BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **25**, No 3
Sofia • 2025 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2025-0019

Comparison of Methods for Handling Imbalanced Data in Customer Churn Prediction with Feature Selection Using SHAP and mRMR Frameworks

Luong Thanh Tam, Luong Gia Vi, Nguyen Manh Tuan

Faculty of Business Information and Technology, University of Economic Ho Chi Minh City, Vietnam
E-mails: tamluong.31231024238@st.ueh.edu.vn viluong.31231024175@st.ueh.edu.vn tuannm@ueh.edu.vn

***Abstract*:** *This study compared methods for handling imbalanced data in predicting customer churn in banking and e-commerce, using datasets with features selected via SHAP and MRMR. Two approaches were evaluated: data-level* (*Oversampling, Undersampling, and Hybrid resampling*) *and algorithm-level. Oversampling excelled on small to medium datasets, while Undersampling improved Recall but reduced Precision, lowering overall performance. Ensemble models outperformed single models, with tree-based Decision Trees showing better learning on imbalanced data among single models. The study recommends ensemble models for churn prediction, proposing the SHAP framework to enhance their interpretability through global and local explanations. Two models, ROS-CatBoost and CW-XGBoost, achieved exceptional results, with metrics like Accuracy, Precision, Recall, F1-score, ROC AUC, and PR AUC all above 0.9, indicating strong predictive accuracy for both churn and retention. These findings highlight the effectiveness of ensemble models and interpretability tools in addressing imbalanced data challenges.*

***Keywords*:** *Machine learning, Customer churn, Imbalanced data, Resampling, Explainable AI.*

1. Introduction

In the context of intense competition, customer churn causes significant losses to a company's profitability, as the cost of acquiring new customers is considerably higher than retaining existing ones [1]. To address this issue, businesses often leverage the vast amount of customer data they collect, combined with Machine Learning models, to transform data into actionable predictions that form the basis for effective customer retention strategies [2].

However, the application of Machine Learning to the churn prediction problem in real-world scenarios currently faces the challenge of data imbalance. In most business settings, the number of customers who actually churn (the minority class)

is significantly smaller than the number of those who remain (the majority class). When the data becomes imbalanced to a certain degree, models tend to become biased, losing sensitivity or becoming entirely "blind" to accurately identifying instances of the minority class – namely, customers who are about to churn, the very target of interest. This leads to a paradox: the model may achieve a very high overall accuracy, yet fail in its primary goal of detecting churn. For example, with an actual churn rate of 1%, a model that simply predicts "all customers will stay" would achieve 99% accuracy, but the recall or true positive rate for churn would be 0. Moreover, misclassifying rare events such as churn can result in substantial losses, rendering the model practically useless for business purposes [3].

The contributions of our research include:

- Enhancing the predictive performance for customer churn – the minority class in imbalanced datasets – by selecting input features using SHAP and mRMR.
- Evaluating the improvement in model evaluation metrics through two levels: Resampling (Data-level) and Cost-sensitive learning – Class Weight (Algorithm-level) on small to medium-sized datasets.
- Identifying that ensemble models and tree-based models, specifically Decision Trees, exhibit strong learning capabilities on imbalanced datasets with high "inherent robustness."
- Determining that two models, ROS-CatBoost and CW-XGBoost, achieve high performance, with evaluation metrics including Accuracy, Precision, Recall, F1-score, ROC AUC, and PR AUC all exceeding 0.9, indicating highly accurate predictions for both churn and retention classes.
- To facilitate the practical deployment of ensemble models, we propose a method to enhance model interpretability using Explainable AI – specifically SHAPley values – at both global and local levels.

2. Literature review

2.1. Customer churn prediction and imbalance data

Customer churn is a fundamental business challenge that has a significant impact in highly competitive, subscription-based industries where maintaining long-term customer value is a top priority, such as telecommunications [4-9], banking [10-13], and e-commerce [14]. To address the problem of predicting customer churn, numerous studies have applied and extensively evaluated a wide range of machine learning algorithm families. Commonly used methods include:

- Naive Bayes: A probabilistic algorithm based on Bayes' theorem with the assumption of feature independence, offering fast and reliable results [7].
- Logistic Regression: A classical statistical method often chosen as a baseline model due to its simplicity and effectiveness [14].
- Support Vector Machines (SVM): A technique that finds the optimal hyperplane to separate data classes, particularly effective in high-dimensional spaces [1].

- Decision Trees: Algorithms such as C4.5 and CART are notable for their clear and interpretable model structures [15].
- Ensemble Learning: Combines multiple weak learners to form a stronger model, including: Bagging: Exemplified by Random Forest, which enhances accuracy and reduces variance [12]; Boosting: Algorithms such as AdaBoost, Gradient Boosting, and XGBoost, which aim to optimize predictive performance [7, 13].

However, [16] it has been noted that in real-world datasets, the proportion of churned customers is often significantly lower than that of retained customers, leading to class imbalance. The accuracy-optimization mechanism of most classification algorithms causes models to focus on learning the characteristics of the majority class while neglecting or misclassifying minority class instances – namely, the customers at risk of churn. As a result, despite achieving high overall accuracy, such models often fail to accurately detect the churn group, which is the critical target that needs to be identified.

2.2. Related work

Table 1 presents a selection of previous studies in the field of customer churn prediction, including information on the models used, data imbalance handling techniques, and evaluation results.

Table 1. Summary literature review

Ref	Year	Best classifier	Imbalance handling techniques	F1-score	Recall/ Sensitivity
[10]	2019	Gradient boost	None	-	-
[11]	2020	Random Forest	Random oversampling, Random undersampling	-	-
[12]	2022	Random Forests	None	0.8225	-
[17]	2022	ExtraTrees	None	0.9286	0.9286
[15]	2022	C5 Tree	None	0.9050	0.8910
[13]	2023	XGBoost	ADASYN, SMOTE, SMOTE-ENN	0.9004	0.9262
[18]	2024	XGBoost	ADASYN, NEARMISS	-	0.8967
[19]	2022	SVM	SMOTE	-	0.9721
[4]	2019	XGBoost	Random oversampling, undersampling	-	-
[5]	2020	Random Forest	None	-	-
[6]	2020	Logit Boost	None	0.919	0.9820
[7]	2021	AdaBoost	Resampling (No detailed information)	0.8060	0.8164
[8]	2021	OWELM	SMOTE	0.9080	-
[20]	2021	XGBoost	None	0.9554	0.9548
[9]	2022	S-RNN	None	-	0.9827
[21]	2025	TriBoost	SMOTE	0.9025	0.8782

In [3] indicate that, despite advancements and evaluations of various methods and techniques for addressing data imbalance, current research still lacks systematic

assessments and comprehensive comparisons among imbalance handling techniques. Another study [16] compared six oversampling techniques but fell short of broader, more diverse comparative analyses across different groups of handling strategies (e.g., Oversampling vs Undersampling vs Hybrid vs Class Weight) across multiple model types and churn datasets. The absence of investigations into the "inherent robustness" and "differential sensitivity" of algorithm families remains a limitation: While several studies, including relatively extensive surveys like [3] have compared the effectiveness of various techniques, a systematic evaluation of the inherent robustness and differential sensitivity of prominent algorithm families (e.g., Boosting, Bagging, Tree-based, Naive Bayes, SVC) when confronted with imbalanced data and under the influence of different handling strategies still requires further exploration.

