JST: Smart Systems and Devices
Volume 34, Issue 1, January 2024, 016-022

Credit Card Service Churn Prediction by Machine Learning Models

Tran Hoang Hai[1*], Vu Van Thieu[1], Doan Minh Hieu[2]
[1]*Hanoi University of Science and Technology, Ha Noi, Vietnam*
[2]*HUS High School for Gifted Students, Ha Noi, Vietnam*
[*]*Corresponding author email: hai.tranhoang@hust.edu.vn*

Abstract

The paper presents a study on the application of basic machine learning models for churn customer classification. Churn prediction is an essential task in customer retention for businesses, and accurate identification of customers who are likely to churn can significantly impact the organization's revenue and customer satisfaction. In this study, we explore the performance of various machine learning models, including K-Nearest Neighbor, Random Forest, Adaboost and a deep learning model which is CNN-1D. We use the BankChurners dataset, then we predict the probability that customers abandoning bank services such as credit card services. We evaluate the models basing on various performance metrics such as accuracy, precision, recall, and F1-score. The result demonstrates the potential of basic machine learning models for churn customer classification and provides insights into the key factors contributing to customer churn.

Keywords: Prediction, business analysis, machine learning, E-commerce.

1. Introduction

Customer churn prediction is a crucial aspect of customer retention for businesses. Churn, in the context of customer retention, refers to the phenomenon of customers discontinuing their relationship with a company or brand. In other words, it is the rate at which customers leave a business over a given period. Customer churn is a costly problem for businesses because it affects their revenue and profitability. Therefore, identifying customers who are likely to churn and taking appropriate measures to retain them can significantly impact a business's bottom line [1, 2]. Predicting customer churn has traditionally been a challenging task for businesses. However, with the increasing availability of data and advances in machine learning algorithms, businesses can leverage customer data to predict churn accurately. Churn prediction models use historical data on customer behavior, such as transactional history, customer demographics, and customer service interactions, to identify patterns that indicate a customer's likelihood to churn. Machine learning algorithms, such as logistic regression, decision trees, random forest, and support vector machines, are commonly used to predict customer churn [3]. These algorithms are trained on historical data to identify patterns and build a model that can predict the likelihood of churn for new customers.

In this paper, we explore the performance of various machine learning models for churn customer classification. We evaluate the models basing on different performance metrics and conduct feature importance analysis to identify the most significant factors contributing to customer churn.

2. Problem Statement

2.1. Business Problem and Related Works

Lending organizations aim to retain customers for as long as possible, as each customer has unique reasons for borrowing. These organizations generate revenue by earning interest on loans and profiting from the difference in interest rates between themselves and their customers. However, borrowers only benefit from the lender's services if they spend a significant amount of time using them. As such, the value of a loan increases as the amount of time a customer spends with the lender also increases. Consequently, lending organizations face the challenging task of analyzing customer abandonment statistics to identify the risk associated with each client's behavior and develop strategies to mitigate any disadvantageous problems. To accomplish this, most lending organizations utilize predictive support systems that analyze two situations. Firstly, they predict applicant credit ratings to establish loan standards. Secondly, they improve their services to offer their customers the best possible satisfaction, thereby retaining existing customers and increasing their lifetime value [4, 5]. While everyone acknowledges the importance of retaining customers, banks are limited in what they can do when they do not see customer churn coming. This is where predicting churn at the right time becomes crucial, particularly in the absence of clear customer feedback. Early and accurate churn prediction allows customer relationship management and customer experience teams to engage with customers proactively and creatively. In fact, by reaching out to customers early enough, 11% of churn

