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 12 variables/columns and 28,501 observations.

2.1 Preparing the Data

Table 1: Variable Information

Variable	Data Type	Definition
PersonAge	int64	Age of the borrower
PersonIncome	int64	Income of the borrower
PersonHomeOwnership	object	Home ownership of the borrower
PersonEmpLength	float64	Employment length of the borrower
LoanIntent	object	Intention of the loan
LoanGrade	object	Loan grade
LoanAmnt	int64	Amount of the loan (USD)
LoanIntRate	float64	Loan interest rate
LoanStatus	int64	Loan status (0 - not defaulted, 1 - defaulted)
LoanPercentIncome	float64	Loan percentage of income
PreviousDefault	object	If the borrower has defaulted before
CredHistory	int64	Credit history length

2.2 Descriptive Statistics

Table 2: Summary Statistics of Numeric Variables

Variable	N	Mean	Median	SD	Min	Max
PersonAge	28500.0	27.7	26.0	6.3	20.0	144.0
PersonIncome	28500.0	66446.2	56000.0	51532.2	4000.0	2039784.0
PersonEmpLength	28500.0	4.8	4.0	4.2	0.0	123.0
LoanAmnt	28500.0	9658.7	8000.0	6329.7	500.0	35000.0
LoanIntRate	28500.0	11.0	11.0	3.2	5.4	23.2
LoanStatus	28500.0	0.2	0.0	0.4	0.0	1.0
LoanPercentIncome	28500.0	0.2	0.2	0.1	0.0	0.8
CredHistory	28500.0	5.8	4.0	4.0	2.0	30.0

