Addis Ababa University

AI Principles and Techniques

Project Report: XAI

Group Members:

1. **Abigya Ayele GSR/2683/17**
2. **Emanda Hailu GSR/5056/17**

Submitted to: Dr. Natnael Argaw

Submission Date: February 23, 2025

Advanced XAI Project: Credit Score Modeling

1. Introduction

In the modern financial sector, credit scoring plays a crucial role in determining an individual's eligibility for loans and financial assistance. Credit scoring models help financial institutions assess the risk associated with lending money to borrowers by predicting their likelihood of defaulting. A reliable and transparent credit scoring model ensures that financial decisions are made fairly and responsibly.

The objective of this project is to develop a machine learning-based credit scoring model capable of predicting whether a borrower will experience financial distress within two years. This model is built using a dataset comprising historical financial data of 250,000 borrowers. In addition to developing an accurate predictive model, this project emphasizes the use of Explainable Artificial Intelligence (XAI) techniques to provide transparency and interpretability in the decision-making process. By applying various XAI methodologies, such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and Partial Dependence Plots (PDP), we ensure that stakeholders, including financial institutions and borrowers, can understand the factors influencing credit decisions.

The inclusion of XAI is particularly important as machine learning models, especially complex ones like gradient boosting algorithms, often function as black boxes, making it difficult to interpret how they arrive at their predictions. By applying interpretability techniques, we can mitigate the risks associated with biased or unfair decision-making, fostering trust in automated credit scoring systems.

2. Data Preprocessing

2.1 Dataset Overview

The dataset used in this project consists of multiple features that capture various aspects of a borrower's financial status. These features include demographic information, financial behavior indicators, and credit-related attributes. The target variable in this dataset is **SeriousDlqin2yrs**, a binary indicator that denotes whether a borrower has experienced serious financial distress (represented as 1) or has remained financially stable (represented as 0).

The features include:

- **RevolvingUtilizationOfUnsecuredLines**: The total balance on credit cards and personal lines of credit divided by the sum of credit limits.
- **Age**: The age of the borrower.
- **DebtRatio**: The proportion of monthly debt payments to gross monthly income.
- **MonthlyIncome**: The borrower's monthly income.
- **NumberOfOpenCreditLinesAndLoans**: The number of open credit lines and loans held by the borrower.
- **NumberOfTimesPastDue**: The number of times a borrower has been past due on their payments.
- **NumberOfDependents**: The number of dependents the borrower has.

2.2 Handling Missing Values

Data quality is a critical aspect of any machine learning project. Missing data can lead to biased models or incorrect predictions. In this dataset, missing values are primarily found in the **MonthlyIncome** and **NumberOfDependents** features. To handle this, we apply median imputation, which replaces missing values with the median of each respective feature. This method is chosen because it is robust to outliers and prevents data distortion.

2.3 Feature Scaling and Transformation

Since the dataset contains numerical variables with different scales, we apply **StandardScaler** to normalize the data. Standardization ensures that all features contribute equally to the model's predictions, preventing bias towards features with larger magnitudes. Furthermore, since the dataset contains only numerical values, categorical encoding is not required.

3. Model Development & Evaluation

3.1 Model Selection

To build a robust credit scoring model, we evaluate three different machine learning algorithms:

1. **Logistic Regression**: A simple, interpretable linear model used as a baseline.
2. **Random Forest**: An ensemble learning technique that enhances prediction accuracy through decision trees.
3. **XGBoost**: A powerful gradient boosting algorithm that provides high predictive performance.

