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 businesses to invest, expand and smooth consumption. However, credit risk remains a fundamental challenge for financial institutions, as loan defaulting can lead to substantial financial losses for both the company and 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, illustrated in Table 1.

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	int64	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

2.1 Preparing the Data

Table 2: Missing Values in Each Variable

Variable	Missing Values
PersonAge	0
PersonIncome	0
PersonHomeOwnership	0
PersonEmpLength	887
LoanIntent	0
LoanGrade	0
LoanAmnt	0
LoanIntRate	3095
LoanStatus	0
LoanPercentIncome	0
PreviousDefault	0
CredHistory	0

Table 2 displays the missing values within the dataset for each variable. The only variables with missing data are *PersonEmpLength* and *LoaanIntRate*, containing 887 and 3095 observations with no values, respectively. Missing data can have a large impact on data analysis if not handled properly and can lead to skewed or incorrect conclusions, making handling this data in the correct way crucial. Due to the negatively skewed nature of *PersonEmpLength*, illustrated in Figure 1, median imputation was deployed in order to maintain the observations and not

impact sample size. *LoanIntRate* saw a high correlation with *LoanGrade*, shown by Figure (? Corr Matrix), therefore regression imputation was used to fill these missing variables and not lose sample size.

2.2 Descriptive Statistics

Table 3: Summary Statistics of Numeric Variables

Variable	N	Mean	Median	SD	Min	Max
PersonAge	32415.0	27.7	26.0	6.3	20.0	144.0
PersonIncome	32415.0	65908.6	55000.0	52533.0	4000.0	2039784.0
PersonEmpLength	32415.0	4.8	4.0	4.1	0.0	123.0
LoanGrade	32415.0	1.2	1.0	1.2	0.0	6.0
LoanAmnt	32415.0	9594.0	8000.0	6322.8	500.0	35000.0
LoanIntRate	32415.0	11.0	11.0	3.2	5.4	23.4
LoanStatus	32415.0	0.2	0.0	0.4	0.0	1.0
LoanPercentIncome	32415.0	0.2	0.2	0.1	0.0	0.8
CredHistory	32415.0	5.8	4.0	4.1	2.0	30.0

Table 3 contains all the summary statistics for all variables within the dataset. *PersonAge* and *PersonEmpLength* show maximum values of 144 and 123 years respectively, which are both above the oldest age a person has lived (122 years), meaning that they are potential errors. To remove these errors from the dataset, both observations for *PersonEmpLength* were removed as to not impact the models. For *PersonAge*, all observations with ages above 122 years were removed. This left *PersonAge* with a maximum value of 94 and *PersonEmpLength* with a maximum value of 41, both which are reasonable.

2.3 Distribution Analysis

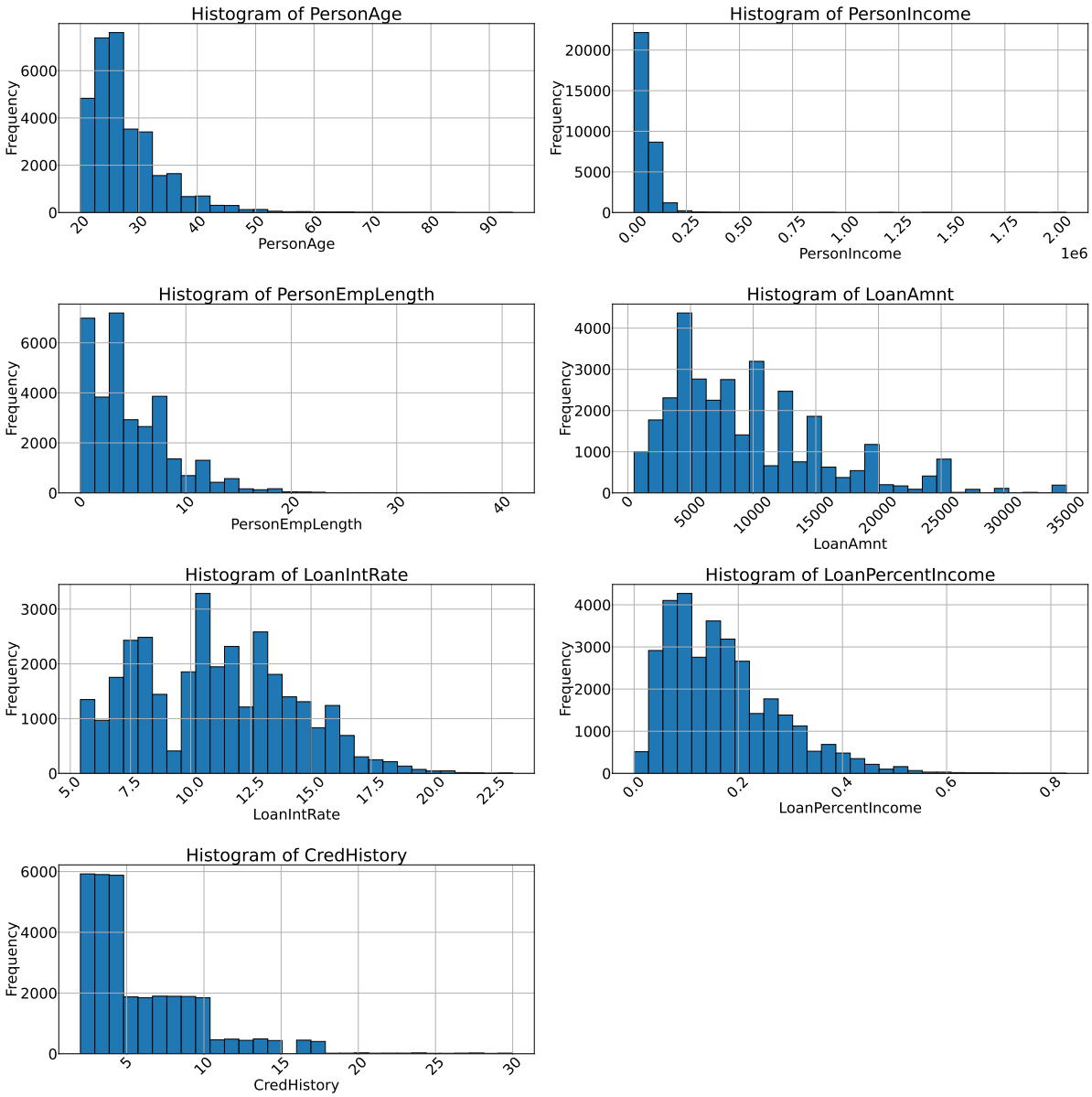

Figure ?: Histograms of all Numeric Variables

The histograms shown in Figure 1 illustrate the distributions for each numeric variable. All of the variables shown have negatively skewed distributions. This is due to individuals with low age likely to have low values in each of these variables. *PersonAge*, *PersonEmpLength* and *CredLength* have very similar distributions, indicating potential correlation between these variables.

