Malaysian Journal of Computing, 9 (1): 1722-1751, 2024

eISSN: 2600-8238

MACHINE LEARNING-BASED APPROACHES FOR CREDIT CARD DEBT PREDICTION

Nurain Ibrahim[1*], Umi Munirah Ishak[2], Nur Nabilah Arina Ali[3], and Norshahida Shaadan[4]

[1,4]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*
[1]Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*
[2]D'Monte Laguna Merbok Sungai Petani 31, Persiaran BLM 1A, Bandar Laguna Merbok, 08000 Sungai Petani, Kedah, Malaysia
[3]Fresenius Medical Care, Axis Technology Centre, 2nd Floor, Lot Petaling Jaya, Jalan 51A/225, Seksyen 13, 46100 Petaling Jaya, Selangor, Malaysia
[1*]nurainibrahim@uitm.edu.my, [2]umishak98@gmail.com, [3]nurnabilaharina@gmail.com, [4]norshahida588@uitm.edu.my

ABSTRACT

The primary concern in the stock market and banks that offer credit cards has been a problem over time. Regardless of their capacity to pay, most card users abuse their credit cards and accrue debt from cash cards. The most significant issue facing cardholders and banks alike is this calamity. Predicting credit card customers' default payments became vital to lowering this risk. Data mining approaches, including decision tree, logistic regression, and Naïve Bayes *with feature selection methods, were applied to secondary credit card debt data to identify the significant factors that impact credit card default and to enhance the prediction of credit card default. As a result, the decision tree with Gini index splitting criteria forward selection wrapper method was identified as the best model with the highest percentages of accuracy, precision, sensitivity, and area under ROC of 76.39%, 72.02%, 85.08%, and 0.891 respectively. Additionally, the significant factors that impact credit card default are gender, education level, repayment status in July 2005, repayment status in August 2005, status of repayment in September 2005, and the amount paid in June 2005 and May 2005. This study may help financial institutions assess creditworthiness and give consumers insights into their financial behaviors.*

Keywords*: Credit Card Debt, Decision Tree, Logistic Regression, Naïve Bayes*

Received for review: 13-03-2024; Accepted: 27-03-2024; Published: 01-04-2024
DOI: 10.24191/mjoc.v9i1.25656

1. Introduction

A credit card is a modern and beneficial financial product that works on the principle of electronic money. The credit card has become a currency transaction system's most important market phenomenon. Credit cards enable customers to pay for products and services at retailers that accept cards. Banks issue various types of credit cards, such as Visa and Mastercard. Nowadays, credit cards are a part of everyday life. People prefer to use credit cards rather than cash to make payments.

According to Statista, by the end of 2020, the total number of credit card transactions in Taiwan from 2008 to 2020 will exceed three trillion dollars. A credit card default happens if the customer fails to pay the minimum amount required for several months. Credit card default can be purposeful or not for various reasons, such as client health, business loss, and employment. Furthermore, a credit card default might result in bank losses. One of the primary forecasts that banks worry about is credit card default, along with credit scoring, to help them understand why consumers are more likely to fail. The default prognosis is based on the customer's credit history and the bankers' experiences. Credit default prediction aims to assist lenders in making loan decisions for their clients. Numerous defaulters will result in a loss of revenue and cash flow problems. Researchers have studied credit card default prediction using data mining techniques. In research by Thomas (2000), discriminant analysis was employed to grade borrower behaviors and credits. Yeh and Lien (2009) compared the predictions of various algorithms. They employed discriminant analysis, logistic regression, Bayes classifier, closest neighbor, artificial neural network, and classification tree to forecast consumer default payments in Taiwan.

