

Exploring effective feature selection methods in logistic regression for credit scoring models in XYZ bank

Brigitta Permadhi, Devi Fitriana

Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Nov 16, 2025

Revised Jun 25, 2026

Accepted Aug 28, 2026

Keywords:

Accuracy

Credit scoring

Feature selection

Hybrid model

Logistic regression

ABSTRACT

This research explores optimized variable selection in credit scoring models, comparing wrapper, embedded, none, and hybrid approaches. Logistic regression is applied to model the dependence of explanatory variables on the target variable while mitigating underfitting, overfitting, and computational challenges. Feature selection methods help mitigate these issues by reducing dimensionality and improving model generalization. The Hybrid model achieved the best overall performance with accuracy = 0.9840 on the test set and accuracy = 0.9839 on the training set (test-train difference = -0.0001), as well as the best recall = 0.9993, F1-score = 0.9842, and the least number of false positives (FP = 47) and false negatives (FN = 1), which suggests good predictive ability and stability with the fewest misclassifications. The wrapper and embedded models have lower accuracies of 0.9793 and 0.9830, respectively, and mild overfitting. The baseline model with no feature selection performed the worst in terms of both accuracy and F1-score (accuracy = 0.5000, F1-score = 0.6667). In conclusion, the results indicate that the combination of wrapper and embedded approaches is the most accurate, stable, and generalisable model for credit scoring tasks.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Brigitta Permadhi

Computer Science Department, BINUS Graduate Program – Master of Computer Science

Bina Nusantara University

Jakarta, Indonesia

Email: brigitta.permadhi@binus.ac.id

1. INTRODUCTION

In today's data-driven world, eligibility scoring models play a crucial role in supporting individuals and organizations in making well-informed decisions [1]. Eligibility scoring is a foundational process for evaluating objects based on pertinent variables, allowing for the prediction of specific characteristics [2]. This capability enables businesses and individuals to assess risks and make well informed choices [3]. It has applications across various sectors, including finance, healthcare, marketing, and insurance [4].

This method assigns numerical scores to objects for comparison, classification, and decision-making, making it essential in fields requiring objective analysis. In credit scoring, predictive techniques such as logistic regression, random forest and decision trees analyze data to assess credit risk, producing scores that indicate levels of risk, quality or probability [2].

A key challenge in credit scoring is overfitting, which occurs when a model becomes too tailored to the training data, reducing its ability to generalize to new data, leading to poor real-world performance. Conversely, underfitting occurs when the model is too simplistic to recognize meaningful data patterns, resulting in weak predictive performance on both training and testing sets [5]. To address these issues, feature selection methods are essential for effectively reducing feature dimensionality [6].

Feature selection models are valuable for data analysts in selecting the most significant variables [7]. By systematically selecting only the most relevant features, feature selection enhances model efficiency, simplifies computations, and reduces overfitting risk, leading to more interpretable and computationally efficient credit scoring models [8].

The wrapper method follows an iterative process which is initializing a feature subset, training the model, evaluating its performance, updating the subset, and repeating the cycle until a stopping condition is met [9]. The embedded method integrates feature selection within the model training process, allowing the algorithm to learn and assign higher importance to the most relevant features while minimizing the influence of less significant ones [8]. Hence, we proposed a new hybrid approach, which combines multiple selection techniques to leverage their respective strengths, enhancing model interpretability, performance, and mitigating overfitting [10].

While hybrid feature selection methods such as Lasso combined with sequential feature selection have been previously explored, this study does not aim to propose a fundamentally new algorithm. Instead, the contribution of this research lies in the contextual adaptation and empirical validation of a hybrid feature selection strategy on a large-scale, real-world banking dataset [10]. Unlike prior studies that focus on methodological novelty, this study emphasizes how the integration of embedded and wrapper methods can be effectively applied to high-dimensional and imbalanced financial data. The results demonstrate that such adaptation leads to a stable, interpretable, and generalizable model, which is critical for practical credit scoring applications.

Logistic Regression is retained in this study due to its strong interpretability and alignment with regulatory requirements commonly found in financial institutions [11]. In credit scoring applications, models must not only achieve high predictive performance but also provide transparent and explainable decision-making processes. Logistic Regression enables practitioners to clearly understand the relationship between input variables and credit risk, making it suitable for audit and compliance purposes [12].

Furthermore, feature selection is not only important for improving model performance but also provides significant business value [5]. By reducing the number of input variables, it simplifies model interpretation for stakeholders and supports regulatory transparency. In addition, fewer variables reduce data collection, storage, and processing costs, making the model more efficient and practical for real-world implementation in banking environments.

