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. 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

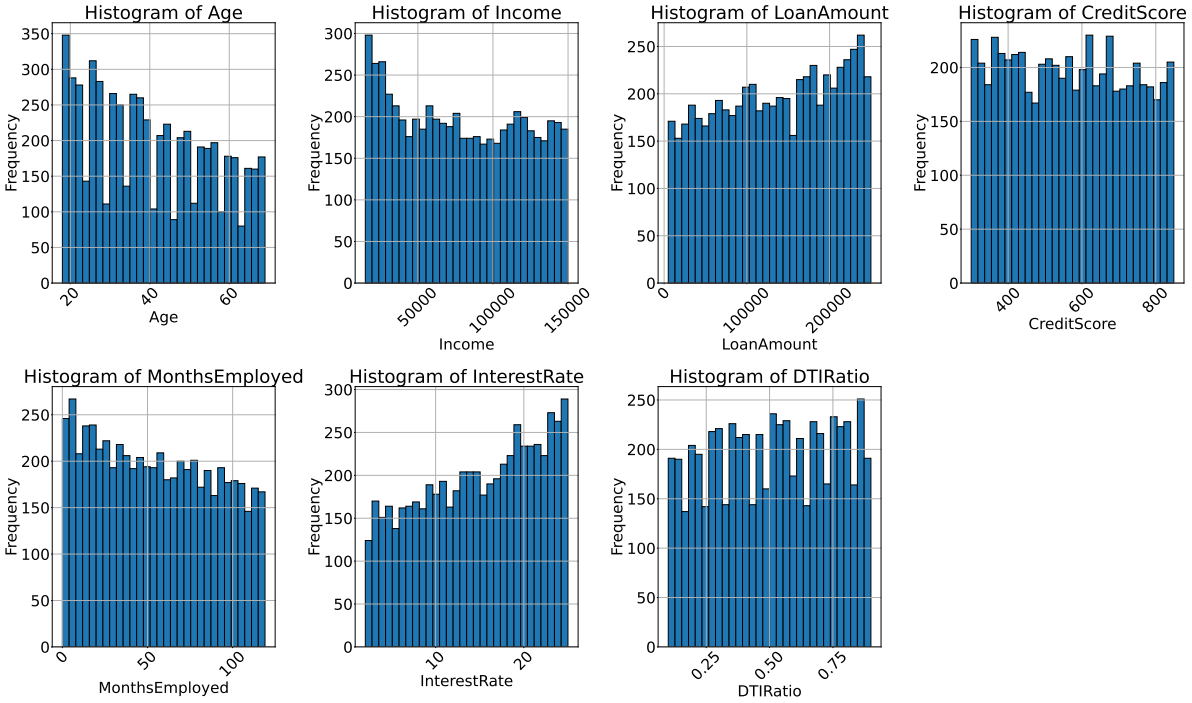


Figure ?: Histograms of all Numeric Variables