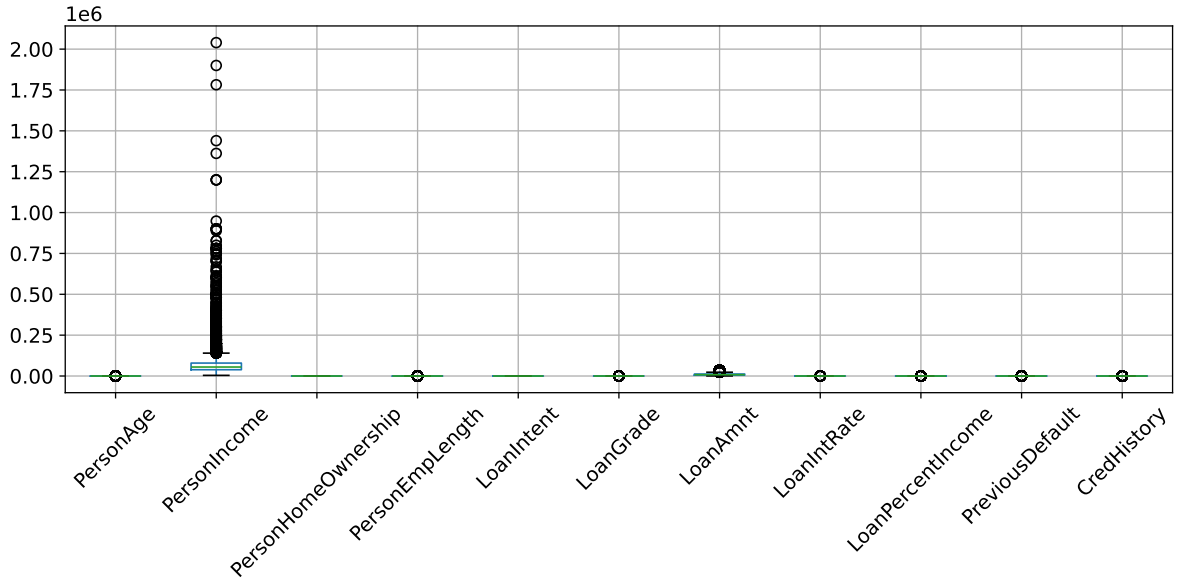

Figure ?: Box Plots of All Variables Before Normalisation

Figure 3 shows that the data isn't scaled proportionally, therefore we need to apply a scaling technique. Due to the skewness of all the variables quantile transformation was deployed, normalised data is shown in Figure 4. The plot shows outliers, however there is no reason for these to be errors meaning they will not be removed. For example, the reason for outliers in *PersonIncome* is due to people earning considerably more than average.