2.3 Distribution Analysis

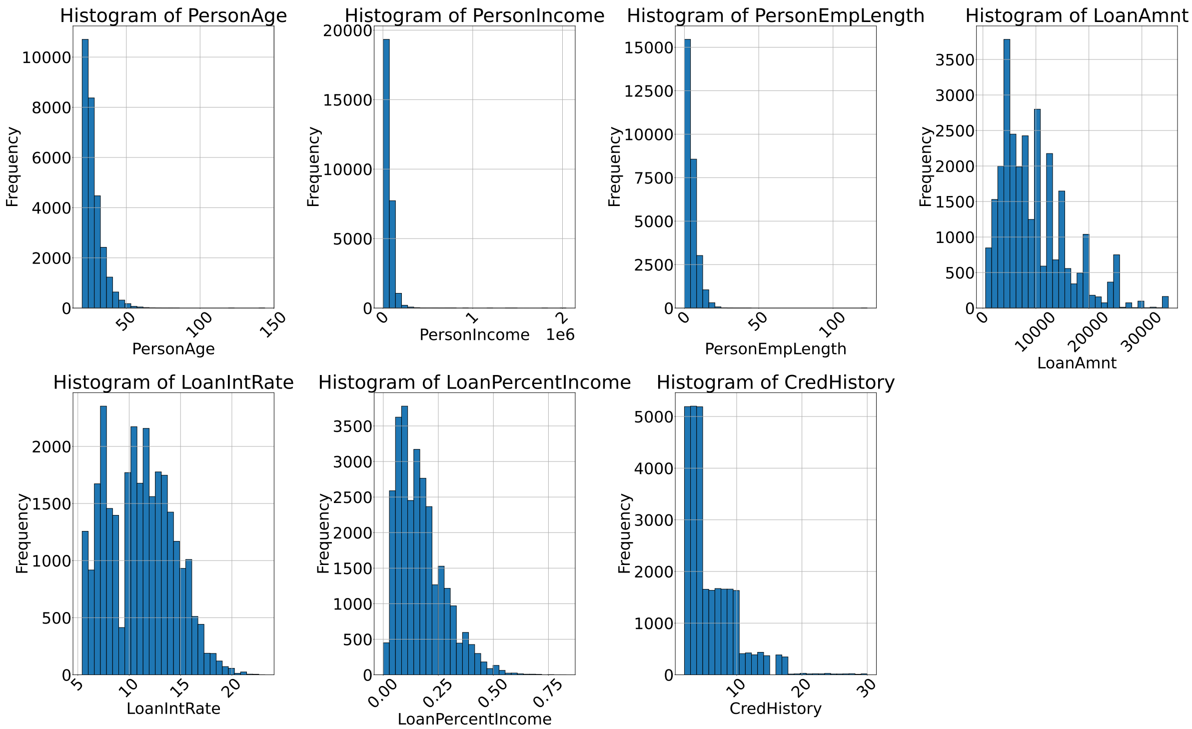

Figure ?: Histograms of all Numeric Variables

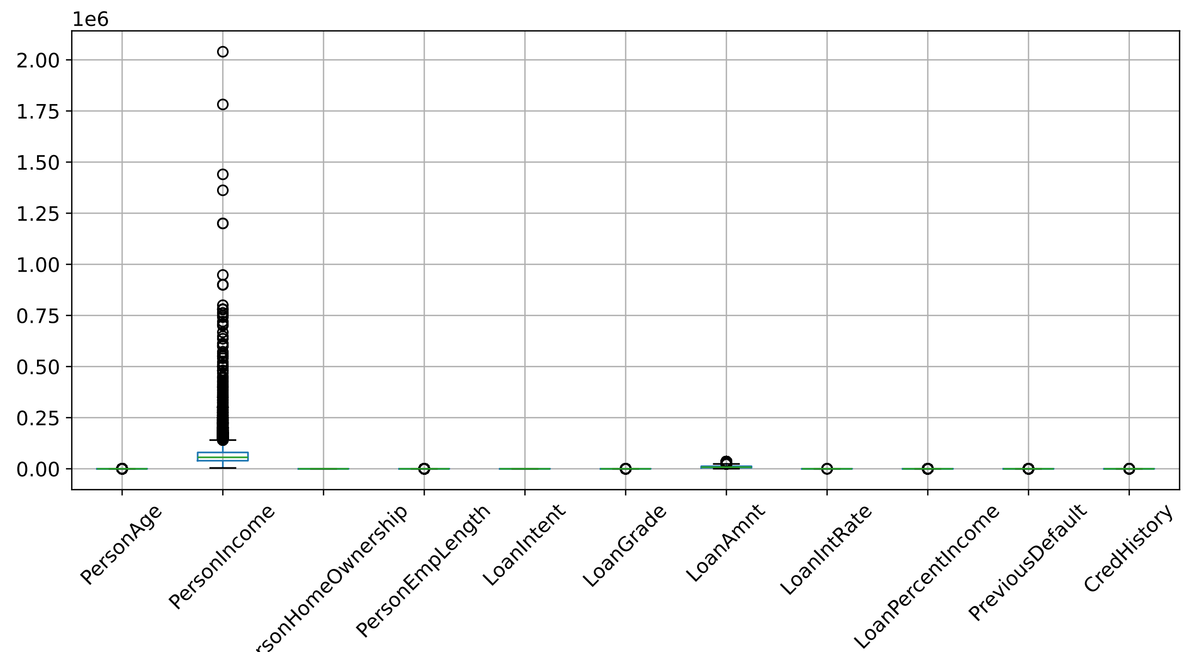

Figure ?: Box Plots of All Variables Before Normalisation

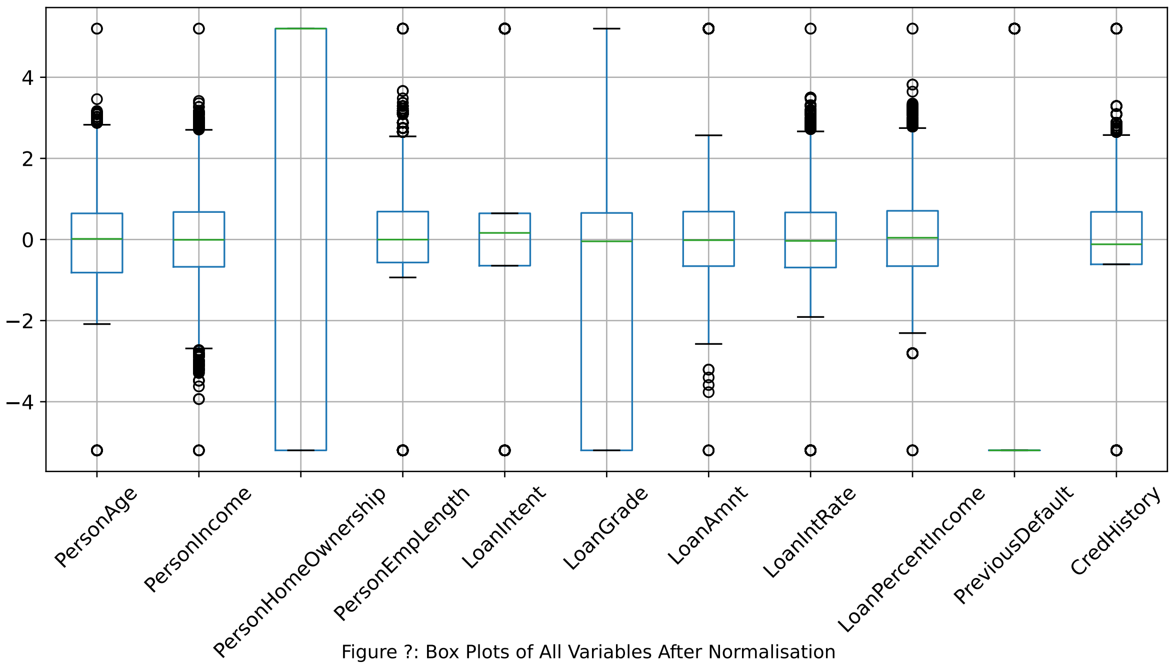

Figure ?: Box Plots of All Variables After Normalisation

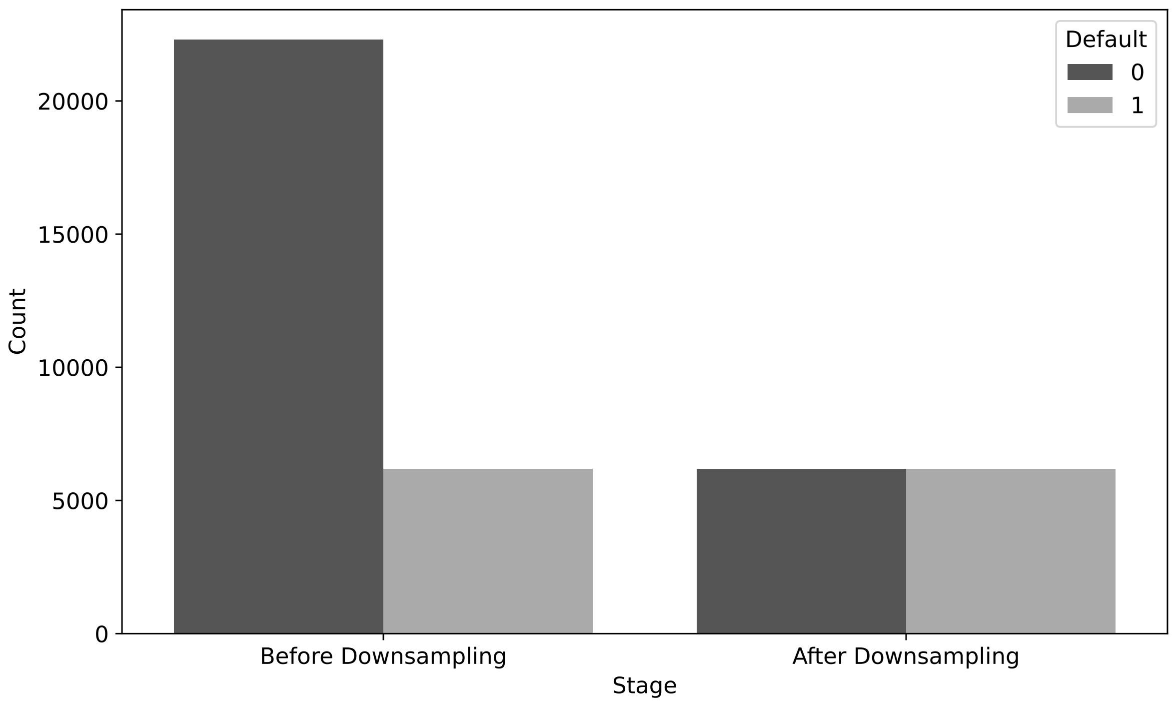

Figure ?: Distribution of Default Before and After Downsampling