Regardless of their capacity to pay, most card users abuse their credit cards and accrue debt from cash cards. The most significant issue facing cardholders and banks alike is this calamity. Taiwan's card issuers are now facing a cash and credit debt problem. Predicting credit card customers' default payments became vital to lowering this risk. Risk prediction models have been developed using a variety of statistical approaches, including discriminant analysis, Bayes classifiers, logistic regression, nearest neighbour, artificial neural networks, and classification methods. Credit risk refers to the potential for delayed payment of extended credit. Using different data mining methods, the results of all the analyses may be found immediately. The computational process known as "data mining" uses techniques from database systems, machine learning, and statistics to extract functional patterns from massive data sets.

Furthermore, there is a shortage of research on the variables influencing credit card default. We attempted to use a decision tree, which is a more understandable approach, in this investigation. Consequently, we evaluate decision trees, logistic regression, and naïve bayes to determine which model performs best for a default credit card.

2. Related Works

Subasi and Cankurt (2019) claimed that random forest is an excellent alternative to forecast the payment default precisely. The study indicates that the model was developed using the Synthetic Minority Over-Sampling Technique (SMOTE) method. Oversampling the minority class in the unbalanced dataset creates synthetic cases. According to the results, Random Forest's SMOTE yields a maximum accuracy of 89.01%. Area Under Receiver Operating Characteristic (AUROC) also acts as one of the indicators for checking model assessment.

Additionally, this impact increased the random forest's AUROC of 0.947, which was greater than that of other classifiers. Furthermore, research by Sayjadah et al. (2018) on the prediction of credit card default using machine learning approaches demonstrates that random forest has a greater accuracy when determining the credit risk of credit card consumers. Three machine learning methods were examined in this study: random forest, logistic regression, and partial decision trees. For logistic regression, part decision tree, and random forest, the corresponding area under the curve (AUROC) values are 75%, 64%, and 77%. The AUROC score is more significant for the random forest.

Yang and Zhang (2018) found that in LightGBM research comparing several data mining techniques for credit card default prediction, LightGBM performs well in predicting the categorical response variable. The study compares logistic regression, SVM, neural networks, Xgboost, and LightGBM to forecast credit card default. In the research, ten-fold cross-validation was used. LightGBM has the most remarkable accurate rate of 89.29% compared to other techniques, suggesting a solid classification impact. Bhattacharyya et al. (2011) also state that logistic regression constantly has good AUROC performance. In the training data, the AUROC of random forests and support vector machines decreases as fraud rates decrease. The effectiveness of support vector machines, random forests, and logistic regression for credit card fraud detection is investigated in this paper. A data set on credit card transactions from January 2006 to January 2007 was utilised in this investigation. Hassan (2020) studied of artificial neural network (ANN) credit card default prediction demonstrates how effectively ANN fared in this regard. A data collection of 40% of testing and 60% of training, consisting of 30,000 bank clients with credit cards, was employed for this investigation. The finding revealed that the Round Mean Square Error (RMSE) is 0.37, and the accuracy is 79%. This study suggested that other methods should be implemented for the credit card debt data to improve the precision and accuracy of the model. Hence, this study implements three methods: logistic regression, decision tree, and naïve bayes.

3. Methodology

This section consists of the methodology of this study, including data acquisition in Section 3.1, data preparation in Section 3.2, data modelling in Section 3.3, and model assessment in Section 3.4.

3.1 Data Acquisition

The credit card data is secondary data taken from an internet web, namely the UCI Machine Learning Repository, which contains 24 features and 30000 sample sizes. Table 1 provides a description of the data.