ISSN 2734-9373
https://doi.org/10.51316/jst.171.ssad.2024.34.1.3
Received: July 11, 2023; accepted: December 10, 2023

can be avoided. Therefore, analyzing customer data and employing new technology to predict these scenarios is vital for maximizing profits of lending organizations. The authors in [6] employed sensitivity analysis and boosting techniques in their churn prediction framework. Their study revealed that the proposed framework achieved high accuracy even when the training and test data were taken from different time windows. In [7], authors recommended the use of a logit model and classification tree to predict churn in the insurance market. The application of artificial intelligence has numerous advantages over traditional methods because these methods are more robust and can accurately predict non-linear data [8]. Other techniques such as ensemble learning and evolutionary learning [9-11] have also been employed by practitioners. In [12], the authors presented two genetic algorithm-based neural network models for predicting customer churn in wireless service subscriptions. Their study concluded that the proposed models outperformed the existing statistical z-score model in terms of accuracy and all other performance criteria. Furthermore, the authors proposed a genetic algorithm-based neural network in [13] and demonstrated that this algorithm exhibited superior prediction accuracy, receiver operating characteristic, and decile lift compared to existing methods.

2.2. Dataset Analysis

Because of the above reasons, there have been a lot of competition on Kaggle, which commercial and credit companies provide their data to encourage researchers around the world seeking solutions for their service's improvement. The data may consist of age, education, and monthly income of customer basing on company's provision. For example, the Home Credit Default Risk which is a competition by home credit group starting from 2018 with the prize is 70.00 dollars. They have provided many information including all client's previous credits provided by other financial institutions that were reported to Credit Bureau, monthly balances of previous credits in Credit Bureau, monthly balance and cash loans that the applicant had with Home Credit, repayment history for the previously disbursed credits in Home Credit related to the loans. The competition attracted 7170 teams to participate and hundreds of submissions. Another dataset in Kaggle is BankChurners. This dataset is collected in [14] which is a company based in US specializing in data analysis. The difference between this dataset and home credit dataset mentioned above is that the home credit dataset focus on loans and lending limit while the BankChurners dataset focus on classificatory user leaved service and improve the quality. In this paper, the BankChurners dataset is utilized for classification churn customers. The shape of dataset is (10271, 21) in which, each row of dataset is customer's information with 21 attributes as shown in Table 1.

Table 1. Feature descriptions

Feature	Description
Clientnum	Client number. Unique identifier for the customer holding the account
Customer_Age	Demographic variable — Customer's Age in Years
Gender	Demographic variable — M=Male, F=Female
Dependent_count	Demographic variable — Number of dependents
Education_Level	Demographic variable — Educational Qualification of the account holder (example: high school, college graduate, etc.)
Income_Category	Demographic variable — Annual Income Category of the account holder
Card_Category	Product Variable — Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book	Period of relationship with bank
Total_Relationship_Count	Total no. of products held by the customer
Months_Inactive_12_mon	Number of months inactive in the last 12 months
Marital_Status	Demographic variable — Married, Single, Divorced, Unknown
Contacts_Count_12_mon	Number of Contacts in the last 12 months
Credit_Limit	Credit Limit on the Credit Card
Total_Revolving_Bal	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Average Card Utilization Ratio
Attrition_Flag	Internal event (customer activity) variable