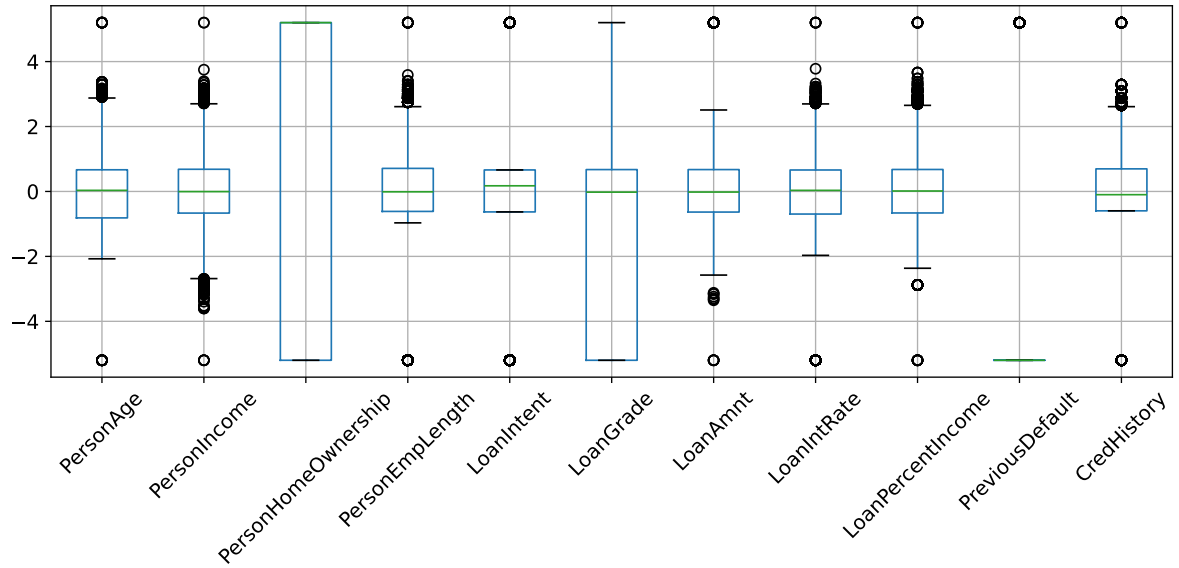

Figure ?: Box Plots of All Variables After Normalisation

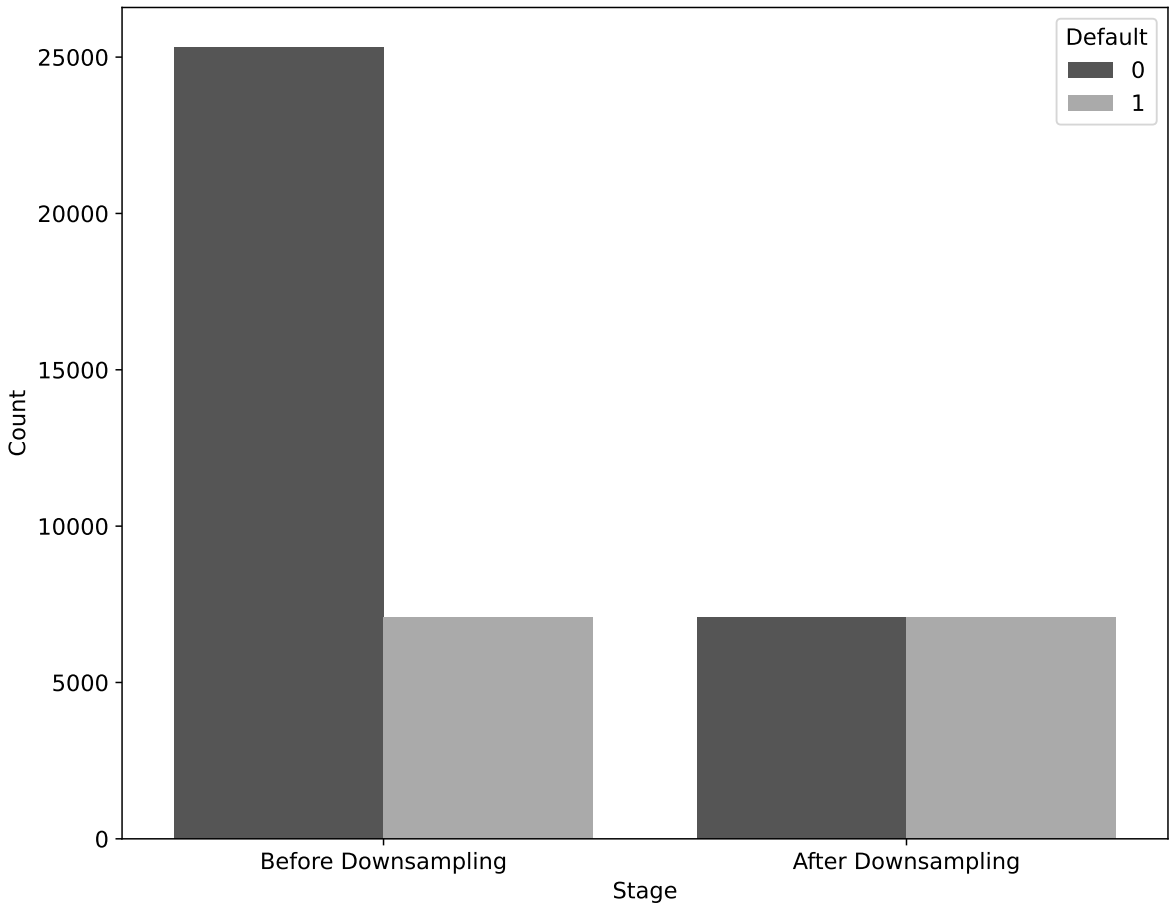

Figure ?: Distribution of LoanStatus Before and After Downsampling

Figure 5 demonstrates the the distribution of *LoanStatus*. Before downsampling there was a large discrepancy between the number of people who defaulted and who didn't. This can cause large impacts on the ML models deployed in the analysis, leading to skewed perforamnce metrics as the models will predict the majority class with high accuracy but the minority class with lower accuracy. To circumvent this issue, downsampling was performed to ensure both outcomes had the same number of observations, shown in Figure 6

2.4 Correlation Analysis

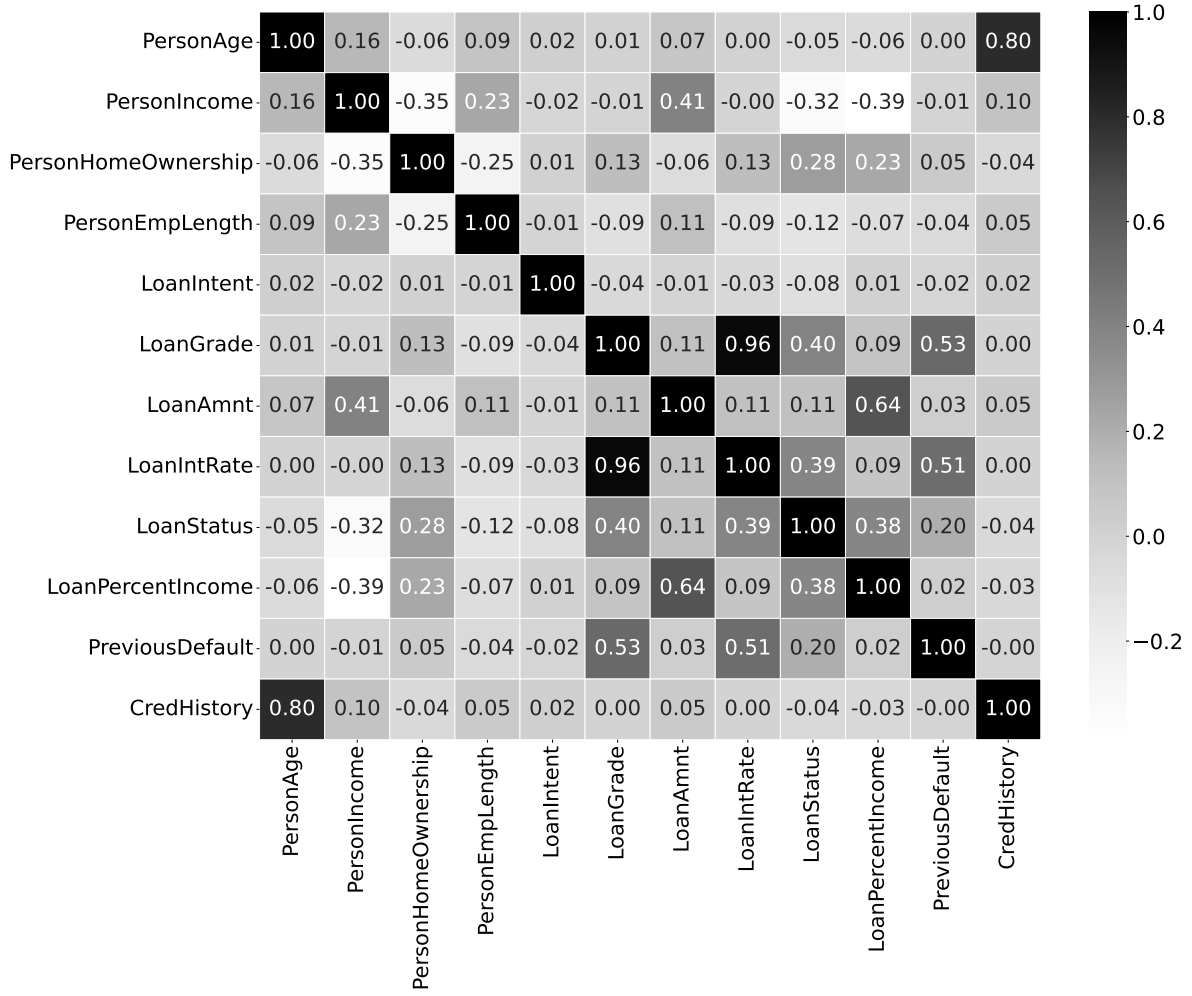

Figure ?: Correlation Plot of All Variables

Figure 7 shows a correlation plot quantifying the relationships between the variables and to the target *LoanStatus*. *LoanGrade* and *LoanIntRate* have a high correlation coefficient (0.96), indicating that they are highly correlated. Also, a similar relationship is shown between *PersonAge* and *CredHistory* (r = 0.80). Both these relationships make logical sense as someone who is older who have a longer credit history and as loan grade increases it is likley that the interest rate does as well. Due to the mullticolliearity in the data, these variables may have to be removed however, futher analysis with variance inflation factor (VIF) is required.

