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

BEE2041 - Data Science in Economics

Student Number - 720017170

Table of contents

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, as loan defaulting can lead to substantial financial losses for both the company and for stakeholders. The ability to predict these defaults is vital for lending institutions to mitigate their risk and make more informed lending predictions. Recent advancements in machine learning (ML) have aided in the development of robust predictive models that outperform traditional credit-scoring methods (Yang, 2024)

Ensemble methods such as Random Forest (RF), XGBoost, and Light Gradient Boosting Machines (LGBM), have shown significant promise in improving classification accuracy over traditional statistical methods (Yadav, 2025). These models offer enhanced predictive capacity due to their ability to capture non-linear relationships in borrower data, providing financial institutions with more reliable risk assessment (Roy, 2025)

This study aims to explore a data-driven approach to credit risk analysis by using ML methods to predict loan defaulting. Logistic regression (LR), RF, XGBoost and LGBM have all been implemented and compared using standard performance metrics such as accuracy, precision, recall, F1-score and area under the curve (AUC). Moreover, exploratory data analysis will be conducted to examine the distribution of important financial variables, identify correlations and allow for optimised feature selection to improve model performance.

Due to the increasing reliance on alternative data sources and advanced computational methods in the financial sector, the results of this study may have significant practical implications. Improved credit risk analysis can help lenders reduce default rates, minimise losses and promote more inclusive access to credit (Ellsworth, 2025). By leveraging the latest ML methods, this project aims to contribute to the growing body of research on predictive analytics in finance and support more robust lending practices (Khoshkhoy Nilash & Esmaeilpour, 2025).

2. Data

Prior to conducting the analysis of credit risk, we need to understand and organise the data. For this analysis we will be using a loan defaulting dataset from Kaggle (reference), consisting of 16 variables/columns and 255,347 observations.