This study aims to optimize feature selection in credit scoring models using data from Bank XYZ, focusing on improving model accuracy and efficiency. The research method includes data collection, setting objective, cleansing data, feature selection, data splitting, inputting preprocessed data into the model, evaluation and validation [11]. This study will contribute to advancing feature selection and emphasize their importance.

2. RESEARCH METHOD

The research process begins with data collection, followed by defining the model's objective. Next, data cleaning is performed to ensure the quality and consistency of the dataset. Feature selection is then applied to identify the most significant variables. In this study, four feature selection strategies are employed: Wrapper, Embedded, Hybrid, and a baseline with no feature selection, chosen to represent a range of computational complexity and generalization ability. To minimize potential bias and enhance reproducibility, the dataset was randomly divided into training and testing subsets. The logistic regression model was fitted using the training data and subsequently evaluated on the test set to assess its predictive performance. Each feature selection approach is assessed using standard classification metrics, including accuracy, precision, recall, and F1-score [13]. The results obtained from these evaluations are analyzed and compared to determine the relative effectiveness of each feature selection method. A central hypothesis of this study is that the Hybrid method will yield superior results by combining statistical relevance and model-specific refinement. The final step involves validating the model's performance on new data [14]. Figure 1 illustrates the entire process.

2.1. Data collection

The first step involves obtaining the dataset for the eligibility scoring model. In addition to data acquisition, this phase also involves identifying the dependent variable (y) and independent variables (x) in the predictive model. It is essential to ensure that the dataset contains relevant features necessary for predicting the desired outcomes [1]. In this study, data is obtained from Bank XYZ and pertains to credit approvals, making it a credit scoring model [15]. In this model, the target variable (y) is credit approval decision and the other columns become the independent variables (x).

