Cost-Sensitive and TPE-Driven Ensemble Models with xAI for E-Commerce Churn Prediction

Luong Thanh Tam[1][0009-0007-8919-6634], Nguyen Phan Truc Ly[1][0009-0003-8744-7429], Than Ngan Tran[1][0009-0000-2499-1134], Tran Quoc Dat[1][0009-0009-9282-0746], and Nguyen Manh Tuan[1][0000-0002-7158-6550]

[1] University of Economics Ho Chi Minh City

tamluong.31231024238@st.ueh.edu.vn

Abstract. In the context of e-commerce businesses facing the challenge of retaining customers, this study focuses on enhancing model performance through the integration of cost-sensitive learning to address imbalanced data, combined with Tree-structured Parzen Estimator (TPE) for hyperparameter optimization and xAI tools for interpretability. The study implements approaches at both the data and algorithm levels, incorporating feature dimension tuning and ensemble modeling to identify an effective hybrid method for handling imbalanced data. The proposed CW-XGBoost model stands out by applying class weights, significantly improving Recall without negatively affecting other metrics; when Recall increases, Precision also rises – and by a larger margin: over 2% in the ensemble group and 0.5% with CW-XGBoost, demonstrating the absence of trade-offs found in resampling methods and resulting in a substantial enhancement of overall model generalization. Furthermore, the TPE-based optimization produced the TPE-CW-XGBoost model, which improved most metrics, achieving a Recall of 0.931 and ROC-AUC of 0.98, demonstrating superior applicability. To address the transparency limitations of black-box models, the study integrates xAI via the SHAP method, providing interpretability at both global and local levels, surpassing traditional methods such as gain importance. The study identifies Tenure and Complain as the two most influential factors driving customers' churn decisions.

Keywords: Customer churn prediction, imbalanced data, hyperparameters optimization, tree-structure parzen estimator, x-ai.

1 Introduction

Customer Relationship Management (CRM) is defined as a process by which companies manage their interactions with customers by integrating data from various sources and analyzing that data [1]. CRM plays a critical role in building customer loyalty, especially in the digital age, where e-commerce has experienced unprecedented growth. This growth has brought numerous opportunities for businesses but also presents significant challenges in maintaining competitive advantage—particularly in non-contractual environments like e-commerce [2], where customers can continuously change their purchasing behavior without notifying the company [3]. In other words, the absence of formal contractual obligations between customers and businesses makes tracking and predicting customer churn behavior particularly complex in such settings [4]. Consequently, this has driven the development of customer retention and churn prediction strategies, with particular emphasis on targeted marketing campaigns aimed at customers who exhibit signs of imminent churn.

With the rapid advancement of technology and the ability to collect large volumes of customer data, machine learning techniques have emerged as powerful tools for analyzing and uncovering hidden patterns in data across various industries such as banking [5], telecommunication [6], and aviation…. These models can effectively identify risk factors leading to churn and predict which customers are likely to leave in the future, often with high levels of accuracy. Early prediction enables businesses to proactively implement retention strategies in a timely and effective manner.

Although previous studies have achieved remarkable results, several critical research gaps remain unaddressed. Current churn prediction models still lack effective approaches for handling class imbalance as well as robust hyperparameter optimization techniques to maximize predictive performance. Moreover, despite the rapid growth of e-commerce, this domain has received limited attention in churn prediction research, resulting in a shortage of valuable insights to support decision-makers.

To fill this research gap, we propose an approach that balances predictive accuracy and model interpretability for predicting customer churn behavior in the e-commerce domain. The method begins with a comprehensive and systematic data preprocessing procedure, followed by a comparison of multiple machine learning algorithms to select the model with optimal performance. We then employ the SHAP method [7] to clarify the role of each feature in churn decisions, both at the global and local levels. This approach not only improves the reliability of model interpretation but also provides practical value by enabling marketing teams to leverage insights on churn likelihood and the specific reasons predicted for each customer [8]. Consequently, businesses can design targeted marketing strategies by segment and implement personalized anti-churn initiatives, thereby enhancing customer retention in the highly competitive e-commerce environment. To further strengthen its practical relevance for managers and strategists, our study also proposes the development of an interactive web-based platform that enables real-time churn monitoring and causal analysis, thereby supporting timely and effective decision-making for maintaining customer relationships.

This study makes the following contributions:

- Evaluation of imbalance-handling techniques: We systematically assess methods at both the data-level and algorithm-level. The findings demonstrate that cost-sensitive learning combined with ensemble models stands out, as it achieves significant gains in Recall with only minimal trade-off in Precision. This balance allows the model to better identify churners while reducing misclassification of non-churners as churners.
- Advanced hyperparameter optimization with TPE: We propose the integration of TPE to optimize CW-XGBoost. This approach improves both Precision and Recall simultaneously, narrowing the trade-off observed in earlier imbalance-handling stages. The optimized model, TPE-CW-XGBoost, achieves Precision and Recall values of approximately 0.94, demonstrating enhanced generalization capability.
- Improving model interpretability with SHAP: To address the “black-box” nature of ensemble models, we employ SHAP for both global and local explanations. This enhances the transparency of the predictive process and strengthens the model’s applicability in practice.

The remainder of this study is organized as follows. The next section presents the literature review, followed by the research methodology. We then describe the experiment and analyze the results. Finally, the study concludes with the main contributions and proposes directions for future research.

2 Related work

2.1 Customer churn prediction

Several studies [4], [9] have indicated that most customer churn prediction models are developed within contractual business environments, where churn is understood as a customer-initiated termination of a service agreement through cancellation or breach of agreed terms. However, this approach does not fully align with non-contractual business models, such as those in the e-commerce sector, where customers may stop using the service at any time without performing any formal cancellation actions or notifying the business.