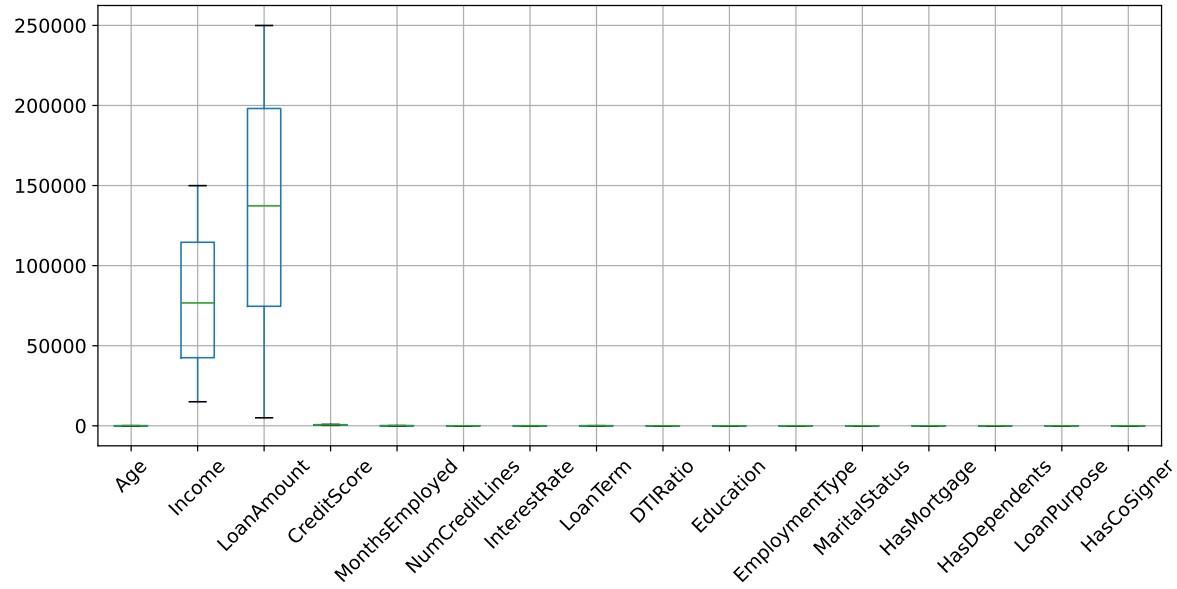


Figure ?: Box Plots of All Variables Before Normalisation

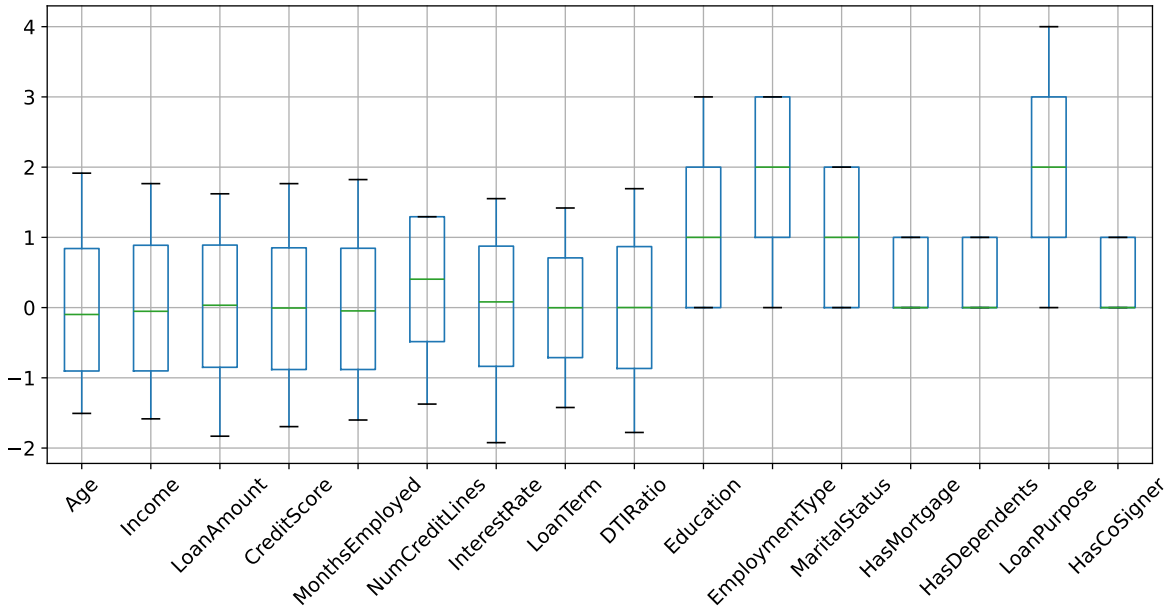


Figure ?: Box Plots of All Variables After Normalisation

