Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute
InformingScience.org

IJIKM.org

Volume 18, 2023

CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED CLASSIFICATION MODELS

Hoang Tran	University of Economics and Law, Ho Chi Minh City, Vietnam Vietnam National University, Ho Chi Minh City, Vietnam	hoangtd21411c@st.uel.edu.vn
Ngoc Le	University of Economics and Law, Ho Chi Minh City, Vietnam Vietnam National University, Ho Chi Minh City, Vietnam	ngocltl21413c@st.uel.edu.vn
Van-Ho Nguyen*	University of Economics and Law, Ho Chi Minh City, Vietnam Vietnam National University, Ho Chi Minh City, Vietnam	honv@uel.edu.vn

* Corresponding author

ABSTRACT

Aim/Purpose	Previous research has generally concentrated on identifying the variables that most significantly influence customer churn or has used customer segmentation to identify a subset of potential consumers, excluding its effects on forecast accuracy. Consequently, there are two primary research goals in this work. The initial goal was to examine the impact of customer segmentation on the accuracy of customer churn prediction in the banking sector using machine learning models. The second objective is to experiment, contrast, and assess which machine learning approaches are most effective in predicting customer churn.
Background	This paper reviews the theoretical basis of customer churn, and customer segmentation, and suggests using supervised machine-learning techniques for customer attrition prediction.
Methodology	In this study, we use different machine learning models such as k-means clustering to segment customers, k-nearest neighbors, logistic regression, decision tree, random forest, and support vector machine to apply to the dataset to predict customer churn.

Accepting Editor Dirk Frosch-Wilke | Received: November 9, 2022 | Revised: January 2, January 28, February 19, February 20, 2023 | Accepted: February 21, 2023.
Cite as: Tran, H., Le, N., & Nguyen, V.-H. (2023). Customer churn prediction in the banking sector using machine learning-based classification models. *Interdisciplinary Journal of Information, Knowledge, and Management*, *18*, 87-105. https://doi.org/10.28945/5086

Contribution	The results demonstrate that the dataset performs well with the random forest model, with an accuracy of about 97%, and that, following customer segmentation, the mean accuracy of each model performed well, with logistic regression having the lowest accuracy (87.27%) and random forest having the best (97.25%).
Findings	Customer segmentation does not have much impact on the precision of predictions. It is dependent on the dataset and the models we choose.
Recommendations for Practitioners	The practitioners can apply the proposed solutions to build a predictive system or apply them in other fields such as education, tourism, marketing, and human resources.
Recommendations for Researchers	The research paradigm is also applicable in other areas such as artificial intelligence, machine learning, and churn prediction.
Impact on Society	Customer churn will cause the value flowing from customers to enterprises to decrease. If customer churn continues to occur, the enterprise will gradually lose its competitive advantage.
Future Research	Build a real-time or near real-time application to provide close information to make good decisions. Furthermore, handle the imbalanced data using new techniques.
Keywords	churn prediction, machine learning, banking industry, classification models, SMOTE

INTRODUCTION

Customer churn is a business term known as customer agitation. In business, when customers are dissatisfied with the service or product that you provide, attrition will occur, or in other words, they will stop connecting or cooperating with your business. Nowadays, more and more businesses are concerned with customer retention. Customer churn hurts businesses since it can result in large premium losses, decreased profit margins, and possibly lost referral business from loyal clients. According to Baran and Galka (2017), "The pioneering work of F. F. Reichheld and W. E. Sasser Jr. found a strong relationship between customer retention and company profits. They found that just a 5 percent increase in customer retention yielded improved profitability in net present value from 20 to 85 percent across a wide range of businesses." Additionally, many types of research demonstrated that keeping current clients costs about five times less than acquiring new ones (Dawes & Swailes, 1999). According to Reichheld (1996), a client's average net present value improves by 35% for software companies and 95% for advertising agencies with a 5% increase in customer retention rates. To limit this, businesses must anticipate specific customers at risk of leaving to adjust their strategies accordingly, such as improving the quality of products and services or increasing their benefits. Therefore, it is essential to create predictive models that could aid in client retention. Machine learning methods can successfully anticipate customer turnover, as Dolatabadi and Keynia (2017) show. Data collection and analysis allow one to identify customers who may be about to leave a business, thereby focusing on customer retention rather than attracting new clients. In addition, machine learning models are also applied in retail and economic sectors. For retailers, understanding the needs of customers is extremely necessary; machine learning models are applied to analyzing shopping behavior based on the customer's purchase history, thereby finding out the relationship between items. This is the key to boosting purchase rates and optimizing profits for businesses (Reddy & Reddy, 2021).

A few studies have also integrated the use of customer segmentation and machine learning techniques to improve the predictive model's accuracy. Customer segmentation is the process of dividing customers into groups based on common target customer characteristics so that companies can mar-

ket to each group effectively and appropriately. This is a pretty important step, a factor that helps increase the conversion rate for businesses. If businesses take this step well, it can help them divide their advertising budgets better and save more. Additionally, customer segmentation will assist them in better understanding their consumers, identifying the target customer category for the business to focus on, and then considering the factors that affect that customer's churn.

Previous research has generally concentrated on identifying the variables that most significantly influence customer churn or has simply used customer segmentation to identify a subset of potential consumers, excluding its effects on forecast accuracy. Consequently, there are two primary research goals in this work. The initial goal is to examine the impact of customer segmentation on the accuracy of customer churn prediction in the banking sector using machine learning models. The second objective is to experiment, contrast, and assess which machine learning approaches are most effective in predicting customer churn. Relationships with customers have a big impact on long-term growth, thus, understanding their behavior allows banks to easily improve existing policies. Building a model to predict credit card discontinuation has important significance for banks. In this study, we use different machine learning models such as k-means clustering (Kanungo et al., 2000), k-nearest neighbors (Cunningham & Delany, 2021), logistic regression (Sperandei, 2014), decision tree (Quinlan, 1990), random forest (Biau & Scornet, 2016), and support vector machine (Jakkula, 2006) to predict credit card churn.

The remainder of this paper is structured as follows. The history of the study and the related studies are covered in the next section. Then the theoretical foundation is presented, followed by the data and methodology. Next, the model results are shown with some discussion. Finally, the paper outlines the research's limitations.

RELATED WORK

CHURN PREDICTION USING MACHINE LEARNING

Various machine learning models, including logistic regression (LR), decision tree (DT), k-nearest neighbor (KNN), random forest (RF), were used in this study (Kaur & Kaur, 2020) to estimate the likelihood that a client will leave. Performance measures including memory, accuracy, and others are compared. Support vector machines (SVM), which might increase system performance, were not taken into account in this study. Inspired by this, we added the SVM model to our work.

