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

BEE2041 Data Science in Economics - Empirical Project

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.

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 32580 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 5, therefore regression imputation was used to fill these missing variables and not lose sample size. Also, any duplicate observations were removed to remove their impact on the models, this reduced the sample size to 32415 observations.

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 where *PersonEmpLength* was 123 were removed as to not impact the models. For *PersonAge*, all observations with ages above 144 years were removed. This left *PersonAge* with a maximum value of 94 and *PersonEmpLength* with a maximum value of 41, which both are reasonable.

2.3 Distribution Analysis

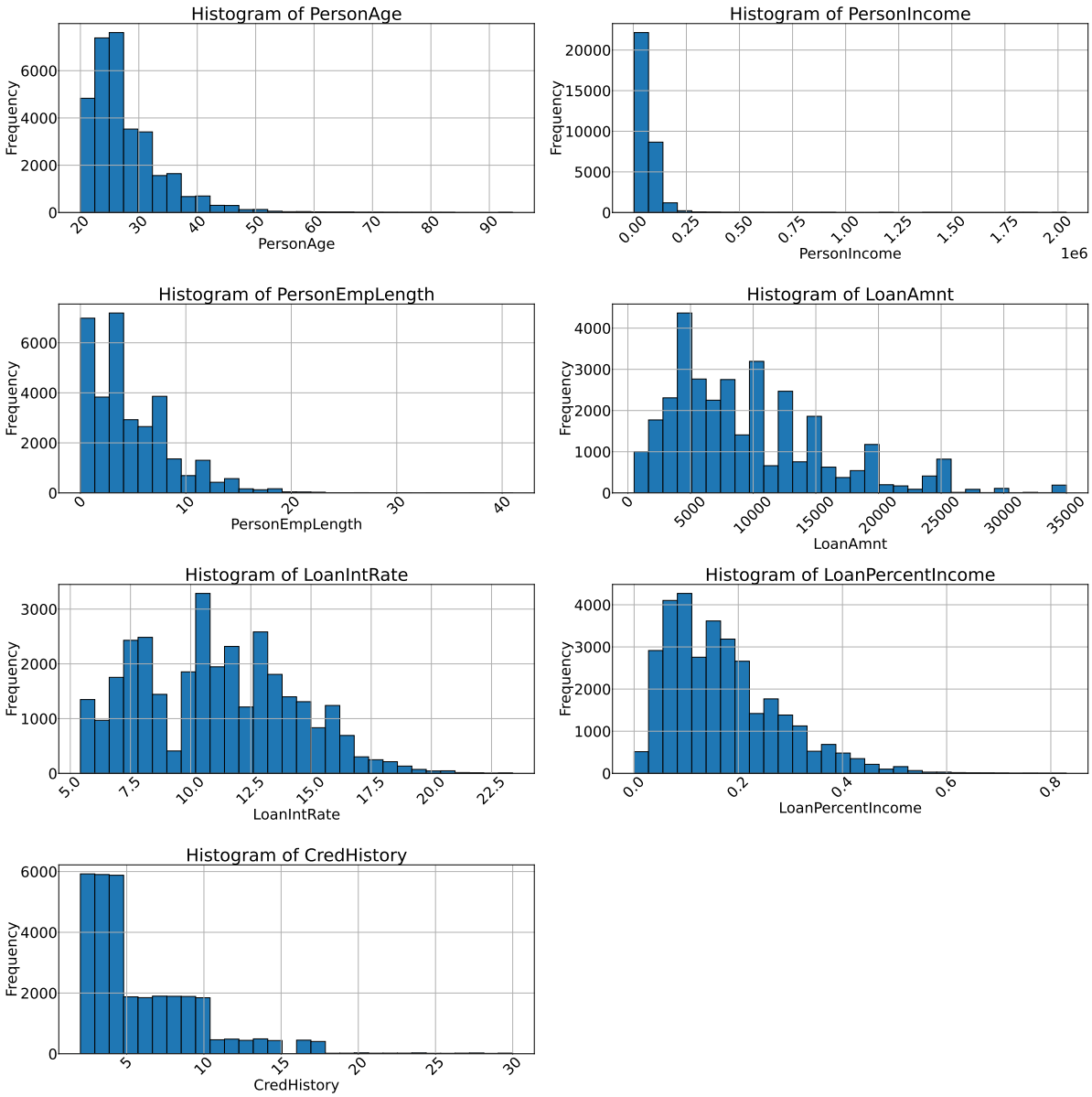

Figure 1: Histograms of all Numeric Variables

The histograms shown in Figure 1 illustrate the distributions for each numeric variable. All of the variables shown have positivley 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.

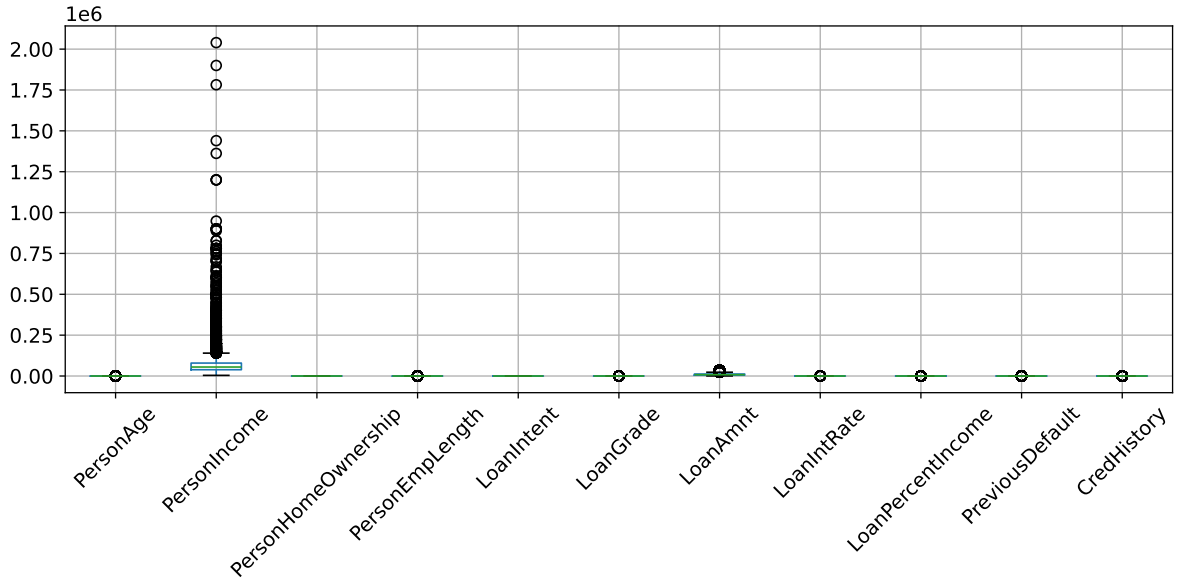

Figure 2: Box Plots of All Variables Before Normalisation

Figure 2 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 3. 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.

