Machine Learning Models in Credit Risk Management: A Comparative Analysis for Loan Default Prediction

Group 12

Chamika
dept. of CSE University of Moratuwa
Colombo, Sri Lanka
chamika.21@cse.mrt.ac.lk

Fathima Hana
dept. of CSE University of Moratuwa
Colombo, Sri Lanka
fathimahana.21@cse.mrt.ac.lk

Subavarshana
dept. of CSE University of Moratuwa
Colombo, Sri Lanka
subavarshanaa.21@cse.mrt.ac.lk

Hanaanee Hana
dept. of CSE University of Moratuwa
Colombo, Sri Lanka
hanaanee.21@cse.mrt.ac.lk

Mamta
dept. of CSE University of Moratuwa
Colombo, Sri Lanka
nallaretnam.21@cse.mrt.ac.lk

***Abstract*—Loan defaults pose a significant financial risk for lending institutions. This study investigates the effectiveness of machine learning models in predicting loan defaults, a crucial aspect of credit risk management.**

We evaluated three machine learning algorithms (CatBoost, LightGBM, and XGBoost) and a voting ensemble model for their ability to classify loan applicants as potential defaulters or non-defaulters. The investigation revealed promising results, with LightGBM achieving a maximum Area Under the ROC Curve (AUC) score during cross-validation. The voting ensemble model also demonstrated comparable performance on a validation fold.

These findings highlight the potential of machine learning to improve loan risk assessment and support informed lending decisions. Additionally, our study emphasizes the growing importance of data science and machine learning in the financial and banking sector for effective risk management and responsible lending practices. Limitations include the need for further evaluation on a separate test set and exploring interpretability techniques for complex models. Future research directions involve incorporating additional data sources, exploring alternative ensemble methods, and addressing potential biases to ensure fair and ethical model application. This research contributes to the advancement of machine learning applications in loan default prediction and responsible financial management.

Keywords—Loan Default Prediction, Machine Learning, Credit Risk Management, CatBoost, LightGBM, XGBoost, Ensemble Learning, Financial Services

I. Introduction and Background

The financial sector hinges on responsible lending practices that balance loan issuance with effective risk management. Loan defaults, where borrowers fail to repay their loans, pose a significant financial threat for lending institutions. These defaults can lead to substantial losses, impacting an institution's stability and overall profitability.

Traditionally, financial institutions employ various credit risk management strategies to mitigate these risks. These strategies involve analyzing borrower financials, credit history, and other relevant data points to assess creditworthiness and determine loan eligibility. However, the ever-increasing volume of financial data available is transforming the landscape of credit risk management.

The emergence of machine learning (ML) presents a powerful tool for financial institutions to bolster their credit risk assessment capabilities. ML models can analyze vast datasets, encompassing traditional credit history information and alternative data sources. By doing so, they can identify complex patterns and relationships that might be missed by conventional methods. These patterns can then be leveraged to predict loan defaults with greater accuracy, enabling lenders to make more informed decisions.

This research investigates the potential of machine learning models in predicting loan defaults. We aim to achieve the following objectives:

- Evaluate the performance of different machine learning models
- Explore the benefits of ensemble learning
- Highlight the significance of data science in finance

II. Related Work

01. Corporate credit risk modeling and the macroeconomy. This paper investigates the influence of macroeconomic factors on corporate credit risk modeling. The authors propose a duration model that incorporates both firm-specific variables and macroeconomic indicators to predict business loan defaults. Their findings show that macroeconomic variables significantly improve the model's ability to explain default risk compared to models using only firm-specific data. Additionally, the study highlights the importance of duration dependence in default risk, suggesting that both default probability and survival time are crucial for accurate risk assessment.

02. This paper examines the application of credit risk models in measuring creditworthiness, particularly

for financial institutions. The authors argue that traditional credit ratings may not be sufficient for accurate risk assessment. They propose using the "distance-to-default" (DTD) value derived from equity market data and volatility as a more precise measure. The study investigates the effectiveness of the Merton model and KMV's EDF model in calculating DTD values for South Korean banks and brokerage firms. The findings suggest that these models can effectively reflect the credit quality of financial institutions. Furthermore, the paper proposes incorporating DTD values alongside credit ratings to determine collateral requirements in over-the-counter (OTC) derivatives transactions. This approach, according to the authors, can lead to a more efficient allocation of resources by minimizing collateral amounts needed without compromising risk mitigation.

