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 . . . 3
2.2 Distribution Analysis . . . 3
2.3 Correlation Analysis . . . 6
3. Results and Discussion . . . 8
3.1 Logistic Regression . . . 8
3.2 Random Forest . . . 11
3.3 XGBoost . . . 13
3.4 Light Gradient Boosting Machine (LGBM) . . . 15
3.5 Model Evaluation and Comparisons . . . 17
4. Conclusion . . . 17

1. Introduction

Access to credit is a important driver of economic growth, allowing households or buisnesses to invest, expand and smooth consumption. However, credit risk remains a fundemental challenge for financial institutions.

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	255347.0	43.5	43.0	15.0	18.0	69.0
Income	255347.0	82499.3	82466.0	38963.0	15000.0	149999.0
LoanAmount	255347.0	127578.9	127556.0	70840.7	5000.0	249999.0
CreditScore	255347.0	574.3	574.0	158.9	300.0	849.0
MonthsEmployed	255347.0	59.5	60.0	34.6	0.0	119.0
InterestRate	255347.0	13.5	13.5	6.6	2.0	25.0
DTIRatio	255347.0	0.5	0.5	0.2	0.1	0.9

2.2 Distribution Analysis

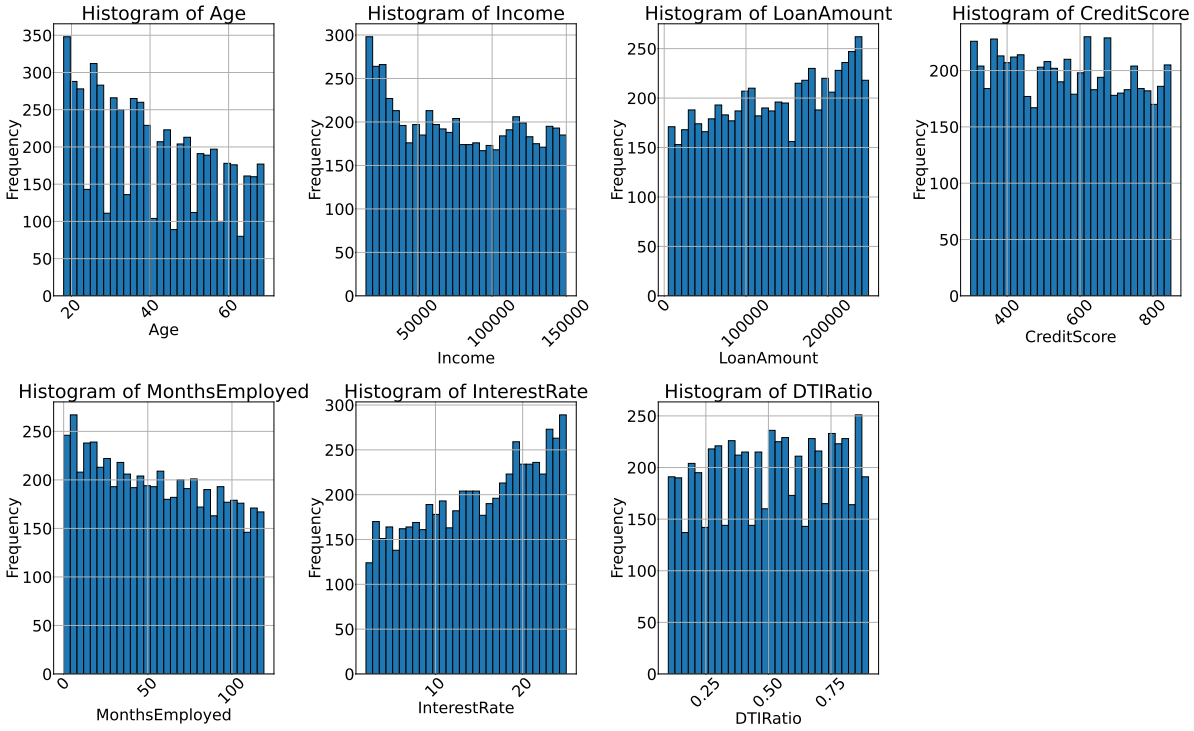

Figure ?: Histograms of all Numeric Variables

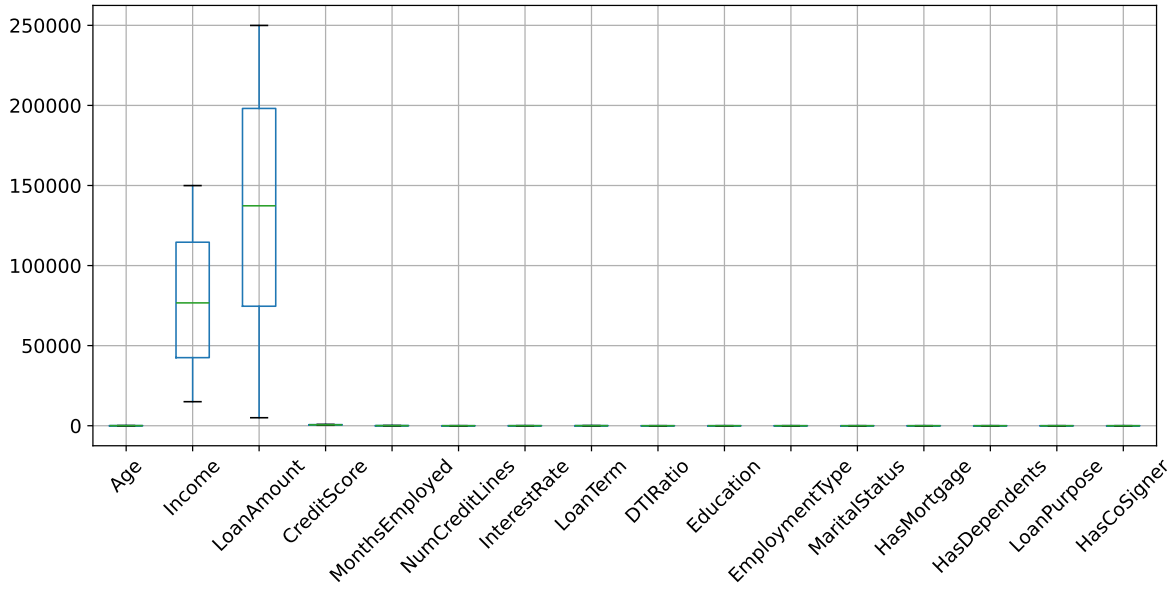

Figure ?: Box Plots of All Variables Before Normalisation

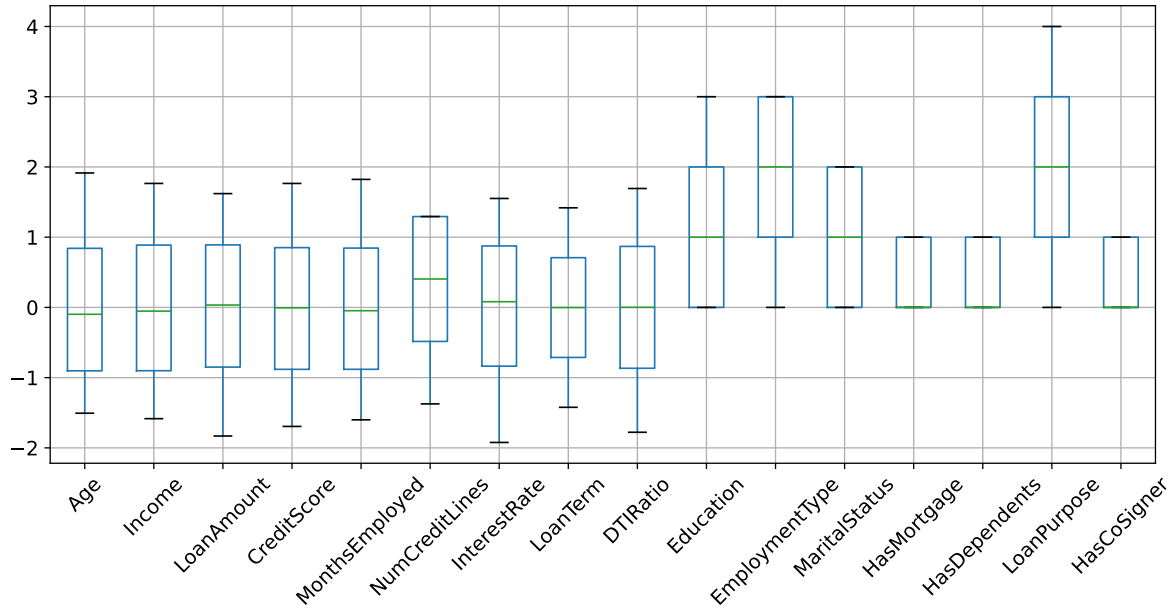

Figure ?: Box Plots of All Variables After Normalisation

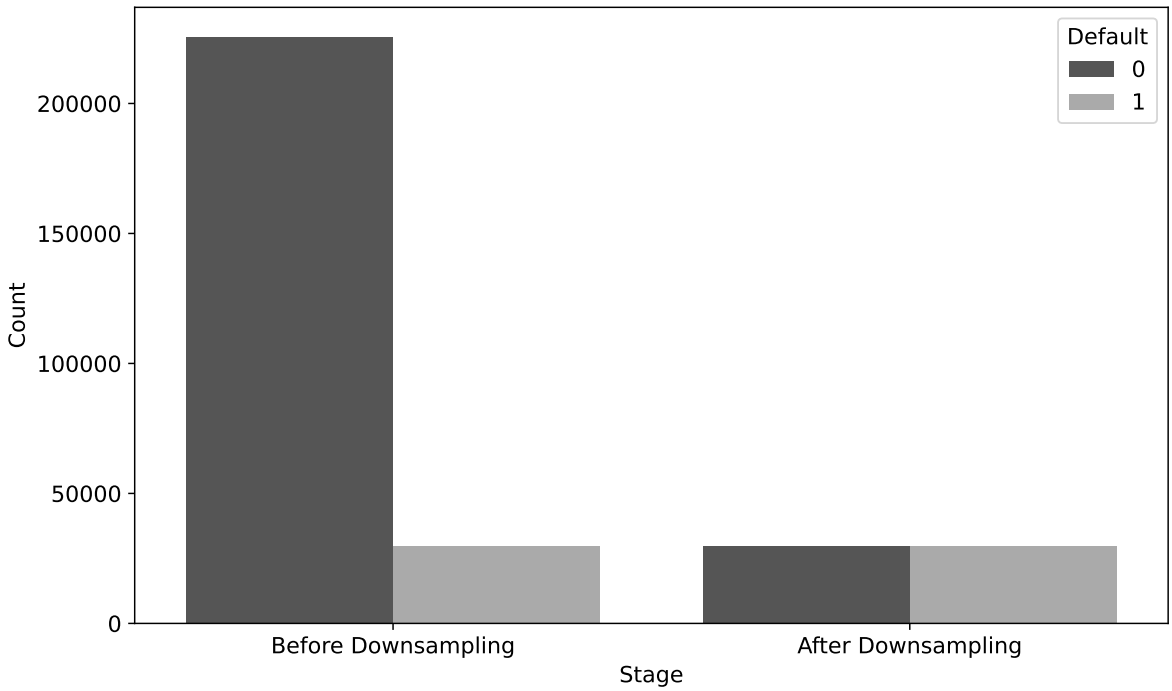

Figure ?: Distribution of Default Before and After Downsampling

Downsampled the dataset to ensure that the models didn't get affected by the magnitude of the majority class = can affect performance metrics.