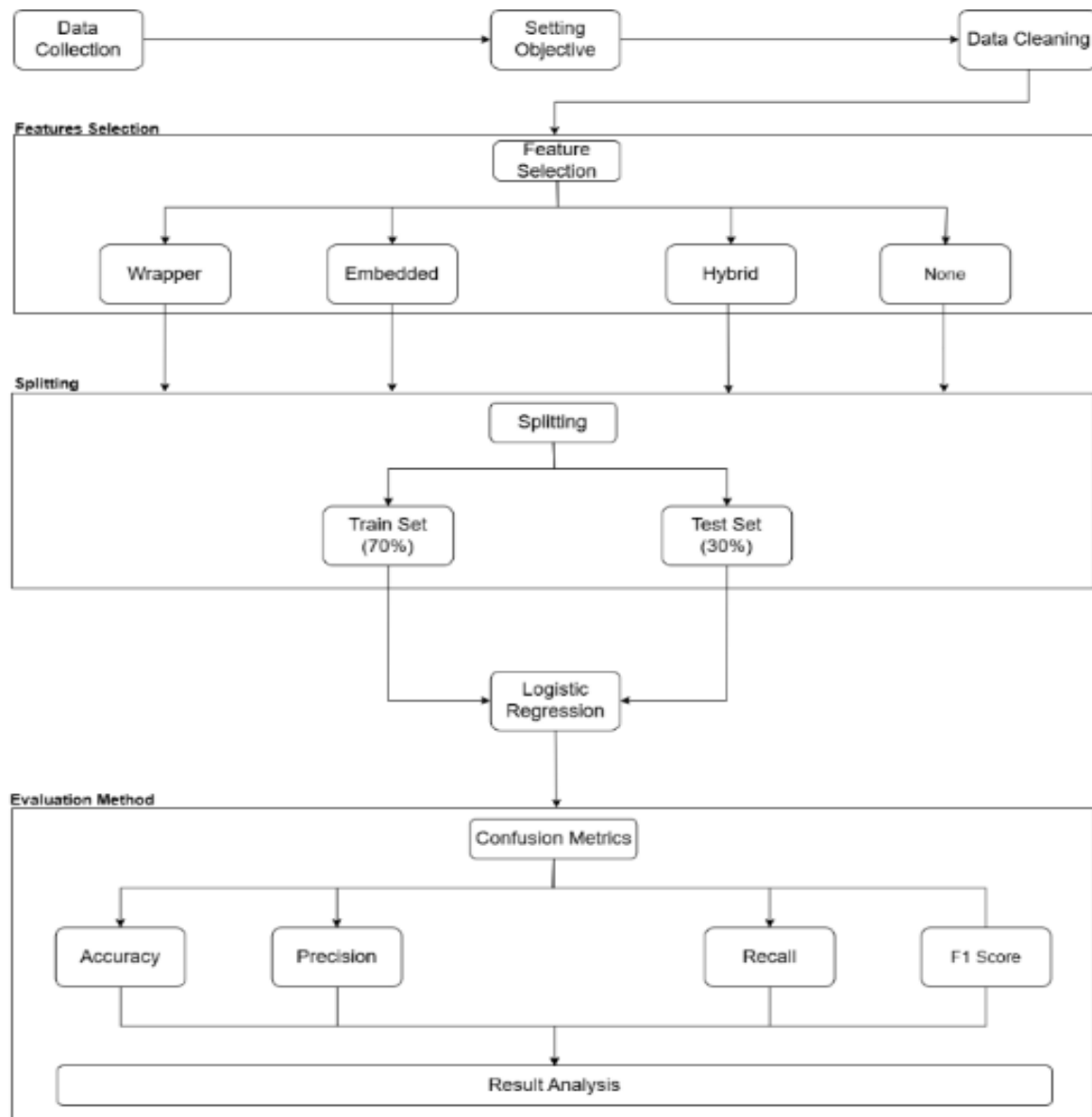


Figure 1. Research method

2.2. Setting objective

In this step, the classification criteria for creditworthiness are defined. Specifically, credit scores of 3, 4, and 5 are categorized as "Bad," while scores of 1 and 2 are considered "Good". These Good/Bad classifications serve as the target variable for training the model. The definitions of good and bad is flexible and can be adjusted based on specific needs and contexts [16].

2.3. Data cleaning

Data cleaning is performed to ensure of high quality and reliable input for the models. This step involves the handling of missing data, outliers, data errors and data types [10]. Additionally, filtering is applied to remove constant variables, which provide no useful information and duplicate variables, which introduce redundancy. Proper data cleaning, including these filtering steps, is necessary for the model to learn from accurate and representative data [1].

2.4. Feature selection

Feature selection is a critical process in machine learning. It involves selecting a relevant subset of input variables from a larger dataset [17]. The primary objective of feature selection is to identify the most

informative and significant features that contribute the most to the predictive ability of a machine learning model [18]. Several feature selection methods exist, including filter, wrapper, and embedded methods [19]. The filter method evaluates each feature individually using statistical techniques to determine its relevance level to the target variable. The filter method helps identify and remove constant, duplicate or highly correlated features in the dataset [10].

The wrapper method is an iterative approach that selects an optimal subset of variables for a predictive model [15]. In each iteration, a model is built using the selected variables and evaluated using metrics like accuracy or F1-score [20]. The process continues until stopping criteria are met. This method accounts for variable interactions and is effective for handling correlated or less impactful variables, ensuring more accurate and optimal feature selection [21]. In addition, it is good at removing redundant or weakly correlated features. This happens because, by testing the effects on a model's performance, this method can cut out all the variables which do not add anything to the model's predictive capability. However, it is time-consuming for large-scale data due to its computational inefficiency. Wrapper is greedy and could easily get stuck in a local optimum. The feature selection process would then be too dependent on validation performance, discard useful features too aggressively, and may cause underfitting. In contrast, the other shortcoming of the wrapper is it might cause overfitting. By repeatedly optimizing for validation performance, it can fit too closely to the training data, capturing noise and unnecessary features, which diminishes the model's ability to generalize to unseen data.

The embedded method integrates feature selection into the training process, allowing the algorithm to automatically select the most important features [22]. Significant features are given higher weights, while less important ones are disregarded [23]. However, this method may struggle with highly correlated features and may not always provide precise importance rankings [7]. A major weakness is over-penalization through Lasso regularization: it forces too many features to be zero, despite potentially having complementary value to other features. This may lead to underfitting as Lasso discards important features. In addition, its emphasis on evaluating features individually means that feature interactions or combined effects of features are overlooked, despite their potential to improve predictive accuracy. This limitation becomes more pronounced when features are highly correlated, since it typically retains only a single variable from the group and removes features that still carry informative signals, leading to underfitting. Furthermore, Lasso struggles to capture non-linear relationships between features and the target variable, limiting its capability to model complex patterns and increasing the risk of underfitting in such scenarios.

The hybrid method in this study combines embedded and wrapper feature selection techniques to improve both model performance and generalization. While embedded methods such as Lasso are effective in reducing dimensionality through regularization, wrapper methods such as Sequential Forward Selection (SFS) are capable of capturing interactions between features. By integrating these approaches, the hybrid method aims to balance statistical relevance and predictive performance [10].