Among interpretable machine learning techniques, logistic regression and decision trees are commonly employed in churn prediction [10], [11] due to their simple operational nature. However, recent research in churn modeling shows that ensemble models tend to outperform other approaches in terms of prediction performance [12].

Notably, the Extreme Gradient Boosting (XGBoost) algorithm has been widely applied to various predictive tasks. XGBoost is popular because of the ability in model nonlinear relationships within data and for its fast-processing speed, which enables researchers to experiment with multiple hyperparameter settings and select the most optimal configuration.[13] proposed an XGBoost model combined with a Genetic Algorithm (GA) to build a churn prediction system based on banking data from Kaggle. The model used the SMOTEENN method to address class imbalance and achieved an F1-score of 90% and an AUC of 99%, outperforming other machine learning models. [14] compares the performance of five ensemble models—AdaBoost, GBT, XGBoost, CatBoost, and LightGBM—on the churn prediction task and shows that XGBoost outperforms the remaining classifiers. Another study [15] reported that XGBoost achieved 99% accuracy on a telecommunication churn dataset.

Other ensemble models have also demonstrated strong performance in various studies. For instance, [8] focused on customer churn in the rental business, particularly at a water purifier rental company in South Korea. Using a large dataset (approximately 84,000 customers), the study developed a churn prediction model using machine learning algorithms. The LightGBM model achieved an F1-score of 93% and an AUC of 88%. Studies [16], [17] demonstrated the superior performance of Random Forest on banking datasets compared to other models.

A major challenge in churn prediction is class imbalance. To address this issue, Rahman and Kuma [16] applied random oversampling to duplicate minority class instances until a near balance was achieved, while study [18] adopted random undersampling to reduce the size of the majority class. Additionally, SMOTE (Synthetic Minority

Over-sampling Technique) is widely used [3], [12], [17] to generate synthetic samples by interpolating between existing minority class instances leading to improved prediction performance [19]. However, despite their effectiveness, most data-level handling techniques such as SMOTE or random oversampling introduce synthetic data or modify existing samples, which can distort the original data distribution and alter underlying patterns in the dataset [20]. This structural modification raises concerns about the generalizability and robustness of models trained on artificially balanced datasets.

Numerous studies have been conducted to improve the performance of churn prediction models, and hyperparameter optimization plays a key role in this effort. For instance, [13] employed Genetic Algorithms to optimize XGBoost, significantly improving model performance. Likewise, Grid Search has been widely adopted in studies [12], [17], [21] and has contributed to improved accuracy. However, traditional hyperparameter optimization methods such as Grid Search is known to be computationally expensive and inefficient, especially when dealing with a large hyperparameter space. Despite these efforts, a clear research gap remains in the exploration of more advanced and efficient hyperparameter optimization techniques for churn prediction models.

2.2 Model interpretation

Investigating consumer decision-making is a central concern in marketing and corporate planning, with various frameworks and approaches established to interpret the reasons behind customer attrition [6]. It is worth noting that most current studies use surveys as the primary data collection method. While such surveys can help identify factors influencing purchase behavior, they often remain at a general level and do not accurately reflect individual differences among customers or align with the specific operations of each business. However, with the advent of big data and advanced analytical tools, modern research has shifted towards data-driven approaches. In particular, model interpretability can offer invaluable insights into the reasons behind customer churn.13 In churn studies, [13] utilized Shapley values to identify key features enabling commercial banks to enhance services, retain customers, and build early warning systems. Similarly, [12] illustrated model interpretation using SHAP and Explainable Boosting Machine to determine influential features in streaming services, supporting customer retention strategies. SHAP offers both global and local interpretability, making model predictions more trustworthy and actionable.

Nevertheless, despite the growing use of SHAP and other techniques in sectors like banking, telecommunications home appliance rental business…. very few studies have explored model interpretability in the context of e-commerce churn prediction, where customer behavior is more volatile and less structured [21]. There is a need for research that interprets churn prediction models in e-commerce, supporting personalized retention strategies and enhancing the practical utility of churn models in real-world commercial platforms.

3 Methodology

3.1 Handling imbalanced data techniques

One of the major challenges in the field of machine learning is the problem of class imbalance, in which the number of samples belonging to the majority class significantly outweighs those of the minority class [22]. This phenomenon often causes learning models to be biased toward predicting the majority class, while neglecting or poorly identifying the minority class—which is typically of critical importance in applications such as fraud detection, rare disease diagnosis, or customer churn prediction. To address this issue, existing studies have proposed two main approaches: data-level methods and algorithm-level methods.

At the data level, techniques directly intervene in the distribution of training samples to rebalance the majority and minority classes [23]. The most common approach is oversampling, which involves replicating or generating synthetic samples for the minority class (e.g., SMOTE), thereby narrowing the gap between the two classes. In some cases, undersampling is applied by removing a portion of majority class data; however, this strategy is often ineffective when the original dataset is small in scale. In addition, another research branch focuses on dimensionality reduction and feature selection. Through methods such as wrapper approaches, the model is repeatedly trained on different feature subsets and evaluated using appropriate metrics, thereby identifying the feature set with the strongest discriminative capacity [24]. Representative techniques include Recursive Feature Elimination (RFE) [25] —which progressively removes less important features based on model-assigned weights [26]—or strategies such as Sequential Forward Selection and Sequential Backward Selection, which iteratively add or remove features to determine the optimal subset. These methods indirectly enhance model performance when dealing with imbalanced data and [27], [28].

