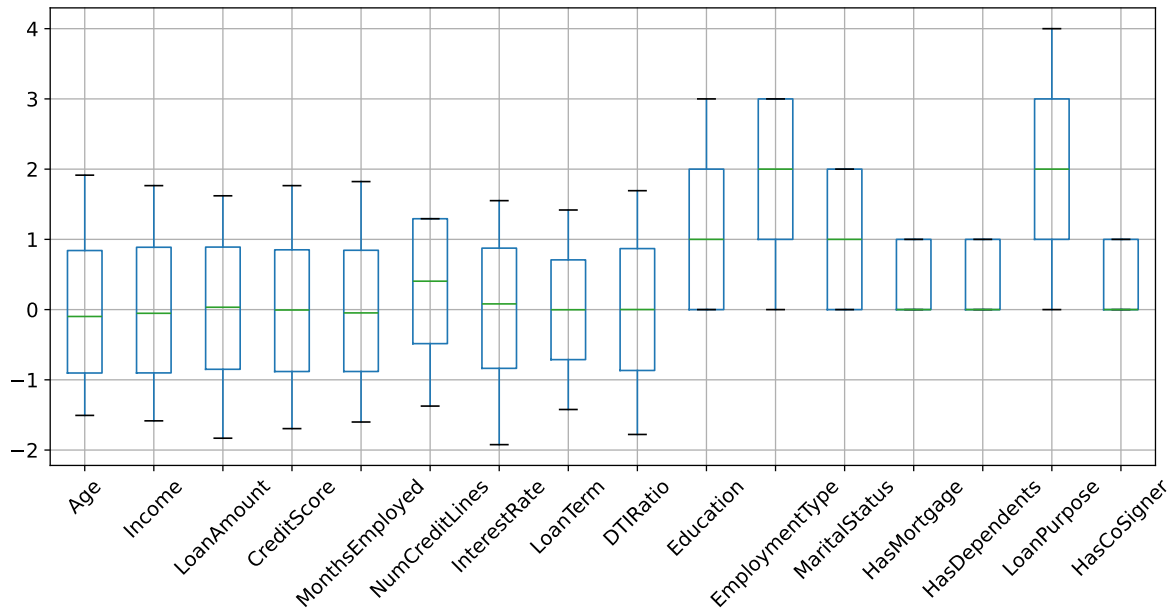


Figure 7: Box Plots of All Variables After Normalisation

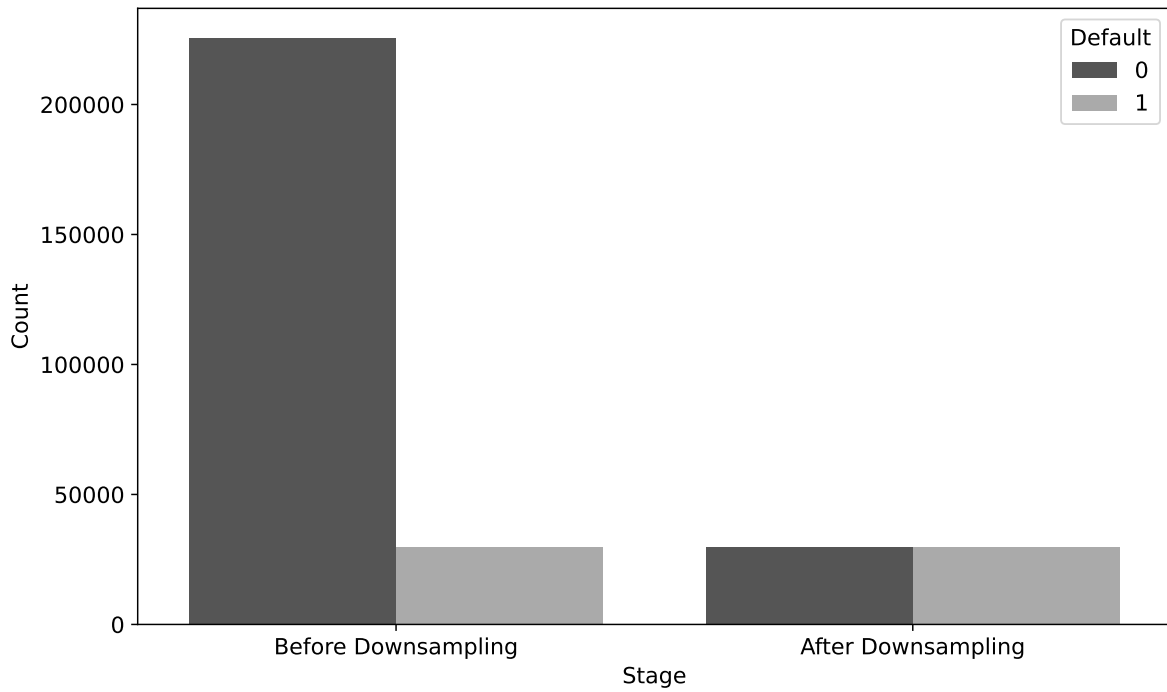


Figure 8: Distribution of Default Before and After Downsampling