In this implementation, the hybrid approach is conducted in two stages. First, L1-based Lasso regularization is applied to the dataset to eliminate less relevant features by shrinking their coefficients toward zero. The regularization strength is controlled to balance between feature sparsity and information retention. Second, Sequential Forward Selection (SFS) is applied to the reduced feature set. In this stage, features are added iteratively based on their contribution to model performance, using Logistic Regression as the base estimator. The selection process continues until no significant improvement in evaluation metrics is observed. This two-stage process ensures that the selected features are both statistically meaningful and empirically effective for prediction [23].

This study applies four feature selection methods: wrapper, embedded, hybrid, and no feature selection. The wrapper method uses SFS to iteratively select optimal features, while the embedded method applies Lasso (L1) regularization to eliminate less important features. The hybrid method, as the proposed approach, is compared against these conventional methods to assess performance. Each selected feature subset is tested using logistic regression, and the results are evaluated using a confusion matrix, accuracy, precision, recall, and F1-score.

2.5. Data splitting

The cleaned dataset is split into training and testing subsets using a Holdout validation approach with a 70:30 ratio. To ensure that the class distribution remains consistent across both subsets, stratified sampling is applied during the splitting process. This approach preserves the proportion of “good” and “bad” credit classes, thereby reducing sampling bias and ensuring a fair evaluation of model performance on unseen data. The training subset is used to build and optimize the model, while the testing subset evaluates its performance on unseen data, assessing how well the model generalizes to real-world situations.

2.6. Model creation

The logistic regression model is built using the training data to predict a binary outcome [20]. Logistic regression models the relationship between input variables (x) and the probability of an event (y) occurring, producing binary outputs (0 or 1) [15]. This is achieved through the sigmoid function, which maps a linear combination of the inputs and their respective weights to a probability value [15]. The resulting probability is then subjected to a threshold to determine the classification output, with adjustments made to balance between accuracy and minimizing false negatives [16].

Training and testing phases are both involved in the logistic regression process. The model is trained on labeled data, adjusting its weights to increase the accuracy of its predicted outcomes. Testing, on the other hand, is performed by applying the learned coefficients to new data. This allows the researcher to generate accurate and interpretable predictions that can be used in a wide range of applications by estimating the probability of a specific outcome.

During model creation, it is crucial to balance overfitting and underfitting [24]. Overfitting occurs when a model is too complex and "memorizes" the training data, capturing noise and outliers instead of general patterns, leading to low training error but high testing error. Addressing overfitting may involve adding data, regularization or reducing model complexity [7].

In contrast, underfitting occurs when the model is not sophisticated enough to identify underlying patterns, leading to high training and testing error [5]. Possible approaches to addressing underfitting include increasing model complexity or incorporating additional relevant features. Maintaining a balance between underfitting and overfitting is essential for reliable model performance. Common strategies used to achieve this balance include regularization and feature selection.

After the logistic regression model is built using the training data and passed the satisfactory evaluation, we can use it to make predictions on new data. By inputting the values of the input variables into the model, we can obtain the predicted values of the output variable or the probabilities.

To ensure reliable model performance, several implementation considerations are applied. Regularization is used in Logistic Regression to prevent overfitting. Class imbalance is handled through stratified sampling during data splitting to preserve class distribution. In addition, preprocessing steps include the removal of constant and highly correlated features to reduce noise and redundancy in the dataset.

2.7. Model evaluation

The evaluation method involves using a confusion matrix to test model performance and accuracy on the test data [19]. This matrix is a summary table that captures the model's correct and incorrect predictions by comparing predicted class labels with actual class labels. The confusion matrix has four essential components: true positive (TP), true negative (TN), false positive (FP) and false negative (FN) [25]. True positives represent the positive cases that are correctly classified and True Negatives is the cases correctly identified as negative. False Positives represent instances incorrectly classified as positive, even though they are negative and False Negatives denote cases incorrectly classified as negative despite being positive [24]. This approach constitutes the primary evaluation method used in this study, providing a standardized framework for assessing classification performance. Using these components of the confusion matrix, we derive key metrics that provide insights into the model's performance such as:

Accuracy is the overall correctness rate, calculated by (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is the proportion of true positive predictions among positive predictions, calculated by (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall or Sensitivity is the proportion of true positive predictions among actual positive instances, calculated by (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score represents the balance between precision and recall, calculated by (4).

$$F1 - score = 2 * \frac{precision*recall}{precision+recall} \quad (4)$$

Comparing the metric values from the selected feature-selection method with those from the conventional models will help clarify their relative performance. This evaluation supports consistent comparison across different model configurations and assists in identifying the most suitable model for the task. It also helps indicate areas that may require further improvement or closer attention.