Guliyev and Tatoğlu (2021) employed SHapley Additive exPlanations (SHAP) values to help the machine learning model evaluation and interpretability for customer churn analysis. It focused on particularly explainable machine learning models. Utilizing actual banking data, the research aimed to estimate the explainable machine learning model and assess a variety of machine learning models using test data. The XG-boost model fared better than other machine learning techniques in categorizing clients who churn, according to the data. According to Yaseen (2021), they used the feature selection technique to determine the most important variables in predicting customer attrition. They used the wrapper-based feature selection approach, where particle swarm optimization (PSO) was applied for searches, and different classifiers, such as decision tree, naive bayes, k-nearest neighbor, and logistic regression, were applied for evaluation to judge the enactment on optimally sampled and condensed datasets. Last, but not least, simulations showed that their recommended strategy did well for forecasting churners and might therefore be helpful for the telecommunications sector's constantly expanding rivalry. Rahman and Kumar (2020) proposed an approach for predicting client turnover in a bank using machine learning methods, a subfield of artificial intelligence by examining consumer behavior, the study encourages investigation into the possibility of churn. In their work, the classifiers KNN, SVM, DT, and RF are employed. Additionally, several feature selection techniques have been used to identify the features that are more pertinent and to assess system performance. On the Kaggle churn modeling dataset, the experiment was run. To discover a suitable model with greater precision and predictability, the results are compared. As a result, the accuracy of the random forest

model following oversampling is superior to other models. The dataset used by Karvana et al. (2019), which consisted of 57 attributes, was used to test 5 different categorization techniques. Using comparisons between several classes, experiments were conducted repeatedly. The most effective approach for forecasting customer attrition at an Indonesian private bank uses support vector machines in comparison to data from class sample split 50:50. The company that plans to take strategic action to stop customer churn might make use of the modeling results. When there is a lack of a lot of data, this research uses k-fold cross-validation to assess machine learning models. This helps to avoid the model being evaluated incorrectly.

Elyusufi and Ait Kbir (2022) analyzed several specific machine learning models that had been put forth in the literature to address this issue and contrasted them with some recently developed models that are based on ensemble learning techniques. As a result, they developed predictive churn strategies that analyzed customer history data, determined who was active after a specific period, and then developed models that pinpointed the points at which a client might stop using a certain firm service. When completing the training step with traditional models like multi-layer perception neural networks, ensemble learning methods were also employed to locate pertinent features to reduce their quantity. The proposed methodologies could obtain accuracy levels of up to 89% when other research studies using the same dataset only manage to reach 86%. According to this study, the SVM model has an accuracy rating of 86.18%, which is rather good. A high accuracy for a particular model would mean that the model is capable of predicting the choice that a client can make (leave the bank/remain with the bank).

With the use of the three intelligence models random forest, adaboost, and support vector machine, Muneer et al. (2022) developed a method for predicting client attrition (SVM). When the synthetic minority oversampling technique (SMOTE) was used to overcome the unbalanced dataset and the combination of undersampling and oversampling, the method produces the best results. Excellent results were obtained utilizing the approach on SMOTED data, with a 91.90% F1 score and an overall accuracy of 88.7% when employing RF. Additionally, the experimental findings demonstrated that RF produced good outcomes for the entire feature-selected datasets. SMOTE was utilized in this study to reduce sample size disparity. This improves the forecast results' accuracy as well.

Churn Prediction Using Machine Learning Combines Customer Segmentation

Sivasankar and Vijaya (2017) compared different unsupervised learning methodologies employing algorithms like fuzzy c-means (FCM), probabilistic fuzzy c-means (PFCM), and k-means clustering (k-means), which aggregated similar types of customers into clusters and forecast improved customer segmentation. By using the Holdout approach, the clusters were sorted into training and testing groups. The training was done using decision trees, and testing was done using the produced models. The experimental findings of this study have demonstrated that the k-means method provides better cluster quality than fuzzy grouping and that the classification accuracy is improved when the decision tree is used in combination with the k-means algorithm. Therefore, in this work, we integrated the decision tree and the k-means algorithm.

Using the clustering technique and the k-means clustering algorithm, Olaniyi et al. (2020) analyzed consumer competency and sector continuity to anticipate customer behavior. The data were grouped into three labels according to the inflow and outflow of transactions. A support vector machine was used to classify the clustering findings, and an accuracy of 97% was reached. This study enabled banking administrators to analyze client behavior, which might lead to appropriate methods for improving consumer engagement and strengthening administrator conducts. Through customer segmentation, Zhang et al. (2022) intended to create a churn prediction model for telecom clients. A telecom customer churn prediction model was created using data gathered from three significant Chinese telecom providers, fisher discriminant equations, and logistic regression analysis. The results indicate that the regression-based telecom customer churn model produced better outcomes and had a

higher prediction accuracy (93.94%). By clustering and classifying the customers according to their features, this study (Routray, 2021) suggested a market segmentation and customer behavior prediction model. As a result, using the rapid-miner-9.07-toolkit, this study offered the random forest algorithm for predicting the output of churn and the decision tree for predicting the choice of the mobile plan of consumers. These studies examined client segments to discover prospective future consumer groups and, based on that information, to create efficient customer retention tactics.

Xiahou and Harada (2022) suggested a loss prediction model that combines customer segmentation using k-means with prediction using a support vector machine. The technique identified the primary customer groups and divides the consumer base into three categories. To forecast customer attrition, the support vector machine and logistic regression were evaluated. The findings demonstrated the importance of k-means clustering segmentation by demonstrating a significant improvement in each prediction index following customer segmentation. SVM predictions were more accurate than those made using logistic regression. To assist businesses in effectively segmenting their customers, Zhuang (2018) developed a customer value model that incorporated the importance of social networks. Then, they forecast customer turnover before and after the subdivision using the machine learning technique XG-boost. The study discovered that consumer segmentation increases prediction accuracy. The XG-boost algorithm was also superior to other algorithms in terms of benefits.

The empirical findings of the above studies demonstrated that consumer segmentation would greatly enhance each predictor, so we experimented with the data from the banking industry to support this.

DATA AND METHODOLOGY

The research method will comprise phases from data collection to data preparation and pre-processing so that it can be used in the model, including customer segmentation to verify the hypothesis and handling imbalanced data using SMOTE. SMOTE is a common sampling technique to enhance the sample size of the minority group in cases when there is an imbalance in the sample size. To increase the sample size, we will select the k nearest neighbor samples for each sample belonging to the minority group, and then perform linear combination to create the simulated sample. This pipeline in Figure 1 demonstrates the step-by-step process we employ in this section to anticipate client attrition.

DATA PREPARATION

Data preparation is crucial for improved data analysis and management across all data-driven applications, not just data science (Hameed & Naumann, 2020). This section provides some background information on the dataset and outlines how it was utilized for the prediction process. The Kaggle website (https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn) provides a freely accessible dataset for the prediction job. The final two columns of the 23 variables should be eliminated because they are useless for categorizing. After removing the final two columns, the dataset now has 21 variables, with a total of 10,127 entries. Table 1 contains and describes a list of the variables in the dataset.

Machine learning scaling is part of data pre-processing as this technique brings data points that are far from each other closer to increase the algorithm effectiveness and speed up the Machine Learning processing. Scaling data enables the model to learn and understand the problem. There are two main data scaling methods commonly used in algorithms when scaling is indispensable techniques are normalization and standardization. As a result, shown in Figure 2, multiple histograms have been generated to show the distribution of the required variables so that a suitable scaling technique may be selected. The variables that were used to determine the distribution for data scaling among the quantitative data have a wide range. Since the remaining quantitative data has a narrow range of variance as shown in Figure 3, it will be maintained.

