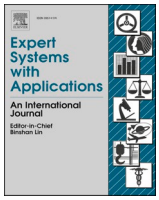
Expert Systems With Applications 247 (2024) 123251

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Imbalanced customer churn classification using a new multi-strategy collaborative processing method

Congjun Rao [a], Yaling Xu [a,*], Xinping Xiao [a], Fuyan Hu [a], Mark Goh [b]

[a] School of Science, Wuhan University of Technology, Wuhan 430070, PR China
[b] NUS Business School & The Logistics Institute-Asia Pacific, National University of Singapore, 119623, Singapore

ARTICLE INFO

Keywords:
Imbalanced customer data
Multi-strategy collaborative processing
IADASYN
FLCatBoost

ABSTRACT

The rapid advancement of big data and artificial intelligence heralds a dual-edged era of opportunities and challenges for the banking sector. Indeed, enhancing a model's capability to accurately classify imbalanced datasets represents a critical challenge within the field of customer churn prediction (CCP). In this paper, to address the challenges presented by the problem of imbalanced customer classification, a new multi-strategy collaborative processing method named IADASYN-FLCatBoost is proposed from dual perspectives: data and algorithm. At the data level, the traditional Adaptive Synthetic (ADASYN) sampling is improved, that is, the LOF (Local Outlier Factor) algorithm is introduced to eliminate outliers, and the classification features are specially processed to synthesize new minority class samples, thus an improved ADASYN (IADASYN) algorithm is obtained. At the algorithm level, the Focal Loss is embedded into the CatBoost ensemble learning framework to form a new Focal Loss-CatBoost (FLCatBoost) to make a focal-aware, cost-sensitive version of imbalanced customer churn prediction. Moreover, the empirical analysis is conducted in conjunction with the credit card customer dataset obtained from the Kaggle platform. The results of the staged comparison experiments show that the proposed method IADASYN-FLCatBoost in this paper shows the best prediction performance. Comparing the proposed method with 5 other imbalanced classification algorithms and 20 classifiers composed of classical sampling methods and ensemble learning algorithms, it is verified that the classification effect of the proposed method performs best, and the values of *Recall*, *F*1 score, *G-mean* and Area under Precision-Recall curve (*AUPRC*) have been significantly improved. In addition, further verification of the model also proves that the proposed method has certain generalizability and is still valid for other banks and customer churn datasets of other industries.

1. Introduction

With the rapid development of emerging technologies such as artificial intelligence, big data and cloud computing, the banking industry, as one of the most important industries in modern society, has also ushered in the *Bank 4.0* era of invisibility and intelligence (Wang, & Ji, 2018). Along with it, there is a fierce competition among major banks. Credit card business, as the main business module of banks, still attracts many consumers due to its high reliability, low interest rate and high credit limit service advantages in the context of smart finance. However, with the increase in the number of cards issued by major banks and the excessive homogenization of products and services, credit card customer churn has emerged as a significant and complex challenge. It is found that reducing the customer churn rate can improve the profitability of banks to some extent. As a result, more and more banking professionals are shifting from a "product-centric" marketing philosophy to a "customer-centric" one in an era of competitor emergence and business diversification (Larivière, & Van den Poel, 2004; Wu, Li, Zhao, & Liu, 2022). The banks are a data-oriented industry, and one of its challenges is how to build an effective prediction model of customer churn based on the historical behavioral data of these customers, so as to help bank managers identify the potential churners and formulate retention strategies in time (Qiu, 2020; Devriendt, Berrevoets, & Verbeke, 2021).

Researchers have realized the importance of customer maintenance since 1990s, and started to address the customer churn problem from the methodological perspective of statistics and machine learning, with

* Corresponding author.
E-mail addresses: cjrao@foxmail.com (C. Rao), ylingxu2020@163.com (Y. Xu), xpxiaowh@163.com (X. Xiao), hufywhut@163.com (F. Hu), bizgohkh@nus.edu.sg (M. Goh).

https://doi.org/10.1016/j.eswa.2024.123251
Received 26 June 2023; Received in revised form 23 December 2023; Accepted 14 January 2024
Available online 23 January 2024

single classification algorithms such as decision tree, logistic regression (Nie, Wei, Zhang, Tian, & Shi, 2011), support vector machine (SVM) (Farquad, Ravi, & Raju, 2014), artificial neural network (ANN) (Keramati, Jafari-Marandi, Aliannejadi, Ahmadian, Mozaffari, & Abbasi, 2014), random forest (RF) (Coussement, & Van den Poel, 2008; Wang, Rao, Goh, & Xiao, 2023), XGBoost (Srikanth, Papineni, Sridevi, Indira, Radhika, & Syed, 2022), CatBoost (Jain, Tomar, & Jana, 2021; Wei, Rao, Xiao, Chen, & Goh, 2023), and other ensemble learning algorithms, which have achieved good prediction results. Some studies have also confirmed that the latter classifies better than the former in the field of customer churn prediction (Tekouabou, Gherghina, Toulni, Mata, & Martins, 2022). CatBoost, as one of the latest algorithms in the GBDT (Gradient Boosting Decision Tree) ensemble learning framework, outperforms other advanced decision tree algorithms in both classification and regression puzzles. However, with the continuous research, scholars gradually realize that customer churn is a typical binary classification problem with nonlinear, high-dimensional and imbalanced data distribution. For example, in the identification of credit card churn customer, most customers will choose to stay and continue to enjoy the service, and only a very small number of people choose to leave, i.e., in the customer data set, churn customers only account for a small proportion, and the positive and negative sample ratios are extremely imbalanced (Xiao, Xie, He, & Jiang, 2012). In this sense, if the imbalanced sample set is directly used for training and learning, even though the overall accuracy of the model is high, the classification accuracy of the minority class samples is low, and the loss caused by misclassification of the minority class samples is much more serious than that brought by the majority class samples (Li, Jia, Wang, Qi, Jin, Chu, & Mu, 2022). Therefore, the study on the imbalanced customer classification problem has important theoretical and practical values.

The characteristics of the imbalanced customer churn classification problem are mainly reflected in two aspects: Imbalanced data distribution and limitations of traditional classification algorithms, and the most common solutions are from dual perspectives: data and algorithm (Chan, Kwong, & Jiang, 2021). At the data level, the class distribution is balanced by resampling methods such as oversampling, undersampling and hybrid sampling. That is, the number of samples in each category in the data set is roughly equal by sampling operation. At the algorithm level, since traditional classification models usually assume a balanced class distribution of the training set data, while in real customer classification problems, the class distribution is often imbalanced. Thus, most scholars approach the solution from the perspective of optimizing existing traditional classification algorithms or redesigning new machine learning algorithms, specifically two types of cost-sensitive learning and integration learning methods. The methods based on the data level exist independently of the classifier because they act on the data processing stage, which make the learning of imbalanced data more convenient. However, improving or designing new algorithms on the algorithm level makes it more universal in the application of classification models. What is more noteworthy is that both data and classifier will affect the final classification result of the model. Among them, there are two main data characteristics that affect the classification results: one is the overlap between classes, and the other is the noise samples. Overlapping between classes means that samples of different classes appear in the same area relatively densely, and noise samples in the data will result in extended training durations, hinder the learning of precise decision boundaries, and contribute to the complexity of the resultant classification model. If the selected samples are noise samples, the synthesized samples are also likely to be noise samples, which is not conducive to classification decision. Therefore, this study comprehensively considers the above two characteristics when sampling the data, and introduces the outlier detection method with density distribution, i.e., LOF algorithm, to eliminate outliers in the ADASYN oversampling method. In addition, combined with the characteristics of the credit card customer index system, the classified variables are specially treated when generating new samples. At the algorithm level, CatBoost algorithm, as the latest algorithm in Boosting integrated learning framework, performs better in classification performance, especially suitable for situations where there are many classification feature variables in the credit card customer index system. However, CatBoost algorithm also has some defects. To be specific, when it is applied to the imbalanced classification scene, the classification performance is greatly limited. Its default cross entropy loss function in this algorithm does not take into account the difference of losses between samples of various categories, that is, the misjudgment loss of the lost customer and the non-lost customer samples is set equal, which leads to the classifier's bias towards the non-lost samples in order to reduce the discriminant loss rate during training (Chen, Wang. Zhang, & Wang, 2021). Therefore, in this paper, the focus loss function proposed to solve the class imbalance problem in target detection is embedded into the CatBoost algorithm, and the default cross entropy loss function is replaced by the focus loss function, which transforms CatBoost into a focus-aware and cost-sensitive imbalanced customer churn prediction version (FLCatBoost). This approach makes the model more focused on lost customer samples and samples that are difficult to classify (Chan, Papaioannou, & Straub, 2023), thus enhancing the credibility of the model and reducing the model's missed judgment on lost customer samples. At the same time, the application of CatBoost and Focal Loss is further extended to the field of customer churn. Moreover, scholars have combined Focal Loss with XGBoost (Chen, & Qin, 2021), LightGBM (Mushava, & Murray, 2022) and other integrated classifiers to solve classification problems such as credit score and bankruptcy prediction. Therefore, considering both the perspectives of data and algorithm, this study proposes a new balanced transformation mechanism based on multi-strategy collaborative processing method to solve the classification problem of imbalanced credit card customer churn. The contributions of this study are summarized as follows:

(1) To solve the imbalanced customer churn classification problem, a new multi-strategy co-processing method IADASYN-FLCatBoost is proposed from dual perspectives: data and algorithm. At the data level, a new minority class sample is synthesized by the IADASYN oversampling method to make the sample class distribution balanced. At the algorithm level, the Focal Loss function is embedded into the CatBoost integration framework to convert CatBoost into a focus-aware, cost-sensitive version of imbalanced customer churn prediction.

(2) A new IADASYN algorithm is proposed by improving the traditional ADASYN algorithm. In this new algorithm, outliers in the training set samples are eliminated based on the LOF algorithm. In addition, considering that the direct calculation of the Euclidean distance measure in the ADASYN algorithm is not suitable for coping with the problem of both discrete and continuous variables in the customer churn index system, a special treatment is made for the categorical feature variables by using the median to punish the difference of classification features.

(3) To comprehensively evaluate the performance of IADASYN-FLCatBoost, five benchmark resampling methods (SMOTE, ADASYN, Borderline-SMOTE, SMOTETomek and IADASYN) and four benchmark ensemble classification models (XGBoost, LightGBM, CatBoost, and FLCatBoost) are compared under the evaluation indexes of *Recall*, *F*1 score, *G-mean*, *Kappa* coefficient and *AUPRC* value. In addition, the proposed method is compared with five popular algorithms (SMOTEBoost, RUSBoost, SMOTEBagging, OverBoost and SPE) in the field of imbalanced data classification. The comparison results show that the classification effect of IADASYN-FLCatBoost performs best, which demonstrates the feasibility of this new method in the field of customer churn.