Additionally, several studies [11, 22, 23] in the field of customer churn prediction have implemented resampling techniques; however, applying these techniques to the entire dataset rather than solely to the training set has led to data leakage in the test set, resulting in outcomes lacking validity.

Our study is specifically designed to directly address these limitations through a series of systematic experiments, encompassing a wide range of data processing techniques, algorithm families, and diverse datasets. Notably, our approach integrates a modern feature selection phase utilizing Explainable AI techniques—specifically SHAP combined with mRMR – to enhance the overall model performance in handling imbalanced data. This integration also facilitates interpretability by uncovering the inner mechanisms that drive the model's predictions. The ultimate goal is to provide deep insights and reliable recommendations for the customer churn prediction problem in the context of imbalanced datasets.

3. Proposed model

In this study, the research team conducted experiments to evaluate the effectiveness of data imbalance handling techniques, including resampling and class weighting, with the technical route outlined in Fig. 1. Following data preparation and preprocessing steps – such as removing outliers, handling missing values, encoding, and feature scaling – the study employed XGBoost, a model previously identified as high-performing for customer churn prediction in numerous studies [11, 22, 23] the base model. Feature selection was performed using SHAP (Shapley Additive Explanations) and mRMR (Maximum Relevance Minimum Redundancy) to determine the training features for the model. SHAP is an interpretability framework designed for black-box models within the domain of XAI (Explainable AI). For each predicted sample, SHAP [24] computes the Shapley value for every feature, which is regarded as a measure of that feature's contribution to the overall prediction outcome. The mRMR [25] method is a feature selection technique that reduces the feature space by maximizing relevance to the class label while minimizing redundancy among features. Consequently, mRMR facilitates the selection of an optimal feature subset, well-suited for predictive tasks.

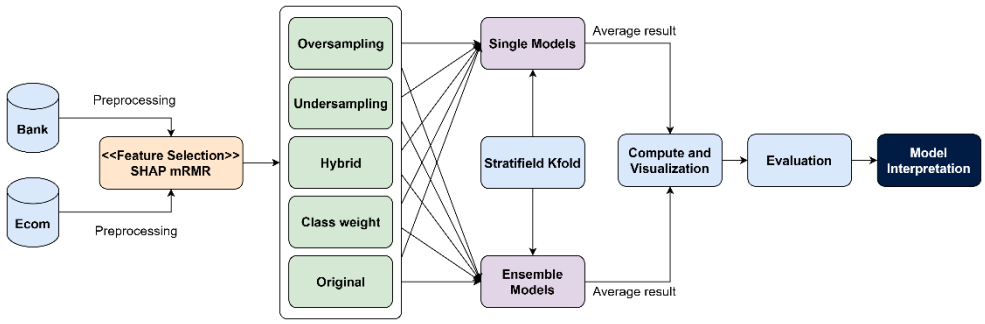


Fig. 1. Technical route

Step 1. Perform data preprocessing on both Dataset 1 and Dataset 2.

Step 2. Conduct feature selection using a combined approach of SHAP-mRMR. The implementation structure of SHAP-mRMR is described in Fig. 2.

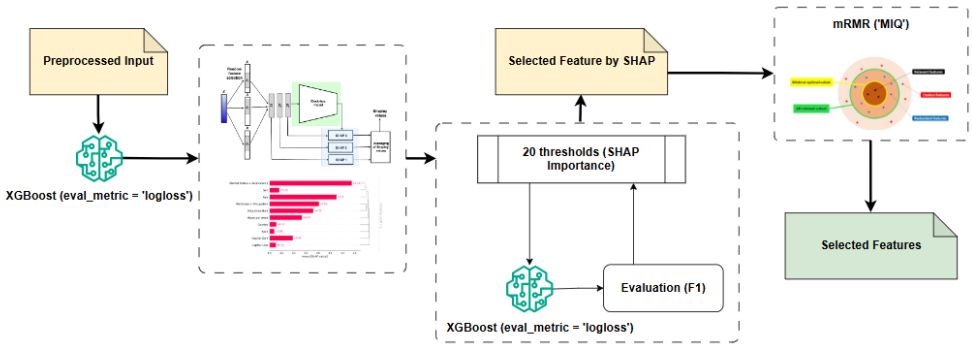


Fig. 2. SHAP-mRMR Selection features framework

In our methodology, XGBoost is implemented as a baseline model to identify highly influential features through the SHAP-mRMR technique. With 20 levels of SHAP importance thresholds used, the study trains the base model and evaluates it through the F1-score. The best feature combination obtained from the SHAP evaluation step will be processed by mRMR with the MIQ metric, prioritizing high relevance to the target variable while minimizing redundancy with other features. The reduced dataset will serve as input for the next step.

Step 3. Develop both single and ensemble models (Fig. 3), incorporating resampling techniques (Fig. 4), and evaluate their performance using Stratified k-Fold cross-validation (k=5). The model parameters and techniques are utilized with default settings from the scikit-learn library.

Step 4. Compute and assess model performance using various evaluation metrics, including Accuracy, Precision, Recall, F1-score, ROC AUC, and PR AUC.

Step 5. Identify and return the best-performing classifier for each dataset, with appropriate handling of class imbalance.

Step 6. Propose a modern model explanation method based on Explainable AI that is suitable for the proposed model.