2.1 Preparing the 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

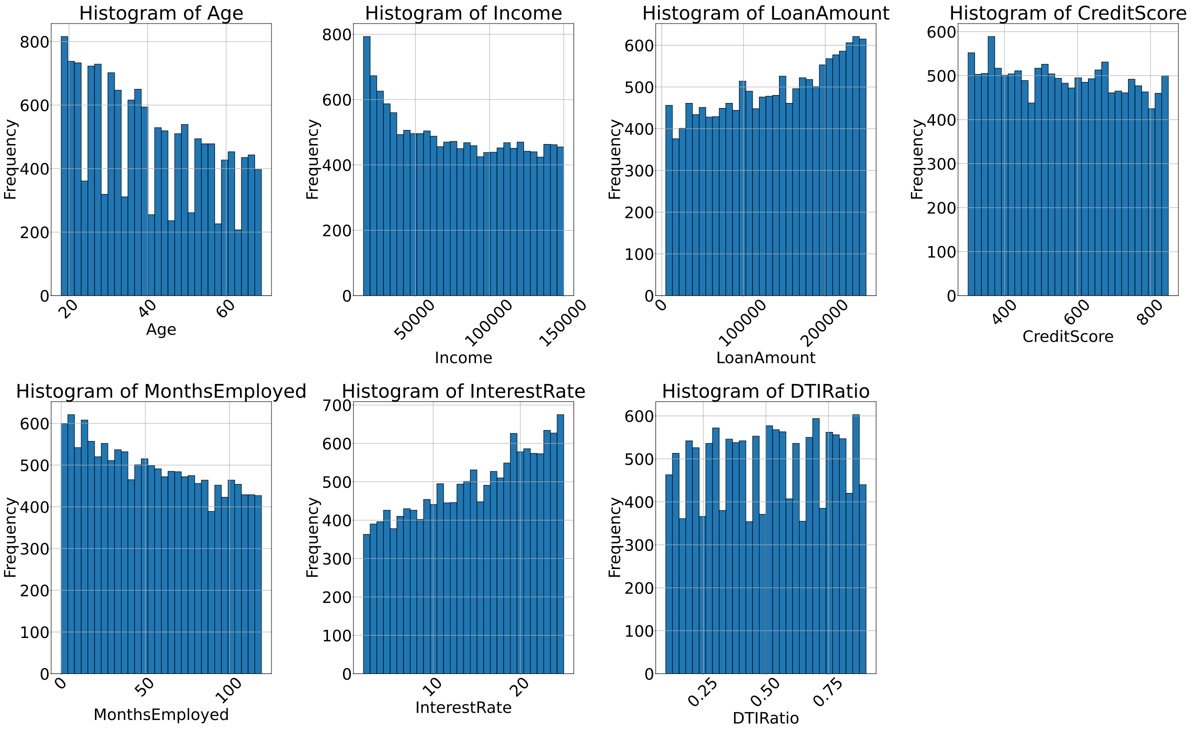


Figure ?: Histograms of all Numeric Variables

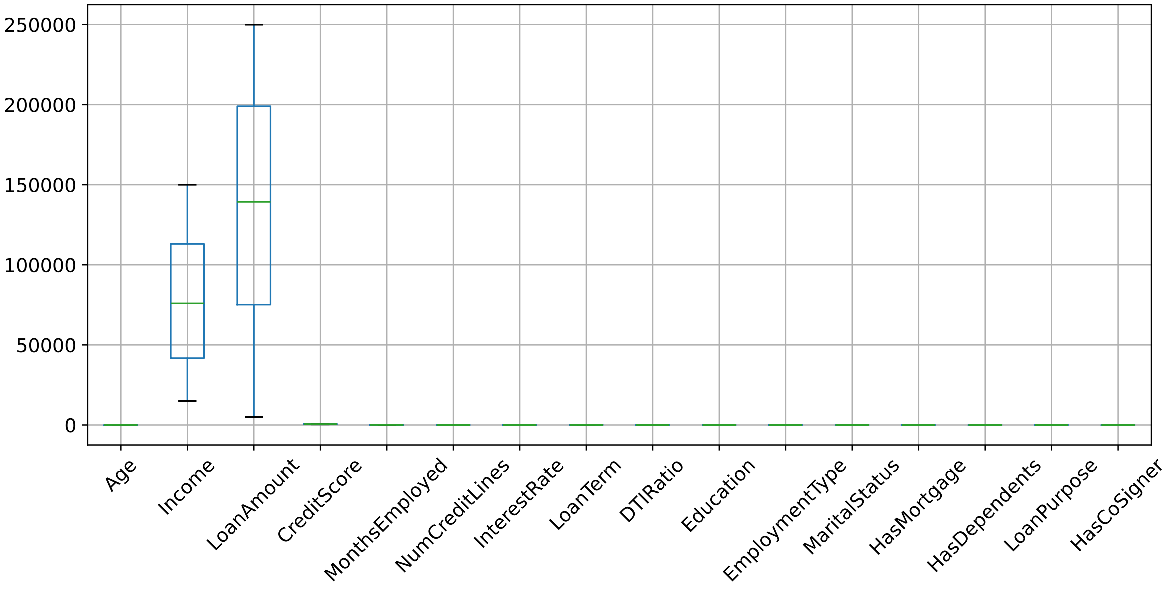


Figure ?: Box Plots of All Variables Before Normalisation

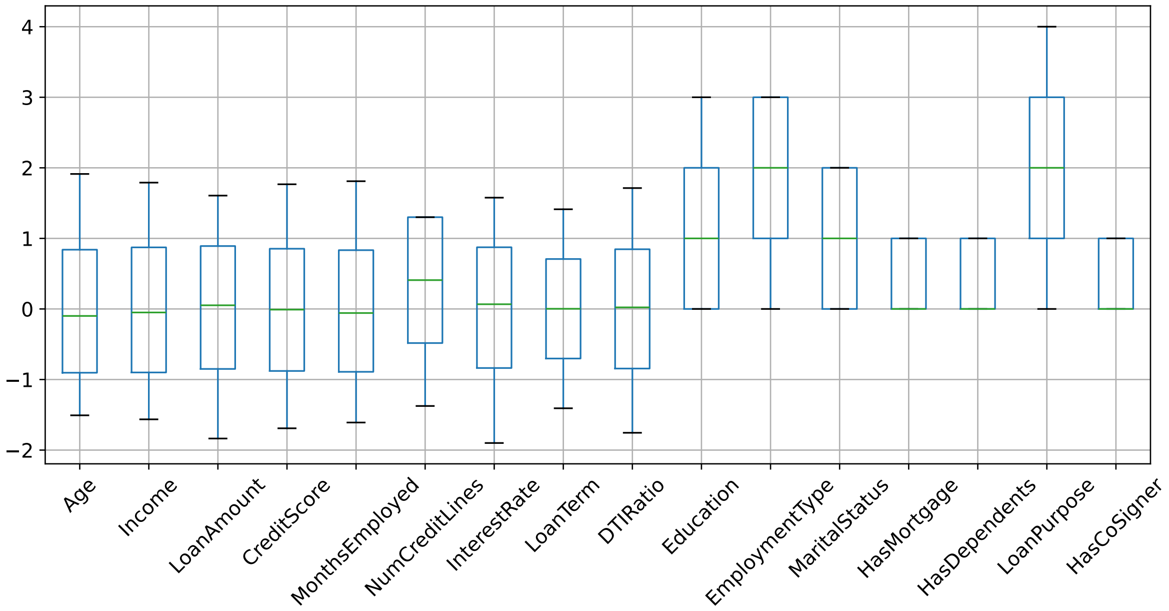


Figure ?: Box Plots of All Variables After Normalisation

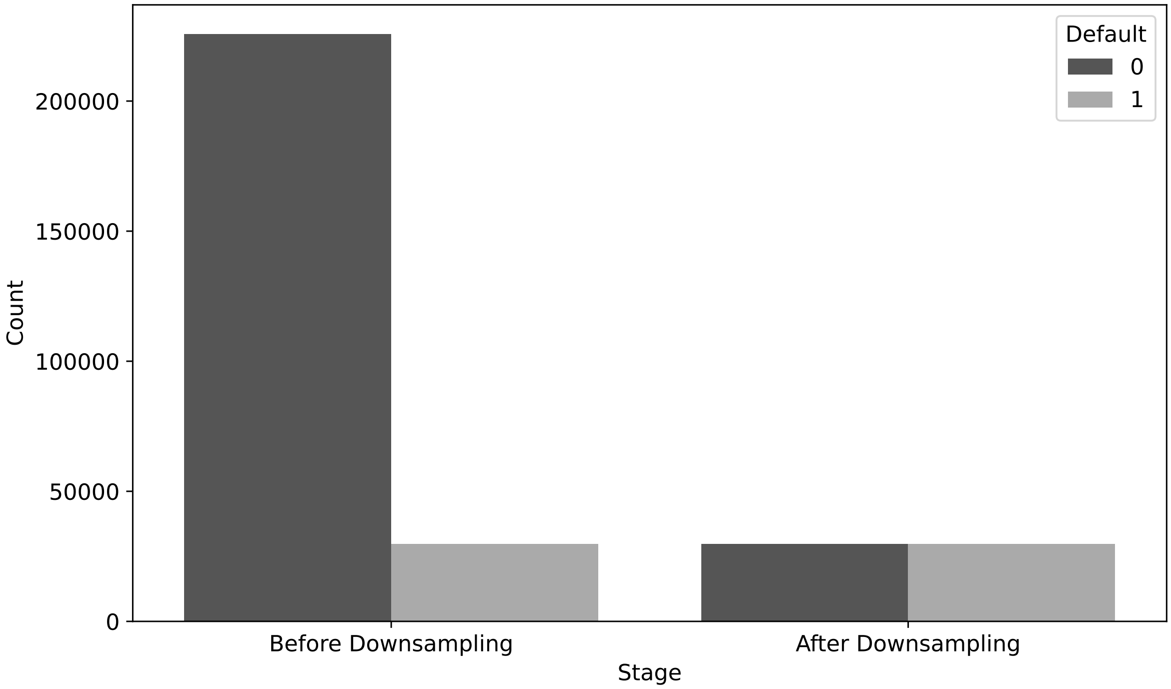
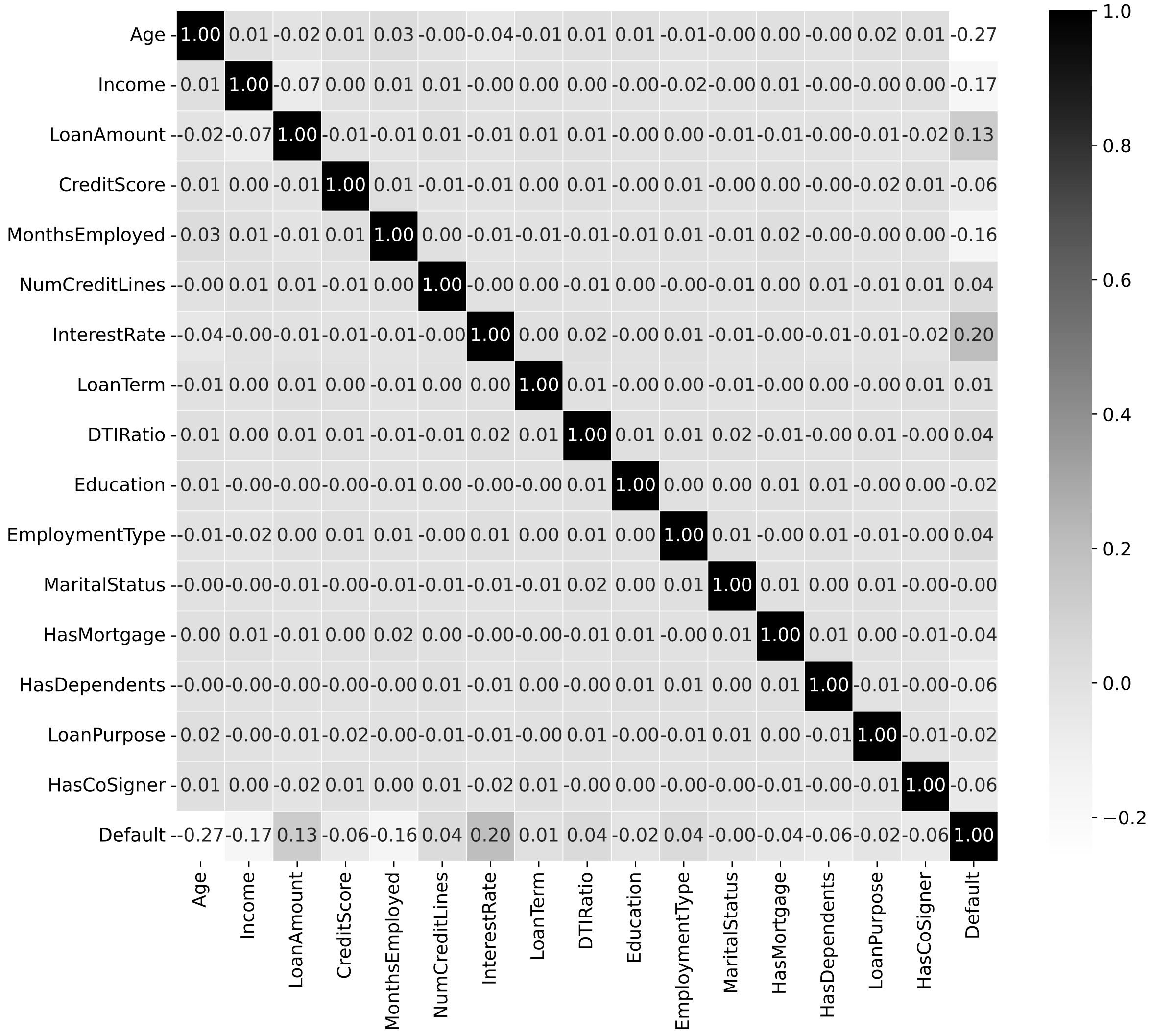


