Interdisciplinary Social Studies
P-ISSN: 2808-0467
E-ISSN: 2808-5051

Comparative Analysis of Machine Learning Models for Predicting Default in Home Credit Companies

Daniel Alexander, Raden Supriyanto
Universitas Gunadarma, Indonesia
Email: daniel.pnjt@gmail.com, supriyanto.r@gmail.com

Keywords:

Credit Scoring,
Machine Learning,
Home Credit

ABSTRACT

Credit scoring is an important process in the financial world to accurately identify the risk of prospective debtors. This study aims to build a credit scoring model by comparing the performance of several machine learning algorithms, namely XGBoost, random forest, and logistic regression. The dataset used consists of training and test data with several ratios: 70:30, 75:25, and 80:20. The preparation process involves variable selection using the 5C principle, missing value imputation, categorical transformation of variables, and creation of derived features. Furthermore, modeling and optimization are performed for each model to improve classification performance, especially in identifying debtors with the potential to default. The evaluation results show that the XGBoost model has the best performance, with an accuracy of 84.8%, a precision of 84.9%, a recall of 84.7%, and an AUC of 92.3%. The main assessment of the character principle is the external credit score variable; the main assessment of the capacity principle is income; the main assessment of the capital principle is car ownership; the main assessment of the collateral principle is the credit financing ratio; and the main assessment of the condition principle is the regional rating.

This is an open access article under the CC BY-SA license.

INTRODUCTION

Credit is the provision of money or equivalent bills based on an agreement or contract between a bank and another party, requiring the financed party to return the money or bills after a certain period in return for compensation or profit sharing (Fahmi, 2015; Halim, 2021; Kasmir, 2016; Salsabila, 2015; Sukamulja, 2022). Simply put, credit can be understood as a financial facility that allows individuals or business entities to obtain funds to meet certain needs, such as purchasing products, investments, or business capital, with the obligation to repay within a predetermined time period. The provision of credit is not only limited to banking institutions but can also be facilitated by various other financial service entities, including finance companies, savings and loan cooperatives, and financial technology, which is increasingly popular in the digital era. Banks generally have a conventional approach with in-depth analysis and a longer application process, while financial technology companies offer a faster and more flexible process. Financial technology also offers a variety of services, ranging from digital payments and online loans to application-based investments, which make it easier for people to access financial services (Aizat & Nazjmi, 2019; Gunawan et al., 2023; Hasanudin & Rahmiyanti, 2023; Jaumotte et al., 2021; Muslih & Supeno, 2022).

Technological advances have expanded access to credit facilities, allowing segments of society previously unreached by traditional banking services, such as micro-entrepreneurs and individuals with no formal credit history, to obtain financing. One of the financial technology products that is currently developing is cash financing. Cash financing in financial technology is preferred over traditional financial services such as banks for several

reasons related to convenience, accessibility, and speed of services. The speed of financing services carried out by financial technology has an impact on the need for the speed of the verification process of prospective borrowers, making it necessary to increase efficiency and accuracy in the decision-making process, especially in assessing creditworthiness.

Home Credit is one of the financial technology companies currently developing in Indonesia. Founded in 1997, Home Credit is an international consumer financing service provider that currently operates in seven countries in Central and Eastern Europe and Asia. Home Credit encourages and expands financial inclusion and spurs economic development through support for domestic consumption. Home Credit has several business products, namely goods financing, cash financing, paylater, and protection.

Based on audited financial statements, Home Credit recorded positive financial performance over the last three years (2022–2024). Net interest income continues to increase from 1.58 trillion rupiah in 2022 to 2.03 trillion rupiah in 2024. Revenue from fees and net commissions also showed an increase to 995.19 billion rupiah in 2024. Although operating expenses also increased, profit before tax managed to grow to 116.17 billion rupiah in 2024, reflecting better efficiency in the company's operations. Net profit was recorded at 65.44 billion rupiah in 2024, almost 2.5 times higher than the previous year. With the growth of Home Credit application users exceeding 17 million, this achievement confirms the importance of implementing an accurate and efficient credit scoring system in the credit granting process. The credit scoring model is a strategic element to maintain the quality of the financing portfolio while encouraging sustainable business growth.