At the algorithm level, instead of altering the data, methods emphasize adjusting the model's learning mechanism to improve its ability to recognize the minority class [29]. An important direction is cost-sensitive learning, in which misclassifying minority class samples is assigned a higher cost compared to the majority class. This approach is often implemented through class weighting and the selection of appropriate evaluation metrics such as F1-score, Recall, or AUC, rather than Accuracy, which is prone to bias. Furthermore, research has also leveraged the power of ensemble models such as Bagging, Boosting, and Random Forest. By combining multiple models into a robust predictive system, ensemble methods significantly improve classification effectiveness for the minority class, particularly when integrated with data balancing techniques [30].

3.2 Tree-structure parzen estimator (TPE) for hyperparameters optimization

Hyperparameter optimization plays a critical role in boosting the performance of machine learning models [31]. The Tree-structured Parzen Estimator, introduced by [32], is a probabilistic, model-based optimization algorithm specifically designed for efficient hyperparameter tuning in machine learning. TPE leverages Bayesian optimization principles to intelligently explore the hyperparameter space. Rather than modeling the objective function directly, TPE constructs two non-parametric probability density functions: one representing the distribution of hyperparameter configurations associated with lower (better) loss values, and the other with higher (worse) loss values. The algorithm then selects new candidate configurations by maximizing the ratio of these densities, thereby focusing the search on promising regions of the parameter space. Key advantages of TPE is its ability to handling hierarchical, and conditional search spaces where certain hyperparameters are only relevant depending on the values of others—scenarios common in deep learning architectures and ensemble models. Moreover, TPE is highly scalable and compatible with parallel evaluation, making it wellsuited for large-scale experiments.

3.3 SHAP for model interpretation

In modern data science, understanding and interpreting how machine learning models generate predictions has become a critical requirement—especially when these systems are deployed in real-world contexts that demand transparency and accountability. One of the most prominent tools addressing this challenge is Shapley Additive Explanations [7], a method grounded in the concept of Shapley values [33] from cooperative game theory. Originally developed to fairly distribute gains among players in cooperative games, the Shapley value has been reinterpreted in the machine learning context to quantify the contribution of each input feature to a model's output [13].

In practical data science workflows, SHAP serves as a post hoc model analysis tool, enabling practitioners to understand why a model produces a particular prediction. Rather than relying solely on global performance metrics, SHAP provides local-level insights by attributing the predicted outcome of an individual data point to specific features. This level of interpretability is essential not only for internal model validation but also for effectively communicating findings to non-technical stakeholders. Beyond visualization, SHAP is also employed to detect data irregularities, evaluate model stability under different input distributions, and assess bias across different user groups.

3.4 Evaluation metrics

In the context of customer churn prediction, the components of the confusion matrix are defined as follows:

- True Positive (TP): Cases in which customers actually churn and the model correctly predicts churn. These represent accurate predictions of churners.
- True Negative (TN): Cases in which customers actually remain and the model correctly predicts non-churn. These represent accurate predictions of loyal customers.
- False Positive (FP): Cases in which customers actually remain but the model incorrectly predicts churn. Such misclassifications may lead businesses to implement unnecessary retention strategies, thereby wasting resources.
- False Negative (FN): Cases in which customers actually churn but the model incorrectly predicts non-churn. These errors are particularly critical as they result in missed opportunities for intervention, leading to customer attrition and potential revenue loss.

To evaluate the performance of the model, we employ established evaluation metrics: Accuracy measures the proportion of correct predictions over the entire dataset. However, this metric can be misleading in imbalanced datasets where one class dominates.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision reflects the proportion of positive predictions that are actually correct. It is important in reducing false alarms (false positives).

$$Precision = TP / (TP + FP)$$

Recall indicates the ability of the model to correctly identify positive cases, which is especially crucial when the cost of missing a positive instance is high. In customer churn prediction, Recall is prioritized because it ensures the model identifies most customers at risk of leaving, allowing businesses to take timely retention actions and minimize losses.

$$Recall = TP / (TP + FN)$$

In practice, Precision and Recall often involve a trade-off: increasing Recall may reduce Precision, and vice versa.

4 Experiment

4.1 Experiment design

In this study, the authors construct an experimental framework to examine the following issues:

- The effectiveness of data imbalance handling techniques, including data-level (class distribution and feature dimension) and algorithm-level (cost-sensitive learning and ensemble models such as bagging and boosting).
- The effectiveness of combining data imbalance handling methods when integrated into hybrid models.
- Evaluation of the effectiveness of the Tree-structured Parzen Estimator in hyperparameter optimization of the proposed model.
- The proposal of applying xAI to explain the model, thereby enhancing model transparency.

The proposed experimental framework is presented in detail in Fig. 1.