Figure ?: Distribution of Default Before and After Downsampling

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

75% in order to decrease the execution time to be reasonable.

2.3 Correlation Analysis



	Age	Income	LoanAmount	CreditScore	MonthsEmployed	NumCreditLines	InterestRate	LoanTerm	DTIRatio	Education	EmploymentType	MaritalStatus	HasMortgage	HasDependents	LoanPurpose	HasCoSigner	Default
Age	1.00	0.01	-0.02	0.01	0.03	-0.00	-0.04	-0.01	0.01	0.01	-0.01	-0.00	0.00	-0.00	0.02	0.01	-0.27
Income	0.01	1.00	-0.07	0.00	0.01	0.01	-0.00	0.00	0.00	-0.00	-0.02	-0.00	0.01	-0.00	-0.00	0.00	-0.17
LoanAmount	-0.02	-0.07	1.00	-0.01	-0.01	0.01	-0.01	0.01	0.01	-0.00	0.00	-0.01	-0.01	-0.00	-0.01	-0.02	0.13
CreditScore	0.01	0.00	-0.01	1.00	0.01	-0.01	-0.01	0.00	0.01	-0.00	0.01	-0.00	0.00	-0.00	-0.02	0.01	-0.06
MonthsEmployed	0.03	0.01	-0.01	0.01	1.00	0.00	-0.01	-0.01	-0.01	-0.01	0.01	-0.01	0.02	-0.00	-0.00	0.00	-0.16
NumCreditLines	-0.00	0.01	0.01	-0.01	0.00	1.00	-0.00	0.00	-0.01	0.00	-0.00	-0.01	0.00	0.01	-0.01	0.01	0.04
InterestRate	-0.04	-0.00	-0.01	-0.01	-0.01	-0.00	1.00	0.00	0.02	-0.00	0.01	-0.01	-0.00	-0.01	-0.01	-0.02	0.20
LoanTerm	-0.01	0.00	0.01	0.00	-0.01	0.00	0.00	1.00	0.01	-0.00	0.00	-0.01	-0.00	0.00	-0.00	0.01	0.01
DTIRatio	0.01	0.00	0.01	0.01	-0.01	-0.01	0.02	0.01	1.00	0.01	0.01	0.02	-0.01	-0.00	0.01	-0.00	0.04
Education	0.01	-0.00	-0.00	-0.00	-0.01	0.00	-0.00	-0.00	0.01	1.00	0.00	0.00	0.01	0.01	-0.00	0.00	-0.02
EmploymentType	-0.01	-0.02	0.00	0.01	0.01	-0.00	0.01	0.00	0.01	0.00	1.00	0.01	-0.00	0.01	-0.01	-0.00	0.04
MaritalStatus	-0.00	-0.00	-0.01	-0.00	-0.01	-0.01	-0.01	-0.01	0.02	0.00	0.01	1.00	0.01	0.00	0.01	-0.00	-0.00
HasMortgage	0.00	0.01	-0.01	0.00	0.02	0.00	-0.00	-0.00	-0.01	0.01	-0.00	0.01	1.00	0.01	0.00	-0.01	-0.04
HasDependents	-0.00	-0.00	-0.00	-0.00	-0.00	0.01	-0.01	0.00	-0.00	0.01	0.01	0.00	0.01	1.00	-0.01	-0.00	-0.06
LoanPurpose	0.02	-0.00	-0.01	-0.02	-0.00	-0.01	-0.01	-0.00	0.01	-0.00	-0.01	0.01	0.00	-0.01	1.00	-0.01	-0.02
HasCoSigner	0.01	0.00	-0.02	0.01	0.00	0.01	-0.02	0.01	-0.00	0.00	-0.00	-0.00	-0.01	-0.00	-0.01	1.00	-0.06
Default	-0.27	-0.17	0.13	-0.06	-0.16	0.04	0.20	0.01	0.04	-0.02	0.04	-0.00	-0.04	-0.06	-0.02	-0.06	1.00

Figure ?: Correlation Plot of All Variables

Table 3: Variance Inflation Factor (VIF) Values

Feature	VIF
Age	1.004023
Income	1.006059
LoanAmount	1.006939
CreditScore	1.001777
MonthsEmployed	1.003127
NumCreditLines	1.001308
InterestRate	1.003661
LoanTerm	1.000665
DTIRatio	1.001484
Education	1.000544
EmploymentType	1.000999
MaritalStatus	1.001289
HasMortgage	1.001233
HasDependents	1.000986
LoanPurpose	1.001842
HasCoSigner	1.001298