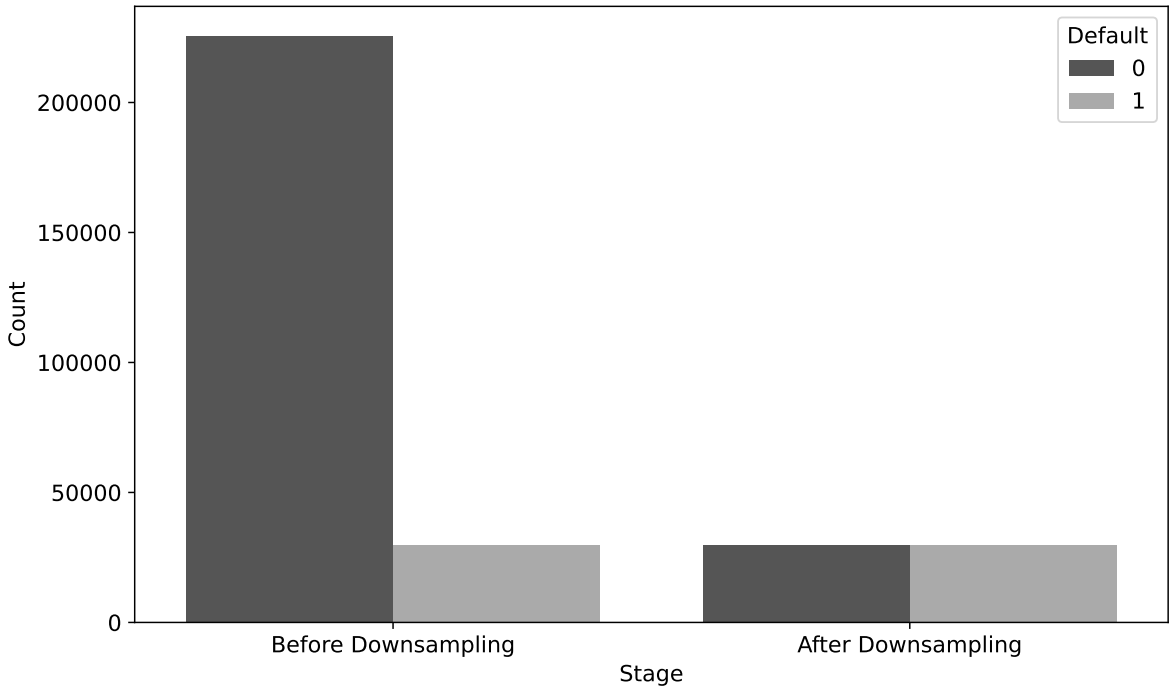


Figure ?: Distribution of Default Before and After Downsampling

2.3 Correlation Analysis

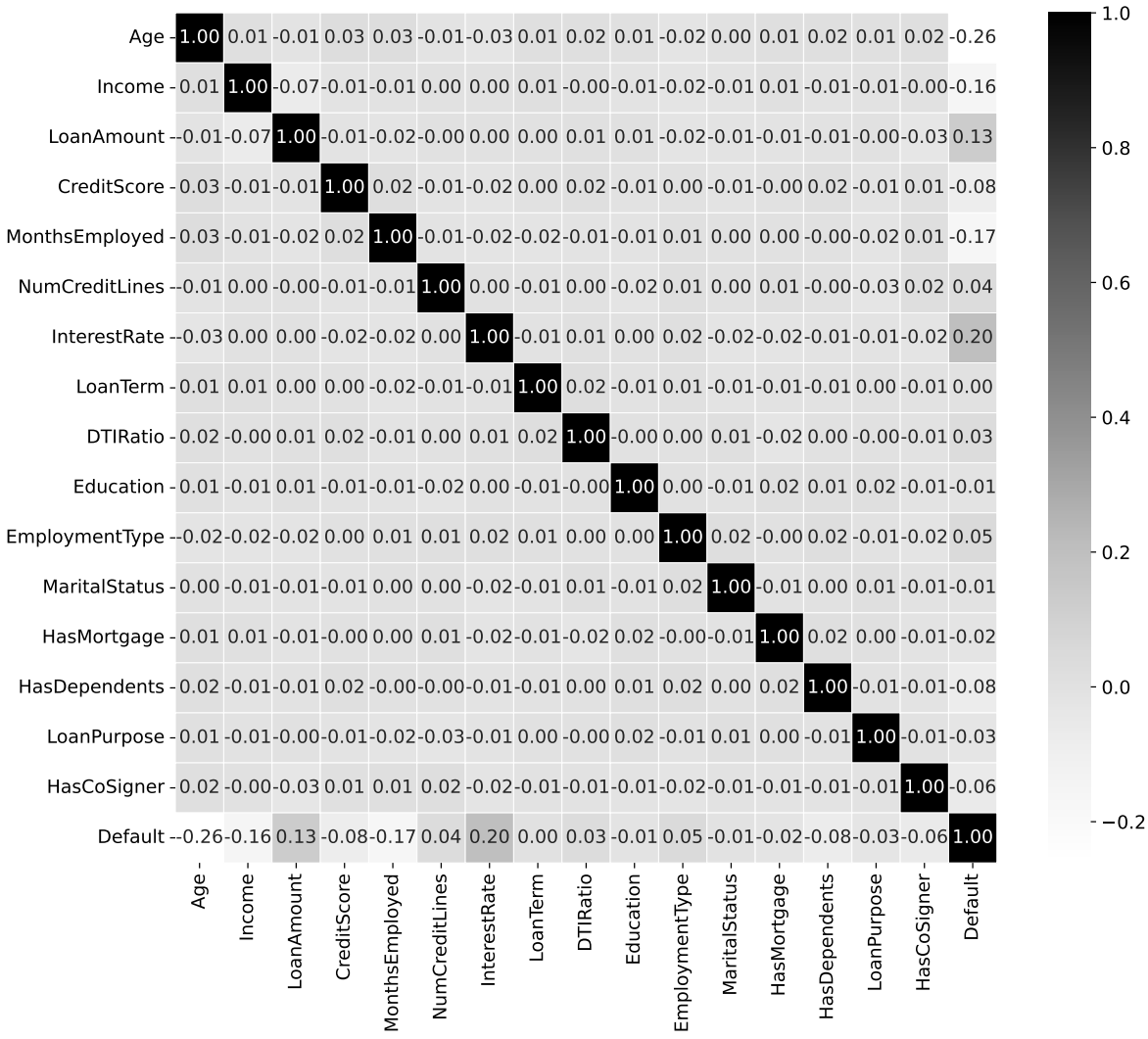


Figure ?: Correlation Plot of All Variables