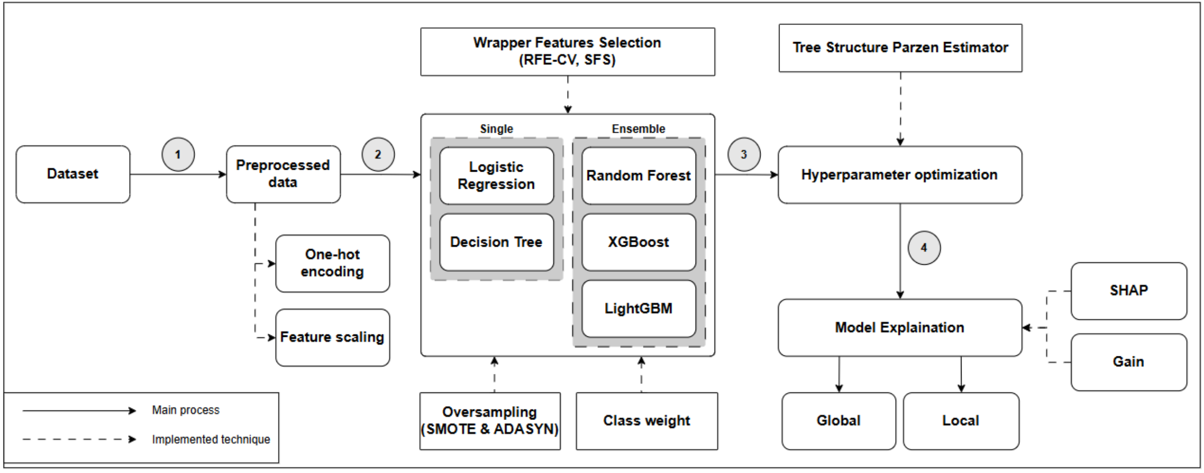


Fig. 1. Proposed Experimental Framework of the Study

Step 1-1: Data Collection: The initial phase involves collecting structured customer-level data from a real e-commerce system. The dataset comprises a diverse combination of demographic, behavioral, and financial attributes. It also includes a binary churn label. This multidimensional dataset serves as the foundation for subsequent steps in building churn prediction models in e-commerce applications.

Step 1-2: Data Preprocessing: To construct an effective model, the raw data is processed through several preprocessing steps as follows:

- Missing values are imputed using the median for continuous numerical data and the mode for categorical features.
- One-hot encoding is applied to categorical variables.
- Numerical continuous features are normalized using Min-Max scaling.

Step 2: Modeling Classifiers: Classifiers, including single models (Logistic Regression, Decision Tree) and ensemble models (Random Forest, XGBoost, and LightGBM), are set up with default parameters and a fixed random state of 42 to ensure reproducibility. Imbalance handling techniques are incorporated, including Resampling-Oversampling (SMOTE and ADASYN), cost-sensitive learning with class weights, and feature dimension adjustment using wrapper methods (RFE with cross-validation and Sequential Feature Selection – SFS). The following cases are considered:

- No processing: No feature selection, resampling, or class weighting applied. This serves to compare the inherent robustness of single models and ensemble models (*).
- Resampling: Oversampling applied using SMOTE and ADASYN techniques (**).
- Class weight: Applying the scale_pos_weight parameter for XGBoost and class_weight='balanced' for the remaining models (***).
- Feature selection: Applying two techniques—RFE with cross-validation and SFS—for all models with default parameters, optimized using Recall as the primary metric (****).
- Feature selection – Resampling: Combination of (**) and (****).
- Feature selection – Class weight: Combination of (***) and (****).

The results are evaluated using StratifiedKFold (k=5) with averaging. The best-performing model within each technique group is selected as the optimal model.

Step 3: Hyperparameter Optimization: The model with the best performance from Step 2 is further tuned using the Tree-structured Parzen Estimator, with optimization focused on Recall to ensure the model prioritizes the churned customer group.

Step 4: Model Explainability with xAI-SHAP: The selected model is explained using xAI-SHAP to enhance the transparency of the black-box model. This is compared with traditional techniques such as the built-in gain importance, providing interpretability at both global and local levels.

4.2 Experimental environment

The experiments were conducted in the Google Colab environment using the Python programming language, with the main libraries including: pandas, numpy, matplotlib, sklearn, xgboost, lightgbm, imbalanced-learn, shap, and optuna.

4.3 Experiment preparation

In this study, the authors utilize a customer churn dataset in the e-commerce domain. The dataset contains structured customer-level information, covering demographic, behavioral, and transactional attributes. Key variables include churn status, tenure, city tier, app usage time, number of registered devices, and satisfaction score. The dataset consists of 5,630 customers with a clear class imbalance, as visualized in Fig. 2, showing a churn-to-non-churn ratio of approximately 1:5. The dataset will be processed according to Step 1-2 presented in the previous section 4.1.

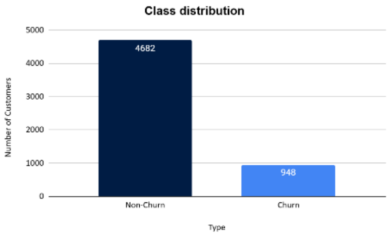


Fig. 2. Class distribution

5 Results

In this section, the study analyzes the results of identifying the most effective hybrid technique for handling imbalanced data, evaluated through recall and its trade-off with precision. Based on this, the best-performing classification models within the group are determined, followed by hyperparameter optimization using TPE, and enhanced model transparency through the application of xAI techniques that provide explanations at both global and local levels.

5.1 Handling imbalanced data

For the process of handling imbalanced data, the study considers the following methods: No processing, Oversampling, Class weight, Feature selection, FS & Class weight, and FS & Oversampling, applied to two groups of models: Single models (Logistic Regression, Decision Tree), with results presented in Fig. 3, and Ensemble models (Bagging – Random Forest, Boosting – XGBoost and LightGBM), with results presented in Fig. 4.

Although there is a clear performance gap showing that ensemble models learn better than single models—both precision and recall of ensemble models are higher, and the gap between these two metrics is smaller—across other methods (except No processing), both groups exhibit a common trend: oversampling and class weight methods prove particularly effective for single models and relatively stable for ensemble models. In contrast, feature selection for adjusting feature dimensions does not show any significant changes for either group of models. Meanwhile, feature selection integrated with oversampling or class weight demonstrates the strongest improvement in recall but suffers from a severe trade-off with precision, leading to excessive misclassification of non-churners as churners. Consequently, deploying models with these techniques in practice may result in substantial costs when implementing CRM strategies for customer retention.