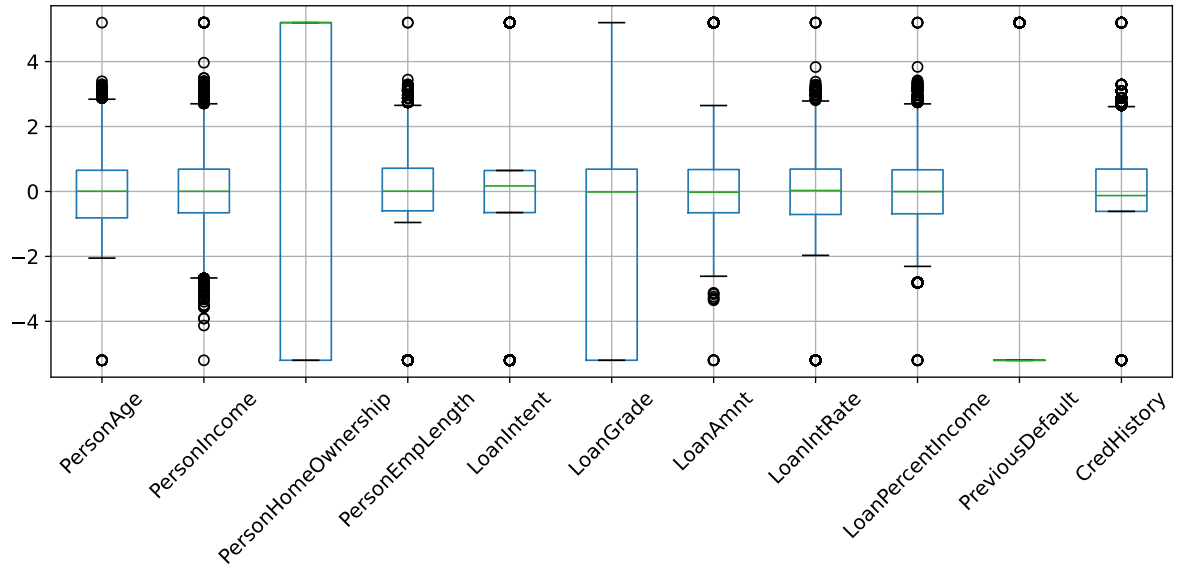

Figure 3: Box Plots of All Variables After Normalisation

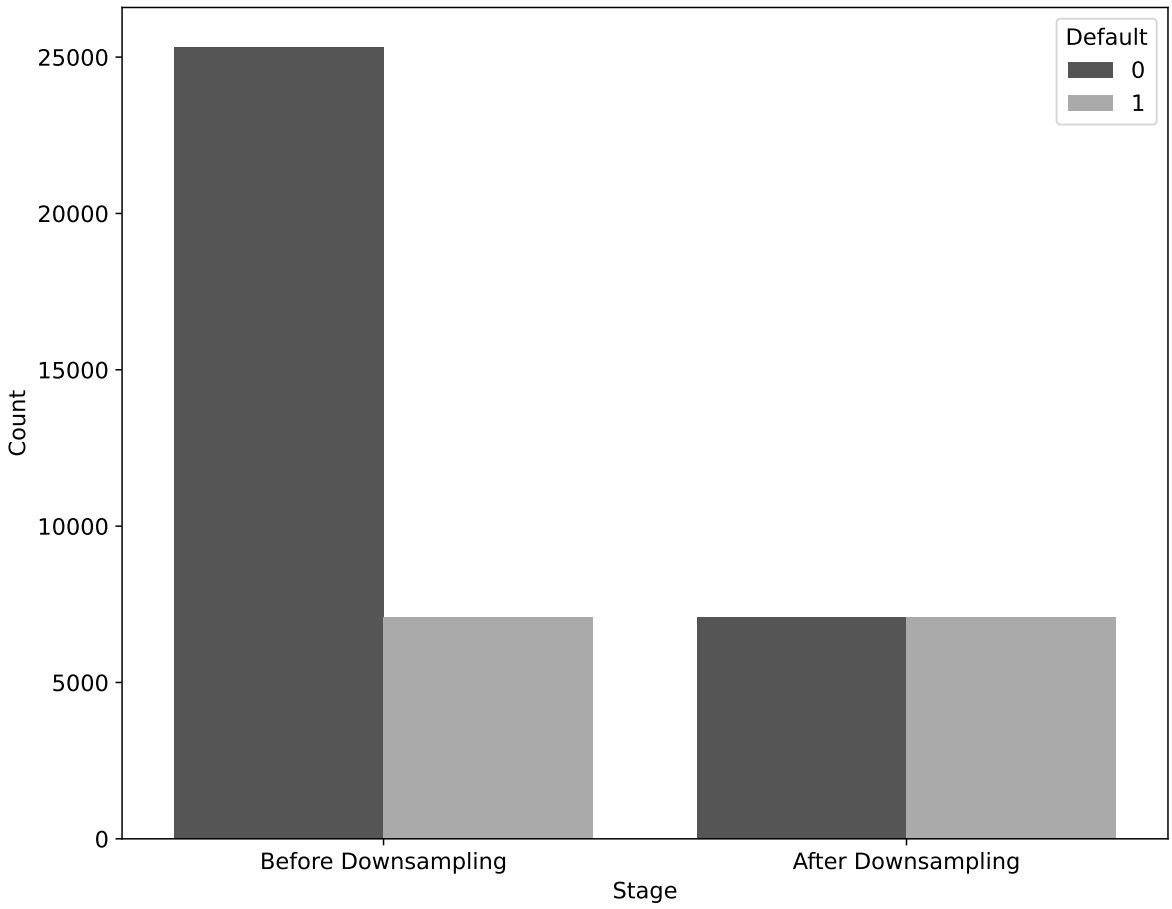

Figure 4: Distribution of LoanStatus Before and After Downsampling

Figure 4 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 4

2.4 Correlation Analysis

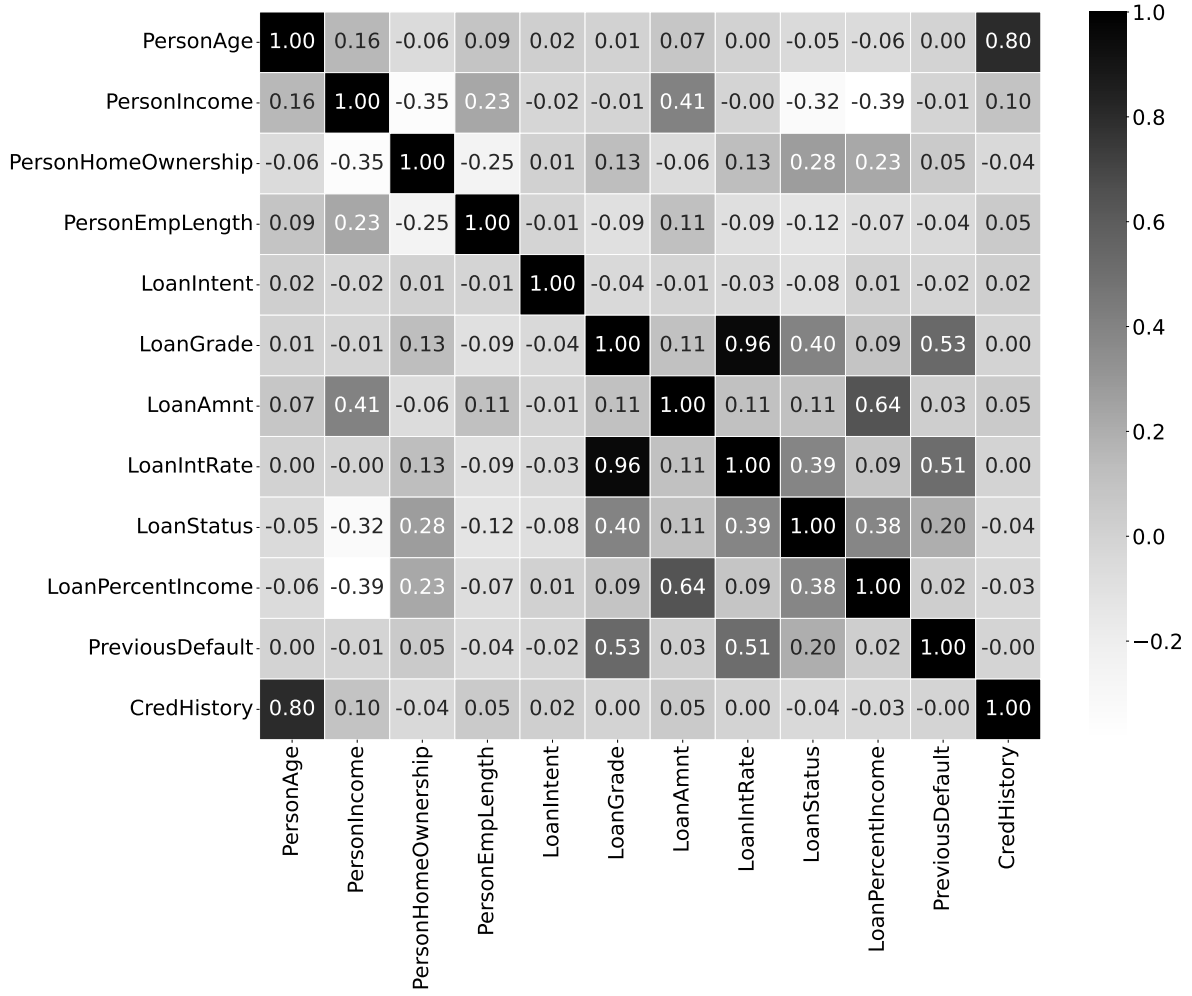

Figure 5: Correlation Plot of All Variables

Figure 5 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 = 0.8). 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	1.512
PersonIncome	9.587
PersonHomeOwnership	1.198
PersonEmpLength	1.064
LoanIntent	1.002
LoanGrade	3.017
LoanAmnt	12.837
LoanIntRate	3.147
LoanPercentIncome	12.357
PreviousDefault	1.253
CredHistory	1.48

VIF values for all the variables are shown within Table 4. In contrast to Figure 5, *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 and 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.