Table 1. Data Description

Variable	Description	Role
Limit_Bal	Credit amount (NT dollars)	Input
Gender	Gender (1=male, 2=female)	Input
Education	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown)	Input
Marital Status	Marital status (1=married, 2=single, 3=others)	Input
Age	Age in years	Input
Pay_0	Status of Repayment in September 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)	Input
Pay_2	Status of Repayment in August 2005 (scale same as above)	Input
Pay_3	Status of Repayment in July 2005 (scale same as above)	Input
Pay_4	Status of Repayment in June 2005 (scale same as above)	Input
Pay_5	Status of Repayment in May 2005 (scale same as above)	Input
Pay_6	Status of Repayment in April 2005 (scale same as above)	Input
Bill_Amt1	Bill statement amount in September 2005 (NT dollars)	Input
Bill_Amt2	Bill statement amount in August 2005 (NT dollars)	Input
Bill_Amt3	Bill statement amount in July 2005 (NT dollars)	Input
Bill_Amt4	Bill statement amount in June 2005 (NT dollars)	Input
Bill_Amt5	Bill statement amount in May 2005 (NT dollars)	Input
Bill_Amt6	Bill statement amount in April 2005 (NT dollars)	Input
Pay_Amt1	Previous payment amount in September 2005 (NT dollars)	Input

Pay_Amt2	Previous payment amount in August 2005 (NT dollars)	Input
Pay_Amt3	Previous payment amount in July 2005 (NT dollars)	Input
Pay_Amt4	Previous payment amount in June 2005 (NT dollars)	Input
Pay_Amt5	Previous payment amount in May 2005 (NT dollars)	Input
Pay_Amt6	Previous payment amount in April 2005 (NT dollars)	Input
Default_payment_next_month	Default payment (1=yes, 0=no)	Output

3.2 Data Preparation

Data preparation is a crucial step in the data analysis and machine learning process. The data collected from the actual data is rarely perfect or pristine. It often contains various issues that need to be addressed during this phase. Feature selection also falls under the data preparation stage. Specifically, forward selection, backward elimination, and optimize selection wrapper feature selection methods were used in this study to enhance the prediction of credit card debt. The ratio of data splitting used in this study is 70:30, where 70% is used for training, and 30% is used for testing the model (Shafie, Peng Ooi, and Khaw, 2023).

3.3 Data Modelling

The data mining approaches applied to the data include decision trees, logistic regression, and naïve bayes and act as classification methods.

3.3.1 Decision Tree

Two common uses of the decision tree approach are creating categorization schemes based on several factors or creating prediction algorithms for a target variable. This method creates an inverted tree with root, internal, and leaf nodes by splitting a population into sections that resemble branches (Chong et al., 2023). The non-parametric technique may handle large, complicated datasets without a problematic parametric framework. Data may be divided into training and validation datasets after the study's sample size is sufficiently large. To create the best possible final model, a decision tree model is built using the training dataset, and the validation dataset is used to determine the appropriate tree size.

3.3.2 Logistic regression

Logistic regression is a valuable approach when assessing a categorical answer variable to get a binary result, such as Yes or No, regarding default payment (Ibrahim and Kamarudin, 2023). A linear function of the observed values of the available response variables is known as a logistic regression. Probability is used in logistic regression to assess the connection between the independent and dependent variables. This method's primary benefit is the ability to create an easy-to-understand classification probability calculation. The credit card client's default payment, represented as "0," meaning no, and "1, meaning yes," is the binary result in this research. Logistic regression analysis was used to identify the relevant factors present in the dataset.

3.3.3 Naive Bayes

Based on Thomas Bayes' theorem and predicated on predictor independence, Naïve Bayes is a statistical approach (Mansur Huang, Ibrahim, and Mat Diah, 2021). The Naïve Bayes classifier assumes that a feature's existence in a class is independent of the existence of any other features. The posterior probability is the likelihood of a Bayes theorem. Using the numbers from (c), $P(x)$, and $P(x|c)$, the posterior probability, $P(c|x)$, may be calculated using the Bayes theorem. The Naïve Bayes classifiers assume that the influence of a predictor's x, value on a particular class (c) is unaffected by the values of other predictors.