2.8. Model validation

After building the model and selecting the best subset, it can be validated on new data to confirm it's optimal. Model validation assesses the model's generalization ability and accuracy with new instances, comparing predictions to actual outcomes [13]. Key evaluation metrics like accuracy, precision, recall and F1-score help determine the model's reliability. If the model performs satisfactorily, it is deemed suitable for real world applications [24].

In addition to the Holdout validation approach, 5-fold cross-validation is conducted to further evaluate model robustness. The dataset is divided into five subsets, and the model is trained and evaluated iteratively across different folds. The average accuracy across folds is used to provide a more reliable estimate of model performance and to reduce sensitivity to data partitioning.

2.9. Result analysis

After the evaluation phase, this study performs an analysis in a structured comparison of the outcomes from the applied feature selection methods. Accuracy, precision, recall and F1-score in the training and testing sets are analyzed, in order to evaluate the predictive quality. Test-train differences are further calculated and compared in order to get a sense for generalization, including over- and underfitting. Misclassification counts are further compared to aid the interpretation of reliability under class imbalance. Finally, selected features are briefly reviewed for relevance and diversity. The model with the best performing generalization is identified as most suitable based on this analysis.

3. RESULTS AND DISCUSSION

In this paper we are evaluating and comparing three of feature selection methods, wrappers, embedded and hybrid methods under one logistic regression model, to apply in credit scoring. The project aims to create a template for performing the feature selection effectively, for numerical data, to build consistent predictive models with no or very little overfitting [17]. The dataset is a credit scoring dataset from a banking information which consists of 66 columns and 153,588 rows, 40 of the features are constant and 11 are highly correlated features removed in the preprocessing phase. The target variable "Kualitas" is converted to two-class classification with "good" (1, 2) and "bad" (3, 4, 5) as approval and denial respectively.

The research employs three feature selection model wrappers, embedded and hybrid to identify key variables for accurate outcome prediction [18]. The wrapper method tests various feature subsets based on performance metrics, optimizing for the best combination [25]. The embedded method integrates feature selection directly into the model training, enhancing efficiency, though its effectiveness can vary across algorithms [21]. The hybrid approach combines filter, wrapper and embedded methods, leveraging their strengths to create a comprehensive and balanced selection process that optimizes model performance [26]. Each approach plays a role in controlling overfitting, with wrapper methods fine-tuning performance through evaluation-based selection, and embedded methods applying regularization to reduce variance. The hybrid model integrates these strengths to enhance both prediction accuracy and generalization.

The original dataset is split into a training dataset and a testing dataset, with the former used for learning the model parameters and the latter used for evaluating predictive accuracy. In this study, we use a 70/30 ratio of train data to test data is used. This allows the model to learn from a large enough sample of the data and to test its performance on unseen data. The choice of a logistic regression model was for the ease of use and its readabilities of interpretation which are two key features in a credit scoring model. Logistic regression is a linear model that uses the logistic function to model the probability of an event, making it a natural choice for binary classification tasks such as credit scoring, where the outcome is either "default" or "non-default". The accuracy of the model is measured by the percentage of correct predictions [25].

The results in Table 1 present a comparative evaluation of three feature selection approaches: the Wrapper method, the Embedded method, and the Hybrid method. The Evaluation was conducted to determine which technique offers the best generalization performance when applied to imbalanced, high-dimensional financial data. In the context of credit risk modeling, selected features must support both high model fidelity during training and sustained predictive reliability when applied to previously unseen data.

Although all three methods demonstrate strong performance across the evaluation metrics, the hybrid approach demonstrates the strongest generalization among all methods, with a minimal performance gap between training and testing sets and the fewest misclassifications, making it the most dependable choice for deployment in practical settings.

The Wrapper has a recall of 1.0000 for both the train and test set. Despite this, the small test-train precision gap (0.0018) indicates some slight overfitting. The Wrapper method assesses the effect of feature subsets on prediction performance directly. For this reason, it selected only one feature for portfolio classification.