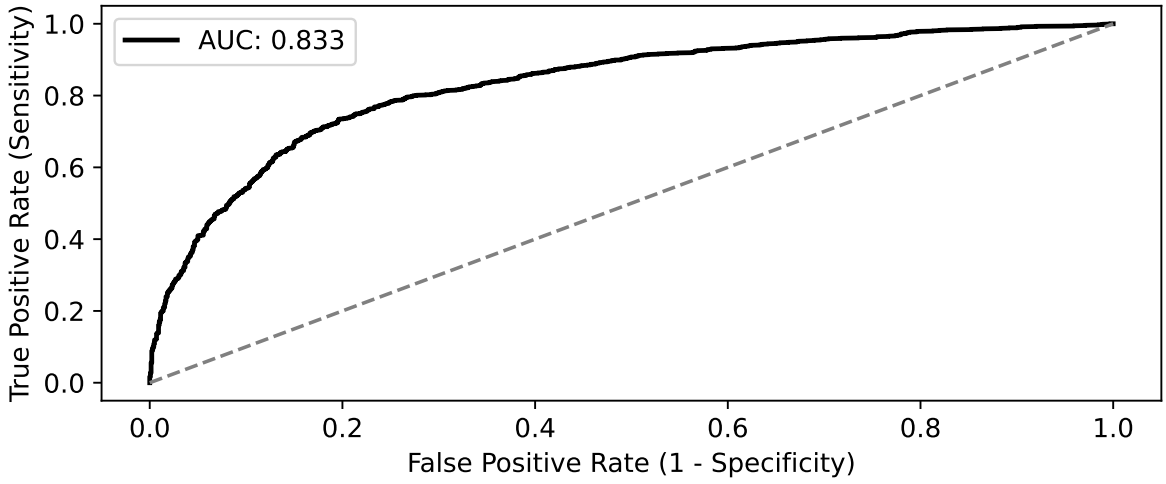

Figure 6: ROC Curve for Logistic Regression Model

Figure 6 shows the ROC curve for the LR model, an indication of the trade-off between sensitivity and specificity of the model. The model achieved a AUC score of 0.833 which is considered considerable (Çorbacıoğlu, 2023), indicating solid predictive performance when distinguishing between positive outcomes. The model's curve lies well above the diagonal reference line (AUC = 0.5), which represents random classification, demonstrating its predictive applications. However, the graph shows room for improvement due to true positive rate (TPR) remaining below 0.9.

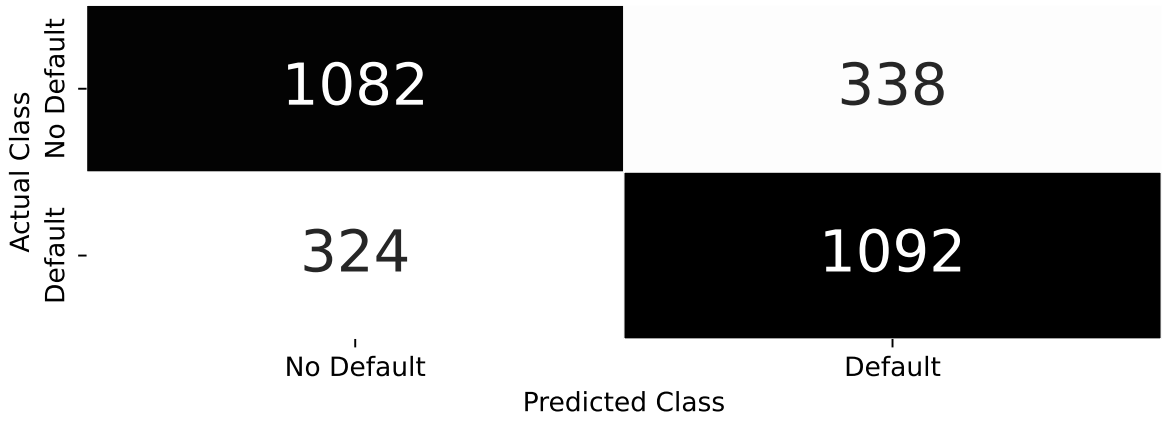

Figure 7: Confusion Matrix for Logistic Regression Model

Figure 7 visualises the error within the classification model. The matrix reveals that the model correctly identified 1082 non-default cases (true negatives) and 1092 default cases (true positives), demonstrating its ability to capture both classes effectively. However, 338 non-default cases were misclassified as defaults (false positives), while 324 default cases were incorrectly predicted as non-defaults (false negatives), potentially leading to losses in revenue for a financial institution.

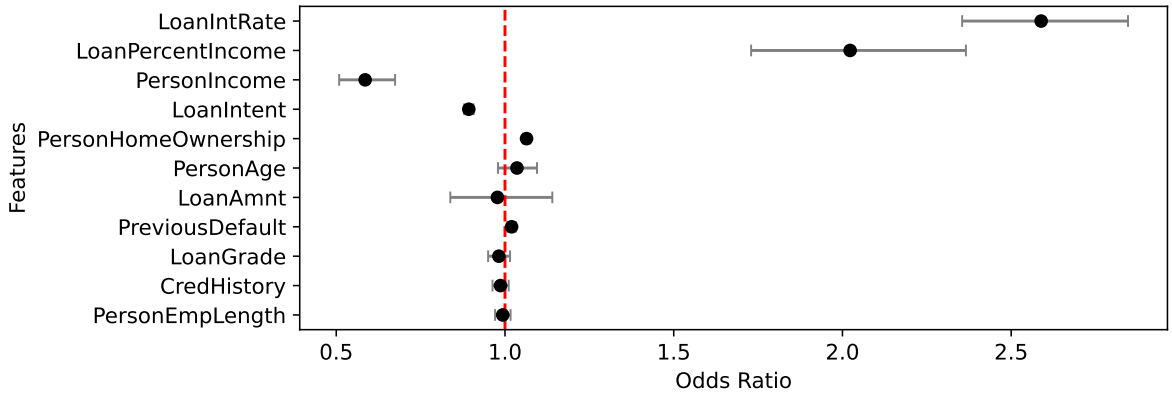

Figure 8: Odds Ratios for Logistic Regression Model

Figure 8 shows the odds ratios for the LR model. Odds ratios were calculated allowing an easy interpretation of the relationships between the individual features and credit risk. The odds ratio indicate the increase in the risk of defaulting for a one-unit increase in that variable. The results indicate that *LoanIntRate* and *LoanPercentIncome* have the strongest positive associations with default, with odds ratios of 2.589 and 2.023 , respectively. This suggests that as interest rates or the proportion of income allocated to the loan increase, the likelihood of default rises significantly. Conversely, *PersonIncome* has an odds ratio of 0.586, implying that higher income levels reduce the probability of default, aligning with expectations in traditional credit risk assessment.

3.2 Random Forest

The second model that was developed and compared with the LR model was an RF as they have been shown to have superior performance than LR models (Couronné et al., 2018). This model was trained on all the standard variables.