The rest of the paper is organized as follows: Section 2 provides the literature review. Section 3 details the construction process of the new customer churn prediction method IADASYN-FLCatBoost. Section 4 conducts an empirical analysis according to the credit card customer dataset obtained from the Kaggle platform and introduces several corresponding evaluation metrics. The experimental parameter settings, experimental result analysis, algorithm performance evaluation and

model validity are provided in Section 5. Section 6 concludes the paper with some future research directions.

2. Literature review

Imbalanced category distribution is a typical feature of the customer churn classification problem. To address this problem, many scholars dealt with it at the data level or algorithm level (Liu, Fan, Xia, & Xia, 2022) (see Table 1). At the data level, the class distribution of the training set is balanced by resampling methods. For example, DoD, Huyn & Vo (2011) proposed the SMOTE (Synthetic Minority Oversampling Technique) oversampling technique to customer churn prediction, which eliminated the data imbalance and improved the prediction effect of the model by generating a few class samples. However, the SMOTE method does not differentially select the minority class samples in the process of generating new samples, and is prone to the problem of sample overlap, generating some samples that cannot provide effective information (Meng & Li, 2022). In this regard, Han, Wang & Mao (2005) proposed a Borderline-SMOTE algorithm, which improved the problem of sample overlap. However, this method only oversamples a few class samples that are in the boundary, which is easy to cause the problem of blurring the boundary of positive and negative classes. Thus, Pustokhina, Pustokhin, Nguyen, Elhoseny and Shankar (2021) correlated different instances in the actual training set with the sampling rate in order to achieve the maximum and optimal classification accuracy for minority class samples, and used an improved multi-objective rainfall optimization algorithm (MOROA) to determine the optimal sampling rate and parameter settings of the WELM for SMOTE. An improved Synthetic Minority Oversampling Technique (ISMOTE) was first used to process the imbalanced telecom customer dataset, and then the WELM model was applied to determine the labels of the classification (Pustokhina, Pustokhin, Nguyen, Elhoseny, & Shankar, 2023). The experimental simulation results showed that the ISMOTE-OWELM model outperformed the other models in terms of prediction accuracy on the applied datasets I, II, and III, with the results of 0.94, 0.92, and 0.909, respectively. There are also scholars who have designed under-sampling methods. For example, Sundarkumar and Ravi (2015) designed a hybrid undersampling method for the imbalanced characteristics of the customer churn data. However, undersampling methods often lead to information loss and are not applicable to datasets with high imbalance rate. Some scholars combined the advantages of oversampling and undersampling to balance the sample distribution through the integrated sampling method. For example, Batista, Prati and Monard (2014) used the SMOTE + ENN algorithm and SMOTE + Tomek algorithm for equalizing 13 UCI datasets respectively, and the result showed that the hybrid sampling algorithm could better predict the minority class samples and significantly improved the model accuracy. However, although the combined sampling method improve the problem of information loss

Table 1
Related research.

Author	Title	Method	Inadequacy/Conclusion
DoD, Huyn & Vo	Customer churn prediction in an Internet service provider	SMOTE	Does not differentially select the minority class samples and is prone to the problem of sample overlap, generating some samples that cannot provide effective information.
Han, Wang & Mao	Border-line-SMOTE: A new over-sampling method in imbalanced data sets learning	Borderline-SMOTE	Only oversamples the samples that are in the boundary cause the problem of blurring the boundary of positive and negative classes.
Pustokhina, Pustokhin, Nguyen, Elhoseny & Shankar	Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector	ISMOTE-OWELM	The performance of the ISMOTE-OWELM model over the compared methods. But the performance can be further enhanced using different feature selection methods.
Sundarkumar, & Ravi	A novel hybrid undersampling method for mining imbalanced datasets in banking and insurance	GMDH OCSVM	Undersampling methods often lead to information loss and are not applicable to datasets with high imbalance rate.
Batista, Prati, & Monard	A study of the behavior of several methods for balancing machine learning training data	SMOTE + ENN SMOTE + Tomek	The classification effect still needs to be improved for datasets with a serious imbalance between the proportion of positive and negative samples.
Wu, Zhao, Guo & Run	Prediction of Online Game User Turnover: Comparison and analysis of sampling methods based on unbalanced data	ADASYN	Verified the effectiveness of the method in the field of customer churn prediction.
Qiu	Research on the application of telecom customer churn prediction based on random forest	Random oversampling, SMOTE and ADASYN	ADASYN had the best performance in classification prediction than the first two methods.
Tang, Xie, Wang, Zhu, & Bie	Predicting typhoon-induced transmission line outages with coordination of static and dynamic data	The denoising adaptive synthetic (ADASYN) sampling algorithm	The original ADASYN algorithm tends to synthesize more minority samples for samples close to the majority class which may have a produce misleading effects.
Wong, Seng & Wong	Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain	CSDNN CSDE	Achieve better performance on the three evaluation metrics (*TPR*, *TNR* and *AUC)*, and the average rankings also showed that CSDE and CSDNN were ranked first and third.
Praveen, Manas, Jasroop & Pratyush	Customer churn prediction system: a machine learning approach	Decision Tree, Logistic Regression, Plain Bayes, AdaBoost, XGBoost	The AdaBoost and XGBoost model had the highest accuracy which were better than the other classifiers.
Swetha, & Dayananda	Improvised—XGBoost Machine learning Algorithm for Customer Churn Prediction	Improvized-XGBoost	The Improvised-XGBoost model had the best performance. XGBoost outperformed other state-of-the-art techniques in a variety of performance metrics.
Thammasiri, Hengpraprohm, Hengpraprohm & Mukviboonchai	Imbalance classification model for churn prediction	SMOTEBagging	Enhance the prediction of the classification model for minority class samples.
Kumar, Biswas & Devi	TLUSBoost algorithm: A boosting solution for class imbalance problem	TLUSBoost	The TLUSBoost was vertified on 16 benchmark data sets and outperformed the other models in terms of *Accuracy* and *F*1.
Pulicherla, Kumar & Abbaraju	Job Shifting Prediction and Analysis Using Machine Learning	CatBoost	When CatBoost is used to classify and predict unbalanced data with high feature dimension, the classification performance is greatly limited.
Xie, Li, Ngai & Ying	Customer churn prediction using improved balanced random forests	IBRF	Experimental results showed that the IBRF method can significantly improve the prediction accuracy compared with traditional classification models and has better prediction performance compared with balanced random forest and weighted random forest algorithms.

due to random undersampling, the classification effect still needs to be improved for datasets with a serious imbalance between the proportion of positive and negative samples. To address this problem, Wu, Zhao, Guo & Run (2016) proposed to equalize the data with ADASYN algorithm and applied it to the field of online gaming customers to verify the effectiveness of the method in the field of customer churn prediction. Qiu (2018) applied the three methods of random oversampling, SMOTE and ADASYN in the study of telecommunication customer churn prediction, and the result showed that ADASYN had the best performance in classification prediction than the first two methods. In order to improve the performance of sampling methods, many scholars have improved the traditional sampling methods. For example, Tang, Xie, Wang, Zhu and Bie (2022) considered that the original ADASYN algorithm tends to synthesize more minority samples for samples close to the majority class, and these samples may have a produce misleading effects, and they proposed a denoising adaptive synthetic sampling algorithm to generate minority class samples adaptively to balance the data set by purposefully selecting target samples. To verify the performance, the SMOTE and cost-sensitive learning methods were used for comparison, and the test results demonstrated that the denoising ADASYN algorithm can overcome the misleading effects of noisy samples and enhance the prediction model's ability to learn decision boundaries.

At the algorithm level, traditional classification models usually assume that the data class distribution of the training set is balanced, while the class distribution is often imbalanced in the actual customer classification problem (Wen, Wang, Ji, & Traore, 2022). Thus, most scholars mainly solve the classification problem of imbalanced data by improving the existing machine learning algorithms or proposing a new classification algorithm, which is mainly cost-sensitive learning and integrated learning methods. The core idea of cost-sensitive learning is to consider the cost difference between the samples of each category when they are misclassified. In the classification problem, by setting the misclassification cost of minority class samples higher than that of majority class samples, the classifier's preference for majority class samples is reduced (Rao, Liu, Goh, & Wen, 2020). For example, Wong, Seng and Wong (2020) proposed two new cost-sensitive methods for class imbalance problems such as churn prediction, fraud detection, etc., namely, cost-sensitive deep neural network (CSDNN) and cost-sensitive deep neural network integration (CSDE). In order to evaluate the performance of the proposed method with more than a dozen other existing methods (including AdaCost, MetaCost, RUSBoost, SMOTEBoost, AdaBoost, LogitBoost, BalanceCascade, Logistic Regression, Neural Network, Support Vector Machine, Decision Tree, etc.), six datasets from different industry domains for the experiments and the results show that CSDE and CSDNN achieve better performance on the three evaluation metrics (*TPR*, *TNR* and *AUC*), and the average rankings also showed that CSDE and CSDNN were ranked first and third, respectively, among all the methods.

Integration learning improves classification by integrating multiple classifiers together in a certain way to improve the generalization ability of the classifiers and to avoid the bias caused by individual classifiers in classifying predictions for imbalanced data. Praveen, Manas, Jasroop & Pratyush (2022) used different algorithms such as Decision Tree, Logistic Regression, Plain Bayes, AdaBoost, and XGBoost to build the telecom customer churn prediction models and evaluated the results on the test set using confusion matrix and Receiver Operating Characteristic (ROC) curve. The results showed that the AdaBoost and XGBoost integrated classifiers had the highest accuracy of 81.71 % and 80.8 %, respectively, which were better than the other classifiers and validated the effectiveness of the model. Further, Swetha and Dayananda (2020) proposed an Improvised-XGBoost model with feature function to predict customer churn in telecommunication industry in 2020. The model combines XGBoost to construct the feature function and then constructs the loss function by iterative method. The study showed that the Improvised-XGBoost model has the best performance in terms of accuracy. XGBoost outperformed other state-of-the-art techniques in a variety of performance metrics such as *Accuracy*, *Precison*, and *Recall*, and the model had better efficiency and could be used for complex datasets. For integrated learning, the research is mainly to combine the sampling technique with the integrated learning framework. The original imbalanced data is processed by data sampling technique, and then the samples are split into different structures to be trained by using the integrated learning classification algorithm (Thabtah, Hammoud, Kamalov, & Gonsalves, 2020). Thammasiri, Hengpraprohm, Hengpraprohm and Mukviboonchai (2018) combined the SMOTE oversampling technique with bagging algorithm and proposed a SMOTEBagging algorithm to enhance the prediction of the classification model for minority class samples. Kumar, Biswas and Devi (2019) proposed Tomek-link undersampling method based boosting algorithm (TLUSBoost), which firstly resampled the data using Tomek-link and redundancy based undersampling (TLRUS) and then boosting using AdaBoost technique. The TLUSBoost method was verified on 16 benchmark data sets and compared with unbalanced algorithms such as EasyEnsemble, BalanceCascade, SMOTEBoost and RUSBoost. The experimental results showed that the model outperformed the others in terms of *Accuracy* and *F*1 score. Pulicherla, Kumar & Abbaraju (2019) applied CatBoost algorithm to the study of employee turnover prediction in a company and compared it with algorithms such as Random Forest, XGBoost, AdaBoost, etc., and the results showed that CatBoost's prediction accuracy, coverage and hit rate were better than other algorithms. However, the algorithm also has some defects, that is, the default cross-entropy loss function in the algorithm does not take into account the difference in loss between churn customer samples and non-churn customer samples, which leads to the classifier being biased towards non-churn customers in order to reduce the discriminative loss rate during training. There are also scholars who combine the multiple methods in studying this problem. For example, Xie, Li, Ngai and Ying (2009) proposed a new improved balanced random forest (IBRF) method for bank customer churn prediction. The method improved the traditional random forest algorithm by combining sampling techniques and cost-sensitive learning to increase the penalty for misclassification of minority class samples while changing the class distribution. Experimental results showed that the IBRF method can significantly improve the prediction accuracy compared with traditional classification models such as decision tree and artificial neural networks, and has better prediction performance compared with balanced random forest and weighted random forest algorithms.