Table 1. Comparative results of feature selection methods

Metric	Wrapper			Embedded			Hybrid			None		
	Train	Test	Difference	Train	Test	Differenc	Train	Test	Difference	Train	Test	Difference
Accuracy	0,9803	0,9793	0,0010	0,9840	0,9830	0,0010	0,9839	0,9840	0,0001	0,5004	0,5000	0,0004
Precision	0,9621	0,9603	0,0018	0,9700	0,9689	0,0011	0,9700	0,9696	0,0004	0,5002	0,5000	0,0002
Recall	1,0000	1,0000	-	0,9989	0,9980	0,0009	0,9986	0,9993	0,0007	1,0000	1,0000	-
F1 Score	0,9807	0,9798	0,0009	0,9842	0,9833	0,0009	0,9842	0,9842	-	0,6667	0,6667	-

We obtain a high classification accuracy but at the expense of an overfitted model with a test-train difference of 0.0174 in accuracy, 0.0069 in precision, and 0.0041 in F1-score. This result is limited by the intrinsic characteristic of the wrapper method that tends to overfit on the training data when performing feature selection, leading to narrow but non-generalisable feature subsets. More specifically, as the evaluation of the features is based on the out-of-sample predictive performance, the information potentially retained by selected features is solely based on the dataset at hand, thus creating oversimplified models that do not account for new or different inputs.

The Embedded method has an accuracy of 0.9830 and small test-train differences of 0.0009 in recall and F1-score, suggesting only a slightly overfitted model. The feature selection is performed directly in the training process of the classifier using regularization, and it selects 5 features that inform about the type of credit, estimated interest income, recency of the loan assessment, the type of region, and provisioning information. This results in a less narrow but still more focused set of features compared to the baseline model. However, being optimized towards the same training objectives, the selection may still encode non-perceivable biases present in the data at hand, thus slightly affecting the model's generalization performance. This result has shown a limitation of the embedded feature selection method as implemented in this study: despite improving efficiency by learning and selecting features simultaneously, its strong dependence on model-specific objectives limits its adaptability when faced with different or unseen data distributions.

The Hybrid method has the best performance with a test accuracy of 0.9840 and a training accuracy of 0.9839, and a small test-train difference of -0.0001. The recall is very high with a value of 0.9993 and the F1-score is the same for both sets at 0.9842. This method also shows the lowest number of misclassifications on the test set. These results can be explained by the two-stage feature selection process that combines a filtering approach in the form of Lasso regularization, an embedded method that penalizes multicollinearity and noise in the data, and Sequential Forward Selection, the wrapper method that ranks features according to their contribution to the predictive objective. The selected features are diverse and informative, covering different aspects of the financial and behavioral characteristics of the loans, such as interest income, credit amount, installment values, duration, and risk assessment. As a result, the model is comprehensive and less prone to overfitting or underfitting, meaning that it is both accurate and generalizable. The results obtained with the Hybrid approach indicate that it is not only well-balanced in terms of model complexity and relevance, but that it is also able to combine different strategies of selection in a synergistic manner, in order to further improve the model's generalization performance, especially when working with high-dimensional and imbalanced financial data.

The model with no feature selection has the worst performance overall on all evaluation metrics. The test-train differences are the lowest among all the methods but the absolute values in accuracy, precision, and F1-score are also the lowest, and there is a very high number of misclassifications on the confusion matrix. This poor performance can be explained by the lack of feature selection that leaves the model with all the available features, including irrelevant, redundant, or noisy ones. In other words, this results in a lot of noise that increases the model's variance and decreases the clarity of the predictions. These results highlight the importance of feature selection when working with complex and imbalanced data.

To further validate model performance, 5-fold cross-validation was conducted on the Logistic Regression model using the selected features. The results show relatively consistent accuracy scores across folds, with the mean performance comparable to the Holdout test results. This suggests that the model performance is not strongly affected by data partitioning, indicating reduced risk of optimistic bias in the evaluation.

Although the model achieves very high recall and minimal differences between training and testing performance, these results are interpreted with caution. The near-perfect recall may be influenced by the

dataset characteristics and class distribution. Therefore, model performance is evaluated not only through accuracy metrics but also through misclassification analysis and feature diversity to ensure that the model remains generalizable and reliable.

4. CONCLUSION

The purpose of this paper is to improve the credit evaluation and verification system used in the banking industry to determine the eligibility criteria for credit card approval. For this purpose, it is essential to select the most relevant features to ensure reliable and successful decisions. The four metrics of accuracy, precision, recall, and F1-score are used as the primary performance indicators for the ML models, and the results indicate that the Hybrid model was the most generalizable to unseen data.

The results for the model that did not perform feature selection were mainly intended to highlight the risks of using all available variables. This algorithm had the lowest performance and highest numbers of false positives and negatives, which can be directly attributed to the data's dimensionality, imbalance, and the presence of noise, redundant, and unrelated information. This clearly indicates that targeted feature selection is a useful technique for removing ambiguity from the data.

