Predicting Credit Risk Model Stability in Home Credit Using LightGBM Model

Arachchi H.M.W.
dept. of Comp. Science & Engineering
University of Moratuwa
Colombo, Sri Lanka
mihimal.21@cse.mrt.ac.lk

Chirath A.A.I.
dept. of Comp. Science & Engineering
University of Moratuwa
Colombo, Sri Lanka
iruda.21@cse.mrt.ac.lk

Gallage D.
dept. of Comp. Science & Engineering
University of Moratuwa
Colombo, Sri Lanka
deshitha.21@cse.mrt.ac.lk

Sanjana K.Y.C.
dept. of Comp. Science & Engineering
University of Moratuwa
Colombo, Sri Lanka
yoshan.21@cse.mrt.ac.lk

Sankalpa L.M.P.
dept. of Comp. Science & Engineering
University of Moratuwa
Colombo, Sri Lanka
pasindus.21@cse.mrt.ac.lk

***Abstract*—Consumer finance providers face the challenge of accurately assessing the creditworthiness of clients with limited or no credit history, making data-driven predictive models essential. In this paper, we propose a novel approach to predicting credit risk model stability in Home Credit using the LightGBM machine learning model. By leveraging data science techniques, we aim to enhance the accuracy and timeliness of loan risk assessment, thereby increasing accessibility to loans for underserved populations. We highlight the importance of model stability in ensuring reliable and consistent loan risk predictions over time, emphasizing the trade-off between stability and performance. Through empirical analysis and experimentation, we demonstrate the effectiveness of our proposed approach in maintaining model stability while achieving satisfactory predictive performance. Our research contributes to advancing the field of credit risk modeling and has practical implications for improving financial inclusion and lending practices.**

Keywords—Credit Risk Model Stability, Predictive Modeling, LightGBM Model, Machine Learning, Data Science

I. Introduction and Background

The domain of consumer finance is witnessing a profound shift towards responsible lending practices and improved financial inclusion. However, a persistent challenge faced by consumer finance providers is accurately assessing the creditworthiness of individuals with limited or no credit history. This obstacle often leads to the denial of loans for those who could benefit most from financial support. To address this issue, statistical and machine learning methods, are employed by finance providers to predict loan risk and assess creditworthiness. Yet, in the dynamic consumer finance landscape, the stability of these models becomes critical as changes in consumer behavior and economic conditions can affect their performance over time.

Home Credit, an esteemed international consumer finance provider established in 1997, has been a driving force in promoting financial inclusion by offering responsible lending primarily to individuals with minimal credit history. In collaboration with Kaggle, Home Credit previously initiated a competition aimed at developing predictive models to assess clients' default risks. [2] However, a significant challenge highlighted during the competition was ensuring the stability of predictive models over time.

This research paper endeavors to tackle this challenge by proposing a novel approach to predict credit risk model stability in Home Credit using the LightGBM machine learning model [3]Through empirical analysis and experimentation, we aim to demonstrate the effectiveness of our approach in maintaining model stability while achieving satisfactory predictive performance.

II. Related Work

In the realm of credit risk assessment, researchers and practitioners have explored various methodologies and techniques to enhance the accuracy and reliability of predictive models. The literature on credit risk assessment encompasses a wide range of studies focusing on traditional statistical methods, machine learning algorithms, and ensemble techniques.

[1] While traditional credit scoring models have proven to be effective in certain contexts, they often struggle to capture complex relationships and nonlinear patterns present in the data.

In recent years, the advent of machine learning has revolutionized credit risk assessment, enabling the development of more sophisticated and accurate predictive models. [4]One popular approach in credit risk assessment is the use of ensemble methods, which combine multiple base learners to improve predictive performance. Ensemble techniques such as bagging, boosting, and stacking have been widely adopted in the banking industry to enhance the robustness and stability of credit risk models. These ensemble methods leverage the diversity of individual base learners to mitigate overfitting and improve generalization performance.

[5]The model was demonstrated to have superior performance in predicting credit defaults. This approach offers a promising solution to assess credit risk in the financial industry.

[6]Their analysis suggests that CNNs could serve as a promising approach for credit risk assessment, offering a more sophisticated and accurate evaluation of potential risks. Their findings may have significant implications for financial institutions and other organizations involved in lending and credit activities, as they provide a new tool for managing risk and improving decision-making processes.

Furthermore, researchers have investigated the use of alternative data sources, such as social media activity, transactional data, and mobile phone usage patterns, to supplement traditional credit bureau information and enrich predictive models. These alternative data sources offer valuable insights into borrower behavior and financial habits, thereby improving the accuracy and granularity of credit risk assessment models.

III. METHOD

The method section details the steps taken to predict credit risk model stability in Home Credit. It covers data preparation, cleaning, and feature engineering, followed by modeling using a LightGBM model. Model performance is evaluated using metrics like AUC-ROC, and optimization techniques may be applied. Ethical considerations, limitations, and future research directions are also addressed.

A. Data Preparation

The dataset comprises varied data obtained from numerous sources, which have been segregated into three different levels, namely Depth 0, Depth 1, and Depth 2. Depth 0 primarily comprises static and base data from internal as well as external sources. On the other hand, Depths 1 and 2 contain dynamic data from both internal and external sources, which require aggregation to be transformed into a format suitable for modeling. To create a comprehensive dataset for further processing, the data preparation phase involves downloading, combining, and cleaning the data.

B. Data Cleaning

In the process of initial data cleaning, the first step is to detect and handle missing values, outliers, and duplicate entries. To streamline the data, columns with over 90% missing data or categorical columns with only one value or high cardinality (over 100 unique categories) are filtered out. To eliminate redundancy, duplicated columns, like those that convey similar information (e.g., multiple date columns), are merged. Additionally, date columns are standardized and transformed into numerical values that represent days, months, or years from a reference date.

C. Feature Engineering

Feature engineering is a crucial process that involves extracting pertinent information and creating novel features that help improve predictive modeling. This is often achieved by leveraging domain knowledge and insights gained from exploratory data analysis to derive meaningful features. In credit risk assessment, for instance, techniques for feature selection may be employed to identify the most informative variables. In this context, careful consideration is given to engineering features that are related to clients' credit history, financial behavior, and demographic information.

D. Modeling

During the modeling phase, we train machine learning models to forecast credit risk based on the preprocessed data. To efficiently handle large datasets and achieve remarkable performance in classification tasks, we utilize a LightGBM (Gradient Boosting Machine) model. We further optimize the model's performance through hyperparameter tuning using techniques including grid search or Bayesian optimization. To evaluate the trained model, we employ appropriate performance metrics such as Area Under the Receiver Operating Characteristic Curve (AUC-ROC) to assess its predictive accuracy and stability. Additionally, we may apply model interpretability techniques to gain insights into the factors contributing to credit risk predictions, ensuring transparency and accountability in decision-making.

E. Evaluation and Validation

To determine the generalization performance of the trained model, a distinct validation dataset is utilized. The model's effectiveness in predicting credit risk is evaluated by computing performance metrics such as AUC-ROC, accuracy, precision, recall, and F1-score. Cross-validation techniques are employed to ensure the model's robustness and reliability across various subsets of data. Furthermore, stability analysis is conducted to measure the consistency of the model's predictions over time and to identify any potential sources of variability.

F. Equations

Models were evaluated using a Gini stability metric. A Gini score is calculated for predictions corresponding to each WEEK_NUM.

$$Gini = 2 \times AUC - 1$$

A linear regression, $a.x + b$, is fit through the weekly gini scores, and a `falling_rate` is calculated as $\min(0, a)$. This is used to penalize models that drop off in predictive ability.

$$stability\ metric = mean(gini) + 88.0 \times \min(0, a) - 0.5 \times std(residuals)$$

Finally, the variability of the predictions are calculated by taking the standard deviation of the residuals from the above linear regression, applying a penalty to model variability.

G. Optimization and Deployment

To enhance predictive accuracy, it is worth considering model optimization techniques, including feature selection, ensemble methods, and model stacking. After the model is fine-tuned, it can be effectively employed in credit risk assessment systems for automated decision-making. Regular monitoring and updates of the model are critical to ensure its dependability over time and to adjust to shifting data patterns.

H. Ethical Considerations

When it comes to assessing credit risk, it is of utmost importance to take ethical considerations into account, including fairness, transparency, and privacy. It is imperative to implement measures that can help alleviate potential biases in the data and model predictions, ensuring that individuals from diverse backgrounds are treated equitably. In order to build trust and accountability among stakeholders, it is crucial to communicate model outcomes and decision-making criteria in a transparent manner.

