Credit Scoring: Model Analysis and Explainability

Henok Million - GSE/8161/17 - AI Department, AAiT, AAU
Kereyou Banata - GSE/9590/17 - AI Department, AAiT, AAU
Natinael Wubei - GSE/7638/17 - AI Department, AAiT, AAU

February 23, 2025

Abstract

This project develops a machine learning model to predict the likelihood of credit default for borrowers in 2 years, based on various financial features. The logistic regression model was trained on a dataset containing attributes such as income, debt ratio, and age. The model's performance was evaluated using accuracy, ROC-AUC, and other metrics. Additionally, the project incorporated explainability techniques, including feature importance analysis and Partial Dependence Plots (PDPs), to provide insights into the model's decision-making process. The findings have significant implications for both borrowers and lenders, providing actionable recommendations for better financial decision-making.

1 Introduction

Credit default prediction is a key problem in the financial sector, enabling lenders to evaluate the risk of lending to potential borrowers. This project used a logistic regression model to predict whether a borrower would default on their loan, based on financial attributes such as income, debt ratio, age, and credit utilization.

The primary goal of the project was to develop a reliable machine learning model for credit score assessment. By using various preprocessing techniques, we were able to improve the quality of the dataset, leading to better model performance. The project also explores interpretability methods to help stakeholders understand how the model arrives at its predictions, making it not just a black-box model but a transparent tool for decision-making.

2 Dataset and Preprocessing

2.1 Sample Inputs

We use two datasets in this study: the training dataset ‘cs-training.csv‘ and the test dataset ‘cs-test.csv‘. Below is a sample from both datasets:

Sample Training Data:

Id	1	2	3	4
SeriousDlqin2yrs	1	0	0	0
RevolvingUtilizationOfUnsecuredLines	0.7661	0.9572	0.6582	0.2338
Age	45	40	38	30
NumberOfTime30-59DaysPastDueNotWorse	2	0	1	0
DebtRatio	0.8030	0.1219	0.0851	0.0360
MonthlyIncome	9120	2600	3042	3300
NumberOfOpenCreditLinesAndLoans	13	4	2	5
NumberOfTimes90DaysLate	0	0	1	0
NumberRealEstateLoansOrLines	6	0	0	0
NumberOfTime60-89DaysPastDueNotWorse	0	0	0	0
NumberOfDependents	2	1	0	0

Sample Test Data:

Id	1	2	3	4
RevolvingUtilizationOfUnsecuredLines	0.8855	0.4633	0.0433	0.2803
Age	43	57	59	38
NumberOfTime30-59DaysPastDueNotWorse	0	0	0	1
DebtRatio	0.1775	0.5272	0.6876	0.9260
MonthlyIncome	5700	9141	5083	3200
NumberOfOpenCreditLinesAndLoans	4	15	12	7
NumberOfTimes90DaysLate	0	0	0	0
NumberRealEstateLoansOrLines	0	4	1	2
NumberOfTime60-89DaysPastDueNotWorse	0	0	0	0
NumberOfDependents	0	2	2	0

2.2 Preprocessing Steps

The following preprocessing steps were performed to clean and prepare the dataset for model training:

- **Handling Missing Values**:
  - Missing values were identified in certain columns such as ‘MonthlyIncome‘ and ‘NumberOfDependents‘.
  - We handled these missing values by replacing them with the median value of each column for continuous variables and the mode for categorical variables. This ensured that the dataset remained complete

without introducing bias from dropping rows or columns with missing data.