3. Results and Discussion

3.1 Random Forest

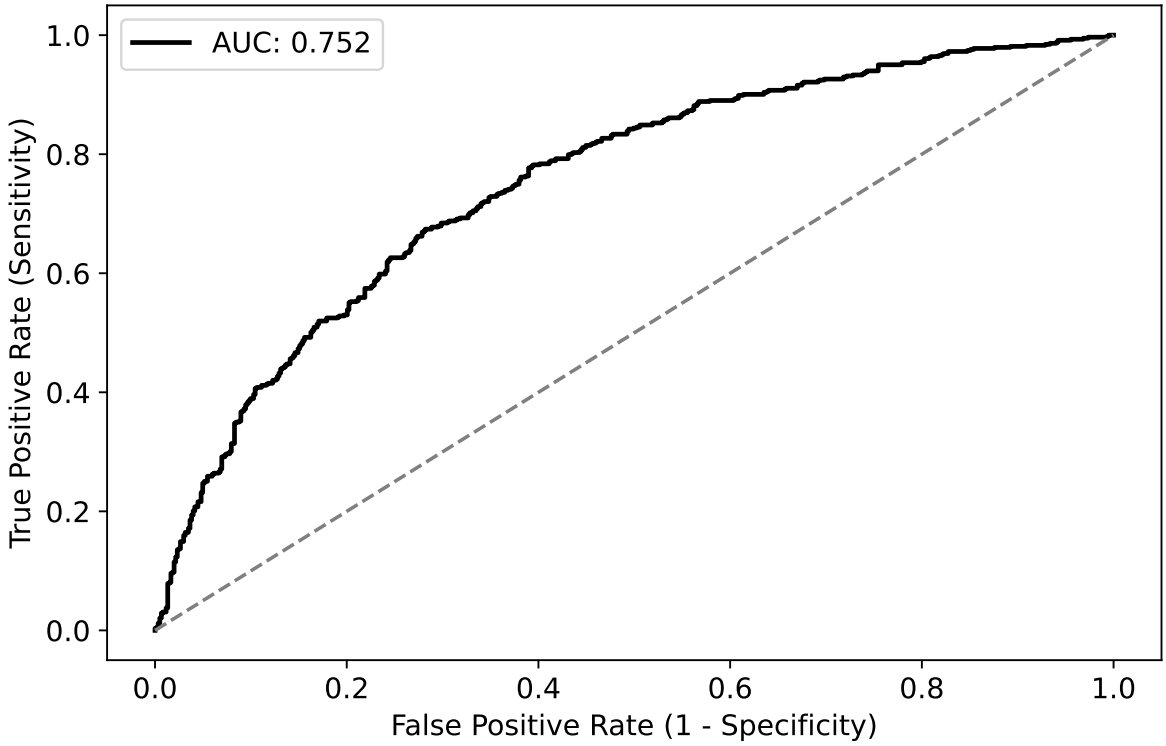


Figure ?: ROC Curve for Random Forest Model

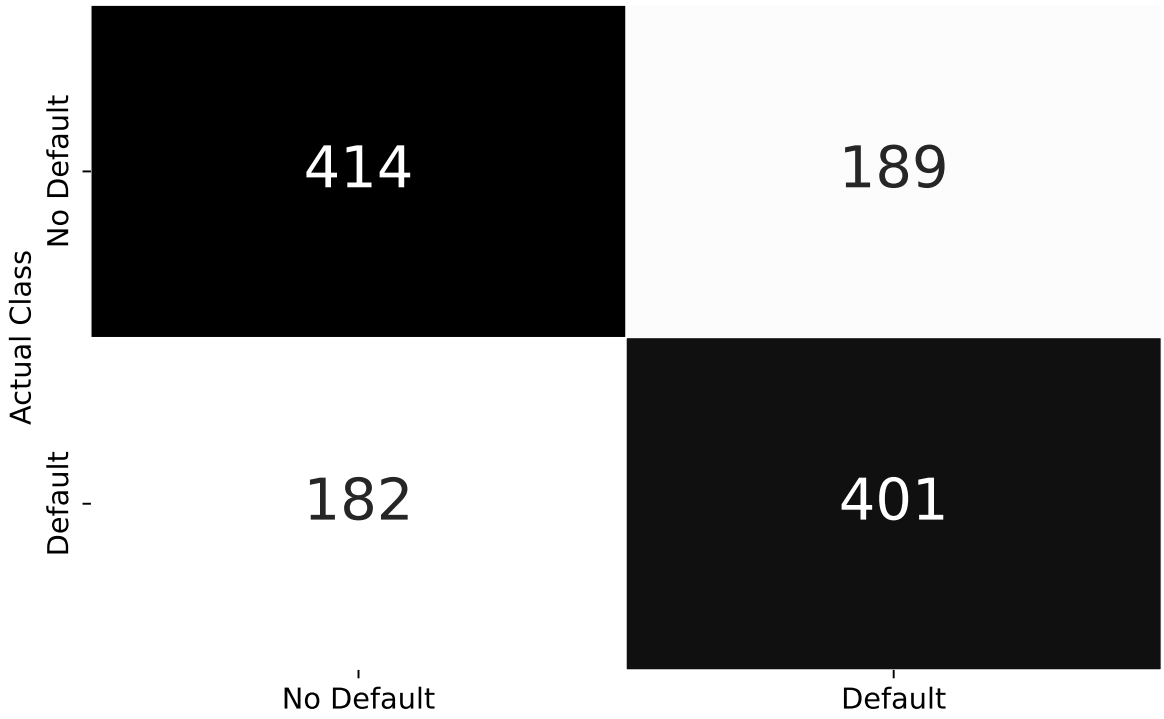
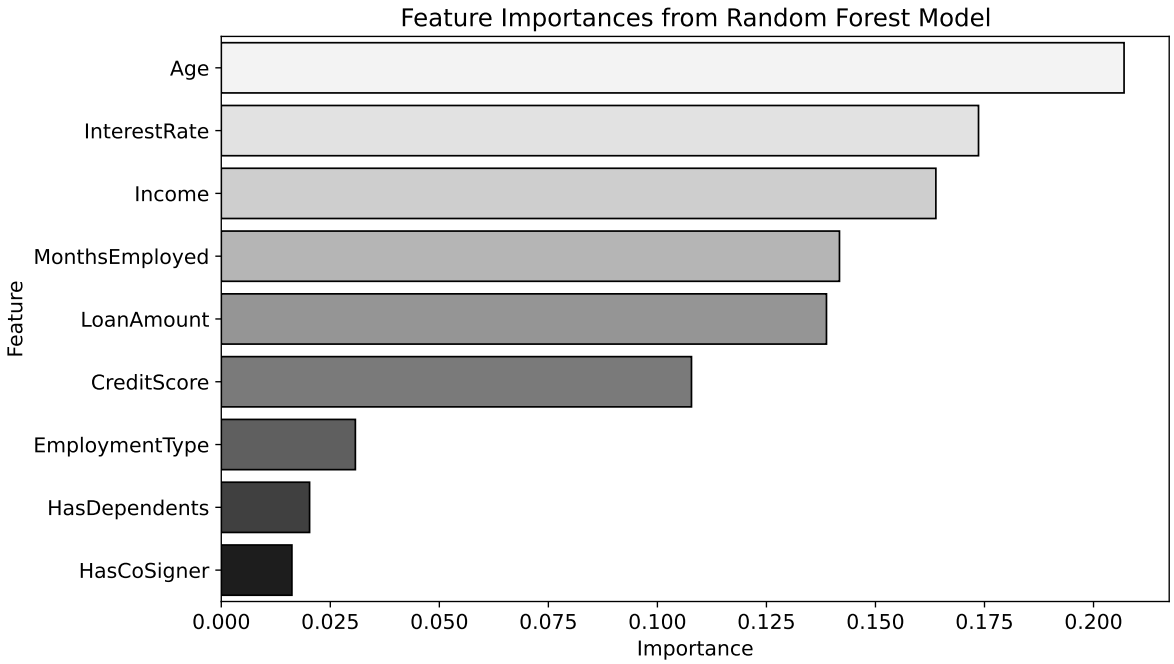