2.3 Correlation Analysis

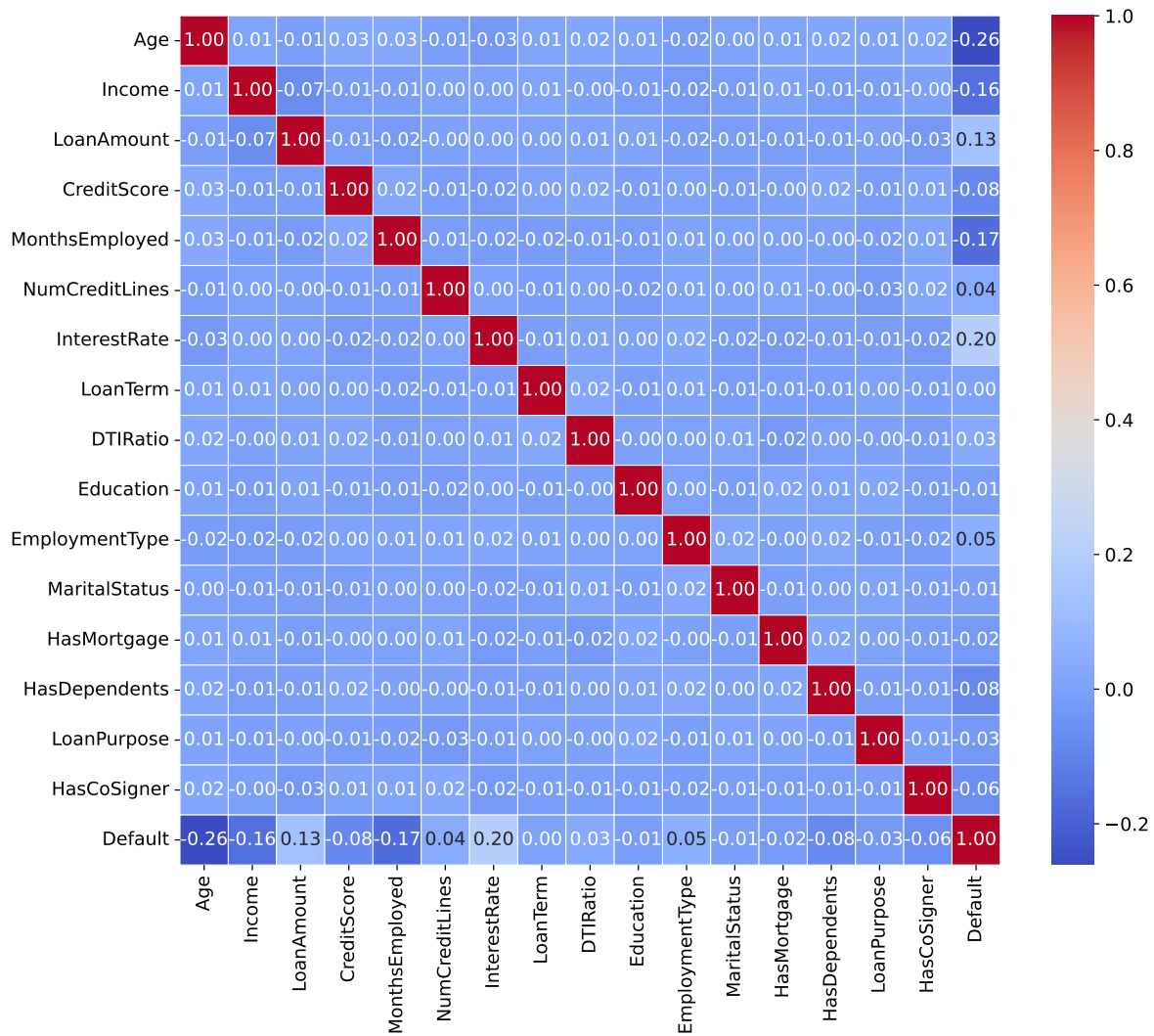


Figure 7: Correlation Plot of All Variables

3. Results and Discussion

3.1 Random Forest

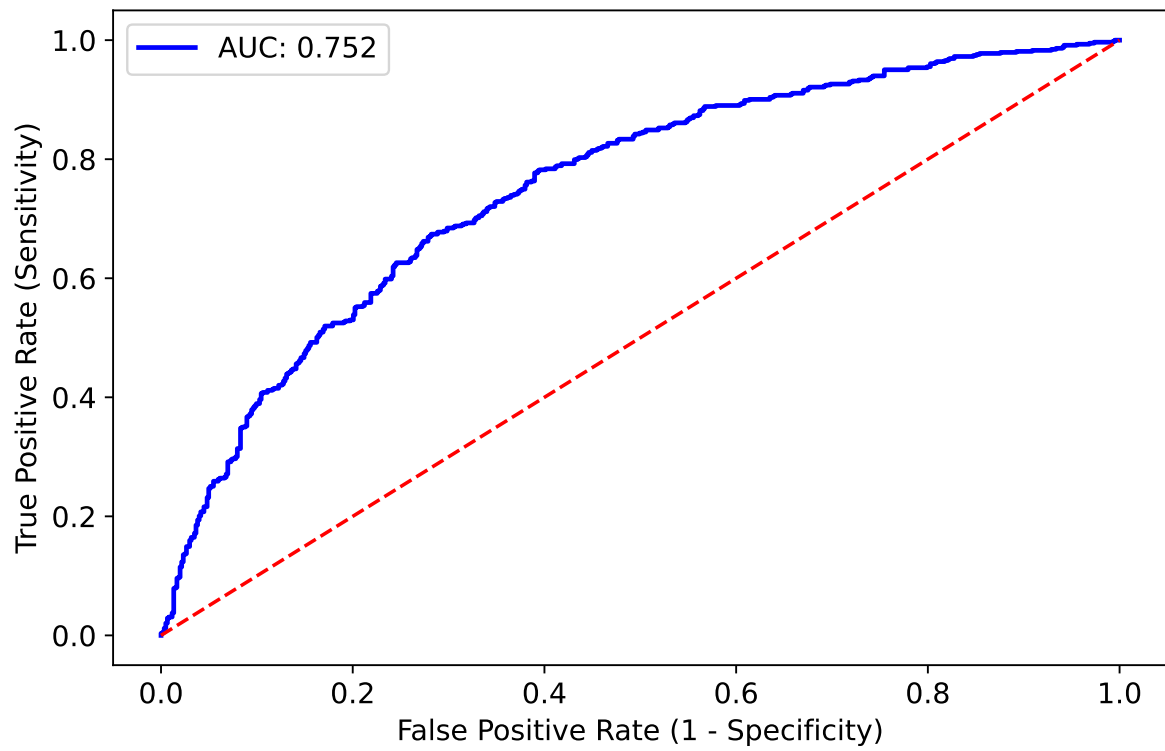


