# Home Loan Credit Risk Predictor

Mona Teja Kurakula
*Thomas Lord Department of Computer Science*
*University of Southern California*
Los Angeles, United States of America
mkurakul@usc.edu

Mohan Sai Ganesh Kanna
*Thomas Lord Department of Computer Science*
*University of Southern California*
Los Angeles, United States of America
mkanna@usc.edu

Akhilaa Sonduri Panthangi
*Thomas Lord Department of Computer Science*
*University of Southern California*
Los Angeles, United States of America
sondurip@usc.edu

Sahithi Namala
*Thomas Lord Department of Computer Science*
*University of Southern California*
Los Angeles, United States of America
namala@usc.edu

Ke-Thia Yao
*Faculty at Thomas Lord Department of Computer Science*
*University of Southern California*
Los Angeles, United States of America
kyao@usc.edu

*Abstract* — In an era marked by exponential population growth and the ubiquitous necessity of loans, the reliance oncredit history as a measure of trust has become deeply entrenched. However, this reliance presents a significant challenge when individuals without credit history seek loans.This research paper delves into the complexities of this issue, elucidating how financial institutions can navigate such scenarios to expand their business and network while minimizing risk and maximizing returns.

This paper underscores the imperative for financial institutionsto develop predictive machine learning models that can accurately assess the risk and likelihood of loan repayment by individuals lacking credit history. It explores the utilization of key data points about individuals and advocates for the implementation of efficient methodologies to forecast loan repayment risk, even in the absence of conventional credit history metrics. Through a comprehensive analysis of existing challenges and emerging solutions, this paper offers insights into the evolving landscape of creditworthiness assessment. It emphasizes the importance of leveraging alternative data sources and innovative machine-learning modelling techniques to bridge the trust gap and enable inclusive lending practices.

This research contributes to the advancement of more equitable and effective lending practices. It serves as a valuable resource for policymakers, financial professionals, and researchers seeking to navigate the intersection of financial inclusion and risk management in an increasingly complex and interconnected world.

*Keywords* — Home Loan, Credit Risk Prediction, Machine Learning, Data Science, Data Analysis, Light Gradient Boosting (LGBM), Categorical Boosting (CatBoost), Hyperparameter Tuning, ROC Curves, Ensemble Methods.

## I. Introduction

Background - Home Credit is a financial institution launched in 1997 that intends to enable more people to have access to financial services, even if they don't have a significant credit history. They accomplish this by identifying better ways to comprehend the risks associated with lending money. This procedure enables them to approve more loan applications from individuals who would have been denied by traditional banks. Finally, they hope to ameliorate the financial situations of groups of individuals who have historically been excluded from the financial system. We aim to build an efficient and robust machine-learning model that helps such kinds of financial institutions. Below is an overview of the procedure that shows our road tobuilding a Home Loan Credit Risk Predictor.

1. Credit History Challenges and Data Science Solutions: This will most likely involve tackling the issues that people with limited or no credit history experience when applying for loans. Data science solutions could include examining alternative data sources other than standard credit reports to determine creditworthiness, such as utility or rent payment histories, mobile phone usage trends, or even social media behaviour.

2. Understanding Score Calculation and Maintaining Stability: This refers to the process of deciding how credit ratings are computed and maintaining the system's fairness and stability. It entails creating algorithms that may reliably predict creditworthiness while avoiding biases and ensuring consistency across time.

3. Building a Machine Learning Model to Predict Creditworthiness: This involves developing a predictive model using machine learning techniques to estimate the possibility of an individual repaying a loan. This methodology would provide a score or risk rating by analyzing different characteristics and data points about an individual such as income, employment history, historical payment behavior, and maybe alternative data sources. In short, these principles emphasize the problems and solutions involved in using data science and machine learning to make fair and accurate lending decisions to those with minimal credit history, ultimately contributing to our goal of increasing financial inclusion.

## II. Exploratory Data Analysis

### *A. Dataset Description*

Data serves as the cornerstone of any machine learning endeavour, shaping the performance and efficacy of our models. The selection and curation of a dataset are pivotal, as they lay the foundation for our model's understanding of underlying patterns. For illustrative purposes, we will utilize the (Home Loan Credit Risk) Dataset to elucidate the process

of model construction. However, the choice of dataset remains flexible, provided careful consideration is given to its relevance and quality, as it can significantly influence the outcome of our machine learning efforts." Uponencountering a dataset, our initial step is to embark on exploratory data analysis (EDA), a fundamental practice aimed at comprehending the data and its structure. Our dataset comprises diverse data sources, necessitating thepresence of unique identifiers such as case_id to establish connections between disparate pieces of information.