3.4 Model Assessment

The data mining approaches are assessed using accuracy, sensitivity, specificity, precision, and area under ROC. These assessment indicators were based on the confusion matrix as shown in Table 2 (Chong et al., 2023; Noh et al., 2023):

Table 2. Data Description

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False negative (FN)	True Positive (TP)

3.4.1 Accuracy

The proportion of accurately predicted occurrences to all instances in the dataset is known as accuracy. The formula for accuracy is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

3.4.2 Sensitivity

The ratio of genuine positive predictions to all positive cases is computed to compute sensitivity. The equation is:

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

3.4.3 Specificity

The ratio of accurate negative predictions to all real negative cases is used to assess specificity. The equation is:

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

3.4.4 Precision

Precision, computed as the ratio of true positive predictions to the total projected positive cases, is sometimes called positive predictive value. The equation is:

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

3.4.5 Area under ROC (AUROC)

AUROC is a statistic used to assess a binary classification model's performance. The ROC curve represents the trade-off between a true positive rate (sensitivity) and a false positive rate at different thresholds.

4. Results

The analyses were conducted using Rapidminer software. Data pre-processing was involved, including checking for missing values and duplicate samples. There are no missing values, and duplicate samples exist in the data. Models of machine learning algorithms named decision trees with different splitting criteria such as Gini index, gain ratio and information gain, logistic regression, and Naïve Bayes with wrapper feature selection such as forward selection, backward elimination, and optimize selection were implemented in this study.

4.1 Descriptive Analysis of Demographic Profiles

Descriptive analysis involves the interpretation and summary of data, which aims to provide an overview of the main characteristics and patterns within the data. The gender, education level, and marital status distributions of the respondents are displayed in Figures 1 to 3. As shown in Figure 1, 18112 (63.11%) respondents are Female, and the remaining 11888 (39.63%) are Male. According to Figure 2, which displays the respondents' educational backgrounds, 14030 (47.31%) studied at a university, followed by 10585 (35.69%) in graduate school and 4917 (16.58%) in high school.