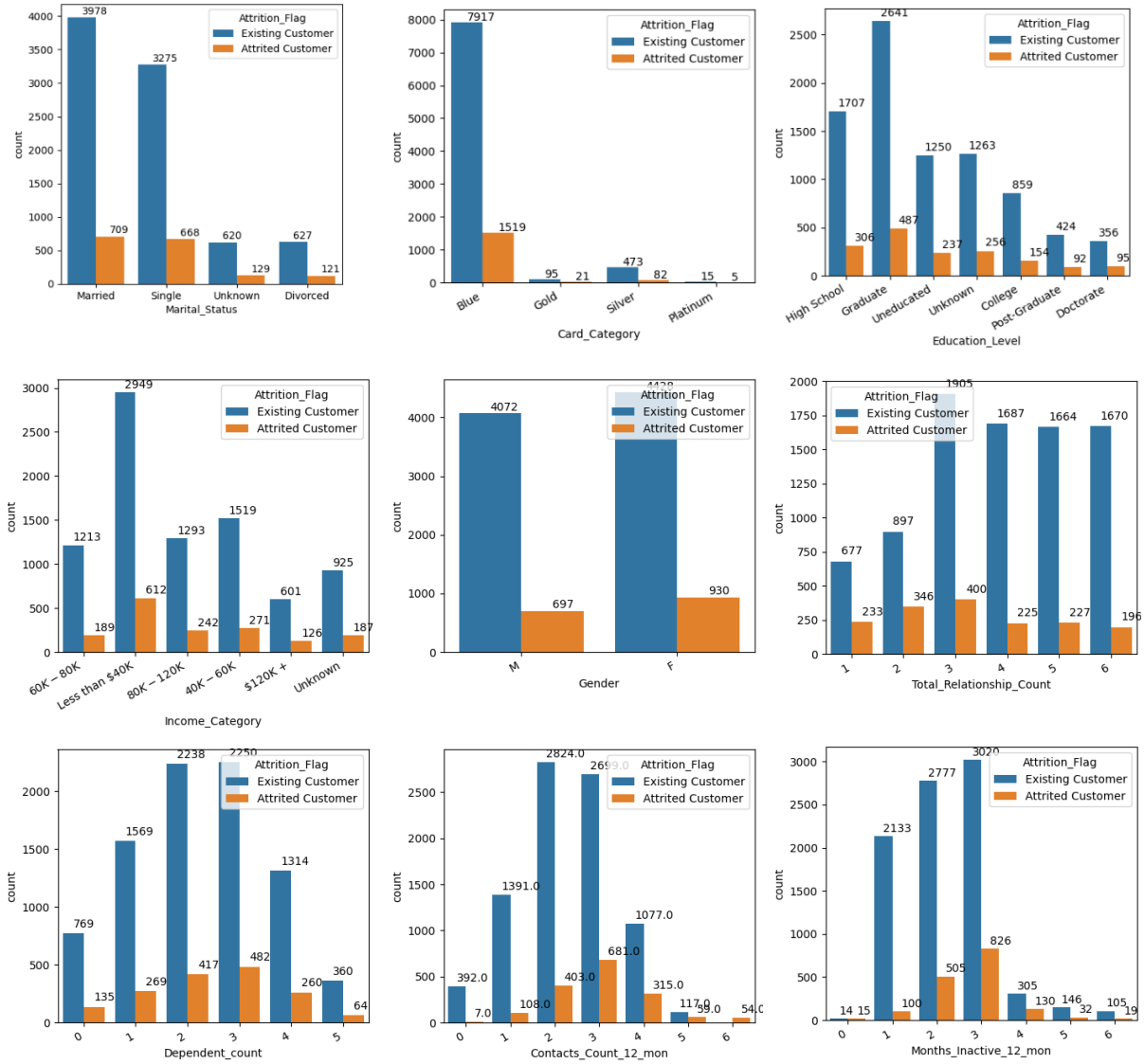

Fig. 1. Data visualization

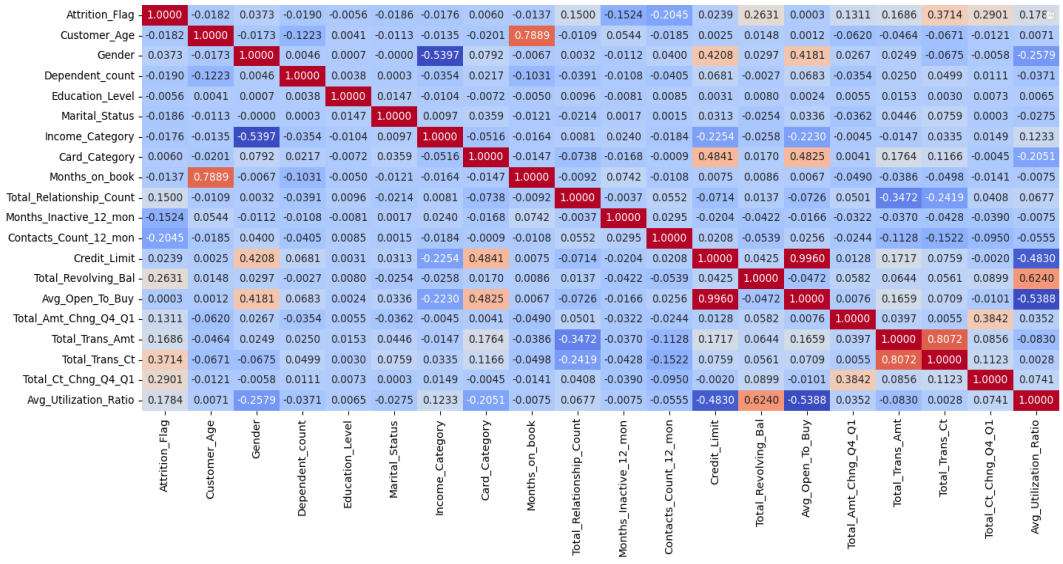

Fig. 2. Correlation matrix

In this dataset, the target feature is Attrition_Flag. Each of data in the dataset is utilized for inputting to classification Attrition_Flag. Nine attributes stored in categories format are visualized. The rest attributes of dataset are stored in randomly discrete distribution. The data visualizations are shown in Fig. 1.

The correlation matric in dataset has shown in Fig. 2.

Basing on the correlation matrix, the relationship of all attributes with Attrition_flag in dataset have the different values. The value is in the range from -0.2045 to 0.3714. In general, these values are the same. This means that dimensionality reduction is very difficult to apply in this dataset. Therefore, all of features are used for classification.

3. Machine Learning Models

In this paper, three algorithms are used to classify churn customers. These algorithms include K-Nearest Neighbor (KNN), Random Forest and Adaboost. There have been a lot of machine learning algorithms and this selection is purely based on performance, ability to save system resources, and ease of installation. The principle of KNN is very simple. The algorithm calculates the Euclid distance from the data point to be classified to all data points in the database, then selects K points with the closest distance. Using majority voting solution, the data point will be classified and labeled. KNN is a simple and intuitive model but still highly effective because it is non-parametric. The model makes no assumptions about the data distribution. Furthermore, it can be used directly for multiclass classification. Random Forest works based on building a lot of decision trees. To ensure that the decision trees have different predictions, or