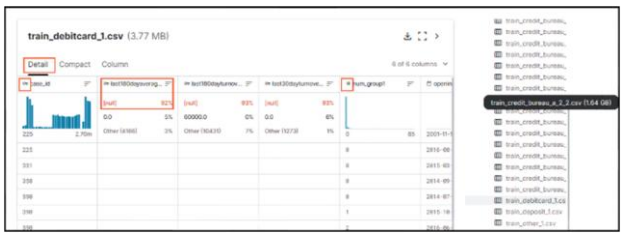


Fig. 1. Kaggle Data Explorer

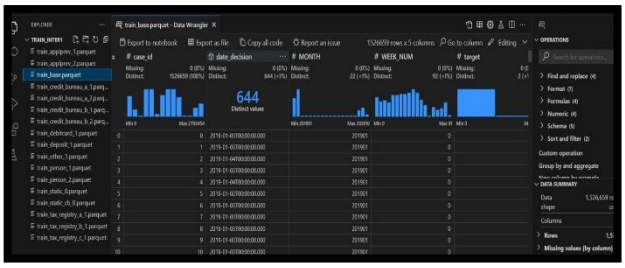


Fig. 2. MS VS Code Data Wrangler Extension

We recommend leveraging tools such as Kaggle Data Explorer when working with datasets sourced from platforms like Kaggle. This facilitates comprehensive exploration, offering insights into key metrics such as standard deviation, mean, minimum, maximum, mode, count, variance, and quartile regions. For offline dataset exploration, Data Wrangler—a Microsoft Visual Studio Code extension—proves invaluable, enabling us to delve deep into data distribution and gain a nuanced understanding of the dataset's intricacies.

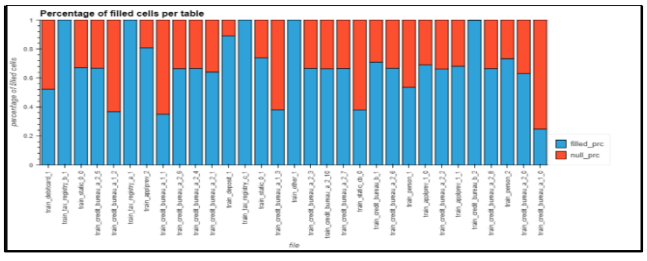


Fig. 3. Completeness of the Dataset

From the above image, we can see that we are dealing with a dataset that has 40.8% missing values. You can also check for other things like feature-wise null calculations and plot the bar charts to see if some features are not at all useful for our analysis. Finding these can help us get rid of redundant data to work on and make our model prediction more accurate. This marks the end of our data-wrangling phase. Now, we shall get started with the data-wrangling phase.

## III. DATA WRANGLING

Data wrangling encompasses numerous steps crucial for preparing data for analysis. Before delving into specifics, it's essential to consider the tools for reading and processing data. Currently, pandas and polars stand as industry standards, each with distinct advantages. Pandas boasts a rich ecosystem and seamless integration with powerful Python libraries, making it ideal for model training. Conversely, polar excels in processing times for dataframe transformations like ‘select”, “filter” and “groupby” operations. To strike an optimal balance between storage efficiency and processing speed, we've adopted a two-pronged approach:

• Storage Optimization: We've minimized the storage footprint by reducing numeric data to the lowest necessary bit size (e.g., from int64 to int8). Additionally, we've opted for parquet files over traditional CSVs for enhanced storage efficiency.

• Processing Efficiency: For data wrangling and processing tasks, we've leveraged polars due to their superior speed compared to pandas. Polars' efficiency ensures the swift execution of transformations, enabling seamless data manipulation and preparation. By implementing this approach, we've optimized both storage utilization and processing times, laying a solid foundation for efficient data wrangling and subsequent analysis.

### *A. Data Integration and Memory Optimization*

In the data integration process, files with identical structures are vertically concatenated into a single file, effectively eliminating redundant versions of the same file type. This consolidation results in a reduced number of files, with each representing a unique file type and its instances. For instance, starting with 32 files, the integration process yields 17 concatenated files, each encapsulating distinct file types and their respective instances.

Entity Relationship Diagram (ERD):

Below is the Entity Relationship Diagram (ERD) depicting the relationships between different entities within the dataset after the integration process. The ERD offers a visual representation of how various data entities are interconnected and provides insights into the dataset's structure and interdependencies.