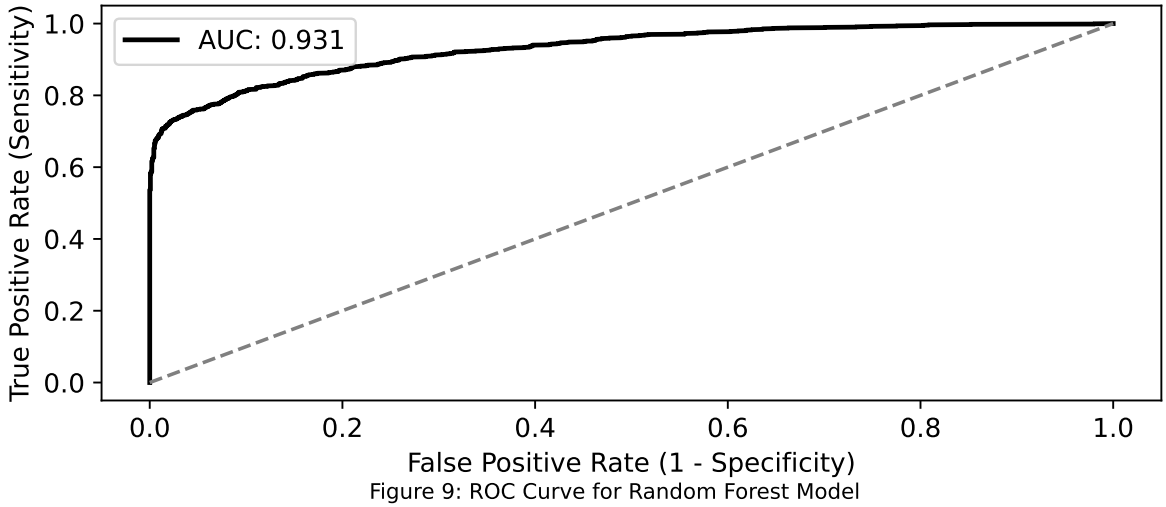

Figure 9: ROC Curve for Random Forest Model

The ROC curve, illustrated in Figure 9 for the RF model showcases its improved classification ability in distinguishing between defaulting and non-defaulting cases when compared to LRs. The model achieved a excellent AUC of 0.931, indicating strong predictive capability and shows that more complex models have the potential to improve credit risk prediction.

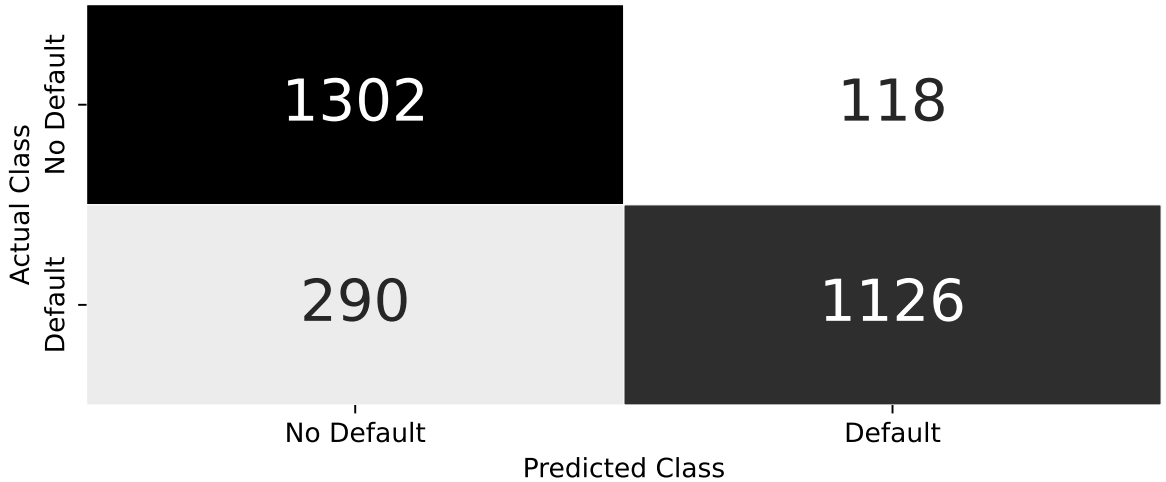

Figure 10: Confusion Matrix for Random Forest Model

The confusion matrix (Figure 10) for the RF model provides a detailed comparison of actual versus predicted default status. In this case, the model correctly predicted Non-default for

1302 instances (True Negatives), and correctly identified "Defaulting" for 1126 instances (True Positives). However, there were 118 False Positives, where the model incorrectly predicted "Defaulting" when the actual class was "No default," and 290 False Negatives. This confusion matrix reiterates the improved performance from the LR as the incorret classification instances have decreased.

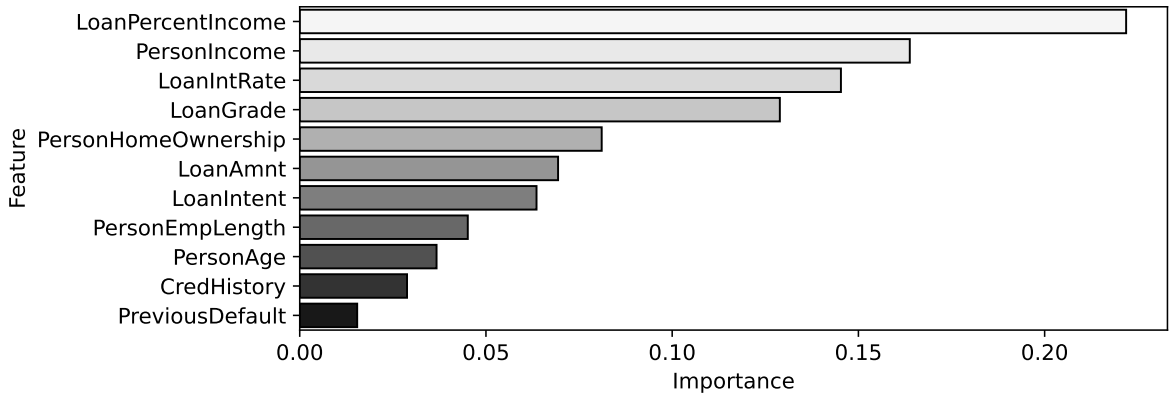

Figure 11: Feature Importances from Random Forest Model

Figure 11 demonstrates the most influential features when predicting credit risk by visualising feature importance calculated using mean decrease in accuracy. *LoanIntRate* is the most important feature suggesting that the proportion of income allocated to a loan has the strongest impact on the model's predictions, supporting the conclusions from the LR which ranked it second. *LoanPercentIncome* and *PersonIncome* are also shown to be within the top 3 most important features as they are in the LR model. Contrastly, to the LR, the RF shows *LoanGrade* to have high importance whereas Figure 8 shows it to have very little impact on credit risk for the LR model, potentially attributed to the differences in model architecture.