03. Research on Credit Risk Evaluation of Commercial Banks Based on Artificial Neural Network Model. This research examines how Artificial Neural Networks (ANNs) can be used to evaluate credit risk for commercial banks. Credit risk is a significant concern for banks, as it refers to the possibility of borrowers defaulting on loans. The authors propose an ANN-based approach to predict the creditworthiness of corporate clients. Their method involves building a credit risk evaluation system using 14 financial indicators. Then, they leverage cluster analysis and factor analysis to assign actual credit ratings to sample data, providing the target variable for the ANN model. Finally, they compare the performance of two traditional credit risk prediction models with three common ANN models. By choosing the best performing model, they aim to achieve accurate credit risk prediction and evaluation for commercial bank clients.

04. A dynamic credit risk assessment model with data mining techniques. This study criticizes static credit risk models in banks, which struggle to adapt to economic shifts. The authors propose a dynamic model using ANFIS, a type of artificial intelligence. This model considers factors influenced by political and economic crises, removing human bias from evaluation. By analyzing historical data and updating monthly based on bad customer lists, the model aims to adapt to changing environments and outperform traditional static models, especially during economic downturns.

05. The purpose of this research is to present a hybrid method for evaluating credit risk of bank customers. Here, after extracting significant financial ratios from balance sheet, Kolmogorove-Smirnov test has been used to specify the kind of financial ratios distribution. Then, a T test has been run to select meaningful variables and DEMATEL method to determine effective ones. Finally, a Fuzzy Expert system has been developed to assess credit risk according to specified effective financial ratios as the system inputs. The presented steps have been studied in an Iranian Bank as empirical study.

III. METHOD

A. Data Collection

Our study utilized data provided by Home Credit, a consumer finance provider specializing in loans for borrowers with limited credit history. The data consisted of anonymized information on loan applicants, including internal and external sources. Internal data encompassed applicant demographics, loan characteristics, and historical interactions with Home Credit. External data sources included credit bureau reports, tax registry information, and alternative data providers. The data was organized into tables with varying depths, reflecting the level of historical information available for each applicant. Depth-0 tables contained static features directly linked to the applicant, while depth-1 and depth-2 tables included historical data indexed by designated groups. Feature definitions provided by Home Credit aided in understanding the transformations applied to specific features.

B. Data preprocessing

The raw data underwent several preprocessing steps to ensure its suitability for modeling. First, data types were assigned based on column naming conventions (e.g., dates, integers). Missing values were addressed using a combination of techniques: features with more than 70% missingness were removed, and for features with informative missing patterns (groups sharing missing values), a representative feature with the most information was chosen within each group. Feature engineering involved creating new features like month and day of decision from the loan decision date. Next, historical data from depth tables (depth 1 and 2) was aggregated by case_id and joined with the base table. To handle dates within these tables, the difference between each date and the decision date was calculated and converted to total days. Finally, potential high cardinality features (string features with many unique values) and constant features (string features with only one unique value) were removed. Additionally, data type conversion to reduce memory usage was applied.

C. Models

1) CatBoost Classifier: CatBoost is a gradient boosting decision tree known for its speed, accuracy, and ability to handle various data types, including categorical features. In this experiment, three CatBoost models were trained using

cross-validation, focusing on maximizing the Area Under the ROC Curve (AUC) for the task of predicting loan default.

2) Light Gradient Boosting Machine (LightGBM): LGBM is another powerful gradient boosting framework known for its efficiency and scalability. Three LightGBM models were trained with similar objectives as the CatBoost models using cross-validation. Early stopping was employed to prevent overfitting.

3) XGBoost: XGBoost is also another gradient boosting technique known for its flexibility and performance. Three XGBoost models were trained with early stopping to optimize for AUC on the loan default prediction task using cross-validation.

4) Ensemble: A voting ensemble model was created to combine the predictions from the nine first-layer models (three CatBoost, three LightGBM, and three XGBoost) trained using cross-validation. The voting ensemble model takes the average prediction from each first-layer model for a given data point.

D. Evaluation Metrics

In this experiment, the primary evaluation metric for model performance was the Area Under the ROC Curve (AUC). A perfect model achieves an AUC of 1, while a random guesser has an AUC of 0.5. AUC was calculated for each fold of cross-validation for all first-layer models (CatBoost, LightGBM, and XGBoost). And then the final performance of the voting ensemble model was evaluated using AUC on a separate hold-out test set (not used for training).

IV. RESULTS

The table below summarizes the AUC scores achieved by the individual first-layer models (CatBoost, LightGBM, and XGBoost) during cross-validation. It also includes the final AUC score obtained by the voting ensemble model on a validation fold.

TABLE 1

MODEL PERFORMANCE SCORES

	Model			
	CatBoost	*LightGBM*	*XGBoost*	*Ensemble*
AUC Score	0.7535	0.7632	0.7454	0.7633

A. OBSERVATIONS

- LightGBM achieved the highest maximum CV AUC score (0.7632) among the individual models, indicating strong performance in predicting loan defaults
- The voting ensemble model obtained a comparable AUC score (0.7633) on a validation fold during cross-validation, suggesting its potential effectiveness.