Although there are many studies that have made important contributions to solving the difficult problem of customer churn prediction, most of them only start from a single level of data or algorithm to improve the classification performance of algorithms, and cannot take into account the characteristics of the imbalance of data distribution structure and classification algorithms at the same time. Moreover, some studies also simply apply or combine the traditional sampling algorithms when resampling minority class samples, and do not fully take into account the problems of overlapping customer sample category regions, data noise, and different losses of minority class samples and hard-to-categorize samples. In addition, many techniques blindly synthesize new samples and are not able to combine the actual problems with reasonable sampling based on the characteristics of class distribution, which still has room for improvement in practical application. At the algorithm level, CatBoost, as one of the latest algorithms in the current boosting integrated learning framework, is especially suitable for the scenario of more categorized feature variables in the credit card customer index system. However, when CatBoost algorithm is applied to imbalanced classification scenarios, the default cross-entropy loss function does not take into account the difference in loss between the samples of each category, which leads to the classifier being biased toward the un-churn customers in order to reduce the discriminative loss rate during training. Combining the above analyses, this paper proposes a new multi-strategy collaborative processing method to solve the classification prediction problem of imbalanced credit card customer

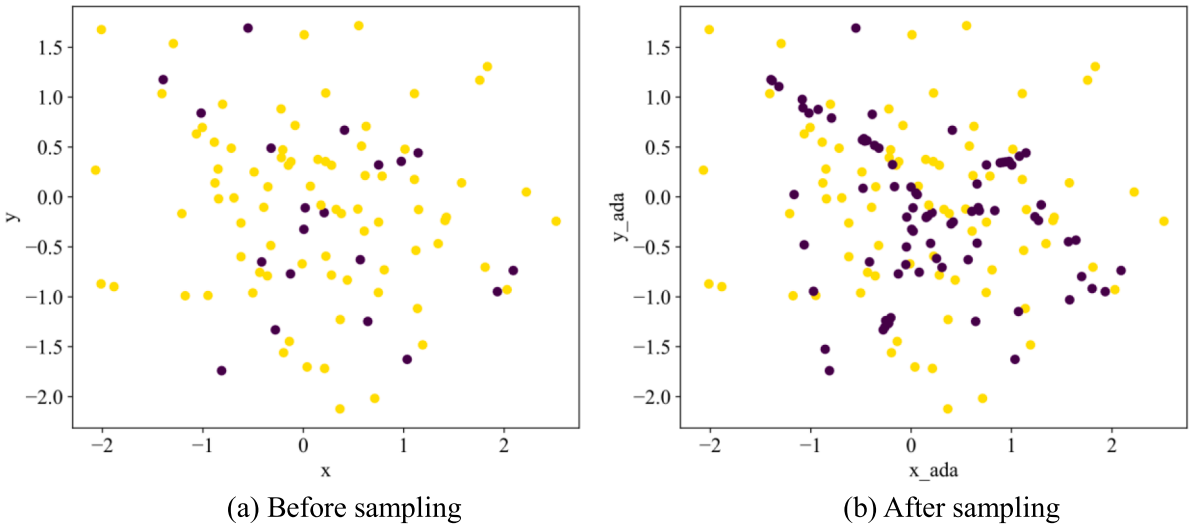


Fig. 1. Comparison of ADASYN before and after oversampling.

churn by combining the data level with the algorithm level, mainly optimizing and improving the original ADASYN sampling algorithm and CatBoost integration algorithm. Thus, the improved model can better deal with the imbalanced classification scenarios that occur most often in life, and provide new ideas for the churn customer prediction problem.

3. The model

In this section, a credit card customer churn prediction model (IADASYN-FLCatBoost) based on multi-strategy collaborative processing approach is constructed. Since the minority class samples are easily discarded as noise when the prediction indicator attributes are rich and the relationship between the indicator attributes is nonlinear, and the commonly used feature selection algorithms tend to select the feature subset that is favorable to the majority class samples, which leads to the poor classification effect of the classification algorithm on the minority class samples. Based on this background, this paper firstly selects the low-dimensional feature subsets that are highly correlated with customer churn and have minimal redundancy with each other based on the EN-XGBoost double-tier feature selection method to ensure the high accuracy and low complexity of the prediction model. Then, a multi-strategy co-processing method is used to solve the imbalanced customer churn classification problem and a new customer churn prediction model named IADASYN-FLCatBoost is constructed. At the data level, the ADASYN algorithm is improved from two aspects by using the obtained IADASYN sampling method combined with data distribution for data balancing operation. At the algorithm level, the Focal Loss function is embedded into the CatBoost ensemble learning algorithm framework to obtain an improved FLCatBoost algorithm, which effectively reduces the impact of data imbalance on classifier performance.

3.1. Methodology

3.1.1. ADASYN

The classical oversampling techniques are usually random sampling with put-back, which only balance the category distribution by simply replicating the minority class samples, and are highly susceptible to model overfitting. The SMOTE algorithm (Zhang, & Ding, 2023) is a representative algorithm in oversampling, and its basic idea is to synthesize new minority class samples randomly using a random function on the concatenation between the minority class samples and their K-nearest neighbor samples for the purpose of oversampling. The ADASYN algorithm, as a derivative of SMOTE, unlike SMOTE which simply copies existing samples, it assigns different weights to each minority class sample and uses some mechanism to automatically determine the number of samples to be synthesized for each minority class sample, making it more likely that minority class samples near the classification boundary will be used to synthesize new samples (Abraham, & Nair, 2018).

Fig. 1 gives the comparison of ADASYN algorithm before and after sampling. As can be seen from Fig. 1. (a) and Fig. 1. (b), the ADASYN algorithm adaptively interpolates different samples according to the distribution of non-churn customer samples around the churn customer samples, so as to expand the data set and balance the data distribution.

Suppose the set of majority class samples in data set D is $T_{\max}$, and the set of minority class samples is $T_{\min}$, then the specific steps of the ADASYN algorithm are as follows.

Step 1. Calculate the total number of lost customers to be synthesized according to Eq. (1).

$$G = (m_b - m_s) \times b \tag{1}$$

where m_b is the number of non-churners, m_s is the number of churners, and b is the number of customers in the interval $[0, 1]$. In this paper, b is set to 1 to make the data distribution balanced.

Step 2. For each customer sample in the set of churners $T_{\min}$, the K nearest neighbors are calculated based on the Euclidean distance, and the number of non-churn customers is denoted as $N_i (i = 1, 2, \cdots, m_s)$. Then, the percentage of non-churners in the K nearest neighbors of each sample is calculated as follows.

$$R_i = N_i / K \tag{2}$$

Step 3. Normalize the R_i obtained from Step 2.

$$\widehat{R}_i = R_i / \sum_{i=1}^{m_s} R_i \tag{3}$$

Step 4. Calculate the number of new samples to be synthesized for each lost customer by using the $\widehat{R}_i$ obtained from Step 3 as sample weights.

$$g = \widehat{R}_i \times G \tag{4}$$

Step 5. Based on the calculated number g to be synthesized for each churn customer sample, a number of customers are randomly selected from the K nearest neighbor samples of the churn customer sample X_i. A new sample S_{new} is generated by random linear interpolation between each selected customer sample X_t and the original customer sample. The interpolation is performed as shown in Eq. (5).

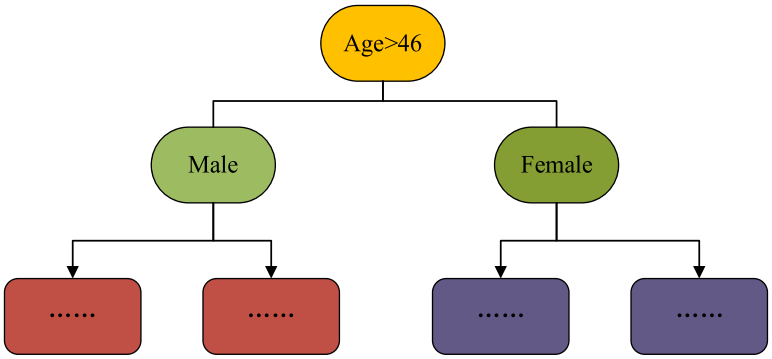


Fig. 2. Fully symmetric binary tree.

$$S_{new} = X_i + rand(0,1) \times (X_t - X_i) \tag{5}$$

where $rand(0,1)$ denotes the random in the interval $(0,1)$.

Step 6. The newly generated churn customer sample S_{new} is added to the credit card customer dataset D to obtain the new balanced dataset.

3.1.2. CatBoost

The traditional machine learning classification models are usually based on the assumption that the distribution among sample data is balanced and the misclassification cost is the same. Thus, these models tend to be biased toward the majority class when solving imbalanced classification problems, resulting in low prediction accuracy for samples from a few classes (Bentejac, Csorgo, & Martinez-Munoz, 2020). In contrast, ensemble learning is a classification integration framework that obtains multiple weak classifiers by dividing the original data set into multiple training subsets, and then weighting the prediction results of each weak classifier to form a high-performance strong classifier in order to obtain the final prediction results. Therefore, the classification effect of the integrated model is significantly better than that of a single classifier in both classification and regression problems. Bagging and Boosting are the two most common ensemble learning frameworks in practical applications. The CatBoost classification algorithm belongs to the gradient boosting decision tree class of machine learning algorithms, and it is also one of the latest boosting ensemble learning frameworks algorithm (Jelen, Podobnik, & Babic, 2021; Zhang, Rao, Xiao, Hu, & Goh, 2024). Compared with earlier ensemble learning algorithms, CatBoost performs better in terms of speed and accuracy, and is able to reduce gradient bias and solve the prediction bias problem by efficiently processing category-based features and the ordered boosting algorithm, thus increasing the reliability and generalization ability of the model (Hussain, Mustafa, Jumani, Baloch, Alotaibi, Khan, & Khan, 2021).