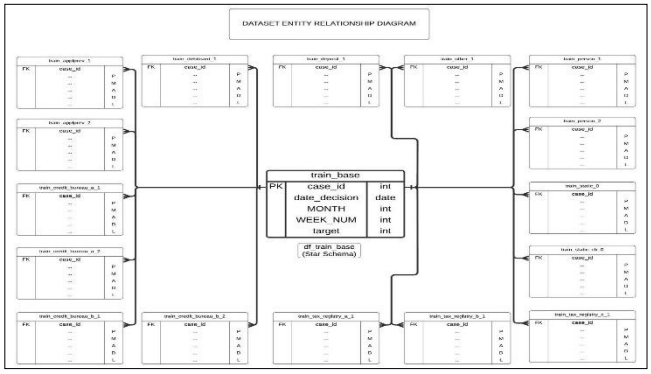


Fig. 4. Completeness of the Dataset

### *B. Data Cleaning*

After completing Data Integration and Memory Optimization, our dataset consists of all file types with one instance each. The next step involves data cleaning, which primarily focuses on removing redundant features to enhance model training efficacy.

- Handling Null Values: We identify and drop features with null values exceeding a predefined threshold (e.g., 70%). Features with a high proportion of missing values introduce noise into the model and are therefore considered redundant.

- Categorical Feature Selection: We scrutinize categorical columns to identify features with only one category, signifying redundancy. Additionally, we assess features with an excessive number of categories (e.g., more than 200) as they can lead to unreliable predictions and computational inefficiency. Such features are also dropped from the dataset.

As per the Entity Relationship Diagram, we designate the Train Base file as our main file due to its star connection with all other data sources in the dataset. This ensures comprehensive coverage of relevant information for model training. By systematically identifying and eliminating redundant features, we streamline the dataset and optimize it for subsequent machine-learning tasks, enhancing the quality and reliability of model predictions.

### *C. Data Aggregation*

Following Data Cleaning, where redundant columns are removed, the next step is Data Aggregation. Leveraging the Entity Relationship Diagram, we identify Train Base as the central file, connected to all other files in the dataset.

Data Aggregation involves joining different files based on a common key (e.g., case_id) and consolidating all relevant information into a single final file. This process ensures that all required columns or features are accessible and organized by case_id.

While there are various approaches to data aggregation, selecting a dataset with a high degree of connectivity as the main file is typical. Subsequently, relevant columns are extracted from other files and merged into this main file.

Post-aggregation, we obtain a final file containing all necessary columns from multiple files, with case_id serving as the unique identifier. However, the issue of missing data persists. Despite dropping columns with high missing values (beyond the 70% threshold), features with a 30% missing rate may still exist in the aggregated file.

After Data Aggregation we end up with a single file with:

1. No duplicate case_id’s
2. No duplicate columns (features)
3. Missing values

### *D. Feature Engineering*

In the context of Home Loan Credit risk prediction, feature engineering plays a crucial role in refining the dataset's structure and introducing new features to enhance predictive capabilities. Here's how we approach feature engineering:

- Date Transformation: Convert date-related columns such as defaulted date and due date into meaningful features by calculating the number of days from loan sanction to default. This transformation unveils underlying patterns and trends, enabling companies to adjust loan policies accordingly.

- Dimensionality Reduction: Identify the redundant features post-aggregation, especially those derived from multiple data sources but representing the same information (e.g., Date_of_Birth and DOB). Utilize correlation analysis to group such features and select a representative feature or column from each group. This ensures balance in training data and reduces redundant processing. For instance, we combine all features related to the applicant's birth date into a single representative feature, eliminating redundancy and improving model efficiency.

- Introduction of New Features: Introduce new features that capture important aspects of the data. For example, create a feature indicating the applicant's credit utilization ratio or the length of their credit history, which are known predictors of credit risk.

- Data Type Transformation: Adjust data types of existing columns or features as necessary. For instance, convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding.

By implementing these feature engineering techniques, we refine the dataset, enhance its predictive power, and streamline the modelling process. This ultimately leads to
more accurate and robust Home Loan Credit risk prediction models, benefiting both financial institutions and borrowers alike.

### *E. Data Imputation*

As the final step in our Data Wrangling phase, data imputation addresses the issue of missing values to ensure the dataset is complete and ready for model training. Our approach to data imputation is as follows:

1. Numeric Data: Replace missing values in numeric columns with the mean of the respective column. This approach provides a reasonable estimate without introducing significant bias into the dataset.