2.3 Correlation Analysis

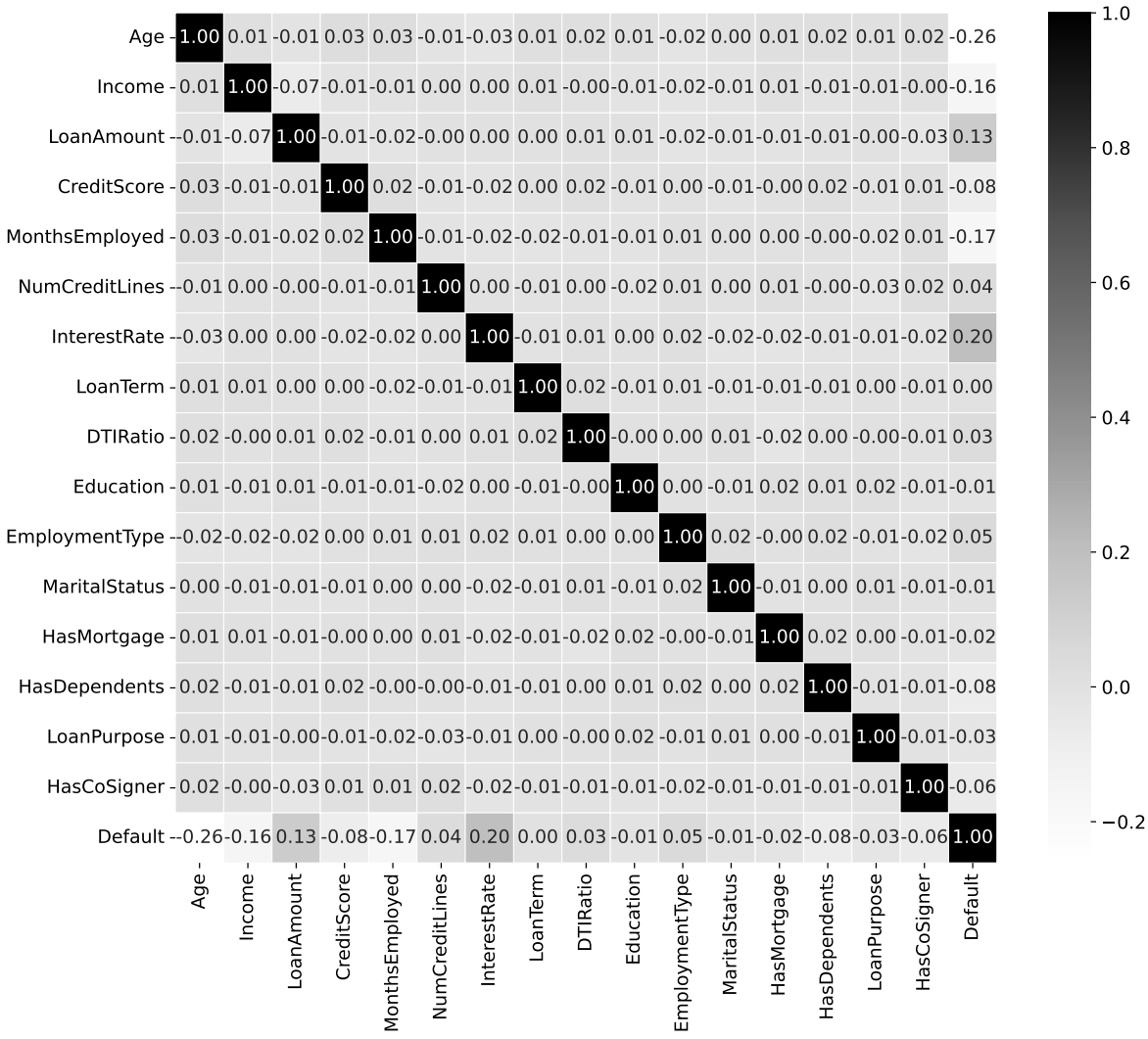

Figure ?: Correlation Plot of All Variables

Table 3: Variance Inflation Factor (VIF) Values

Feature	VIF
Age	1.004797
Income	1.006811
LoanAmount	1.008369
CreditScore	1.002868
MonthsEmployed	1.003494
NumCreditLines	1.002799
InterestRate	1.005943
LoanTerm	1.001839
DTIRatio	1.003853
Education	1.002508
EmploymentType	1.002343
MaritalStatus	1.002549
HasMortgage	1.004015
HasDependents	1.003485
LoanPurpose	1.002860
HasCoSigner	1.004757

Selected Features with a magnitude of correlation to class above 0.05 to remove any variables with low correlation likely to reduce predictive performance