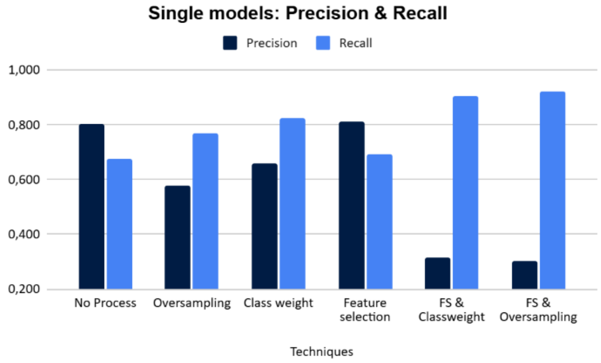


Fig. 3. Performance of single models: Precision & Recall

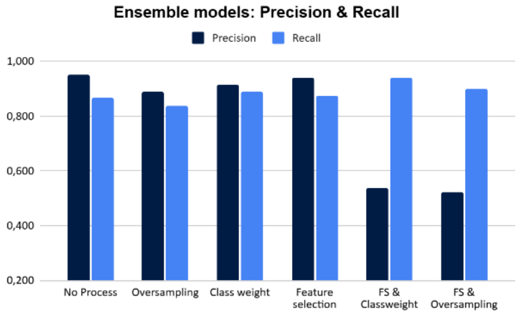


Fig. 4. Performance of ensemble models: Precision & Recall

The results indicate that the most effective techniques for handling data imbalance are No Processing and Class weight. These approaches are optimal in terms of time efficiency, as they do not involve expanding the dataset and thus avoid increasing modeling time, while also mitigating the precision trade-off observed in feature selection combined with oversampling or class weight.

5.2 Modeling and hyperparameters with tree-structure parzen estimator

For the two groups, No Processing and Class weight, applied to ensemble models and identified as the most effective in handling imbalanced data, the study compares the models under these settings, with results presented in Fig. 5. The findings reveal that XGBoost demonstrates outstanding generalization capability both before and after the application of class weighting. Specifically, under the No Processing setting, XGBoost achieves the highest Recall among the three models compared to LightGBM and Random Forest. After applying Class weight, LightGBM shows a substantial improvement in Recall but at the expense of a significant loss in Precision. In contrast, XGBoost achieves a balanced convergence of both Precision and Recall at 0.925, representing a more optimal outcome compared to LightGBM.

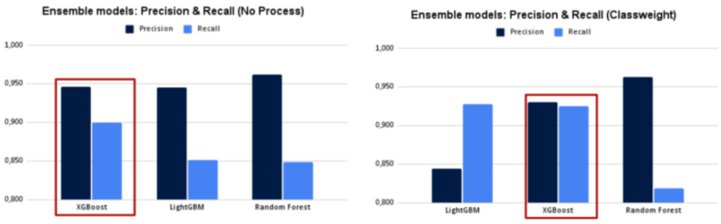


Fig. 5. Performance of ensemble models: Precision & Recall (No Process, Classweight)

In addition to the scale_pos_weight parameter, XGBoost contains multiple other hyperparameters that require careful optimization, including *n_estimators*, *max_depth*, *learning_rate*, *subsample*, and *colsample_bytree*, among others. This study applies the TPE algorithm, supported by the Optuna library, to conduct hyperparameter tuning. The results demonstrate that the TPE-CW-XGBoost model delivers superior performance, not only enhancing the key objective metric Recall but also mitigating the trade-off in Precision typically introduced by class weight adjustments. Fig. 6 illustrates the comparative performance of XGBoost across different optimization stages, from baseline to CW-XGBoost and finally to TPE-CW-XGBoost. The corresponding optimal hyperparameters are detailed in Table 1.

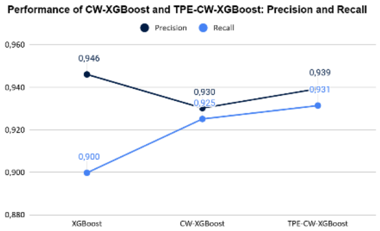


Fig. 6. Performance of CW-XGBoost and TPE-CW-XGBoost: Precision and Recall

Table 1. Optimal Hyperparameters of TPE-CW-XGBoost with Tree-Structured Parzen Estimator

Hyperparameter	Value
n_estimators	323
max_depth	9
learning_rate	0.035
subsample	0.946
colsample_bytree	0.79698
scale_pos_weight	4.939
random_state	42
eval_metric	logloss

For the operation process of TPE, several key pieces of information are visualized in Fig. 7, which indicate that the two most important hyperparameters are max depth and learning rate. Also in Fig. 7, the Contour plot illustrates the interaction between these parameters, with the background color representing the magnitude of the model evaluation metrics.

Following the identification of the TPE-CW-XGBoost model as a promising high-performance approach for e-commerce customer churn prediction, evaluation using the confusion matrix (Fig. 8) and the changes in TP, FN, and FP (Table 2) indicates that the model improvement process effectively targets the minority class – churners – through class weighting (CW) while mitigating trade-offs via hyperparameter optimization with TPE. Specifically, while applying cost-sensitive learning in XGBoost increased TP, it also resulted in a higher FP rate, misclassifying non-churners as churners, as reflected in the decline of precision shown in Fig. 5. In contrast, the integration of TPE reduced FP while further increasing TP, representing a comprehensive enhancement of overall model performance.