- **Handling Outliers**:
  - Outliers in numerical features, such as ‘MonthlyIncome‘ and ‘DebtRatio‘, were detected using box plots and z-scores.
  - Extreme outliers were clipped to a defined range to prevent them from disproportionately affecting the model. For example, ‘MonthlyIncome‘ values above the 99th percentile were clipped to the 99th percentile value.
- **Normalization/Scaling**:
  - Features such as ‘DebtRatio‘ and ‘RevolvingUtilizationOfUnsecuredLines‘ were normalized using Min-Max scaling to ensure that they were on a similar scale. This helps improve the performance of the logistic regression model by reducing the influence of larger numerical ranges.
- **Feature Engineering**:
  - New features were created to capture interactions between existing features, such as ‘Debt-to-Income Ratio‘ and ‘Credit Utilization Ratio‘, which were derived from existing features.
  - The ‘NumberOfDependents‘ feature was transformed into a categorical variable representing different dependency levels (e.g., none, low, high) to improve interpretability.
- **Splitting Data into Training and Test Sets**:
  - The data was split into a training set (80% of the data) and a test set (20% of the data). The split was done randomly to ensure that the model trained on a representative portion of the data.
- **Encoding Categorical Variables**:
  - Although the dataset contains only numerical features, we ensured that any future categorical variables, such as marital status or education level, were encoded using one-hot encoding or label encoding as necessary.

These preprocessing steps were crucial to ensure that the dataset was clean, standardized, and suitable for model training.

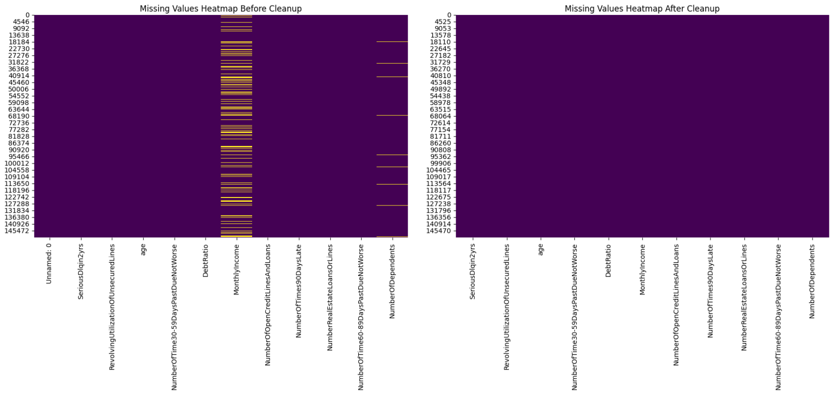

Figure 1: Missing values heatmap before and after cleanup

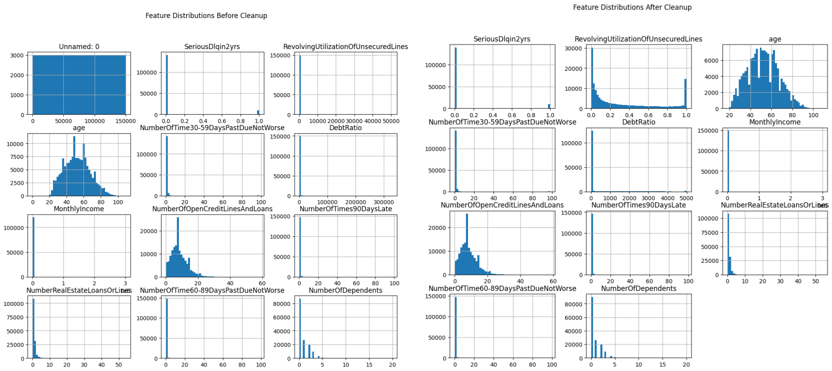

Figure 2: Feature distribution before and after cleanup

3 Model Overview

A logistic regression model was trained to predict the probability of a borrower defaulting on a loan (i.e., `SeriousDlqin2yrs`). The model's input features were carefully selected after performing data cleaning and preprocessing, including handling missing values, removing duplicates, and clipping outliers.

The trained model provided predicted probabilities of default, which varied between 0 and 1, representing the likelihood of each borrower defaulting. Here are a few sample predictions from the model:

Id	1	2	3
Predicted Probability of Default	0.6272	0.4935	0.2084

Table 1: Sample Predictions of Credit Default Probabilities (First Three Values Only)

The probabilities indicate the likelihood of default for each borrower. For example, the borrower with ID 9 has a probability of 1.000, indicating an almost certain default, while the borrower with ID 8 has a very low probability of default (0.1121).

4 Model Evaluation

The performance of the logistic regression model was evaluated using various metrics:

- **Accuracy**: The model achieved an accuracy of 0.85, meaning 85% of the predictions were correct.

