Machine learning in Business: A study of predicting customer churn for banking sector

Thong Cao[1,2], Nhat Nguyen[1,2], Doanh-Chinh Luong[1,2], Nguyen Quang Tuan[1,2], Nghia Nguyen[1,2], Van-Ho Nguyen[1,2,*]

[1]University of Economics and Law, Ho Chi Minh City, Vietnam

[2]Vietnam National University, Ho Chi Minh City, Vietnam

*Corresponding author, Email: honv@uel.edu.vn

Abstract. Customer churn prediction (CCP) is one of the challenging problems in the service industry. The banking industry is no exception. Soon awareness about who is likely to churn helps companies have the right plan for the right customer, which keeps their customers. Furthermore, it brings profit to the business because it costs less to retain existing customers than to acquire new customers. This study aims to learn from data mining algorithms by using them to build the optimal model for customer churn prediction. For this purpose, we carried out this project using credit card churn customer data, around 10,000 records from the Kaggle repository. In the first stage, we conducted an exploratory data analysis (EDA). Next, we applied Recursive Feature Elimination (RFE) method to find the most relevant features. At the final stage of model selection and evaluation, we implemented four models: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). After that, we performed a detailed comparison between model results using metrics calculated from confusion matrix. The result shows that RF is outperformed compared to the others. The proposed model can be applied for businesses to find the customers who intend to not using their products or services.

Keywords: Classification Algorithms, Customer Churn Prediction, Financial Technology, Machine Learning in Business, Recursive Feature Elimination.

1 Introduction

Every organization is facing a problem in a competitive environment is how to predict their customers who are likely to churn from the company. Mainly, in the banking industry, customers can choose among multiple service providers and easily exercise their right to switch from one service provider to another. Soon awareness helps the banking companies have the right plan and promotion to the right customers and enable them to reduce the churn rate. The previous study indicated that improving the retention rate by up to 5% can increase a bank's profit by up to 85% [1]. Moreover, one of the major reasons for this is that it costs less to retain existing customers than to acquire new customers [2]. As can be seen, reducing customer attrition has a significant impact

on increasing profits for commercial banks and enhancing their core competitiveness. Therefore, commercial banks must improve their abilities to predict customer attrition.

In the study, we conducted and compared four machine learning models, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). Moreover, to reduce complexity time by reducing the dimensionality of the input dataset, we used Recursive feature elimination (RFE) to find the most relevant features.

The remainder of this study is organized as follows: Section 2 discusses the related works in terms of customer churn, which has provided much knowledge about the algorithms and helped to design a suitable research method for the study. Research methodology is outlined in Section 3, while conducting experimental processes with Python language and evaluating the results are presented in Section 4. The last Section is the conclusion of the study.

2 Related works

This section provides the conventional definition as well as the formula in terms of customer churn. Followed by the discussion on some related studies based on the purpose of this study.

2.1 Customer churn

Customers churn is defined as customers who stop contracting or using a company's services over a period of time like a month/quarter/year.

The churn rate is a ratio based on the amount of customer churn and the total number of existing customers for a given period of time. The Churn Rate formula is demonstrated as (1) below.

$$\text{Churn Rate} = \frac{Number\ of\ users\ at\ the\ beginning - Number\ of\ users\ at\ the\ end}{Number\ of\ users\ at\ the\ beginning}\ x\ 100\% \quad (1)$$

In reality, attracting new customers is much more expensive than caring for and holding on to existing ones. The business is at a safe level if the churn rate is low. Conversely, the company is in the alarm zone if the churn rate is medium or high.

2.2 Customer churn prediction using machine learning

In the same dataset for developing a churn prediction system with our study, Muneer et al. (2022) applied three intelligent models: Random Forest, AdaBoost, and Support Vector Machine. Besides that, due to an imbalanced dataset, a method was proposed by the authors to solve this issue [3]. The experimental results show that RF raises the best result after imbalanced dataset handling method was applied with F-measure of 91.90% and accuracy of 88.7%. However, accuracy and time complexity can be further improved using the optimization algorithms for the feature selection process. In this study, we applied an additional feature selection method as Recursive feature elimination (RFE) to choose the most relevant features or predictors in predicting the target variable, which is the attrition flag in this situation.