the tree structure of the decision trees is different, serving for majority voting solution, Random Forest uses two techniques including bootstrapping and attribute sampling to randomly select samples and features. This enables the algorithm to increase accuracy and objectivity. AdaBoost is a boosting algorithm. This algorithm combines weak classifiers. Each of the following algorithms corrects the errors of the previous algorithm, which makes the final output as correct as possible. One of the disadvantages of boosting algorithms is that it is easy to be overfitting. The reason is that the algorithm tries to fit each specific data point in the data set. However, in general, Adaboost uses very little system resources, being easy to implement, and the algorithms that make up Adaboost are also very simple, making these algorithms widely applied. A convolutional neural network (CNN) is a type of artificial neural network used primarily for image recognition and processing, due to its ability to recognize patterns in images. A CNN is a powerful tool but it requires a lot of labelled data points for training.

4. Proposal Model

4.1. Pre-Process Data

There are 3 steps for data processing as shown in Fig. 3.

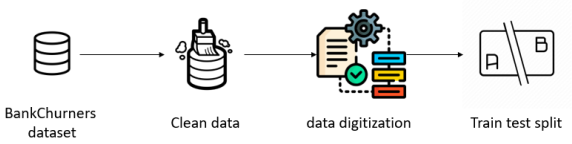

Fig. 3. Data processing

4.1.1. Clean data

The dataset has 21 attributes. Some of which are unused and these attributes contain Null or undefined values such as NaN or Inf. This step aims to remove the NaN and Inf value or replace this value by another value. In this dataset, the CLIENTNUM attribute must be removed.

4.1.2. Data digitization

The dataset has 5 features including Gender, Education_Level, Marital_Status, Income_Category. These features are object category. Therefore, Data digitization plays an important role to pre-process data for building model AI.

4.1.3. Train test split

The dataset is divided into 2 subsets including training and testing set.

4.2. Proposed Model

In Fig. 4, The AI model is proposed. Training data is used for training model through KNN, Random Forest, AdaBoost, and CNN. After modeling, testing data is utilized for evaluating model by measuring accuracy, precision, recall, and F1.

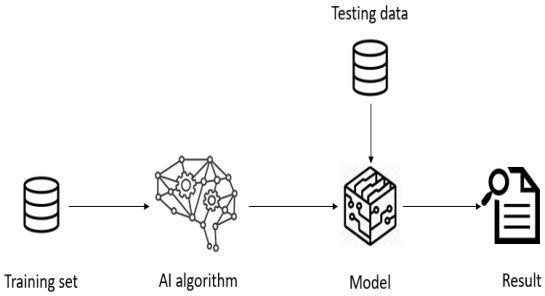

Fig. 4. Proposed model

4.3. Evaluate

Accuracy

The simplest and most common performance metric is Accuracy. The accuracy is calculated as the following:

$$accuracy = \frac{number\ of\ right\ prediction}{number\ of\ data} \quad (1)$$

This evaluation simply calculates the ratio between the number of correctly predicted points and the total number of points in the test data set. Although there are many limitations, the accuracy generally reflects the prediction on the entire testing data set, which is very suitable for the overall evaluation of the model.

(True Positive) TP/(False Positive) FP/(True Negative) TN/(False Negative) FN

For each label, there are 4 quantities to measure how well the model predicts on that label. Specifically with label X:

TP - True Positive: This quantity tells us the number of correctly predicted data on label X.

FP - False Positive: This quantity tells us the number of data that is predicted to be label X but is not actually label X. In this case the model returns a wrong prediction.

TN - True Negative: This quantity tells us how many predicted data are not labeled X and they are really not label X. In this case the model correctly predicted because it did not predict label X.

FN - False Negative: This quantity tells us the number of predicted data that are not labeled X but in fact this is label X. In this case the prediction is wrong because it did not predict label X.

Thus, by evaluating each label through the above 4 quantities, we can know when that label is predicted well by the model. However, each label has up to 4 quantities, which makes deciding which model is better still not easy.

Precision and Recall

Precision and Recall are two metrics used to measure the performance of a classifier in binary and multiclass classification problems.

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

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

Precision demonstrates the ability of the model to correctly predict label X, we see that in (2), the component that makes Precision increase or decrease is not TP but FP. Therefore, when Precision is high, it means that the FP is small or the number of labels mistakenly predicted to label X is low. Recall represents the possibility that the predict model does not miss the label X, just like Precision, Recall depends on FN or in other words it depends on the possibility that the model incorrectly predicts the correct label X. Hence we see that TP and TN do not play any role here. In fact, there are differeent metrics, example Sensitive, that have the same meaning as Precision and Recall. However, only with Precision and Recall we can focus on minimizing FN and FP, the 2 components that make our model less accurate prediction.

F1 score

F1 score is a measure of the harmonic mean of precision and recall.

$$\frac{2}{F1} = \frac{1}{Precision} + \frac{1}{Recall} \quad (4)$$

What we want is that both Precision and Recall parameters to be high. Unfortunately, there is always a trade-off between them, when higher Precision often leads to lower Recall and vice versa. The reason is that if the Precision parameter is high, it means that the model has to be very certain to predict the label X, but this makes the model predicted to lack data that is actually label X. Therefore, we need to combine these 2 metrics in to 1, to tune the model in a single direction without worrying too much about Precision or Recall, hence we use F1 score as the overall measure of the model.