CatBoost uses a fully symmetric binary tree as the base learner. As can be seen from Fig. 2, in this fully symmetric binary tree, the first node is split by whether the age is greater than 46 years old, with the left node being older than 46 years old and the right node being younger than 46 years old. At this point, suppose there are two leaf nodes A and B, both of which continue to search for the node with the greatest splitting gain among all features. Suppose A finds an optimal feature, such as "gender" for male, the gain value of the optimal splitting point is x_1, and B finds an optimal feature, such as "education" for college, the gain value of the optimal splitting point is x_2. Then x_1 and x_2 are compared, and for both A and B, both nodes are split using the best node with the optimal feature corresponding to the largest split gain value. When modeling, CatBoost first binarizes all features and encodes the index of each leaf as a binary vector with length equal to the tree depth, and the evaluation results are calculated by using the binary features in prediction. In addition, it converts the gradient boosting algorithm in traditional GBDT to Ordered Boosting algorithm (Bileki, Barboza, Silva, & Bonato, 2022): Train a separate integrated model M_i for each sample x_i. The M_i obtained by training the sample dataset except x_i. Then use the model M_i to estimate the gradient on x_i and use this estimate to score the final obtained tree model, which not only obtains an unbiased estimate of the gradient, but also overcomes the problem of prediction bias and improves the accuracy of fitting to the data distribution. In addition, the model can also show good adaptive ability when the customer distribution changes.

3.2. The multi-strategy collaborative processing approach

3.2.1. Data level: Oversampling based on IADASYN algorithm

When considering the sample distribution to synthesize new samples, the outlier points, whether minority samples or majority samples, are far away from normal data points, and synthesizing samples near them may not only be detrimental to improving the classification performance of the model, but also affect the true distribution of the data and the complexity of synthesizing samples. Therefore, before oversampling the churn customer samples in the training set, we first introduce the Local Outlier Factor (LOF) algorithm into ADASYN to eliminate the outliers. Then, considering that the ADASYN algorithm needs to calculate the Euclidean distance between samples when generating new samples, and it is not suitable to deal with the situation that there are both discrete and continuous variables in the churn indicator system of this paper, we specially deal with the classification feature variables.

(1) Removing outliers based on LOF algorithm

The LOF algorithm is an outlier detection method based on the density distribution (Wang, & Chen, 2021), which determines whether the samples are outliers by the relative distance between them and the local density, and is suitable for data with very different density dispersions between the samples. The schematic diagram of the LOF algorithm is shown in Fig. 3.

For a sample $X_i(i = 1, 2, \cdots, n)$ in the credit card customer data set D, firstly, find its K nearest neighbor samples $\{X_1, X_2, \cdots, X_k\}$, and calculate the k-th distance $d_k(X_i)$. Then, the maximum of the k-th distance of point X_i and the actual distance from it to point $X_j(j = 1, 2, \cdots, k)$ is taken as the reachable distance of X_i.

$$d_k(X_i, X_j) = \max(d_k(X_i), \|X_i - X_j\|) \tag{6}$$

Secondly, the local reachable density $\rho_k(X_i)$ of point X_i is calculated according to Eq. (9), and the corresponding local reachable density $\rho_k(X_j)$ is calculated for each nearest neighbor sample $\{X_1, X_2, \cdots, X_k\}$.

$$\rho_k(X_i) = \frac{K}{\Sigma_{k=1}^{K} d_k(X_i, X_j)} \tag{7}$$

Finally, the LOF of point X_i is expressed as

$$LOF_k(X_i) = \frac{1}{K} \sum_{k=1}^{K} \frac{\rho_k(X_j)}{\rho_k(X_i)} \tag{8}$$

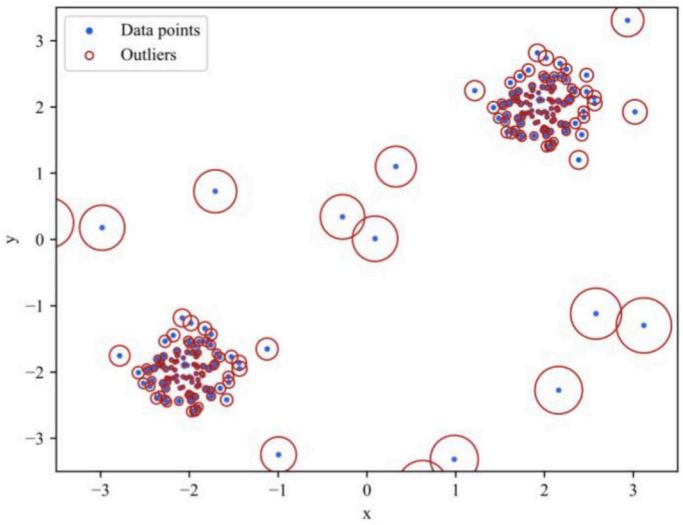


Fig. 3. Schematic diagram of LOF algorithm.

where $LOF_k(X_i)$ represents the average of the ratio of the locally reachable density of all data points in the k-th distance neighborhood of point X_i to the locally reachable density of point X_j. If point X_i is in the same cluster with point X_j, $\rho_k(X_i)$ should be close to $\rho_k(X_j)$, and $LOF_k(X_i)$ is approximately equal to 1, i.e., the density of point X_i and its domain points are comparable, indicating that X_i is a normal point. While if X_i is an outlier, $LOF_k(X_i)$ should be much larger than 1. This paper uses this LOF algorithm to output n data points with the largest degree of outliers in the credit card customer training set, laying the foundation for the next step of data oversampling.

The LOF algorithm considers both local and global attributes of data sets, and determines outliers through the density of neighborhood points, which is suitable for customer data sets with unknown data distribution in this paper. Fig. 4 shows the abnormal situation of the algorithm in the data set with the assumption of $K = 3$.

(2) Special treatment on classification features

Since there are a large number of categorical feature variables in the customer churn indicator system, the traditional ADASYN algorithm for generating new samples is not suitable for the situation when there are both discrete and continuous variables in the customer churn indicator system. Therefore, this paper draws on the advantages of the SMOTENC algorithm (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to perform a special treatment of the categorical features in the credit card customer dataset.

First, median calculation is performed, i.e., the median of the standard deviation of all their continuous feature variables is calculated for the sample of churn customers. The size of the penalty is tightly correlated with the differences between the continuous feature values by penalizing the differences in the categorical features using the median. Then, a nearest neighbor calculation is performed, where the Euclidean distance between the churn customer samples for which the nearest neighbor is being identified and the other churn samples is calculated using the continuous feature variables. For each different nominal feature between the considered feature vector and its potential nearest neighbors, the median of the previously calculated standard deviations is included in the Euclidean distance calculation. That is, if a sample differs from its nearest neighbor in the value of a categorical feature, the median of the previously calculated standard deviation is included in the Euclidean distance calculation. The median is included as many times as there are different categorical features. Finally, when synthesizing a new sample of lost customers, for the continuous feature variables, the same steps as described previously for the ADASYN algorithm are used to obtain them. In contrast, for the categorical features, the value with the highest number of occurrences in the K nearest neighbors is assigned.

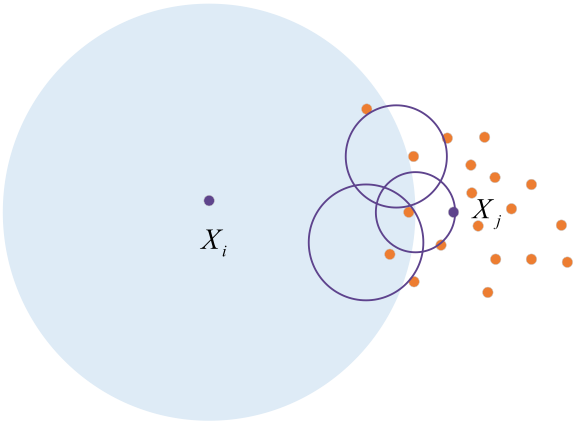


Fig. 4. The abnormal situation at $K = 3$.

Therefore, the specific steps of the IADASYN algorithm obtained from the above two improvements are as follows. The flow of IADASYN sampling algorithm is shown in Fig. 5.

Step 1. Calculate the local reachable density for each data point in the credit card customer dataset according to Eq. (9).

Step 2. Combine the local reachable density calculated in Step 1, and calculate the local outlier factor for each data point by Eq. (10).

Step 3. Determine and eliminate the corresponding outliers according to the judgment conditions in the LOF algorithm.

Step 4. Calculate the total number of churn customer samples to be synthesized.

Step 5. Calculate the median of the standard deviation of all the continuous characteristic variables for each churn customer sample.

Step 6. Calculate the Euclidean distance between the churn samples for which a nearest neighbor is being identified and the other churn samples using the continuous feature variables.

Step 7. Determine whether the sample has the same categorical feature value as its K nearest neighbors and penalize it using the median standard deviation obtained in Step 5, and determine how many categorical features are different and how many times they are included.

Step 8. For each sample in the set of attrition samples, derive its nearest neighbors according to the Euclidean distance formula.

Step 9. Calculate the percentage of non-churn samples in the K

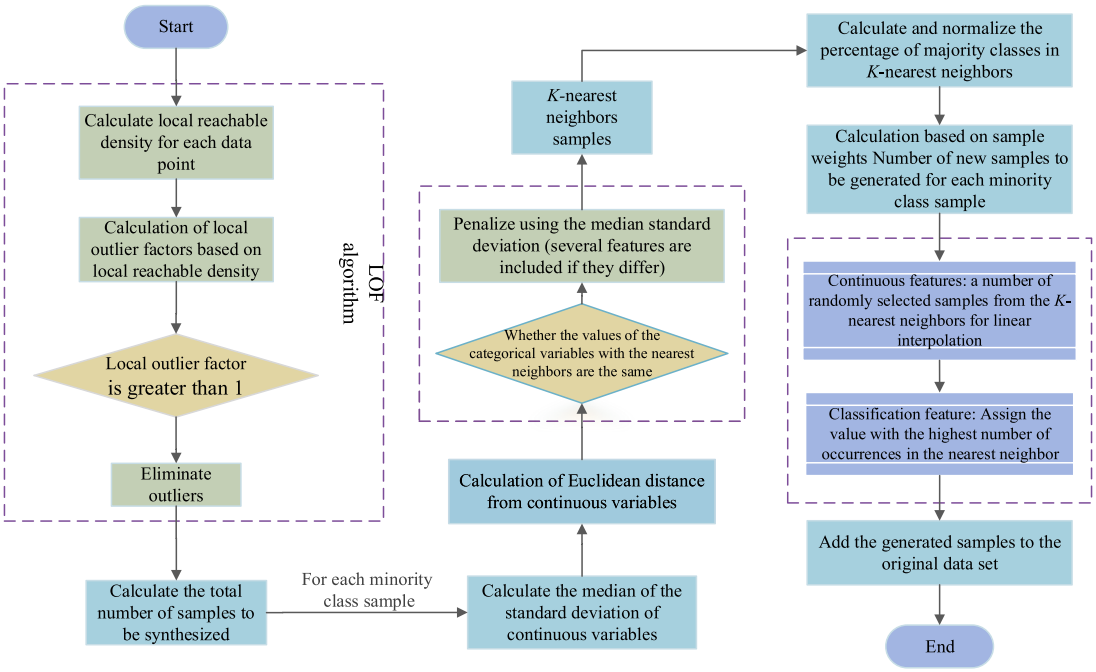


Fig. 5. The flow of IADASYN sampling algorithm.

nearest neighbors and normalize them.