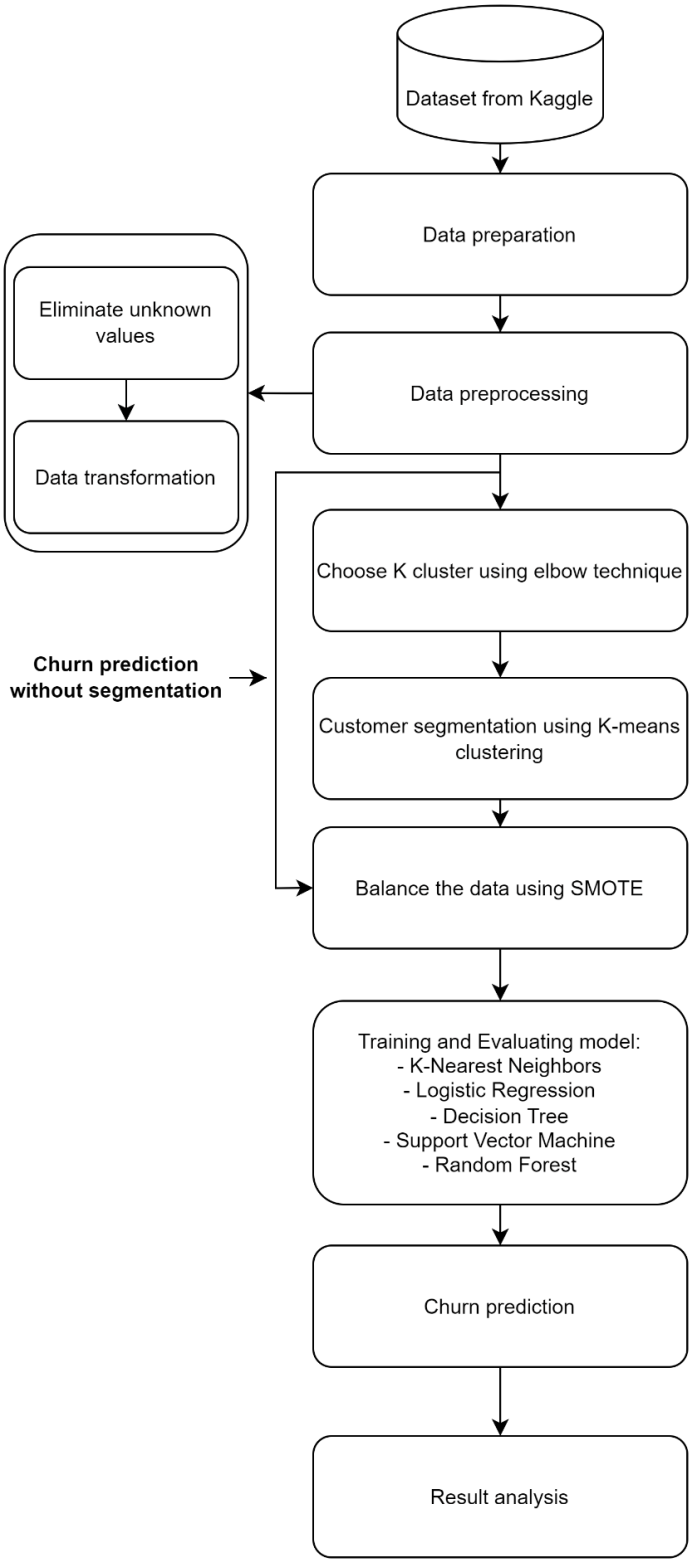
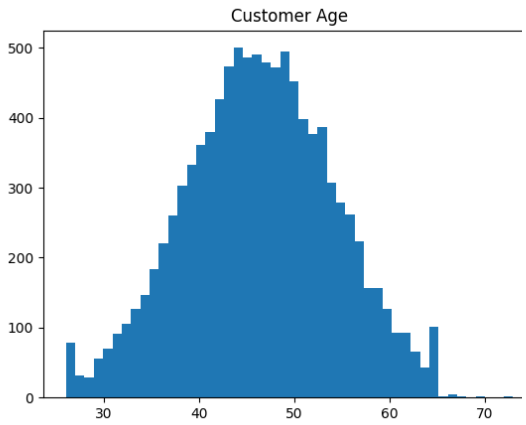
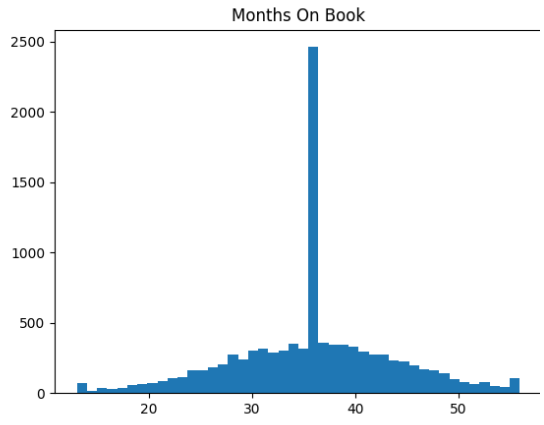


Figure 1. Process of research and implementation

Table 1. Data description

Feature	Description
CLIENTNUM	Client identification number
Attrition_Flag	Account status, Existing = stay, Attrited = churn
Customer_Age	Customer Age
Gender	Customer Gender, M=Male, F=Female
Dependent_count	Number of dependents
Education_Level	Customer education level
Marital_Status	Customer marital status
Income_Category	Customer's Annual Income Category
Card_Category	Type of card
Months_on_book	Relationship duration with the bank
Total_Relationship_Count	Total number of goods owned by the consumer
Months_Inactive_12_mon	Number of inactive months in the last 12 months
Contacts_Count_12_mon	Number of Contacts in the last 12 months
Credit_Limit	The credit limit on the credit card
Total_Revolving_Bal	Total revolving balance on the credit card
Avg_Open_To_Buy	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	The average credit card use ratio