5. Results

According to Fig. 5, 6, 7 and 8, the accuracy, precision, recall and F1-score from those algorithms are different. In particular, the accuracy of AdaBoost is the biggest. It is higher 0.03% than Random Forest, 6.68% than KNN and 9.9% than CNN1-D. The precision of AdaBoost is 0.35% bigger than Random Forest, 7.9% than KNN and 13.95% than CNN1-D. However, the recall of CNN is 1.09% higher than AdaBoost, 1.41% than Random Forest and 4.19% than KNN. The F1-Score of AdaBoost is 0.01% higher than Random Forest, 3.87% than KNN and 5.29% than CNN1-D. Based on these results, we can conclude that Random Forest and AdaBoost have a large correct prediction rate while CNN1-D is an algorithm with a small error prediction rate. Using CNN may not cause any discomfort for customers, but this could miss some of customers who need consulting while other two algorithms AdaBoost and Random Forest do not miss any churn customer but sometimes cause problems.

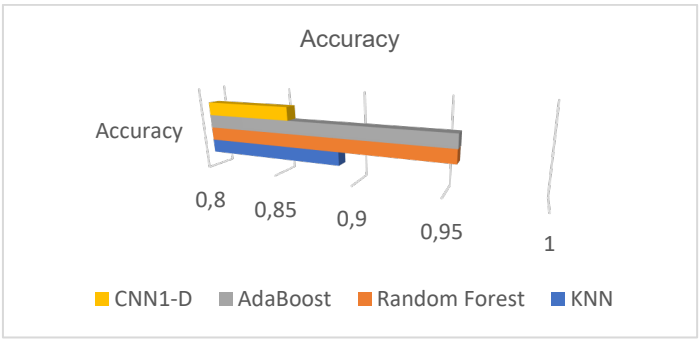

Fig. 5. Comparison of accuracy between three machine learning algorithms and CNN-1D

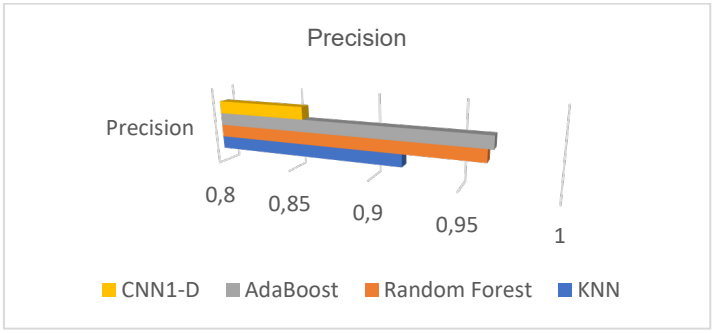

Fig. 6. Comparison of Precision between three machine learning algorithms and CNN-1D

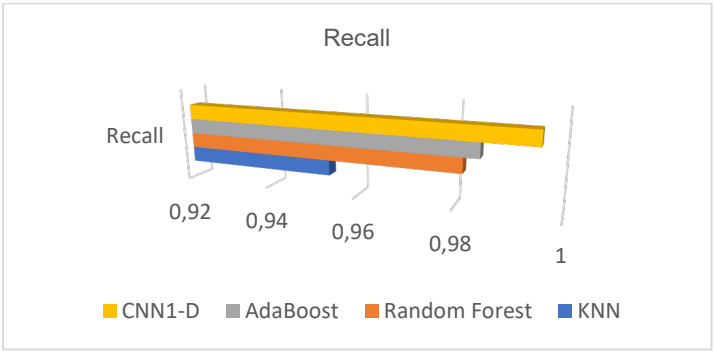

Fig. 7. Comparison of Recall between three machine learning algorithms and CNN-1D

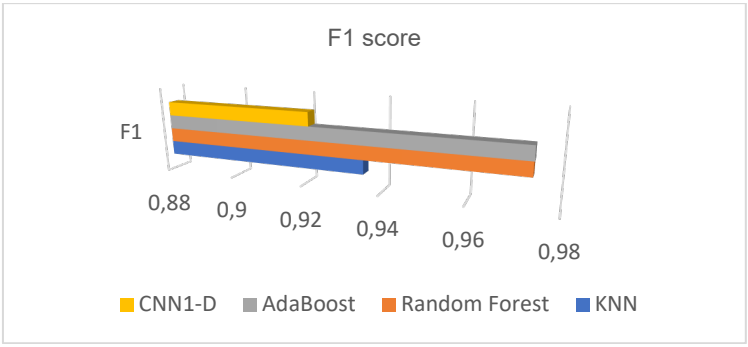

Fig. 8. Comparison of F1-Score between three machine learning algorithms and CNN-1D