Numerous studies have explored machine learning applications in credit scoring. Wang, Zhang, Lu, & Yu (2020) conducted a comparative assessment of various machine learning models, demonstrating the superiority of ensemble methods in credit risk prediction. Similarly, Chen & Guestrin (2016) introduced XGBoost as a scalable tree boosting system that consistently outperforms other algorithms in classification tasks. In the Indonesian context, Andriani, Gunawan, & Naja (2025) compared machine learning algorithms for credit eligibility prediction in Bank BRI Tegal, finding that tree-based methods achieved higher accuracy than traditional statistical approaches. However, these studies primarily focused on traditional banking institutions rather than fintech environments where data characteristics and customer profiles differ significantly. Research by Suhadolnik, Ueyama, & Silva (2023) specifically addressed machine learning for enhanced credit risk assessment but did not comprehensively compare multiple algorithms across different data splitting ratios while incorporating the *5C* principle analysis.

This study offers several novel contributions to the existing literature. First, it provides a comprehensive comparison of three distinct machine learning algorithms (XGBoost, Random Forest, and Logistic Regression) specifically applied to Home Credit's unique dataset, which represents Indonesian fintech borrowers. Second, the research employs multiple data splitting ratios (70:30, 75:25, and 80:20) to evaluate model consistency and robustness, an approach not commonly addressed in previous studies. Third, the study integrates the traditional *5C* principle of credit assessment with modern machine learning interpretation, bridging conventional financial wisdom with advanced analytics. Fourth, the research implements SMOTE resampling to address class imbalance, enhancing model performance for default prediction. Finally, the study provides specific practical

recommendations for Home Credit's operational implementation of machine learning-based credit scoring, moving beyond theoretical comparison to actionable insights.

Technological advances have opened up new opportunities to improve the accuracy of predictive models in various domains, including credit scoring models using machine learning methods. Machine learning methods offer the effectiveness of modeling complex interactions in data (Varian, 2014). Using these methods, we can extract important patterns from data, thereby minimizing errors in predicting customer default. The main advantage of machine learning lies in its ability to process large amounts of high-dimensional data to discover patterns, make predictions, and produce accurate classifications (Devi & Raju, 2024). Through machine learning, financial institutions can make more accurate, faster, and objective credit decisions while expanding access to financing for previously unreachable segments of society. Various models in machine learning methods are capable of classification to create credit scoring models with advantages and disadvantages for each model, such as the logistic regression model, which has the advantage of being easy to understand compared to other models but tends to have lower accuracy on unbalanced data because it is susceptible to underfitting (Rianto & Wahono, 2015). The random forest model has advantages in overcoming noise, missing values, and overfitting with maximum accuracy results even when processing large datasets, but the training time with this algorithm is longer compared to other algorithms (Awangga & Khonsa, 2022). The XGBoost model shows the best performance in handling large-scale datasets, with an optimal balance between accuracy and computational efficiency (Lusiyanti, Musdalifah, Sahari, & Al Fajri, 2025).

This research focuses on the application of machine learning to compare the accuracy of predicting customer default using a credit scoring model developed through machine learning methods. The data used were obtained from the Home Credit Default Risk dataset on the Kaggle website, accessed in November 2024. To address the imbalanced classes in the data, this research will use a resampling process, namely SMOTE. The analysis is conducted by considering various model performance metrics, including accuracy, precision, recall, and area under the curve, to ensure that the machine learning methods provide more reliable predictions. The machine learning algorithms used in this research are the XGBoost model, the Random Forest model, and the Logistic Regression model. These findings are expected to provide in-depth insights into how to optimize the application of machine learning in the real world, especially in the financial industry. Thus, this research aims not only to provide theoretical contributions but also to develop practical recommendations for financial companies wishing to adopt this technology in their credit scoring systems, thereby supporting operational efficiency, expanding access to financing, and better managing risk.

METHOD

The sequence of data analysis techniques in the implementation of this research research is described in the flowchart as follows

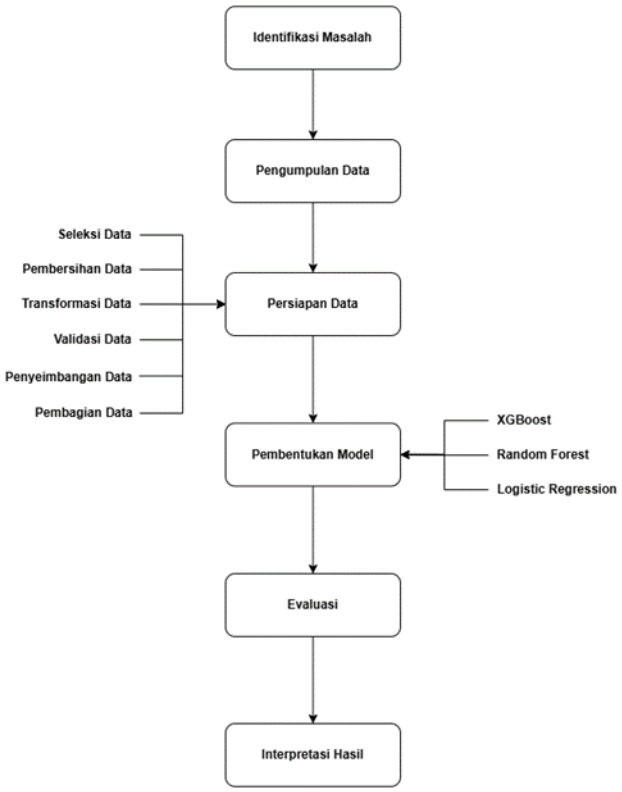


Figure 1. Research Stage Flowchart
Source: Author's Preparation (2024)

In this study, several stages of research were carried out to analyze machine learning methods and traditional methods in credit scoring modeling. The research begins with the identification of the problem. This initial stage provides an overview of how research is conducted. The research was then continued with data collection for research. Data collection was carried out using secondary data obtained from the Kaggle website accessed in November 2024. After the data is collected, the researcher proceeds to the data preparation stage. At this stage, the researcher prepares the data into viable data to proceed to the modeling stage. The model testing stage is carried out to train each model in recognizing patterns. The machine learning method is the XGBoost model, the random forest model, and the logistic regression model. The training results of each model are validated and analyzed at a later stage by comparing the accuracy value of each model in creating a credit scoring model. Next, the result interpretation stage, where an analysis will be carried out related to the influence of each variable on the prediction results.

Research data collection techniques are the methods or means used to gather the information needed for a study or research. The choice of data collection technique is crucial because it can affect the validity and reliability of research results, which must be aligned with the research objectives (Dr. Zainuddin Iba, 2023). In this study, the data used is secondary data obtained from the Kaggle website, which provides the Home Credit Default Risk dataset. The dataset can be accessed and downloaded through the following link: https://www.kaggle.com/competitions/home-credit-default-risk/data.