Step 10. Calculate the number of new samples to be generated for each churn sample based on the sample weights obtained from Step 9.

Step 11. For the continuous features, a number of randomly selected samples from the nearest neighbors are linearly interpolated to create new sample features. For the categorical features, they are assigned the value that occurs most frequently in the potential nearest neighbors of the few samples.

Step 12. Add the churn customer samples obtained from Step 11 to the original training dataset, so that the samples of each category in the dataset are balanced.

3.2.2. *Algorithm level: Improve CatBoost based on Focal Loss function*

The loss function describes the difference between the predicted and true labeled values. The default loss function used in the CatBoost algorithm in the binary classification problem is the Cross Entropy loss (CE), as shown in Eq. (9) (Wang, Deng, & Wang, 2020). It can be seen that for positive class samples, i.e., churn customers, the higher the output probability, the smaller the loss. For negative class samples, i.e., non-churn customers, the smaller the output probability, the smaller the loss. Therefore, the essence of the CE loss function is to penalize misclassification. In the field of customer churn prediction, when the distribution of customer samples is imbalanced, i.e., the number of non-churn customers is much larger than the number of churn customers, the non-churn samples in the loss function will dominate, making the model poorly classified.

$$CE = -y\log(1-y') - (1-y)\log(1-y') = \begin{cases} -\log y', y = 1 \\ -\log(1-y'), y = 0 \end{cases} \tag{9}$$

where y denotes the label of the real sample and y' denotes the predicted probability value.

FLCatBoost is based on CatBoost, and introduces the Focal Loss (FL) function as its loss function to make the customer churn prediction model focus more on minority class samples and difficult to classify samples, improve the sample class imbalance problem at the algorithm level, and increase the model accuracy. FL is a loss function proposed by Lin, Goyal, Girshick, He, and Dollar (2020) by improving the CE loss function in order to solve the problem of severe imbalance in data distribution in image detection, as shown in Eq. (10). This loss function adjusts the loss weights of samples in different categories by introducing category weighting factors, and balances the importance of positive and negative samples by introducing dynamic modulation factors to increase the weights of positive and hard-to-classify samples, making the model pay particular attention to churn customers and hard-to-classify customer samples during the training process.

$$FL = \begin{cases} -\alpha(1-y')^{\gamma}\log(y'), y = 1 \\ -[(1-\alpha)y']^{\gamma}\log(1-y'), y = 0 \end{cases} \tag{10}$$

where α denotes the category weight factor, and γ denotes the dynamic modulation factor. For the churn customer samples, the closer the prediction result of the classification algorithm is to 1, the easier the samples can be distinguished, and the corresponding modulation factor $(1-y')^{\gamma}$ will be small, which reduces the misclassification loss and makes the algorithm pay more attention to the hard-to-classify samples during training. Since the experiments that the best result is obtained when the value of $\alpha = 0.25$ and $\gamma = 2$, this paper also follows their settings.

3.3. *The IADASYN-FLCatBoost model*

As when the predicted index attributes are abundant and the index attributes are nonlinear, the minority samples are easily regarded as noise and discarded, while common feature selection algorithms tend to choose feature subsets that are beneficial to majority samples, which leads to poor classification effect of classification algorithms on minority samples. Therefore, this paper first selects a subset of low-dimensional features that are highly related to customer churn and have the least redundancy based on the EN-XGBoost index selection method to ensure the high accuracy and low complexity of the early warning model. Then new samples are generated by IADASYN method, and added to the original training set to balance the data set. Then, the improved FLCatBoost classification algorithm is used to train the new training set according to the selected variable characteristics, and finally a customer churn early warning model based on multi-strategy collaborative processing method is proposed.

3.3.1. EN-XGBoost double-layer customer churn indicator selection

(1) First-tier based on Elastic Network (EN) regression

Linear regression analysis models, such as least square estimation, ridge regression and lasso regression, have been widely used in the fields of economics, management, and healthcare (Amini, & Hu, 2021). When using regression models for feature selection, lasso regression tends to select one variable at random if the feature variables are correlated with each other, while ridge regression selects all the features. Considering the limitations of the above methods, some scholars combined the advantages of ridge regression and lasso regression and proposed a regression model for elastic network with L1 and L2 paradigms as prior regular terms in 2005 (Cui, Bai, Wang, Jin, & Hancock, 2021). This model can automatically perform feature screening while continuously shrinking the coefficients, which not only removes invalid features as in lasso regression, but also continues to uphold the stability of ridge regression (Liu, Yu, Chen, Han, & Yu, 2020; Xiao, Gao, Chen, & Jiang, 2023; Rao, Huang, Chen, Goh, & Hu, 2023). Therefore, in this paper, the regression model for elastic network is chosen for the first level of indicator selection for customer churn predictors, which can maintain a high accuracy even if there is multicollinearity among variables.

Suppose the total number of samples in the credit card customer dataset is n, the number of feature variables is p, the response variable (whether the customer churns or not) is $y_i(i = 1, 2, \cdots, n)$, the predictor variable is $x_{ik}(k = 1, 2, \cdots, p)$, i.e., the customer's personal information, and the response and predictor variables are centralized and normalized, respectively. Then the cost function of the algorithm is expressed as follows:

$$Cost(w) = \underset{w}{\operatorname{argmin}}\left(\sum_{i=1}^{N}(y_i - w^T x_i)^2 + \lambda \times \left[\rho\|w\|_1 + \frac{(1-\rho)}{2}\|w\|_2^2\right]\right) \quad (11)$$

where $w = (w_0, w_1, \cdots, w_p)^T$ is the regression coefficient matrix, λ is the complex parameter that controls the degree of compression, $\rho(0 \leqslant \rho \leqslant 1)$ is the elastic network mixing parameter, $\|w\|_1$ is the L1 regularization parameter of the lasso regression, and $\|w\|_2^2$ is the L2 regularization parameter of the ridge regression. Therefore, the elastic network regression controls the size of the penalty term through the sum of two parameters, that is, the size of the regression coefficient of each predictive variable to minimize the cost function. The value range of ρ is 0–1, which means adjusting the combination of L1 and L2. If ρ is close to 1, it means that L1 is dominant; If ρ is close to 0, it means that L2 is dominant. That is, when $\rho = 0$, the elastic network is equal to the ridge regression; When $\rho = 1$, it is equal to the lasso regression.

(2) Second-tier based on XGBoost algorithm

XGBoost algorithm is an advanced integrated version of the decision tree-based method proposed by T. Q. Chen in 2014, which has received high attention and achieved good performance as a feature selection method in academia (Chen, Zhang, Yu, Yu, Lawrence, Ma, & Zhang, 2020). The algorithm uses the frequency of occurrence of variables to classify the importance of features when segmenting data in all classification trees, i.e., it calculates which feature is selected as a segmentation point based on the gain value of the structure score. Thus, the importance score of a feature is equal to the sum of its occurrences in all decision trees (Cao, Xie, Shi, & Wang, 2022). It judges the importance of a feature based on the final output feature importance score. The higher the score, the more important the corresponding feature is. The gain score for each split of the decision tree is calculated as follows.

$$Gain = \frac{1}{2} \times \left[\frac{(\sum_{\in I_L} g_i)^2}{\sum_{\in I_L} h_i + \lambda} + \frac{(\sum_{\in I_R} g_i)^2}{\sum_{\in I_R} h_i + \lambda} + \frac{(\sum_{\in I} g_i)^2}{\sum_{\in I} h_i + \lambda}\right] - \gamma \quad (12)$$

where g_i and h_i denote the first-order gradient and second-order gradient of the logarithmic loss function used by XGBoost algorithm in the classification problem, respectively; $I = I_L \cup I_R$, I_L and I_R are the number of samples in the separated left and right node spaces, respectively; λ and γ are the regularization parameters.

3.3.2. The modeling steps of the customer churn early warning model

In this paper, we propose a model based on multi-strategy co-processing method. The modeling idea is as follows: use EN-XGBoost algorithm to filter features, generate new samples by IADASYN method and add them to the original training dataset to balance the data set, use the improved FLCatBoost classification algorithm to train on the new training set based on the filtered variable features, and then the customer churn prediction model based on multi-strategy collaborative processing method is established. The modeling process is shown in Fig. 6, and the specific implementation process is as follows.

Step 1. Input the data set, and make data preprocess including missing value processing and normalization processing.

Step 2. Obtain a subset of key features that affect customer churn through the EN-XGBoost double-tier feature selection method to form the new dataset.

Step 3. Apply the IADASYN algorithm on the training dataset to synthesize new minority samples and add them to the original dataset to form the new dataset.

Step 4. Train the new dataset using the FLCatBoost classification algorithm to obtain the final classification model.

Step 5. Test the model with the test set data to verify the classification effect.

4. Empirical analysis

This section conducts an empirical analysis based on a credit card customer dataset[1] obtained from the data platform Kaggle. Since it is inconvenient to use the original dataset directly for modeling analysis, the data are firstly pre-processed, including operations such as removing redundant features, missing value processing, standardization processing and numeralization of categorical variables. Secondly, the EN-XGBoost double-tier indicator selection algorithm is used to reduce the feature dimensionality to ensure high accuracy and low complexity of the prediction model. Thirdly, the IADASYN oversampling algorithm is used to balance the data so that the distribution between churn and non-churn customers is balanced. Finally, the processed training data set is input in FLCatBoost algorithm to establish a credit card customer churn prediction model based on multi-strategy collaborative processing method.

4.1. Data source and processing

The dataset used in this paper is the historical consumption record of credit card customers of a foreign bank from 2015 to 2017, with a total of 10,127 sample information. Among them, each sample involves 21 characteristic variables, including customer demographic information, accounting information, behavior information and the target variable, as shown in Table 2. Moreover, the target variable "Attrition_Flag" is presented in the form of binary classification. If the customer has churn, it is marked as "Attributed Customer", otherwise, it will be marked as "Existing Customer". Among them, there are 1,627 credit card customers who have churned, and 8,500 customers who have remained with the service.

The characteristic variable "CLIENTNUM" is only to distinguish different customers and is not related to the target variable, so we delete this redundant variable. Since there are missing values in this dataset, and the missing percentages of Education_Level, Income_Category, and Marital_Status are 15 %, 10.98 %, and 7.4 %, respectively, we use the

[1] https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers.