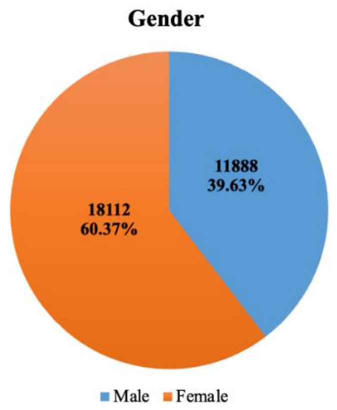

Figure 1. Gender Distribution

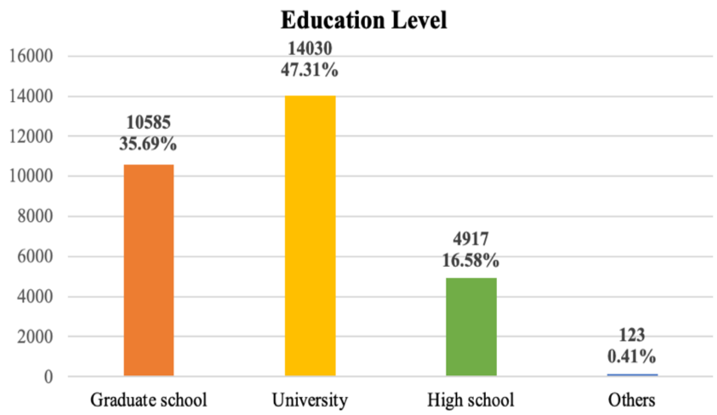

Figure 2. Education Level Distribution

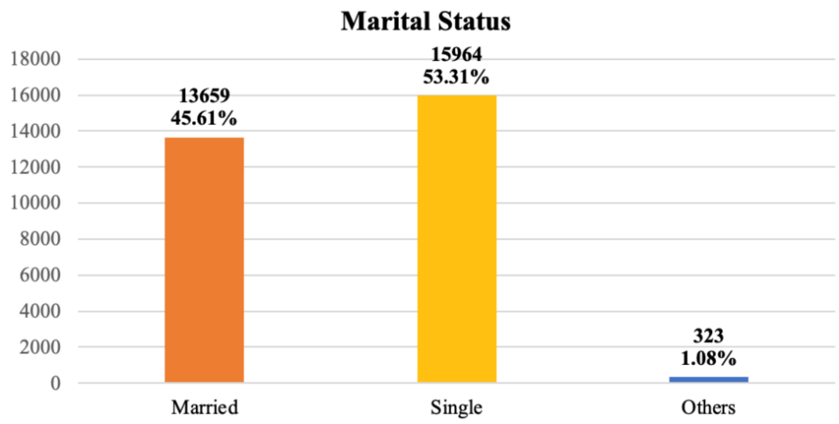

Figure 3. Marital Status Distribution

4.2 Model performances of machine learning approaches

This analysis is run to compare the performance of the three predictive models using a decision tree, logistic regression, and Naïve Bayes in determining the best predictive method for predicting credit card default. This study used three different splitting criteria: the Gini index, gain ratio, information gain, and accuracy for the decision tree. Besides, for logistic regression and Naïve Bayes, this study used three wrapper feature selection methods such as forward selection, backward elimination, and optimize selection. Table 4 displays the percentages of classification accuracy, precision, sensitivity, specificity, and AUROC values. Meanwhile, Figure 4 shows all percentages of model performances except AUROC. According to the results for the decision tree with different splitting criteria, the Gini index has the highest values of accuracy, precision, and specificity, which are 75.58%, 73.20%, and 70.44%, respectively. The gain ratio has the highest sensitivity value, 90.31%, AUROC is 0.942, while the Gini index has the lowest sensitivity value, 60.72%. The output shows that the Gini index has the best performance for decision tree since it has the highest value of accuracy, precision, and specificity.

Once the best splitting criteria were determined, the feature selection method was applied to the credit card data. As seen from Table 4, forward selection has the best performance for the decision tree with the Gini index since it has the highest value of accuracy, precision, and AUROC, 76.39%, 72.02%, and 0.891, respectively. Optimize selection has a higher value on sensitivity with 88.63% and specificity with 64.05% compared to backward elimination and optimize selection.

Meanwhile, for logistic regression supervised machine learning algorithm, backward elimination is the best because it has the highest values on precision at 67.72%, specificity at 66.34%, and AUROC with 0.740. The backward elimination accuracy value is lower by 5.25% compared to the optimize selection. The output shows that backward elimination has the best performance for logistic regression since it has the highest precision, specificity, and AUROC value. Additionally, for the Naïve Bayes model, optimize selection has the highest values on accuracy, precision, and sensitivity, which are 71.25%, 67.50%, and 60.55%, respectively. The specificity value for forward selection is higher than backward elimination and optimize selection, which is 84.31% compared to backward elimination and optimize selection. The output shows that optimized selection performs best for Naïve Bayes since it has the highest value of accuracy, precision, and sensitivity.

As we compared all models, we concluded that the decision tree with a Gini index using forward selection is the best since it has the highest values on the accuracy, precision, sensitivity, and AUROC, which are 76.39%, 72.02%, 85.08%, and 0.891 respectively. Naïve

Bayes with optimized selection has the highest value on specificity with 81.95% compared to decision tree and logistic regression. From the output, it can be concluded that a decision tree with a Gini index using forward selection is the best compared to logistic regression and Naïve Bayes.