Figure 7: ROC Curve for Random Forest Model

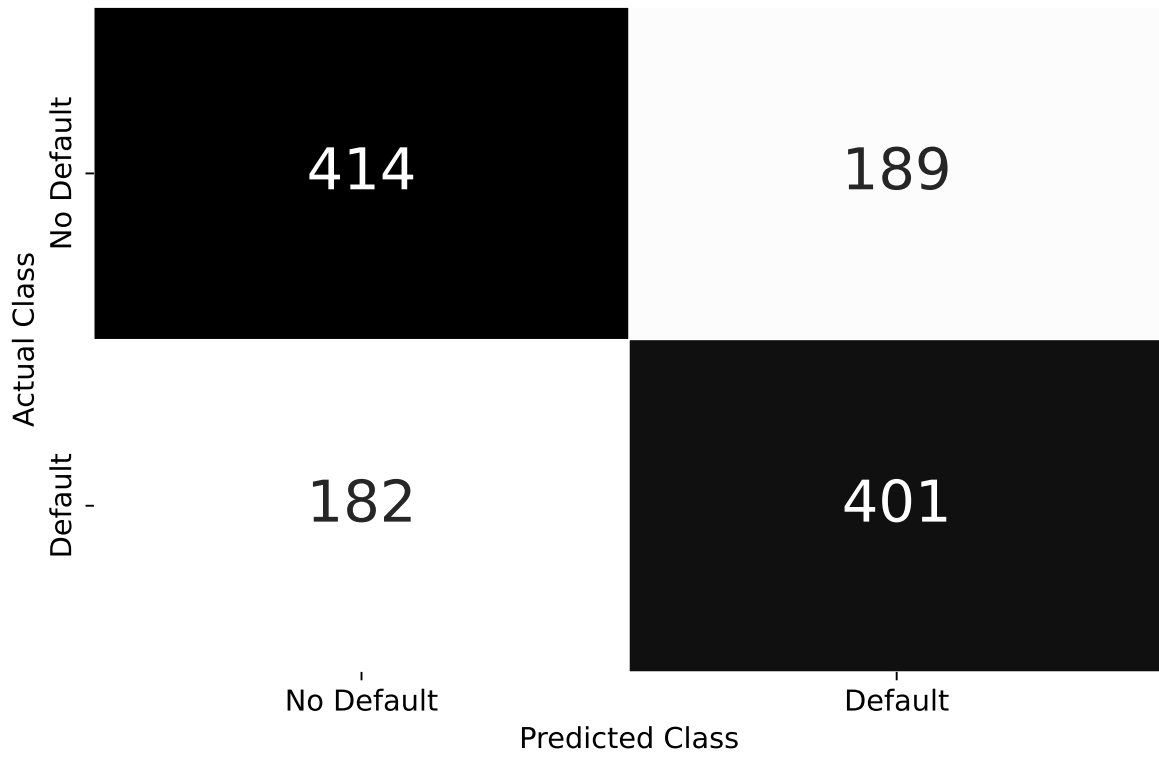


Figure 7: Confusion Matrix for Random Forest Model

3.2 Naive Bayes

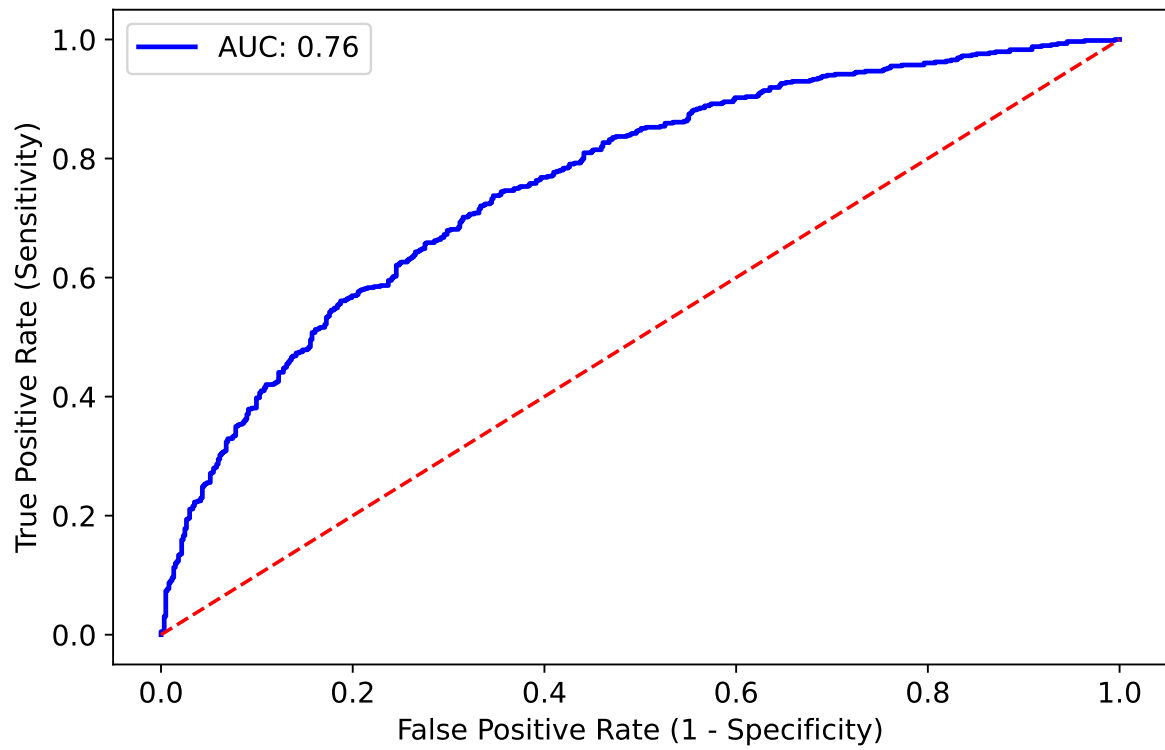