RESULTS AND DISCUSSION

Data Collection

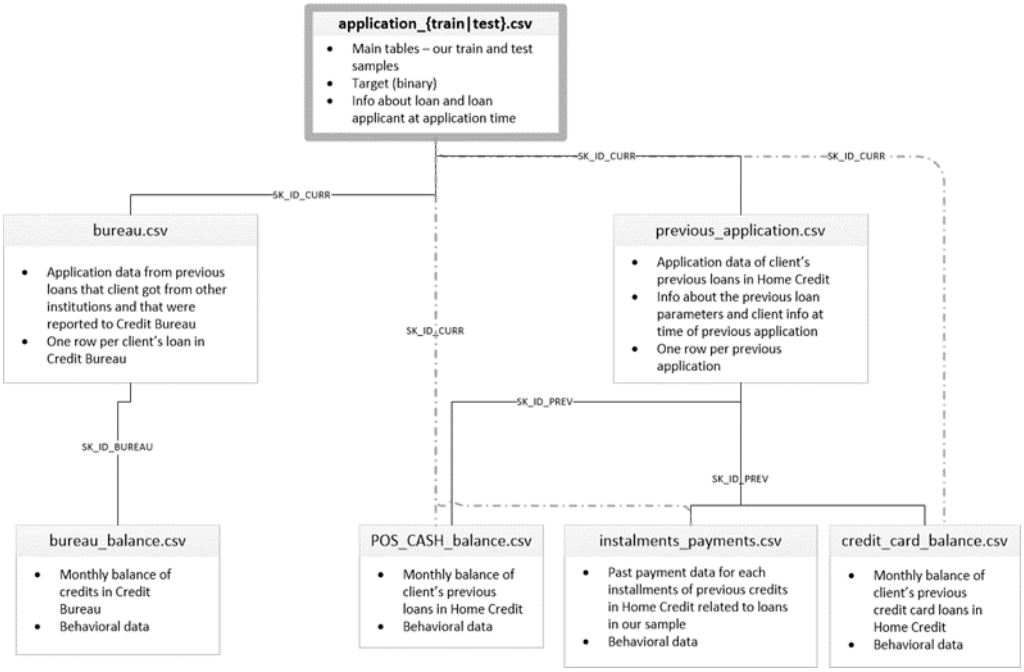


Figure 2. Relational Data Diagram Images

Source: Processed Data (2024)

There is an outlier in the income variable because the maximum value is far different from other values, so it is necessary to clean the outlier on this variable. In addition, it is necessary to combine data to combine the three credit history indicators so that it will produce one indicator for credit history. Furthermore, it is necessary to change the variables of marital status, type of work, and home ownership because these variables are not categorical variables that have rankings. In addition, it is necessary to change the data type for categorical variables to numerical ones.

The variables of debtor age and credit history have a decreasing curve which means that the lower the variable, the lower the risk of default. Meanwhile, the *credit utilization ratio* has an increasing curve which means that the higher the value of the variable will make an increase in the risk of default. On the other hand, it is shown that the income variables, *debt to income ratio*, and *loan to value ratio* are unclear curves, so it is necessary to change from numerical data types to group data.

Each parameter of the evaluation results to measure the feasibility of the model shows that each data sharing ratio produces a non-significant difference. This shows that the model is consistent in doing the classification correctly. The average *accuracy* value of the *XGBoost* model is 0.848 which means that it can correctly predict data in 84.8% of the total cases. The average *AUC* value of the *XGBoost* model is 0.923 which means that it can distinguish defaulting debtors from current debtors very well. In addition, this model is very fast in data training, where the average data training is 3,912 seconds.

Table 1. Results of Random Forest Model Evaluation on Various Data Sharing Ratios

Parameter	70:30	75:25	80:20	Average
Accuracy	0,798	0,796	0,798	0,797
Precision	0,788	0,784	0,787	0,786
Recall	0,817	0,816	0,817	0,817
AUC	0,878	0,877	0,878	0,878
Training time	25,909 second	26,432 second	39,347 second	30,563 second

Source: Processed Data (2024)

Each parameter of the evaluation results to measure the feasibility of the model shows that each data sharing ratio produces a non-significant difference. This shows that the model is consistent in doing the classification correctly. The average accuracy value of the random forest model is 0.797 which means that this model can correctly predict data in 79.7% of the total cases. The average AUC value of the random forest model is 0.878 which means that this model can distinguish defaulting debtors from current debtors very well. In addition, this model is not fast enough in data training, where the average data training is 30.563 seconds.

Each parameter of the evaluation results to measure the feasibility of the model shows that each data sharing ratio produces a non-significant difference. This shows that the model is consistent in doing the classification correctly. The average accuracy value of the logistic regression model is 0.812 which means that this model can correctly predict the data in 81.2% of the total cases. The average AUC value of the logistic regression model is 0.889 which means that this model can distinguish defaulting debtors from current debtors very well. In addition, this model is very fast in data training, where the average data training is 6.129 seconds.

Based on the results of the model evaluation, it can be concluded that the XGBoost model provides the best overall performance compared to random forest and logistic regression. The XGBoost model recorded the best accuracy parameter with a value of 0.866 which means that the value of the model's accuracy in making predictions is 84.8%. The XGBoost model also excels in precision, recall and AUC parameters which make the XGBoost model more accurate in recognizing defaulting debtors.

