Credit Risk Assessment with Open Banking: An Application of Machine Learning

Yueying Li
[1]University College London, [2]Salad Money

Background and Motivation

Credit risk refers to the likelihood of a borrower failing to meet their loan repayment obligations. This directly impacts the financial stability and profitability of lending institutions. **Traditional credit scoring systems** often exclude or misrepresent applicants with limited credit history ('thin' files), thereby restricting access to loans for otherwise creditworthy individuals. Recent advances in **machine learning** provide the ability to analyse complex financial behaviours, which traditional methods may overlook. These algorithms learn from vast datasets to predict credit risk more accurately.

Open Banking enables real-time financial data sharing, with customer consent, offering a richer view of borrower behaviour. This improves the accuracy of credit risk models by providing more comprehensive insights. Financial institutions like Salad Money, which cater to populations with impaired credit scores, utilise open banking and machine learning to assess creditworthiness more inclusively, thereby contributing to **financial inclusion** and **ethical lending**.

Problem Statement

The problem this study aims to address is the **prediction of credit risk** among populations not well-represented by traditional credit metrics. This includes applicants from both the public and private sectors, who may exhibit irregular income patterns or atypical financial behaviours.

Data and Methodology

Data was sourced from internal company systems via **Open Banking APIs**. This dataset includes detailed transactional data on credit applicants, offering insights into their financial behaviour. The dataset comprises **12,123 credit applications** and **4,333 features**, spanning from May 2022 to August 2023. It includes features like salary, loan payments, benefits, and risk markers such as CCJs.

Initial data review focused on understanding the structure, distribution of features, and identifying significant patterns and trends. A notable finding was the **imbalance** in the dataset, with a small percentage of 'never payers' (approx. 3.18%). Demographic and financial behaviour analyses identified key trends. For instance, **younger applicants** and those **living with parents** were more likely to be never payers.

The **preprocessing and feature engineering** steps involved data cleaning, feature transformation, feature selection, and data balancing, resulting an final clean dataset with 110 features. The refined feature set allowed the models to focus on the **most impactful variables**, enhancing their predictive accuracy.

In the classification model development phase, various algorithms, including **Logistic Regression**, **SVM**, **Random Forest**, **GBM**, and **Multi-Layer Perceptron (MLP)**, were tested. To address the class imbalance, techniques such as cost-sensitive learning and Balanced Random Forest were employed. The models were evaluated using performance metrics like accuracy, precision, recall, F1 score, and AUC.

Key Findings

The results of the study highlight the performance of various machine learning models in predicting 'never payers.' The **Balanced Random Forest** model outperformed other models by effectively addressing the dataset's inherent class imbalance. It achieved the highest **AUC score of 0.64**, with a **recall of 0.58**, indicating significant improvement in identifying never payers compared to other models. Although the accuracy of this model was lower than others at 0.61, it demonstrated **better sensitivity** in identifying the minority class, which was crucial for the task.

Feature importance analysis revealed that financial behaviours such as returned direct debits were key predictors of credit risk. Overall, the Balanced Random Forest model was found to be the most suitable for predicting never payers in this context.

Value of the Research

This research provides valuable insights into **enhancing credit risk assessment** by integrating **open banking data** with **machine learning models**, particularly for underrepresented populations. By addressing the limitations of traditional credit scoring systems and improving predictions for applicants with limited credit histories, the study supports more **inclusive and ethical lending practices**. The findings can help financial institutions **reduce default risks** while **promoting financial inclusion**, making the research both commercially and socially impactful.