3.3 XGBoost

The third model that I deployed to improve upon the RF model was an XGBoost

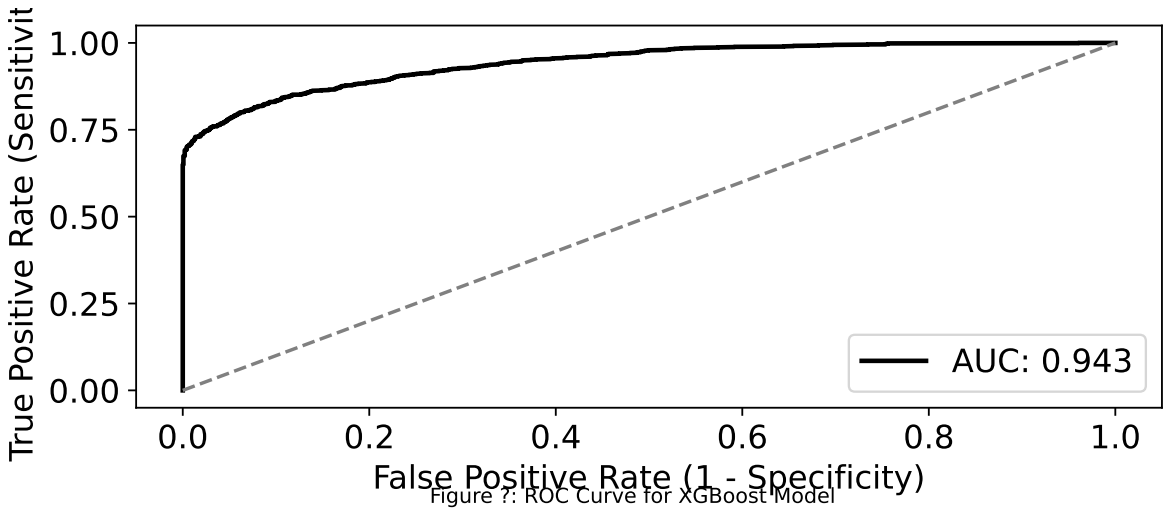

Figure ?: ROC Curve for XGBoost Model

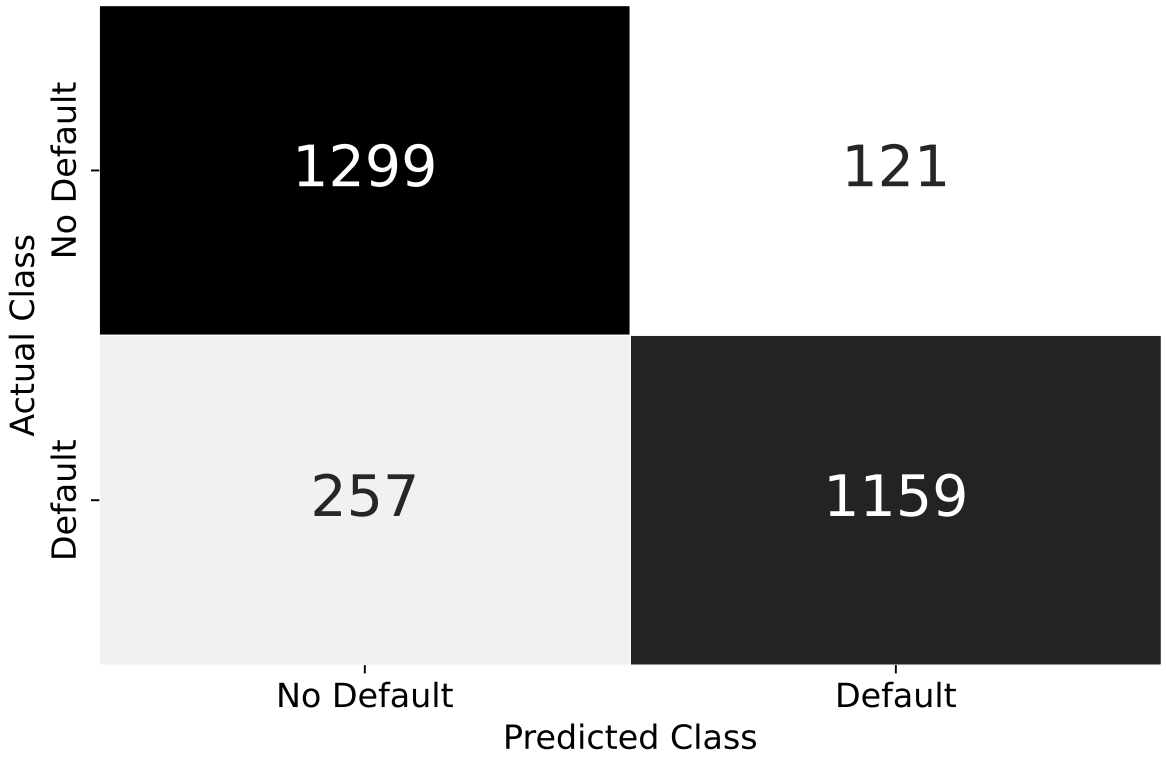

Figure ?: Confusion Matrix for XGBoost Model

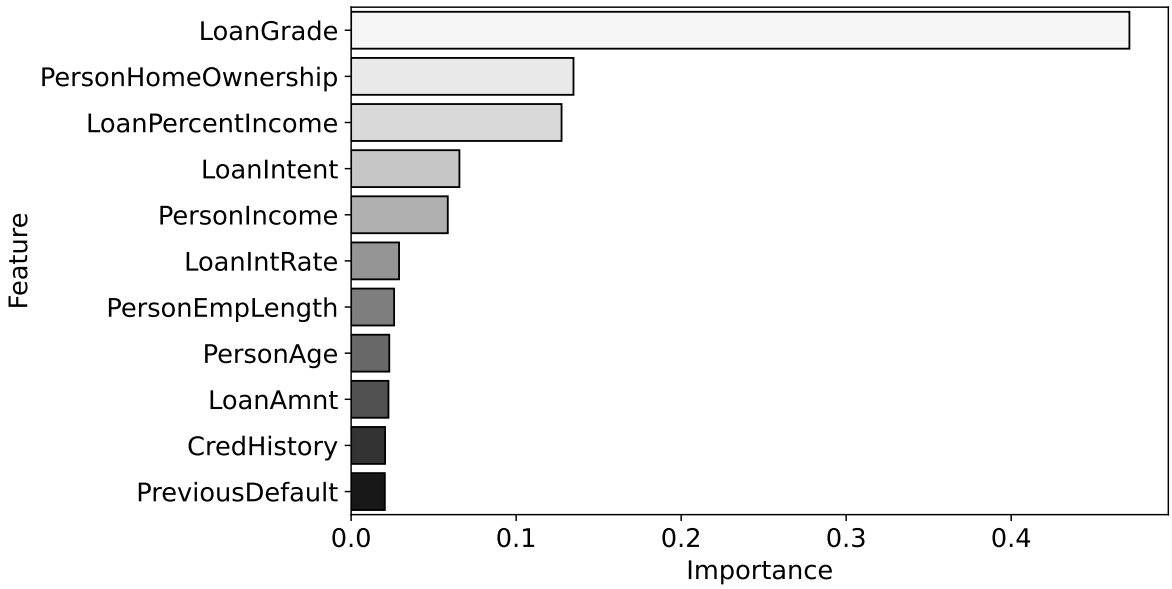

Figure ?: Feature Importances from XGBoost Model

3.4 Light Gradient Boosted Machine

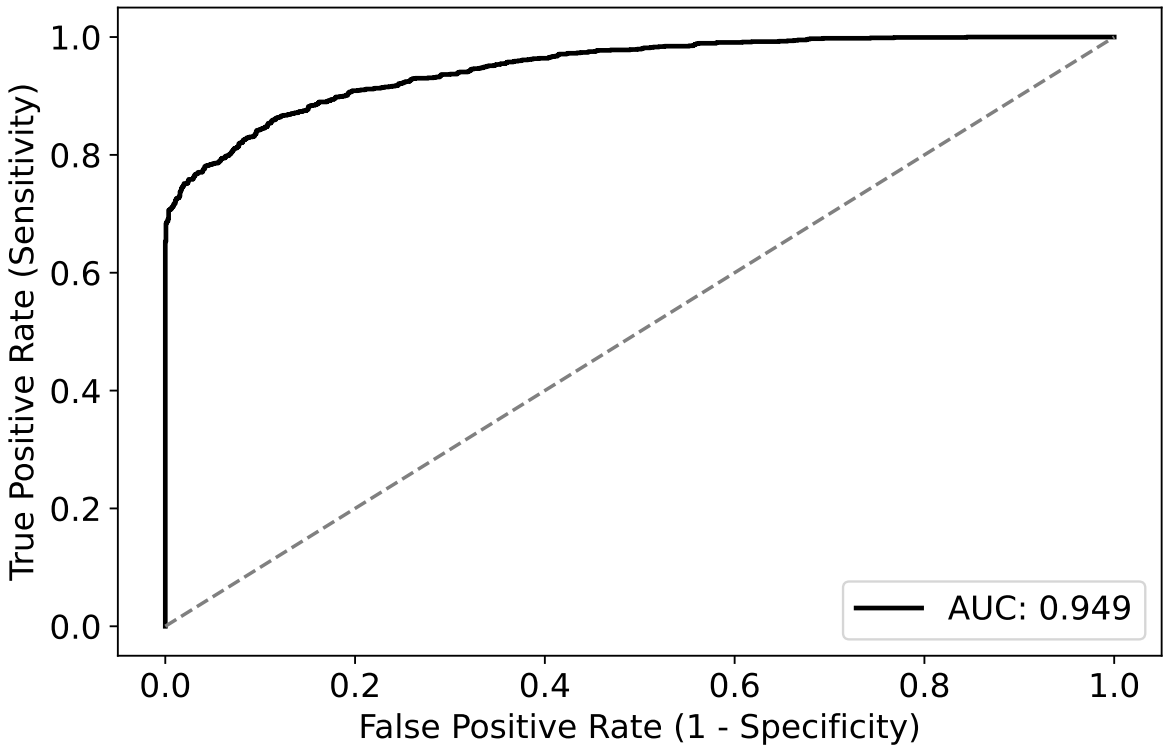