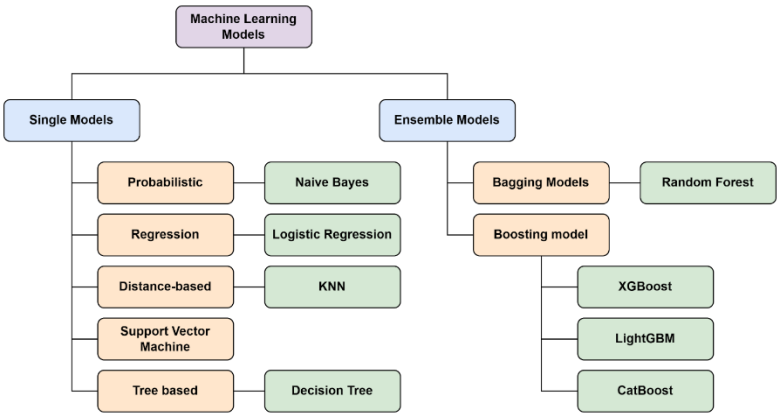


Fig. 3. Machine learning model used in this study

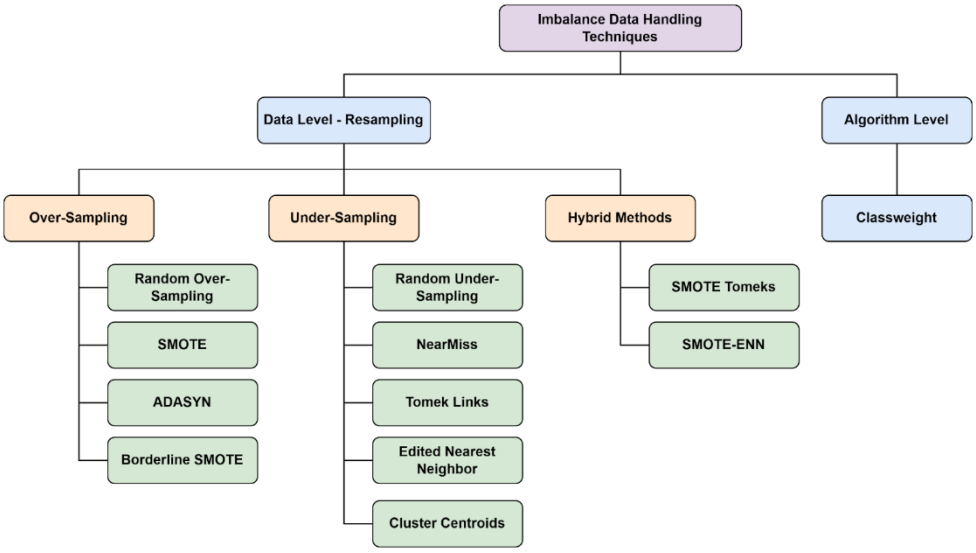


Fig. 4. Handling imbalance data techniques

Data-level approaches aim to adjust the distribution of the training dataset through resampling techniques, creating a more balanced dataset before input into machine learning classification models. Some notable examples include:

Oversampling [26]: These techniques increase the number of samples.

- Random OverSampling (ROS) [26] is a straightforward technique that balances the dataset by randomly duplicating instances from the minority class. While easy to implement, this method can introduce the risk of overfitting due to repeated information.
- SMOTE (Synthetic Minority Over-sampling TEchnique) [26]: generates synthetic minority samples by performing linear interpolation between a given minority instance and its k-Nearest Neighbors. This helps introduce variability and reduce redundancy.

- Variants of SMOTE, such as ADASYN (ADAptive SYNthetic Sampling) [27], adaptively focus on difficult-to-learn minority instances by generating more synthetic samples in these areas, improving the model's ability to generalize across challenging patterns, while Borderline-SMOTE [28] concentrates sample generation near the decision boundary between classes, where misclassification is more likely, aiming to strengthen the classifier's discrimination power in ambiguous regions.

Undersampling [29, 30]: These techniques reduce the number of samples in the dataset.

- Random UnderSampling (RUS) [29]: randomly removes instances from the majority class to balance the dataset. While computationally efficient and simple to implement, this technique may lead to the loss of potentially important data and degrade model performance, especially when the majority class contains informative samples.
- NearMiss [30] is a family of undersampling techniques with three main variants – NearMiss-1, NearMiss-2, and NearMiss-3 – that select majority class samples based on different distance criteria to minority instances. These methods aim to retain the majority samples that are most informative for distinguishing between classes, typically by focusing on those nearest to minority samples, either on average or individually.
- Tomek Links [31]: identifies and removes borderline majority class samples that form a pair (Tomek Link) with a minority class sample – i.e., they are each other's nearest neighbors but belong to different classes. Removing the majority class member of the Tomek Link helps clean class boundaries and reduce class overlap.
- Edited Nearest Neighbors (ENN) [32]: removes samples (typically from the majority class) that disagree with the majority class among their k nearest neighbors (commonly $k = 3$). This technique reduces noise and refines decision boundaries, often used as a post-processing step after resampling.

Hybrid Methods [29]: combine both oversampling and undersampling techniques to leverage the advantages of each. For example, SMOTE-Tomek Links and SMOTE-ENN first apply SMOTE to synthetically increase the number of minority class samples, followed by Tomek Links or Edited Nearest Neighbors (ENN) to remove overlapping or noisy instances from the majority class. These hybrid approaches aim to enhance class balance while also refining decision boundaries by reducing noise and redundancy.

Alternatively, if altering the data structure is undesirable, machine learning algorithms can be directly modified to become more sensitive to the minority class. A prominent example is Class Weight [33], an aspect of cost-sensitive learning, which improves the model by modifying the loss function. Instead of using the standard 0-1 loss function (which only considers correct or incorrect predictions), each sample i is assigned a weight w_{y_i} based on its class y_i, and the model is optimized accordingly:

$$\mathcal{L} = \sum_{i=1}^{\mathbb{N}} w_{y_i} l(f(x_i), y_i). \quad (1)$$

With N representing the total number of samples, $l(f(x_i), y_i)$ denoting the standard loss function (e.g., cross-entropy), and w_{y_i} indicating the weight assigned to class y_i, the approach assigns greater weight to errors occurring in the minority class. Consequently, the training process compels the classifier to pay increased attention to instances from this class. This adjustment shifts the decision boundary further from the minority class instances, thereby enhancing classification performance for these classes.

To evaluate the performance of the model and each method, this study proposes using Precision, Recall, and F1-score as the primary evaluation metrics. In particular, the study aims to improve the F1-score – the harmonic mean of Precision and Recall – with a focus on enhancing Recall, which measures the proportion of correctly predicted churned customers among all actual churned customers (Recall = TP / (TP + FN)). The goal is to maximize the increase in Recall while minimizing the corresponding decrease in Precision.

4. Data preparation and its preprocessing

The study utilizes a credit card customer dataset (Dataset 1) from Kaggle, provided by Sakshi Goyal, used in [13, 23, 34]. After preprocessing, the dataset comprises 10,127 observations and 20 variables. The dependent variable, "Attrition_Flag," is binary, where a value of 1 indicates that a customer has churned, and 0 indicates continued credit card usage. A detailed description of this dataset is provided in Table 2. Similarly to Dataset 1, the study employs Dataset 2 in the e-commerce domain, used in [19, 35], including 5630 customers and 19 features, with its features described in Table 3.

Table 2. Dataset 1 description

Variable name	Description
Customer_Age	Age of the customer (in years)
Gender	Gender of the customer (Male or Female)
Dependent_Count	Number of dependents in the customer's family
Education_Level	Customer's education level
Marital_Status	Marital status of the customer
Income_Category	Annual income range of the customer
Months_on_book	Duration of customer's relationship with the bank (in months)
Total_Relationship_Count	Total number of products the customer holds
Contacts_Count_12_mon	Number of interactions between the customer and the bank in the past 12 months
Card_Category	Type of credit card held by the customer (Blue, Silver, Gold, Platinum)
Credit_Limit	Credit card limit
Total_Revolving_Bal	Total revolving balance on the credit card
Avg_Open_To_Buy	Average available balance on the credit card over the past 12 months
Total_Trans_Amt	Total credit card spending (past 12 months)
Avg_Utilization_Ratio	Average credit card utilization ratio (Amount used / Credit limit)
Total_Amt_Chng_Q4_Q1	Change in total credit card spending (Q4 vs Q1)
Total_Trans_Ct	Total number of transactions made in the last 12 months
Total_Ct_Chng_Q4_Q1	Change in the number of transactions between Q4 and Q1
Months_Inactive_12_mon	Number of months the card was inactive in the past 12 months

Table 3. Dataset 2 description

Variable name	Description
Churn	Whether the customer has churned (1) or not (0)
Tenure	Number of months the customer has been active
CityTier	Development level of the customer's city (Tier 1-3)
WarehouseToHome	Distance from the warehouse to the customer's home (km)
HourSpendOnApp	Average weekly time spent on the mobile app
NumberOfDeviceRegistered	Total number of devices registered by the customer
SatisfactionScore	Customer satisfaction rating
NumberOfAddress	Number of saved delivery addresses
Complain	Whether the customer has submitted complaints
OrderAmountHikeFromLastYear	Increase in order value compared to the previous year
CouponUsed	Number of discount coupons redeemed
OrderCount	Total number of orders placed
DaySinceLastOrder	Days since the customer's most recent order
CashbackAmount	Total cashback earned
PreferredLoginDevice	The device is most often used to log in
PrefferedPaymentMode	Most frequently used payment method
Gender	Gender of the customer
PrefferedOrderCat	Most commonly purchased product category
MarialStatus	Gender of the customer (Male or Female)

In this paper, categorical variables are transformed into a numerical format using label encoding and one-hot encoding to ensure effective utilization by machine learning models. Following the one-hot encoding process, continuous features in the dataset are normalized using Z-score normalization, which converts the original data into a Gaussian distribution. The specific preprocessing transformations applied to the two datasets are detailed in Appendix A.