Figure ?: Confusion Matrix for Random Forest Model



3.2 XGBoost

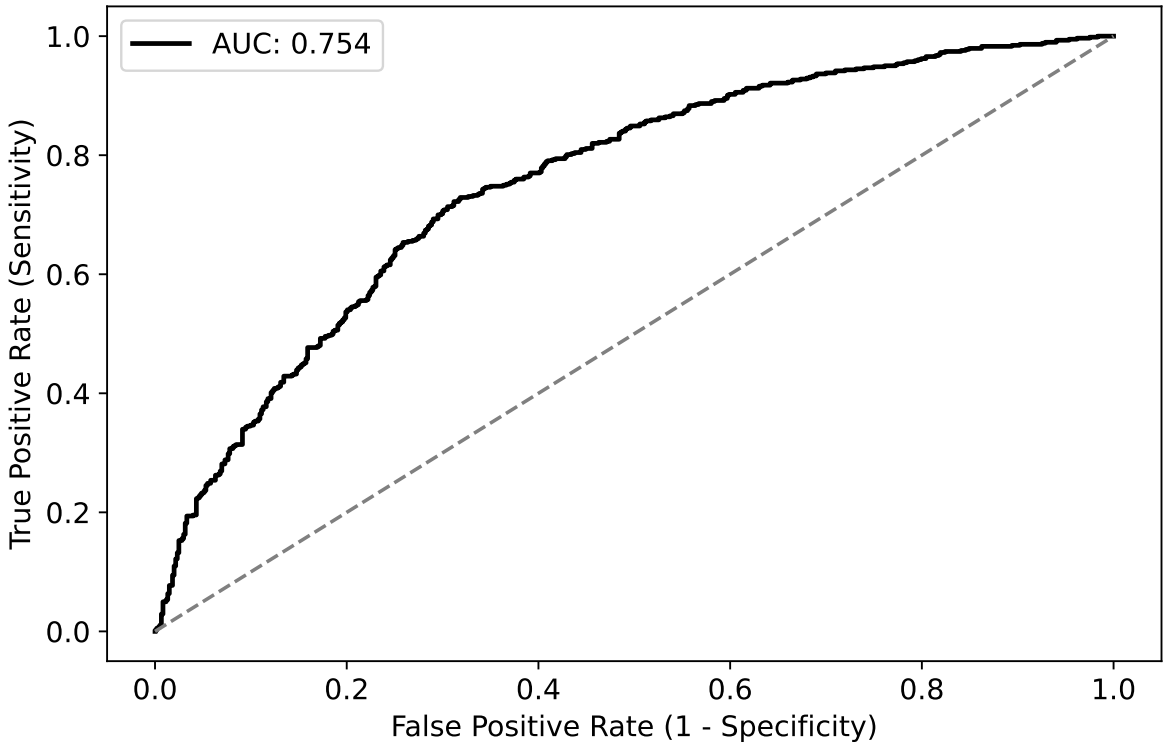


Figure ?: ROC Curve for XGBoost Model