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

2.4 Correlation Analysis

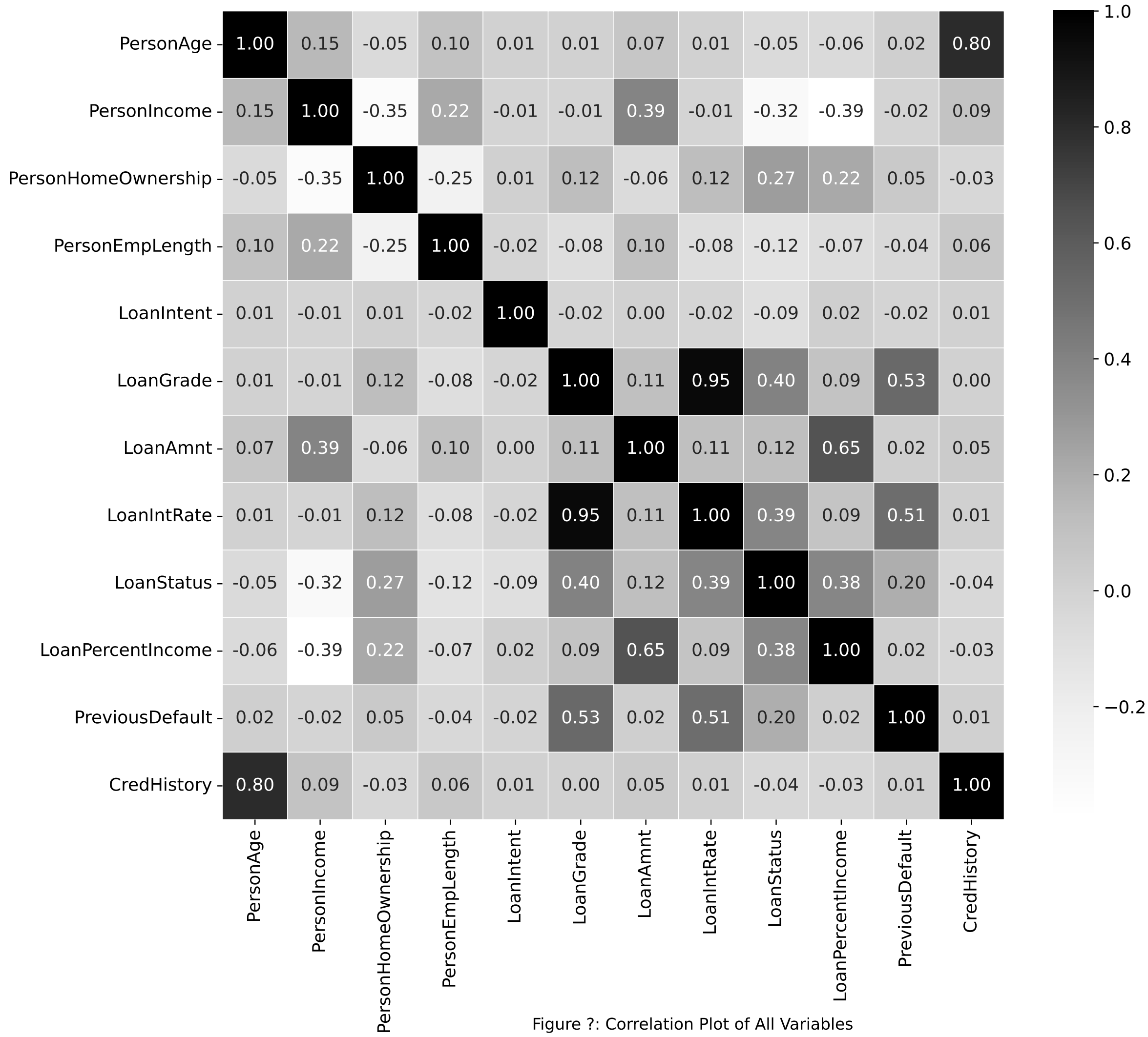

Figure ?: Correlation Plot of All Variables

Table 3: Variance Inflation Factor (VIF) Values

Feature	VIF
PersonAge	1.494000
PersonIncome	9.179000
PersonHomeOwnership	1.190000
PersonEmpLength	1.059000
LoanIntent	1.002000
LoanGrade	2.922000
LoanAmnt	12.343000
LoanIntRate	3.035000
LoanPercentIncome	12.229000
PreviousDefault	1.252000
CredHistory	1.458000