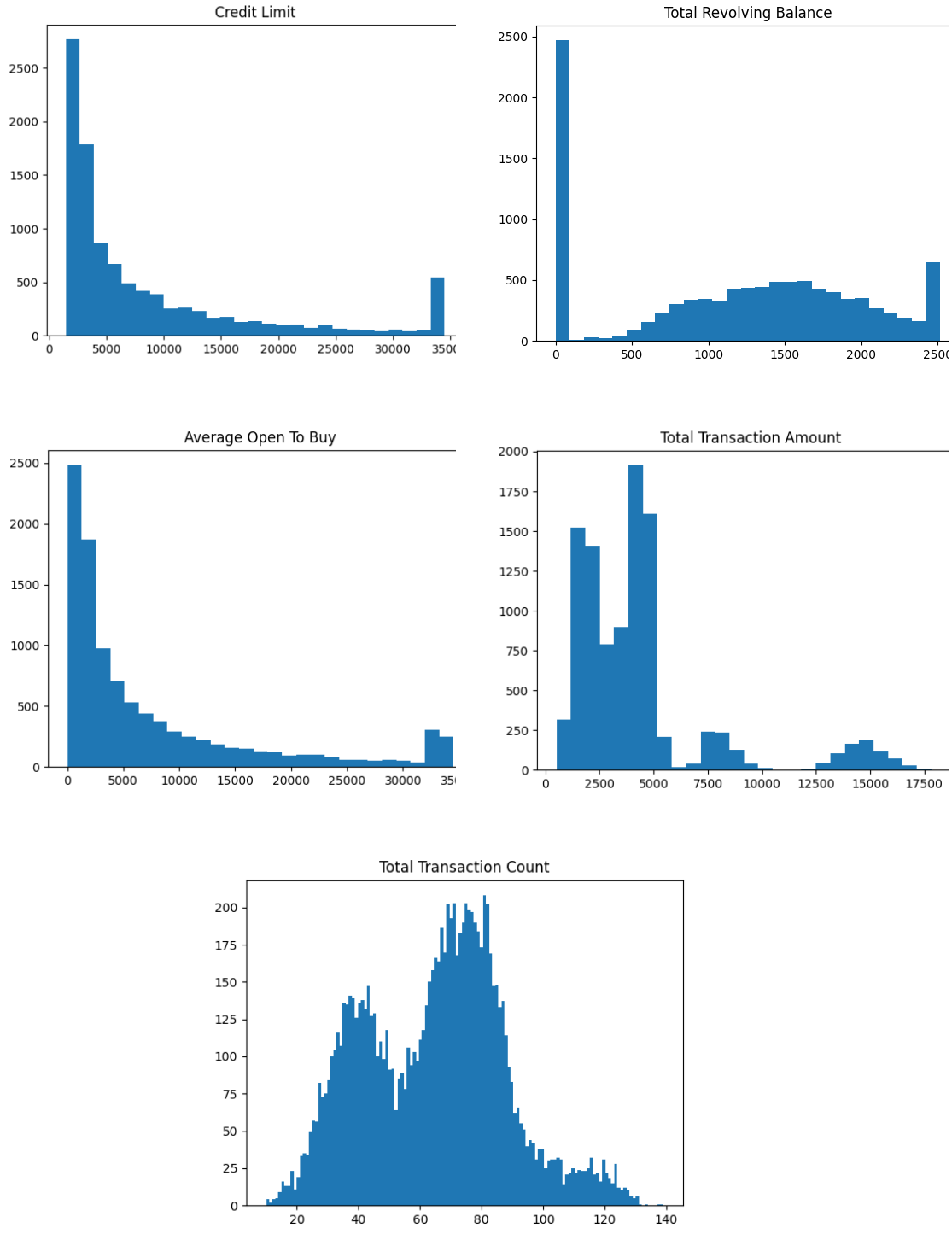


Figure 2. Data distribution

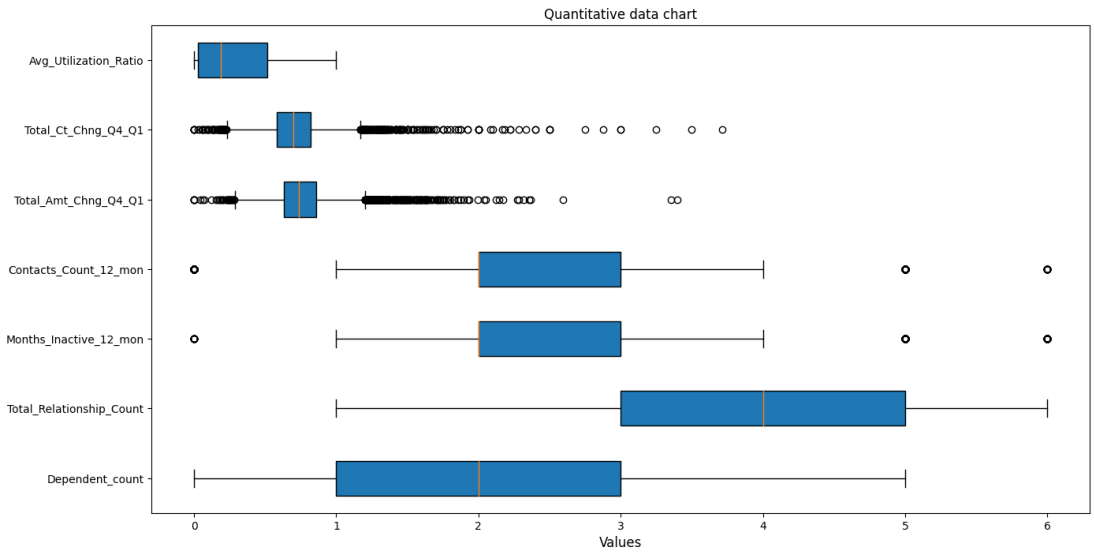


Figure 3. Quantitative data chart

Data Preprocessing

Data preprocessing is a critical stage in machine learning that improves the quality of the data to encourage the extraction of valuable insights from the data (Alexandropoulos et al., 2019). Preparing (cleaning and arranging) raw data to make it acceptable for creating and training Machine Learning models is known as data preprocessing in machine learning. Data preprocessing in machine learning is, to put it simply, a data mining approach that converts raw data into a format that is legible and intelligible. First, the Unknown values discovered in the data preparation phase will be eliminated from the dataset. In the same stage, Customer_Age and Months_on_Book were figured out to have a distribution similar to the normal distribution. Thus, the standardization method will be applied to Customer_Age and Months_on_Book. With the rest of the quantitative data that needs to be transformed, normalization will be used. For the categories with two values, such as Gender and Attrition_Flag, label encoding will then be used. For category data with numerous values, such as Education_Level, Marital_Status, Income_Category, and Card_Category, one hot encoding will be utilized. The final steps include replacing the Divorced value with the Single value and the College value with the Graduate value, as was specified during the data preparation process. A list of the transformation methods we employ for each variable in the dataset is provided in Table 2.

Table 2. Transformation method

Feature	Transformation method
CLIENTNUM	Not used in prediction
Attrition_Flag	Label encoding
Customer_Age	Standardization
Gender	Label encoding
Dependent_count	Unchanged
Education_Level	One-hot encoding
Marital_Status	One-hot encoding
Income_Category	One-hot encoding
Card_Category	One-hot encoding

Feature	Transformation method
Months_on_book	Standardization
Total_Relationship_Count	Unchanged
Months_Inactive_12_mon	Unchanged
Contacts_Count_12_mon	Unchanged
Credit_Limit	Normalization
Total_Revolving_Bal	Normalization
Avg_Open_To_Buy	Normalization
Total_Amt_Chng_Q4_Q1	Unchanged
Total_Trans_Amt	Normalization
Total_Trans_Ct	Normalization
Total_Ct_Chng_Q4_Q1	Unchanged
Avg_Utilization_Ratio	Unchanged

K-MEANS AND CUSTOMER SEGMENTATION

In numerous research, they use k-means clustering to separate customers into several segments (Pradana & Ha, 2021). To increase accuracy, they frequently combine elbow techniques to determine the ideal number of clusters for data division. As a result, the number of clusters is 6 as shown in Figure 4:

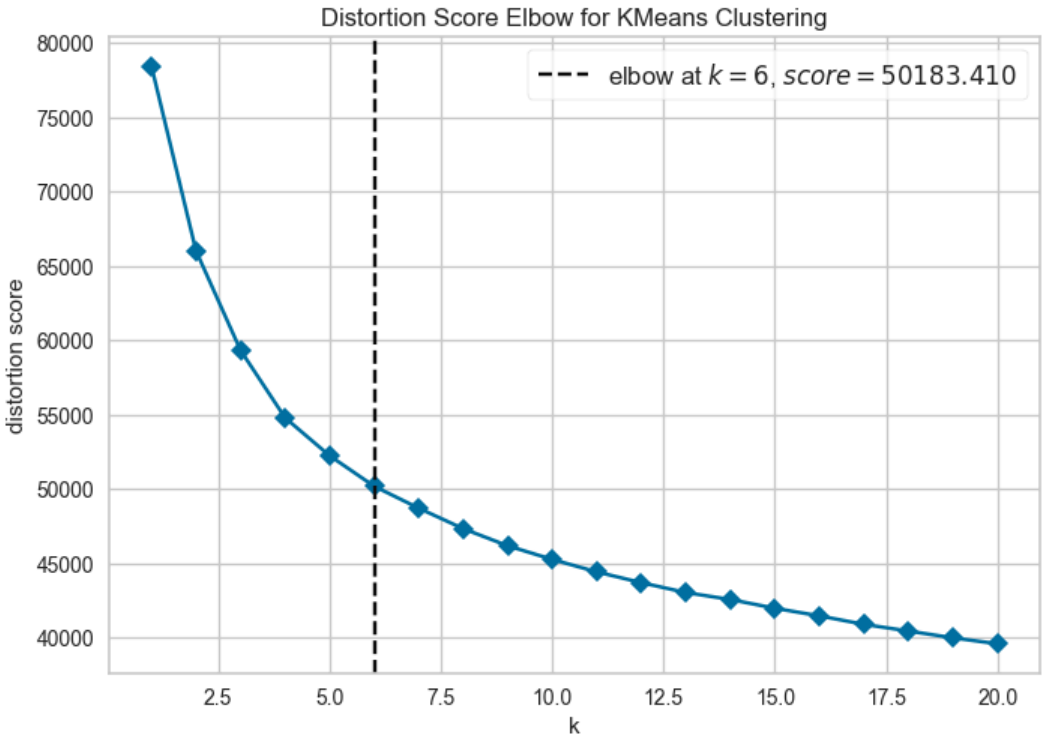


Figure 4. Distortion score elbow for k-means clustering

DATA BALANCING

The results before and after using SMOTE are shown in Tables 3 and 4:

Table 3. The ratio of data before using SMOTE

Dataset	Churn	Non-churn	Ratio
Original	1113	5968	1:5.4
Cluster I	249	840	1:3.4
Cluster II	354	1323	1:3.4
Cluster III	104	930	1:8.9
Cluster IV	182	934	1:5.1
Cluster V	126	820	1:6.5
Cluster VI	98	1121	1:11.4

Table 4. The ratio of data after using SMOTE

Dataset	Churn	Non-churn	Ratio
Original	5968	5968	1:1
Cluster I	840	840	1:1
Cluster II	1323	1323	1:1
Cluster III	930	930	1:1
Cluster IV	934	934	1:1
Cluster V	820	820	1:1
Cluster VI	1121	1121	1:1

RESULT AND DISCUSSION

Data from each cluster and the sample dataset were balanced, then divided into a 70/30 ratio and stratified k-fold cross-validation with k equal to 10 was used to evaluate the model before applying the test set to the models for prediction. The result of each cluster is shown in Table 5.

Remember that accuracy is the proportion of correctly categorized data. The findings shown in Figure 5 demonstrate that the results for the sample and the cluster average using the hold-out approach is 86.26% and 87.57% respectively, indicating that logistic regression has the lowest accuracy. In contrast, random forest produced the greatest results of the five models, with sample accuracy and cluster mean accuracy of 97.4% and 97%, respectively, for the hold-out approach. Three out of the five models had sample accuracy that is higher than the cluster average, which lends credence to the conclusion.

The disadvantage of the accuracy evaluation metric is that it does not explain how each class is classified or which class is classified most properly, it just provides the percentage of data that is correctly classified. Therefore, some other metrics for evaluating the performance of the model that have been used are precision, recall, and the F1 score.

The cluster average precision outcomes for each model are shown in Figure 6. With random forest and support vector machine scores of 96.62% and 96.15% for the sample, and 96.56% and 96.63% for the cluster average respectively, these two models demonstrate their efficacy in accurately identify-

ing clients with churn inclinations when the percentage of actual churn to predicted churn is relatively high. An accuracy of 83.08% for the sample and 84.86% for the cluster average is achieved by the k-nearest neighbor, the least precise of the five models.

However, the k-nearest neighbors had a superior outcome with a recall of 99.15% for the sample and 98.81% for the cluster average in Figure 7, which shows the cluster average recall statistics. Additionally, even if the outcome is only second with a recall of 97.94% and 97.45% corresponding to the sample and cluster average, the random forest still demonstrates that it is a strong model for this data set.

The cluster average F1 score findings, which are the harmonic mean of the model's precision and recall, are displayed in Figure 8. The random forest performed the best on the dataset with a score of 97.43% and 97% matching the sample and cluster average, while the support vector machine came in second with a score of 93.89% and 93.36%.

Table 5. Clusters results with hold out

No.	Model	Cluster	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	KNN	I	89.68	83.78	98.41	90.51
		II	89.8	84.08	98.51	90.72
		III	92.83	88.89	98.63	93.51
		IV	87.52	81.14	98.61	89.03
		V	92.68	88.18	99.62	93.55
		VI	89.6	83.08	99.1	90.38
2	LR	I	88.69	88.24	89.29	88.76
		II	90.81	90.72	90.49	90.6
		III	84.59	85.17	85.17	85.17
		IV	84.67	85.07	85.07	85.07
		V	86.18	86.83	85.43	86.12
		VI	90.49	90.06	91.62	90.83
3	DT	I	92.66	93.68	91.86	92.76
		II	93.45	92.54	94.42	93.47
		III	91.76	89.57	93.61	91.54
		IV	90.02	88.49	92.76	90.57
		V	95.33	95.98	94.84	95.41
		VI	93.76	92.25	96.37	94.26
4	RF	I	95.24	94.47	95.98	95.22
		II	97.48	97.18	97.68	97.43
		III	98.57	98.56	98.56	98.56
		IV	93.76	92.63	94.96	93.78
		V	98.17	98.04	98.43	98.23
		VI	98.81	98.46	99.07	98.77

No.	Model	Cluster	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
5	SVM	I	93.06	96.61	89.41	92.87
		II	91.44	94.97	86.29	90.42
		III	95.7	97.45	94.01	95.7
		IV	91.44	96.33	85.82	90.77
		V	93.9	97.02	90.84	93.83
		VI	96.43	97.4	95.74	96.56