Nie et al. (2011) provided a comparative analysis of machine learning models for credit card churn prediction between logistic regression and decision tree [1]. After

processing multicollinearity using variance inflation factor (VIF) method and selecting variables, the authors built six models with different variable combinations. However, regarding imbalanced processing, the study solved the issue by randomly selecting the number of samples in the churners equal to that in the non-churners to be a training set. So that the training size is too small which may affect the generalization ability of the classifier.

In 2020, Rahman & Kumar applied various classifiers for churn prediction in banking, namely k-Nearest Neighbor, Support Vector Machine, Decision Tree, and Random Forest [4]. The researchers also used feature selection methods, which are Minimum Redundancy Maximum Relevance (mRMR) and Relief, to find the more relevant features. As a result, using the Random Forest model after oversampling is better than other models in terms of accuracy. However, random oversampling can raise overfitting because the generated samples are exact replications of samples.

All those machine learning algorithms above are popular in terms of churn prediction, in this study we choose the well-known algorithms, namely Logistic Regression, Decision Tree, Random Forest, Support Vector Machine to build churn prediction models.

3 Methodology and proposed research model

In this section, the demonstration of our proposed methodology is shown in the Figure 1. According to the EDA phase, we focus on analyzing simple descriptive statistics of some main features in demographic and information transaction group variables to determine the distribution and character extracted from these variables.

Regarding data preprocessing, the checked null value showed that the dataset has no missing or null values. In this phase, feature selection is taken into consideration using RFE. Next, the data has been split into two parts: train and test set in 75% and 25% respectively. The most popular predictive models have been applied in the prediction process, namely, Logistic Regression, Random Forest, Decision tree, Support Vector Machine on the train set. Finally, the obtained results on the test set have been evaluated using confusion matrix such as F-measure, Accuracy, Specificity, Sensitivity, Precision, Recall.

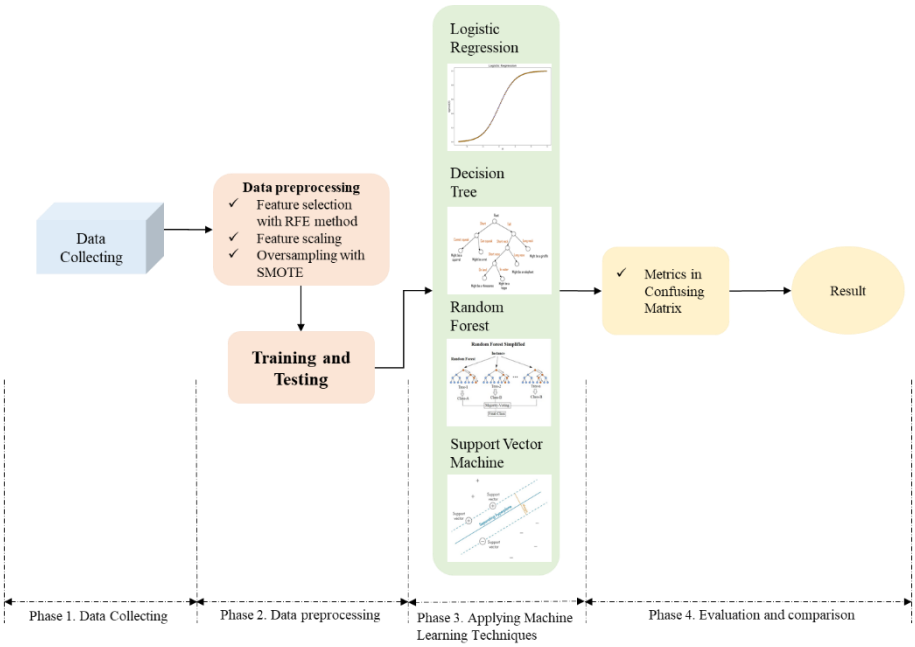

Fig. 1. Overview of the proposed research model

4 Empirical study

In this section, we conduct EDA to extract some main features of the dataset. Followed by the preprocessing stage, RFE chooses the relevant features for the prediction. Scaling was still carried out in this stage. Next, the machine learning algorithms will be discussed in the modeling stage. Finally, the model's performance are demonstrated in experiment results.