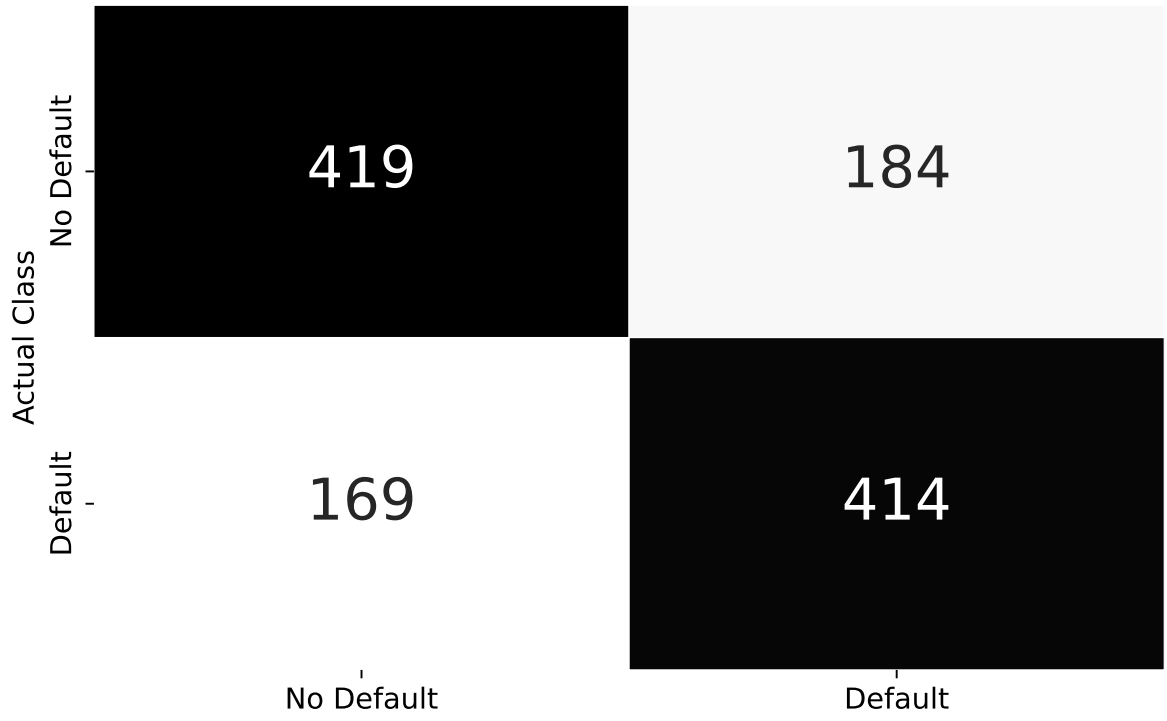
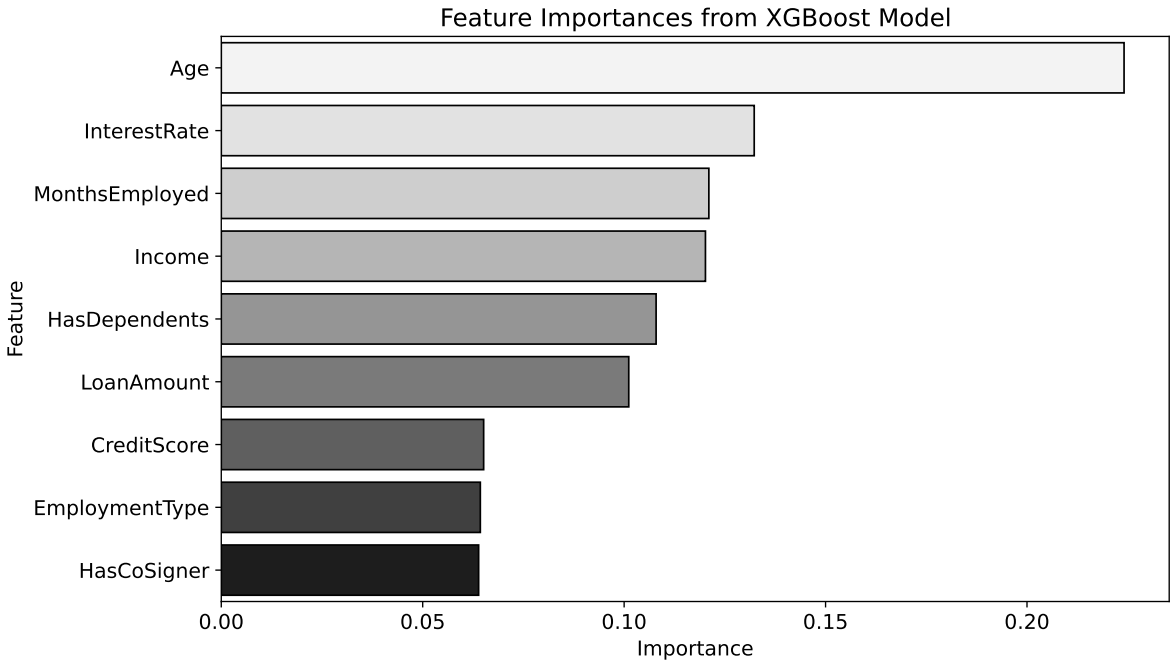
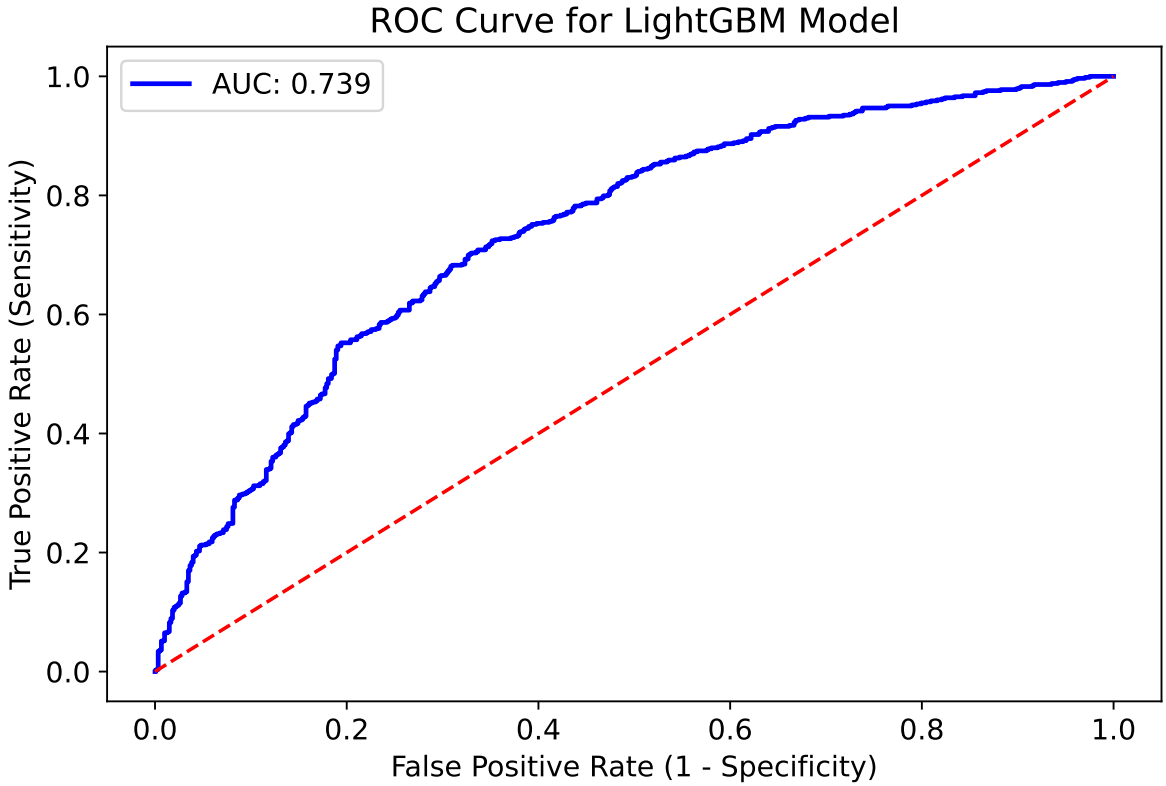