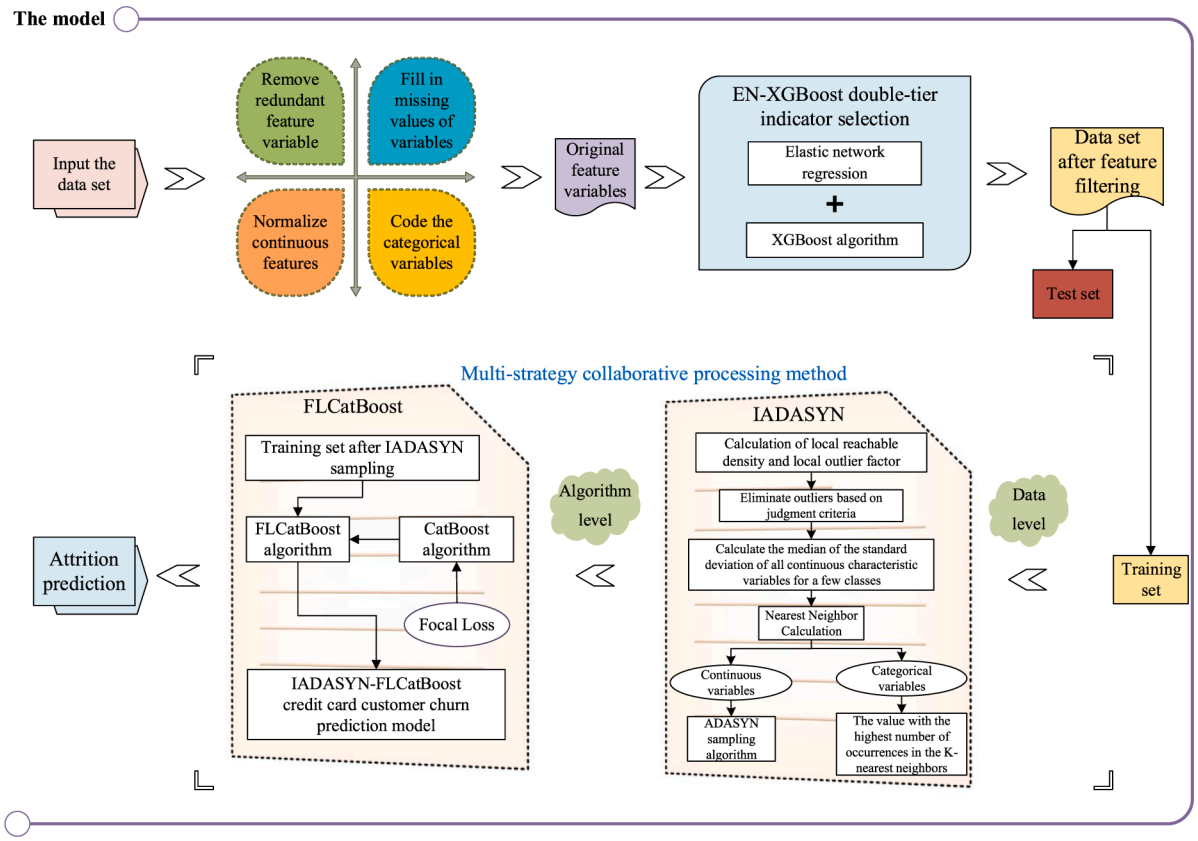


Fig. 6. The flow chart of IADASYN-FLCatBoost model.

random forest regression algorithm to fill the missing data. Then, in order to make the data comparable with each other, the continuous variables are standardized by Z-score to eliminate the difference in magnitude between variables. Finally, the classification variables such as Gender and Marital_Status are coded to facilitate the establishment of the subsequent model.

4.2. Indicator selection

For the remaining 19 predictor variables affecting credit card customer churn, the EN-XGBoost double-tier indicator selection algorithm is used to reduce the dimensionality. In this paper, the data set is divided into a training set and a test set according to the ratio of 7:3. First, the random grid search method is used to determine the optimal combination of the parameters of λ and ρ in the elastic network algorithm. The search range of λ is set to [0.1, 1.0], the step size is set to 0.1, the search range of ρ is set to [0.005, 0.3], and the step size is set to 0.001. Then, we use mean squared error (MSE) as the evaluation criterion and the ten-fold cross-validation is performed to obtain the optimal combination of parameters: $\lambda = 0.3$, $\rho = 0.008$, and the optimal prediction score (MSE): 0.0847. The regression model of the elastic network is established according to the optimal combination of parameters, and the results of the regression coefficients of each predictor variable are obtained as shown in Table 3. Therefore, two variables "Avg_Open_To_Buy" and "Avg_Util_Ratio" with regression coefficient of 0 are eliminated, and the first layer feature subset is obtained.

Then, the first layer feature subset is continued into the XGBoost algorithm with default parameter settings, and the importance scores of the variables obtained are shown in Fig. 7. From Fig. 7, we can see that Total_Trans_Amt, Customer_Age and Total_Amt_Chng have a very significant impact on credit card customer churn prediction, while other factors such as Card_Category and Gender have less impact on customer churn. Therefore, this paper takes 500 as the cut-off point, and again removes the three features of Card_Category, Gender and Marital_Status that are ranked low, and finally keeps 14 most key features to construct the customer churn indicator system.

4.3. Evaluation indexes

Upon completing the model's training, it's imperative to assess its performance. Continuous adjustments based on the evaluation outcomes are crucial to ensure the achievement of satisfactory results. Accuracy is the simplest and most commonly used index to evaluate a model, but using accuracy as an evaluation index without any premise often cannot reflect the performance of a model. For example, in an imbalanced data set, the number of positive samples accounts for 5 % of the total number of samples, and the number of negative samples accounts for 95 % of the total number of samples. Then there is a model that judges all samples as negative, and the model can also achieve 95 % accuracy, but this model has no predictive ability. Because for unknown samples, it always judges these samples as negative classes. Therefore, for a model, we need to choose appropriate evaluation indexes to judge its performance. Nevertheless, when contrasting various models, using different evaluation indexes frequently results in different evaluation results. This indicates that the quality of the model depends not only on the algorithm and data, but also on the specific business objectives and requirements. For instance, in a model designed for identifying lost customers, the primary aim is to minimize customer attrition rates. Therefore, the model's objective is to maximize the detection of all potential lost customers, even at the risk of a higher rate of misclassification.

In the imbalanced classification problem, confusion matrix (Pang, Peng, Chen, Yang, & Zhang, 2019) can reflect the performance of the model intuitively, which is not only common in binary classification problems, but also applicable to multi-classification problems. It is a practical visualization method, as shown in Table 4. We suppose that 1 represents the positive category (the lost customers), 0 represents the

Table 2
Description of characteristic variables and their meanings.

Variable	Explanation	Value
CLIENTNUM	Customer's ID	[708082083, 828343083]
Customer_Age	Customer's age	[26, 73]
Gender	Customer's gender	Male, Female
Dependent_count	Number of family members	[0,5]
Education_Level	Educational background	Doctorate, Post-Graduate, Graduate, College, High School, Uneducated
Marital_Status	Marital status	Single, Married, Divorced
Income_Category	Income category	Less than 40 K, 40 K-60 K, 60 K-80 K, 80 K-120 K, 120 K+
Card_Category	Card category	Blue, Gold, Silver, Platinum
Months_on_book	Number of months booked	[13, 56]
Total_Relation_Count	Count of total relation	[1, 6]
Months_Inactive	Count of inactive months	[0, 6]
Contacts_Count	Count of contacts	[0, 6]
Credit_Limit	Current credit limit	[1438.3, 34516]
Total_Revolving_Bal	Total turnover balance	[0, 2517]
Avg_Open_To_Buy	Average available credit limit	[3, 34516]
Total_Amt_Chng	Change rate of the transaction amount	[0, 3.397]
Total_Trans_Amt	Transaction amount	[510, 18484]
Total_Trans_Ct	Total transaction counts	[10, 139]
Total_Ct_Chng	Change rate of the transaction counts	[0, 3.714]
Avg_Util_Ratio	Average utilization rate of credit cards	[0, 0.999]
Attrition_Flag	Attrition flag	Attrited Customer, Existing Customer

Table 3
Results of regression coefficients.

Variable	Coefficient	Variable	Coefficient
Customer_Age	0.0481	Credit_Limit	−0.0181
Gender	0.1463	Months_Inactive	0.0668
Dependent_count	0.0264	Total_Ct_Chng	−0.4352
Education_Level	0.0163	Total_Revolving_Bal	−0.1958
Marital_Status	0.0538	Avg_Open_To_Buy	0
Contacts_Count	0.0718	Total_Amt_Chng	0.0225
Income_Category	0.0404	Total_Trans_Amt	0.3742
Card_Category	0.0257	Total_Trans_Ct	−0.8417
Months_on_book	0.0905	Avg_Util_Ratio	0
Total_Relation_Count	−0.0263		

negative category (the retained customers), *TP* represents the number of lost customers judged by the model as lost, *FN* represents the number of lost customers judged by the model to be retained, *TN* represents the number of retained customers judged as retained by the model to be retained, *TN* represents the number of retained customers judged as retained by the model, and *FP* represents the number of lost customers judged by the model to be retained. In this matrix, our primary focus is the number of "*TP*" quadrant, as it signifies the number of users the model successfully identifies as at risk of churning. In order to measure the performance of the established customer churn early warning model more intuitively, this paper introduces *Recall*, *Precision*, *F*1 score, *G-mean*, *Kappa* coefficient and *AUPRC* value on the basis of confusion matrix. The evaluation indexes selected for this analysis range from 0 to 1, with higher values denoting superior classification performance of the models.

(1) *Precision, Recall* and *Specificity*

Precision refers to the correct proportion in all the results predicted by the model as customer churn, which reflects the accuracy of the positive sample predicted by the model. However, this index cannot reflect how many positive samples are wrongly predicted as negative samples (*FN*).

Recall is also called True Positive Rate (*TPR*) and *Sensitivity*, among which *Sensitivity* is commonly used in the medical field. *Recall* represents the proportion of customers who are correctly identified, and reflects the comprehensiveness of the model to predict positive samples. Contrary to *Precision*, *Recall* can't reflect how many negative samples are wrongly predicted as positive samples (*FP*). The higher the *Recall*, the better the model's ability to identify.

Specificity refers to the correct proportion predicted by the model among all customers who have not churn. If a model judges all samples as positive categories, $Recall = 0$ at this time, but *Specificity* is very low, which is also unreasonable. Although there is no case of missing judgment, it greatly increases the burden on bank staffs and the trouble of customers who have no intention of churning.

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

(2) *F*1 score and *G-mean*

*F*1 score combines the common characteristics of *Precision* and *Recall*, which can reflect the evaluation effect of both *Precision* and *Recall*. Only when the *Precision* and *Recall* are high can the classification model get a higher *F*1 score, and the calculation formula is as follows.