4.1 Dataset description

The dataset used for the prediction process task is publicly available on the Kaggle website [5]. The variables included in the dataset are listed in Table 1. Total dataset has 23 variables. Removing the last two unrelated columns, the dataset now consists of 20 predictor variables, and one target variable. It contains 10,127 records, of which 8,496 (84%) are non-churners and 1,630 (16%) are churners. Therefore, the dataset is highly unbalanced in terms of the proportion of churners and non-churners.

Table 1. Description of variables in the dataset.

No	ColumnName	Description
1	CLIENTNUM	Unique identifier for the customer
2	Attrition Flag	The target variable showing customer churns or not
3	Customer_Age	Customer's Age in Years
4	Gender	M=Male, F=Female
5	Dependent count	Number of dependents
6	Education_Level	Educational Qualification of the account holder

7	Marital Status	Marital Status of the account holder
8	Income_Category	Annual Income Category of the account holder
9	Card_Category	Product Variable - Type of Card
10	Months_on_book	Period of relationship with the bank
11	Total_Relationship_Count	Total of products held by the customer
12	Months_Inactive_12_mon	Number of months inactive in the last 12 months
13	Contacts Count_12_mon	Number of Contacts in the last 12 months
14	Credit_Limit	Credit Limit on the Credit Card
15	Total_Revolving_Bal	Total Revolving Balance on the Credit Card
16	Avg_Open_To_Buy	Open to Buy Credit Line (Average of last 12 months)
17	Total_Amt_Chng_Q4_Q1	Change in Transaction Amount (Q4 over Q1)
18	Total_Trans_amt	Total Transaction Amount (Last 12 months)
19	Total_Trans_ct	Total Transaction Count (Last 12 months)
20	Total_ct_Chng_Q4_Q1	Change in Transaction Count (Q4 over 01)
21	Avg_Utilization_Ratio	Average Card Utilization Ratio

For each customer in the dataset, there are a lot of detailed features information and their meanings are described in Table1. Those features could be divided into two groups as demographic customers and transaction information. Demographic variables include age, gender, education level, number of dependents, income, marital status, while transaction information consists of the rest.

4.2 Explore data analysis

Figure 2 indicates that 16.1% of the total customer base has left, the disparity between the number of persons still using the service and those leaving shows that the data set is unbalanced.

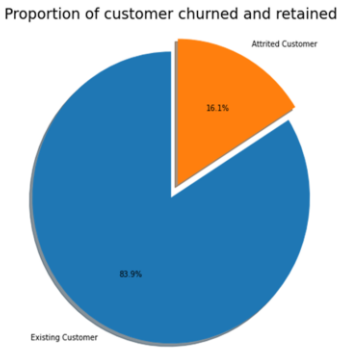

Fig. 2. Proportion of customers who churned and retained.

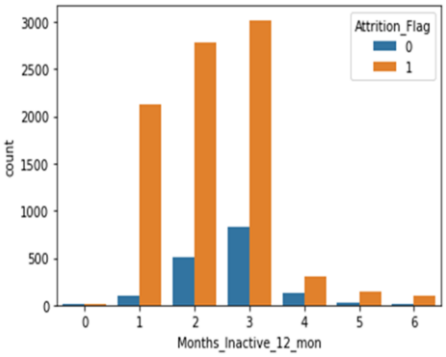

Fig. 3. Proportion of customer inactive 12 months and attrition.