For both datasets, statistical analysis presented in Fig. 5 reveals a severe class imbalance, with churn cases occurring at a ratio of 1:5 compared to retained customers in both datasets.

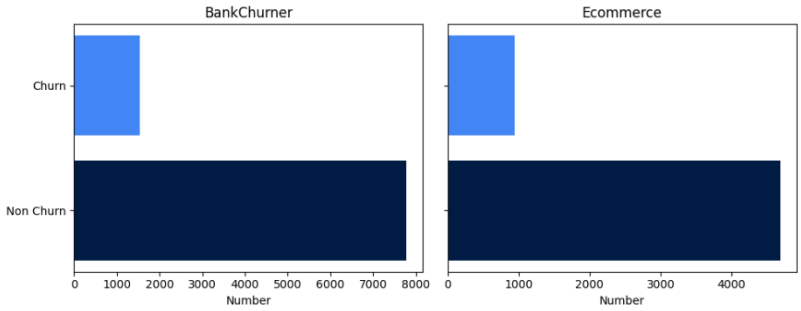
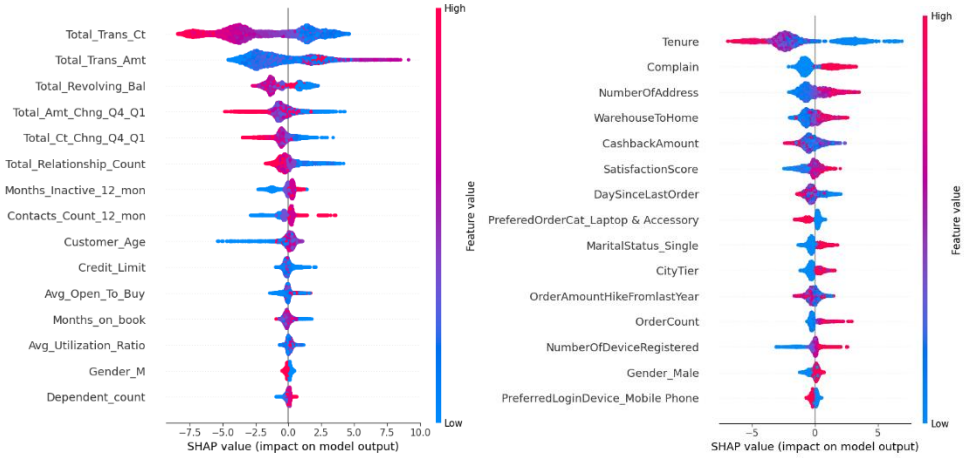


Fig. 5. Number of churn and non-churn customers

Before the training and evaluation phases, the study employed XGBoost in conjunction with SHAP values and the MRMR method to perform feature selection. The results of the model interpretation using SHAP are illustrated in Fig. 6 (left panel: BankChurners dataset; right panel: Ecommerce dataset). By evaluating 20 different threshold values in combination with the F1-score metric, the study identified the most influential features contributing to the model's predictive

performance. These features were subsequently refined using the MRMR method with the "MIQ" criterion, and the final list of selected features is presented in Table 3.



(a) (b)

Fig. 6. Summary SHAP value plots of BankChurners (a) and Ecommerce datasets (b)

Table 4. Selected features after using SHAP-MRMR

Selected feature – BankChurners	Ratio	Selected feature – Ecommerce	Ratio
Contacts_Count_12_mon	12/23	Tenure	12/30
Total_Revolving_Bal		Complain	
Total_Relationship_Count		MaritalStatus_Single	
Avg_Open_To_Buy		PreferedOrderCat_Laptop & Accessory	
Total_Trans_Amt		SatisfactionScore	
Total_Amt_Chng_Q4_Q1		DaySinceLastOrder	
Total_Ct_Chng_Q4_Q1		CityTier	
Months_Inactive_12_mon		CashbackAmount	
Credit_Limit		NumberOfAddress	
Months_on_book		WarehouseToHome	
Customer_Age		OrderAmountHikeFromlastYear	
Total_Trans_Ct		OrderCount	

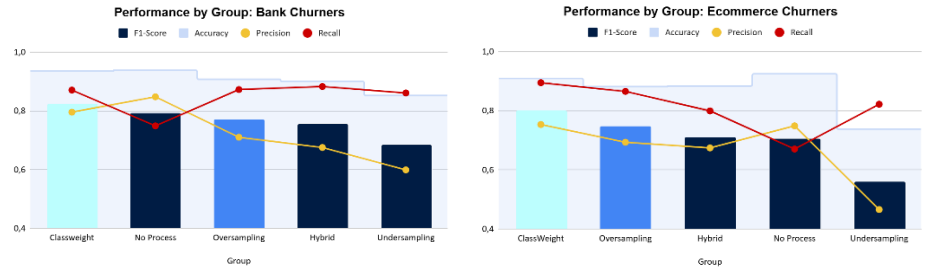
5. Result

5.1. Evaluation of group techniques: Oversampling, Undersampling, Hybrid resampling, and Class weight

After completing the model training and evaluation, the study visualizes and analyzes the experimental outcomes.

Specifically, the research conducts a comprehensive evaluation of the effectiveness of various imbalanced data handling techniques based on empirical data. These include: No Processing, Resampling methods (comprising

Oversampling, Undersampling, and Hybrid Resampling), and Cost-Sensitive Learning through class weight adjustments in the loss function. The results are consolidated across two datasets from the banking and e-commerce domains in the context of customer churn prediction, as illustrated in Fig. 7.



(a) (b)

Fig. 7. Summary performance by group techniques for Bankchurners dataset (a) and Ecommerce dataset (b)

According to Fig. 7, the evaluation metrics results are averaged based on the trained and tested models, and the application of imbalanced data handling techniques leads to negligible changes in Accuracy and F1-score. However, all methods improve Recall over Precision, enhancing the model's ability to identify customer churn cases (the positive class). This is a significant finding, as the minority class is the focal point of the study and is typically difficult to predict in real-world scenarios.

Undersampling significantly increases Recall, enabling the model to better identify churned customers. However, when the minority class contains too few samples, removing data from the majority class may eliminate important information, thereby impairing model learning. Therefore, this technique should only be applied when the dataset is sufficiently large – with enough churn records – to avoid losing essential features. In the subsequent parts of this study, Undersampling will be excluded from model performance comparisons, as the two datasets used are of moderate size with relatively few positive churn observations.