2. Categorical Data: For categorical columns, impute missing values with the mode of the column, representing the most frequently occurring category. This method preserves the distribution of categorical variables while filling in missing values.

3. Date Data: When dealing with date-type data, such as the last modified date or last status update, replace missing values with the last available date. Since the most recent information is typically the most relevant, this approach ensures that missing dates are filled with the latest available data.

By employing these imputation methods, we ensure that our refined dataset is free from missing values and ready for model training. This marks the culmination of our Data Wrangling phase, resulting in a single file with no duplicate Case_IDs, no duplicate columns (features), and no missing values. The refined data is now primed for further analysis and model development, contributing to more accurate and reliable predictions in Home Loan Credit risk assessment.

## IV. Model Selection,Training And Tuning Procedure

In tackling the Home Loan Credit risk prediction, our objective is to classify whether users will default on their loans. We have experimented with various classification techniques, with a focus on tree-based models due to their efficiency. Specifically, we have explored Random Forest and Light GBM models for this task.

Given the imbalanced nature of real-time data, where a significant proportion of users repay loans on time, this skew is reflected in the training dataset. This class imbalance poses a challenge for model training and can lead to biased predictions favoring the majority class. To address this imbalance, we've explored oversampling techniques such as Random Oversampling, SMOTE, and ADASYN, as well as Stratified K-fold cross-validation. While oversampling methods aim to create synthetic data to balance the classes, they may perform poorly due to reliance on artificially generated samples. Conversely, Stratified K-fold cross-validation ensures that class proportions remain consistent across folds, resulting in more reliable model evaluation.

Initial experiments with LightGBM using only non-categorical data resulted in decreased model performance, highlighting the significance of categorical features in loan default prediction. To address this, we turned to CatBoost, known for its efficient handling of categorical data. Additionally, we revisited Random Forest and LightGBM, this time incorporating all features, including categorical variables. By leveraging both categorical and non-categorical data, we aimed to optimize model performance in Home Loan Credit risk prediction. Through these iterations, we seek to identify the most effective approach for predicting loan defaults accurately.

In our Ablation Study, we've evaluated model performances across multiple epochs using cross-validation to assess predictive outcomes comprehensively. This iterative approach provides insights into model stability and generalization capabilities over time. Through careful experimentation and evaluation, we aim to identify the most effective approach for addressing class imbalance and building robust Home Loan Credit risk prediction models.

## V. Ablation Study

In our Ablation Study, we are showcasing the results with different parameters, opting for an iterative approach instead of exhaustive methods like GridSearchCV and Cross-Validation using Stratified K fold.

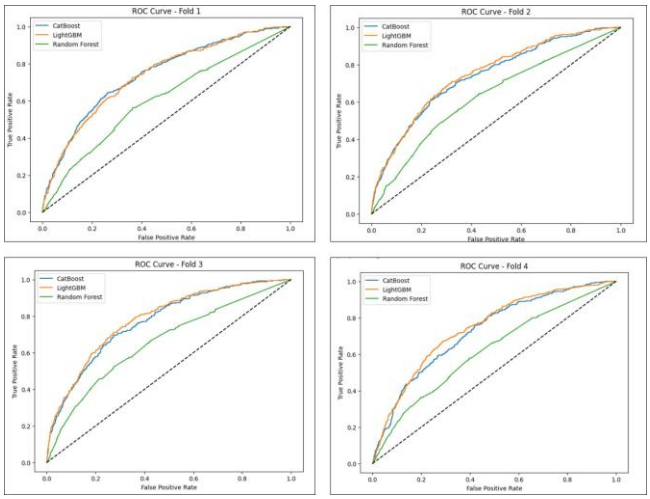


Fig 5. ROC curve

Also, Due to the dataset's size, exhaustive hyperparameter tuning can be computationally intensive though they might perform better. Hence, we focus solely on experimenting with basic parameters like n_estimators for now, providing insights into their impact on model performance. This pragmatic approach balances computational efficiency with parameter exploration, allowing us to glean valuable insights into model behavior and identify optimal configurations for Home Loan Credit risk prediction. Results when changing in modeling steps