Figure ?: ROC Curve for LightGBM Model

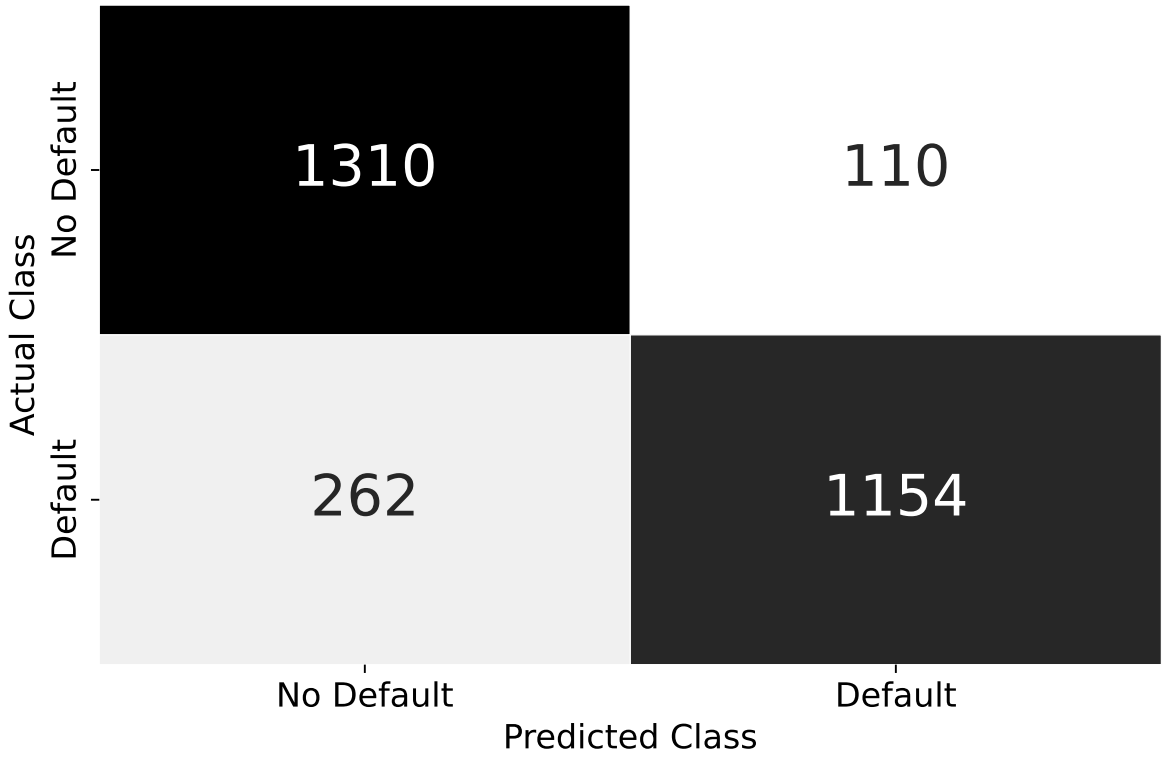

Figure ?: Confusion Matrix for LightGBM Model

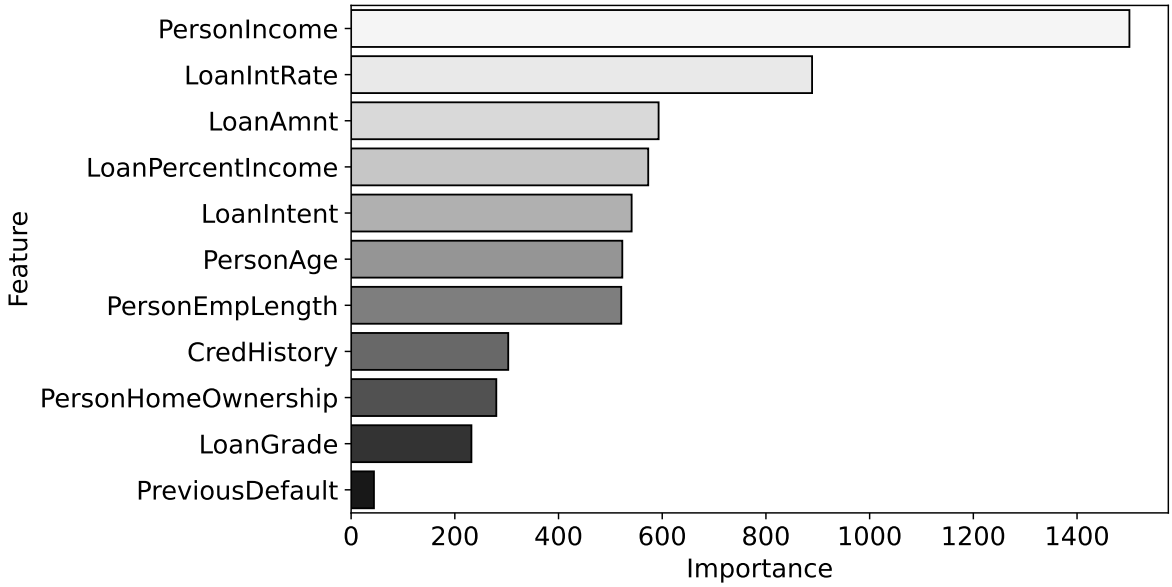

Figure ?: Feature Importances from LightGBM Model

3.5 Model Evaluation and Comparisons

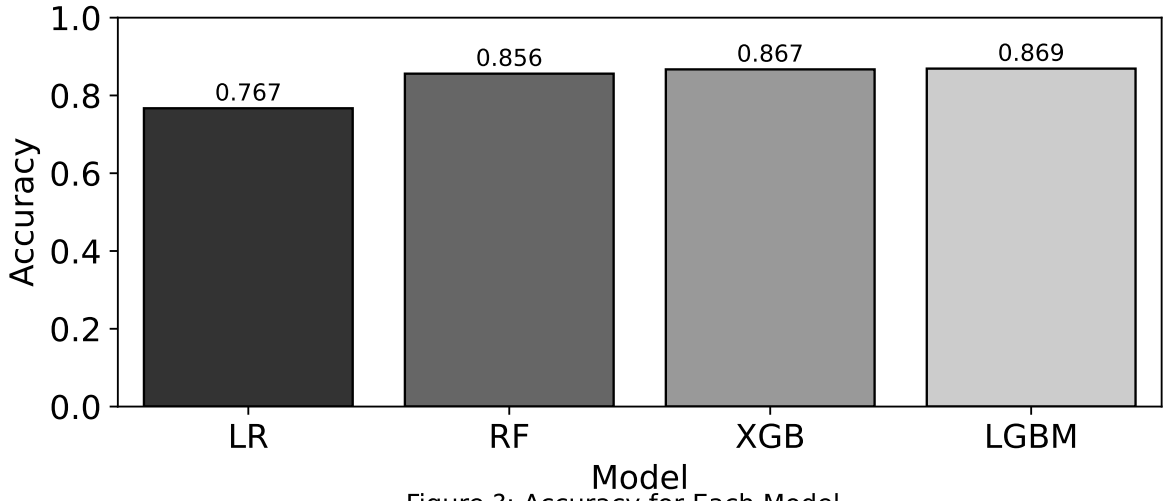

Figure ?: Accuracy for Each Model

Table 5: Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1 Score	AUC	Log Loss	Brier Score
LR	0.767	0.764	0.771	0.767	0.833	8.414	0.233
RF	0.856	0.905	0.795	0.847	0.931	5.185	0.144
XGB	0.867	0.905	0.819	0.86	0.943	4.804	0.133
LGBM	0.869	0.913	0.815	0.861	0.949	4.728	0.131

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	PersonHomeOwnership	LoanIntRate
Feature 3	LoanIntRate	LoanIntRate	LoanPercentIncome	LoanAmnt

4. Conclusion

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