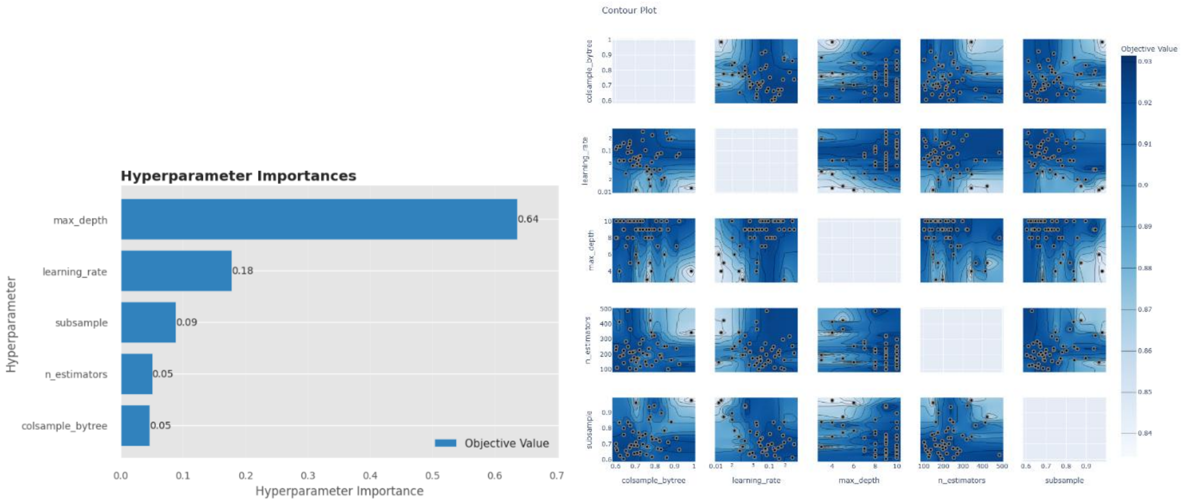


Fig. 7. Hyperparameter importance and Contour plot - Interaction between Parameters

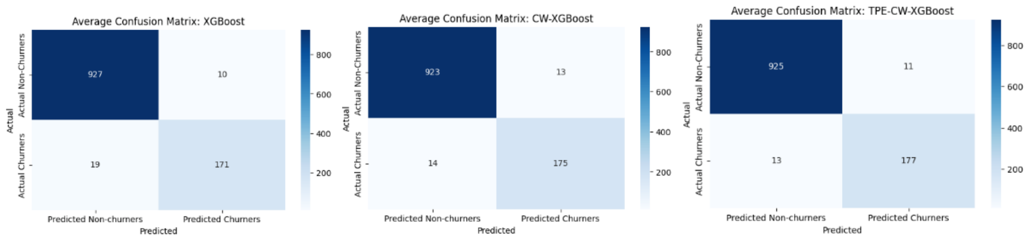


Fig. 8. Average Confusion Matrix: XGBoost, CW-XGBoost, TPE-CW-XGBoost

Table 2. Changes in TP, FN, FP from XGBoost to TPE-CW-XGBoost

Model	TP	FN	FP
XGBoost	-	-	-
CW-XGBoost	↑	↓	↑
TPE-CW-XGBoost	↑	↓	↓

To examine the potential overfitting of the TPE-CW-XGBoost model, a learning curve was constructed using two evaluation metrics, namely Recall and F1-Score. The results, visualized in Fig. 9, indicate that the cross-validation scores for both metrics exhibit an upward trend and gradually converge toward the training scores, suggesting improved generalization capability of the model and the absence of pronounced overfitting.

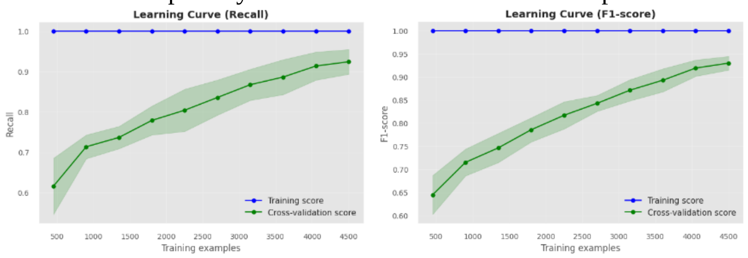
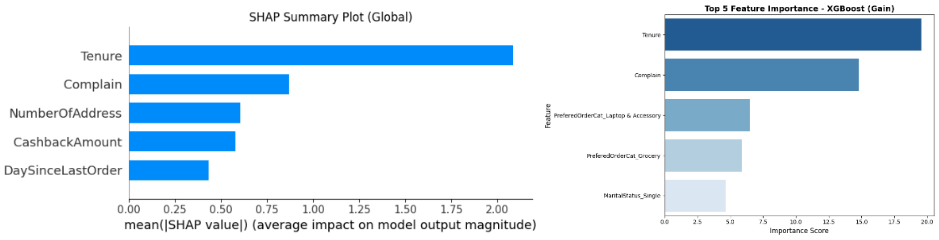


Fig. 9. Learning curve of TPE-CW-XGBoost with Recall and F1-Score

5.3 Model interpretation with x-AI

In this paper, we employ the SHAP method to explain the learning process of the model, specifically the contribution level of each feature in enhancing the model's predictive capability. This provides a more intuitive representation of the degree of influence that features exert on the model. In previous studies, authors often explained models through built-in importance measures, such as coefficients from logistic regression or gain importance from tree-based models. This study proposes an explanatory framework using SHAP. Figure 11 compares feature importance as measured by gain importance and SHAP. Although differences exist, the two most important features identified remain *Tenure* and *Complain*. However, SHAP stands out compared to gain importance thanks to its ability to identify and explain the direction of feature effects through the SHAP Summary Plot (beeswarm), thereby enhancing model transparency. Figure 11 illustrates the direction of influence of the two variables *Tenure* and *Complain*: these variables exert opposite effects, where *Tenure* shows a negative relationship with the predicted outcome, while *Complain* exhibits a positive relationship. This implies that customers with longer engagement periods and no complaints are more likely to remain with the e-commerce business, aligning with economic intuition.