TABLE I. Results when changing in modeling steps

| Type | Condition | Best AUC Score | | |
|---|---|---|---|---|
| | | ***RF*** | ***CAT*** | ***LGBM*** |
| ML | RF Estimators: 60<br>CAT Estimators: 4000<br>LGB Estimators: 1000 | MaxAUC score: 0.6102 | Max AUC score: 0.7624 | Max AUC score: 0.7701 |
| ML | RF Estimators: 60<br>CAT Estimators: 4000<br>LGB Estimators: 1000 | Max AUC score: 0.6573 | Max AUC score: 0.7716 | Max AUC score: 0.7844 |
| ML | RF Estimators: 60<br>CAT Estimators: 4000<br>LGB Estimators: 1000 | Max AUC score: 0.6495 | Max AUC score: 0.7733 | Max AUC score: 0.7820 |

a. ML = Examining Modeling Changes

In our experimentation, we have varied the thresholds for null value frequency and categorical frequency to observe their effects on model performance. By adjusting these thresholds, we aim to optimize the preprocessing steps and enhance the predictive accuracy of our models.

Below, we present the results of these experiments, highlighting the impact of different threshold values on model performance metrics.

TABLE II. RESULTS WHEN CHANGING IN DATA WRANGLING STEPS

| Type | Condition | Best AUC Score | | |
|---|---|---|---|---|
| | | ***RF*** | ***CAT*** | ***LGBM*** |
| DW | Null Frequency Threshold: 0.5 | MaxAUC score: 0.6523 | Max AUC score: 0.7736 | Max AUC score: 0.7856 |
| DW | Null Frequency Threshold: 0.5 | Max AUC score: 0.6573 | Max AUC score: 0.7716 | Max AUC score: 0.7844 |
| DW | Null Frequency Threshold: 0.5 | Max AUC score: 0.6429 | Max AUC score: 0.7747 | Max AUC score: 0.7834 |

a. DW = Examining Changes in Data Wrangling

TABLE III. RESULTS WHEN CHANGING IN DATA WRANGLING STEPS 2

| Type | Condition | Best AUC Score | | |
|---|---|---|---|---|
| | | ***RF*** | ***CAT*** | ***LGBM*** |
| DW | Categorical Frequency Theshold: 190 | MaxAUC score: 0.6557 | Max AUC score: 0.7674 | Max AUC score: 0.7841 |
| DW | Categorical Frequency Theshold: 200 | Max AUC score: 0.6605 | Max AUC score: 0.7714 | Max AUC score: 0.7844 |
| DW | Categorical Frequency Theshold: 210 | Max AUC score: 0.6673 | Max AUC score: 0.7716 | Max AUC score: 0.7861 |

a. DW = Examining Changes in Data Wrangling

This structured and comparative approach in applying Random Forest (RF), CatBoost (CAT) and LightGBM (LGBM) models enabled us to robustly assess their applicability and performance in financial risk assessment tasks, ensuring that our findings are grounded in rigorous empirical evaluation. Below are the best performance results over a few epochs.

TABLE IV. BEST AUC SCORE

| Best AUC Score | | |
|---|---|---|
| ***RF*** | ***CAT*** | ***LGBM*** |
| 0.6605 | 0.7747 | 0.7856 |

a. Above is for a limited set of parameters. You tune them as you wish for different observations

As observed, Random Forest has less AUC curve when compared to LGBM and CatBoost.

## VI. KEY ACCOMPLISHMENTS

This predictive model, based on the Light Gradient Boosting Machine (LightGBM) framework, achieved commendable recognition in the Kaggle competition. It demonstrated robust performance metrics, securing a competitive score of 0.585. This performance not only validates the effectiveness of LightGBM in predicting client defaults but also highlights its potential for practical applications in financial risk assessments. Furthermore, the model attained an impressive Area Under the Curve (AUC) score of 0.7856, indicating a high level of accuracy in distinguishing between default and non-default cases. These results underscore the model's capabilities and its applicability in the domain of risk management.

## VII. CONCLUSION

Based on the ROC curves observed across different folds and epochs, it's evident that both CatBoost and LightGBM exhibit impressive performance, surpassing the Random Forest Classifier. While both CatBoost and LightGBM excel, their relative performance may vary slightly across different folds and epochs.

To capitalize on the strengths of both models, we employ ensemble methods to merge predictions from LightGBM and CatBoost. By combining predictions from multiple models, we aim to leverage the diverse strengths and tendencies of each model, resulting in a more robust and reliable prediction. The ensemble method aggregates predictions by taking an average output, ensuring a balanced and comprehensive approach to Home Loan Credit risk prediction.

Through this fusion of predictions, we harness the complementary strengths of CatBoost and LightGBM, enhancing the overall predictive accuracy and reliability of our model ensemble.