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{16}$$

In the field of customer churn prediction, *Specificity* and *Recall* need to be considered at the same time. *G-mean* represents the geometric average of the two, which can measure the average performance of the classification model on the two groups of customers who have lost and those who have not churned. *G-mean* is expressed as

$$G\text{ - }mean = \sqrt{Specificity \times Recall} \tag{17}$$

(3) *Kappa*

Kappa coefficient is a comprehensive measure of classification accuracy.

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{18}$$

where p_o is the sum of the number of correctly classified samples in each category divided by the total number of samples, i.e., the overall classification accuracy, see Eq. (19). The formula for calculating p_e is given in Eq. (20), and n is the total number of samples.

$$p_o = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

$$p_e = \frac{(TP + FP) \times (FN + TP) + (FN + TN) \times (TN + FP)}{n \times n} \tag{20}$$

(4) PR curve and *AUPRC* value

Taking the False Positive Rate (*FPR*) as the abscissa and the True Positive Rate (*TPR*) as the ordinate, the ROC (Receiver operating characteristic) curve can be drawn after the output results are truncated by different thresholds to measure the performance of the classification model on test sets with different decision thresholds. And *AUC* (Area under curve) value is the area covered by ROC curve. As can be seen from Eqns. (21) and (22), the calculation of *FPR* is only related to the customer samples that are really negative, and the size of *FPR* is only related to the samples that are really positive. Therefore, when the imbalance ratio in the data set changes greatly, the ROC curve will still not change greatly. The abscissa of PR (*Precision-Recall*) curve is the recall rate and the ordinate is the accuracy rate, which is generated by moving the threshold from high to low. Because the numerator of the accuracy rate is *TP* and the denominator contains *FP*, the PR curve is

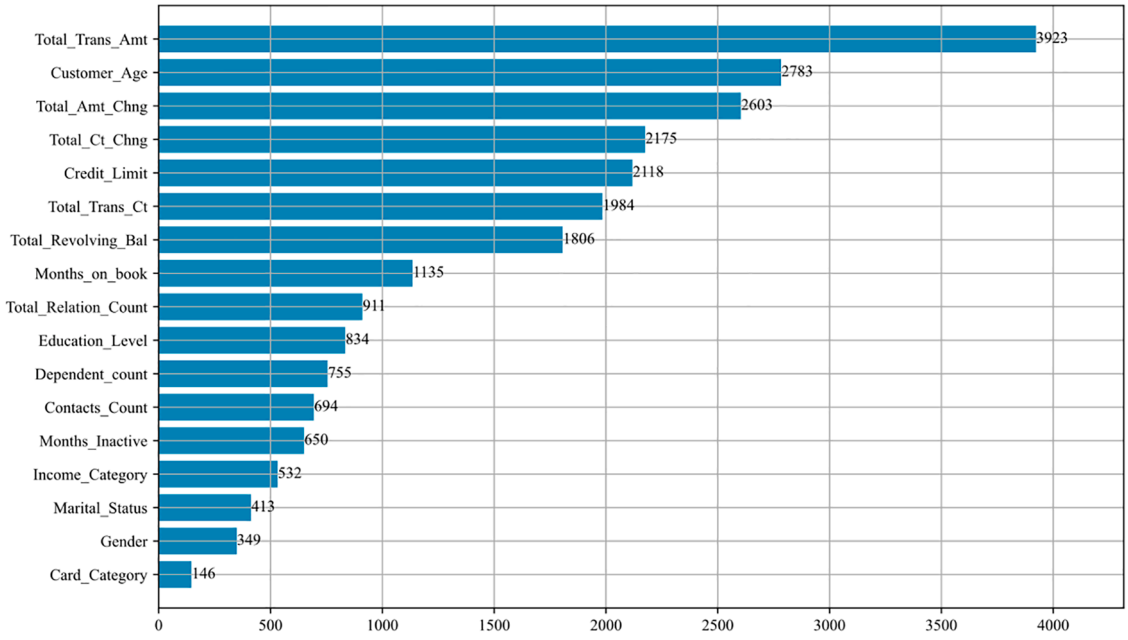


Fig. 7. XGBoost importance score.

Table 4
Confusion matrix.

Predicted value True Value	Positive category	Negative category
Positive category	*TP*	*FN*
Negative category	*FP*	*TN*

easily influenced by the proportion of positive and negative samples and is very sensitive to the imbalanced proportion, which is more suitable for evaluating the classification performance of the model in imbalanced customer data sets. *AUPRC* (Area under *Precision-Recall* curve) is the area value under PR curve, and this index is selected as the index to measure the classifier, which is helpful to compare the performance of each model.

$$FPR = \frac{FP}{FP + TN} \tag{21}$$

$$TPR = \frac{TP}{TP + FN} \tag{22}$$

5. Experimental results and analysis

To verify the effectiveness and applicability of the multi-strategy collaborative processing method IADASYN-FLCatBoost proposed in this paper, the experimental results are analyzed in this section in conjunction with the pre-processed dataset. Firstly, the staged experimental comparison of the improved method is carried out to verify the feasibility of the two-stage improvement at the data and algorithm level. Secondly, the proposed model is compared with the combinations of different resampling algorithms (SMOTE, ADASYN, Borderline-SMOTE, SMOTETomek and IADASYN) and integrated classifiers (XGBoost, LightGBM, CatBoost and FLCatBoost). Then it is compared with some classical imbalanced classification algorithms to further verify the classification effect of the IADASYN-FLCatBoost model. Finally, other bank customer churn dataset *Churn* and telecom customer churn dataset *Telechurn* on Kaggle platform are selected to further verify the generalizability of the model. In this paper, the prediction process of the model on the real imbalanced dataset is simulated in the experiment. On the basis of dividing the data set into training set and test set in a ratio of 7:3, the training set is balanced to make the model trained on the balanced data set, and then the results are evaluated based on the imbalanced data to make the results more reliable. In order to make different models comparable, the random number is set the same in all experiments.

5.1. The results of the staged comparison experiment

In order to verify the improvement effects of the IADASYN-FLCatBoost model at the data level and algorithm level respectively, staged experimental tests are conducted, including the original CatBoost model, the IADASYN-CatBoost model after sampling and the IADASYN-FLCatBoost model. In IADASYN oversampling method, the number of nearest neighbors in the LOF algorithm is set to 10 and the classification feature is set to [1, 2, 3, 5, 6, 7]. In CatBoost algorithm, the hyper-parameters are set to the default values. And in FLCatBoost algorithm, the α and γ in the Focal Loss function is set $\alpha = 0.25$ and $\gamma = 2$, respectively.

The *Recall*, *Precision*, *F*1 score, *G-mean*, *Kappa*, and *AUPRC* value corresponding to each stage of processing are shown in Table 5. In addition, Table 5 also shows the prediction results of two other popular algorithms in Boosting family ensemble learning algorithm: XGBoost and LightGBM. According to the results in Table 5, compared with XGBoost and LightGBM algorithms, the *Precision*, *F*1 score, *G-mean*, *Kappa* and *AUPRC* value of CatBoost model are all the highest, which shows that compared with other classifiers, CatBoost has certain advantages in predicting customer churn, and also proves the rationality of choosing CatBoost as the basic model for customer classification prediction.

In the staged experimental tests, compared with the original CatBoost model, after data processing, that is, balancing the distribution of customer data by IADASYN method, the evaluation indexes of this model have been improved to some extent, such as *Recall*, *G-mean*, *Kappa* and *AUPRC* value, among which the *Recall* is the most significant, with an increase of 2.49 %, while the *G-mean*, *Kappa* and *AUPRC* value have increased by 0.89 % and 0.1 respectively. It can be seen that the *Recall*, *G-mean*, *Kappa* and *AUPRC* value of IADASYN-FLCatBoost model are increased to 94.2 %, 95.93 %, 97.15 % and 97.51 % respectively after the loss function in CatBoost algorithm is further replaced by the focus loss function. On the whole, IADASYN-FLCatBoost model has

Table 5
Comparison results of the staged experiments.

Model	*Recall*	*Precision*	*F*1	*G-mean*	*Kappa*	*AUPRC*
XGBoost	0.9047	0.9258	0.9152	0.9446	0.9691	0.9455
LightGBM	0.9130	0.9265	0.9197	0.9447	0.9590	0.9497
CatBoost	0.9068	0.9359	0.9211	0.9467	0.9699	0.9623
IADASYN-CatBoost	0.9317	0.8982	0.9146	0.9556	0.9709	0.9649
IADASYN-FLCatBoost	0.9420	0.8852	0.9127	0.9593	0.9715	0.9751

obviously achieved the best results in terms of *Recall*, which is 3.52 % higher than that of the original CatBoost model. Moreover, the *G-mean* of CatBoost and IADASYN-CatBoost are 94.67 % and 95.56 % respectively, and the *G-mean* obtained by IADASYN-FLCatBoost model is the highest, reaching 95.93 %. It can also be seen from Fig. 8 that the area under the PR curve of IADASYN-FLCatBoost model is the largest, that is, the *AUPRC* value is the largest, and IADASYN-FLCatBoost is improved by 0.26 % and 1.02 % respectively compared with IADASYN-CatBoost and CatBoost algorithms. To sum up, the early warning model of customer churn based on IADASYN-FLCatBoost proposed in this paper is practical and meaningful, and can effectively improve the forecasting effect.

5.2. Experimental comparison with other classification models

5.2.1. Comparison of the combinations of other sampling algorithms and classifiers

In order to further verify the classification effect of IADASYN-CatBoost proposed in this paper, it is compared and analyzed with some classifiers combining the classical sampling methods and ensemble learning algorithms. To make the experiment more representative, other sampling methods selected include the classical SMOTE (Fonseca, & Bacao, 2023), ADASYN (Mitra, Bajpai, & Biswas, 2023), Borderline-SMOTE (Li, Zhu, Wu, & Fan, 2021) and SMOTETomek (Lui, Gregory, Anderson, Lee, & Cowling, 2022) comprehensive sampling method. The ensemble algorithms include the popular XGBoost, LightGBM (Truong, Tangaramvong, & Papazafeiropoulos, 2024), CatBoost and the FLCatBoost algorithm improved in this paper. In the experiment, all sampling methods have the same sampling ratio and are verified under the same test set. The test results of classifiers under the above 20 different combinations are shown in Table 6. In order to compare the performance of these methods more intuitively, all the results are presented in the form of line charts in Fig. 9.