Oversampling demonstrates superior performance in handling class imbalance. However, it significantly increases computational cost for large datasets, as the number of samples grows and prolongs training time. This trade-off must be carefully considered in resource-constrained environments.

Hybrid resampling techniques, which combine Oversampling and Undersampling, result in a noticeable reduction in Precision, thereby diminishing the model's ability to predict the majority class. Although this effect is less severe than that of UnderSampling alone, it remains a concern. Moreover, this approach requires both Oversampling and Undersampling steps, increasing preprocessing time. Nevertheless, the resulting dataset is smaller than that produced by Oversampling alone, reducing training time. Researchers must consider the risk of discarding observations containing crucial features when applying this method.

The Classweight technique is limited to models that support weighted loss functions and may require manual configuration for certain algorithms. However, it performs strongly in supported models. Fig. 7 demonstrates the outstanding performance of Classweight, which may be attributed to the exclusion of K-Nearest Neighbors (KNN) and Naive Bayes models – traditional models with poor generalization capabilities – from the training process. To test this hypothesis, the study also excluded KNN and Naive Bayes from the evaluation of other techniques, with the results visualized in Fig. 8. These findings confirm that ClassWeight outperforms resampling methods even when evaluated on the same set of models, underscoring its effectiveness in handling imbalanced data.

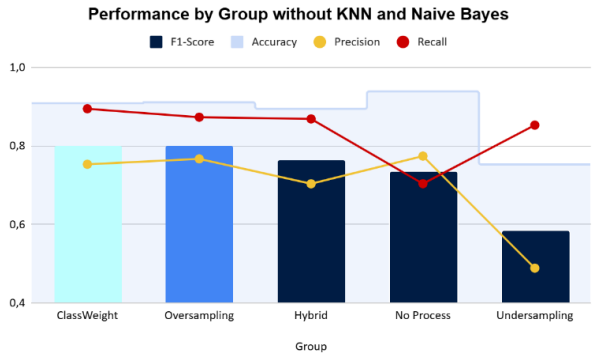


Fig. 8. Performance of two datasets without KNN and Naive Bayes

5.2. Model – imbalance handling techniques analysis

Concurrently, based on experiments conducted on imbalanced data without preprocessing, across a variety of model types, including Bagging, Boosting, and single models (categorized into Tree-based, Probability-based, Distance-based, and Margin-based groups), the aggregated results in Fig. 9 – averaged from two datasets, Ecommerce and Banking Churners – demonstrate that ensemble models exhibit significantly superior performance compared to standalone models. However, there remains a preference for the majority class (Precision > Recall).

For the Ensemble model group, the ability to withstand imbalanced data can be attributed to the following reasons:

- Boosting enhances performance by sequentially training models, where each subsequent model focuses on correcting the errors of its predecessor, particularly emphasizing difficult-to-classify instances [36] – making it highly suitable for imbalanced datasets.
- Bagging improves accuracy and mitigates overfitting by training multiple independent models on random data samples and aggregating their results [37] – thereby increasing model stability.
- Among the single models, Decision Tree and Naïve Bayes stand out by achieving a relatively balanced performance between Precision and Recall. This originates from the nature of the algorithms: Naïve Bayes performs classification based on conditional probability under the assumption of feature independence,

while Decision Tree optimizes data partitioning based on criteria such as Gini index or Entropy to create clear classification boundaries. Both algorithms exhibit minimal bias toward any specific class, which contributes to reducing the gap between Precision and Recall. However, although Naïve Bayes is less affected by imbalanced data, it tends to have a relatively low F1-score due to its limited generalization capability.

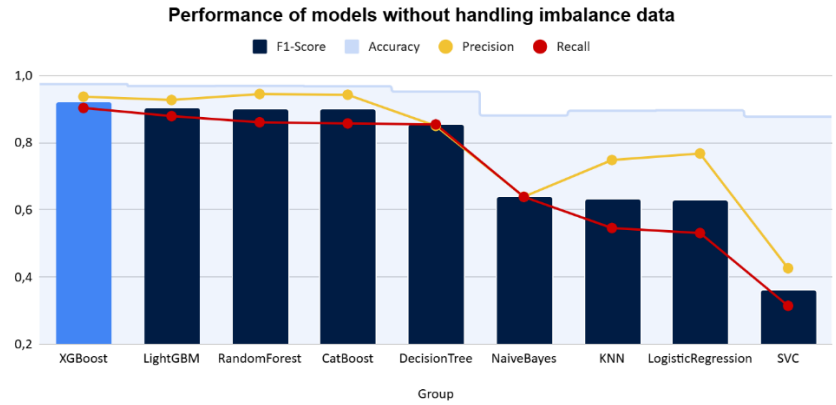


Fig. 9. Performance of models without handling imbalance data

After applying imbalance handling techniques and visualizing the outcomes in Fig. 10, the study observed significant improvements in the ensemble model group, particularly in narrowing the gap between Precision and Recall. For the single model group, except for Decision Tree (which already demonstrated good generalization and low sensitivity to imbalanced data from the outset), the remaining models recorded a noticeable increase in Recall, indicating enhanced capability in detecting churned customers. However, this improvement came at the cost of a relative decrease in Precision, reflecting a trade-off between the two metrics. Fig. 11 illustrates the detailed changes in Precision and Recall within the single model group before and after applying imbalance treatment techniques.

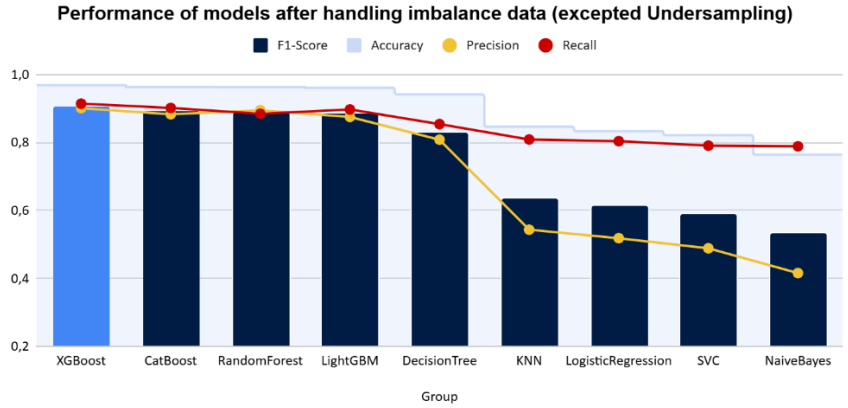
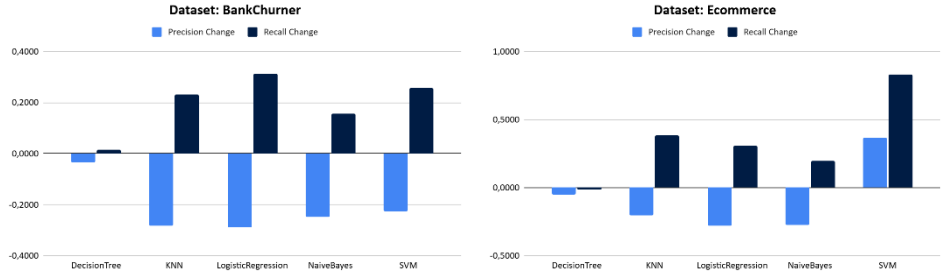


