Proposal: Predicting Loan Default Probability Using Transaction Data

Brandon Dioneda bdioneda@ucsd.edu
Mert Ozer mozer@ucsd.edu
Qianjin Zhou q9zhou@ucsd.edu

Brian Duke (Mentor) brian.duke@prismdata.com
Kyle Nero (Mentor) kyle.nero@prismdata.com
Berk Ustun (Mentor) berk@ucsd.edu

1 Introduction

In the financial services industry, accurately predicting the likelihood of a loan applicant defaulting is crucial for risk management and financial stability. This project aims to develop a machine learning model that predicts default probabilities by analyzing individuals' bank transaction data.

2 Problem Statement

2.1 For a General Audience

We are working on creating a system that helps financial service providers decide whether to grant loans to individuals. By analyzing people's spending and income patterns from their bank transactions, we can predict how likely they are to repay a loan. This helps lenders make better decisions and reduces the risk of financial loss.

2.2 For a Domain Expert

The objective is to build a predictive model that estimates the probability of default for loan applicants by leveraging categorized inflow and outflow transaction data. We will engineer features related to balance fluctuations, income stability, expenditure categories, and their temporal changes. The model will be trained using appropriate machine learning algorithms, optimized for performance and efficiency while keeping the model unbiased and avoiding discriminatory practices.

3 Data Acquisition and Quality

- **Data Availability:** The necessary transaction data will be provided by our mentor (PrismData), ensuring we have access to relevant and comprehensive datasets.
- **Data Suitability:** The data includes categorized inflow and outflow transactions, sufficient to extract features like income levels, spending habits, and balance changes.
- **Data Quality:** Preliminary assessments indicate the data is of high quality, with no missing values, suitable for training robust machine learning models.

4 Project Goals and Impact

Investing 10 weeks in this project is justified due to its potential to significantly improve lending decisions. The project's success could lead to more reliable credit risk assessments and improved profitability. We will:

- Engineer meaningful financial features from transaction data.
- Select and train machine learning models (e.g., logistic regression, decision trees, gradient boosting) considering performance and computational efficiency.
- Validate the model using appropriate metrics (e.g., ROC AUC, accuracy) to ensure reliability.

5 Previous Work and Technical Approach

While traditional credit scoring models exist, they might rely on limited financial indicators and may not fully utilize transactional data. This would cause a barrier between those with non-traditional credit histories from getting a loan when they would otherwise be eligible for one. Our approach aims to fill this gap by:

- **Feature Engineering:** Creating detailed features capturing income patterns, spending categories, and their temporal dynamics.
- **Model Training:** Using advanced algorithms suited for tabular data, such as XGBoost or LightGBM, known for their performance in classification tasks.
- **Model Evaluation:** Employing cross validation and hyperparameter tuning to optimize model performance and prevent overfitting.

This methodology aligns with best practices in predictive modeling and is feasible within the project timeline.