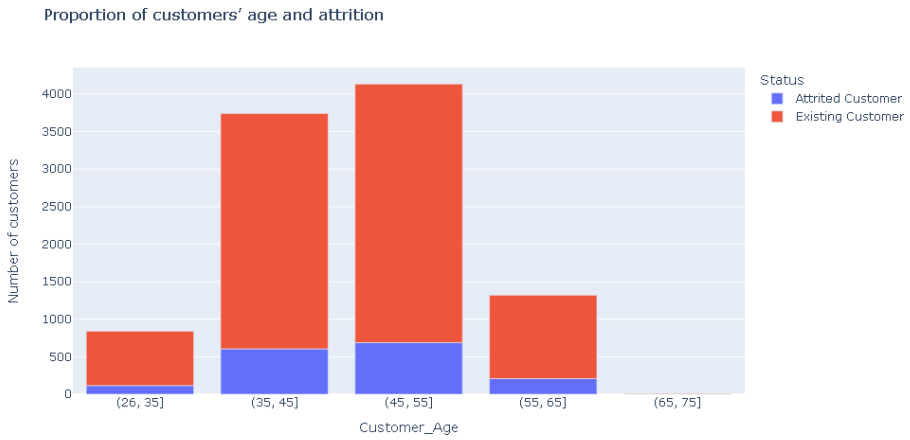

Fig. 4. Proportion of customers' age and attrition.

Figure 3 demonstrates that most users not interacting with the services for two to three months are likely to stop using them. Businesses can then offer more promotions and other services for those customers to retain them.

Figure 4 shows a variation in service preferences between the age groupings. We can see that the main customer group of the company is mainly in the 35-55 age group, but this is also the group with the most significant number of customers leaving, showing the urgency of building better services aimed at this group. The bank might need to evaluate its target market or its retention strategy for the various age groups.

4.3 Preprocessing

Feature selection. Recursive Feature Elimination is a widely used algorithm for selecting features that are most relevant in the predicting of the target variable. First, it builds a model based on all features and calculates the importance of each feature in the model. Then, it rank-orders the features and removes the one with the least importance iteratively based on model evaluation metrics. This process continues until a smaller subset of features is retained in the model [6]. In this study, we will apply the RFE method combined with cross-validation to find out the number of variables that bring the highest results.

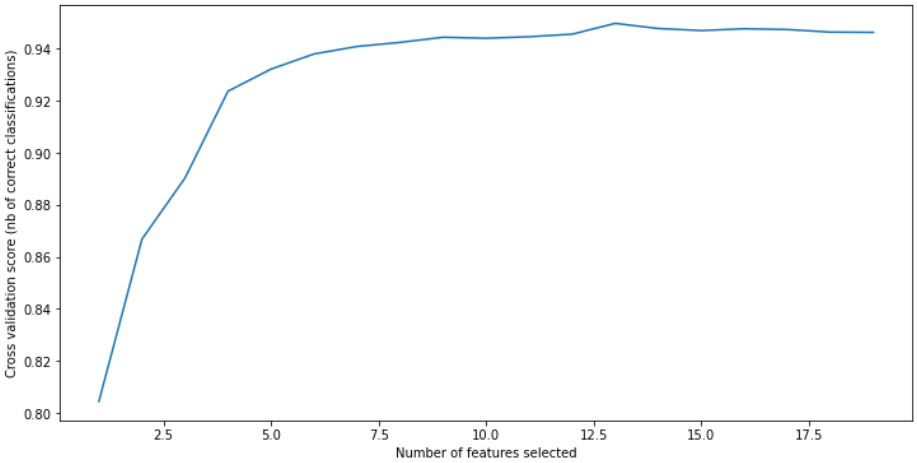

Fig. 5. Result of RFE method.