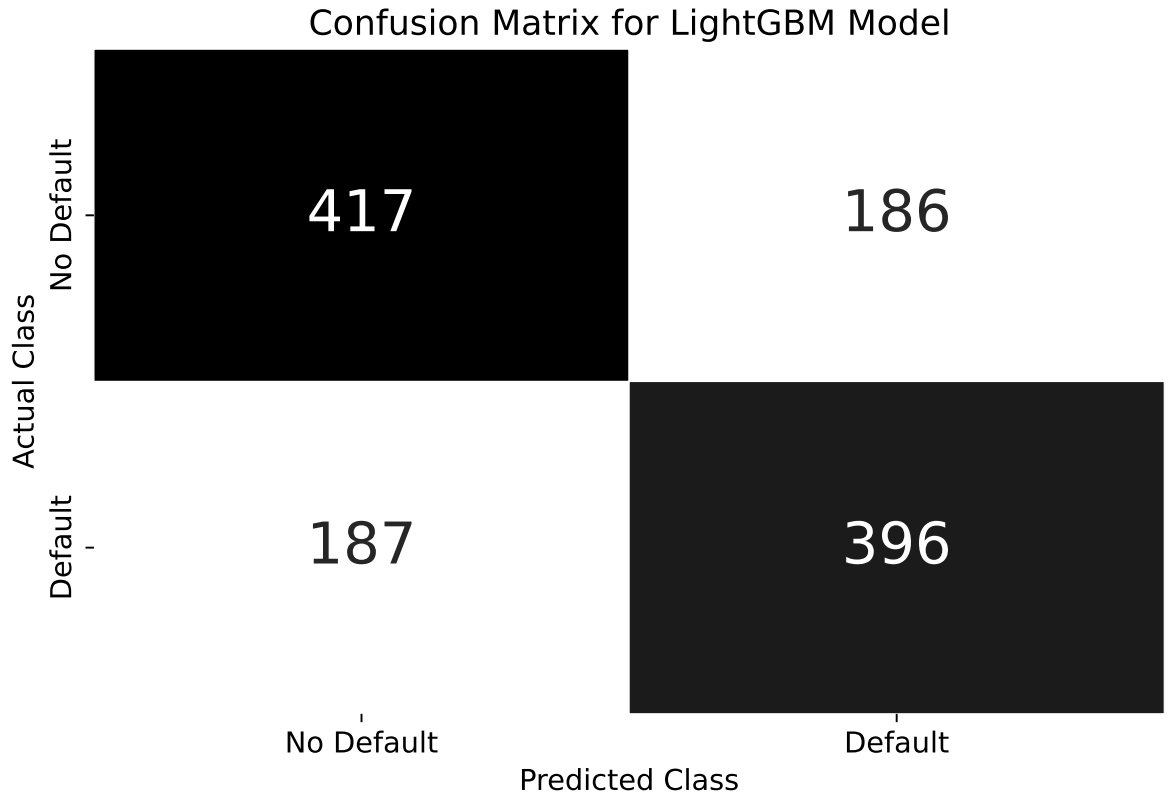
Figure ?: Confusion Matrix for SVM Model