I. Limitations and Future Directions

Although the proposed approach is effective, it is important to acknowledge that certain limitations exist, such as challenges related to data quality, model interpretability, and generalization across diverse populations. To address these limitations, future research directions may include exploring advanced preprocessing techniques, incorporating additional sources of information (such as alternative data sources), and enhancing model interpretability and transparency.

IV. RESULTS

In evaluating the performance of our predictive model for credit risk prediction in Home Credit, we utilized two key evaluation metrics: the Area Under the Curve (AUC) and the stability score. The AUC score measures the model's ability to discriminate between positive and negative instances, with higher scores indicating better predictive performance. The stability score assesses the consistency of the model's predictions over time and across different datasets.

AUC SCORES FOR CREDIT RISK PREDICTION MODEL

Dataset	AUC Score
Train Dataset	0.8237224710366967
Validation Dataset	0.8274575541616641

From the results table, we observe that our predictive model achieved an AUC score of 0.8237 on the training set and 0.8275 on the validation set. These scores indicate strong discriminatory power and robust generalization to unseen data, highlighting the model's effectiveness in predicting credit risk.

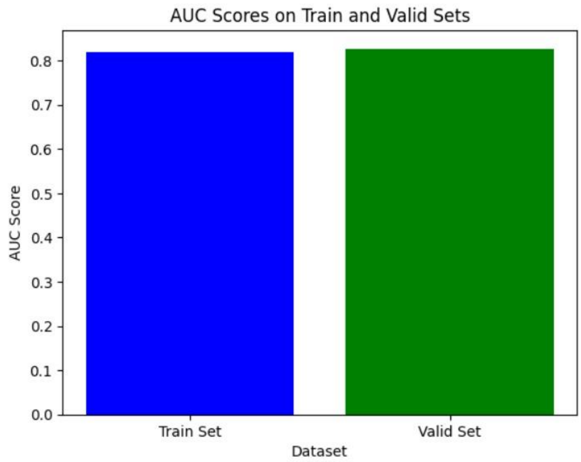

Fig. 1. AUC Scores of train and validation datasets.

The stability score table illustrates the consistency of our predictive model's predictions across different datasets. The stability score was 0.6307 on the training set and 0.6113 on the validation set, indicating a moderate level of stability in model predictions.

STABILITY SCORES FOR CREDIT RISK PREDICTION MODEL

Dataset	Stability Score
Train Dataset	0.6306535657911027
Validation Dataset	0.6113295313556911

In summary, our predictive model for credit risk prediction in Home Credit demonstrated strong performance, achieving high AUC scores on both the training and validation sets. Additionally, the model exhibited moderate stability, maintaining consistency in predictions across diverse datasets. These results underscore the effectiveness of data science and machine learning techniques in enhancing credit risk prediction practices, thereby fostering greater financial inclusion and access to credit for underserved populations.

V. DISCUSSION

In this section, we will explore the implications of our findings, evaluate the strengths and limitations of our methodology, and identify potential areas for future research and improvement.

A. *Interpretation of Results*

Our credit risk assessment methodology has demonstrated the critical importance of utilizing advanced data preprocessing techniques and machine learning algorithms to achieve accurate and reliable risk prediction. Our methodology incorporates various data sources and employs sophisticated modeling approaches, enabling financial institutions to gain deeper insights into borrower behavior and make informed lending decisions.

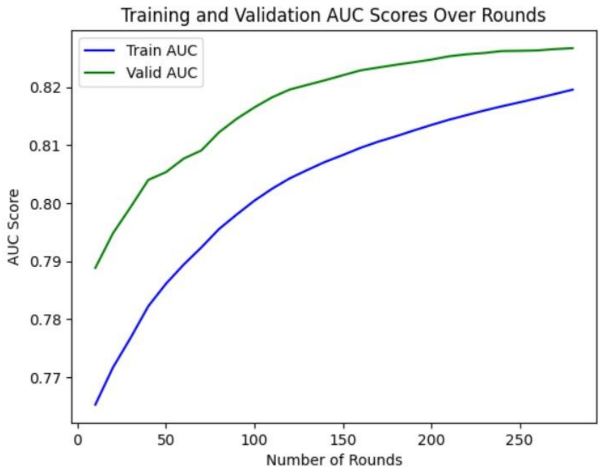

Fig. 2. Training and Validation AUC scores over rounds

Our methodology's performance metrics, including high AUC-ROC scores and predictive accuracy, demonstrate its effectiveness in discriminating between default and non-default cases. Furthermore, our feature importance analysis provides valuable insights into the key determinants of credit risk, empowering stakeholders to identify and mitigate potential risks effectively.