The result of RFE is shown as Figure 5, the best results with 13 variables including: 'Customer_Age', 'Months_on_book', 'Total_Relationship_Count', 'Credit_Limit','Avg_ Open_ To_Buy','Total_Amt_Chng_Q4_Q1','Total_Ct_Chng_Q4_Q1','Avg_Utilization_Ratio', 'Total_ Revolving_Bal', 'Total_Trans_Ct', 'Total_Trans_Amt', 'Contacts_Count_12_mon'.

Scaling. In this process, we scale the features from our dataset to standard normally distributed data by using StandardScaler() function. The reason is to avoid features in greater numeric ranges dominating those in smaller numeric ranges. Moreover, it performs numerical calculations during the iterative process of model building [7]. The values of the data set are changed after scaling, however, they do not change the distribution of the data and the numerical gap in each variable is significantly reduced.

4.4 Modeling

In this section we built four predictive models, the concept of those algorithms were demonstrated below.

Logistic regression. LR is a supervised statistical technique that estimates the probability of an event occurring, such as churn or non-churn. Since the outcome is a probability which is bounded between 0 and 1 as two classes of the dependent variable.

Decision tree. DT is a machine learning algorithm, used for classification and prediction problems, builds tree models by dividing samples according to the values of independent variables into a lot of branches in which each branch represents the outcome of the test, and each leaf node (the last node) represents a class label (decision taken after computing all features).

Random forest. RF is an extension algorithm of Decision Tree. DT creates a tree, then Random Forest generates a series of decision trees to build a model, it will initially select many different samples from the dataset by the method of selection, each sample will generate a decision tree, finally the final result of the algorithm is the summation

of many different decision trees, so the information of one tree will complement the other, bringing good prediction results for the model.

Support vector machine. SVM is a machine learning method based on statistical learning theory developed by [8]. The SVM aims to find an optimal linearly separable SVM classification surface. In other words, it divides the prediction into two parts –1 that is the left side of the hyperplane and +1 that is the right side of the one. In terms of binary classification, the optimal separating line required not only separate two class labels correctly but also maximize the margin between the two classes. In the practice case, we must face the fact that the optimization becomes a trade-off between a large margin and a small error penalty [9]. Besides that, it still has a solution for nonlinear dataset by applying kernel trick, which transform input data to a high-dimensional feature space in which the one becomes linearly separable [10].

4.5 Experimental results

After building the machine learning models, we evaluate the performance of each model and compare them with each other through measures calculated from the confusion matrix. Applied to the problem, the target variable contains two label classes, churners and non-churners, represented as positive and negative classes, respectively. If the model predicts correctly data points will belong to two groups as TP (True Positive) and TN (True Negative) otherwise belong to the FN (False Negative) and FP (False Positive) group. Table 2 demonstrates the structure of confusion matrix [11].

Table 2. Classifier evaluation.

		Predicted class	
		Churners	Non-churners
Actual class	Churners	TP	FN
	Non-churners	FP	TN

Next, there are some metrics in the Confusion matrix that we applied in the article, which are Accuracy, Precision, Recall, F1-score (F-measure). Accuracy indicates the accuracy of the entire model.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

Precision shows how many percent correctly the model predicted the variables of the Positive class.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

Recall shows what percentage of the variables are actually Positive in all the Positive variables in the Actual Class.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

F1-score (F-measure) is the harmonic average of two quantities, Precision and Recall.

$$\text{F1-score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

Table 3. Summary of models' performance of four algorithms.

Model	Accuracy	Precision	Recall	F1-Score
LR	0.90	0.81	0.76	0.78
DT	0.94	0.90	0.89	0.89
RF	0.96	0.95	0.90	0.92
SVM	0.94	0.90	0.83	0.86

Table 3 demonstrates the performance of four algorithms applied for the dataset. As can be seen, the result predictions have high accuracy in all models (all above 0.9). However, these degree of accuracy cannot be the best measures showing how good model performances are due to the prediction model could be predicted well in the majority class and predicted badly in minority. The visualization of the evaluation shown in Figure 6 below.