3. Results and Discussion

3.1 Logistic Regression

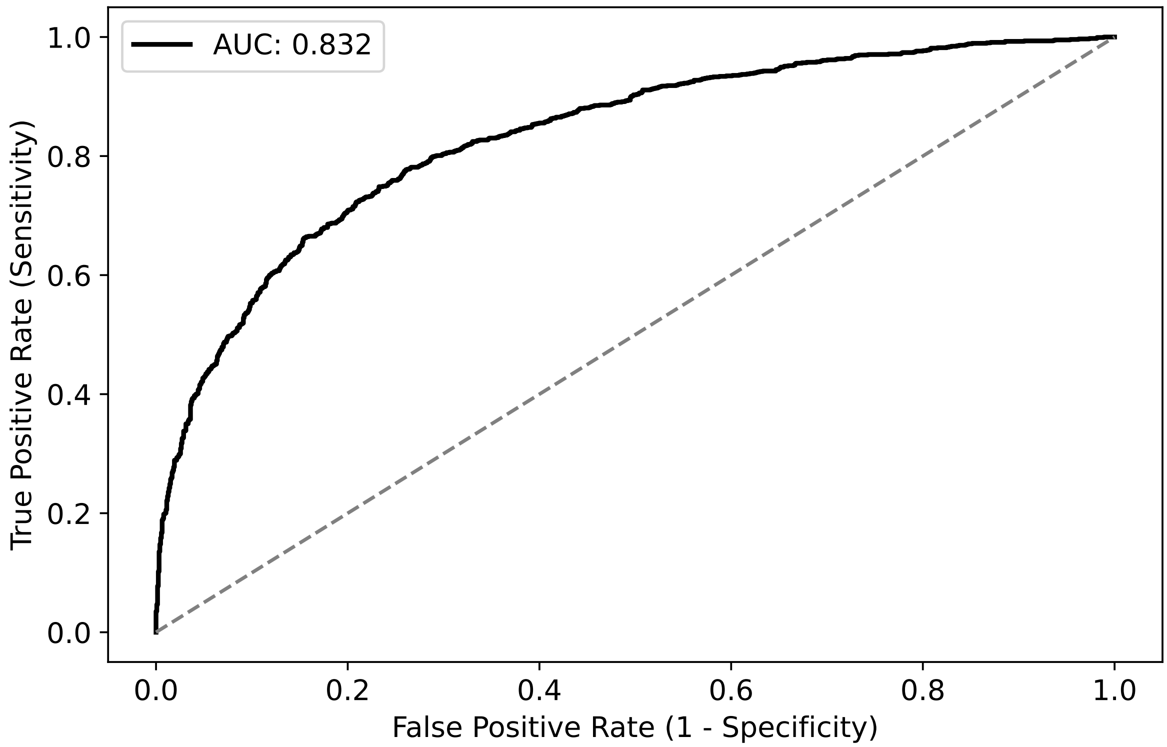

Figure ?: ROC Curve for Logistic Regression Mod

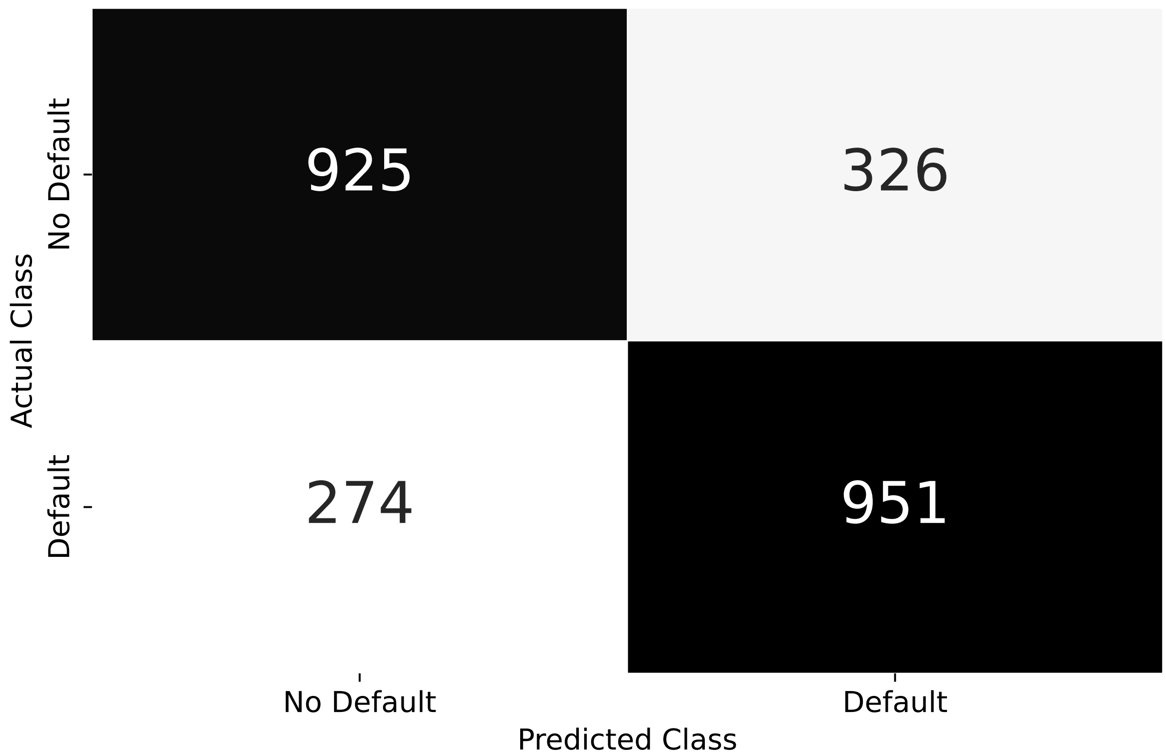

Figure ?: Confusion Matrix for Logistic Regression Model

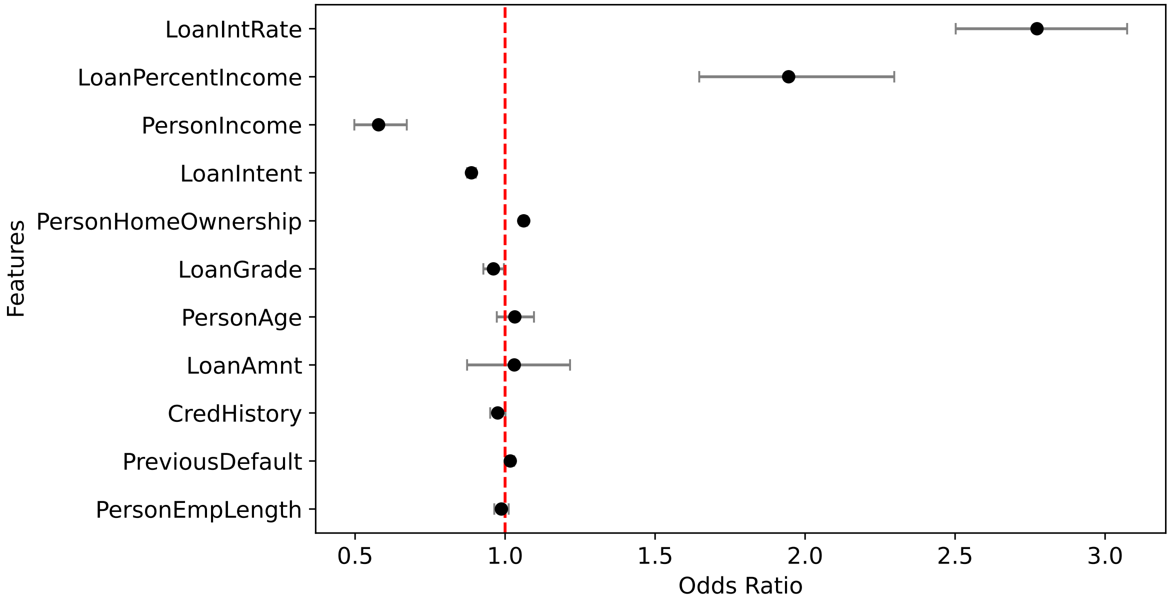

Figure ?: Odds Ratios with 95% Confidence Intervals