It is visible that the three methods have significant advantages and disadvantages: the high recall of the Wrapper method directly impacts the low performance and generalization of the results. Since this algorithm only had one feature in the final selection, the training results were subject to a single important influence and had a low level of bias. The embedded method has a greater number of features, and while some bias is also visible from the training data, the predictive and related features improve the final balance of the two metrics. Both of these drawbacks and advantages are present in the hybrid method, as it uses the Embedded technique for a quick initial selection and the Wrapper method for a secondary and thorough sequential elimination of the least predictive variables. Thus, the obtained model has a sufficient number of features that are diverse and contextually unique, making this algorithm the most successful and the most deployable in real-world credit scoring.

In addition to its strong predictive performance, the proposed Hybrid model demonstrates practical applicability in real-world banking environments. By leveraging logistic regression, the model remains highly interpretable and aligned with regulatory expectations, enabling transparent decision-making. Furthermore, the reduced number of selected features contributes to computational efficiency and lowers data collection costs. Therefore, this model is not only accurate but also implementable, making it a viable decision-support tool for credit approval processes in financial institutions.

This study focuses primarily on the predictive modeling aspect of credit scoring and does not explicitly incorporate regulatory frameworks such as weight of evidence (WOE), information value (IV), or fairness constraints. However, the use of Logistic Regression supports interpretability and aligns with common regulatory expectations in financial institutions. Future research may extend this work by integrating explainability techniques and fairness considerations.

ACKNOWLEDGMENTS

This paper also contains proprietary data, including a dataset and materials under a non-disclosure agreement (NDA). For that reason, the dataset cannot be distributed to the public. If there are any data inquiries for academic purposes, please send the inquiry to the author via brigitta.permadhi@binus.ac.id.

The authors would like to acknowledge the use of artificial intelligence tools for paraphrasing and light debugging assistance during the preparation of this paper. All final interpretations, analyses, and conclusions are the sole responsibility of the authors. The prompts used were: i) Please paraphrase the following paragraph to make it clear and understandable and maintain the same meaning. ii) Please help me improve the fluency and coherence of this part. iii) Please help improve code clarity and add comments.

FUNDING INFORMATION

The authors state that no funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Brigitta Permadhi	✓	✓	✓		✓	✓	✓		✓	✓	✓			
Devi Fitrianah	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓

C : **C**onceptualizationM : **M**ethodologySo : **S**oftwareVa : **V**alidationFo : **F**ormal analysisI : **I**nvestigationR : **R**esourcesD : **D**ata CurationO : Writing - **O**riginal DraftE : Writing - Review & **E**editingVi : **V**isualizationSu : **S**upervisionP : **P**roject administrationFu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [initials, BP]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

- [1] T. Kobayashi *et al.*, "Development of a scoring model based on objective factors to predict gangrenous/perforated appendicitis," *BMC Gastroenterology*, vol. 23, no. 1, 2023, doi: 10.1186/s12876-023-02767-7.
- [2] O. P. Rud, "Data mining cookbook: modeling data for marketing, risk, and customer relationship management," *John Wiley & Sons*, 2001.
- [3] R. Damiata and D. Fitrianah, "Prediction analysis of Kartu Jakarta Pintar (KJP) awardees in vocational high school XYZ using C4.5 algorithm," *International Journal of Machine Learning and Computing*, vol. 10, no. 1, pp. 44–50, Jan. 2020, doi: 10.18178/ijmlc.2020.10.1.896.
- [4] M. Momeni, M. Rashidifar, F. H. Balam, A. Roointan, and A. Gholaminejad, "A comprehensive analysis of gene expression profiling data in COVID-19 patients for discovery of specific and differential blood biomarker signatures," *Scientific Reports*, vol. 13, no. 1, Apr. 2023, doi: 10.1038/s41598-023-32268-2.
- [5] J. Li *et al.*, "Feature selection: a data perspective," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–45, Dec. 2017, doi: 10.1145/3136625.
- [6] R. Salman, A. Alzaatreh, and M. T. Al Bataineh, "Feature selection of the respiratory microbiota associated with asthma," *Journal of Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00767-8.
- [7] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, Springer Berlin Heidelberg, pp. 315–324.
- [8] V. Kumar, "Feature selection: a literature review," *The Smart Computing Review*, vol. 4, no. 3, 2014, doi: 10.6029/smartcr.2014.03.007.
- [9] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26, Mar. 2019, doi: 10.2478/cait-2019-0001.
- [10] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.
- [11] D. A. V. de Paula, R. Artes, F. Ayres, and A. M. A. F. Minardi, "Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques," *RAUSP Management Journal*, vol. 54, no. 3, pp. 321–336, 2019, doi: 10.1108/rausp-03-2018-0003.
- [12] R. Gai and H. Zhang, "Prediction model of agricultural water quality based on optimized logistic regression algorithm," *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, Feb. 2023, doi: 10.1186/s13634-023-00973-9.
- [13] D. Fitrianah, "Feature exploration for prediction of potential tuna fishing zones," *International Journal of Information and Electronics Engineering*, 2015, doi: 10.7763/ijeee.2015.v5.543.
- [14] M. J. Ershadi and D. Omidzadeh, "Customer validation using hybrid logistic regression and credit scoring model: a case study," *Calitatea*, vol. 19, no. 167, pp. 59–62, 2018.
- [15] Z. Ye, H. Zhang, J. Liang, S. Yi, and X. Zhan, "Significance of logistic regression scoring model based on natural killer cell-mediated cytotoxic pathway in the diagnosis of colon cancer," *Frontiers in Immunology*, vol. 14, Jan. 2023, doi: 10.3389/fimmu.2023.1117908.
- [16] A. Rahman, "Statistics-based data preprocessing methods and machine learning algorithms for big data analysis," *International Journal of Artificial Intelligence*, vol. 17, no. 2, pp. 44–65, 2019.
- [17] M. Bannasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, Dec. 2015, doi: 10.1016/j.eswa.2015.07.007.
- [18] J. Subramanian and R. Simon, "Overfitting in prediction models – Is it a problem only in high dimensions?," *Contemporary Clinical Trials*, vol. 36, no. 2, pp. 636–641, Nov. 2013, doi: 10.1016/j.cct.2013.06.011.
- [19] O. Mehrpour *et al.*, "Comparison of decision tree with common machine learning models for prediction of biguanide and sulfonylurea poisoning in the United States: an analysis of the National Poison Data System," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, Apr. 2023, doi: 10.1186/s12911-022-02095-y.
- [20] K. H. Hui, C. S. Ooi, M. H. Lim, M. S. Leong, and S. M. Al-Obaidi, "An improved wrapper-based feature selection method for machinery fault diagnosis," *PLOS ONE*, vol. 12, no. 12, p. e0189143, Dec. 2017, doi: 10.1371/journal.pone.0189143.