Table 4. Model performances of all machine learning approaches

Model	Accuracy	Precision	Sensitivity	Specificity	AUROC
Decision Tree – Gini Index splitting criteria	75.58	73.20	60.72	70.44	0.836
Decision Tree – Gain Ratio splitting criteria	68.40	62.79	90.31	46.48	0.942
Decision Tree – Information Gain splitting criteria	74.61	69.49	87.74	61.48	0.843
Decision Tree – Gini Index – Forward Selection	76.39	72.02	85.08	59.57	0.891
Decision Tree – Gini Index – Backward Elimination	75.67	70.68	87.74	63.60	0.843
Decision Tree – Gini Index – Optimize Selection	76.34	71.14	88.63	64.06	0.845
Logistic Regression – Forward Selection	69.67	64.77	86.27	53.07	0.720
Logistic Regression – Backward Elimination	64.48	67.72	70.61	66.34	0.740
Logistic Regression – Optimize Selection	69.73	66.29	80.28	59.18	0.724
Naïve Bayes – Forward Selection	70.44	66.00	56.57	84.31	0.722
Naïve Bayes – Backward Elimination	70.43	66.94	50.14	80.72	0.743
Naïve Bayes – Optimize Selection	71.25	67.50	60.55	81.95	0.741

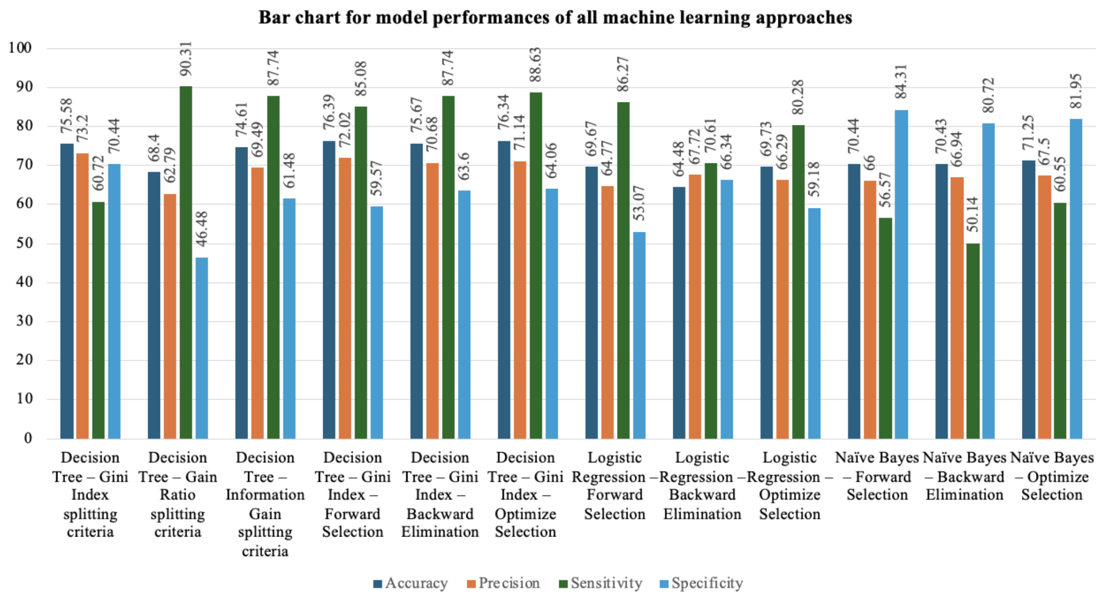

Figure 4. Bar chart for model performances of all machine learning approaches

4.3 Identification of significant variables that impact credit card

As mentioned in the previous section, the best wrapper feature selection when it was wrapped with a decision tree, logistic regression, and Naïve Bayes are displayed in this section. Table 5 depicts the significant factors impacting all respective models' credit cards. Value "1" indicates the variable is a significant variable obtained from that model, and value "0" indicates otherwise. From Table 5, gender, education, and Repayment status in September, August, and July 2005, the mentioned models determine the amounts of previous payments in June and July 2005 (NT dollars).