3.2 Random Forest

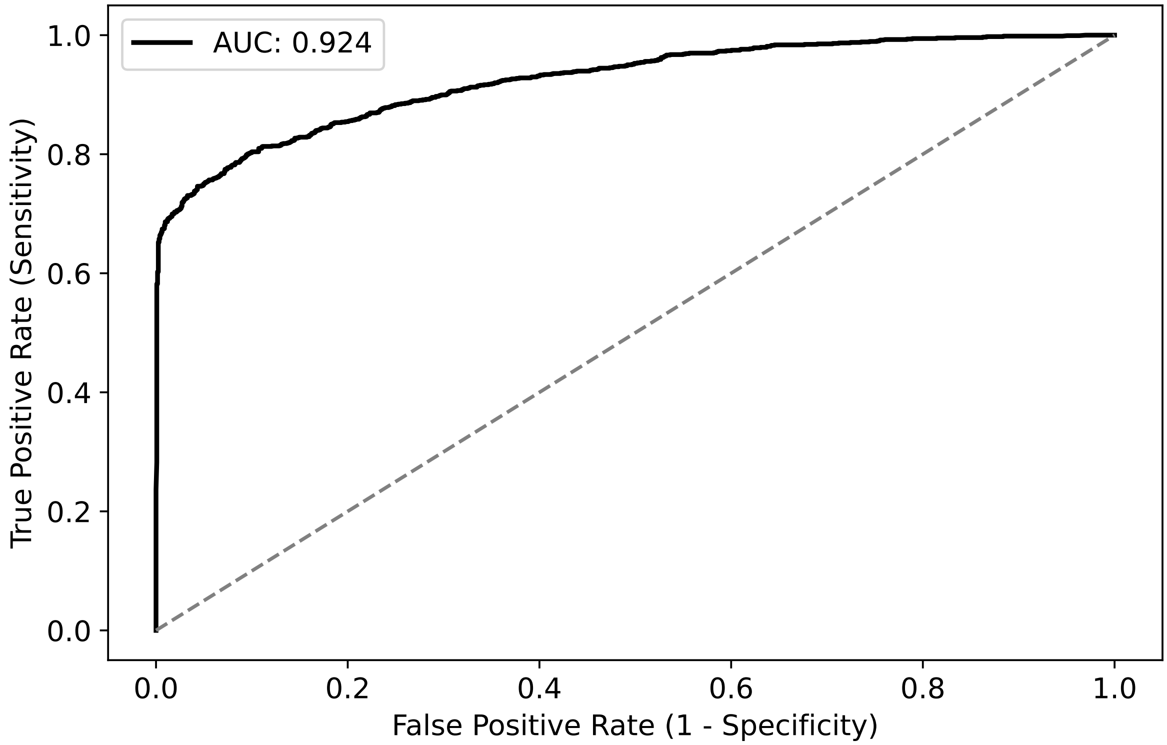

Figure ?: ROC Curve for Random Forest Model

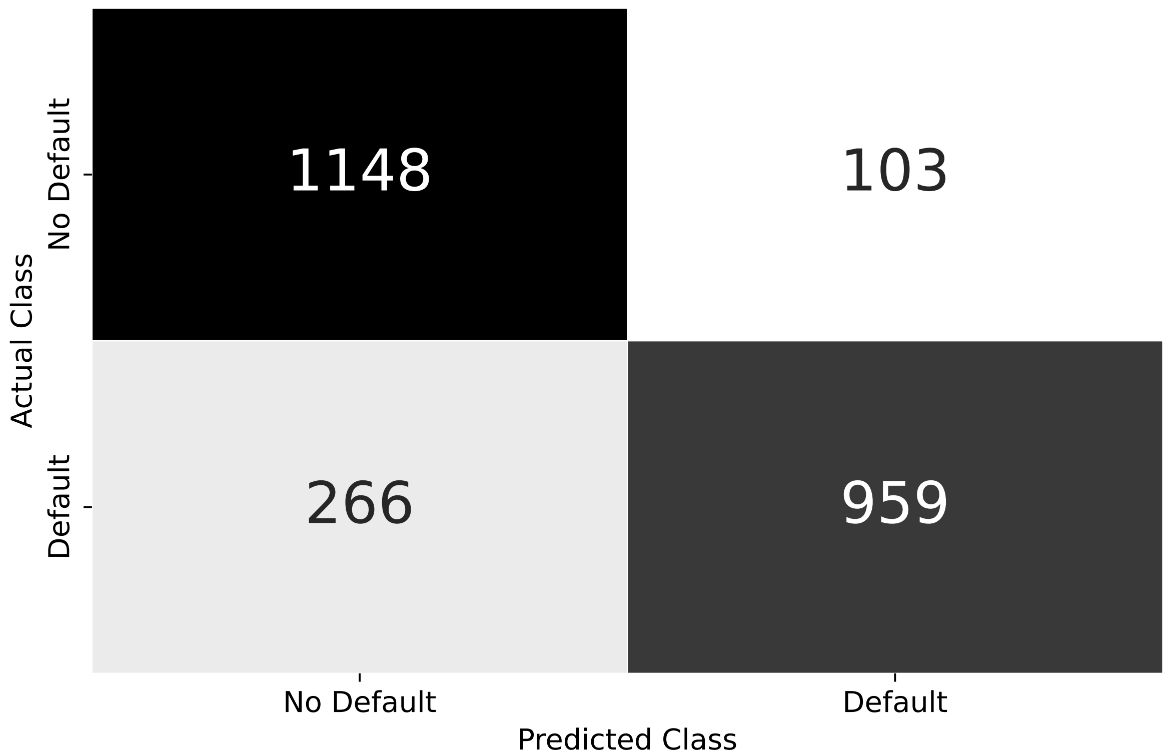

Figure ?: Confusion Matrix for Random Forest Model

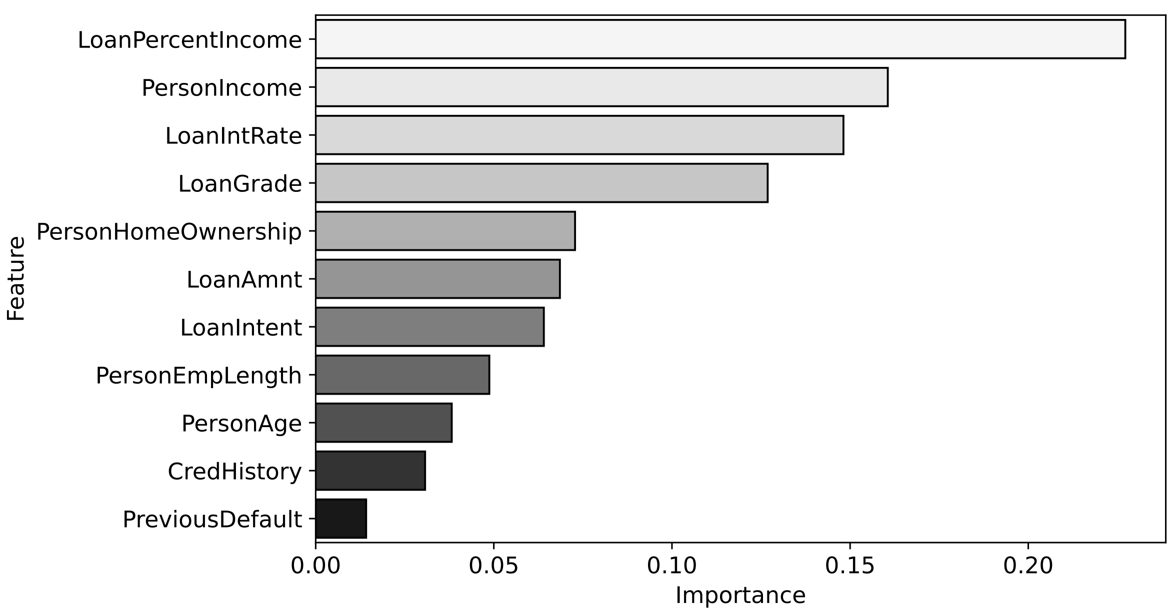

Figure ?: Feature Importances from Random Forest Model

3.3 XGBoost