Table 4: Variance Inflation Factor (VIF) Values

Feature	VIF
PersonAge	1.505000
PersonIncome	9.344000
PersonHomeOwnership	1.198000
PersonEmpLength	1.065000
LoanIntent	1.002000
LoanGrade	2.998000
LoanAmnt	12.480000
LoanIntRate	3.118000
LoanPercentIncome	11.993000
PreviousDefault	1.250000
CredHistory	1.472000

VIF values for all the variables are shown within Table 4. In contrast to Figure 7, *LoanGrade*, *LoanIntRate*, *PersonAge*, *CredHistory* have low VIF values, indicating low levels of multicollinearity. However, *LoanAmnt* and *LoanPercentIncome* have VIF values greater than 10 which shows multicollinearity and actions need to be taken to ensure they don't affect the models. For the logistic regression, L1 and L2 regularisation was deployed to reduce the affects of multicollinearity. Due to the other models being tree based why handle multicollinearity well, therefore no futher processing is needed.

Within this analysis, LR, RF, XGboost and LGBM models will be trained to predict *LoanStatus* using *PersonAge*, *PersonIncome*, *PersonHomeOwnership*, *PersonEmpLength*, *LoanIntent*, *LoanGrade*, *LoanAmnt*, *LoanIntRate*, *LoanPercentIncome*, *PreviousDefault* and *CredHistory*.

3. Results and Discussion

3.1 Logistic Regression

The first model deployed was an LR trained on all the standard variables, this model acts as a baseline to compare all more complex models with.

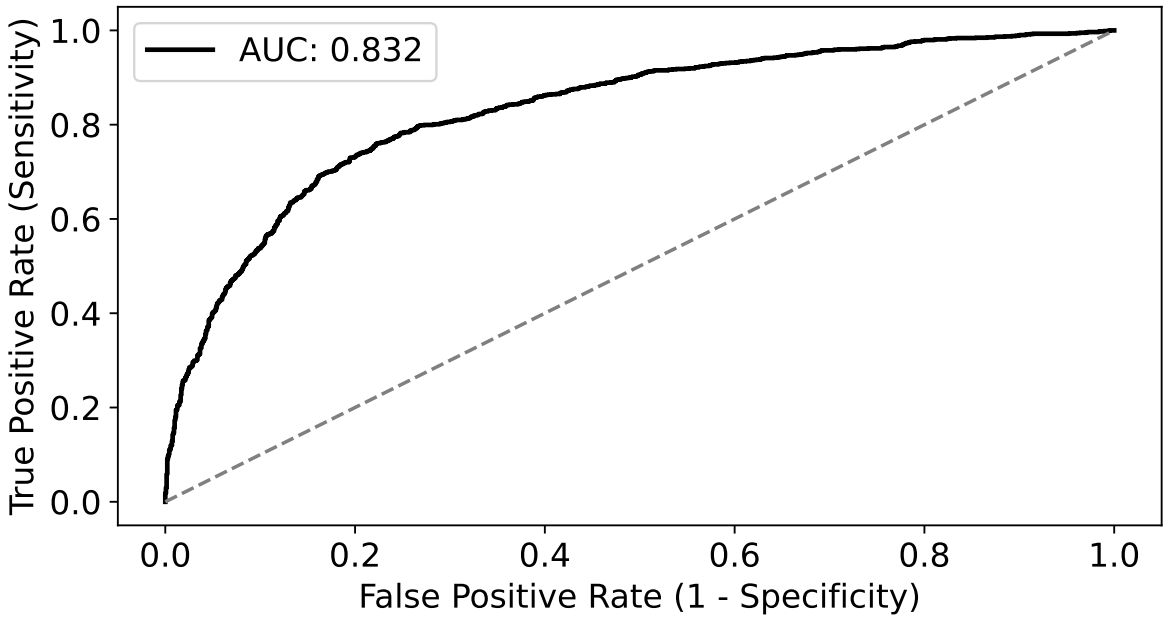

Figure ?: ROC Curve for Logistic Regression Model

Figure ? shows the ROC graph for the LR model

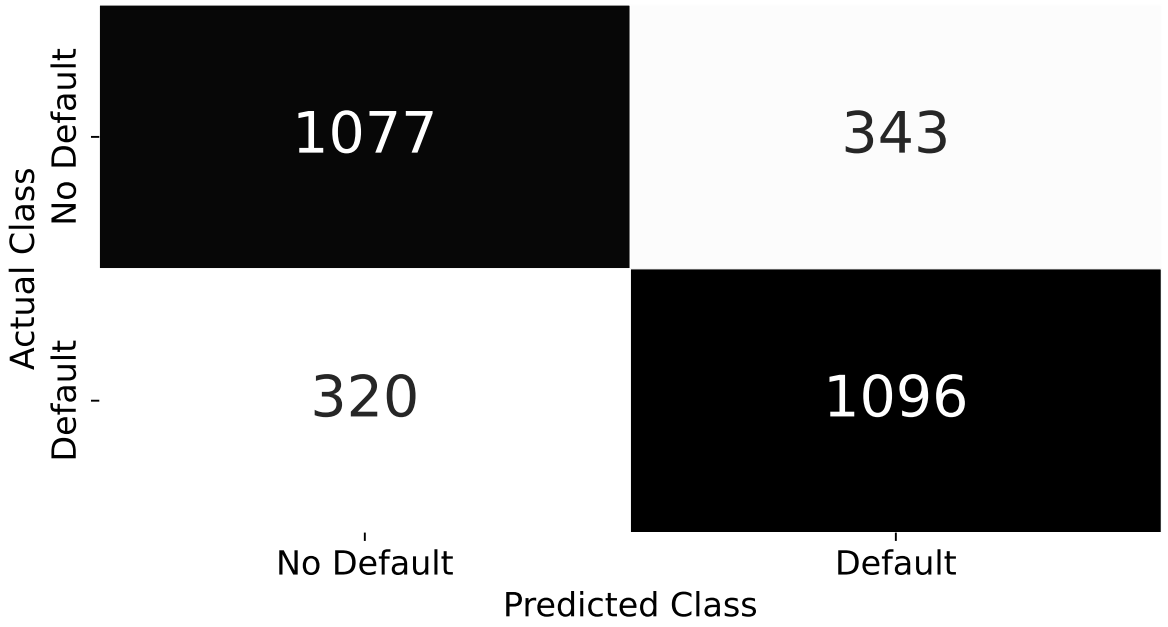

Figure ?: Confusion Matrix for Logistic Regression Model

Conf Matrix...