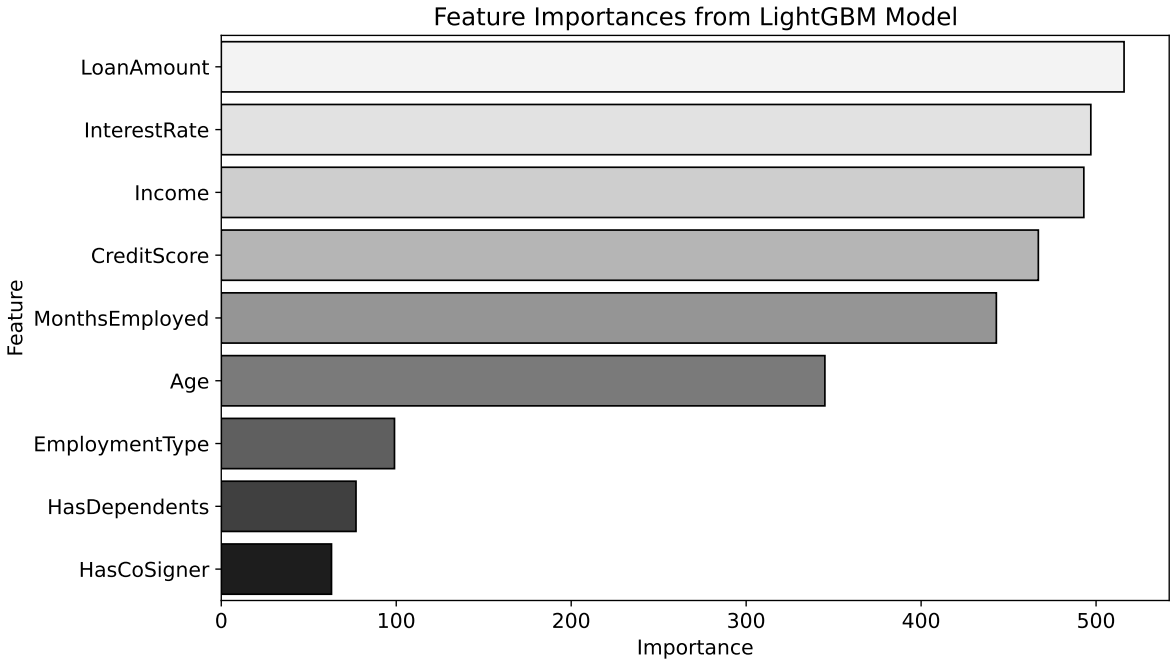


3.3 Light Gradient Boosting Machine (LGBM)

```
[LightGBM] [Info] Number of positive: 2382, number of negative: 2362
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.0003
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 1198
[LightGBM] [Info] Number of data points in the train set: 4744, number of used features: 9
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.502108 -> initscore=0.008432
[LightGBM] [Info] Start training from score 0.008432
```




Confusion Matrix for LightGBM Model
417
186
187
396
Actual Class
No Default
Default
No Default
Default
Predicted Class


Feature Importances from LightGBM Model
LoanAmount
InterestRate
Income
CreditScore
MonthsEmployed
Age
EmploymentType
HasDependents
HasCoSigner
Feature
0
100
200
300
400
500
Importance

3.4 Model Evaluation and Comparisons

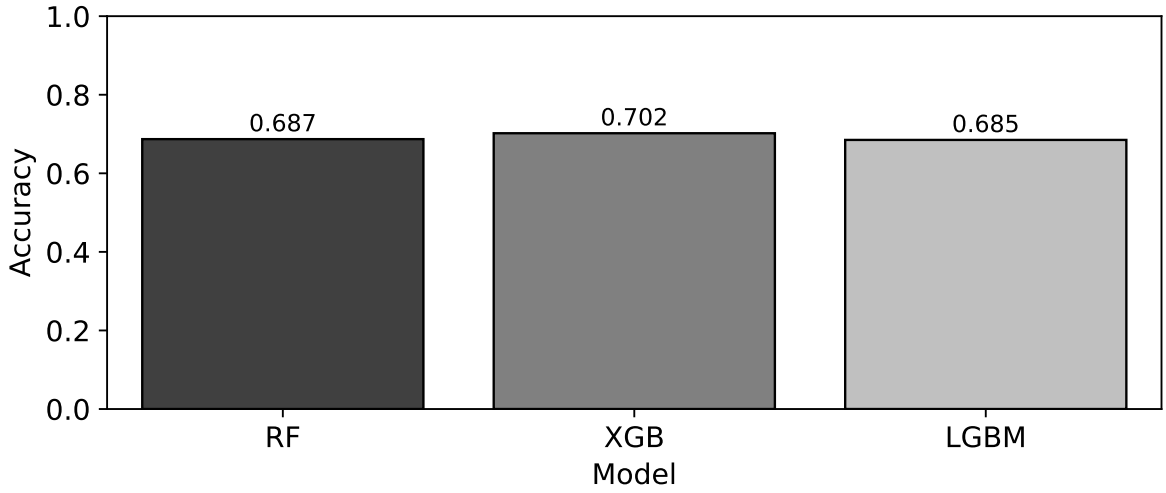


Figure ?: Accuracy for Each Model

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
XGB	0.702	0.692	0.71	0.701	0.754
LGBM	0.685	0.68	0.679	0.68	0.739

4. Conclusion

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