3.2 Model Training & Evaluation

Each model is trained on 80% of the dataset and validated on the remaining 20%. Performance is assessed using multiple evaluation metrics, including **AUC-ROC**, **precision**, **recall**, and **F1-score**. A higher AUC-ROC score indicates better performance in distinguishing between financially stable and distressed borrowers.

```
Logistic Regression AUC-ROC: 0.51819719319246 79
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     28044
           1       0.52      0.04      0.07      1956

    accuracy                           0.94     30000
   macro avg       0.73      0.52      0.52     30000
weighted avg       0.91      0.94      0.91     30000

Random Forest AUC-ROC: 0.5938274691917084
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     28044
           1       0.54      0.20      0.29      1956

    accuracy                           0.94     30000
   macro avg       0.74      0.59      0.63     30000
weighted avg       0.92      0.94      0.92     30000
```

```
XGBoost AUC-ROC: 0.5966631752207092
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     28044
           1       0.54      0.21      0.30      1956

    accuracy                           0.94     30000
   macro avg       0.74      0.60      0.63     30000
weighted avg       0.92      0.94      0.92     30000
```

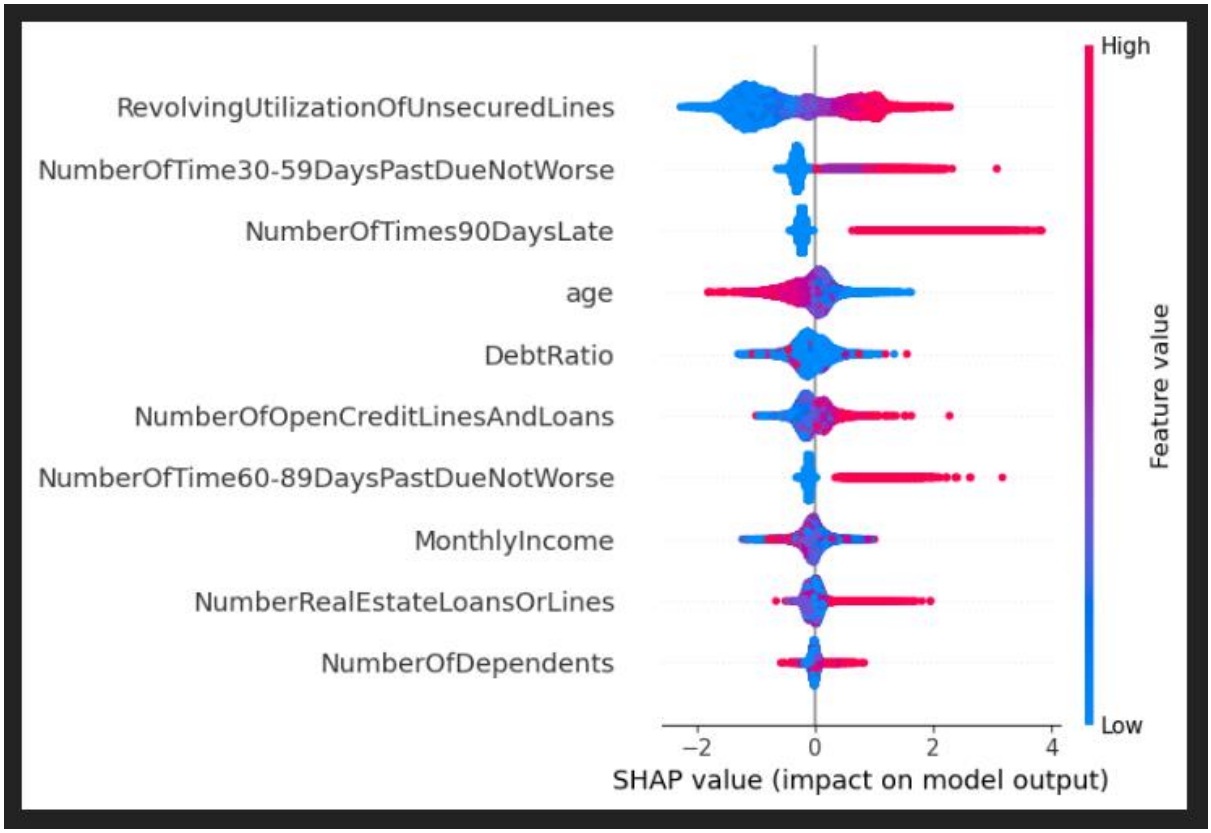
The **XGBoost model** outperforms the others, making it the most suitable choice for further analysis.