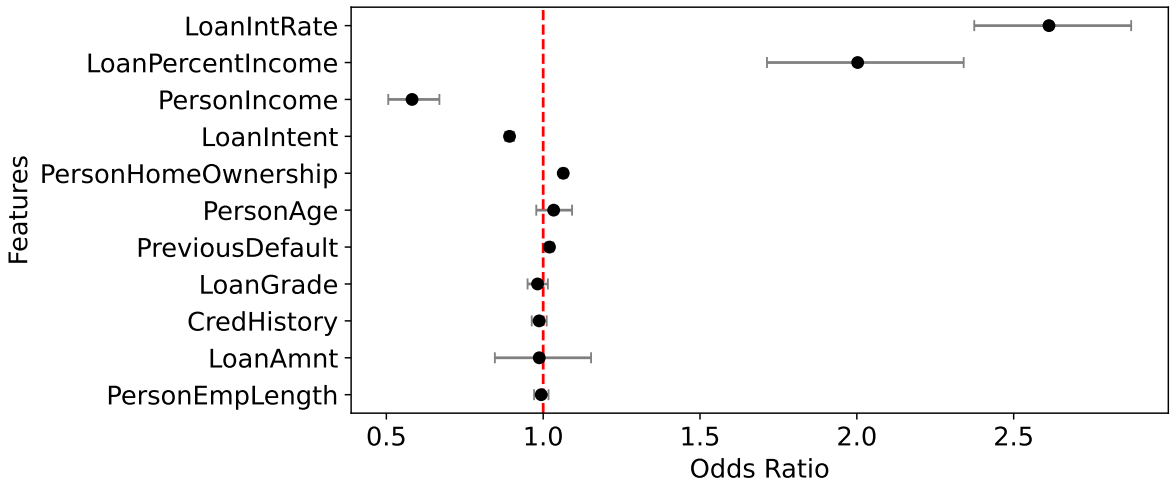

Figure ?: Odds Ratios for Logistic Regression Model

Odds ratios were calculated allowing an easy interpretation of the relationships between the factors and credit risk. The odds ratio indicates the increase in the risk of defaulting for a one-unit increase in that variable.