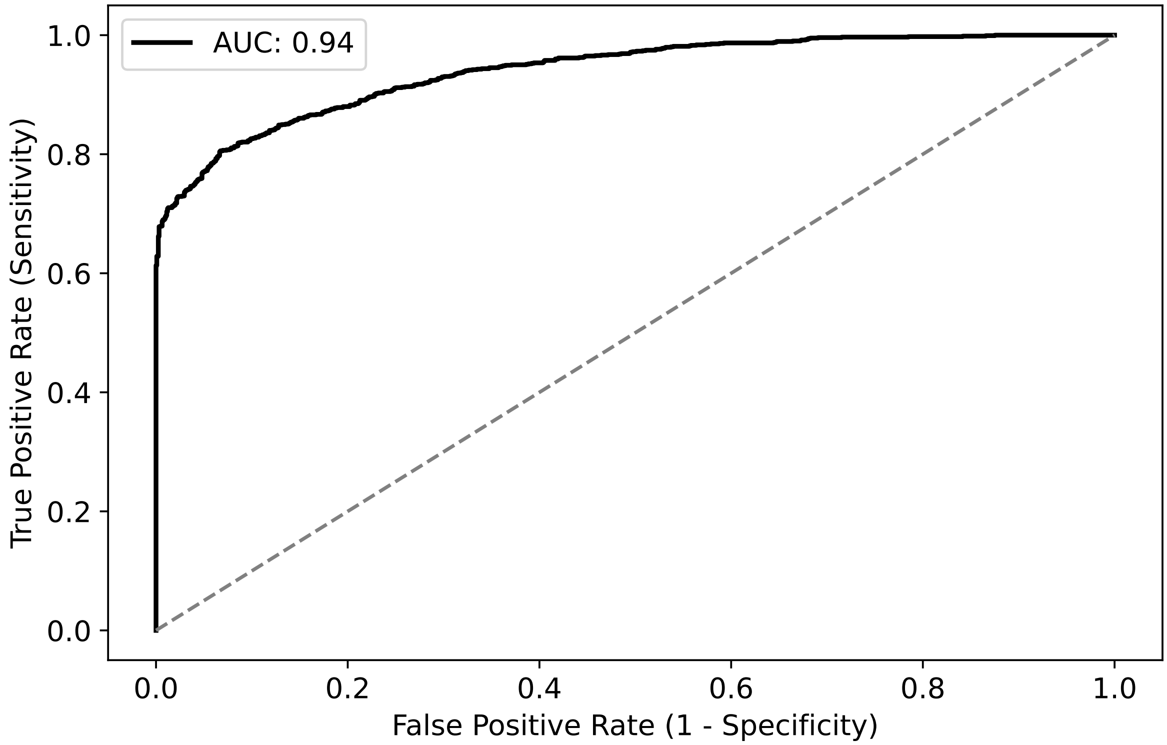

Figure ?: ROC Curve for XGBoost Model

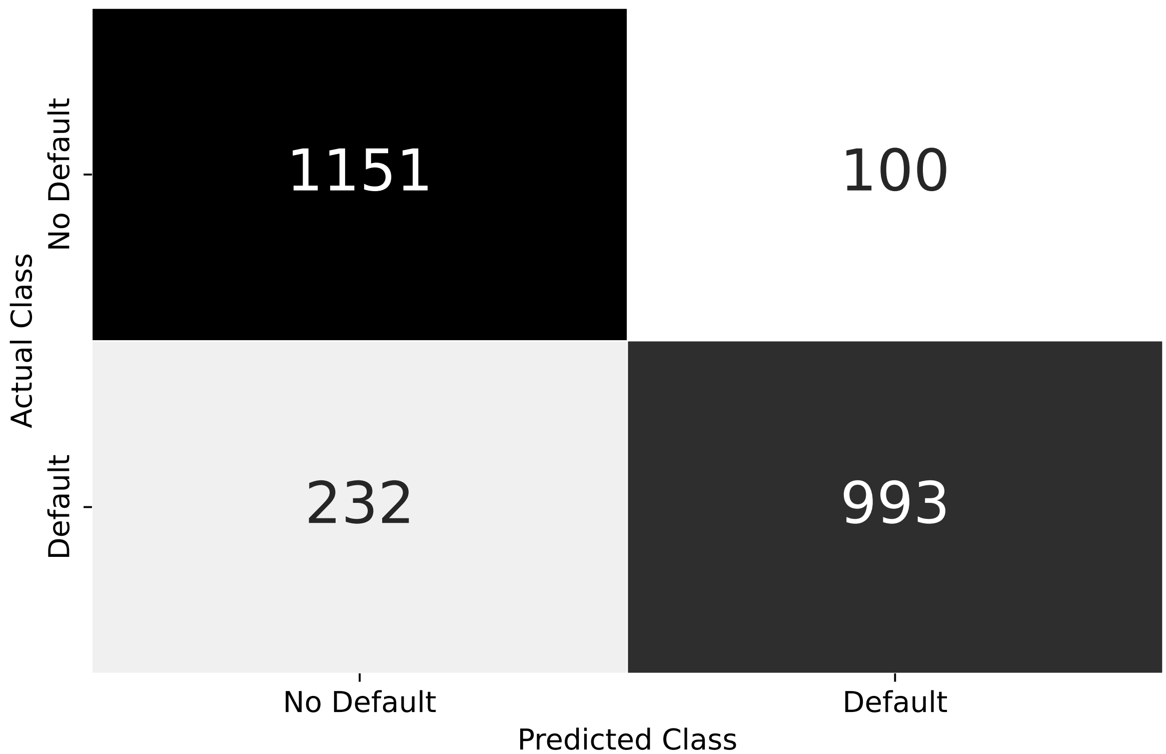

Figure ?: Confusion Matrix for XGBoost Model

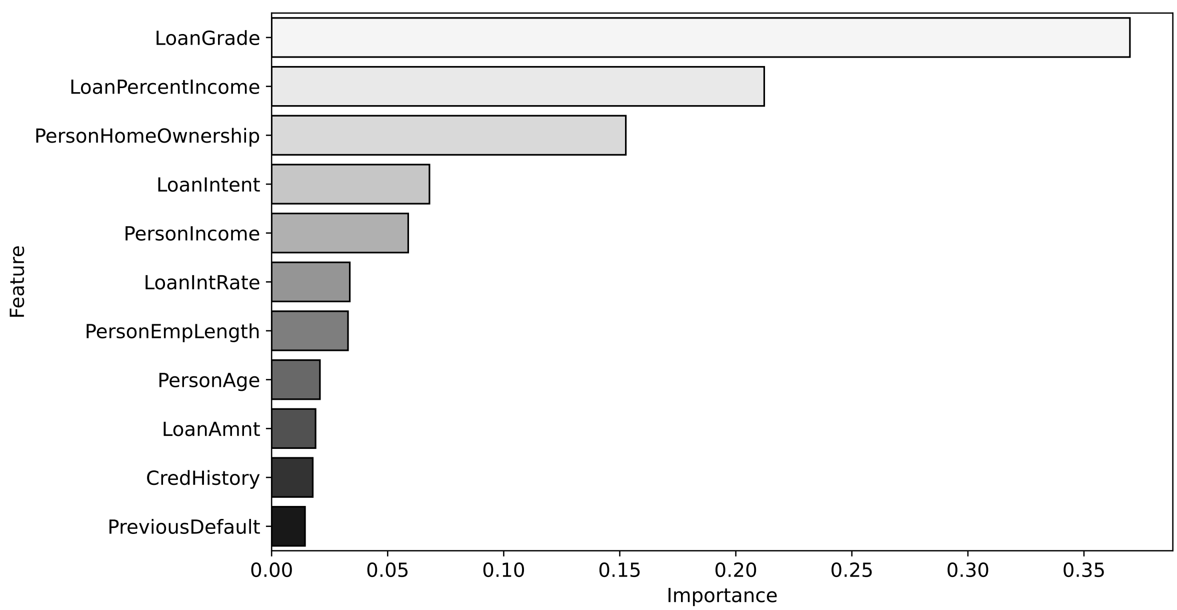

Figure ?: eature Importances from XGBoost Model

3.4 Light Gradient Boosted Machine

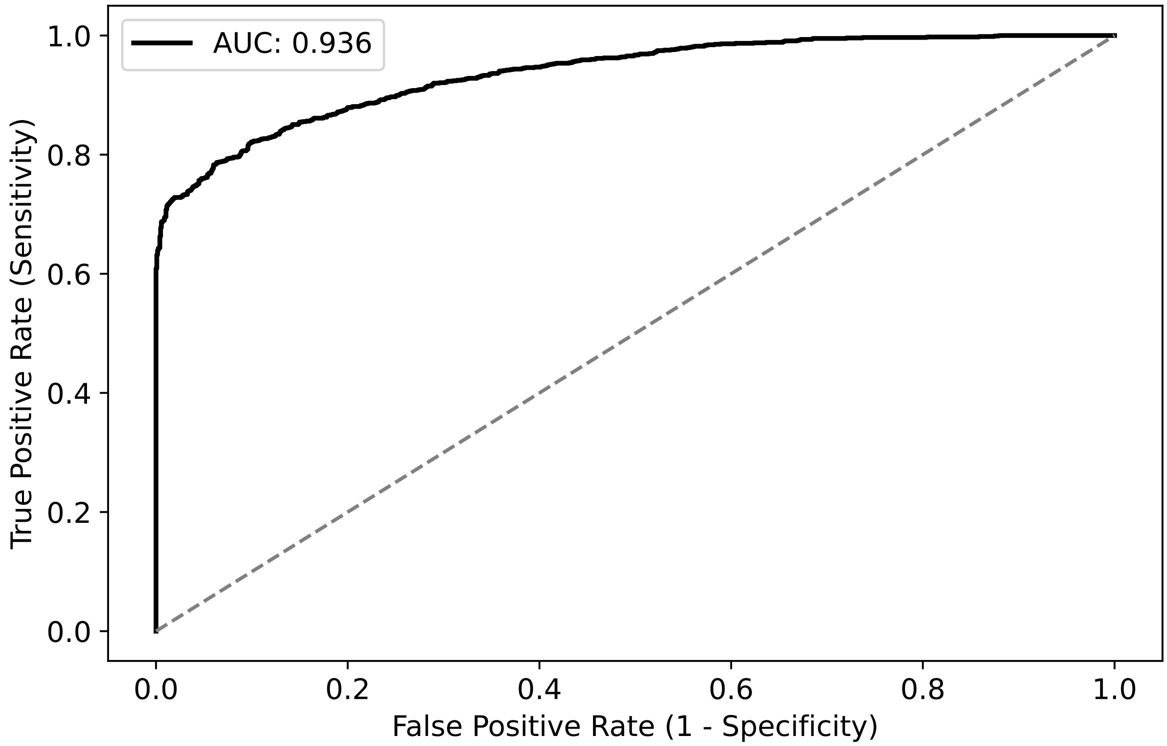

