Credit Risk Modeling Using Machine Learning

Rituj Upadhyay (2020570), Hardik Sachdeva (2022193), Mananya Kohli (2022275), Akshat Tokas (2022056)
IIIT Delhi
Email: {rituj, hardik, mananya, akshat
@iiitd.ac.in

***Abstract*—Credit risk assessment is essential in the financial industry, particularly in today's economic landscape, where institutions are facing increasing non-performing loans (NPLs) and defaults. Traditional credit risk models struggle to capture the complexities of evolving borrower behavior. This project leverages machine learning (ML) techniques to improve credit risk prediction. The goal is to build a flexible model that can better predict defaults, mitigate financial losses, and offer insights to refine lending strategies.**

I. Introduction: Problem Statement

In today's rapidly evolving financial landscape, credit risk assessment is more crucial than ever. With increasing levels of consumer debt and economic uncertainty due to global events like the COVID-19 pandemic, traditional credit risk models are struggling to adapt to new dynamics. Machine learning (ML) offers a solution by providing more flexible and accurate models that can quickly respond to changes in borrower behavior and economic conditions.

A. Current Situation

The financial sector is facing unprecedented challenges due to economic fluctuations, increasing non-performing loans (NPLs), and a rise in defaults. Financial institutions are under pressure to improve their credit risk management practices to mitigate potential losses and maintain profitability.

B. Impact of Machine Learning

ML models can analyze vast amounts of data in real-time, uncovering patterns and correlations that traditional models might miss. By integrating ML, financial institutions can not only predict defaults more accurately but also offer more personalized lending products, reduce the risk of financial crises, and enhance overall financial stability. Additionally, ML can help in reducing biases in credit risk assessments, leading to fairer lending practices and expanding access to credit for underserved populations.

With the COVID-19 pandemic exacerbating financial uncertainty, the limitations of traditional credit risk models—often relying on historical data and static rules—have become more apparent. This project aims to develop a machine learning-based model that improves the prediction of credit risk, offering insights into lending decisions and improving financial outcomes.

II. Literature Survey

A. Predicting Credit Risk for Unsecured Lending

The paper titled ”Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach” investigates the creation of a machine learning (ML)-based credit scoring model aimed at predicting defaults in unsecured lending, specifically credit card lending. Due to the imbalance in the datasets, where the majority of customers do not default, the authors employ SMOTE (Synthetic Minority Oversampling Technique) to balance the data.

After evaluating seven different ML models, the Light Gradient Boosting Machine (LGBM) classifier was determined to be the most effective for handling large datasets and providing greater accuracy. The research underscores that utilizing such models can enhance the prediction of credit defaults, allowing financial institutions to manage credit risk more efficiently.

The study highlights the increasing demand for credit cards, which has led to a rise in defaults, making it necessary to use modern and robust methods for assessing credit risk.

B. Machine Learning-Driven Credit Risk: A Systemic Review

The paper explores the significant impact that machine learning and artificial intelligence have had across various sectors, with a particular focus on finance and credit risk estimation. Machine learning has become crucial for predicting credit risk, a key factor in determining the probability of a debtor defaulting. The accurate estimation of credit risk is vital for maintaining financial stability, as misjudgments can lead to systemic crises like the 2008 subprime mortgage collapse.

Historically, statistical methods such as Linear Discriminant Analysis and Logistic Regression have been used for credit risk estimation, but they struggle with handling large datasets. Advances in computing power and the availability of big data have paved the way for more efficient AI-driven approaches. Modern machine learning techniques, including k-Nearest Neighbor, Random Forest, Support Vector Machines, and particularly deep learning, have proven to be more flexible and accurate than traditional statistical models.

III. Dataset: Data Preprocessing Techniques

A. Different Attributes and Visualizations

The dataset consists of multiple numerical and categorical attributes. Numerical attributes include loan data such as ”Total_TL”, ”Tot_Closed_TL”, and ”NetMonthlyIncome”. Visualizations like histograms, box plots, and correlation matrices were used for analysis. Categorical attributes, such as ”Marital Status” and ”Education”, were visualized using bar charts.