Fig. 10. Performance of models after handling imbalance data (except Undersampling)



(a) (b)

Fig. 11. Precision and Recall Changes in Single Models Post-Treatment for BankChurners dataset (a) and Ecommerce dataset (b)

5.3 Model evaluation and interpretation with XAI-SHAP

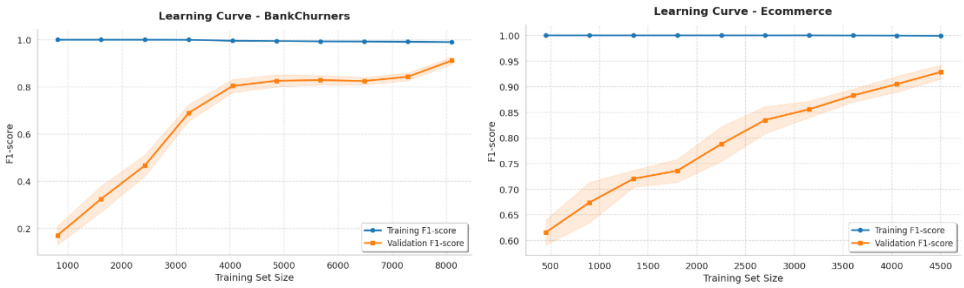
In this section, the study aims to identify optimal methods for handling data imbalance tailored to each model across different datasets, determine the most effective model for each dataset, and provide an interpretive analysis of the results. To select the most suitable imbalanced data handling technique for each model, we evaluate performance based on the average of F1-score and Recall, intending to maximize the balance between Recall and Precision. The goal is to enhance the F1-score by increasing Recall while ensuring Precision either improves or decreases minimally. This approach seeks to minimize the trade-off between Precision and Recall through effective handling of imbalanced data. The results are presented in detail in Table 5.

Table 5. Model performance for churn prediction with Imbalanced data techniques

Dataset	Model	Technique	Acc.	Pre.	Rec.	F1	ROC AUC	PR AUC
BankChurner	CatBoost	Random oversampling	0.97	0.89	0.94	0.92	1.00	0.98
	LightGBM	Random oversampling	0.97	0.89	0.94	0.91	0.99	0.97
	XGBoost	SMOTE	0.97	0.91	0.92	0.91	0.99	0.97
	RandomForest	ADASYN	0.96	0.85	0.90	0.88	0.99	0.94
	DecisionTree	SMOTE ENN	0.92	0.68	0.89	0.77	0.91	0.63
	SVC	Random oversampling	0.90	0.64	0.88	0.74	0.96	0.84
	LogisticRegression	ADASYN	0.83	0.48	0.87	0.62	0.92	0.72
	KNN	ADASYN	0.84	0.49	0.85	0.63	0.89	0.59
	NaiveBayes	Random oversampling	0.79	0.42	0.83	0.56	0.87	0.71
Ecommerce	XGBoost	Classweight	0.98	0.94	0.92	0.93	0.99	0.97
	RandomForest	Random oversampling	0.98	0.95	0.91	0.93	0.99	0.98
	CatBoost	Random oversampling	0.97	0.87	0.93	0.90	0.99	0.96
	LightGBM	Classweight	0.96	0.84	0.93	0.89	0.99	0.95
	DecisionTree	No Process	0.97	0.90	0.91	0.90	0.94	0.83
	KNN	ADASYN	0.84	0.51	0.91	0.65	0.93	0.78
	LogisticRegression	Random oversampling	0.81	0.47	0.84	0.60	0.89	0.71
	SVC	ADASYN	0.72	0.36	0.85	0.50	0.86	0.65
	NaiveBayes	Random oversampling	0.71	0.35	0.82	0.49	0.84	0.63

The results in Table 5 reinforce the conclusions drawn in Section 5.1, indicating that the majority of the most effective detailed techniques for the algorithms across both datasets – medium-sized datasets – belong to the

Oversampling group. Notable exceptions include Decision Tree with No Process and SMOTE-ENN, as well as XGBoost with Classweight. The best-performing models across the two datasets are Random Oversampling-CatBoost (ROS-CatBoost) for the BankChurner dataset and Classweight-XGBoost (CW-XGBoost) for the Ecommerce dataset. To evaluate the research's selection regarding the potential issue of overfitting, we assess these two models using learning curves based on the F1-score metric, as presented in Fig. 12.



(a) (b)

Fig. 12. Learning curve of ROS-CatBoost (a) and CW-XGBoost (b)

The evaluation results show that the performance of our model outperforms previous studies presented in Table 6. For [13, 34] in the banking sector, our study shows a clear improvement in F1 and Recall, while Precision remains superior at 0.89. In the E-commerce sector, our research ensures both Precision and F1-score > 0.92, guaranteeing high model accuracy, whereas [19], in an attempt to increase Recall, sacrificed too much Precision, which dropped to around 0.8, thereby degrading the overall performance of the model.

Table 6. Cross-validation with previous studies

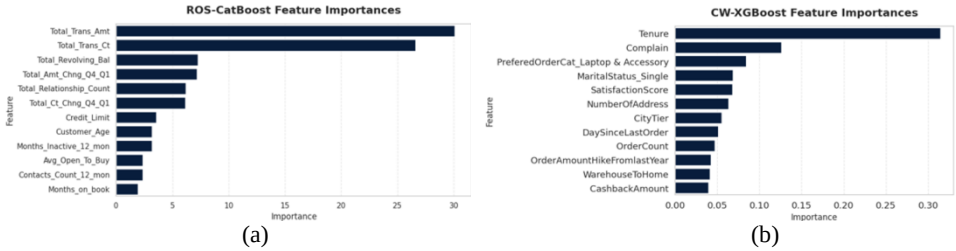
Ref.	Dataset	Model	Acc.	Pre.	Rec.	F1
[34]	BankChurners	C5 Tree	0.94	0.81	0.86	0.84
[13]	BankChurners	SMOTEENN – XGBoost	0.97	0.88	0.93	0.90
Ours	BankChurners	ROS – CatBoost	0.97	0.89	0.94	0.92
[19]	Ecommerce	SMOTE – SVM	0.91	0.82	0.99	-
Ours	Ecommerce	CW – XGBoost	0.98	0.94	0.92	0.93

5.4 Model interpretation with XAI-SHAP for ensemble model

The study indicates that in the problem of customer churn prediction, ensemble models outperform individual models across evaluation metrics such as accuracy, precision, recall, F1-score, ROC AUC, and PR AUC. However, due to the "black-box" nature of these models, feature importance derived from sklearn provides only global-level explanations. In this research, we propose the use of the SHAP framework to interpret models at both global and local levels, elucidating the magnitude and direction (positive or negative) of each variable's impact on prediction outcomes, thereby addressing the transparency limitations of ensemble models.