3. Results and Discussion

3.1 Logistic Regression

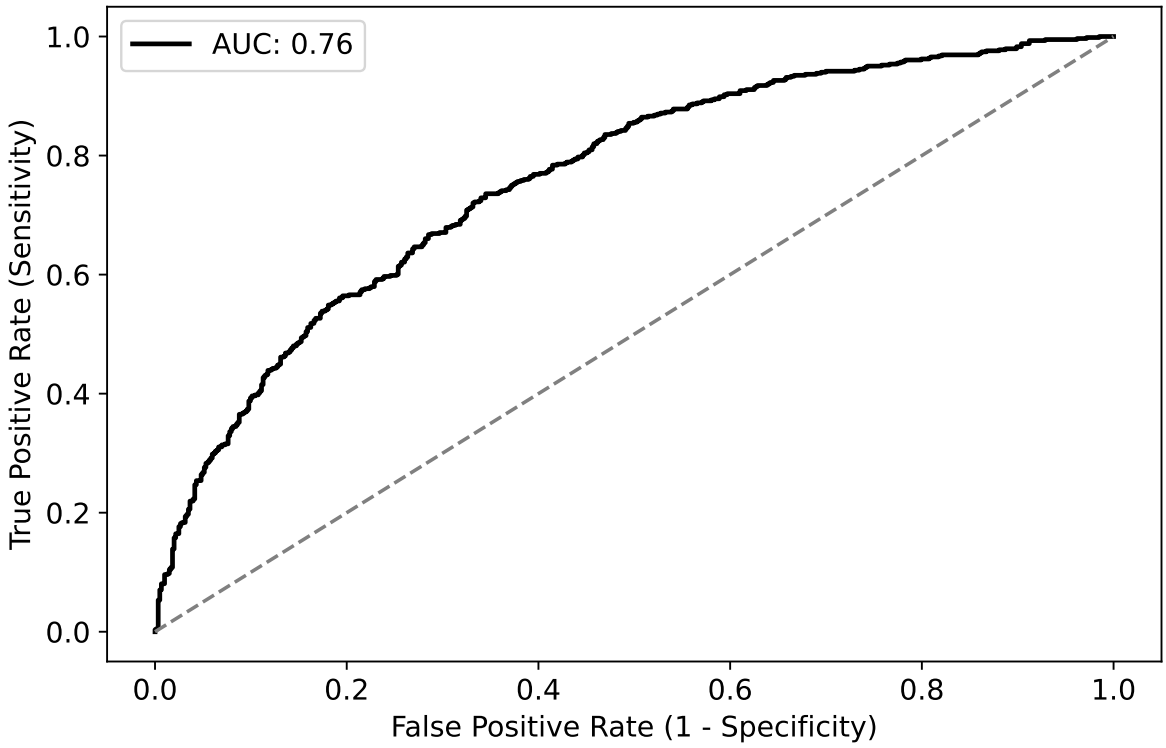

Figure ?: ROC Curve for Logistic Regression Mod

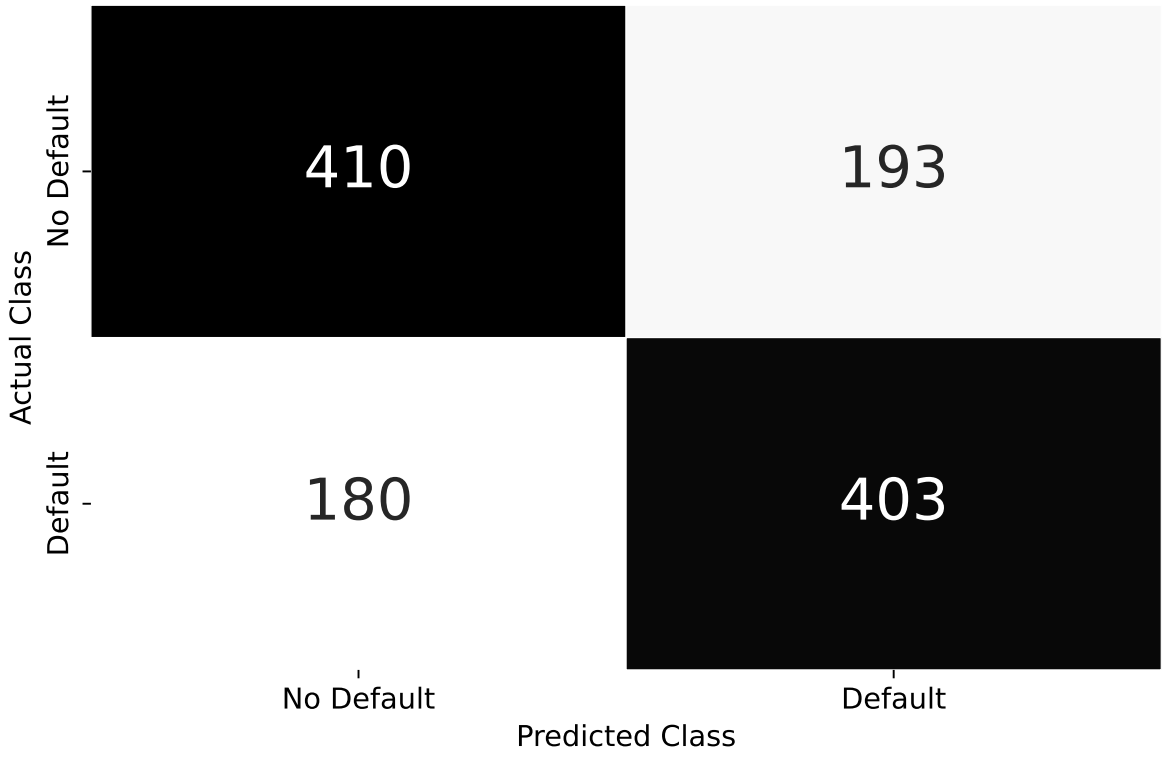

Figure ?: Confusion Matrix for Logistic Regression Model

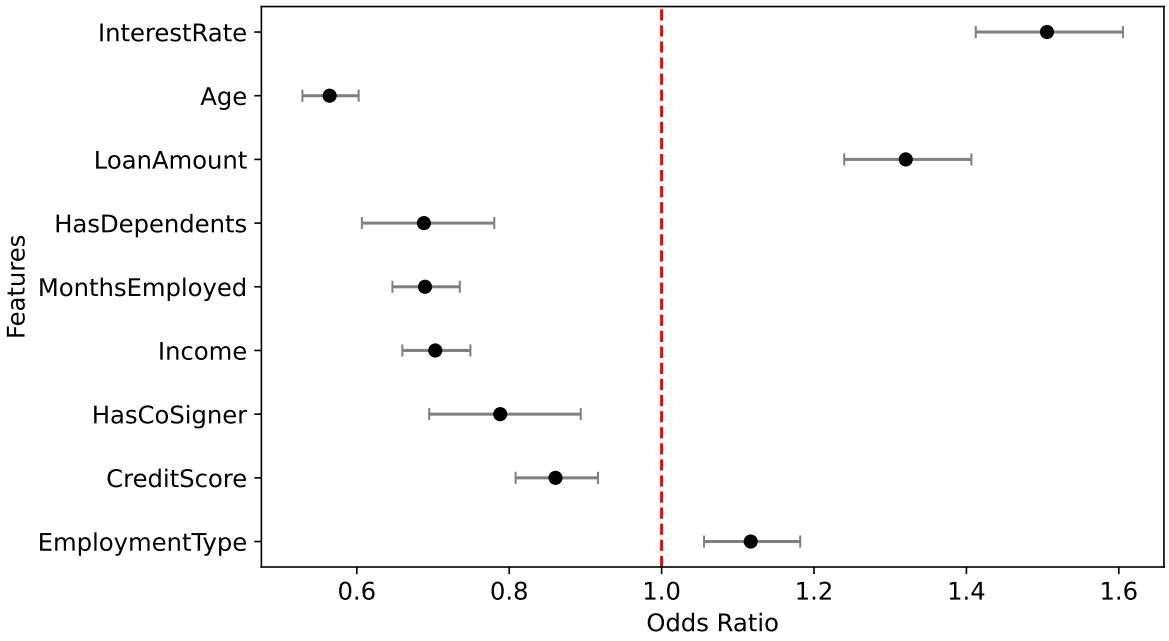

Figure ?: Odds Ratios with 95% Confidence Intervals