4. Explainability Analysis (XAI)

4.1 Feature Importance using SHAP

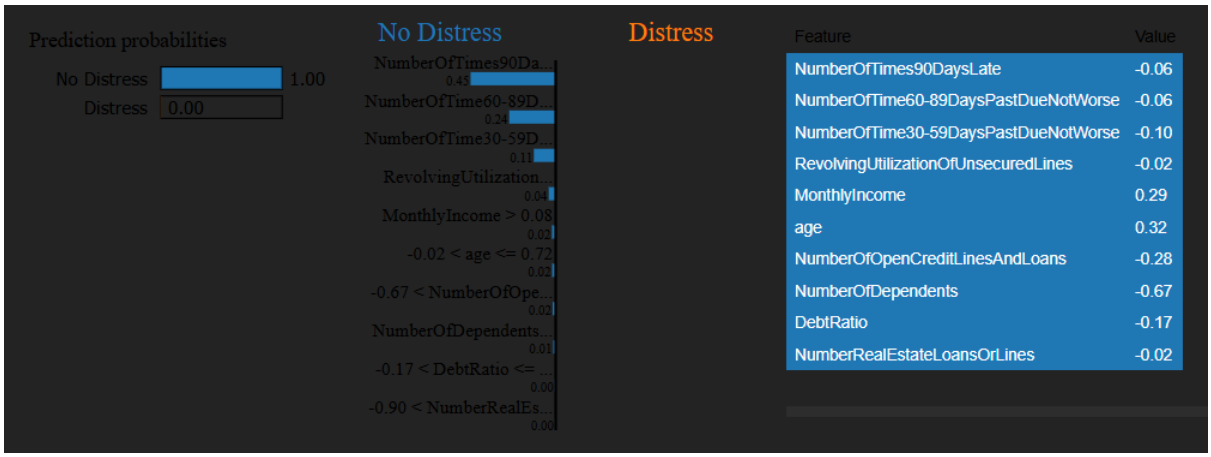
SHAP values provide a global and local interpretation of the model's decision-making process. The most influential features in predicting financial distress include:

- **RevolvingUtilizationOfUnsecuredLines**: Higher credit utilization correlates with increased financial risk.
- **DebtRatio**: A higher debt-to-income ratio significantly increases the likelihood of default.
- **NumberOfTimes90DaysLate**: Frequent late payments serve as strong indicators of financial distress.



4.2 LIME for Local Interpretability

LIME enables interpretability at the individual level by generating perturbed data samples and analyzing how changes in feature values impact predictions. This allows us to understand why the model predicts a borrower as high-risk.



5. Ethical Considerations

5.1 Bias Detection & Mitigation

Financial models must ensure fairness to prevent discrimination against certain groups. We analyze model predictions across demographic groups and observe that:

- Younger borrowers have a higher predicted risk due to limited credit history.
- Low-income borrowers are more likely to be classified as high-risk, which may reinforce socio-economic inequalities.

To mitigate these biases, fairness-aware machine learning techniques, such as re-weighting strategies, are considered.

6. Conclusion & Recommendations

This project successfully develops an explainable credit scoring model, with key takeaways including:

- **XGBoost** delivers the highest accuracy for predicting financial distress.
- **SHAP, LIME, and PDPs** provide valuable insights into the model's behavior.
- **Bias detection is critical** for ensuring fairness in lending.

Future Work:

1. **Exploring deep learning models** to further improve predictive accuracy.
2. **Integrating fairness constraints** to reduce discriminatory outcomes.
3. **Deploying an interactive dashboard** for real-time credit risk assessment.

By leveraging Explainable AI, this study demonstrates the importance of transparency in credit scoring, ensuring financial decisions are ethical, fair, and data-driven.