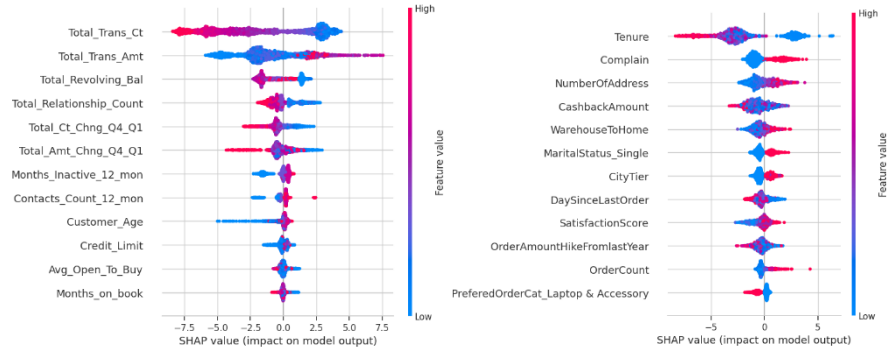
Upon examining the model interpretation results from sklearn with Gini importance (Fig. 13) and the SHAP framework (Fig. 14), the study observes that, despite differences in the ranking of feature importance, both methods yield

consistent results when considering the most influential features. For instance, in the ROS-CatBoost model, both approaches identify the top five features as Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count. Similarly, for the CW-XGBoost model, both methods concur that Tenure and Complain are the two most impactful features. Furthermore, SHAP proves highly effective in elucidating the relationships between features and predictions. For example, Total_Trans_Ct exhibits a negative correlation with customer churn, indicating that higher transaction counts are associated with lower churn rates. Conversely, Complain shows a positive correlation, suggesting that an increase in complaints is linked to a higher likelihood of churn.



(a) (b)

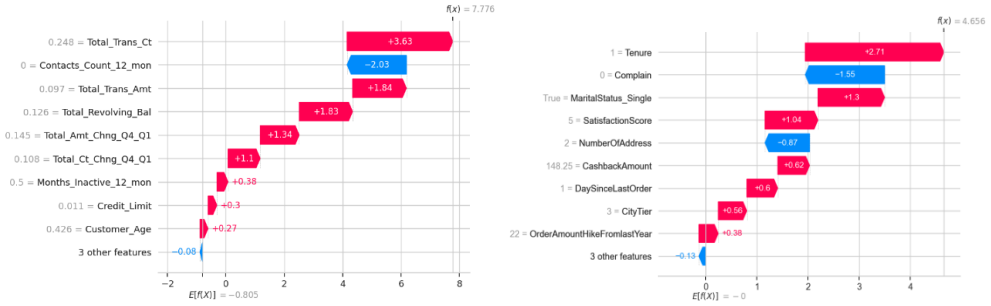
Fig. 13. Feature importance of ROS-CatBoost (a) and CW-XGBoost (b) by sklearn



(a) (b)

Fig. 14. Summary chart of SHAP feature analysis by ROS-CatBoost (a) and CW-XGBoost (b)

Additionally, SHAP excels in its ability to provide local-level model interpretations, enabling the identification of specific reasons for churn for individual customers. Customers with a prediction value $f(x)$ exceeding the expected value $E[f(x)]$ are classified as likely to churn. Fig. 15 presents a waterfall plot illustrating examples of two churned customers across two distinct datasets. The color of the SHAP values indicates the positive or negative impact of each feature on the predicted outcome, with blue representing features that decrease the likelihood of churn and red representing features that increase it.



(a) (b)

Fig. 15. Churn prediction SHAP waterfall for the churner sample by ROS-CatBoost (a) and CW-XGBoost (b)

6. Conclusion and future work

This study conducted a comparative analysis of methods for addressing data imbalance through two approaches: data-level approaches and algorithm-level approaches, applied to two datasets concerning customer churn in the banking and e-commerce sectors, with feature selection performed using SHAP and MRMR. Regarding data-level methods, resampling techniques, including Oversampling, Undersampling, and Hybrid resampling, were implemented. Conversely, algorithm-level methods focused on adjusting class weights in the loss function, based on cost-sensitive learning. The results indicate that, for small to medium-sized datasets, Oversampling demonstrated superior performance, while Undersampling improved Recall but reduced Precision, leading to an overall decline in model performance. Class weight adjustment also proved effective in cases where models supported such techniques.

In terms of models, the study compared ensemble models and single models. The findings revealed that bagging and boosting models exhibited robust resilience to imbalanced data. Among single models, Decision Trees performed most effectively in learning from imbalanced data, whereas other models were significantly affected, often exhibiting bias toward the majority class compared to ensemble models. Consequently, applying imbalance handling techniques is essential for models such as Logistic Regression, KNN, and SVM. To optimize the accurate prediction of customer churn, the study recommends the use of ensemble models in practical applications.

However, the adoption of ensemble models is often accompanied by limitations in transparency. To address this, the study proposes the application of the SHAP framework – an Explainable AI tool – to provide both global and local explanations, offering clear insights into the model's mechanisms and the impact of individual features on specific customers.

The research team plans to extend the study in the future to larger-scale real-world datasets to enhance the robustness of the conclusions. Additionally, the application of data imbalance handling techniques in modern deep learning models will be a significant direction for future research.

***Acknowledgments*:** This research is funded (supported) by University of Economics Ho Chi Minh City, Vietnam (UEH).

References

1. Y, N. N., T. V. Ly, D. V. T. Son. Churn Prediction in Telecommunication Industry Using Kernel Support Vector Machines. – PLOS ONE, Vol. **17**, 2022, No 5, e0267935.
2. Burez, J., D. Van den Poel. Handling Class Imbalance in Customer Churn Prediction. – Expert Systems with Applications, Vol. **36**, 2009, No 3, Part 1, pp. 4626-4636.
3. Zhu, B., B. Baesens, S. K. L. M. van den Broucke. An Empirical Comparison of Techniques for the Class Imbalance Problem in Churn Prediction. – Information Sciences, Vol. **408**, 2017, pp. 84-99.
4. Ahmad, A. K., A. Jafar, K. Aljoumaa. Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. – Journal of Big Data, Vol. **6**, 2019, No 1, 28.
5. P. Bhuse, A. Gandhi, P. Meswani, R. Muni, N. Katre, Eds. Machine Learning Based Telecom-Customer Churn Prediction. – In: Proc. of 3rd International Conference on Intelligent Sustainable Systems (ICISS'20), 3-5 December 2020.
6. Jain, H., A. Khunteta, S. Srivastava. Churn Prediction in Telecommunication Using Logistic Regression and Logit Boost. – Procedia Computer Science, Vol. **167**, 2020, pp. 101-112.
7. Lalwani, P., M. K. Mishra, J. S. Chadha, P. Sethi. Customer Churn Prediction System: A Machine Learning Approach. – Computing, Vol. **104**, 2022, No 2, pp. 271-294.
8. Pustokhina, I. V., D. A. Pustokhin, P. T. Nguyen, M. Elhoseny, K. Shankar. Multi-Objective Rain Optimization Algorithm with WELM Model for Customer Churn Prediction in Telecommunication Sector. – Complex & Intelligent Systems, Vol. **9**, 2023, No 4, pp. 3473-3485.
9. Sudharsan, R., E. Ganesh. A Swish RNN Based Customer Churn Prediction for the Telecom Industry with a Novel Feature Selection Strategy. – Connection Science, Vol. **34**, 2022, No 1, pp. 1855-1876.
10. Long, H. V., L. H. Son, M. Khari, K Arora, S. Chopra, R. Kumar et al. A New Approach for Construction of Geodemographic Segmentation Model and Prediction Analysis. – Computational Intelligence and Neuroscience, Vol. **2019**, 2019, No 1, 9252837.
11. M. Rahman, V. Kumar, Eds. Machine Learning Based Customer Churn Prediction in Banking. – In: Proc. of 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA'20), 5-7 November 2020.
12. De Lima Lemos, R. A., T. C. Silva, B. M. Tabak. Propension to Customer Churn in a Financial Institution: a Machine Learning Approach. – Neural Computing and Applications, Vol. **34**, 2022, No 14, pp. 11751-11768.
13. Peng, K., Y. Peng, W. Li. Research on Customer Churn Prediction and Model Interpretability Analysis. – PLOS ONE, Vol. **18**, 2023, No 12, e0289724.
14. Xiahou, X., Y. Harada. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. – Journal of Theoretical and Applied Electronic Commerce Research, Vol. **17**, 2022, No 2, pp. 458-475.
15. AL-Najjar, D., N. Al-Rousan, H. AL-Najjar. Machine Learning to Develop Credit Card Customer Churn Prediction. – Journal of Theoretical and Applied Electronic Commerce Research, Vol. **17**, 2022, No 4, pp. 1529-1542.
16. Amin, A., S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir et al. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. – IEEE Access, Vol. **4**, 2016, pp. 7940-7957.
17. Bharathi, S. V., D. Pramod, R. Raman. An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers. – Data, Vol. **7**, 2022, No 5, 61.
18. Brito, J. B. G., G. B. Bucco, R. Heldt, J. L. Becker, C. S. Silveira, F. B. Luce et al. A Framework to Improve Churn Prediction Performance in Retail Banking. – Financial Innovation, Vol. **10**, 2024, No 1, 17.