Table 5. Significant variables that impact credit card

Variable	Decision Tree – Gini Index – Forward Selection	Logistic Regression – Backward Elimination	Naïve Bayes – Optimize Selection
Limit_Bal	0	1	0
Gender	1	1	1
Education	1	1	1
Marital Status	1	1	0
Age	1	0	1
Pay_0	1	1	1
Pay_2	1	1	1
Pay_3	1	1	1
Pay_4	1	0	1
Pay_5	1	1	0
Pay_6	1	1	0
Bill_Amt1	0	0	0
Bill_Amt2	0	1	0
Bill_Amt3	0	1	0
Bill_Amt4	0	1	0
Bill_Amt5	0	1	0
Bill_Amt6	0	1	0
Pay_Amt1	1	1	0
Pay_Amt2	0	1	0
Pay_Amt3	0	0	0
Pay_Amt4	1	1	1
Pay_Amt5	1	1	1
Pay_Amt6	0	1	0

5. Discussion

Secondary credit card data from Kaggle was examined in this study using machine learning-based models. All analyses were analysed using RapidMiner software. In conclusion, the decision tree using three splitting criteria, which are Gini index, gain ratio, and information gain, were examined in this study. Additionally, wrapper feature selection methods were wrapped with a decision tree, logistic regression, Naïve Bayes, forward selection, backward elimination, and optimize selection. The output shows that the decision tree with Gini index splitting wrapped with forward selection is the best model since it has the highest percentages of performance evaluation compared to other models. These findings are consistent with those of Afriyie et al. (2023), who discovered a decision tree as the best model for credit card prediction. This study's findings have the same conclusion as Joshi et al. (2021), where both studies concluded that the decision tree is the best model for credit card fraud detection compared to other machine learning approaches. In terms of significant variables, this study found a few significant variables, including gender and education level. This finding is supported by the previous research by Gan and Maysami (2006), which found that education level is related to credit card selection criteria. Hence, the identified significant variables may help the bankers identify the target customers affected by credit card default. Meanwhile, Xu et al. (2022) also supported the findings that gender affects customers' credit card possession.

6. Conclusion and Recommendation

In conclusion, for the decision tree using four splitting criteria, which are Gini index, gain ratio, information gain, and accuracy, the output shows that the Gini index is the best performance since it has the highest accuracy, precision, and specificity values. This study uses feature selection for the Gini index, and the result found that forward selection has the best performance since it has the highest value of accuracy, precision, and AUROC. For logistic regression, backward elimination is the best because it has the highest precision, specificity, and AUROC value. Naïve Bayes with optimize selection is the best because it has the highest values on accuracy, precision, and sensitivity. A decision tree with a Gini index using forward selection is the best for comparing prediction models since it has the highest accuracy, precision, sensitivity, and AUROC values. To summarize, a decision tree with a Gini index using forward selection is the best compared to logistic regression and Naïve Bayes. The most significant factors in this study are gender, education, Repayment status in September, August, and July 2005, and the number of previous payments in June and July 2005 (NT dollars). The significant variables contribute to all the models. As a recommendation, a decision tree should be used to score clients rather than other data mining techniques like logistic regression. Therefore, future studies should focus on other factors that might affect the default credit card since clients from other countries might have different preferred characteristics or factors that could affect the default of credit cards.

Acknowledgement

The authors are thankful to the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Research Nexus UiTM (ReNeU), College of Computing, Informatics and Mathematics UiTM, and Research Management Centre (RMC) UiTM for the support of this research publication.

Funding

The author(s) received no specific funding for this work.

Author Contribution

Umi Munirah Ishak and Nur Nabilah Arina Ali prepared the literature review and conducted the statistical analysis. Nurain Ibrahim oversaw the article writing and research methodology. Nurain Ibrahim and Norshahida Shaadan interpreted the results.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Afriyie, Jonathan Kwaku, Kassim Tawiah, Wilhemina Adoma Pels, Sandra Addai-Henne, Harriet Achiaa Dwamena, Emmanuel Odame Owiredu, Samuel Amening Ayeh, and John Eshun. 2023. "A Supervised Machine Learning Algorithm for Detecting and Predicting Fraud in Credit Card Transactions." *Decision Analytics Journal* 6(January):100163. doi: 10.1016/j.dajour.2023.100163.