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

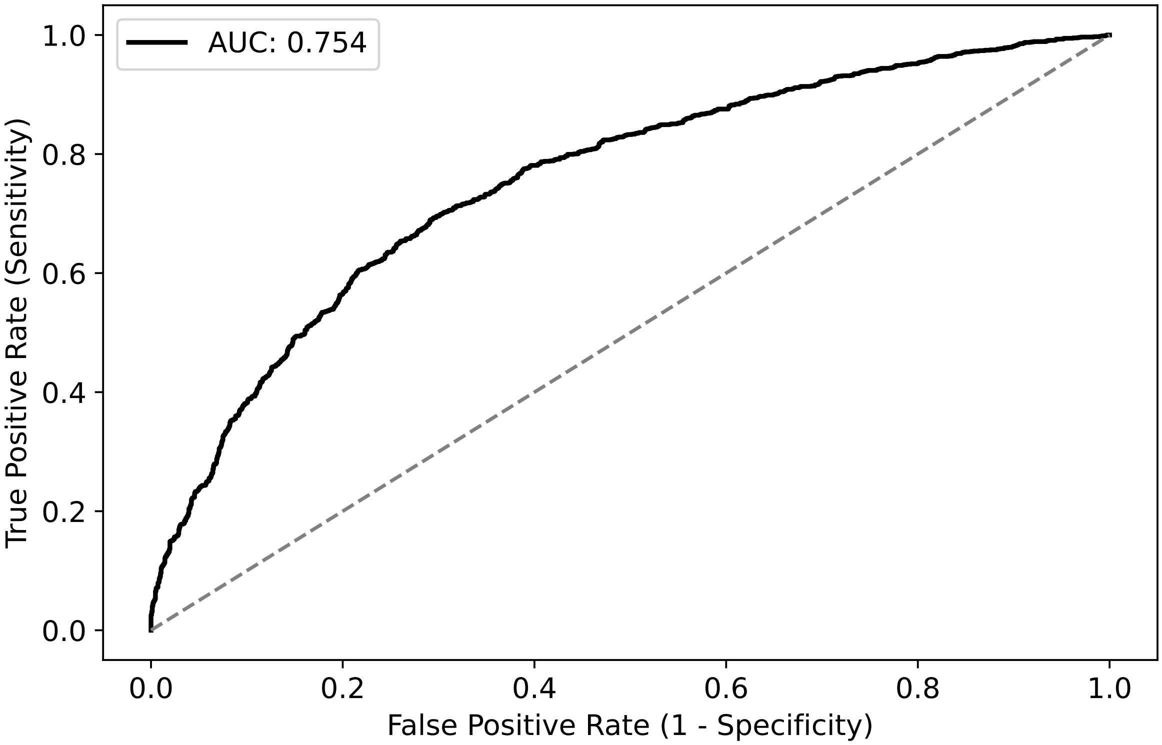


Figure ?: ROC Curve for Logistic Regression Mod

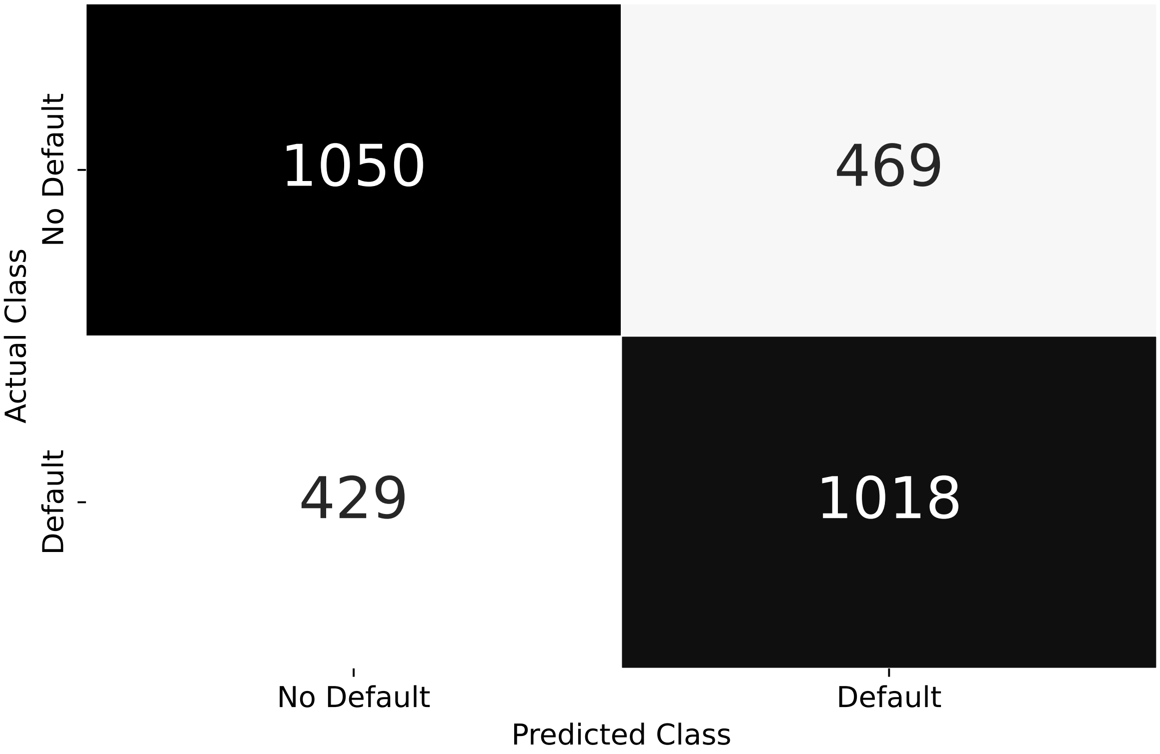


Figure ?: Confusion Matrix for Logistic Regression Model

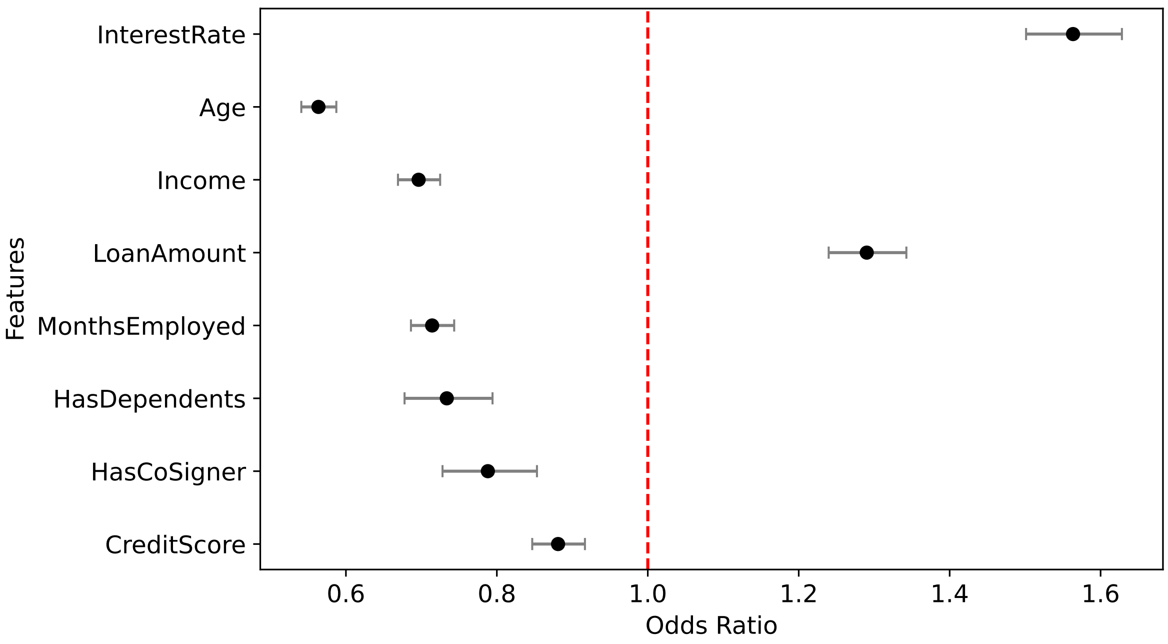


Figure ?: Odds Ratios with 95% Confidence Intervals