V. DISCUSSION

In this section we will discuss the key findings of the experiment, explore their implications for the credit risk model, and identify potential avenues for further improvement.

A. Discussion and Interpretation of Results

Our investigation focused on evaluating the effectiveness of three machine learning models (CatBoost, LightGBM, and XGBoost) and a voting ensemble model for predicting loan defaults. LightGBM emerged as the strongest individual model, achieving a maximum AUC score of 0.7632 during cross-validation. This suggests that LightGBM effectively learned patterns in the training data to distinguish between loan defaulters and non-defaulters.

The voting ensemble model, which leverages the combined strengths of the individual models, achieved a comparable AUC score of 0.7633 on a validation fold. This indicates that ensemble learning holds promise for improving loan default prediction by mitigating potential biases inherent in any single model.

B. Potential Improvements for the Credit Risk Model

This experiment lays a foundation for further refinement of the credit risk model. Here are some key areas which could be improved.

1) Incorporating Additional Data: Enriching the data with additional features like borrower demographics, credit bureau scores, and alternative data sources (e.g., social media activity) could potentially improve the model's ability to capture diverse risk factors and enhance prediction accuracy.

2) External Validation: While cross-validation provides a robust internal evaluation, external validation on a completely independent dataset from a different time period or institution strengthens the model's generalizability and ensures its effectiveness in real-world scenarios.

3) Ensemble Learning Exploration: The current ensemble used by us is a voting approach. Exploring alternative ensemble methods like stacking or blending, where the predictions from different models are used as features to train a final meta-model, could potentially lead to further improvements in prediction accuracy.

4) Model Calibration: While AUC is a valuable metric, it doesn't guarantee well-calibrated probabilities. Calibration techniques like Platt scaling can be employed to ensure the predicted probabilities accurately reflect the true risk of default. This is crucial for setting appropriate loan interest rates and making informed lending decisions.

5) Interpretability Improvements and Ethical Considerations : For complex models like LightGBM and XGBoost etc which we used in this experiment, interpreting feature interactions and the overall model behavior can be challenging.

   While machine learning offers advantages, complex models like LightGBM and XGBoost can be challenging to interpret. This lack of transparency raises ethical concerns, as it's difficult to understand how these models arrive at credit risk decisions.

   To address this, exploring techniques like SHAP and LIME can provide insights into the models' reasoning. Implementing such methods fosters trust and transparency in the credit risk assessment process, ensuring fair and responsible application of these powerful models.

   - Enterprise Shapley Additive exPlanations (SHAP): SHAP assigns credit to individual features for a particular prediction, offering a more nuanced understanding of feature interactions and the model's reasoning.

   - LIME (Local Interpretable Model-Agnostic Explanations): LIME locally approximates the complex model around a specific prediction, providing human-interpretable explanations for individual loan default classifications.

VI. CONCLUSION

This study explored the potential of machine learning models for predicting loan defaults, a critical aspect of credit risk management in the financial sector. We investigated the performance of three machine learning models (CatBoost, LightGBM, and XGBoost) and a voting ensemble model.

The investigation revealed promising results. LightGBM emerged as the strongest individual model, achieving a maximum AUC score of 0.7632 during cross-validation. The voting ensemble model also demonstrated comparable performance, obtaining an AUC score of 0.7633 on a validation fold. These findings highlight the effectiveness of machine learning in identifying loan risk and supporting informed lending decisions.

This study underscores the growing significance of data science and machine learning in the financial and banking industry. By leveraging these technologies, financial institutions can gain valuable insights from vast datasets to enhance credit risk assessment, promote responsible lending practices, and mitigate financial risks.

However, it's important to acknowledge the limitations of this study. The final evaluation of the ensemble model's generalizability requires testing on a separate hold-out test set not used for training. Additionally, exploring interpretability techniques for complex models like LightGBM and XGBoost is crucial for building trust and transparency in their decision-making processes.

Future research directions include incorporating additional data sources, exploring alternative ensemble learning methods, etc. By continuously refining and improving machine learning models for loan default prediction, financial institutions can make significant strides towards more robust risk management and responsible lending practices.

References

[1] Kenneth Carling, Tor Jacobson, Jesper Lindé, Kasper Roszbach. Corporate credit risk modeling and the macroeconomy
[2] Joocheol Kim, Eunhwan Kim . Journal of Economic and Financial Studies
[3] Yong Hu, Jie Su. Research on Credit Risk Evaluation of Commercial Banks Based on Artificial Neural Network Model
[4] Saba Moradi & Farimah Mokhatab Rafiei. A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks
[5] Sanaz Pourdarab, Ahmad Nadali and Hamid Eslami Nosratabadi. A Hybrid Method for Credit Risk Assessment of Bank Customers