3.2 Random Forest

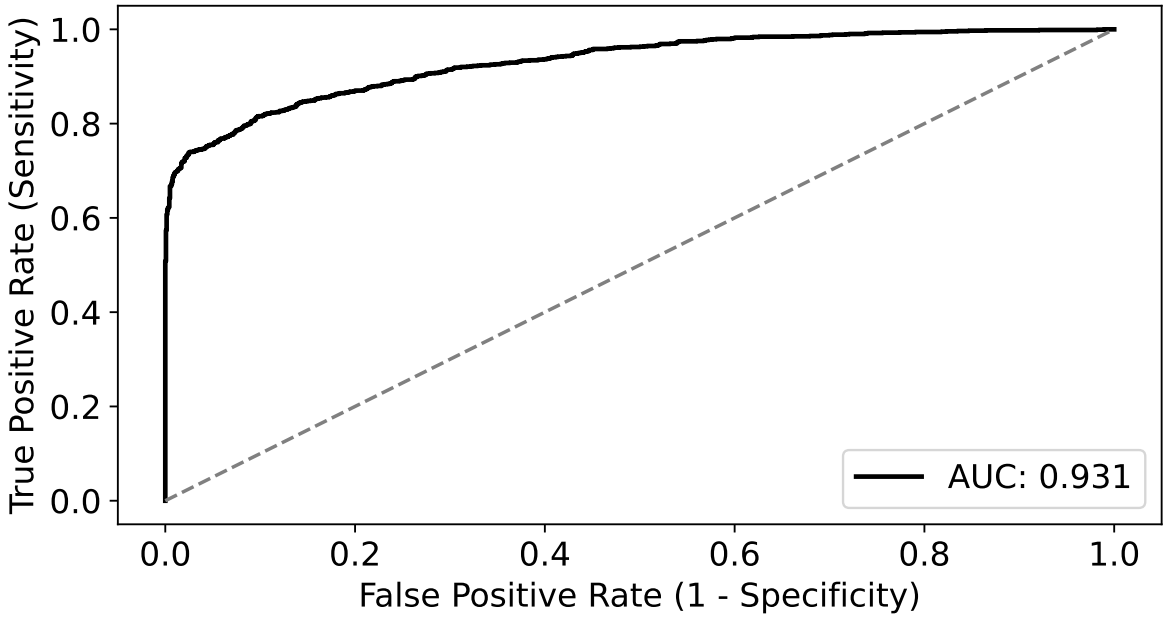

Figure ?: ROC Curve for Random Forest Model

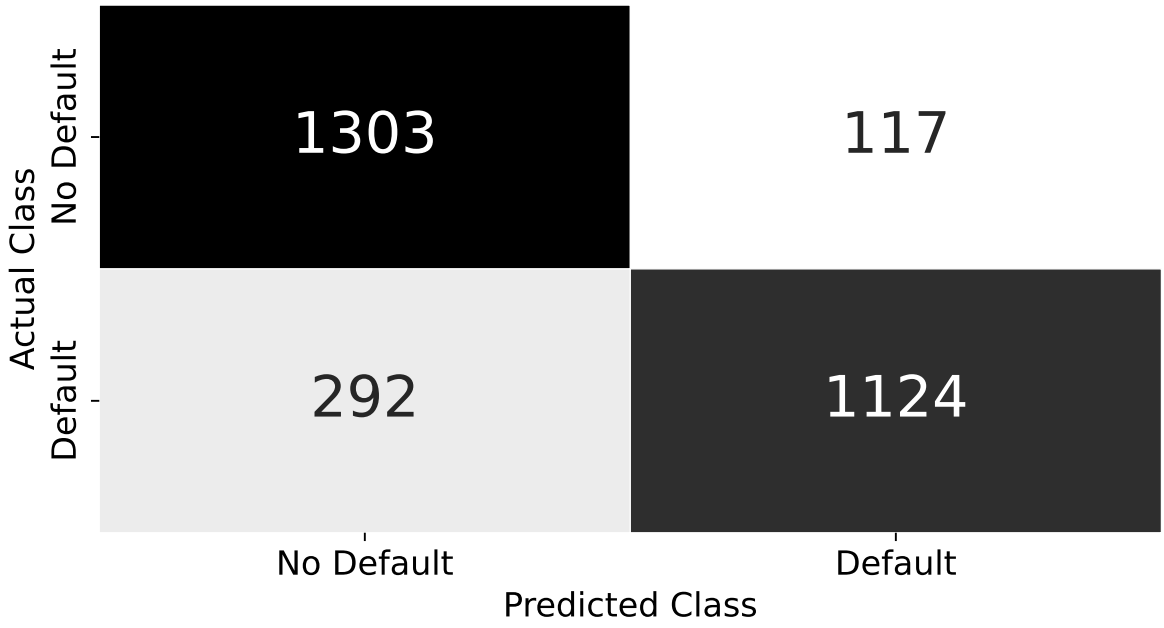

Figure ?: Confusion Matrix for Random Forest Model

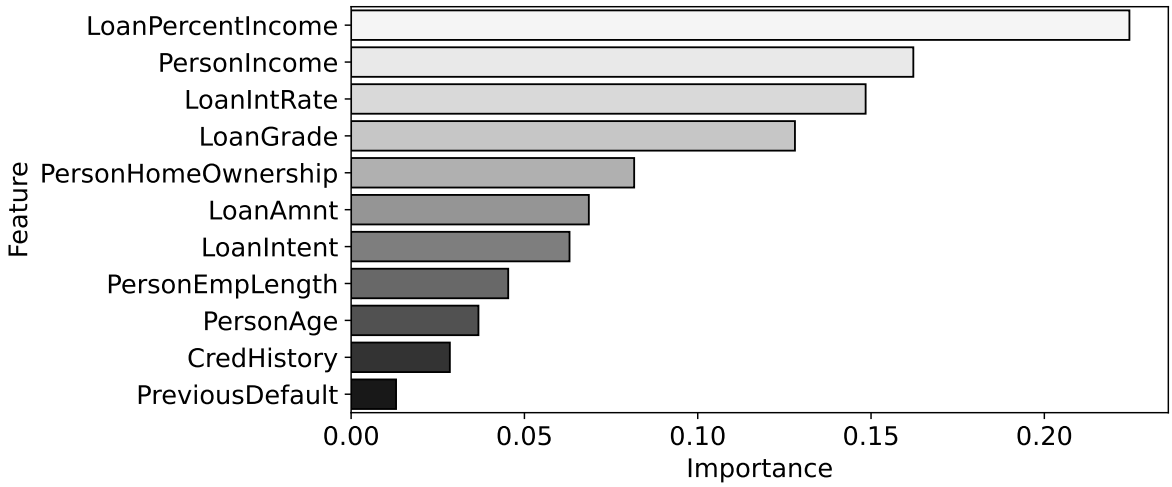

Figure ?: Feature Importances from Random Forest Model

3.3 XGBoost

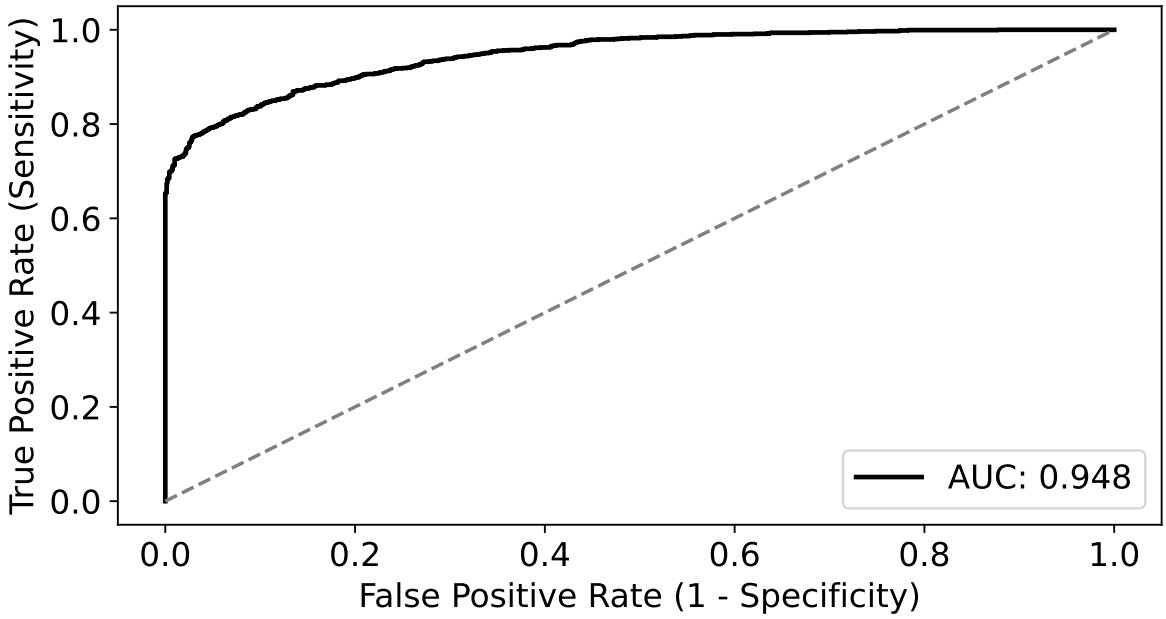

Figure ?: ROC Curve for XGBoost Model

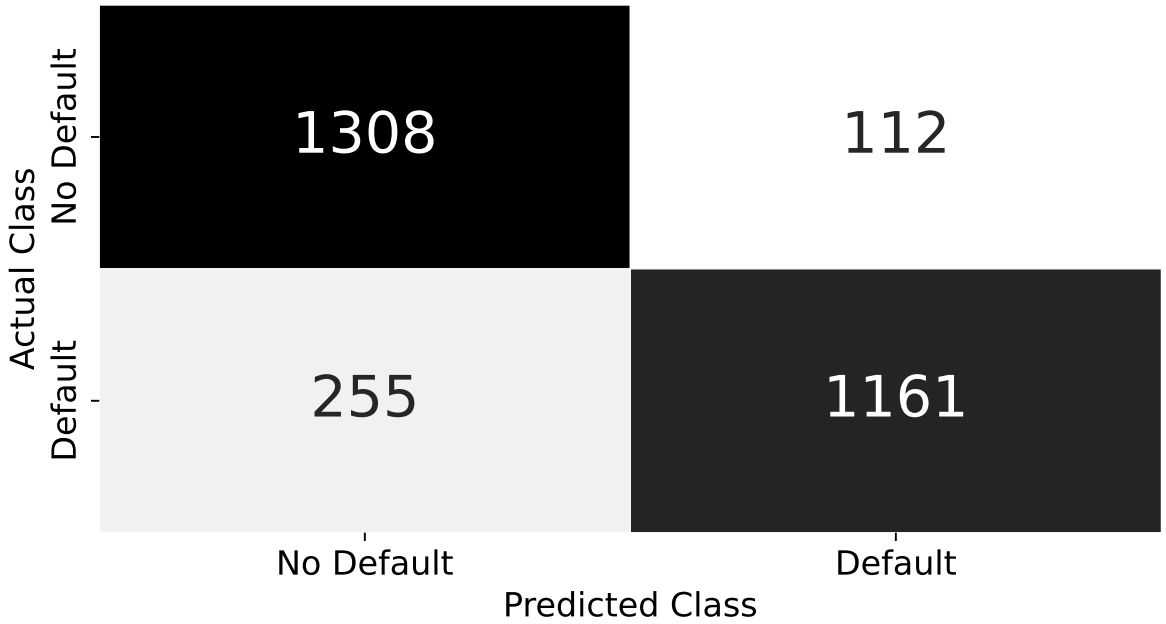

Figure ?: Confusion Matrix for XGBoost Model

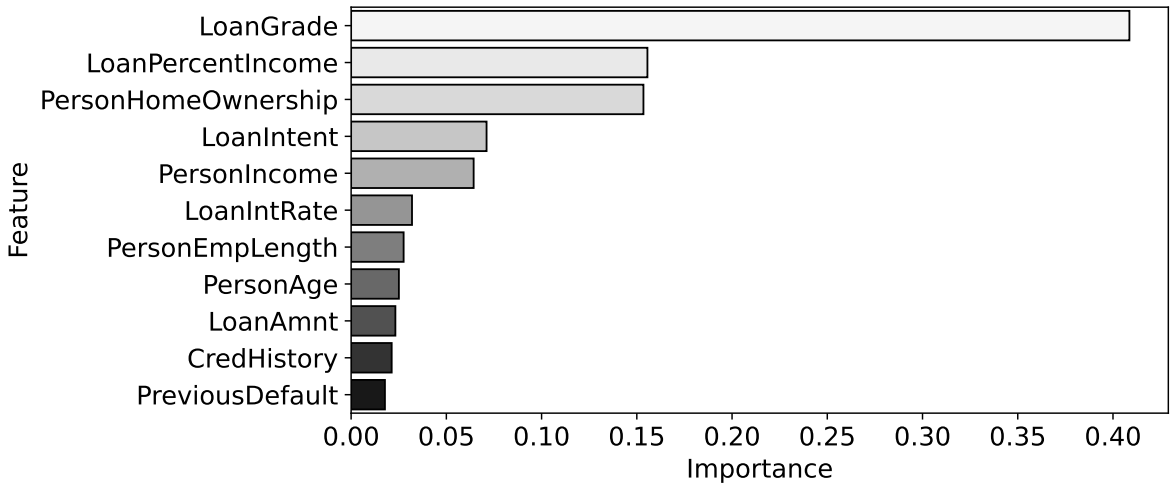

Figure ?: Feature Importances from XGBoost Model