Bhattacharyya, Siddhartha, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. 2011. "Data Mining for Credit Card Fraud: A Comparative Study." *Decision Support Systems* 50(3):602–13. doi: 10.1016/j.dss.2010.08.008.

Chong, Angela Yi Wen, Khai Wah Khaw, Wai Chung Yeong, and Wen Xu Chuah. 2023. "Customer Churn Prediction of Telecom Company Using Machine Learning Algorithms." *Journal of Soft Computing and Data Mining* 4(2):1–22. doi: 10.30880/jscdm.2023.04.02.001.

Gan, Lydia L., and Ramin Cooper Maysami. 2006. "Credit Card Selection Criteria: Singapore Perspective." *Economic Growth Centre Working Paper Series*.

Hassan, Malik Mubasher. 2020. "Credit Card Default Prediction Using Artificial Neural Networks." *GIS Science Journal* 7(7):383–90.

Ibrahim, Nurain, and Adina Najwa Kamarudin. 2023. "Assessing Time-Dependent Performance of a Feature Selection Method Using Correlation Sharing T-Statistics (CorT) for Heart Failure Data Classification." *AIP Conference Proceedings* 2500(February). doi: 10.1063/5.0109918.

Joshi, Aditya, Anuj Singh, Shikha Chauhan, and Anupama Sharma. 2021. "Decision Tree Algorithm for Credit Card Fraud Detection." *Webology* 18(4):2055–61. doi: 10.29121/web/v18i4/103.

Mansur Huang, Nur Shahellin Mansur, Zaidah Ibrahim, and Norizan Mat Diah. 2021. "Machine Learning Techniques for Early Heart Failure Prediction." *Malaysian Journal of Computing* 6(2):872–84.

Noh, S. S. M., Ibrahim, N., Mansor, M. M., & Yusoff, M. (2023). Hybrid filtering methods for feature selection in high-dimensional cancer data. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(6), 6862-6871.

Sayjadah, Yashna, Ibrahim Abaker Targio Hashem, Faiz Alotaibi, and Khairl Azhar Kasmiran. 2018. "Credit Card Default Prediction Using Machine Learning Techniques." *Proceedings - 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018* (April 2020):1–6. doi: 10.1109/ICACCAF.2018.8776802.

Shafie, Shahidan, Soek Peng Ooi, and Khai Wah Khaw. 2023. "Prediction of Employee Promotion Using Hybrid Sampling Method With Machine Learning Architecture." *Malaysian Journal of Computing* 8(1):1264–86. doi: 10.24191/mjoc.v8i1.18456.

Subasi, A., and S. Cankurt. 2019. "Prediction of Default Payment of Credit Card Clients Using Data Mining Techniques." Pp. 115–20 in *2019 International Engineering Conference (IEC)*.

Thomas, Lyn C. 2000. "A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers." *International Journal of Forecasting* 16(2):149–72. doi: 10.1016/S0169-2070(00)00034-0.

Xu, Changqin, Alexander Unger, Chongzeng Bi, Julie Papastamatelou, and Gerhard Raab. 2022. "The Influence of Internet Shopping and Use of Credit Cards on Gender Differences in Compulsive Buying." *Journal of Internet and Digital Economics* 2(1):27–45. doi: 10.1108/jide-11-2021-0017.

Yang, Shenghui, and Haomin Zhang. 2018. "Comparison of Several Data Mining Methods in Credit Card Default Prediction." *Intelligent Information Management* 10(05):115–22. doi: 10.4236/iim.2018.105010.

Yeh, I. Cheng, and Che-hui Lien. 2009. "The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients." *Expert Systems with Applications* 36(2, Part 1):2473–80. doi: https://doi.org/10.1016/j.eswa.2007.12.020.