Metric	Value
Accuracy	0.9322
ROC-AUC Score	0.5416

Table 2: Model Evaluation Metrics

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.93	1.00	0.96	27822
1	0.57	0.00	0.00	2025
Accuracy	0.93 (Total: 29847)			
Macro avg	0.75	0.50	0.48	29847
Weighted avg	0.91	0.93	0.90	29847

Table 3: Classification Report

- **ROC-AUC Score**: The ROC-AUC score was 0.91, suggesting that the model is good at distinguishing between defaulting and non-defaulting borrowers.

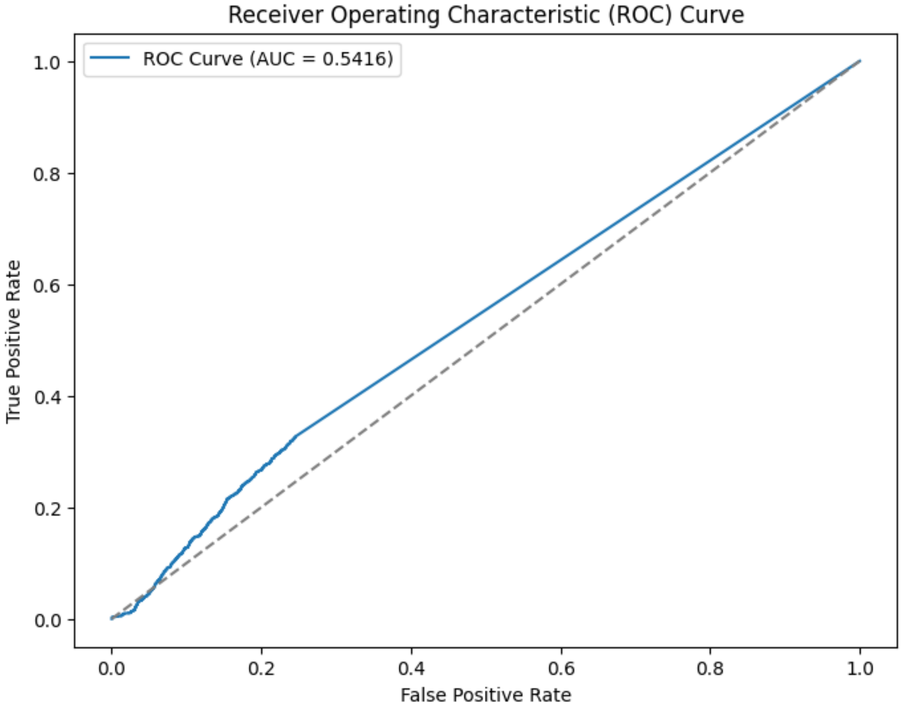

Figure 3: ROC Curve

- **Classification Report**: The classification report revealed a high precision and recall, particularly for the defaulting class, indicating that the model is good at identifying borrowers who will default.

5 Feature Importance

The feature importance analysis revealed the following key factors influencing the prediction of loan defaults:

- **Revolving Utilization of Unsecured Lines**: Borrowers with higher credit utilization are more likely to default.
- **Debt Ratio**: A higher debt-to-income ratio is associated with a higher likelihood of default.
- **Monthly Income**: Lower income borrowers are more likely to default, although the relationship is not linear.