(a) (b)

Fig. 10. Feature Importance Comparison: SHAP Bar (a) and Gain (b)

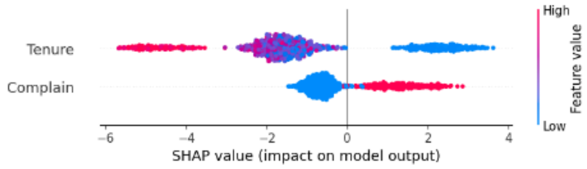


Fig. 11. Global explaination for Tenure and Complain by SHAP

In addition, SHAP stands out due to its ability to provide local explanations for individual customers, allowing the identification of the contribution of each feature to the expected value E[f(x)]—that is, the average prediction value across the entire dataset—thereby forming the individual prediction value f(x) for each specific customer. If $f(x) > E[f(x)]$, the customer is classified as a churner, whereas if $f(x) \leq E[f(x)]$, the customer is considered a non-churner.

Fig. 13 illustrates the results of local explanation by SHAP for two sample customers: one non-churner and one churner. For example, for a customer classified as a churner, the global expected value is E[f(x)]=0.666. Among the features, low Tenure (value = 1) contributes the most with +2.56, while a high NumberOfAddress adds +0.79, pushing the individual prediction value up to f(x)=2.965. Although this customer has no complaints (a feature that decreases f(x) with a contribution of −0.76), the cumulative effect of the other features still leads the model to predict that this customer has a high likelihood of churn.

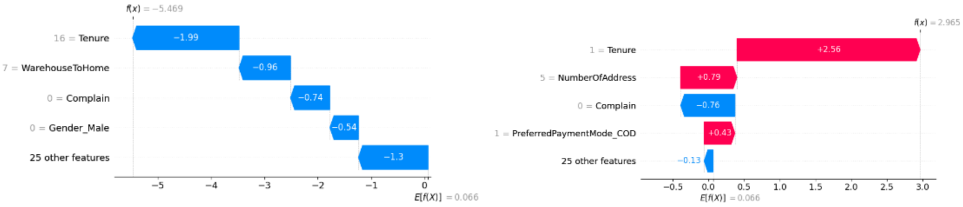


Fig. 12. Local explaination for sample churners and non-churners by SHAP

6 Conclusion and future works

In recent years, numerous studies have been conducted on customer churn prediction. However, the issue of data imbalance remains a significant challenge, particularly with respect to the trade-off between precision and recall in improving model performance. Moreover, widely applied machine learning models are predominantly ensemble models, which are often regarded as “black boxes” with limited interpretability, thereby intensifying the trade-off between model performance and explainability.

Against this backdrop, our study addresses three key issues:

- Handling data imbalance: We categorize existing approaches into two main groups: Data-level (resampling, feature dimension reduction) and Algorithm-level (ensemble models, cost-sensitive learning). Experimental results demonstrate that a hybrid model combining ensemble models with classweight yields superior performance. In particular, CW-XGBoost achieves stable Recall and Precision values greater than 0.9, with only minimal discrepancies between the two. Furthermore, applying Classweight provides computational efficiency, as it does not increase the training sample size nor require iterative feature selection processes such as wrapper methods.
- Enhancing model generalization: The study employs the TPE technique from Optuna for hyperparameter optimization, thereby upgrading CW-XGBoost into TPE-CW-XGBoost. The optimized model achieves Recall = 0.931 and Precision = 0.939, reflecting notable improvements compared to using classweight alone. Although the model already demonstrates very high performance, validation through the learning curve indicates no evidence of overfitting. Visualization of the optimization process further identifies max_depth and learning_rate as the two most influential hyperparameters.
- Improving model transparency: To strengthen interpretability, the study introduces and applies SHAP as the primary explanatory framework. Compared with traditional methods such as gain importance from tree-based models, SHAP demonstrates superior capabilities by providing both global and local explanations, while explicitly capturing the direction of influence for individual features on prediction outcomes.

Although the results are highly promising, with Recall and Precision approaching 0.94 in the TPE-CW-XGBoost model, the study still has certain limitations. Specifically, the current investigation is restricted to traditional machine learning models and has not yet explored advanced neural network architectures such as MLP, CNN, or LSTM. Moreover, the dataset employed consists solely of independent records, without incorporating time-series or unstructured data. In future work, we aim to extend the study of imbalanced data handling to modern architectures applied to unstructured and temporal data, while integrating state-of-the-art explainability techniques to further enhance model transparency.

Acknowledgement

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

References

1. M. BARDICCHIA, Digital CRM: Strategies and Emerging Trends: Building Customer Relationship in the Digital Era. Amazon Digital Services LLC - Kdp, 2022.
2. K. Coussement, S. Lessmann, and G. Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry," Decis. Support Syst., vol. 95, pp. 27–36, Mar. 2017, doi: 10.1016/j.dss.2016.11.007.
3. X. Xiahou and Y. Harada, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," J. Theor. Appl. Electron. Commer. Res., vol. 17, no. 2, Art. no. 2, June 2022, doi: 10.3390/jtaer17020024.
4. A. Tamaddoni, M. M. Sepehri, S. Choobdar, and B. Teimourpour, "Modeling customer churn in a non-contractual setting: the case of telecommunications service providers," 2013.
5. T.-T. Luong, V.-G. Luong, A. H. T. Tran, and T. M. Nguyen, "Application of Machine Learning Techniques for Customer Churn Prediction in the Banking Sector," Interdiscip. J. Inf. Knowl. Manag., vol. 20, p. 009, Mar. 2025.
6. P. Boozary, S. Sheykhan, H. GhorbanTanhaei, and C. Magazzino, "Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction," Int. J. Inf. Manag. Data Insights, vol. 5, no. 1, p. 100331, June 2025, doi: 10.1016/j.jjimei.2025.100331.
7. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 25, 2017, arXiv: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.
8. Y. Suh, "Machine learning based customer churn prediction in home appliance rental business," J. Big Data, vol. 10, no. 1, p. 41, Apr. 2023, doi: 10.1186/s40537-023-00721-8.
9. K. Matuszelański and K. Kopczewska, "Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach," J. Theor. Appl. Electron. Commer. Res., vol. 17, no. 1, Art. no. 1, Mar. 2022, doi: 10.3390/jtaer17010009.
10. M. Óskarsdóttir, K. E. Gísladóttir, R. Stefánsson, D. Aleman, and C. Sarraute, "Social networks for enhanced player churn prediction in mobile free-to-play games," Appl. Netw. Sci., vol. 7, no. 1, Art. no. 1, Dec. 2022, doi: 10.1007/s41109-022-00524-5.
11. T. Gattermann-Itschert and U. W. Thonemann, "How training on multiple time slices improves performance in churn prediction," Eur. J. Oper. Res., vol. 295, no. 2, pp. 664–674, Dec. 2021, doi: 10.1016/j.ejor.2021.05.035.