3.2 Random Forest

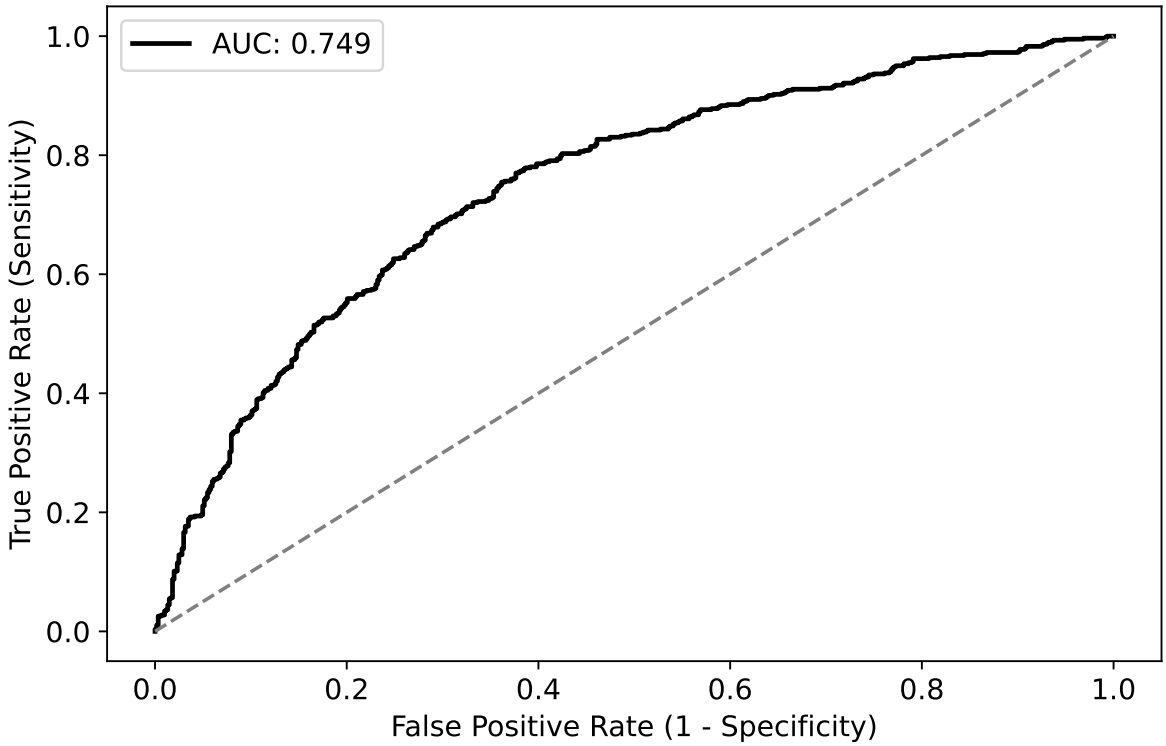

Figure ?: ROC Curve for Random Forest Model

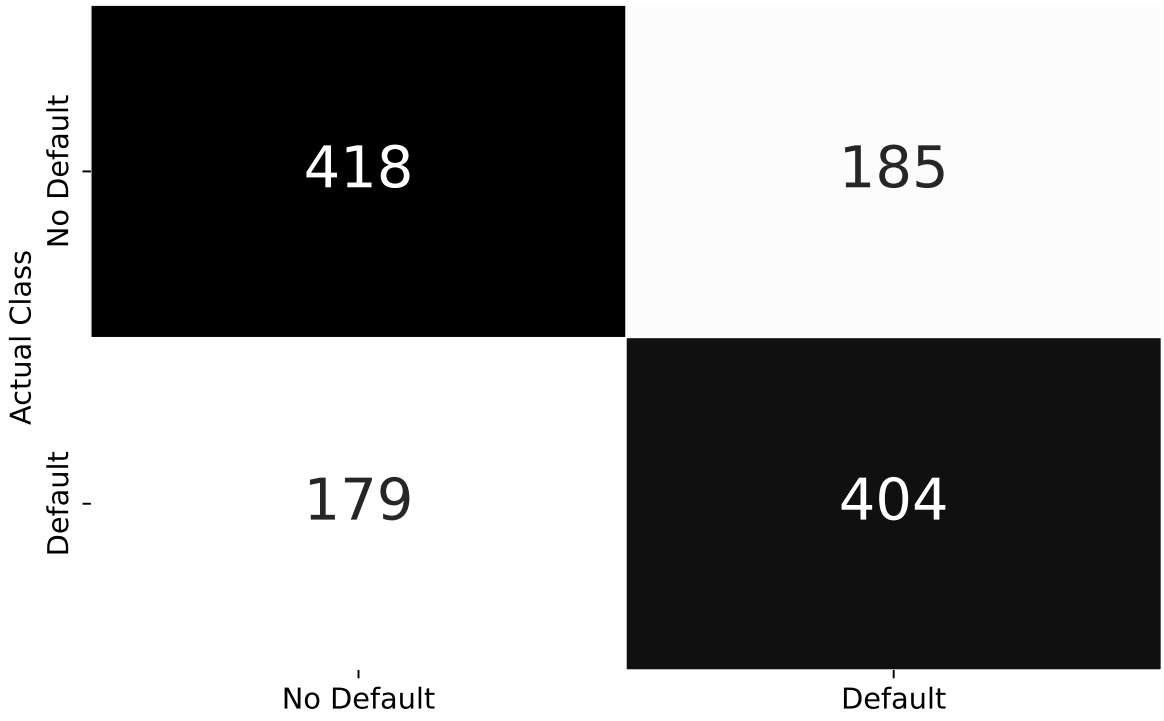

Figure ?: Confusion Matrix for Random Forest Model

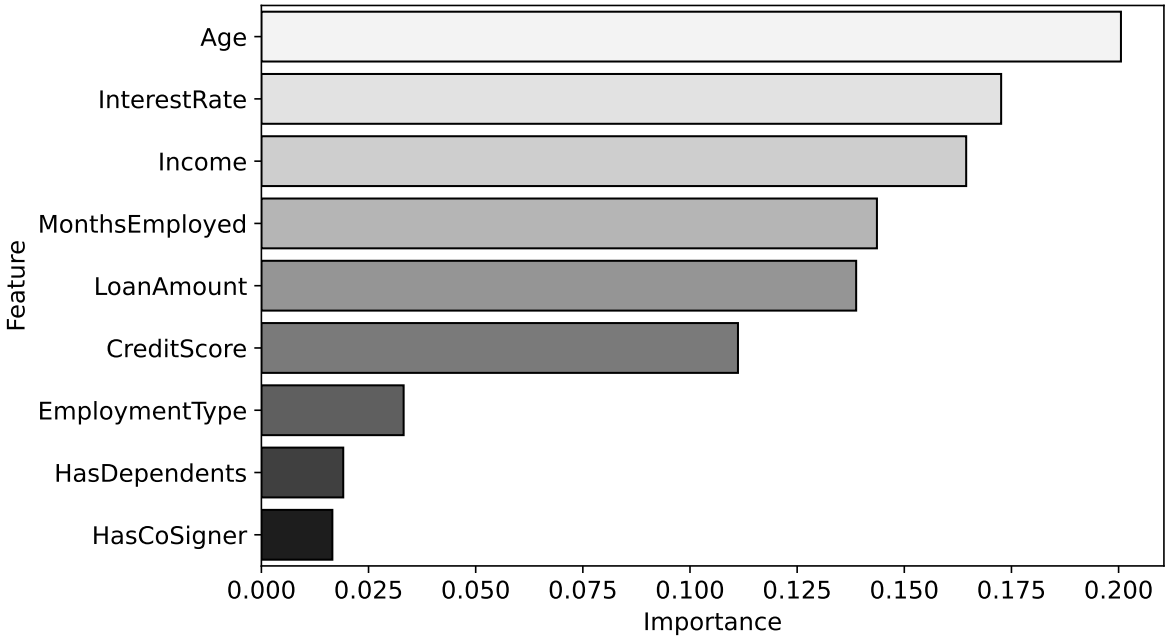

Figure ?: Feature Importances from Random Forest Model

3.3 XGBoost