B. Preprocessing Details

Several preprocessing steps were essential before proceeding with the modeling phase:

- **Loading and Cleaning Datasets:** Two datasets were loaded from Excel files. Invalid values represented by -99999 were identified and removed. Columns with a high number of invalid entries (greater than 10,000) were dropped to avoid introducing noise into the analysis.
- **Merging Datasets:** The datasets were merged using an inner join on the *PROSPECTID* column. This ensured that only matching records were retained, effectively eliminating any null values in the final merged dataset.
- **Categorical Feature Analysis:** Chi-Square tests were performed on categorical features such as *MARITALSTATUS*, *EDUCATION*, *GENDER*, *last_prod_enq2*, and *first_prod_enq2*. Only categorical features with a significant relationship to the target variable (*Approved_Flag*) were retained (p-value 0.05).
- **Multicollinearity Check on Numerical Features:** Variance Inflation Factor (VIF) was calculated to detect multicollinearity in numerical features. Features with a VIF ¿ 6 were sequentially removed to avoid redundancy and improve model performance. After filtering, the remaining numerical features were validated using ANOVA to ensure their relevance to the target variable.
- **Encoding Categorical Variables:** Ordinal encoding was applied to the *EDUCATION* feature, following the specified hierarchy:
  - *SSC: 1, 12TH: 2, GRADUATE: 3, UNDERGRADUATE: 3, PROFESSIONAL: 3, POST-GRADUATE: 4, OTHERS: 1.*
  - Business input may be needed to verify the grouping of “OTHERS.”

  One-hot encoding was applied to other categorical features such as *MARITALSTATUS*, *GENDER*, *last_prod_enq2*, and *first_prod_enq2*.
- **Feature Selection:** Statistical techniques such as Chi-Square tests were used to select the most relevant categorical features, and VIF was used for numerical features.
- **Outlier Detection and Removal:** Outliers were handled using the IQR method or other statistical techniques to prevent skewed predictions.
- **Final Dataset Preparation:** The cleaned and transformed dataset included both encoded categorical and validated numerical features. This version of the dataset ensures optimal feature representation for future model experimentation and training.

This structured process guarantees a robust feature set with minimal multicollinearity and maximum relevance to the target variable, setting a solid foundation for model building.

C. Exploratory Data Analysis (EDA)

Exploratory data analysis was performed to understand the structure of the data, detect missing values, and identify relationships between features.It refers to the process of analyzing and visualizing data to grasp its main features, uncover trends, detect outliers, and recognize relationships between variables. This method involves exploring datasets to understand their key characteristics and identify patterns.

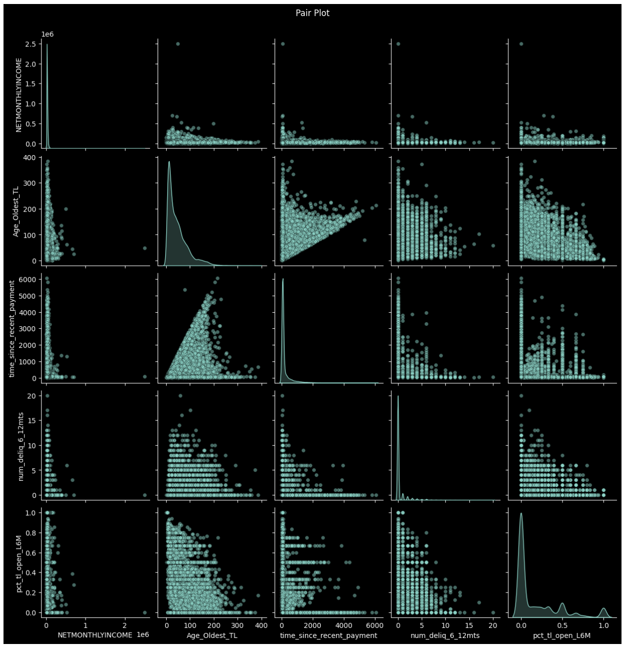

Fig. 1. Pair Plot

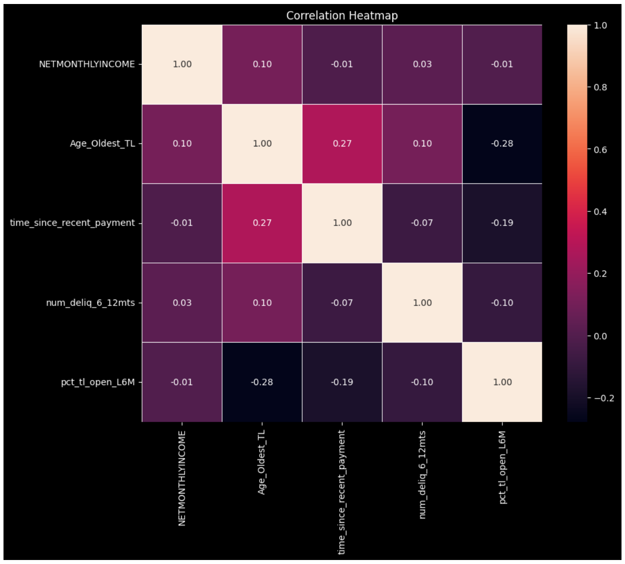

Fig. 2. Correlation Heatmap

IV. Methodology and Model Details

The goal of this phase is to evaluate the feasibility of various machine learning models for predicting the *Approved_Flag*.