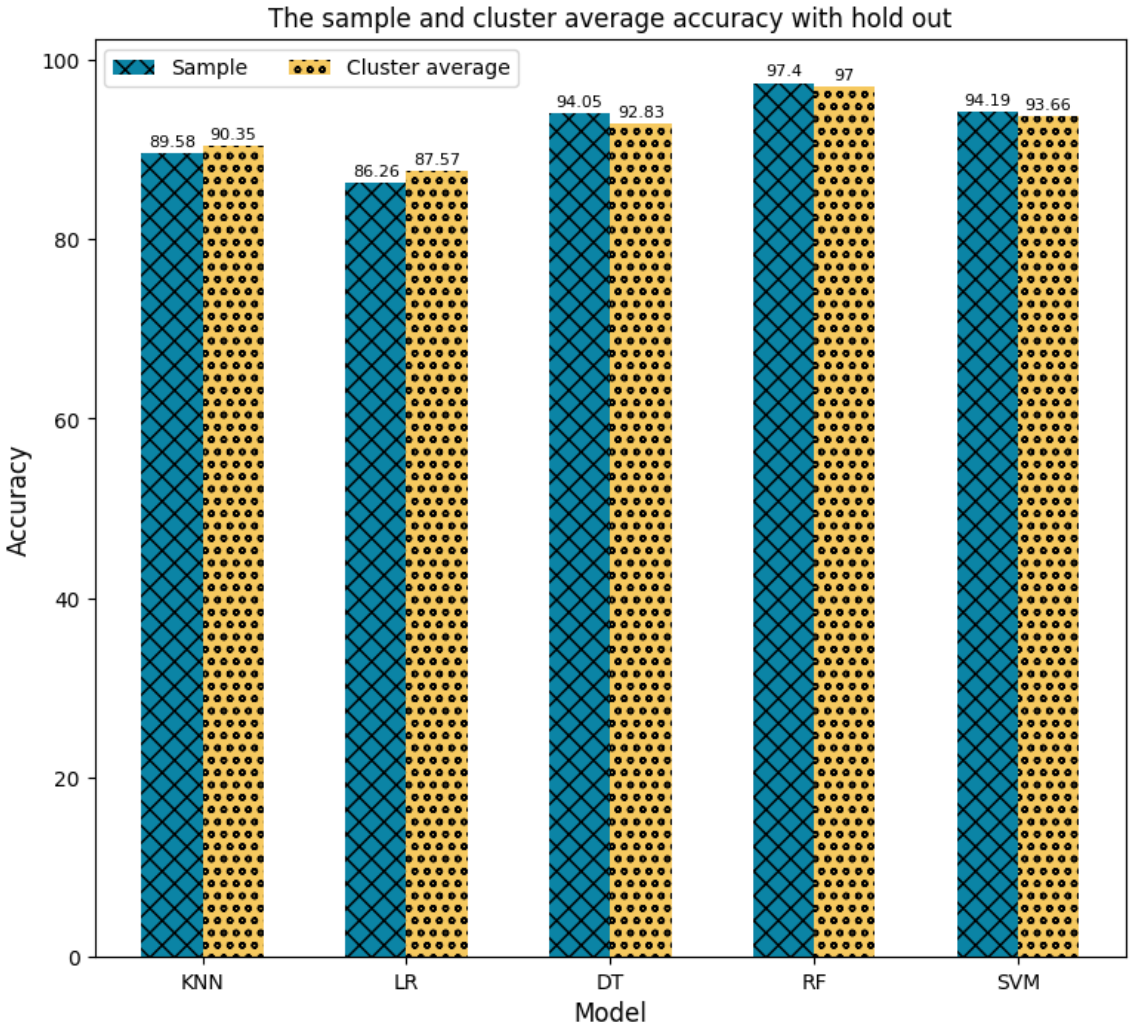


Figure 5. The sample and cluster average accuracy with hold out

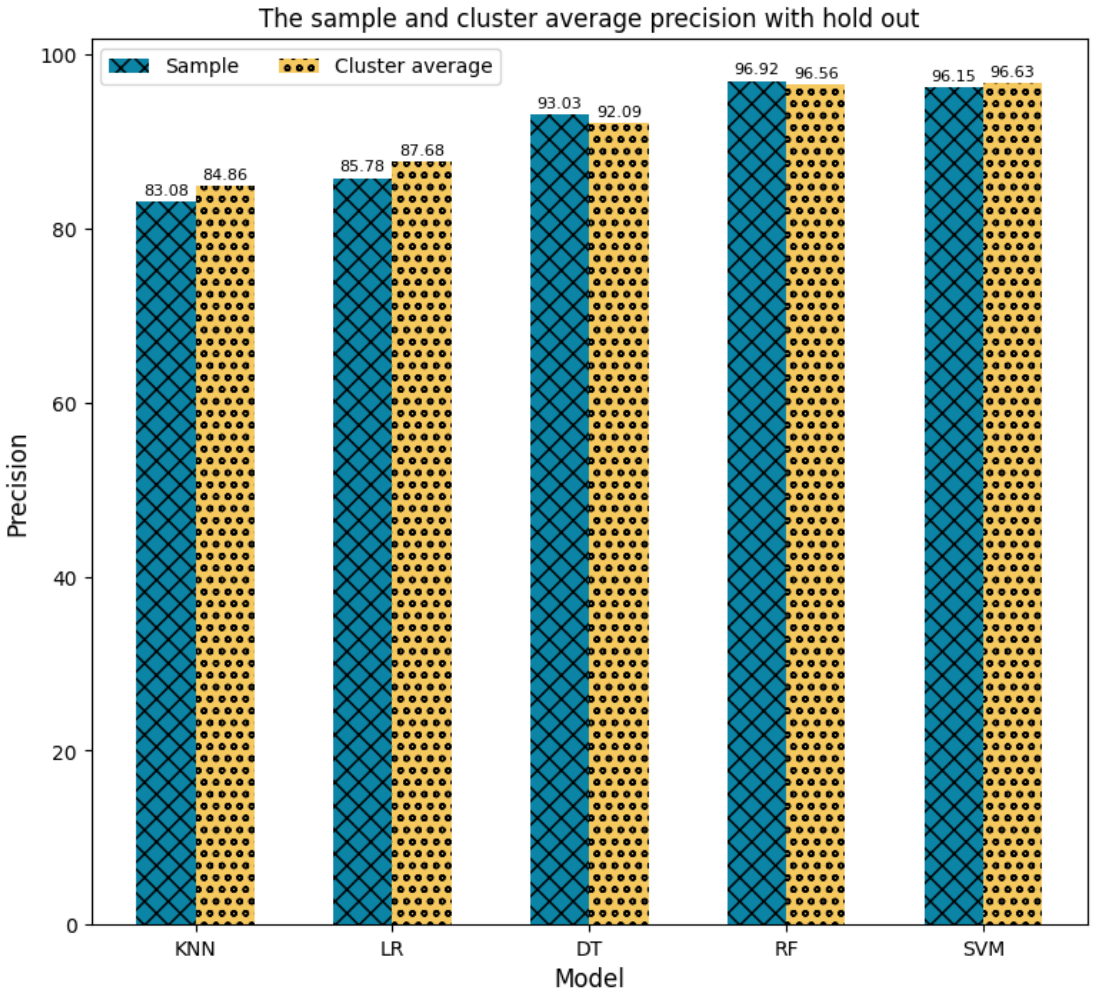


Figure 6. The sample and cluster average precision with hold out

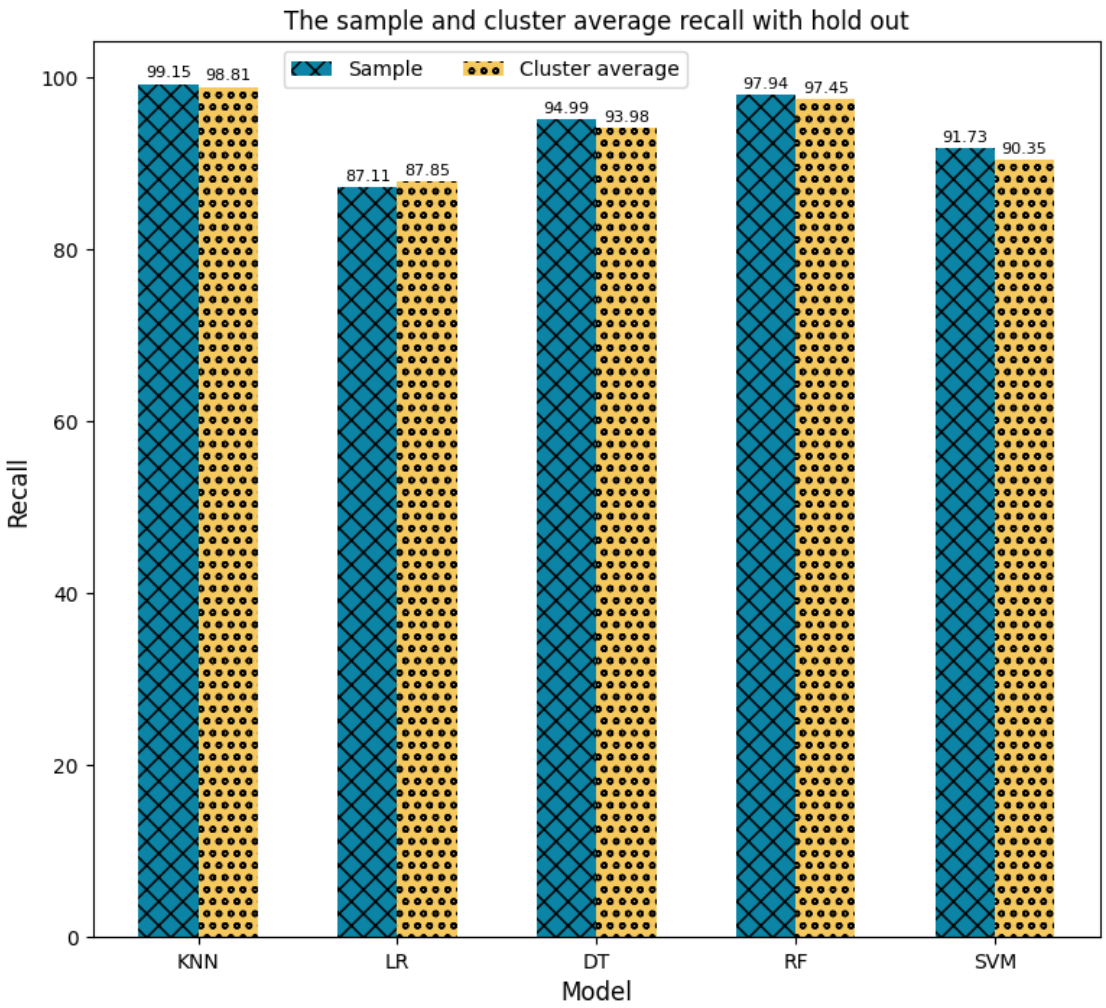


Figure 7. The sample and cluster average recall with hold out

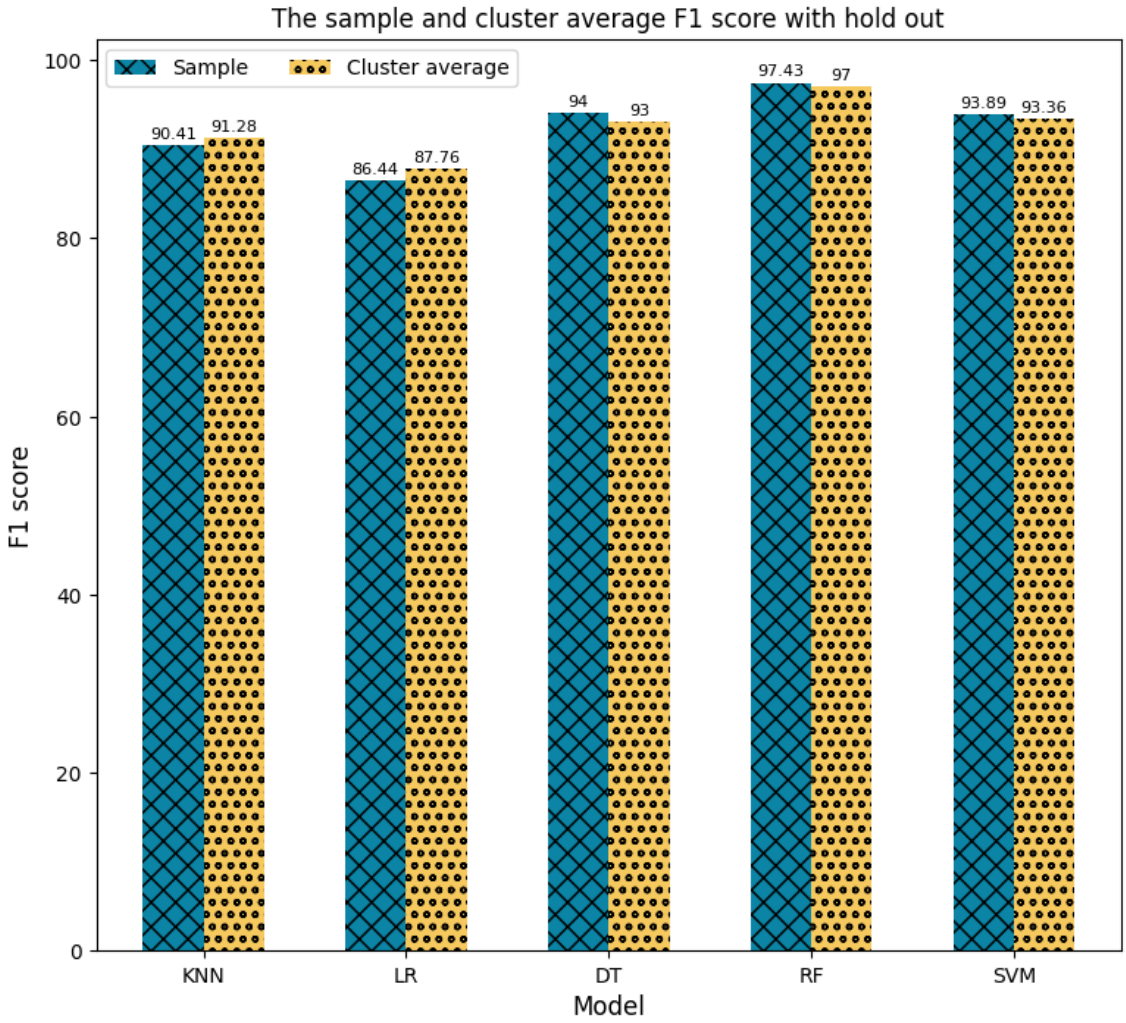


Figure 8. The sample and cluster average F1 score with hold out

The results from this research only go as far as testing, contrasting, and deciding which machine learning technique is best for categorizing customer attrition and how the accuracy of churn prediction is impacted by customer segmentation in the banking industry. Xiahou and Harada's (2022) findings demonstrate that each prediction index improved considerably following customer segmentation, demonstrating the necessity of k-means clustering segmentation. The research, however, only examines the performance of LR and SVM and just on a certain dataset. As a result, it is still not quite clear that segmentation enhances churn prediction. When comparing the outcomes of churn forecasting before and after applying customer segmentation, Zhuang (2018) employed just one model, XG-boost, and that study mainly concentrated on forecasting possible customer groups after clustering. As a result, it cannot be conclusively shown that customer segmentation truly improves the accuracy of churn prediction. Our study examined the effects of customer segmentation on churn prediction using a variety of classification techniques. The results show that the proportion of models with lower churn prediction results after using segmentation with evaluation metrics of accuracy, precision, recall, and F1-score are 3:5, 2:5, 4:5, and 3:5. Thus, it is clear that the churn prediction result depends on the dataset, the model's tuning, and the model's intended usage.