6. Conclusion

This paper resolves problems of previous work by propose a new dataset with different machine learning models. This study highlights the effectiveness of machine learning algorithms in predicting customer churn and provides valuable insights into the factors that drive customer churn in the banking industry. Overall, these findings can help businesses develop targeted strategies to retain customers and minimize customer churn, ultimately leading to increased customer satisfaction and revenue.

Acknowledgement

This research was funded by Vietnam Ministry of Education and Training under BKA-04-2022 grant entitled “Research and Development of Network

Intrusion Detection Framework Based on Combination and Improvement of Machine Learning and Deep Learning Techniques". The authors declare no conflict of interest and Dr. Tran Hoang Hai is the corresponding author.

References

[1] J. Burez and D. Van den Poel, CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services, Expert Syst. Appl. 32 (2), 2007, 277-288, https://doi.org/10.1016/jeswa.2005.11.037

[2] J. Burez and D. Van den Poel, Handling class imbalance in customer churn prediction, Expert Syst. Appl. 36 (3), 2009, 4626-4636, https://doi.org/10.1016/jeswa.2008.05.027

[3] Panimalar, S. A. and Krishnakumar, A., A review of churn prediction models using different machine learning and deep learning approaches in cloud environment, Journal of Current Science and Technology, 13(1), 2023, 136-161.

[4] N. Gordini, A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy, Expert Syst. Appl. 41 (14), 2014, 6433-6445, https://doi.org/10.1016/j.eswa.2014.04.026

[5] A. Lemmens and C. Croux, Bagging and boosting classification trees to predict churn, J. Market. Res. 43 (2), 2006, 276-286,

[6] Michael C. Mozer, Richard Wolniewicz, David B. Grimes, Eric Johnson, and Howard Kaushansky, Churn reduction in the wireless industry, In Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99). MIT Press, Cambridge, MA, USA, 1999, 935-941.

[7] Risselada Hans, Verhoef Peter C and Bijmolt Tammo H.A, Staying power of churn prediction models, Journal of Interactive Marketing, Elsevier, vol. 24(3), 2010, 198-208.

[8] Martínez-López Francisco and Casillas Jorge, Artificial intelligence-based systems applied in industrial marketing: An historical overview, current and future insights, Industrial Marketing Management, 42, 2013, 489-495.

[9] Tamaddoni Ali, Stakhovych Stanislav, and Ewing Michael, Comparing churn prediction techniques and assessing their performance: A contingent perspective, Journal of Service Research. 19, 2015, https://doi.org/10.1177/1094670515616376.

[10] Coussement Kristof and De Bock Koen, Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning, Journal of Business Research, 66, 2013, 1629-1636, https://doi.org/DOI: 10.1016/j.jbusres.2012.12.008.

[11] Wai-Ho Au, K. C. C. Chan, and Xin Yao, A novel evolutionary data mining algorithm with applications to churn prediction, in IEEE Transactions on Evolutionary Computation, vol. 7, no. 6, 2003, 532-545, https://doi.org/10.1109/TEVC.2003.819264.

[12] P. C. Pendharkar, Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services, Expert Syst. Appl. 36 (3), 2009, 6714-6720, https://doi.org/10.1016/j.eswa.2008.08.050.

[13] A. Sharma and P. K. Panigrahi, A neural network based approach for predicting customer churn in cellular network services, Int. J. Comput. Appl. 27 (11), 2011, 26-31, https://doi.org/10.5120/3344-4605.

[14] Churn for Bank Customers. [Online] Available at: https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers (Accessed 25/12/2023).