- [21] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy," *Pertanika Journal of Science and Technology*, vol. 26, no. 1, p. 329, 2018.
- [22] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/j.jksuci.2019.06.012.
- [23] X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, and H.-S. Ye, "A hybrid genetic algorithm with wrapper-embedded approaches for feature selection," *IEEE Access*, vol. 6, pp. 22863–22874, 2018, doi: 10.1109/access.2018.2818682.
- [24] J. Liu *et al.*, "Development and validation of predictive model based on deep learning method for classification of dyslipidemia in Chinese medicine," *Health Information Science and Systems*, vol. 11, no. 1, Apr. 2023, doi: 10.1007/s13755-023-00215-0.
- [25] N. Kushmerick, "Wrapper induction: Efficiency and expressiveness," *Artificial Intelligence*, vol. 118, no. 1–2, pp. 15–68, Apr. 2000, doi: 10.1016/s0004-3702(99)00100-9.
- [26] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997, doi: 10.1016/s0004-3702(97)00043-x.

BIOGRAPHIES OF AUTHORS



Brigitta Permadhi    has a Bachelor of Computer Science (S.Kom.), with the class of 2024 and a Master of Computer Science (M.Kom.) from Bina Nusantara University (Binus), Indonesia. Brigitta Permadhi's field of study is computer science with an emphasis on data science and natural language processing (NLP). Her research interests include data science, natural language processing, and applications within the field of computer science. She has been involved in research projects or academic work related to hybrid approaches for text summarization using BERT and T5 models, utilizing the transformers library in Python. She can be contacted at email: brigitta.permadhi@binus.ac.id.



Devi Fitriana    is a lecturer and researcher in Master of Computer Science Department in Bina Nusantara University (Binus), Indonesia. She received Bachelor's degree in Computer Science major in Computer Science from Bina Nusantara University in 2000, Master's degree in Information Technology from Universitas Indonesia in 2008, and Doctor of Philosophy (Ph.D.) in Computer Science from Universitas Indonesia in 2015. She received research placement (Sandwich Program) in Laboratory for Pattern Recognition and Image Processing and GIS (PRIPGIS Lab), Department of Computer Science, Michigan State University, East Lansing, Michigan, United States, in 2014. She is currently active in the lecturing and researches in the field of data mining, big data analytics, and artificial intelligence. The courses she teaches include data mining, big data analytics, and topics in artificial intelligence. Her research interests include data mining, machine learning, artificial intelligence, and applied remote sensing. She can be contacted at email: devi.fitriana@binus.ac.id.