Figure 7: ROC Curve for Naive Bayes Model

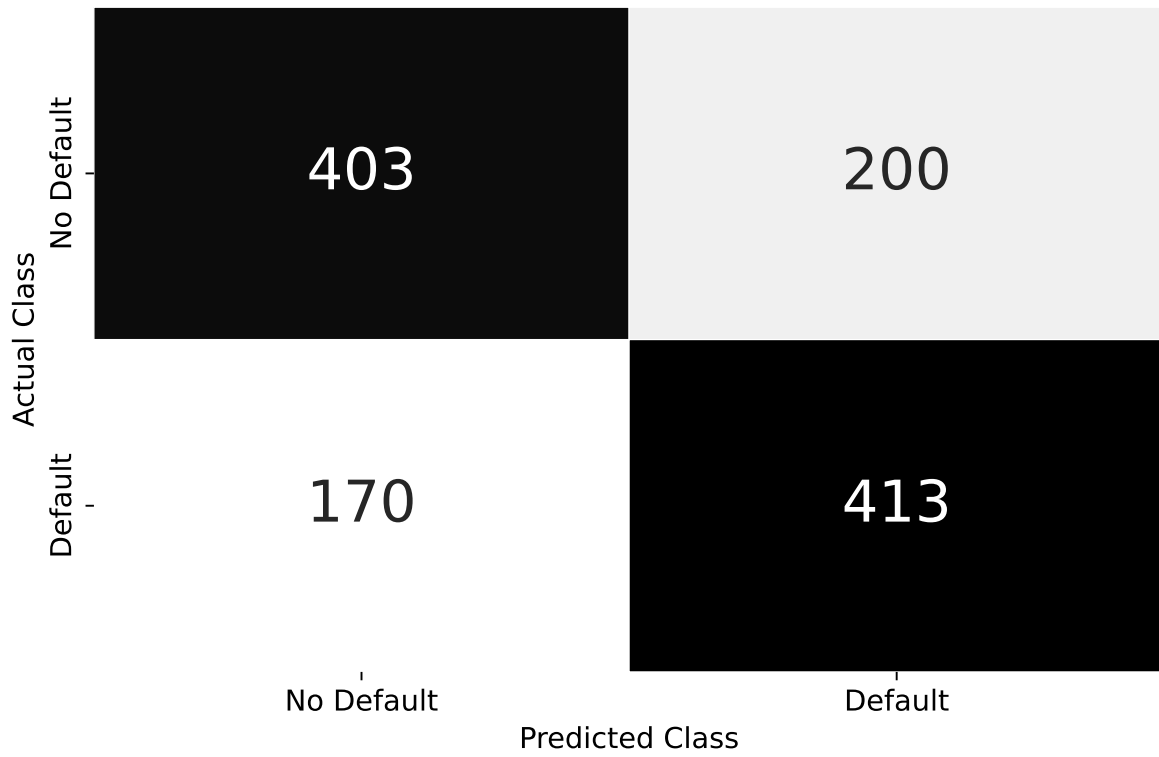


Figure 7: Confusion Matrix for NB Model

3.3 XGBoost

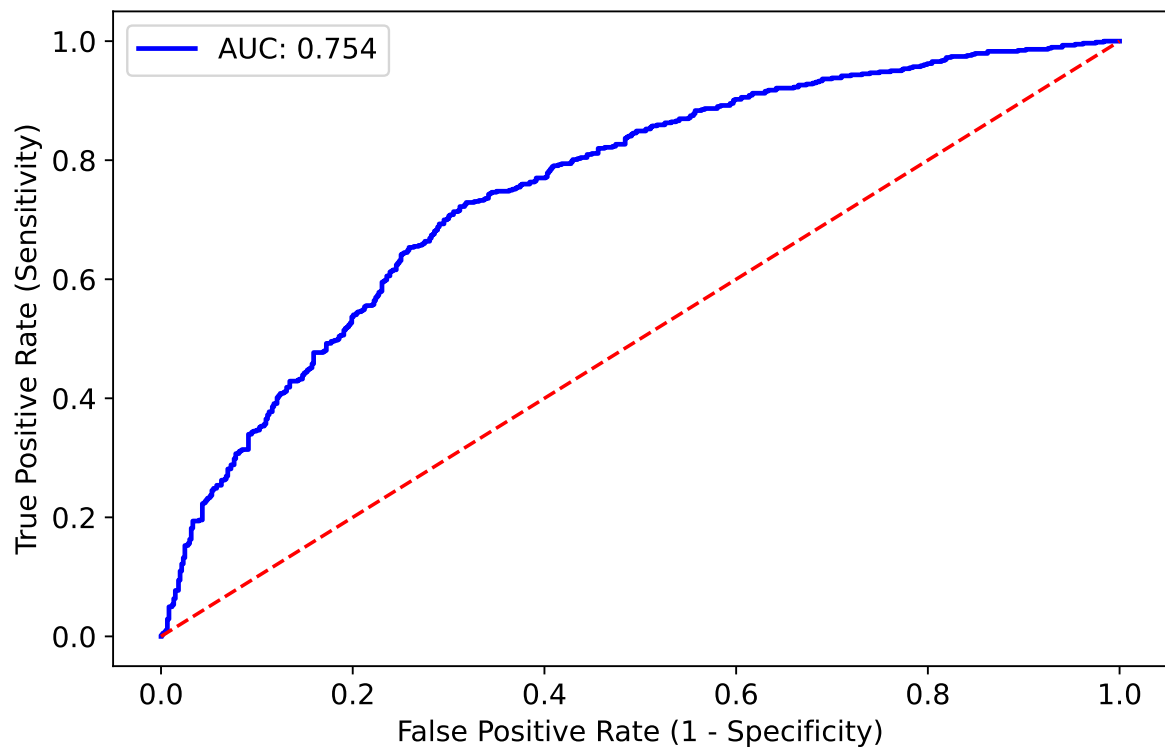


Figure 7: ROC Curve for XGBoost Model