CONCLUSION

This paper reviewed the theoretical basis of customer churn, and customer segmentation, and suggested the use of supervised machine-learning techniques for customer attrition prediction. The experimental findings indicate that the two best training methods are random forest and support vector machines and discovered that customer segmentation has no impact on the ability to anticipate customer attrition. This study serves as a benchmark for the implementation of customer churn prediction in industries including banking, e-commerce, telecommunications, and insurance. However, there are still limitations in this study that may be continued or studied in the future.

First, this study exclusively uses the Kaggle website to collect data for its analysis. By linking to a local bank and assisting them in analyzing their customer attrition, research may also extend to gathering the client's transaction history or customer information.

Second, the study is restricted to determining whether or not a consumer quits the business. The next study will focus on determining which feature is most responsible for customer attrition.

Finally, this research simply looks at how attrition is classified and how customer segmentation influences that outcome. The increased research approach will concentrate on developing an application to help a financial company, such as a bank, keep its customers.

ACKNOWLEDGMENT

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

REFERENCES

Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. *The Knowledge Engineering Review, 34*, 1–33. https://doi.org/10.1017/S026988891800036X

Baran, R. J., & Galka, R. J. (2017). *Customer relationship management: The foundation of contemporary marketing strategy.* Routledge.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Cunningham, P., & Delany, S. J. (2021). k-Nearest neighbour classifiers – A tutorial. *ACM Computing Surveys, 54*(6), 1–25. https://doi.org/10.1145/3459665

Dawes, J., & Swailes, S. (1999). Retention sans frontieres: Issues for financial service retailers. *International Journal of Bank Marketing, 17*(1), 36–43. https://doi.org/10.1108/02652329910254037

Dolatabadi, S. H., & Keynia, F. (2017, July). Designing of customer and employee churn prediction model based on data mining method and neural predictor. *Proceedings of the 2nd International Conference on Computer and Communication Systems, Krakow, Poland*, 74–77. https://doi.org/10.1109/CCOMS.2017.8075270

Elyusufi, Y., & Ait Kbir, M. (2022). Churn prediction analysis by combining machine learning algorithms and best features exploration. *International Journal of Advanced Computer Science and Applications, 13*(7), 615–622. https://doi.org/10.14569/IJACSA.2022.0130773

Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal of Applied Microeconometrics, 1*(2), 85–99. https://doi.org/10.53753/jame.1.2.03

Hameed, M., & Naumann, F. (2020). Data preparation: A survey of commercial tools. *ACM SIGMOD Record, 49*(3), 18–29. https://doi.org/10.1145/3444831.3444835

Jakkula, V. (2006). *Tutorial on support vector machine (SVM).* School of EECS, Washington State University. https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000). The analysis of a simple *k*-means clustering algorithm. *Proceedings of the 16th Annual Symposium on Computational Geometry*, 100–109. https://doi.org/10.1145/336154.336189

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019, October). Customer churn analysis and prediction using data mining models in banking industry. *Proceedings of the International Workshop on Big Data and Information Security, Bali, Indonesia,* 33–38. https://doi.org/10.1109/IWBIS.2019.8935884

Kaur, I., & Kaur, J. (2020, November). Customer churn analysis and prediction in banking industry using machine learning. *Proceedings of the Sixth International Conference on Parallel, Distributed and Grid Computing*, *Waknaghat, India,* 434–437. https://doi.org/10.1109/PDGC50313.2020.9315761

Muneer, A., Ali, R. F., Alghamdi, A., Taib, S. M., Almaghthawi, A., & Abdullah Ghaleb, E. A. (2022). Predicting customers churning in banking industry: A machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science, 26*(1), 539–549. https://doi.org/10.11591/ijeecs.v26.i1.pp539-549

Olaniyi, A. S., Olaolu, A. M., Jimada-Ojuolape, B., & Kayode, S. Y. (2020). Customer churn prediction in banking industry using k-means and support vector machine algorithms. *International Journal of Multidisciplinary Sciences and Advanced Technology, 1*(1), 48–54.

Pradana, M. G., & Ha, H. T. (2021). Maximizing strategy improvement in mall customer segmentation using k-means clustering. *Journal of Applied Data Sciences, 2*(1), 19–25. https://doi.org/10.47738/jads.v2i1.18

Quinlan, J. R. (1990). Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics, 20*(2), 339–346. https://doi.org/10.1109/21.52545

Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology*, *Coimbatore, India,* 1196–1201. https://doi.org/10.1109/ICECA49313.2020.9297529

Reddy, V. N., & Reddy, P. S. S. (2021). Market basket analysis using machine learning algorithms. *International Research Journal of Engineering and Technology, 8*(7), 2570–2572.

Reichheld, F. F. (1996). *The loyalty effect: The hidden force behind growth, profits, and lasting value*. Harvard Business School Press.

Routray, S. K. (2021). Marketing strategy through machine learning techniques: A case study at telecom industry. *International Journal of Innovation Engineering and Science Research, 5*(3), 21–30.

Sivasankar, E., & Vijaya, J. (2017). Customer segmentation by various clustering approaches and building an effective hybrid learning system on churn prediction dataset. In H. Behera, & D. Mohapatra (Eds.), *Computational intelligence in data mining* (pp. 181–191). Springer. https://doi.org/10.1007/978-981-10-3874-7_18

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica, 24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Xiahou, X., & Harada, Y. (2022). B2C e-commerce customer churn prediction based on k-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research, 17*(2), 458–475. https://doi.org/10.3390/jtaer17020024

Yaseen, A. (2021). Next-wave of e-commerce: Mobile customers churn prediction using machine learning. *Lahore Garrison University Research Journal of Computer Science and Information Technology, 5*(2), 62–72. https://doi.org/10.54692/lgurjcsit.2021.0502209

Zhang, T., Moro, S., & Ramos, R. F. (2022). A data-driven approach to improve customer churn prediction based on telecom customer segmentation. *Future Internet, 14*(3), 94. https://doi.org/10.3390/fi14030094

Zhuang, Y. (2018). Research on e-commerce customer churn prediction based on improved value model and XG-boost algorithm. *Management Science and Engineering, 12(3)*, 51-56. https://doi.org/10.3968/10816

AUTHORS

Hoang Tran is currently a sophomore student majoring in E-commerce at the Faculty of Information Systems, University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam. His current research interests include business analytics, business intelligence, data analytics, and machine learning. He can be contacted by email at hoangtd21411c@st.uel.edu.vn

Ngoc Le is currently a sophomore at the Faculty of Economic Mathematics, University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam. Her current research interests include business analytics, data analytics, and machine learning. You can contact her via email at ngocltl21413c@st.uel.edu.vn

Van-Ho Nguyen received a B.S. degree in Management Information System (MIS) from the Faculty of Information Systems, University of Economics and Law (VNU–HCM), Vietnam in 2015, and a Master's degree in MIS from the University of Economics Ho Chi Minh City, Vietnam in 2020. He is currently a lecturer at the Faculty of Information Systems, University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam. His research was published in international journals and conferences, such as the *Journal of Information Processing Systems*, *Business Research Systems*, and *Journal of Hospitality and Tourism Technology*, *9th NAFOSTED Conference on Information and Computer Science* in 2022, and the *4th International Conference on Business (ICB)* in 2021. He is currently a reviewer for the *Journal of Marketing Analytics and Business Research Systems*. His current research interests include business analytics, business intelligence, data analytics, and machine learning. He can be contacted by email at honv@uel.edu.vn or nguyenvanho.uel@gmail.com