The models considered in this analysis are:

A. Random Forest

The Random Forest algorithm is an ensemble machine learning technique used for both classification and regression tasks. It builds multiple decision trees during training, where each tree is trained on random subsets of the data and features. The randomness introduced by using different subsets reduces overfitting and increases model robustness. For classification, it aggregates the results by voting across trees, and for regression, it averages the predictions. This collective decision-making from multiple trees makes Random Forest highly accurate, capable of handling complex data, and resilient to noise in the dataset. Random Forest is considered because it can handle imbalanced datasets, reduces overfitting through bagging, and provides feature importance.

B. XGBoost

XGBoost, short for ”Extreme Gradient Boosting,” is a highly efficient and scalable machine learning algorithm designed for both classification and regression tasks. It is an ensemble method that combines multiple weak models to produce a more accurate and robust prediction. XGBoost stands out for its ability to handle large datasets and missing values efficiently, without requiring extensive pre-processing. With built-in support for parallel processing, XGBoost accelerates training on large datasets, making it a popular choice for achieving state-of-the-art performance in various machine learning tasks. XGBoost is explored due to its high performance with tabular data and ability to handle multi-class classification with fine-tuning options.

C. Decision Tree

A decision tree is a supervised learning algorithm used in machine learning for both classification and regression tasks. It models decisions through a tree-like structure, where each internal node represents a test on an attribute, each branch corresponds to the outcome of that test, and each leaf node signifies a final decision or prediction. Decision trees split data into subsets based on feature values, making decisions at each step until a final outcome is reached. Known for being intuitive and interpretable, decision trees are widely applied in various domains to model complex decision-making processes. The Decision Tree classifier is examined as a simpler model, offering easy interpretation but with potential risks of overfitting.

D. Scaling and Further Considerations

Additionally, the use of standard scaling for numerical features is considered to ensure consistent performance, especially for distance-based algorithms that may be explored in future steps. This exploratory analysis allows us to align the model selection with the data structure, understand potential challenges, and determine which models are worth further experimentation based on preliminary metrics such as accuracy, precision, and recall.

V. Results and Analysis

The EDA indicated strong predictive power in certain features but revealed imbalances in the dataset. The analysis reveals several imbalances and patterns in the dataset. The *Approved_Flag* is skewed toward P2, suggesting a higher likelihood of this category being approved compared to P1, P3, and P4. In terms of demographics, *MARITALSTATUS_Married* and *GENDER_F* show significant imbalances, with 73.4% of individuals being married and only 11.2% being female, indicating the dataset is predominantly male and married. Loan ownership data, especially for home and personal loans, show that most individuals don't have these loans, as reflected in *HL_Flag* and *PL_Flag*. The *NETMONTHLYINCOME* distribution is heavily skewed, pointing to a middle-income segment. The *Age_Oldest_TL* suggests most individuals have a relatively young credit history, while delinquency-related variables such as *recent_level_of_deliq* indicate low recent delinquency. Overall, the dataset shows notable imbalances across different features, which could impact predictive modeling.

VI. Conclusion

A. Learning from the Project

The preprocessing steps, such as handling missing values and balancing data, significantly improved model performance. Feature selection techniques such as ANOVA and Chi-square tests helped identify the most impactful features, boosting prediction accuracy. It underscores the importance of data preprocessing, such as addressing data imbalances and selecting relevant features. Various algorithms, including Random Forest, XGBoost, and Decision Trees, are evaluated for their effectiveness in managing complex datasets, ultimately aiming to improve lending strategies and financial decision-making in a changing economic environment.

B. Work Left

The upcoming steps for the project focus on training a range of machine learning models, such as various Regression, Random Forest, and various Boosting, to determine which model performs best in predicting credit risk. Once identified, the best-performing model will be validated on a separate test dataset and subsequently deployed for real-time predictions. In addition, efforts will be made to optimize hyperparameters for enhanced performance and to tackle data imbalances. A comprehensive analysis will follow to interpret the model results and derive meaningful insights for credit risk assessment.

C. Contribution of Each Team Member

- **Mananya Kohli & Akshat Tokas:** Handled data collection and preprocessing.
- **Hardik Sachdeva & Rituj Upadhyay:** Led the exploratory data analysis and feature engineering.
- **Akshat Tokas & Rituj Upadhyay:** Focused on model development and evaluation.
- **Mananya Kohli & Hardik Sachdeva:** Will handle model interpretation and deployment.

References

1) https://arxiv.org/abs/2110.02206 Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach
2) https://link.springer.com/article/10.1007/s00521-022-07472-2 Machine Learning-Driven Credit Risk: A Systematic Review