3.2 Random Forest

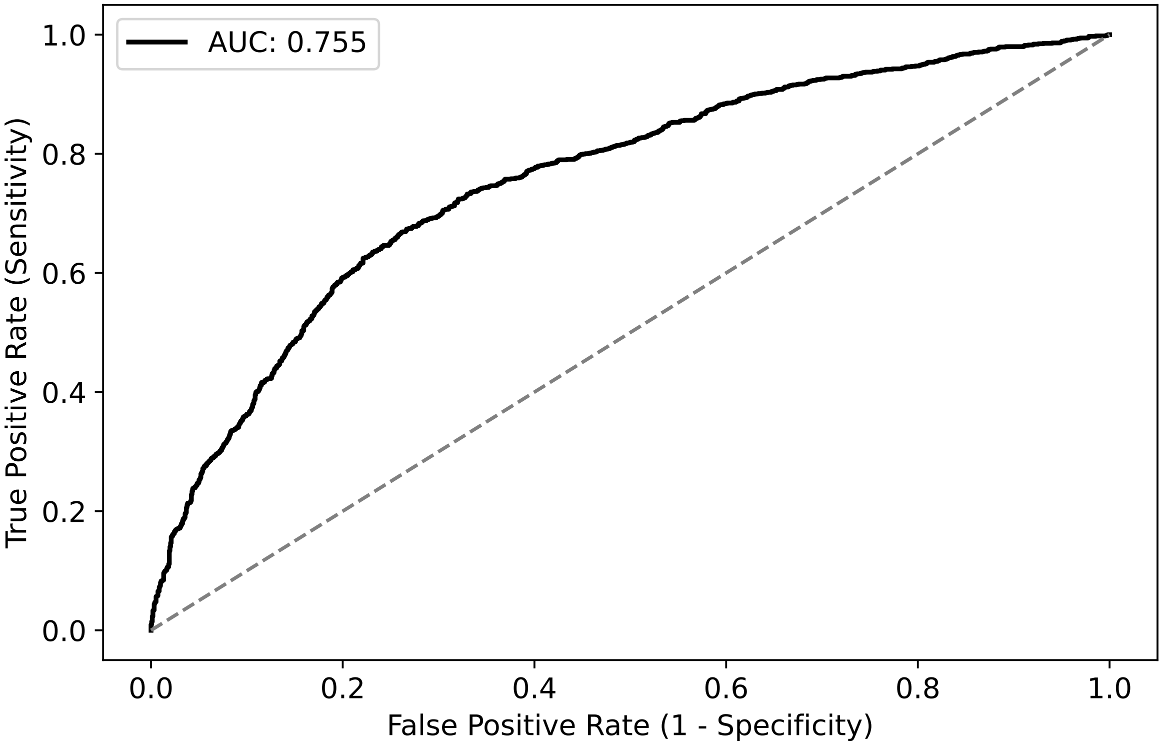


Figure ?: ROC Curve for Random Forest Model

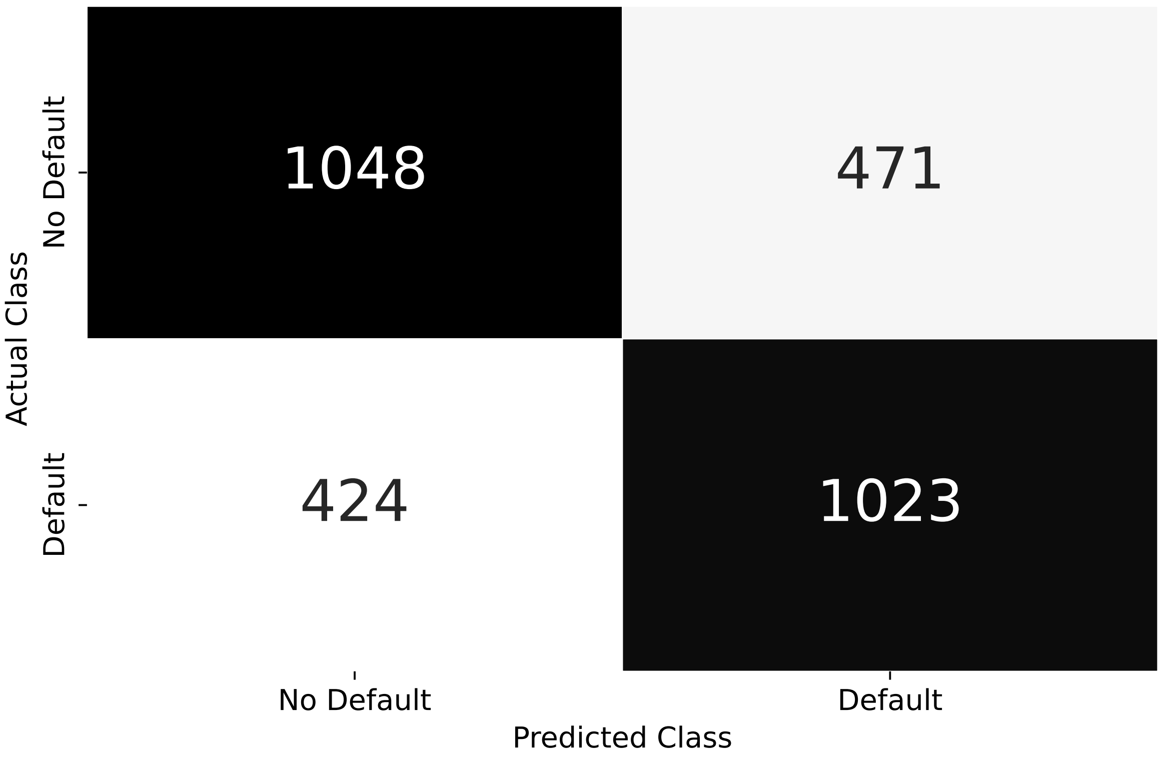


Figure ?: Confusion Matrix for Random Forest Model

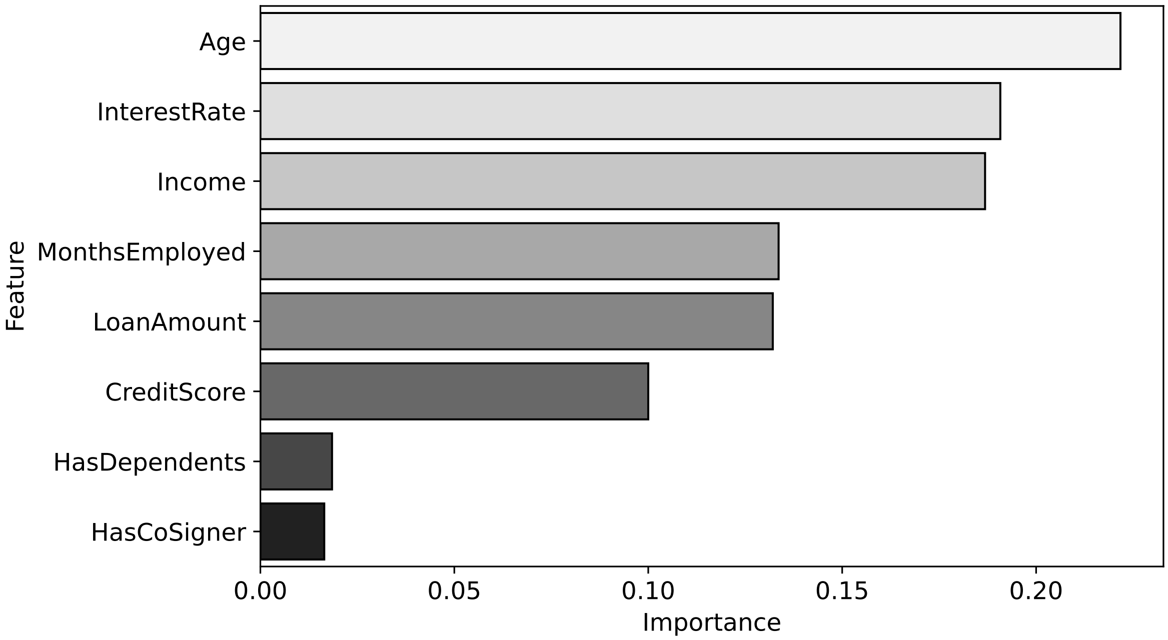


Figure ?: Feature Importances from Random Forest Model

3.3 XGBoost

