Loan defaulter classification system

Mansi Lakhani (40197791), Apekshaba Gohil (40203058), Akshaya Bushan (40220139), Anitha Ramakrishnan (40231724)

I. INTRODUCTION

When financial institutions approve loans, it is important for them to assess the creditworthiness of borrowers to minimize the risk of defaults. The ability to accurately identify potential defaulters can help lenders make informed decisions, manage their loan portfolios effectively, and maintain the stability of their financial operations.

Developing an application to address the problem of predicting loan defaulters comes with its own set of challenges. Some of these challenges include:

1. Data Quality: Obtaining comprehensive and reliable loan data, including borrower information, credit history, financial statements, and repayment behavior, can be challenging.

2. Feature Selection: Identifying the most relevant features that capture the borrower's creditworthiness while avoiding unnecessary or redundant variables is a challenge. Additionally, handling missing data or dealing with categorical variables requires careful preprocessing.

3. Imbalanced Data: Loan default datasets often suffer from class imbalance, where the number of non-defaulters significantly outweighs the number of defaulters. This can affect model performance and result in biased predictions.

Throughout the development of the loan defaulter prediction application, the goals and expectations would be:

1. Accurate Predictions: The primary objective would be to develop a model that accurately predicts loan defaulters, minimizing false positives. This would help lenders make more informed decisions and reduce the risk of financial losses.

2. Risk Mitigation: Financial institutions can proactively manage and mitigate the risk of financial losses associated with unpaid loans.

3. Responsible Lending: Predicting defaulters allows lenders to ensure responsible lending practices by evaluating borrowers' creditworthiness and offering loans to those who are more likely to repay them.

II. DATA SELECTION

The dataset utilized in this project was obtained from Kaggle. The dataset contains information related to loan applications and borrowers' details. It consists of 39,000 samples with 111 features. The features include borrower attributes such as annual income, employment length, loan amount, loan status, credit history, and various financial indicators. Additionally, the dataset provides information on delinquencies, credit utilization, loan descriptions, and other factors that can help predict loan defaulters. The dataset consists of various types of features. It includes numerical features such as loan amount and interest rate, categorical features like home ownership and loan status, text features such as the title and description of the loan, and date and time features such as the last payment date. These different types of features provide diverse information that can be used to analyze borrowers' creditworthiness and make informed decisions regarding loan approvals and risk management.

III. METHODOLOGY

To solve the problem of predicting loan defaulters, several possible methods can be considered. The methods we are using to solve the loan defaulter prediction problem are decision tree and deep neural networks (CNN).

During the preprocessing stage, numerical features undergo scaling using methods like min-max scaling or standardization. This standardizes their values and prevents any single feature from dominating the learning process. Categorical features are transformed through techniques like one-hot encoding or entity embeddings, which represent them as binary vectors or continuous representations, respectively. Text features, such as loan descriptions, are tokenized into words or sub words and then embedded using pre-trained word embeddings like Word2Vec or GloVe. Padding may also be applied to ensure consistent input dimensions across all samples.

By following these preprocessing steps, the data is prepared in a format that allows the deep learning models to effectively learn patterns and relationships. The scaled numerical values, transformed categorical features, and embedded text representations provide the necessary information for the CNN or RNN model to make accurate predictions.

Evaluation metrics for the deep learning pipeline can include accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics provide insights into the model's performance in correctly identifying loan defaulters and non-defaulters.

The obtained results from different models can be compared and analyzed based on their performance metrics. Visualizations like ROC curves or precision-recall curves can aid in understanding the trade-offs between different models.

Analyzing and comparing the performance of the models can offer valuable insights into the efficacy of various methods for predicting loan defaulters. These insights can assist financial institutions in making more informed judgments when evaluating creditworthiness and managing loan portfolios.

REFERENCES

[1] Dataset: "Loan Classification dataset" from Kaggle. https://www.kaggle.com/datasets/abhishek14398/loan-dataset

[2] Gupta, Kanishk and Chakrabarti, Binayak and Ansari, Aseer Ahmad and Rautaray, Siddharth Swarup and Pandey, Manjusha, Loanification - Loan Approval Classification using Machine Learning Algorithms (April 24, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021

[3] Rokach, L., Maimon, O. (2005). Decision Trees. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA.