Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 20, 2025

APPLICATION OF MACHINE LEARNING TECHNIQUES FOR CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR

Tam-Thanh Luong*	University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam	lthanhtam915@gmail.com
Vi-Gia Luong	University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam	viluong.31231024175@st.ueh.edu.vn
Anh Hoang Tuan Tran	University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam	anhtran.31231024605@st.ueh.edu.vn
Tuan Manh Nguyen	University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam	tuannm@ueh.edu.vn

* Corresponding author

ABSTRACT

Aim/Purpose	Previous studies have primarily focused on comparing predictive models without considering the impact of data preprocessing on model performance. Therefore, this study sets two main objectives. The first objective is to investigate the effect of resampling methods for handling imbalanced data on model effectiveness. The second objective is to compare and evaluate machine learning methods to identify the optimal model for each resampling technique, thereby determining the model that achieves the highest performance.
Background	In the highly competitive banking industry, attrition of customers is a major challenge for banks trying to improve customer retention. While many studies have focused on building and evaluating models to predict customer churn, they often miss addressing the problem of imbalanced data, which can significantly affect the model's accuracy.

Accepting Editor Dimitar Grozdanov Christozov | Received: November 24, 2024 | Revised: January 31, February 9, February 10, February 20, March 1, 2025 | Accepted: March 7, 2025.
Cite as: Luong, T.-T., Luong, V.-G., Tran, A., & Nguyen, T. M.. (2025). Application of machine learning techniques for customer churn prediction in the banking sector. *Interdisciplinary Journal of Information, Knowledge, and Management, 20,* Article 9. https://doi.org/10.28945/5469

Methodology	In this study, following exploratory data analysis (EDA), we apply various techniques to address data imbalance and use a range of machine learning models, including Naïve Bayes, Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM, to predict customer churn using the dataset.
Contribution	The contribution of this research lies in its comprehensive evaluation and comparison of various techniques for handling imbalanced data in churn prediction models. The study identifies SMOTE-ENN as the most effective method for resampling imbalanced data. Among the models tested, LightGBM (accuracy = 0.979) achieves the highest performance based on evaluation metrics. Additionally, the research highlights that tree-based machine learning models generally perform better when trained on imbalanced datasets.
Findings	Tree-based and ensemble models perform better than regression and probability-based methods when dealing with imbalanced data. SMOTE-ENN has been shown to improve the performance of machine learning models greatly.
Recommendations for Practitioners	Practitioners can deploy high-performance models, such as XGBoost and LightGBM, combined with effective resampling methods like SMOTE-ENN to predict customer churn in banking, marketing, and human resources.
Recommendations for Researchers	To optimize the predictive model in the study, researchers can focus on feature selection, dimensionality reduction, or hyperparameter tuning.
Impact on Society	Customer churn reduces revenue and threatens competitive advantage, so businesses need effective retention strategies to maintain sustainable growth. High-performance customer churn prediction models can be an effective solution to address this issue.
Future Research	Deploy the model on real-world datasets while further optimizing the feature selection process and hyperparameter tuning, combined with SHAP values analysis to identify key features that significantly influence the model's predictions.
Keywords	churn prediction, machine learning, imbalanced data, classification models, oversampling, undersampling, hybrid method, banking industry

INTRODUCTION

Customer churn, also known as customer attrition, is a critical issue in the banking sector, especially in the context of Vietnam's rapidly growing credit card market, driven by digital transformation and the trend of cashless payments. Churn occurs when customers stop using services or switch to other providers, usually due to dissatisfaction with service quality or being attracted by better offers from competitors.

According to Yuan and Tao (2023), the average annual percentage rate (APR) on credit cards in the United States ranges between 16% and 30%. In another study (Gerritsen & Bikker, 2020), it was found that, on average, individuals transfer 6.1% of their total savings to a bank offering an interest rate that is 1% higher than their current bank, highlighting the competitive nature and the high risk of customer churn in this industry.

Churn affects revenue and increases the cost of acquiring new customers, which is five times higher than retaining existing ones (Dawes & Swailes, 1999). Additionally, according to Reichheld and Sasser (1990), a mere 5% increase in customer retention can boost profits by 25% to 95%, depending on the industry. In the face of growing competition, predicting churn to proactively intervene, improve services, and optimize promotional policies has become critical.

To address this issue, many previous studies have compared and evaluated machine learning models to predict customer churn. Commonly used models include Logistic Regression, Random Forest, XGBoost, and Gradient Boosting. These tools help to accurately identify customers at risk of churn and serve as a foundation for developing effective customer retention strategies, from improving service quality to personalizing promotions. However, due to the impact of imbalanced data structures (where the majority class consists of retained customers), previous studies often achieved low evaluation metrics.

This study focuses on addressing the issue of data imbalance in the dataset to enhance the performance of the prediction model. The resampling methods used include OverSampling (Random OverSampling, SMOTE, Border-SMOTE, and ADASYN), UnderSampling (Random UnderSampling, NearMiss), Hybrid Methods (SMOTE-ENN, SMOTE Tomek Links), and Class Weight.

The contributions of this paper are described as follows:

1. Identifying the most suitable classification algorithm for each resampling method.
2. Developing a comprehensive model to support customer churn prediction in the banking sector.
3. Analyzing and explaining the robustness of tree-based models in handling imbalanced data.
4. The SMOTE-ENN resampling method demonstrates the highest effectiveness due to the combination of SMOTE and ENN.

Research on customer churn not only helps increase profits but also contributes to improving service quality, minimizing risks, and promoting sustainable development for the banking sector in Vietnam. Advances in machine learning not only support churn prediction but also open up opportunities to better understand customer behavior, thereby optimizing long-term business strategies.

The structure of this paper is as follows. The next section reviews the background of the study along with related research. This is followed by an overview of the theoretical framework, data, and methodology. Then, the results of the model are presented, accompanied by a discussion. Finally, the paper addresses the limitations of the research.

LITERATURE REVIEW

RELATED WORK

Previous studies on customer churn prediction using machine learning have explored various techniques and applications across different sectors, such as banking, telecommunications, and retail. One study by Rahman and Kumar (2020) utilized behavioral data from 10,000 customers on Kaggle, including 2,037 churn samples. Feature selection methods like mRMR and Relief were employed to improve model performance. The Random Forest algorithm, combined with Random Oversampling, achieved the highest accuracy of 95.74%, demonstrating the effectiveness of handling imbalanced data in churn prediction.

In the telecommunications sector, Aggarwal and Vijayakumar (2024) used data from 7,043 customers on Kaggle, with 1,869 churn samples. SMOTE-ENN was applied to balance the data, and algorithms like Random Forest and Decision Tree were tested. The results showed that Random Forest had the best performance, achieving an accuracy of 99%. In another study, Amin et al. (2020) used a Just-in-Time (JIT) prediction method, applying models like SVM and kNN, along with ensemble models such as stacking, achieving high flexibility and performance from data across companies in the same industry.

Peng et al. (2023) applied the GA-XGBoost model combined with SHAP techniques to analyze the impact of features on churn. The SMOTE-ENN technique was used to handle imbalanced data, achieving an AUC of 0.9912 and enhancing the interpretability of prediction results. Long et al.

(2019) also combined demographic segmentation and customer behavior. The Gradient Boosting algorithm demonstrated the best fit for churn prediction on banking data, improving accuracy and reducing overfitting risks. Other studies, along with the data resampling method, are presented in Table 1.

Table 1. Previous research

Reference	Year	Resampling method
(Keramati et al., 2016)	2016	Boostrap
(Rajamohamed & Manokaran, 2018)	2017	None
(Amin et al., 2020)	2017	Random Under-Sampling
(Vijaya & Sivasankar, 2019)	2017	Random Under-Sampling
(Lalwani et al., 2022)	2021	Resampling
(Xiahou & Harada, 2022)	2022	SMOTE
(Ahmad et al., 2019)	2019	Smote
(Dong et al., 2020)	2019	None
(Domingos et al., 2021)	2021	None
(Vo et al., 2021)	2021	SMOTE, SMOTEENN, SMOTETomek, SVM-SMOTE, and Borderline-SMOTE
(Pustokhina et al., 2023)	2021	ISMOTE
(Al-Najjar et al., 2022)	2022	None
(Amin et al., 2019)	2018	None
(de Lima Lemos et al., 2022)	2022	None
(Bharathi et al., 2022)	2022	None
(Y et al., 2022)	2022	SMOTE Tomek, SMOTE ENN
(Peng et al., 2023)	2023	ADASYN, SMOTE, SMOTEENN
(Wagh et al., 2024)	2024	SMOTE & ENN
(Rahman & Kumar, 2020)	2020	Over-Sampling and Under-Sampling

Research Gap

Studies related to customer churn prediction in the banking sector are summarized in Table 1. While numerous studies and machine learning models have been used to predict customer churn, handling imbalanced data has not received adequate attention. Previous research has relied mainly on random methods for addressing imbalanced data without solid theoretical foundations. There is a need for more specific research that compares and analyzes these techniques to assess their effectiveness in handling imbalanced data in churn prediction.

Resampling Methodology

Table 2 presents the resampling methods employed in this study.

Table 2. Resampling methodology

Methodology	Detailed methodology	Description
Over-Sampling	Random Over-Sampling	Random Over-Sampling is a technique for handling class imbalance by increasing the number of samples in the minority class through random duplication of existing samples from this class (He & Garcia, 2009).

Methodology	Detailed methodology	Description
	SMOTE	SMOTE (Synthetic Minority Over-Sampling Technique) is a technique that generates synthetic data samples for the minority class based on an interpolation method (Nitesh, 2002).
	Border - SMOTE	Borderline-SMOTE is an advanced variant of SMOTE that enhances data generation in regions prone to misclassification (decision boundary) rather than randomly generating new samples across the entire minority class (Han et al., 2005).
	ADASYN	Adaptive Synthetic (ADASYN) is a technique for increasing minority class samples based on density distribution, thereby generating synthetic samples from the minority class (He et al., 2008).
Under-Sampling	Random Under-Sampling	Random Under-Sampling (RUS) is a sampling technique used to improve class imbalance by randomly removing samples from the majority class (Zuech et al., 2021).
	NearMiss	This is a method based on k-nearest neighbors (k-NN). The Euclidean distance can be used as a distance metric to remove majority class samples (Zuech et al., 2021).
Class-Weight		Class-Weight refers to assigning different weights to classes in a dataset, particularly when the dataset is imbalanced (Ghosh et al., 2024).
Hybrid Methods	SMOTE – ENN	This sampling technique combines oversampling and under-sampling methods by increasing minority class samples through interpolation and then removing redundant samples using the ENN (Edited Nearest Neighbor) method (Muntasir Nishat et al., 2022).
	SMOTE Tomek Link	This technique combines SMOTE and Tomek Link, where T-link removes majority-class samples close to the minority class using the nearest neighbor rule to select the samples (Swana et al., 2022).

PROPOSED RESEARCH MODEL

All experiments were conducted on a Dell Inspiron 5510 running Windows 11, a 64-bit operating system, with a Core i5-11300H processor, 12GB RAM, and a 512GB SSD. All source code was implemented in Python 3.11, and Visual Studio Code was used for coding. The libraries used include NumPy, Pandas, scikit-learn, imbalanced-learn, SciPy, Matplotlib, and Seaborn. The entire dataset is available at https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers.

According to Figure 1, the research methodology will encompass several stages, including data collection, preparation, preprocessing, feeding data into machine learning models, and evaluating the results. Data preprocessing will involve removing missing or noisy values, converting categorical data into numerical format (using One-Hot Encoding), and normalizing or standardizing the data to bring it to a consistent scale. Before the data is fed into the prediction models, various techniques for handling imbalanced data will be applied, including OverSampling (Random Oversampling, SMOTE, BorderLine-SMOTE, ADASYN), UnderSampling (Random UnderSampling, NearMiss), Class Weight adjustment, and hybrid methods (SMOTE TomekLink, SMOTEENN). The machine learning models employed will include Naive Bayes, Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM. The results will be evaluated using a confusion matrix and metrics such as accuracy, precision, recall, and F1 score.