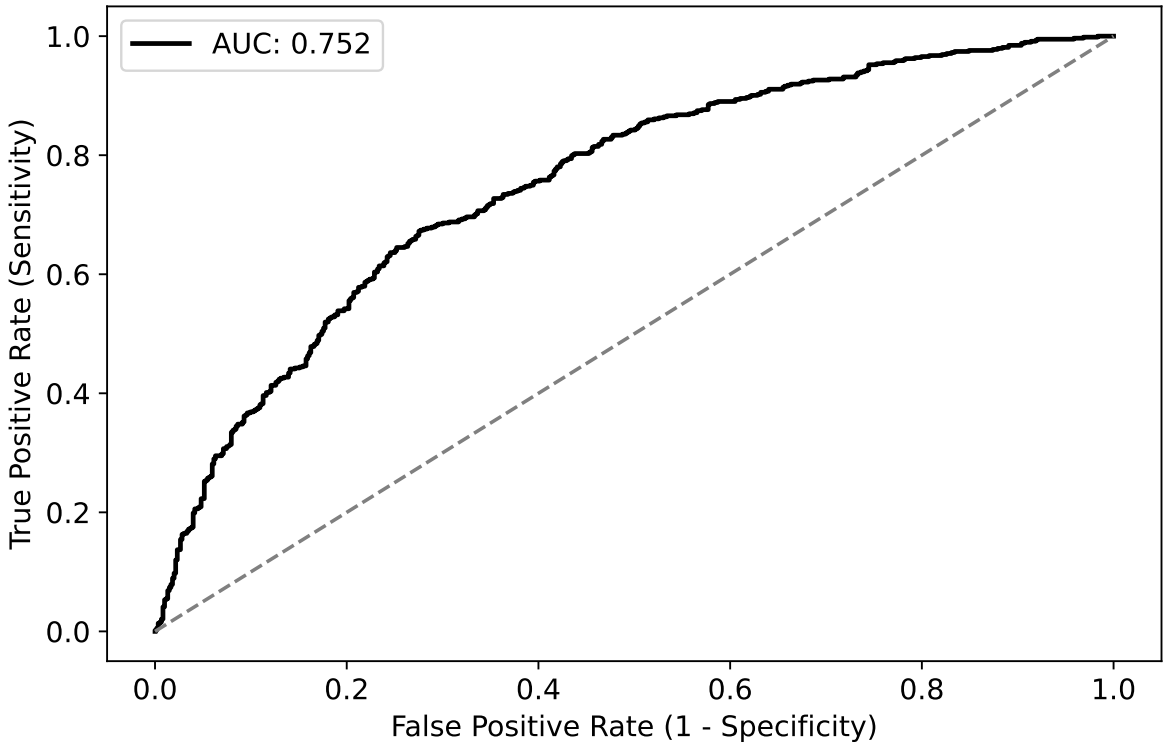

Figure ?: ROC Curve for XGBoost Model

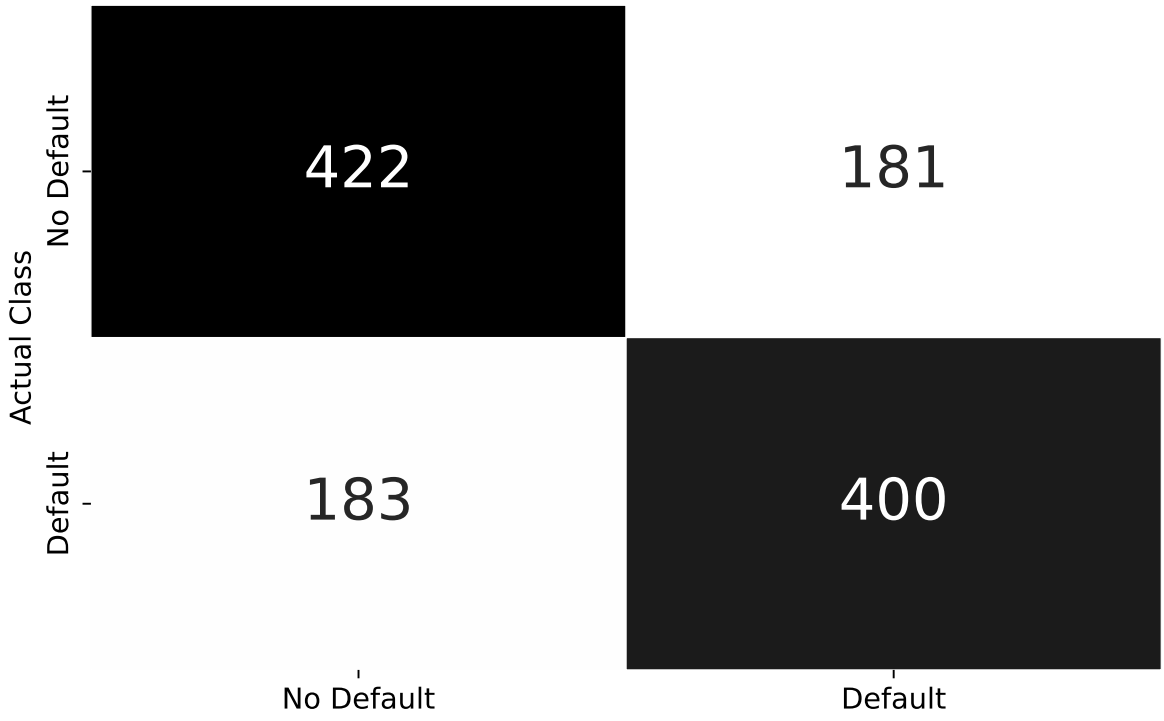

Figure ?: Confusion Matrix for XGBoost Model

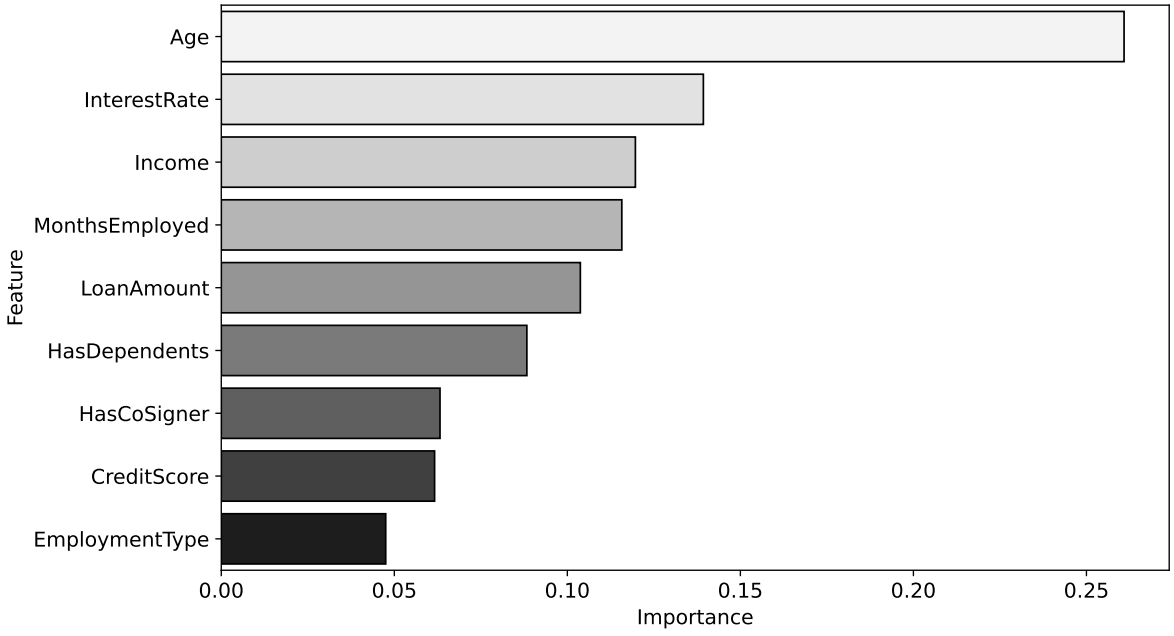

Figure ?: eature Importances from XGBoost Model

3.4 Light Gradient Boosting Machine (LGBM)

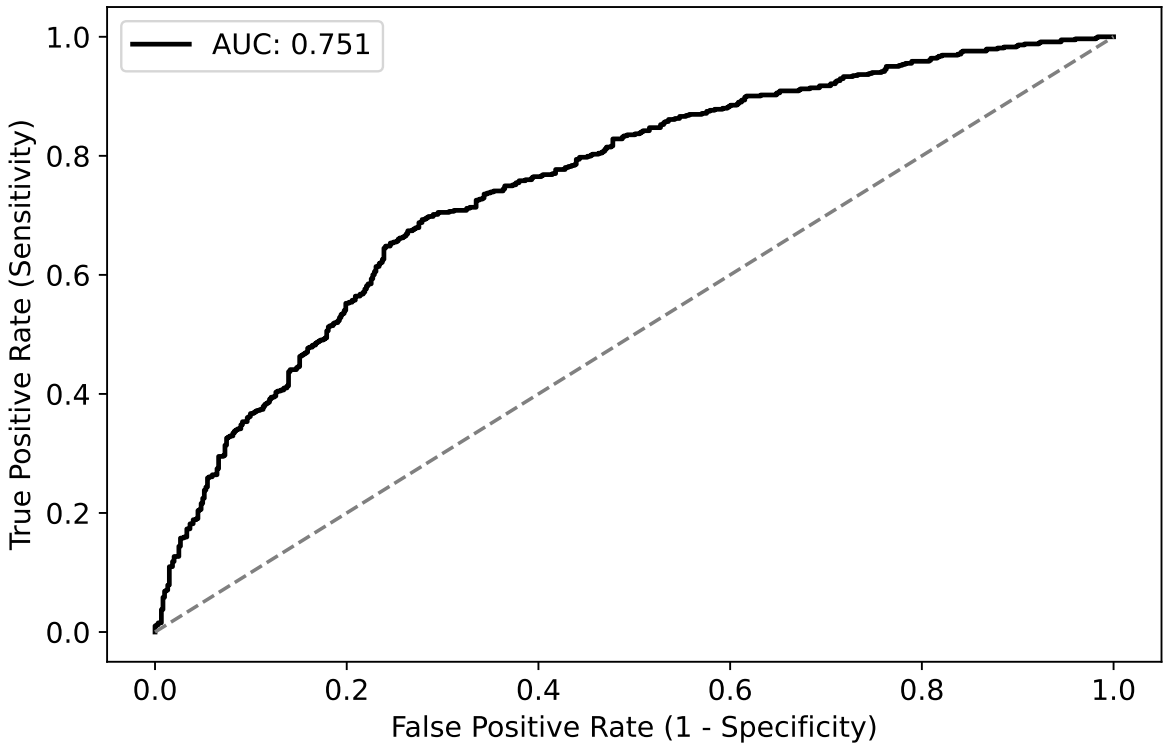