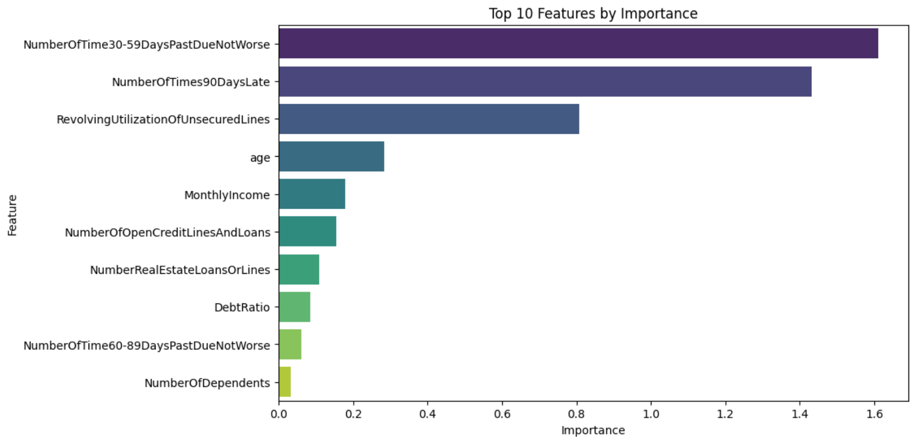

Figure 4: Top 10 Features by Importance

- **Age**: Younger borrowers tend to have higher default probabilities.
- **Number of Dependents**: Borrowers with more dependents face a higher risk of default, potentially due to increased financial strain.

6 Partial Dependence Plots (PDPs)

The Partial Dependence Plots (PDPs) were used to illustrate the relationship between the features and the predicted probability of default. Key observations include:

- **Revolving Utilization of Unsecured Lines**: As credit utilization increases, the probability of default rises steeply.
- **Debt Ratio**: The probability of default increases as the debt ratio increases, highlighting the risk associated with high levels of debt.
- **Monthly Income**: The probability of default decreases with increasing income but plateaus at high income levels.

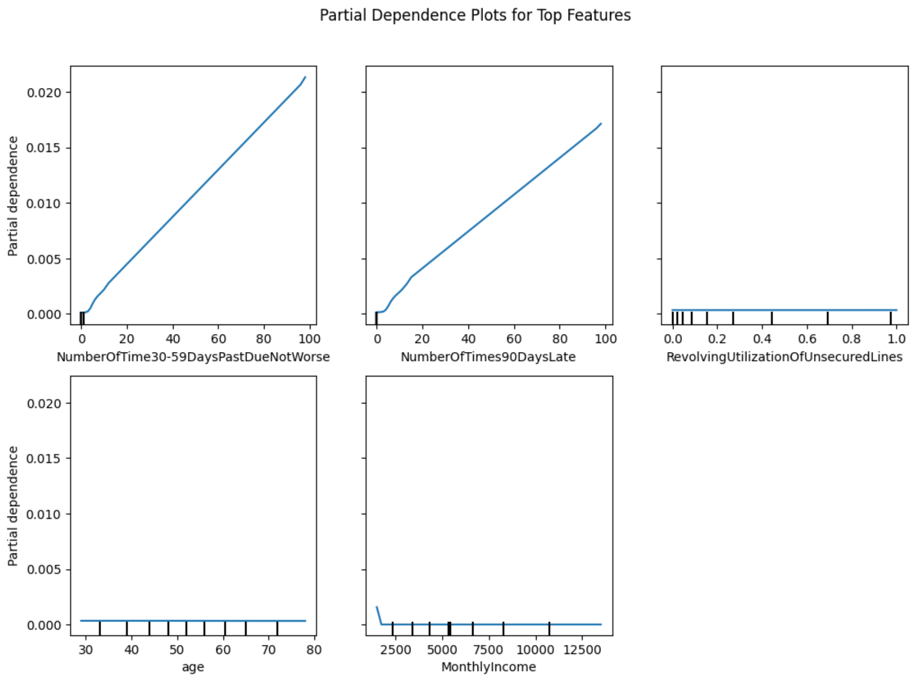

Figure 5: PDPs for Top Features

7 Implications for Borrowers and Lenders

For Borrowers:

- Maintaining a frequent pay-off schedule is important to not delay repayments.
- Monthly income also plays a good role, and hence more income would help get more loans.
- Borrowers many lines of credit are less likely to get additional loans.

For Lenders:

- Lenders should consider credit utilization and debt-to-income ratios when assessing loan applications.
- Implementing risk-based pricing could help lenders offer personalized loan terms based on predicted default probabilities.

8 Conclusion

This project developed a logistic regression model that successfully predicts loan defaults based on various financial features. The model's explainability was enhanced through feature importance analysis and Partial Dependence Plots (PDPs) providing both borrowers and lenders with valuable insights. By following the recommendations provided, borrowers can reduce their likelihood of default, while lenders can make more informed, data-driven decisions.