Based on the results of the modeling that has been carried out, it is known that the XGBoost model is relevant to the logistic regression model because the values are not much different, so an analysis will be carried out to determine the amount of influence of each variable on the target variable.

Debtors who live in areas and cities with a moderate rating. This shows that the rating of a region with a moderate value requires a loan compared to a low or high rating of a region. The interesting thing is that the rating of regions and cities with low ratings have a greater chance compared to regions that have a medium rating. After further analysis, it was found that many debtors have houses in low ratings, this means that the rating of the area is directly proportional to the price of houses in the area.

Overall, the target of Home Credit is debtors who are already economically established. This makes Home Credit at a safe point to run its business. The principle of character is able to see the characteristics of a debtor with the characteristics of a debtor who

is married, a minimum level of high school or vocational education, and the debtor has a good credit record in another institution will make his loan application acceptable to Home Credit. The target of Home Credit debtors who come from the upper middle economy and the majority of debtors already own houses and cars as well as the ability of debtors to manage debts makes credit applications run smoothly as they should.

The formation of the random forest model is carried out using the parameters described earlier. This model is then evaluated based on several scenarios of the ratio of the distribution of training data and test data, namely 70:30, 75:25, and 80:20 as follows:

1. Ratio 70:30

   A random forest model was carried out using a 70:30 ratio, where 70% of the total data will be used as training data and 30% of the total data will be used as data testing.

2. Ratio 75:25

   A random forest model was carried out using a ratio of 75:25, where 70% of the total data will be used as training data and 30% of the total data will be used as testing data.

3. Ratio 80:20

   A random forest model was carried out using an 80:20 ratio, where 70% of the total data will be used as training data and 30% of the total data will be used as testing data.

Based on the results of the evaluation of three machine learning models, namely XGBoost, random forest, and logistic regression, it can be concluded that the three machine learning methods can be used to form a credit scoring model. All three have the ability to handle binary classification, which is to distinguish between risky and non-risky debtors. Based on the results of the evaluation of performance metrics with various data sharing ratios, the three models showed relatively stable performance in classification.

Based on the training time comparison, the XGBoost model shows the most optimal performance compared to the other two models. The XGBoost model has a training time of 3.9 seconds, while random forest has a training time of 30.6 seconds and logistic regression has a training time of 6.1 seconds. Rapid model training and superior classification results show that XGBoost has great potential to be implemented in a real-time debtor screening system at Home Credit.

Overall, the XGBoost model showed the best performance with an accuracy value of 84.8%, precision of 85.3%, recall of 84.3%, and AUC of 92.3%. This indicates that XGBoost is not only accurate, but also able to distinguish classes in a balanced way between risky and non-risky debtors. Meanwhile, the random forest model produces an accuracy of 79.7%, precision of 78.6%, recall of 81.7%, and an AUC of 87.8%, which shows a fairly good performance despite being below XGBoost. The logistic regression model showed the lowest performance, with an accuracy of 81.2%, accuracy of 81.0%, recall of 81.5%, and an AUC of 88.9%.

Based on the results of the model interpretation, it is obtained that important variables in determining creditworthiness can be categorized into the 5C principle. For the principle of character, marital status is the main indicator in assessing the character of prospective debtors. In addition, credit history is the second indicator to assess the character of

prospective debtors, where the greater the external credit score, the greater the likelihood of receiving credit applications. In the principle of capacity, the type of work is the main indicator in assessing the capacity of prospective debtors. In addition, the amount of debtor's income greatly affects the ability to pay loan installments where the greater the income, the greater the likelihood of the credit application being accepted. For the principle of Capital, the type of home ownership is an important indicator in assessing the debtor's capital. In addition, car ownership is a representation of the capital strength possessed by the debtor. In the Collateral principle, credit card ownership is an important variable in assessing this principle. In addition, the credit utilization ratio reflects how the debtor manages his existing loans where the lower this value, the greater the likelihood of his credit application being accepted. Meanwhile, in the principle of Condition, the assessment of the debtor's area of residence through regional ratings shows an influence on the decision to grant credit. Thus, the XGBoost model not only excels in terms of technical performance, but is also able to represent the principles of comprehensive credit evaluation based on the 5C approach.