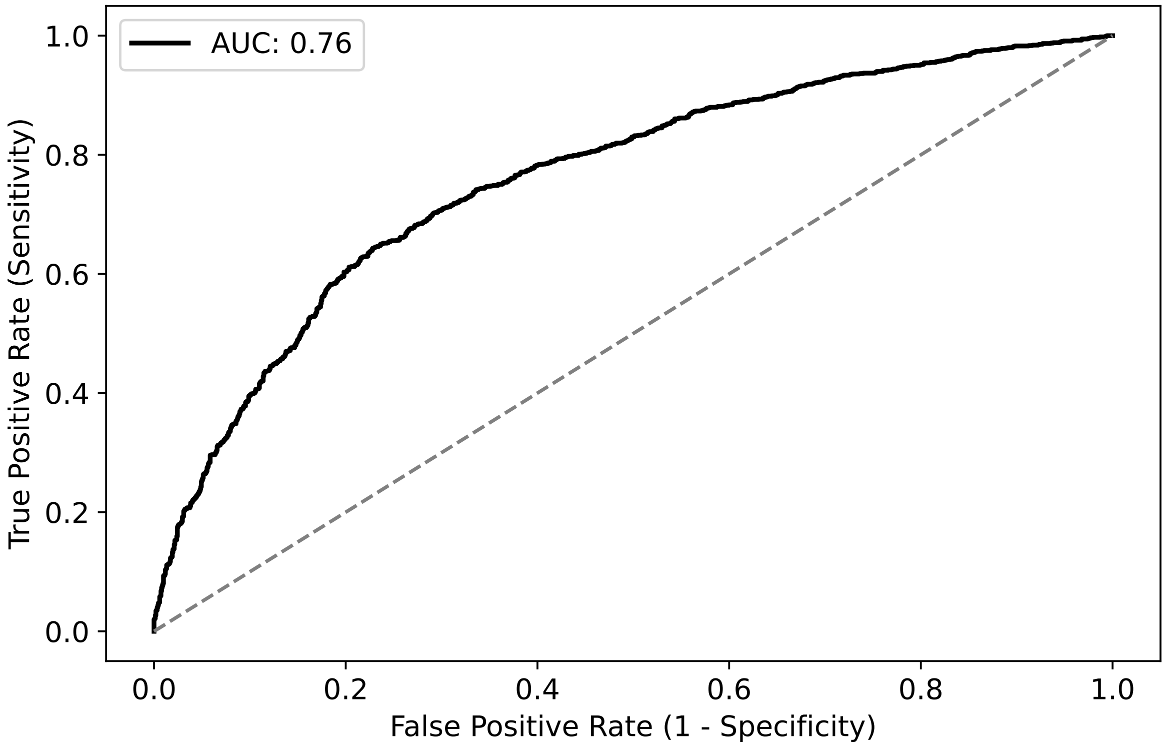


Figure ?: ROC Curve for XGBoost Model

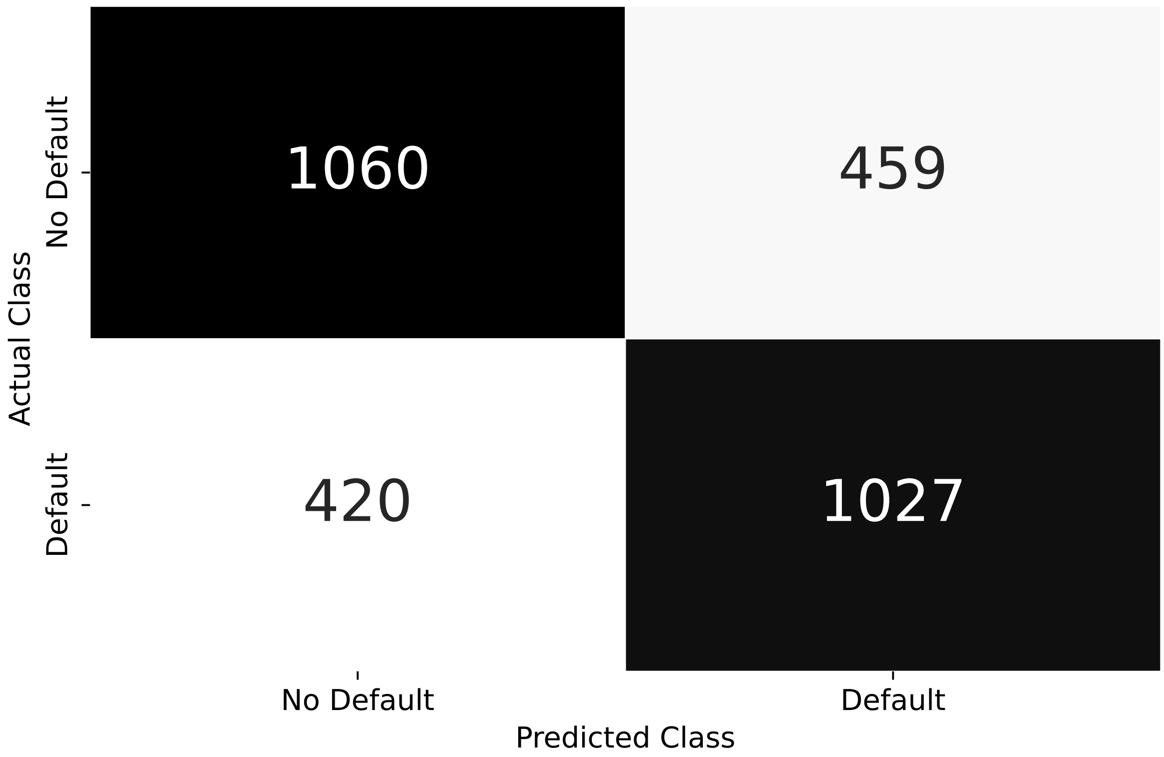


Figure ?: Confusion Matrix for XGBoost Model

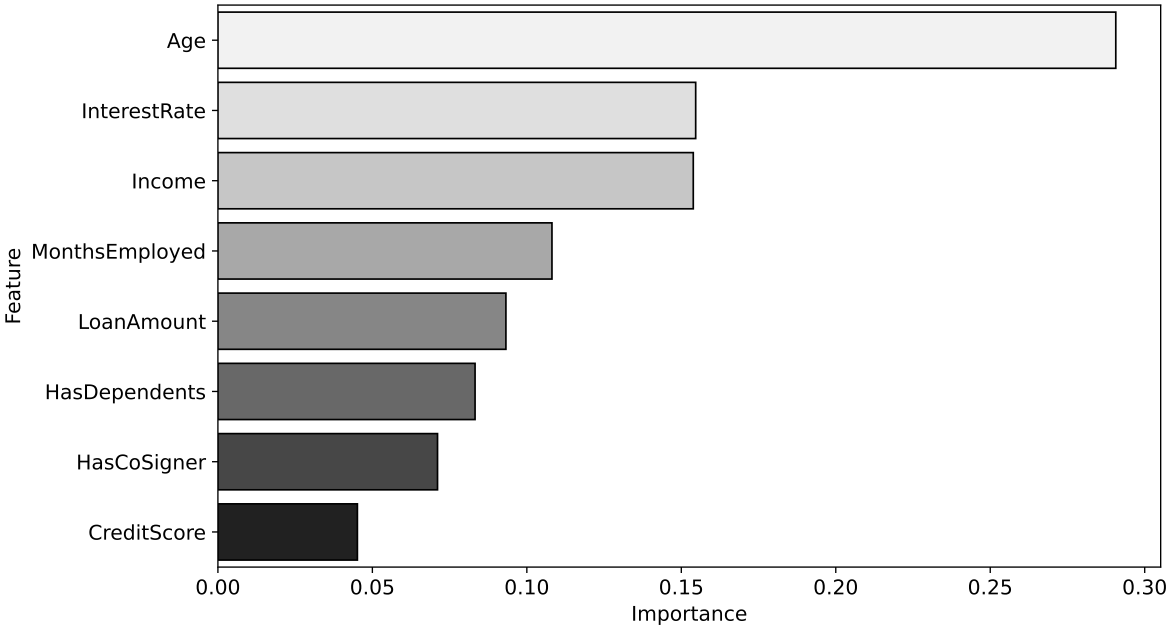


Figure ?: eature Importances from XGBoost Model

3.4 Light Gradient Boosting Machine (LGBM)