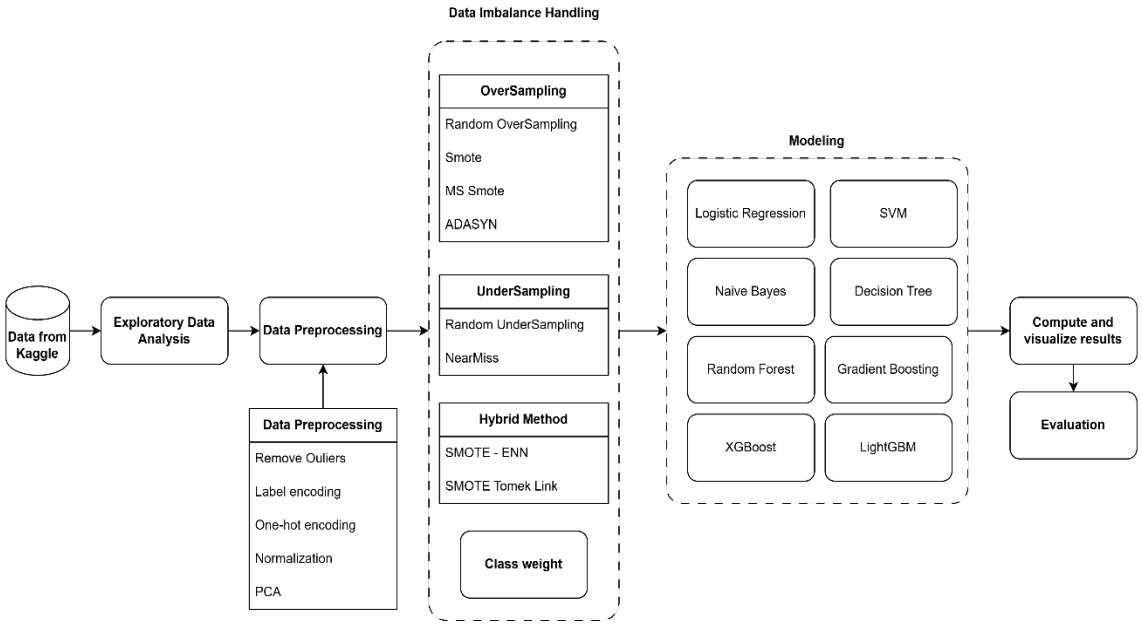

Figure 1. Process of research and implementation

DATA PREPARATION

The study uses a credit card customer dataset published on Kaggle by Sakshi Goyal (Goyal, 2021). After the preprocessing step, the dataset contains 10,127 observations with a total of 20 variables, including one dependent variable and 19 independent variables. The dependent variable is the Attrition Flag, a binary variable where a value of 1 indicates the customer has stopped using the credit card, and a value of 0 indicates the customer continues to use the credit card. Detailed data descriptions are provided in Appendix A.

Once the data is collected, an exploratory data analysis (EDA) will be performed to uncover patterns and insights. Following this, data preprocessing will take place, which includes handling missing values, removing highly correlated variables, normalizing the data, applying one-hot encoding, and carrying out normalization and standardization. After preparing the data, it will be fed into the model for training. The results will then be evaluated, visualized, and compared using key performance metrics such as the confusion matrix, accuracy, precision, recall, and F1-score.

EXPLORATORY DATA ANALYSIS

Customers use four types of credit cards: Blue, Silver, Gold, and Platinum. The bar chart (Figure 2) illustrates the distribution of customers across these card types, with the Blue card being the most commonly used by 9,436 customers. The Silver card has 555 customers, the Gold card has 116 customers, and the Platinum card is the least common, with only 20 customers. This uneven distribution indicates that the Blue card is the most popular choice, while higher-tier cards like Gold and Platinum have significantly fewer customers.

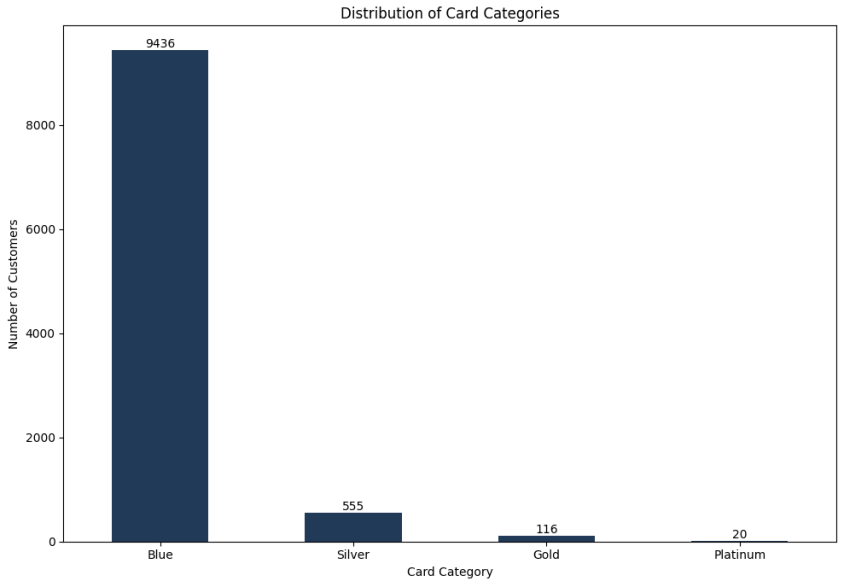

Figure 2. Distribution of card categories

The data in Table 3 indicates that 53% of customers are female, while 47% are male. The most common number of dependents is 3, followed by 2 and 1. Graduate is the most prevalent education level, followed by High School and No Education. The most common marital status is Married, followed by Single. In terms of income, the most common group is Less Than $40K, followed by the group Between $40K and $60K.

Table 3. Demographic variables description

Variables	Description	Values	Frequency	Percentage
Gender	F = Female M = Male	F	5358	53
		M	4769	47
Dependent_count	Number of dependents	0	904	8.93
		1	1838	18.15
		2	2655	26.22
		3	2732	26.98
		4	1574	15.54
		5	424	4.19
Education_Level	Educational qualification	Doctorate	451	4.45
		Postgraduate	516	5.1
		Graduate	3128	30.89
		College	1013	10
		High School	2013	19.88
		Uneducated	1487	14.68
		Unknown	1519	15
Marital_Status		Married	4687	46.28
		Single	3943	38.94
		Divorced	748	7.39
		Unknown	749	7.4

Variables	Description	Values	Frequency	Percentage
Income_Category	Annual income	$120K +	727	7.18
		$80K - $120K	1535	15.16
		$60K - $80K	1402	13.84
		$40K - $60K	1790	17.68
		Less than $40K	3561	35.16
		Unknown	1112	10.98

Figure 3 shows that the number of female customers slightly exceeds that of male customers, with no significant difference in attrition rates between the two genders. The group with two dependents comprises the majority, and the attrition rates within these groups are relatively uniform. The "Graduate" education level holds the highest proportion, with the highest attrition rate observed within this group. Regarding marital status, the "Married" group accounts for a larger customer base, and its attrition rate is higher than that of the "Single" group. The income group under $40K has the largest number of customers, but attrition rates decrease as income increases. Finally, the "Blue" card is the most widely used, with a substantial number of both departing and retained customers. Premium cards such as "Gold" and "Platinum" exhibit lower attrition rates.