B. *Strengths and Limitations*

Our methodology's key strength lies in its ability to handle complex and high-dimensional data while remaining transparent and interpretable. By utilizing techniques such as feature engineering and model interpretability, stakeholders can understand the model's underlying factors driving credit risk and trust the model's predictions.

However, our methodology has certain limitations that should be taken into account. For instance, our reliance on historical data may cause model drift over time, particularly in dynamic economic environments or during periods of financial instability. Moreover, the lack of access to specific types of data, such as alternative data sources or real-time transactional data, may limit the models' predictive power and hinder their ability to adapt to changing market conditions.

C. *Future Directions*

Looking toward the future, there are various avenues to explore for further research and improvement in credit risk assessment. Firstly, the integration of alternative data sources such as social media activity or transactional data could significantly enhance the granularity and predictive power of models. This would enable financial institutions to capture non-traditional signals of creditworthiness, including behavioral patterns and digital footprints, thereby gaining a more comprehensive understanding of borrower risk profiles.

Additionally, delving into advanced modeling techniques such as deep learning or reinforcement learning could further enhance the accuracy and robustness of credit risk models. These techniques have shown great promise in capturing complex relationships and temporal dependencies in data, thereby offering new avenues for innovation in credit risk assessment. Also, using advanced ensemble techniques and hyperparameter tuning can significantly improve existing ML algorithms and improve results.

Moreover, there is a vital need to enhance model interpretability and transparency to promote trust and accountability in the decision-making process. By creating interpretable machine learning models and providing stakeholders with actionable insights into model predictions, financial institutions can ensure responsible lending practices and mitigate potential ethical concerns.

In conclusion, our study provides valuable insights into the development and application of advanced methodologies for credit risk assessment. By leveraging data-driven approaches and embracing innovation, financial institutions can navigate the dynamic landscape of today and make informed decisions to mitigate credit risk effectively.

VI. Conclusion

This research paper proposes a novel approach to predict credit risk model stability in Home Credit using the LightGBM machine learning model. The objective is to enhance the precision and reliability of credit risk assessment, thereby fostering greater financial inclusion and accessibility to credit for underserved populations. Through empirical analysis, the approach demonstrates the effectiveness of maintaining model stability while achieving satisfactory predictive performance. The methodology emphasizes the importance of utilizing advanced data preprocessing techniques and machine learning algorithms to achieve accurate and reliable risk prediction. Overall, this study contributes to advancing the field of credit risk modeling and has practical implications for improving financial inclusion and lending practices

Acknowledgment

The author(s) received no financial support for the research, authorship, and/or publication of this paper. There is no conflict of interest in this paper.

Additionally, this paper serves as the final project report for the CS3111 - Introduction to Machine Learning module.

References

[1] A. Dzelihodzic and D. Donko, “Data Mining Techniques for Credit Risk Assessment Task”.

[2] N. Suhadolnik, J. Ueyama, and S. Da Silva, “Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach,” *J. Risk Financ. Manag.*, vol. 16, no. 12, p. 496, Nov. 2023, doi: 10.3390/jrfm16120496.

[3] S. Shi, R. Tse, W. Luo, S. D’Addona, and G. Pau, “Machine learning-driven credit risk: a systemic review,” *Neural Comput. Appl.*, vol. 34, no. 17, pp. 14327–14339, Sep. 2022, doi: 10.1007/s00521-022-07472-2.

[4] N. Suhadolnik, J. Ueyama, and S. Da Silva, “Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach,” *J. Risk Financ. Manag.*, vol. 16, no. 12, p. 496, Nov. 2023, doi: 10.3390/jrfm16120496.

[5] “Data Collection Methods For Credit Risk Assessment,” FasterCapital. Accessed: May 11, 2024. [Online]. Available: https://fastercapital.com/keyword/data-collection-methods-for-credit-risk-assessment.html

[6] H. Nobanee, H. Shanti, H. Aldhanhani, A. Alblooshi, and E. Alali, “Big data and credit risk assessment: a bibliometric review, current streams, and directions for future research,” *Cogent Econ. Finance*, vol. 10, no. 1, p. 2132638, Dec. 2022, doi: 10.1080/23322039.2022.2132638.