Figure ?: ROC Curve for LightGBM Model

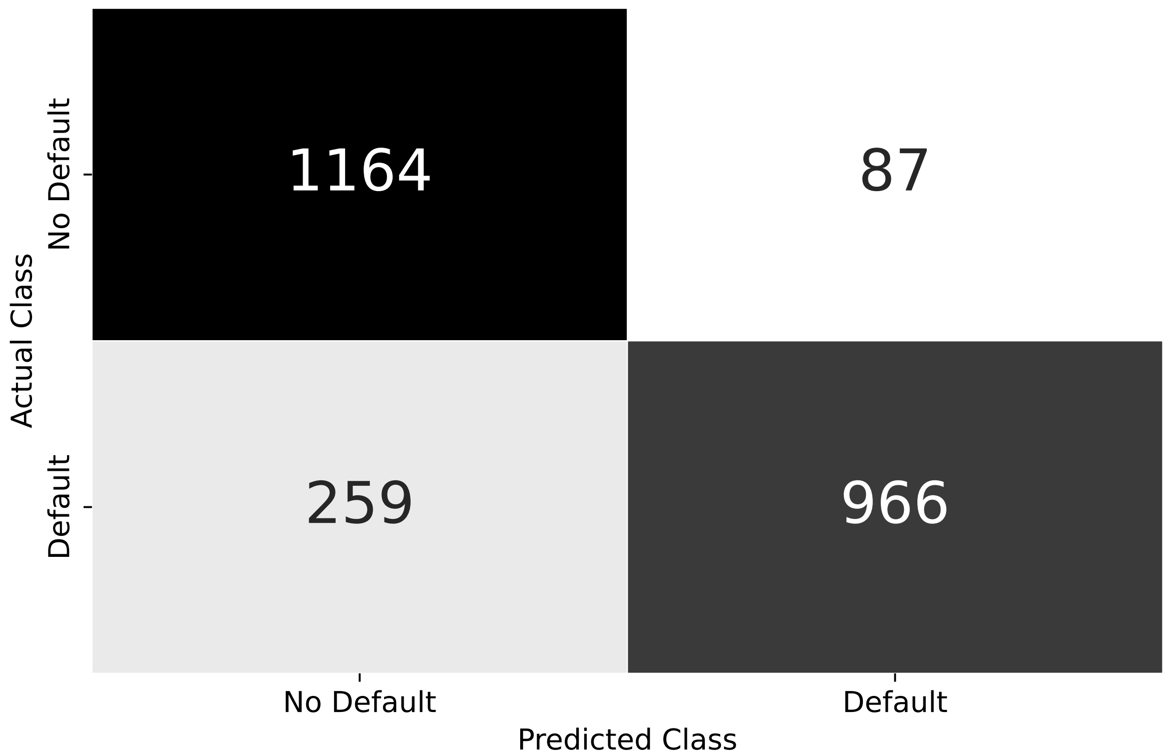

Figure ?: Confusion Matrix for LightGBM Model

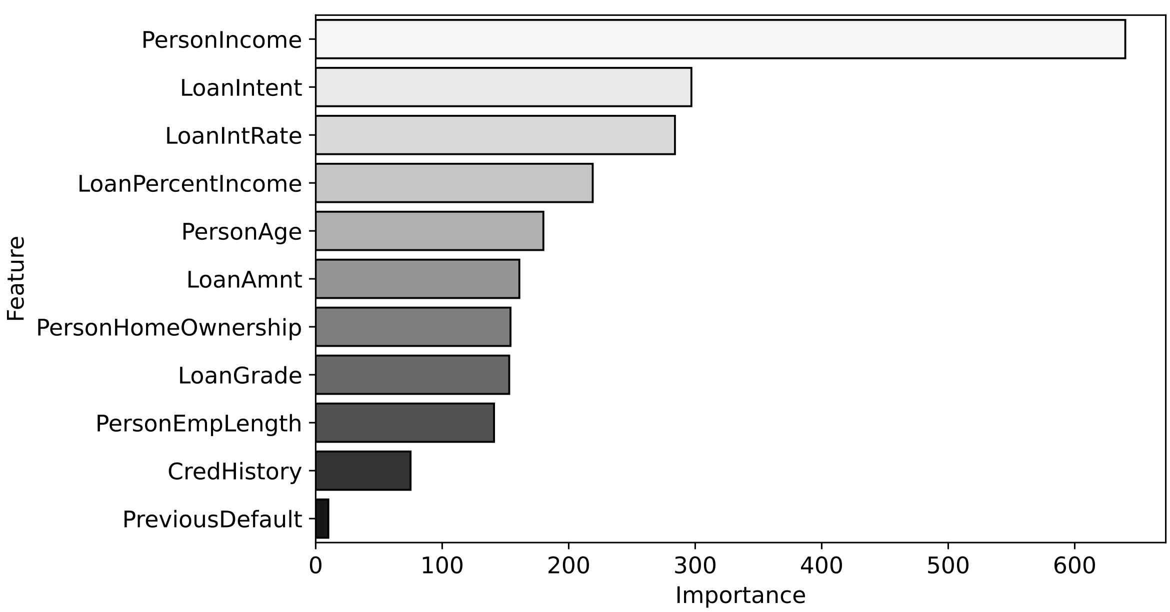

Figure ?: Feature Importances from LightGBM Model

3.5 Model Evaluation and Comparisons

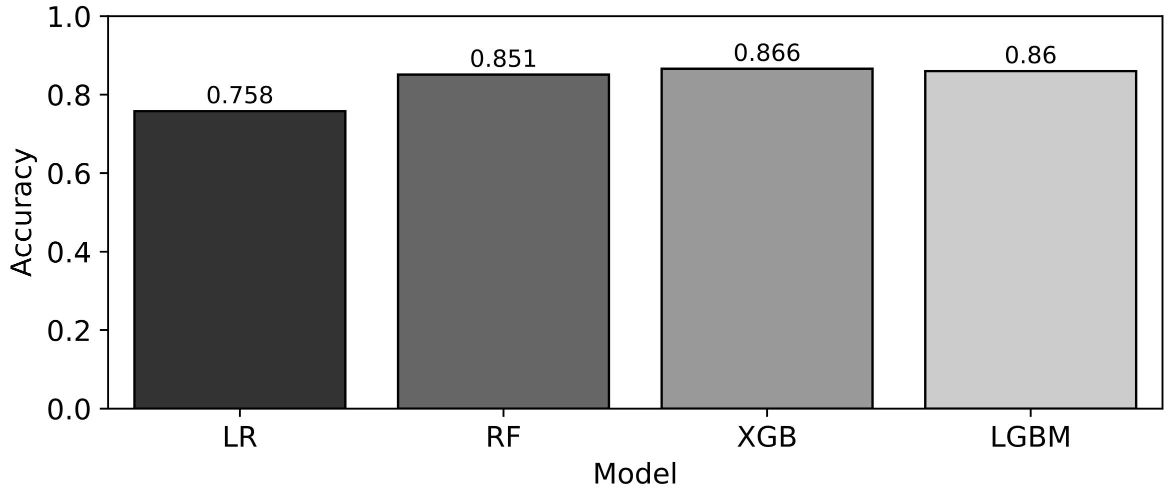

Figure ?: Accuracy for Each Model

Table 4: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss	Brier Score
LR	0.758	0.745	0.776	0.76	0.832	8.734	0.242
RF	0.851	0.903	0.783	0.839	0.924	5.372	0.149
XGB	0.866	0.909	0.811	0.857	0.94	4.833	0.134
LGBM	0.86	0.917	0.789	0.848	0.936	5.037	0.14

4. Conclusion

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