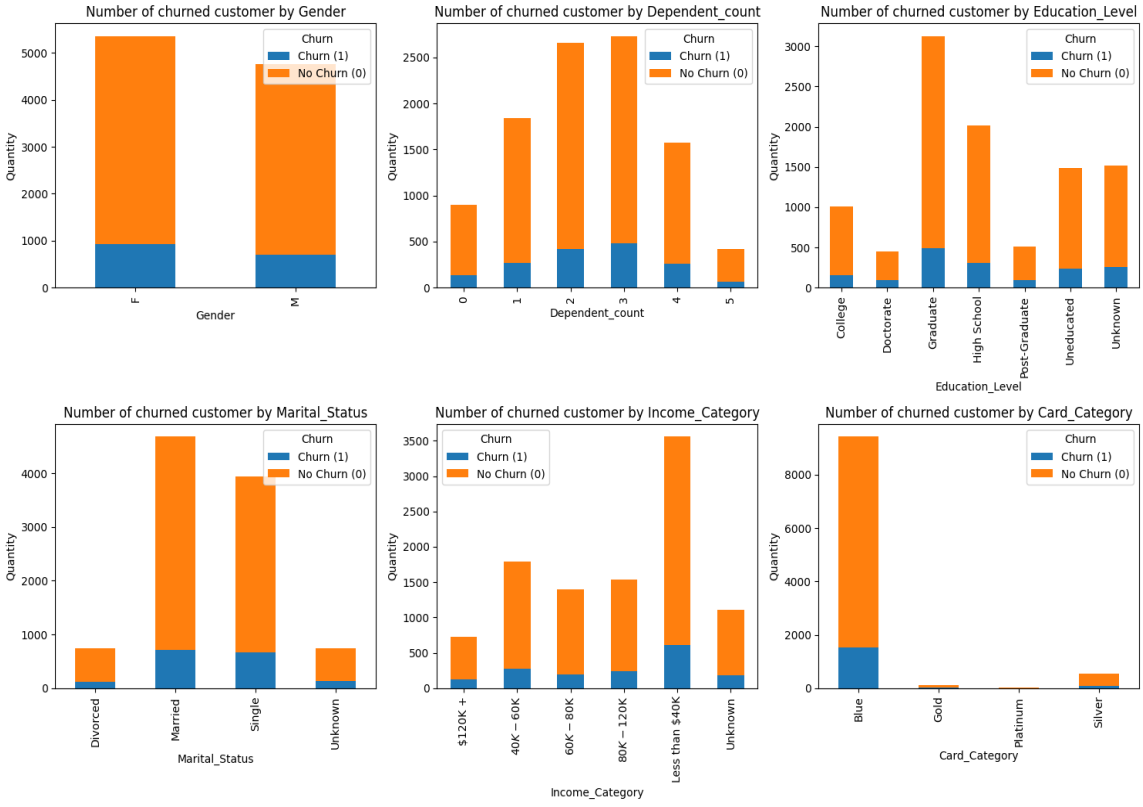

Figure 3. Customer churn by demographics

The scatter plot (Figure 4) reveals a distinct distribution between the two customer groups: existing and churned customers. Notably, churned customers (orange) tend to exhibit higher values for the change in transaction count (Total_Ct.Chng_Q4_01) compared to existing customers (blue). This suggests that churned customers often experience greater fluctuations in their transaction behavior, which may indicate dissatisfaction or changing service needs. In contrast, existing customers, who

exhibit less variation, demonstrate more stability in their transaction patterns. This visualization highlights the distinct differences between the two customer groups, providing insights that banks can leverage to refine and enhance customer retention strategies.

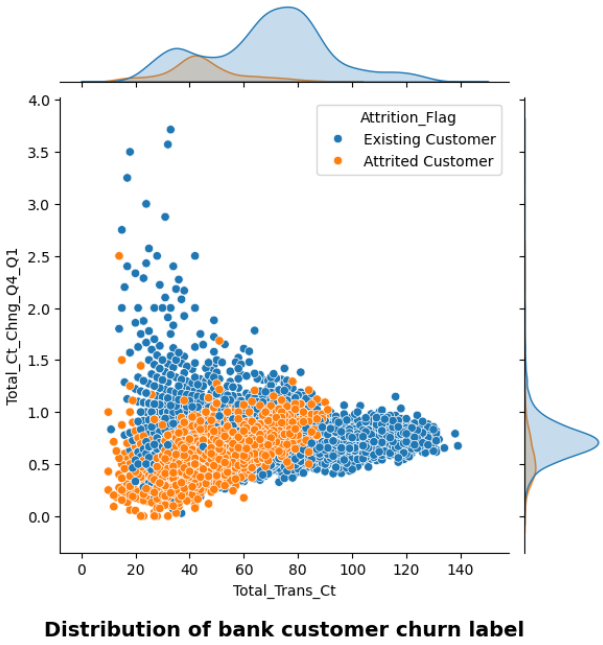

Figure 4. Distribution of bank customer churn label

DATA PREPROCESSING

Data preprocessing is an essential step in the machine learning process, helping to improve data quality and, in turn, enhance the ability to extract useful information (Jain et al., 2018). This process involves cleaning and preparing raw data to be applied in the construction and training of machine learning models. Simply put, data preprocessing is a data mining method aimed at transforming raw data into a form that is easier to understand and process. First, undefined values during the data preparation stage will be removed from the dataset. Then, the variables Customer_Age and Months_on_Book, which have distributions close to a normal distribution, will be normalized.

For the remaining quantitative variables, the normalization method will be applied. Categorical variables with two values, such as Gender and Attrition_Flag, will be label encoded, while categorical variables with multiple values, such as Education_Level, Marital_Status, Income_Category, and Card_Category, will use one-hot encoding. Table 4 lists the transformation methods applied to each variable in the dataset.

Table 4. Data preprocessing methods

Feature	Transformation method
CLIENTNUM	Not used in prediction
Attrition_Flag	Label encoding
Customer_Age	Standardization
Gender	Label encoding
Dependent_count	Unchanged

Feature	Transformation method
Education_Level	One-hot encoding
Marital_Status	One-hot encoding
Income_Category	One-hot encoding
Card_Category	One-hot encoding
Months_on_book	Standardization
Total_Relationship_Count	Unchanged
Months_Inactive_12_mon	Unchanged
Contacts_Count_12_mon	Unchanged
Credit_Limit	Not used in prediction
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

In Figure 5 and Figure 6, "Credit_Limit" and "Avg_Open_To_Buy" have an almost perfect positive correlation (0.996), indicating that when one value increases, the other also increases. This is logical as the available credit limit will increase when the credit limit is higher. The decision to remove the "Credit_Limit" variable will help mitigate multicollinearity issues with "Avg_Open_To_Buy," which has a high correlation. This will improve the stability of the model and increase the accuracy of predictions by reducing redundant information. Retaining "Avg_Open_To_Buy" makes sense, as this variable may provide more direct insight into the customer's available credit, which is likely more relevant to our analysis.

The data was prepared to enhance the performance of the prediction model. Two unnecessary variables were removed, reducing the number of variables from 23 to 21. A data normalization process was performed to bring the data points closer together, improving the speed and efficiency of the machine learning algorithm. Normalization was applied to scale the data within the range [0, 1], while Z-score standardization was used to center the data at 0 with a standard deviation of 1. Quantitative variables with wide distributions were normalized, while variables with narrow ranges were left unchanged.

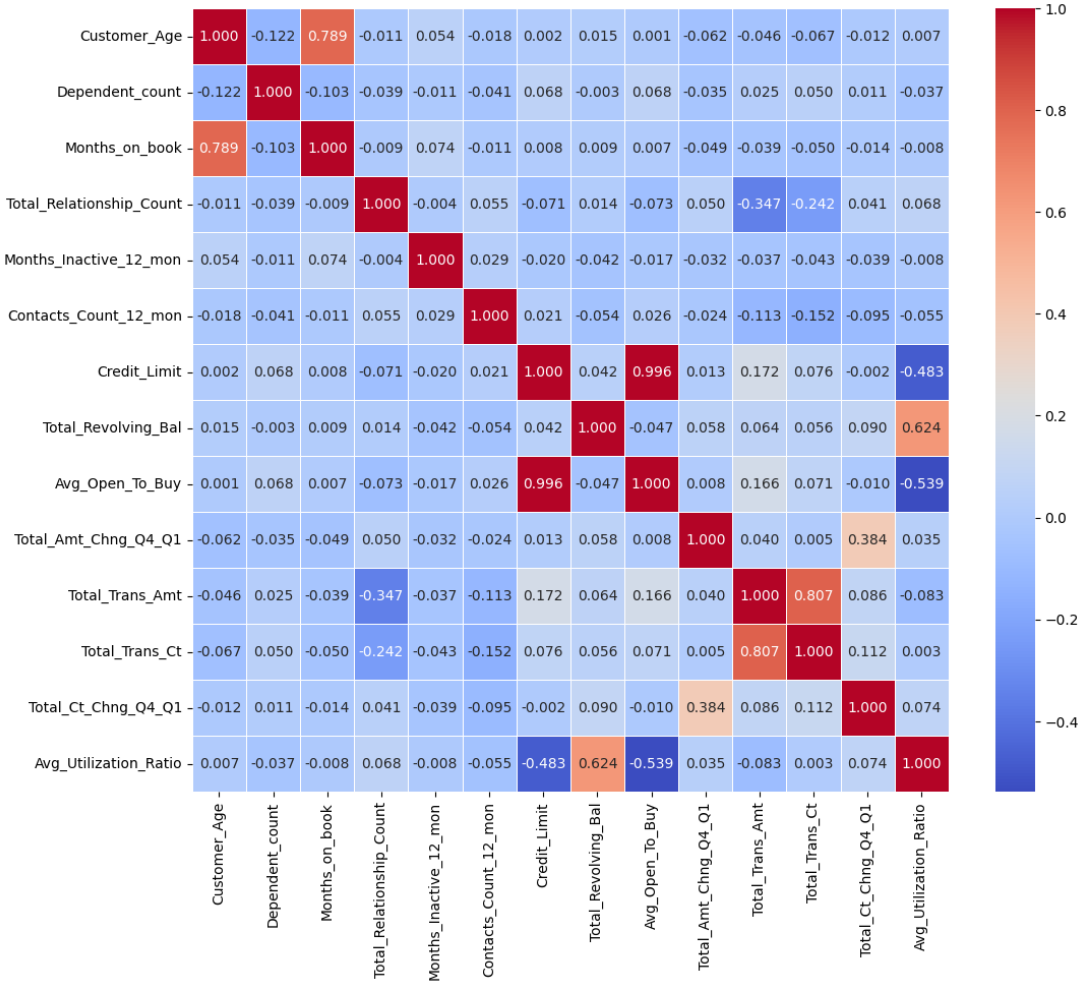

Figure 5. Correlation of variables

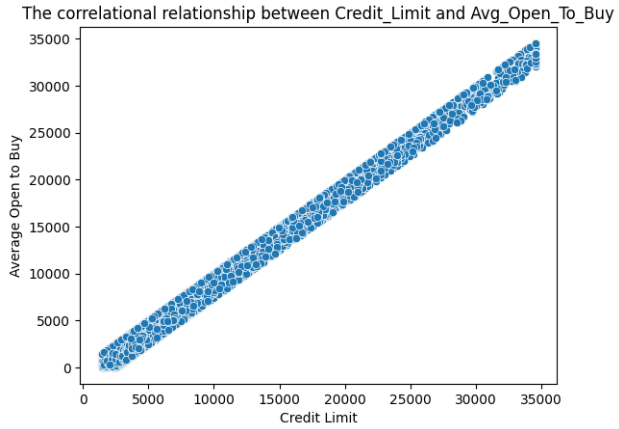

Figure 6. Correlation between "credit limit" and "average open to buy"

RESULT AND DISCUSSION

Table 5 provides the list of variables included in the model, where Attrition_Flag (the classification variable) is the dependent variable, and the other variables serve as independent variables.

Table 5. Dataset features

STT	Feature	Description
1	Attrition_Flag	Dependent Variables
2	Customer_Age	Independent Variables
3	Gender	Independent Variables
4	Dependent_count	Independent Variables
5	Education_Level	Independent Variables
6	Marital_Status	Independent Variables
7	Income_Category	Independent Variables
8	Card_Category	Independent Variables
9	Months_on_book	Independent Variables
10	Avg_Open_To_Buy	Independent Variables
11	Total_Relationship_Count	Independent Variables
12	Months_Inactive_12_mon	Independent Variables
13	Contacts_Count_12_mon	Independent Variables
14	Total_Revolving_Bal	Independent Variables
15	Total_Amt_Chng_Q4_Q1	Independent Variables
16	Total_Trans_Amt	Independent Variables
17	Total_Trans_Ct	Independent Variables
18	Total_Ct_Chng_Q4_Q1	Independent Variables
19	Avg_Utilization_Ratio	Independent Variables

After training and evaluating the models using K-Fold cross-validation, the study compares and assesses the results based on three aspects:

1. Evaluating resampling methods for handling data imbalance to identify the most effective approach.
2. Comparing the average evaluation metrics to determine the optimal model across all scenarios.
3. Analyzing the performance differences between tree-based and non-tree-based models in the context of imbalanced data without resampling.

EVALUATION OF RESAMPLING METHODOLOGY

According to Figure7, the results show that the Hybrid method outperforms other methods, such as Over-Sampling, Under-Sampling, and Classweight, across all three metrics: Accuracy, Recall, and F1-Score. Specifically, all three metrics exceed 0.93. Additionally, based on Figure 7 and Table 6, among the Hybrid methods, SMOTE-ENN demonstrates better performance than SMOTE Tomek Link across all four metrics: Accuracy, Precision, Recall, and F1-Score, with values of 0.941, 0.948, 0.955, and 0.952, respectively. Notably, the runtime of SMOTE-ENN is only one-third of SMOTE Tomek Link's (9.3 seconds vs. 27.6 seconds).

In the case of no imbalance handling and the Class weight method, the results show instability. Although it achieves high Accuracy, the Recall and Precision metrics for the unbalanced data indicate that, without resampling, the model prioritizes overall Accuracy and tends to make incorrect predictions for the majority class. This leads to the model being unable to accurately predict the churned customers, reducing overall effectiveness.

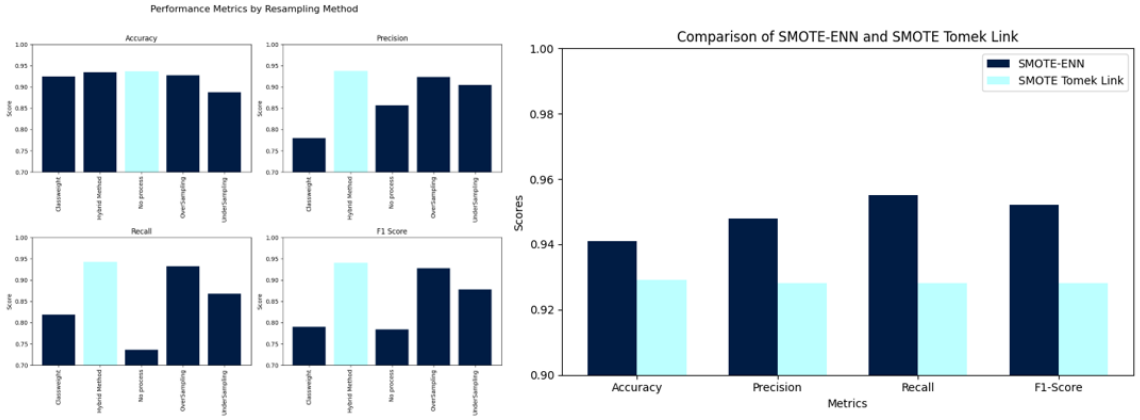

Figure 7. Average metrics:
accuracy, precision, recall, F1-score (imbalance handling methods)

Table 6. Average metrics of SMOTE-ENN and SMOTE Tomek Link

Method	Accuracy	Precision	Recall	F1-Score	Runtime(s)
SMOTE-ENN	0.941	0.948	0.955	0.952	9.3
SMOTE Tomek Link	0.929	0.928	0.928	0.928	27.6

Evaluation of Machine Learning Model

Based on Appendix A, the study synthesizes the average results of the metrics presented in Appendix B, visualized in Figure 8. The results indicate that in most cases, XGBoost delivers stable and superior performance compared to other models, achieving an accuracy of 0.9723, with other metrics such as Precision, Recall, and F1-Score being 0.9596, 0.9605, and 0.9599, respectively. Specifically, when combined with SMOTE – ENN, XGBoost outperforms with an accuracy of 0.982, and all metrics, including Precision, Recall, and F1-score are >0.98. Following XGBoost, LightGBM shows stable results with all metrics >0.97, demonstrating its capability if training time is a priority.

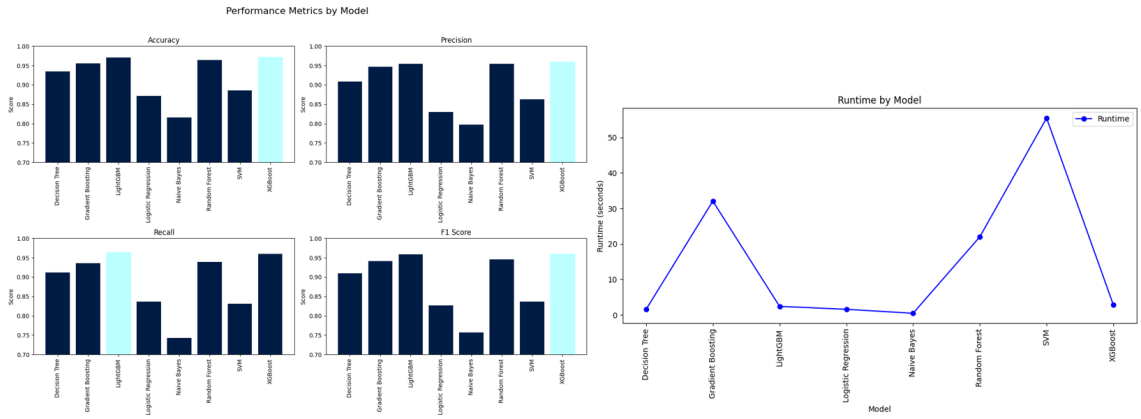

Figure 8. Average metrics:
accuracy, precision, recall, F1-Score, runtime (machine learning models)

On the other hand, Naïve Bayes yields lower results with evaluation metrics all below 0.8, where Accuracy reaches 0.81, while Precision, Recall, and F1-Score are 0.798, 0.743, and 0.757, respectively, making it unsuitable for practical deployment.

Evaluation of Tree-Based & Non-Tree-Based Models

When no data imbalance handling methods (such as oversampling, under-sampling, hybrid methods, or class weight adjustment) are applied, tree-based models (such as Random Forest, XGBoost, LightGBM) typically outperform non-tree-based models (such as Logistic Regression, Naive Bayes, SVM). Tree-based models have a better ability to self-adjust when handling the dominant class due to their mechanism of splitting data and selecting important features. However, even though resampling is not required, these models can still be affected by data imbalance if parameters such as class weight are not adjusted. On the other hand, non-tree-based models, despite their fast computation and ease of implementation, often struggle to accurately classify the minority class because they tend to prioritize the dominant class and lack the capability to detect complex patterns in the minority class. Therefore, in cases where data imbalance is not addressed, tree-based models remain a more effective choice due to their superior self-adjustment and ability to handle imbalanced data.

Specifically, in this study, the results are illustrated in Figure 9. Tree-based algorithms, including Decision Tree, Random Forest, XGBoost, Gradient Boosting, and LightGBM (represented by blue lines), achieve higher accuracy and sensitivity compared to the remaining models: Naïve Bayes, SVM, and Logistic Regression (represented by red lines).

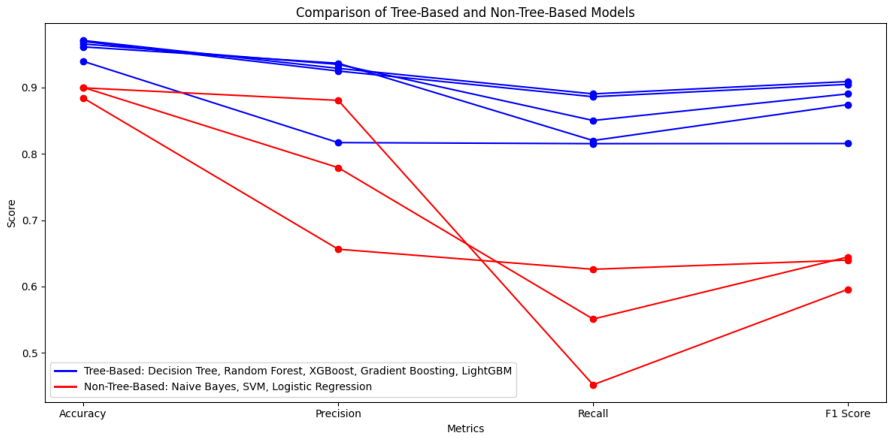

Figure 9. Comparison of tree-based and non-tree-based models

Conclusion and Recommendation

This paper summarized the theoretical basis of customer churn while suggesting the use of supervised machine-learning techniques for predicting customer attrition. Table 7 presents the machine learning models for predicting customer churn, tailored to different imbalance data handling techniques. Each method is applied to improve the class distribution in the data, enhancing the performance of machine learning models in predicting customers likely to churn.

Table 7. Performance Insights: Impact of Imbalance Handling on Models

Imbalance technique	Detailed method	Model(s)	Remarks
No process	No process	LightGBM, XGBoost	Excellent prediction without imbalance handling
UnderSampling	Under-sampling	Random, LightGBM, XGBoost	
	NearMiss	LightGBM, XGBoost	

Imbalance technique	Detailed method	Model(s)	Remarks
OverSampling	Random oversampling	Random Forest, XGBoost	
	Smote	XGBoost, LightGBM	
	Borderline Smote	XGBoost, LightGBM	
Hybrid methods	ADASYN	XGBoost, LightGBM	
Class weight	Smote Tomek Link	XGBoost, LightGBM, Random Forest	XGBoost, LightGBM, Random Forest
	SmoteENN	XGBoost, LightGBM, Logistic, SVM	
	Class weight	XGBoost, LightGBM	SVM decreases precision

The contributions of this paper are described as follows:

1. The research demonstrates that SMOTEENN is the most effective method for handling imbalanced data. This technique combines SMOTE, which generates synthetic samples for the minority class, and ENN, which removes misclassified samples, improving both class balance and data quality. This leads to better model generalization. The average evaluation results of SMOTE-ENN through K-fold cross-validation show that the Accuracy, Precision, Recall, and F1-Score are 0.941, 0.948, 0.955, and 0.952, respectively.
2. LightGBM and XGBoost perform well with imbalanced data due to their gradient-boosting approach. LightGBM's histogram-based method and leaf-wise growth enhance performance on imbalanced datasets, while XGBoost's regularization prevents overfitting, making both algorithms robust. When combined with SMOTEENN, these models achieve significant improvements in Precision, Recall, and F1 score. Specifically, XGBoost achieves an accuracy of up to 0.982 and a sensitivity of 0.990.
3. In cases where data imbalance is not addressed, tree-based models, ensemble bagging, and boosting methods should be preferred due to their ability to effectively split and learn features, delivering superior performance compared to algorithms such as Naïve Bayes (probability-based), Logistic Regression (regression-based), and SVM.

This study serves as a benchmark for implementing customer churn prediction in industries such as banking, e-commerce, telecommunications, and insurance. However, there are several limitations in this research that could be addressed or further explored in future studies.

First, this study exclusively uses data from the Kaggle website for analysis. To enhance practical relevance, future studies could apply this methodology to data collected from customer transaction histories at banks. Second, this study is limited to determining methods for handling imbalanced data and identifying suitable machine-learning models for each technique. Future projects could utilize real-world data to identify which features truly influence the model's prediction outcomes. Third, in future studies, hyperparameter optimization methods such as Grid Search, Random Search, Bayesian Optimization, and feature selection techniques like Recursive Feature Elimination (RFE), Regularization (L1-Lasso Regression and L2-Ridge Regression) can be applied to enhance prediction performance and improve evaluation metrics.

Finally, this research focuses solely on classifying customer churn and examining how data imbalance handling techniques affect the results. Advanced research methods will aim to develop an application to help financial companies, such as banks, retain their customers.

ACKNOWLEDGMENT

This research is funded by the University of Economics Ho Chi Minh City, Vietnam (UEH).

REFERENCES

Aggarwal, P., & Vijayakumar, V. (2024, May). Customer churn prediction in the Telecom sector. *Proceedings of the 3rd International Conference on Artificial Intelligence for Internet of Things, Vellore, India*, 1-6. https://doi.org/10.1109/AIIoT58432.2024.10574660

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*, Article 28. https://doi.org/10.1186/s40537-019-0191-6

Al-Najjar, D., Al-Rousan, N., & Al-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(4), 1529-1542. https://doi.org/10.3390/jtaer17040077

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*, 290-301. https://doi.org/10.1016/j.jbusres.2018.03.003

Amin, A., Al-Obeidat, F., Shah, B., Tae, M. A., Khan, C., Durrani, H. U. R., & Anwar, S. (2020). Just-in-time customer churn prediction in the telecommunication sector. *The Journal of Supercomputing*, *76*, 3924-3948. https://doi.org/10.1007/s11227-017-2149-9

Bharathi, S. V., Pramod, D., & Raman, R. (2022). An ensemble model for predicting retail banking churn in the youth segment of customers. *Data*, *7*(5), 61. https://doi.org/10.3390/data7050061

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

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, *34*, 11751-11768. https://doi.org/10.1007/s00521-022-07067-x

Domingos, E., Ojeme, B., & Daramola, O. (2021). Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. *Computation*, *9*(3), 34. https://doi.org/10.3390/computation9030034

Dong, N. P., Long, H. V., & Khastan, A. (2020). Optimal control of a fractional order model for granular SEIR epidemic with uncertainty. *Communications in Nonlinear Science and Numerical Simulation, 88*, 105312. https://doi.org/10.1016/j.cnsns.2020.105312

Gerritsen, D. F., & Bikker, J. A. (2020). Bank switching and interest rates: Examining annual transfers between savings accounts. *Journal of Financial Services Research*, *57*, 29-49. https://doi.org/10.1007/s10693-018-0305-x

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning*, *113*, 4845-4901. https://doi.org/10.1007/s10994-022-06268-8

Goyal, S. (2021). *Credit card customers*. https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D. S. Huang, X. P. Zhang, & G. B. Huang (Eds.), *Advances in intelligent computing*. Springer. https://doi.org/10.1007/11538059_91

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the IEEE International Joint Conference on Neural Networks, Hong Kong*, 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Jain, S., Shukla, S., & Wadhvani, R. (2018). Dynamic selection of normalization techniques using data complexity measures. *Expert Systems with Applications*, *106*, 252-262. https://doi.org/https://doi.org/10.1016/j.eswa.2018.04.008