Figure ?: ROC Curve for LightGBM Model

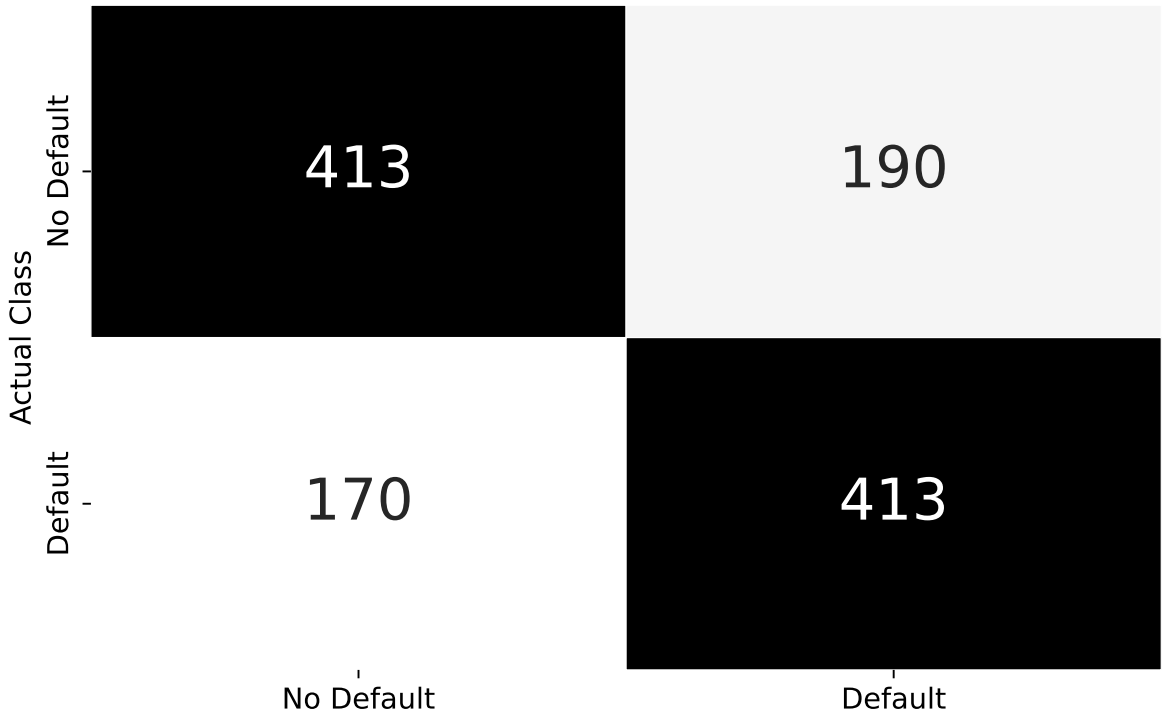

Figure ?: Confusion Matrix for LightGBM Model

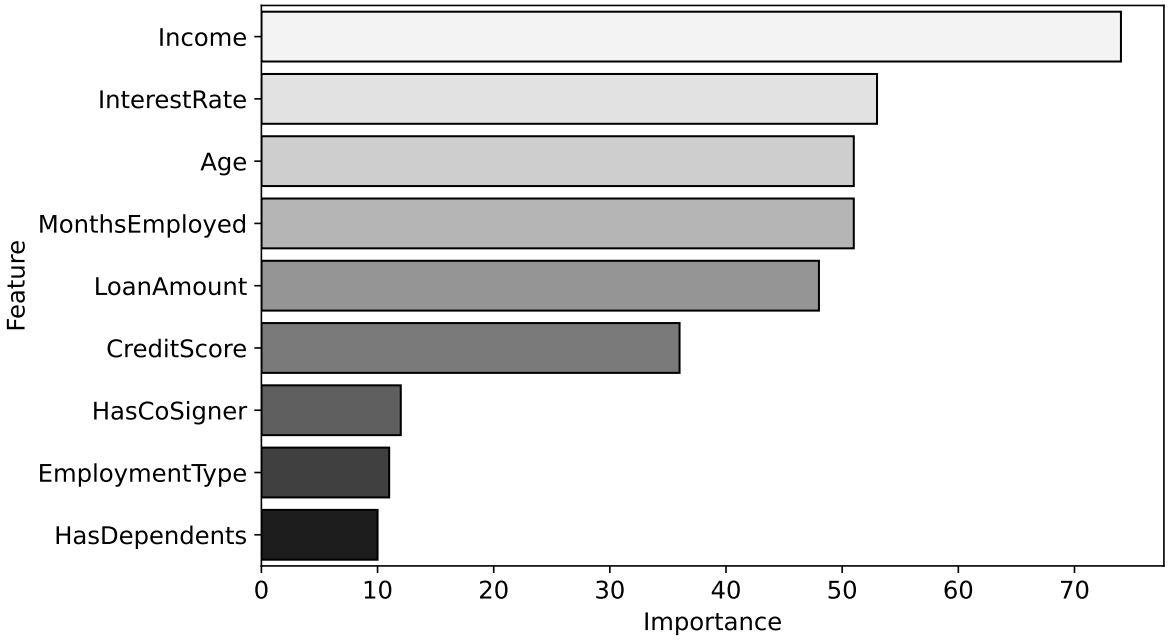

Figure ?: Feature Importances from LightGBM Model

3.5 Model Evaluation and Comparisons

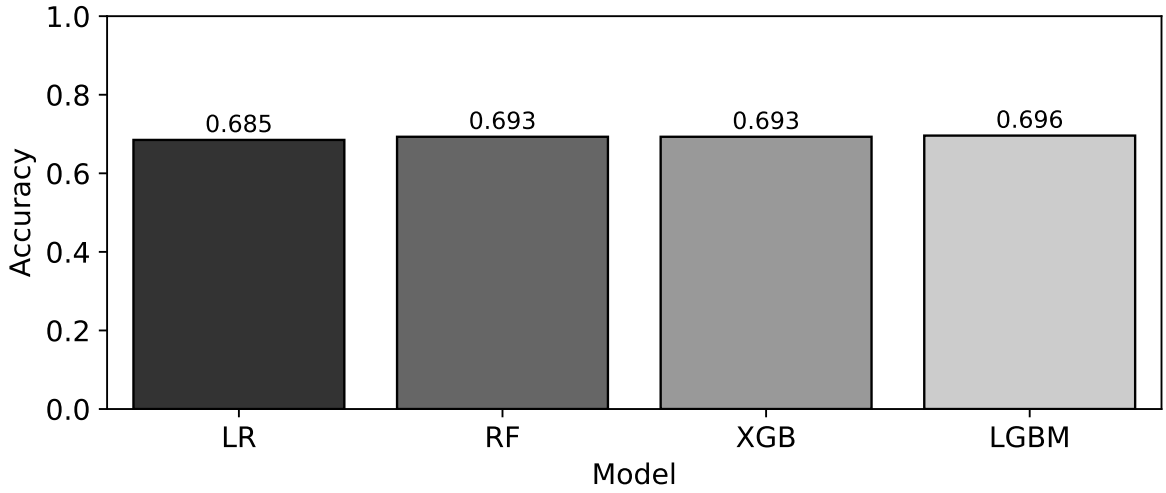

Figure ?: Accuracy for Each Model

Table 4: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
LR	0.685	0.676	0.691	0.684	0.76	11.336
RF	0.693	0.686	0.693	0.689	0.749	11.062
XGB	0.693	0.688	0.686	0.687	0.752	11.062
LGBM	0.696	0.685	0.708	0.696	0.751	10.941

4. Conclusion

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