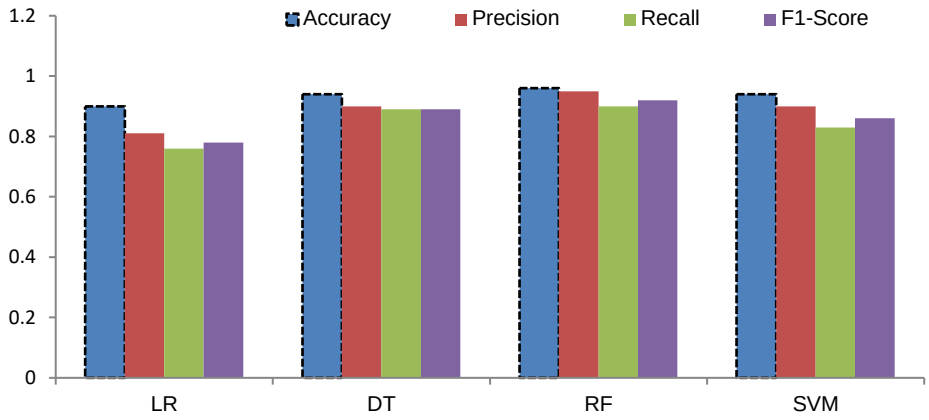

Fig 6. Visualization of models' performance of four algorithms

5 Conclusion and Future works

Experimental results show that the RFE method helps to reduce the number of data variables but still keeps the high prediction results, thereby not also saving the model processing cost but still keeping the high efficiency. The evaluation using some metrics calculated from the confusion matrix demonstrates that RF processed with the dataset gave the highest accuracy prediction results. We suggest that the RF model can be further developed for predicting customer churn in the banking sector.

Acknowledgments

This research is funded by University of Economics and Law, Vietnam National University Ho Chi Minh City.

References

1. Nie, G., Rowe, W., Zhang, L., Tian, Y., and Shi, Y.: 'Credit card churn forecasting by logistic regression and decision tree', Expert Systems with Applications, 2011, 38, (12), pp. 15273-15285
2. Roberts, J.H.: 'Developing new rules for new markets', Journal of the Academy of Marketing Science, 2000, 28, (1), pp. 31-44
3. Muneer, A., Ali, R.F., Alghamdi, A., Taib, S.M., Almaghthawi, A., and Abdullah Ghaleb, E.: 'Predicting customers churning in banking industry: A machine learning approach', Indonesian Journal of Electrical Engineering and Computer Science, 2022, 26, (1), pp. 539-549
4. Rahman, M., and Kumar, V.: 'Machine learning based customer churn prediction in banking', in Editor (Ed.)^(Eds.): 'Book Machine learning based customer churn prediction in banking' (IEEE, 2020, edn.), pp. 1196-1201
5. Credit Card customers. (2020). Retrieved from Kaggle, last accessed 2022/04/27, https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers
6. Bulut, O.: 'Effective Feature Selection: Recursive Feature Elimination Using R' (Towards Data Science, 2021, edn.). Retrieved from Towards Data Science: https://towardsdatascience.com/effective-feature-selection-recursive-feature-elimination-using-r-148ff998e4f7.
7. Sharda, R., Delen, D., and Turban, E.: 'Analytics, data science, & artificial intelligence: Systems for decision support' (Pearson Education Limited, 2021), p.308
8. Vapnik, V.N.: 'The nature of statistical learning', Theory, 1995
9. Cortes, C., and Vapnik, V.: 'Support-vector networks', Machine learning, 1995, 20, (3), pp. 273-297
10. Boser, B.E., Guyon, I.M., and Vapnik, V.N.: 'A training algorithm for optimal margin classifiers', in Editor (Ed.)^(Eds.): 'Book A training algorithm for optimal margin classifiers' (1992, edn.), pp. 144-152
11. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., and Chatzisavvas, K.C.: 'A comparison of machine learning techniques for customer churn prediction', Simulation Modelling Practice and Theory, 2015, 55, pp. 1-9