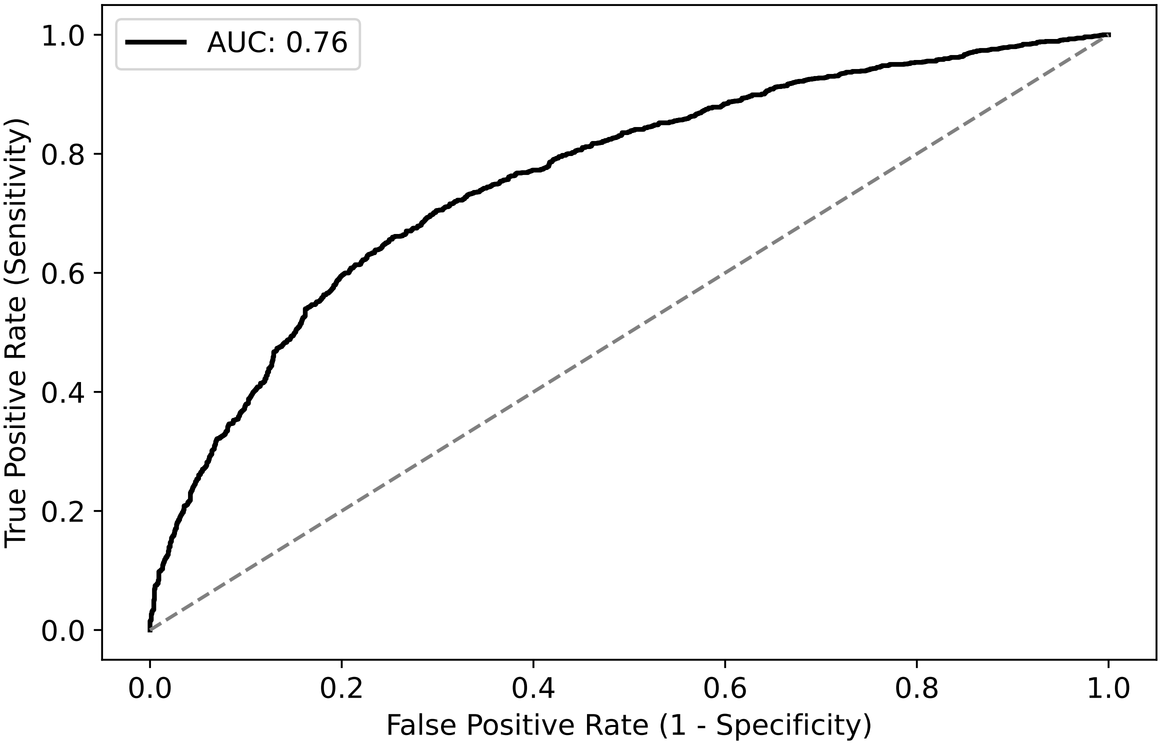


Figure ?: ROC Curve for LightGBM Model

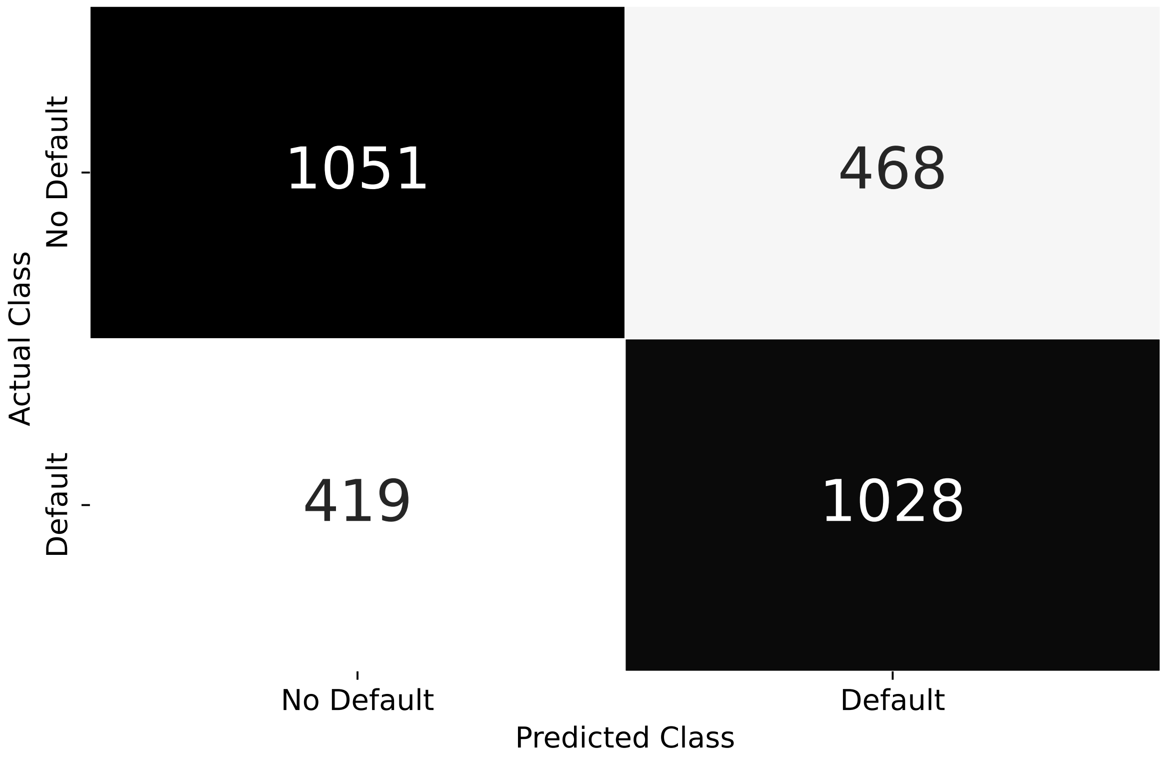


Figure ?: Confusion Matrix for LightGBM Model

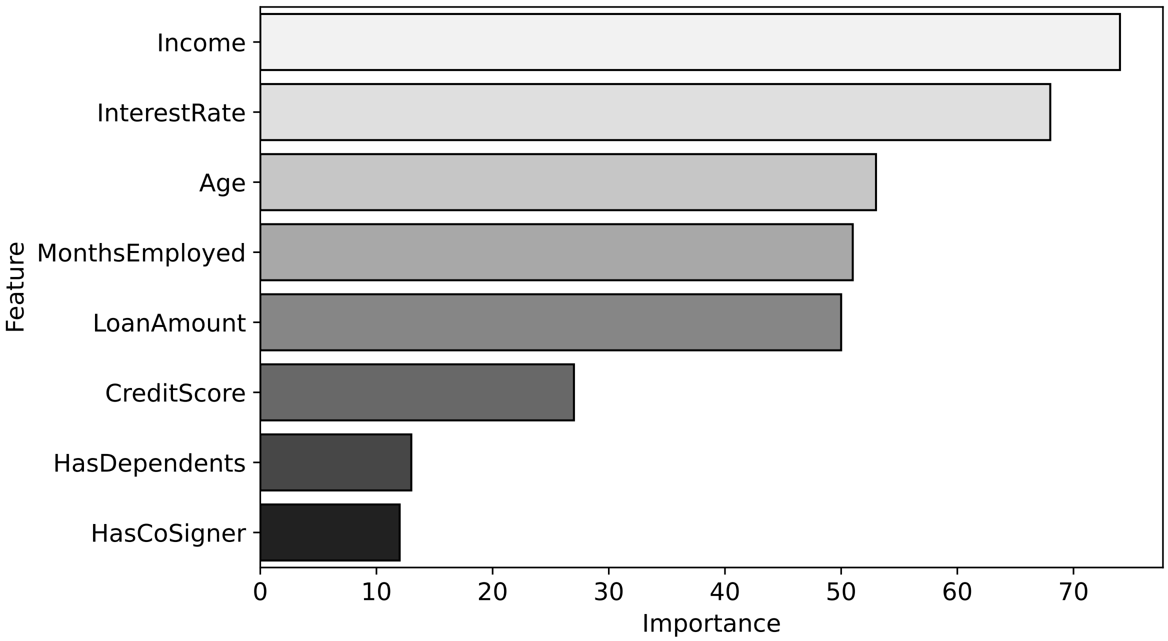


Figure ?: Feature Importances from LightGBM Model

3.5 Model Evaluation and Comparisons

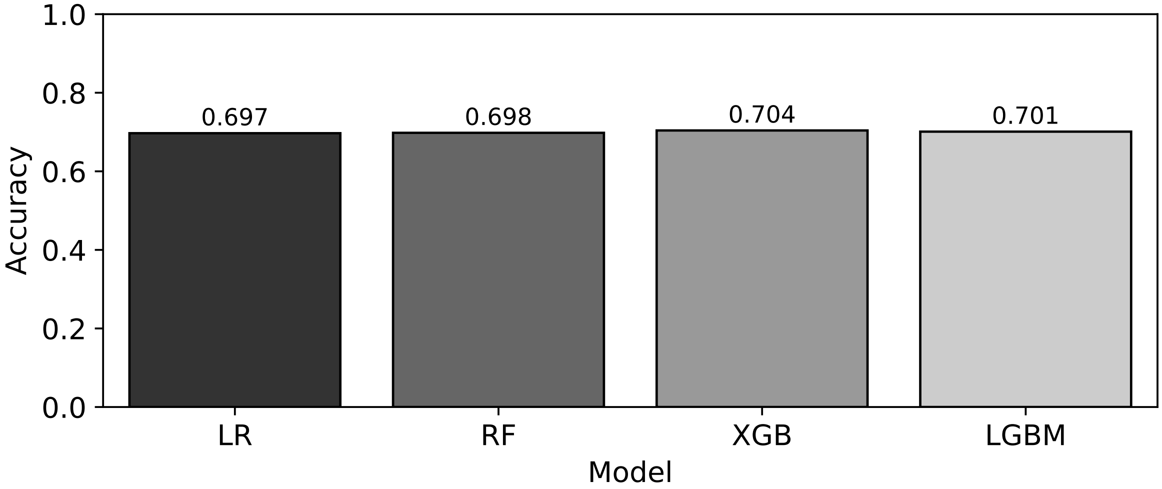


Figure ?: Accuracy for Each Model

Table 4: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss
LR	0.697	0.685	0.704	0.694	0.754	10.913
RF	0.698	0.685	0.707	0.696	0.755	10.876
XGB	0.704	0.691	0.71	0.7	0.76	10.682
LGBM	0.701	0.687	0.71	0.699	0.76	10.779

4. Conclusion

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