12. U. Gani Joy, K. E. Hoque, M. Nazim Uddin, L. Chowdhury, and S.-B. Park, “A Big Data-Driven Hybrid Model for Enhancing Streaming Service Customer Retention Through Churn Prediction Integrated With Explainable AI,” IEEE Access, vol. 12, pp. 69130–69150, 2024, doi: 10.1109/ACCESS.2024.3401247.
13. K. Peng, Y. Peng, and W. Li, “Research on customer churn prediction and model interpretability analysis,” PLOS ONE, vol. 18, no. 12, p. e0289724, thg 12 2023, doi: 10.1371/journal.pone.0289724.
14. R. P. Sari, F. Febriyanto, and A. C. Adi, “Analysis Implementation of the Ensemble Algorithm in Predicting Customer Churn in Telco Data: A Comparative Study,” Informatica, vol. 47, no. 7, Art. no. 7, July 2023, doi: 10.31449/inf.v47i7.4797.
15. P. Swetha and R. B. Dayananda, “A customer churn prediction model in telecom industry using Improved_XGBoost,” Int. J. Cloud Comput., vol. 12, no. 2/3/4, p. 277, 2023, doi: 10.1504/IJCC.2023.130903.
16. M. Rahman and V. Kumar, “Machine Learning Based Customer Churn Prediction In Banking,” in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Oct. 2020, pp. 1196–1201. doi: 10.1109/ICECA49313.2020.9297529.
17. P. P. Singh, F. I. Anik, R. Senapati, A. Sinha, N. Sakib, and E. Hossain, “Investigating customer churn in banking: a machine learning approach and visualization app for data science and management,” Data Sci. Manag., vol. 7, no. 1, pp. 7–16, Mar. 2024, doi: 10.1016/j.dsm.2023.09.002.
18. T. M. Alam et al., “Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World,” Comput. J., vol. 64, no. 11, pp. 1731–1746, Nov. 2021, doi: 10.1093/comjnl/bxaa056.
19. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, June 2002, doi: 10.1613/jair.953.
20. H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), June 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.
21. S. S. Poudel, S. Pokharel, and M. Timilsina, “Explaining customer churn prediction in telecom industry using tabular machine learning models,” Mach. Learn. Appl., vol. 17, p. 100567, Sept. 2024, doi: 10.1016/j.mlwa.2024.100567.
22. G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” Expert Syst. Appl., vol. 73, pp. 220–239, May 2017, doi: 10.1016/j.eswa.2016.12.035.
23. B. Nikpour, F. Rahmati, B. Mirzaei, and H. Nezamabadi-pour, “A comprehensive review on data-level methods for imbalanced data classification,” Expert Syst. Appl., vol. 295, p. 128920, Jan. 2026, doi: 10.1016/j.eswa.2025.128920.
24. M. Walowe Mwadulo, “A Review on Feature Selection Methods For Classification Tasks,” Int. J. Comput. Appl. Technol. Res., vol. 5, no. 6, pp. 395–402, June 2016, doi: 10.7753/IJCATR0506.1013.
25. [25] M. Kuhn and K. Johnson, “Data Pre-processing,” in Applied Predictive Modeling, M. Kuhn and K. Johnson, Eds., New York, NY: Springer, 2013, pp. 27–59. doi: 10.1007/978-1-4614-6849-3_3.
26. M. Awad and S. Fraihat, “Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems,” J. Sens. Actuator Netw., vol. 12, no. 5, Art. no. 5, Oct. 2023, doi: 10.3390/jsan12050067.
27. E. C. Blessie and E. Karthikeyan, “Sigmis: A Feature Selection Algorithm Using Correlation Based Method,” J. Algorithms Comput. Technol., vol. 6, no. 3, pp. 385–394, Sept. 2012, doi: 10.1260/1748-3018.6.3.385.
28. R. L. Michel, “The Keats Bicentennial,” Arion J. Humanit. Class., vol. 28, no. 3, pp. 1–1, 2021, doi: 10.2307/arion.28.3.0001.
29. G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” SIGKDD Explor Newsl, vol. 6, no. 1, pp. 20–29, Tháng Sáu 2004, doi: 10.1145/1007730.1007735.
30. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches,” IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 42, no. 4, pp. 463–484, July 2012, doi: 10.1109/TSMCC.2011.2161285.
31. R. Islam, A. Sultana, and MD. N. Tuhin, “A comparative analysis of machine learning algorithms with tree-structured parzen estimator for liver disease prediction,” Healthc. Anal., vol. 6, p. 100358, Dec. 2024, doi: 10.1016/j.health.2024.100358.
32. J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for Hyper-Parameter Optimization,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2011. Accessed: Aug. 20, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html
33. L. S. Shapley, “Contributions to the Theory of Games, Volume II,” H. W. Kuhn and A. W. Tucker, Eds., Princeton University Press, 1953, pp. 307–318. doi: 10.1515/9781400881970-018.