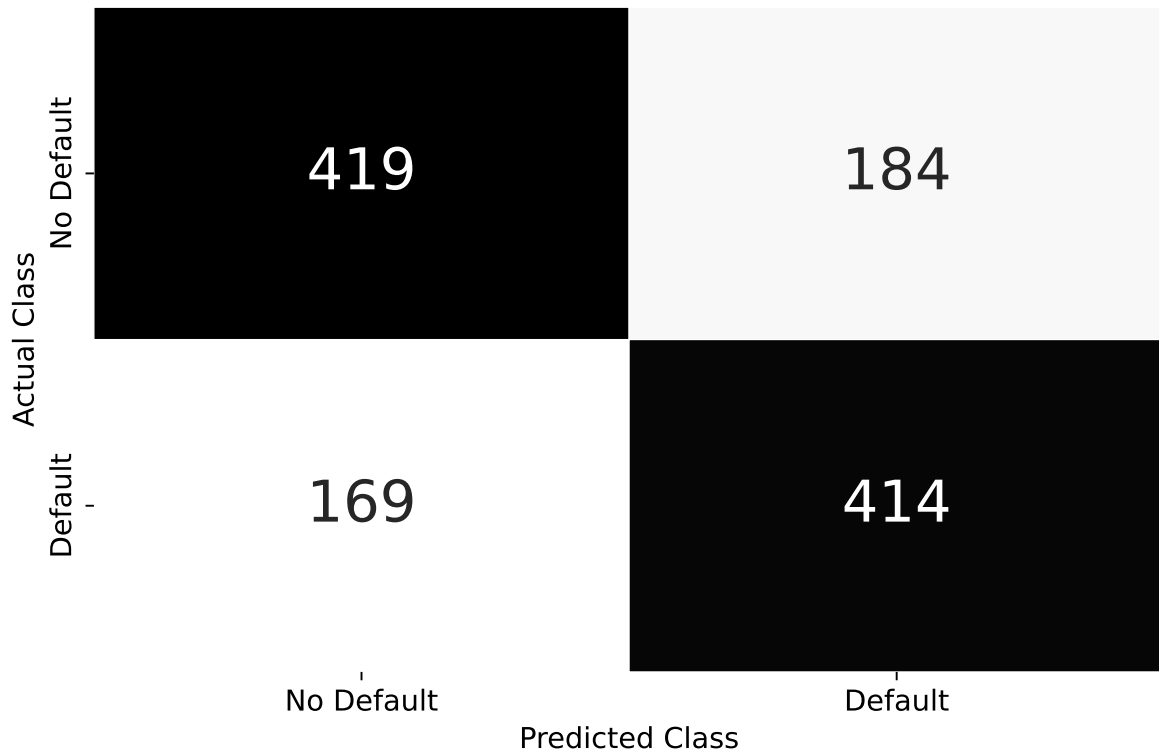


Figure ?: Confusion Matrix for SVM Model

3.4 Model Evaluation and Comparisons

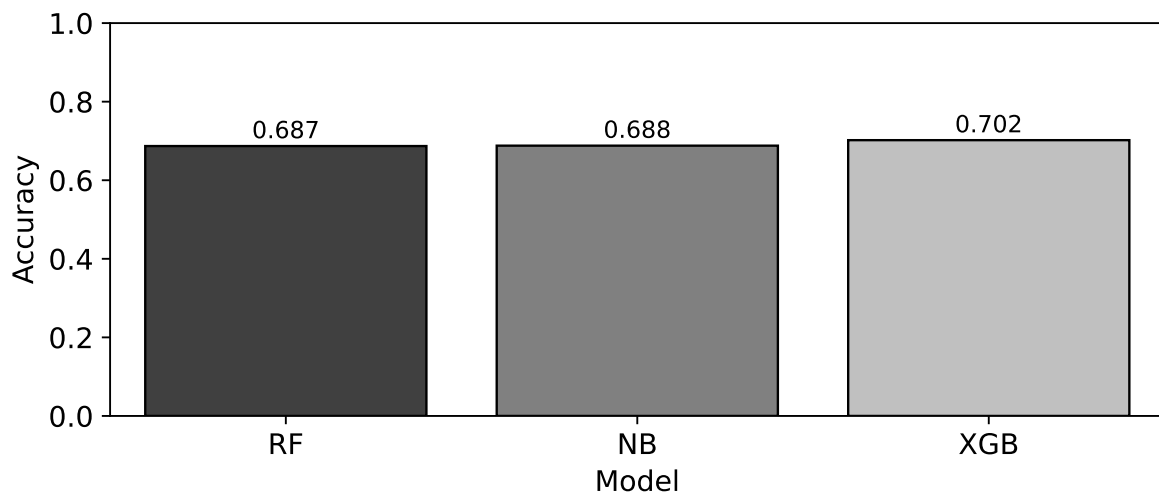


Figure ?: Accuracy for Each Model

Table 3: Performance Metrics for Each Model

| Model | Accuracy | Precision | Recall | F1_Score | AUC |
|-------|----------|-----------|--------|----------|-------|
| RF | 0.687 | 0.68 | 0.688 | 0.684 | 0.752 |
| NB | 0.688 | 0.674 | 0.708 | 0.691 | 0.76 |
| XGB | 0.702 | 0.692 | 0.71 | 0.701 | 0.754 |

4. Conclusion

Link to Github Repository = <https://github.com/JoshLG18/DSE-EMP-Project>