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*, Article 10. https://doi.org/10.1186/s40854-016-0029-6

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing*, *104*, 271-294. https://doi.org/10.1007/s00607-021-00908-y

Long, H. V., Son, L. H., Khari, M., Arora, K., Chopra, S., Kumar, R., Le, T., & Baik, S. W. (2019). A new approach for construction of geodemographic segmentation model and prediction analysis. *Computational Intelligence and Neuroscience*, *2019*(1), Article 9252837. https://doi.org/10.1155/2019/9252837

Muntasir Nishat, M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., Ar-Rafi, A. M., Nasrullah, S. M., Reza, M. T., & Khan, M. R. H. (2022). A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Scientific Programming*, *2022*(1), Article 3649406. https://doi.org/10.1155/2022/3649406

Nitesh, V. C. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *PloS ONE*, *18*(12), e0289724. https://doi.org/10.1371/journal.pone.0289724

Pustokhina, I. V., Pustokhin, D. A., Nguyen, P. T., Elhoseny, M., & Shankar, K. (2023). Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. *Complex & Intelligent Systems*, *9*, 3473-3485. https://doi.org/10.1007/s40747-021-00353-6

Rahman, M., & Kumar, V. (2020). Machine learning based customer churn prediction in banking. 2020 4th international conference on electronics, communication and aerospace technology (ICECA). https://doi.org/10.1109/ICECA49313.2020.9297529

Rajamohamed, R., & Manokaran, J. (2018). Improved credit card churn prediction based on rough clustering and supervised learning techniques. *Cluster Computing*, *21*, 65-77. https://doi.org/10.1007/s10586-017-0933-1

Reichheld, F. F., & Sasser, W. E., Jr. (1990). Zero defections: Quality comes to services. *Harvard Business Review*. https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Swana, E. F., Doorsamy, W., & Bokoro, P. (2022). Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. *Sensors*, *22*(9), 3246. https://doi.org/10.3390/s22093246

Vijaya, J., & Sivasankar, E. (2019). An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing. *Cluster Computing*, *22*(Suppl 5), 10757-10768. https://doi.org/10.1007/s10586-017-1172-1

Vo, N. N. Y., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems*, *212*, 106586. https://doi.org/10.1016/j.knosys.2020.106586

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, *14*, 100342. https://doi.org/10.1016/j.rico.2023.100342

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

Y, N. N., Ly, T. V., & Son, D. V. T. (2022). Churn prediction in telecommunication industry using kernel Support Vector Machines. *PL0S ONE*, *17*(5), e0267935. https://doi.org/10.1371/journal.pone.0267935

Yuan, Y., & Tao, R. (2023). Prepayment and credit utilization in peer-to-peer lending. *Managerial Finance*, *49*(12), 1849-1864. https://doi.org/10.1108/MF-02-2023-0136

Zuech, R., Hancock, J., & Khoshgoftaar, T. M. (2021). Detecting web attacks using random undersampling and ensemble learners. *Journal of Big Data*, *8*, Article 75. https://doi.org/10.1186/s40537-021-00460-8

APPENDICES

APPENDIX A: DATA DESCRIPTION

Variable group	Variable name	Data type	Description
Demographics	Customer_Age	Int64	Age of the customer (in years)
	Dependent_Count	Int64	Number of dependents in the customer's family
	Education_Level	Object	Customer's education level
	Marital_Status	Object	Customer's marital status
	Income_Category	Object	Customer's annual income
Relationship with the bank	Months_on_book	Int64	Length of relationship with the bank
	Total_Relationship_Count	Int64	Total number of products the customer holds
	Contacts_Count_12_mon	Int64	Number of contacts between the customer and the bank in the past 12 months
Credit card transaction history	Card_Category	Object	Type of credit card the customer uses (Blue, Silver, Gold, Platinum)
	Credit_Limit	Float64	Credit card limit
	Total_Revolving_Bal	Int64	Total revolving balance
	Avg_Open_To_Buy	Int64	Average available balance on the credit card over the past 12 months
	Total_Trans_Amt	Int64	Total credit card spending (past 12 months)
	Avg_Utilization_Ratio	Float64	Average credit card utilization ratio (Amount used/Credit limit)
	Total_Amt_Chng_Q4_Q1	Float64	Change in total credit card spending (Q4 vs Q1)
	Total_Trans_Ct	Int64	Total number of transactions (past 12 months)
	Total_Ct_Chng_Q4_Q1	Float64	Change in total number of transactions (Q4 vs Q1)
	Months_Inactive_12_mon	Int64	Number of months the card was inactive in the past 12 months

APPENDIX B: AVERAGE RESULTS OF IMBALANCED DATA HANDLING METHODS

Imbalance method	Accuracy	Precision	Recall	F1 Score	Runtime
Classweight	0.9242	0.7801	0.8185	0.7897	6.4500
Hybrid Method	0.9349	0.9379	0.9428	0.9403	18.4250
No process	0.9361	0.8571	0.7363	0.7842	6.9500
OverSampling	0.9272	0.9236	0.9322	0.9277	23.2469
UnderSampling	0.8876	0.9050	0.8673	0.8776	2.3813

AUTHORS

Tam-Thanh Luong is a sophomore student majoring in e-commerce in Business Information Technology at the University of Economics, Ho Chi Minh City, Vietnam. His current research interests include data analytics and machine learning. He can be contacted via email at lthanhtam915@gmail.com

Vi-Gia Luong is a sophomore student majoring in e-commerce in Business Information Technology at the University of Economics, Ho Chi Minh City, Ho Chi Minh City, Vietnam. His current research interests include data analytics and digital marketing. He can be contacted via email at viluong.31231024175@gmail.com

Anh Hoang Tuan Tran is a sophomore student majoring in e-commerce in Business Information Technology at the University of Economics. Ho Chi Minh City, Ho Chi Minh City, Vietnam. His current research interests include business analytics and machine learning. He can be contacted via email at anhtran.31231024605@st.ueh.edu.vn

Tuan Manh Nguyen received a B.S. degree in Information Technology from the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (VNU-HCM), Vietnam, in 2005 and a Master's degree in Management Information Systems from the Ho Chi Minh City University of Technology (VNU-HCM), Vietnam in 2010. He is currently a lecturer in the Faculty of Business Information Technology, College of Technology and Design, University of Economics, Ho Chi Minh City, Vietnam. His research has been published in international journals and conferences, such as the Foundations of Management, 2nd International Conference - Resilience by Technology and Design (RTD 2024), 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter). His primary research areas include: (i) architectural models and digital technology solutions for organizations, (ii) software technology, digital transformation, and information security, and (iii) AI, machine learning, and big data. He can be contacted via email at tuannm@ueh.edu.vn