19. X i a h o u, X., Y. H a r a d a. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. – Journal of Theoretical and Applied Electronic Commerce Research [Internet], Vol. **17**, 2022, No 2, pp. 458-475.
20. X u, T., Y. M a, K. K i m. Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. – Applied Sciences, Vol. **11**, 2021, No 11, 4742.
21. A s i f, D., M. S. A r i f, A. M u k h e i m e r. A Data-Driven Approach with Explainable Artificial Intelligence for Customer Churn Prediction in the Telecommunications Industry. – Results in Engineering, Vol. **26**, 2025, 104629.
22. Z h o u, Y., W. C h e n, X. S u n, D. Y a n g. Early Warning of Telecom Enterprise Customer Churn Based on Ensemble Learning. – PLOS ONE, Vol. **18**, 2023, No 10, e0292466.
23. N g o, V.-B., V.-H. V u. Multi-Level Machine Learning Model to Improve the Effectiveness of Predicting Customers Churn Banks. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 3, pp. 3-20.
24. L u n d b e r g, S. M., S.-I. L e e. A Unified Approach to Interpreting Model Predictions. – Advances in Neural Information Processing Systems, Vol. **30**, 2017.
25. H a n c h u a n, P., L. F u h u i, C. D i n g. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. – IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. **27**, 2005, No 8, pp. 1226-1238.
26. N i t e s h, V. C. SMOTE: Synthetic Minority Over-Sampling Technique. – J. Artif. Intell. Res., Vol. **16**, 2002, No 1, 321.
27. H. Haibo, B. Yang, E. A. Garcia, L. Shutao, Eds. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. – In: Proc. of IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1-8 June 2008.
28. H. Han, W.-Y. Wang, B.-H. Mao, Eds. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. – In: Proc. of International Conference on Intelligent Computing, Springer, 2005.
29. B a t i s t a, G. E., R. C. P r a t i, M. C. M o n a r d. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. – ACM SIGKDD Explorations Newsletter, Vol. **6**, 2004, No 1, pp. 20-29.
30. I. Mani, I. Zhang, Eds. kNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. – In: Proc. of Workshop on Learning from Imbalanced Datasets, 2003, ICML United States.
31. Two Modifications of CNN. – IEEE Transactions on Systems, Man, and Cybernetics. SMC-Vol. **6**, 1976, No 11, pp. 769-772.
32. W i l s o n, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. – IEEE Transactions on Systems, Man, and Cybernetics, SMC- Vol. **2**, 1972, No 3, pp. 408-421.
33. F e r n á n d e z, A., S. G a r c í a, M. G a l a r, R. C. P r a t i, B. K r a w c z y k, F. H e r r e r a. Cost-Sensitive Learning. – In: A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, F. Herrera, Eds. Learning from Imbalanced Data Sets. Cham, Springer International Publishing, 2018, pp. 63-78.
34. A l-N a j j a r, D., N. A l-R o u s a n, H. A l-N a j j a r. Machine Learning to Develop Credit Card Customer Churn Prediction. – Journal of Theoretical and Applied Electronic Commerce Research [Internet], Vol. **17**, 2022, No 4, pp. 1529-1542.
35. W u, Z., L. J i n g, B. W u, L. J i n. A PCA-AdaBoost Model for E-Commerce Customer Churn Prediction. – Annals of Operations Research, 2022.
36. J o h n L u, Z. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Oxford University Press, 2010.
37. B r e i m a n, L. Bagging Predictors. – Machine Learning, Vol. **24**, 1996, pp. 123-140.

Appendix A. Data preprocessing methodology

Feature BankChurners	Transfomation BankChurners	Feature Ecommerce	Transformation Ecommerce
Attrition_Flag	Label encoding	Churn	No transformation applied
Customer_Age	Feature scaling	Tenure	No transformation applied
Gender	Label encoding	PreferredLoginDevice	One-hot encoding
Dependent_count	No transformation applied	CityTier	No transformation applied
Education_Level	One-hot encoding	WarehouseToHome	No transformation applied
Marital_Status	One-hot encoding	PreferredPaymentMode	One-hot encoding
Income_Category	One-hot encoding	Gender	One-hot encoding
Card_Category	One-hot encoding	HourSpendOnApp	One-hot encoding
Months_on_book	Feature scaling	NumberOfDeviceRegistered	No transformation applied
Total_Relationship_Count	No transformation applied	PreferedOrderCat	One-hot encoding
Months_Inactive_12_mon	No transformation applied	SatisfactionScore	No transformation applied
Contacts_Count_12_mon	No transformation applied	MaritalStatus	One-hot encoding
Credit_Limit	Feature scaling	NumberOfAddress	No transformation applied
Total_Revolving_Bal	Feature scaling	Complain	No transformation applied
Avg_Open_To_Buy	Feature scaling	OrderAmountHikeFromlastYear	No transformation applied
Total_Amt_Chng_Q4_Q1	No transformation applied	CouponUsed	No transformation applied
Total_Trans_Amt	Feature scaling	OrderCount	No transformation applied
Total_Trans_Ct	Feature scaling	DaySinceLastOrder	No transformation applied
Total_Ct_Chng_Q4_Q1	No transformation applied	CashbackAmount	No transformation applied
Avg_Utilization_Ratio	No transformation applied	–	–