CONCLUSION

A comprehensive evaluation of three machine learning models determined that XGBoost is the best for credit scoring at Home Credit, achieving an accuracy of 85.3%, recall of 84.3%, AUC of 92.3%, and the fastest training time of 3.9 seconds, enabling real-time implementation. Beyond its technical superiority, XGBoost effectively identifies key credit assessment variables aligned with the *5C* principles, such as external credit history (character), income (capacity), car ownership (capital), credit utilization ratio (collateral), and regional rating (condition). Implementing this model is expected to enhance the accurate identification of at-risk debtors, accelerate credit decisions, reduce defaults, and improve the quality of Home Credit's loan portfolio. Future research could explore integrating additional alternative data sources and testing deep learning models to further improve predictive accuracy and robustness in fintech credit scoring.

REFERENCES

Aizat, S. M., & Nazjmi, M. F. W. (2019). Financial technology (fintech) services in Islamic financial institutions. *International Postgraduate Conference*, November.

Andriani, W., Gunawan, G., & Naja, N. N. (2025). Analisis perbandingan machine learning untuk prediksi kelayakan kredit perbankan pada Bank BRI Tegal. *Jurnal Penerapan Teknologi Informasi dan Komunikasi*, 82–92.

Awangga, R. M., & Khonsa, N. H. (2022). Analisis performa algoritma random forest dan naive bayes multinomial pada dataset ulasan obat dan ulasan film. *Jurnal Telekomunikasi dan Komputer*, 60–70.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Devi, & Raju. (2024). A comparative analysis of machine learning algorithms for big data. *International Journal of Management Science and Engineering Management*, 1608–1630.

Fahmi, I. (2015). *Analisis laporan keuangan*. Alfabeta.

Gunawan, A., Jufrizen, & Pulungan, D. R. (2023). Improving MSME performance through financial literacy, financial technology, and financial inclusion. *International Journal of Applied Economics, Finance and Accounting, 15*(1). https://doi.org/10.33094/ijaefa.v15i1.761

Halim, I. (2021). *Analisis laporan keuangan*.

Hasanudin, A. I., & Rahmiyanti, S. (2023). The influence of financial literacy, financial technology, and inclusion finance on MSME financial performance in Cilegon City. *Jurnal Manajemen, 12*(1).

Iba, Z. (2023). *Metode penelitian*. Eureka Media Aksara.

Jaumotte, F., Lall, S., & Papageorgiou, C. (2021). Rising income inequality: Technology, or trade and financial globalization? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1266518

Kasmir. (2016). Pengertian laporan keuangan, tujuan laporan keuangan dan analisis laporan keuangan. *Jurnal Akuntansi, 53*(9).

Lusiyanti, D., Musdalifah, S., Sahari, A., & Al Fajri, I. (2025). Evaluasi kinerja algoritma machine learning. *Jurnal Matematika*, 84–92.

Muslih, M., & Supeno. (2022). Financial technology: Digital legal challenges and Indonesia's economic prospects after Covid-19 outbreak. *Legality: Jurnal Ilmiah Hukum, 30*(2). https://doi.org/10.22219/ljih.v30i2.22784

Rianto, H., & Wahono, R. S. (2015). Resampling logistic regression untuk penanganan ketidakseimbangan class pada prediksi cacat software. *Journal of Software Engineering*, 46–53.

Salsabila, A. (2015). Pengertian laporan keuangan, tujuan laporan keuangan dan analisis laporan keuangan. *Jurnal Akuntansi, 53*(9).

Sukamulja, S. (2022). *Analisis laporan keuangan: Sebagai dasar pengambilan keputusan investasi* (Edisi revisi). Penerbit Andi.

Suhadolnik, N., Ueyama, J., & Silva, S. D. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management, 16*(7), 496. https://doi.org/10.3390/jrfm16070496

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A comparative assessment of credit risk model based on machine learning. *Procedia Computer Science, 174*, 141–149. https://doi.org/10.1016/j.procs.2020.06.120