3.4 Light Gradient Boosted Machine

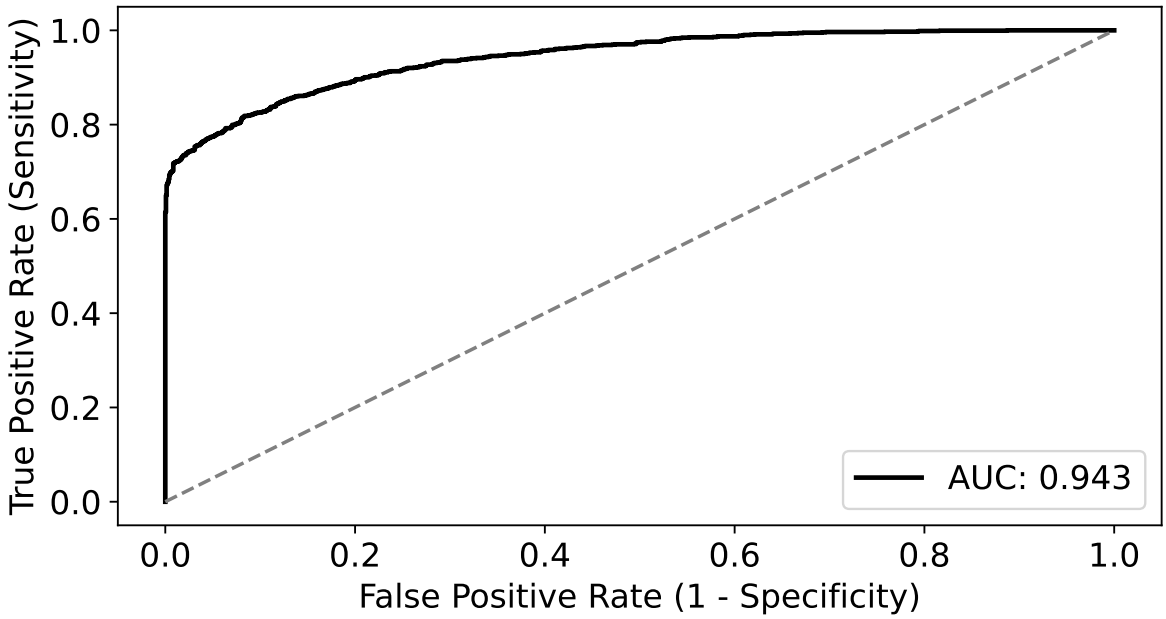

Figure ?: ROC Curve for LightGBM Model

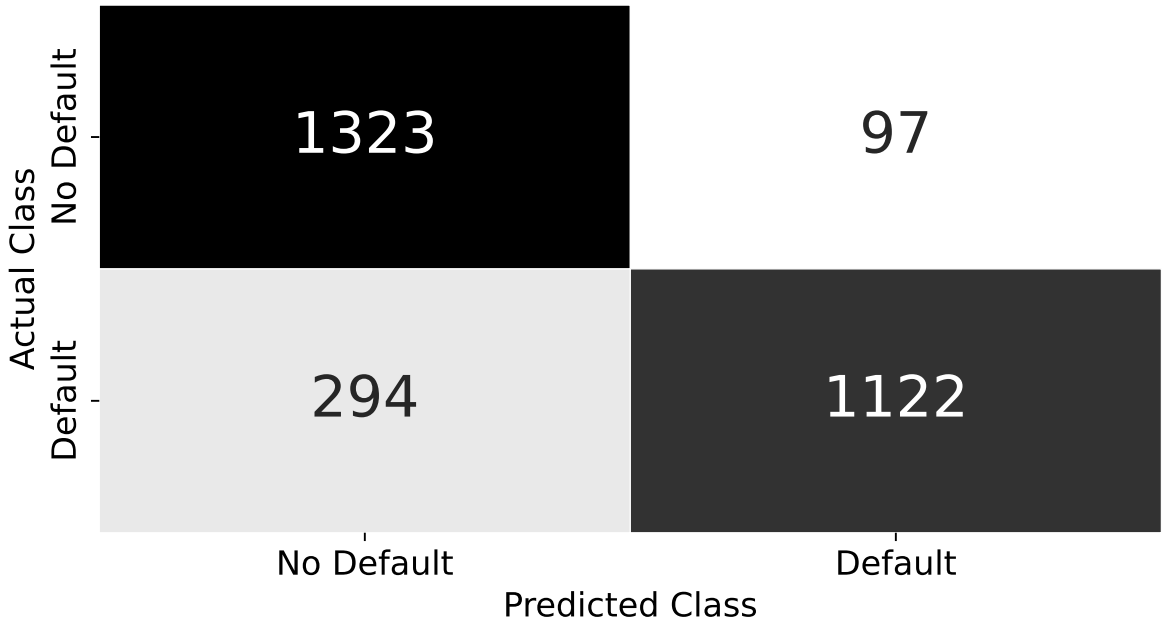

Figure ?: Confusion Matrix for LightGBM Model

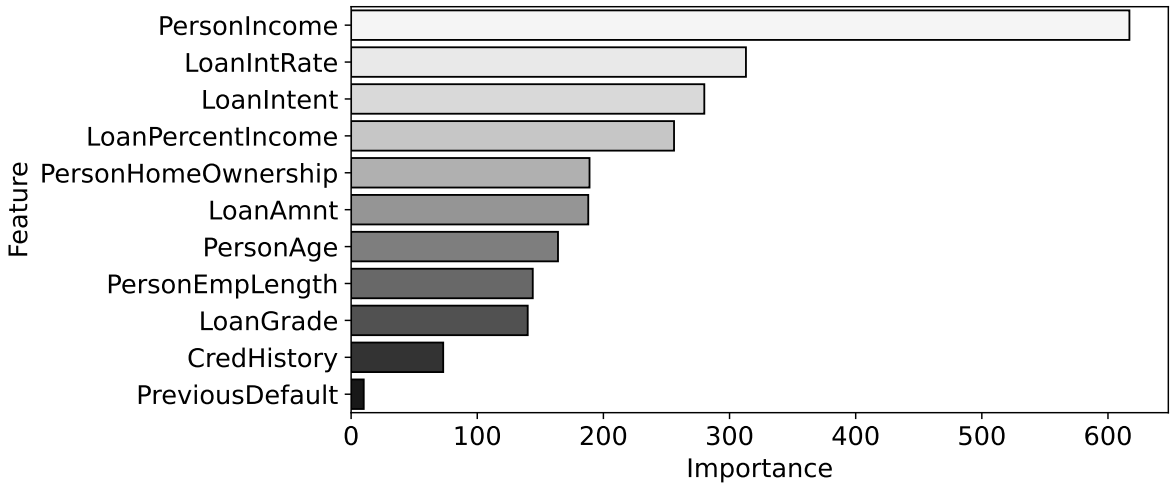

Figure ?: Feature Importances from LightGBM Model

3.5 Model Evaluation and Comparisons

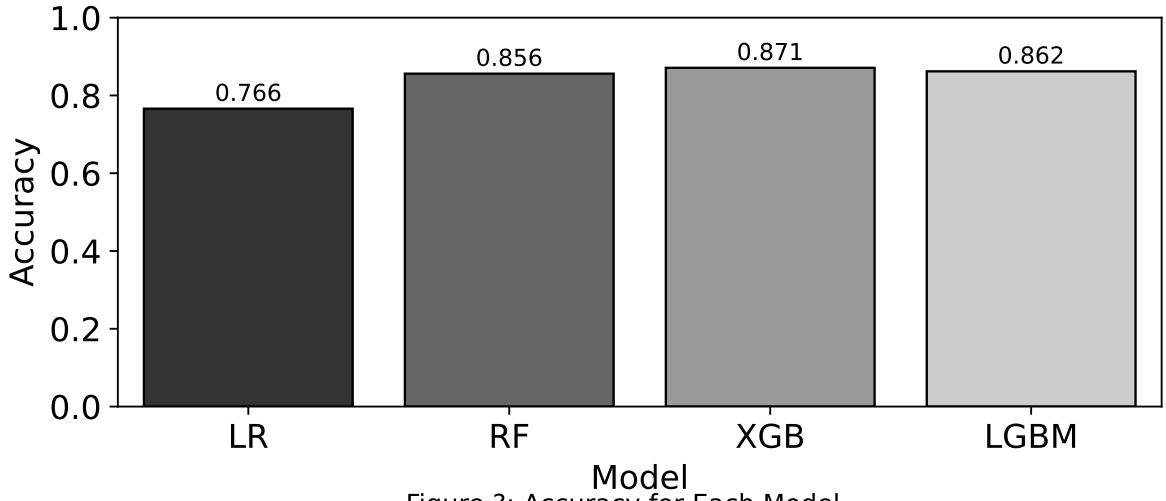

Figure ?: Accuracy for Each Model

Table 5: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss	Brier Score
LR	0.766	0.762	0.774	0.768	0.832	8.426	0.234
RF	0.856	0.906	0.794	0.846	0.931	5.198	0.144
XGB	0.871	0.912	0.82	0.864	0.948	4.664	0.129
LGBM	0.862	0.92	0.792	0.852	0.943	4.969	0.138

Table 6: Top 3 Most Important Variables for Each Model

	Logistic Regression	Random Forest	XGBoost	LightGBM
Feature 1	PersonIncome	LoanPercentIncome	LoanGrade	PersonIncome
Feature 2	LoanPercentIncome	PersonIncome	LoanPercentIncome	LoanIntRate
Feature 3	LoanIntRate	LoanIntRate	PersonHomeOwnership	LoanIntent

4. Conclusion

Link to Github Repository = BEE2041 Data Science In Economics Empirical Project