## VIII. FUTURE SCOPE

A viable strategy for data imputation is to use K-Nearest Neighbors (KNN). KNN imputation estimates missing values by considering the values of surrounding data points. This method uses the similarity of data points to anticipate missing values, making it applicable to a variety of datasets

Ensemble methods: You can enhance your model's predictive performance by investigating other ensemble approaches including stacking, Random Forest, and Gradient Boosting. Ensemble approaches, which frequently outperform individual models, aggregate predictions from several models to provide forecasts that are more accurate.

Optimizing strategies for handling categorical data: Special attention must be paid to categorical data during preprocessing. Improved model performance might result from investigating various encoding strategies (such as label encoding and one-hot encoding) and feature engineering methods tailored to categorical variables.

XGBoost, or Extreme Gradient Boosting, is a powerful machine learning method known for its excellent performance and efficiency, particularly with large datasets. It offers built-in regularization techniques like L1 and L2 regularization to prevent overfitting and improve generalization. XGBoost is highly flexible, allowing customization of the model to suit specific problem domains and evaluation criteria. It provides feature significance scores to identify predictive features and has been widely adopted in both industry and academia, winning numerous machine learning competitions. In summary, XGBoost's robustness, adaptability, and feature-rich capabilities make it a valuable tool for developing reliable and accurate machine-learning models.

## IX. References


[1] Yosza, D., Faris, M., & Muslim, M. A. (2023). Credit risk assessment in P2P lending using LightGBM and particle swarm optimization. *Register Jurnal Ilmiah Teknologi Sistem Informasi*, March 2023.

[2] Baesens, B., & Smedts, K. (2023). Boosting credit risk models. *Faculty of Economics and Business*, August 1, 2023.

[3] Wanga, Y., Zhanga, Y., Lua, Y., & Yua, X. (2020). A comparative assessment of credit risk model based on machine learning. *IIKI 2019: 8th International Conference on Identification, Information & Knowledge in the Internet of Things*.

[4] Taş, C. (2023). Comparison of machine learning and standard credit risk models. *A Thesis Submitted to the Graduate School of Informatics, Middle East Technical University*, July 2023.

[5] Saha, T., Sanyal, S., Biswas, S. K., & Patro, B. (2023). Credit risk prediction using machine learning analytics: An ensemble model. *International Journal of Management and Applied Science*, 9(9), September 2023. ISSN: 2394-7926.

[6] Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A comparative assessment of credit risk model based on machine learning: A case study of bank loan data. *Procedia Computer Science*, 174, 141–149.

[7] Coskun, S. B., & Turanli, M. (2023). Credit risk analysis using boosting methods. *JAMSI*, 19(1).

[8] Pillai, S. G., Woodbury, J., Dikshit, N., Leider, A., & Tappert, C. C. (2019). Credit risk analysis applying machine learning classification models. *Proceedings of the Future Technologies Conference (FTC) 2019: Advances in Intelligent Systems and Computing*, 1069, 107–126.

[9] Manisha, N., Raj, M. R., Manimala, S., Soniya, D., & Kiran, K. S. (2022). Credit risk assessment for Home Credit Group. *International Journal of Creative Research Thoughts (IJCRT)*, April 2022. Retrieved from www.ijcrt.org.

[10] Dasril, Y., Arisandy, Y., Salahudin, S. N., Mahmud, N., Zinnah, K. I., Ar Rahman, Y., & Ahmed, N. (2023). Home credit default risk assessment using embedded feature selection and stacking ensemble technique. *Journal of Numerical Optimization and Technology Management*, 1(2).

[11] Li, Y. (2019). Credit risk prediction based on machine learning methods. *14th International Conference on Computer Science & Education (ICCSE 2019)*, Toronto, Canada, August 19–21, 2019, pp. 1011–1013.

[12] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

[13] Mahmudi, H., Bhargava, R., & Das, R. (2022). Evaluation of gradient boosting algorithms on balanced home credit default risk. *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT 2022)*, pp. 1–6. doi: 10.1109/TQCEBT54229.2022.10041584.

[14] Li, S., Dong, X., Ma, D., Dang, B., Zang, H., & Gong, Y. (2024). Utilizing the LightGBM algorithm for operator user credit assessment research. March 2024.

[15] Kaggle. (n.d.). Home Credit: Credit Risk Model Stability. Retrieved from https://www.kaggle.com/competitions/home-credit-credit-risk-model-stability.