Combining the results in Table 6 and Fig. 9, it can be found that, when the selected integrated classification algorithm is the same, the classifier based on the improved IADASYN method in this paper has a certain improvement in evaluation indexes such as *Precision*, *F*1 score, *Kappa* and *AUPRC* value compared with SMOTE, ADASYN, Borderline-SMOTE and SMOTETomek sampling methods. Especially, in FLCatBoost algorithm, the *Recall* of IADASYN-FLCatBoost and SMOTE-FLCatBoost models is equal, with the highest value of 94.2 %. Compared with other models, the *Precision* is improved by 3.15 %-4.32 %, the *F*1 score is improved by 1.7 %-2.53 %, and the *G-mean* is also slightly improved. The IADASYN-FLCatBoost model has achieved the highest value. On the *AUPRC* index, the IADASYN-FLCatBoost model proposed in this paper also performed the best (97.51 %), which is 1.32 % higher than that of Borderline-SMOTE-FLCatBoost model. In particular, compared with the original ADASYN sampling algorithm, the improved IADASYN sampling algorithm significantly improves the classification performance of FLCatBoost model, and has achieved good performance in six evaluation indexes. By training the data sampled by IADASYN algorithm, XGBoost algorithm also slightly improves the *Precision*, *F*1 score and *AUPRC* value, with the highest increases of 2.44 %, 1.12 % and 1.05 %, respectively. In the CatBoost algorithm, the *Precision*, *F*1 score and *Kappa* of the IADASYN-CatBoost model also improved to some extent, with the highest increases of 3.46 %, 1.61 % and 0.42 % respectively. In the LightGBM algorithm, IADASYN-LightGBM model performs best in *Precision*, *F*1 score, *Kappa* and *AUPRC* value, with the highest values of 86.97 %, 90.35 %, 96.64 % and 90.96 % respectively. It is proved that IADASYN sampling method can effectively remove the influence of outliers far from the classification boundary, and the special treatment of classification features in customer data sets further ensures the rationality of synthetic samples.

And when the same sampling methods are used, it can be found that SMOTE and CatBoost algorithm have the best combination effect, with the highest *Recall*, *F*1 score, *G-mean* and *Kappa*, which are increased by 1.45 %, 0.64 %, 0.65 % and 0.24 % respectively compared with other models. The *Precision*, *F*1 score and *Kappa* of ADASYN-CatBoost model

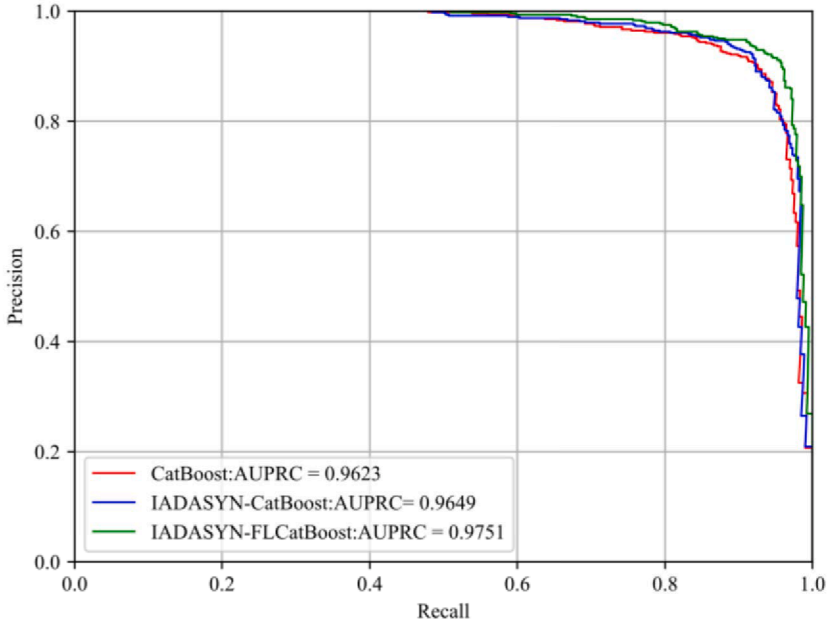


Fig. 8. Comparison of PR curves.

Table 6
Comparison of experimental results of different oversampling methods.

	Recall	*Precision*	*F*1	*G-mean*	*Kappa*	*AUPRC*
SMOTE-FLCatBoost	0.9420	0.8537	0.8957	0.9557	0.9634	0.9664
ADASYN-FLCatBoost	0.9400	0.8454	0.8902	0.9536	0.9513	0.9651
Borderline-SMOTE-FLCatBoost	0.9379	0.8420	0.8874	0.9522	0.9603	0.9619
SMOTETomek-FLCatBoost	0.9379	0.8499	0.8917	0.9532	0.9620	0.9641
IADASYN-FLCatBoost	0.9420	0.8852	0.9127	0.9593	0.9699	0.9751
SMOTE-XGBoost	0.9296	0.8702	0.8989	0.9514	0.9651	0.9055
ADASYN-XGBoost	0.9337	0.8526	0.8913	0.9514	0.9620	0.8984
Borderline-SMOTE-XGBoost	0.9317	0.8704	0.9000	0.9525	0.9655	0.9065
SMOTETomek-XGBoost	0.9379	0.8662	0.9006	0.9551	0.9655	0.9069
IADASYN-XGBoost	0.9296	0.8770	0.9025	0.9522	0.9664	0.9089
SMOTE-LightGBM	0.9420	0.8650	0.9019	0.9570	0.9658	0.9081
ADASYN-LightGBM	0.9462	0.8400	0.8900	0.9560	0.9610	0.8974
Borderline-SMOTE-LightGBM	0.9379	0.8483	0.8909	0.9530	0.9617	0.8980
SMOTETomek-LightGBM	0.9379	0.8612	0.8979	0.9545	0.9644	0.9045
IADASYN-LightGBM	0.9400	0.8697	0.9035	0.9565	0.9664	0.9096
SMOTE-CatBoost	0.9441	0.8636	0.9021	0.9579	0.9658	0.9653
ADASYN-CatBoost	0.9358	0.8743	0.9040	0.9550	0.9669	0.9617
Borderline-SMOTE-CatBoost	0.9255	0.8730	0.8985	0.9497	0.9651	0.9607
SMOTETomek-CatBoost	0.9358	0.8659	0.8995	0.9540	0.9651	0.9626
IADASYN-CatBoost	0.9317	0.8982	0.9146	0.9556	0.9709	0.9649

are 87.43 %, 90.40 % and 96.69 % respectively, which are also the highest. The classifier composed of Borderline-SMOTE method and FLCatBoost algorithm is obviously the best in *Recall* and *AUPRC* value. Compared with the other three models, the classifier SMOTETomek-XGBoost composed of SMOTETomek and XGBoost algorithm performs better on five evaluation indexes except the *AUPRC* value. As for the IADASYN sampling method proposed in this paper, the classifier combined with the improved FLCatBoost algorithm, that is, the IADASYN-FLCatBoost model proposed in this paper, has the best prediction effect, with the highest *Recall* increased by 1.24 %, the highest *G-mean* increased by 0.71 %, and the highest *AUPRC* value increased by 6.62 % compared with the IADASYN-XGBoost model. Therefore, by comparing with the above 20 classifiers from different angles, it is concluded that the IADASYN-FLCatBoost model proposed in this paper is effective from two aspects, which further shows the superiority of IADASYN-FLCatBoost model in predicting customer churn tendency.

5.2.2. Comparison with imbalanced classification algorithms

To further demonstrate the effectiveness of the model in this paper, this section compares and analyzes IADASYN-FLCatBoost with other SMOTEBoost (Zhao, Jin, Chen, Zhang, Yu, & Liu, 2020), RUSBoost (Tarkocin, & Donduran, 2024), SMOTEBagging (Shen, Liu, Wang, & Zhou, 2020), OverBoost (Wang, Li, Tang, Wang, & Xun, 2019) and SPE (Dai, Liu, & Liu, 2022) algorithms, which are popular in the field of imbalanced data classification, to verify the effectiveness and feasibility of the method. The experiments are conducted under the same test set, and the *Recall*, *Precision*, *F*1 score, *G-mean*, *Kappa* coefficient and *AUPRC* value of each algorithm are obtained as shown in Table 7.

According to Table 7, it can be found that compared with other imbalanced classification algorithms, the IADASYN-FLCatBoost model proposed in this paper also performs best. In terms of the *Recall*, OverBoost achieved the highest value, followed by the performance of this model. The *Precision* obtained by IADASYN-FLCatBoost is the highest, reaching 88.52 %, which is 4.18 %-18.99 % higher than other algorithms. Combining the above two indexes, the model finally improved the *F*1 score by 4.55 %-11.34 %, among which RUSBoost algorithm performed worst, while the model proposed in this paper performed best. Compared with the five existing imbalanced classification algorithms, the IADASYN-FLCatBoost model has improved by 1.52 %-3.39 % and 2.00 %-5.59 % respectively, in which SMOTEBagging algorithm has the lowest *G-mean* (92.54 %) and RUSBoost algorithm has the worst performance on *Kappa* (91.95 %). Fig. 10 shows the comparison of *Recall*, *Precision*, *F*1 score and other indicators more intuitively. On the whole, SMOTEBagging has the worst classification performance in this imbalanced credit card customer training set, and the classification prediction effect of the model proposed in this paper is the best. It can also be seen from Table 6 and Fig. 9 that the *AUPRC* value of IADASYN-FLCatBoost is the highest (97.51 %), which is 3.23 %-5.53 % higher than other models. Fig. 11 plots the PR curves of the above six models. It can be seen that the area under the PR curve of IADASYN-FLCatBoost model is the largest, that is, the *AUPRC* value is the highest. To sum up, the IADASYN-FLCatBoost model has certain advantages in the field of credit card customer churn prediction.

5.3. Further validation of the model

To further verify the validity and generalizability of the IADASYN-FLCatBoost model on other imbalanced customer datasets, the other bank customer churn dataset[2] *Churn* and telecom customer churn dataset[3] *Telechurn*, are selected for testing. Among them, the dataset *Churn* has 10,000 customer data and 10 feature variables, containing 5 categorical features, and the distribution ratio of churn customers to non-churn customers is about 1:5. The dataset *Telechurn* has 3,333 customer data and 10 feature variables, containing 2 categorical features, and the data imbalance ratio is about 1:6.

First, the redundant features “CustomerId” (customer’s ID information) and “Surname” (customer’s name) are removed from the dataset *Churn*, and the categorical features “ Geography” and “Gender” are coded and standardized for continuous variables, while the dataset *Telechurn* only requires Z-score standardization for eight continuous variables. Since these two datasets have fewer feature variables, feature selection is not required. Then, they are divided into training and test sets in the ratio of 7:3, respectively, and both training sets are sampled using the IADASYN algorithm to balance the sample distribution. Finally, the FLCatBoost algorithm with default parameter settings is trained to establish the IADASYN-FLCatBoost model. The same staged experimental tests are conducted here, including the original CatBoost model, the sampled IADASYN-CatBoost model, and the IADASYN-FLCatBoost model. The parameters in the experiments are set the same as in Section 5.1, and the test results obtained are shown in Tables 8 and 9.

Combined the experimental results in Tables 7 and 8, it can be found that the IADASYN-FLCatBoost model has achieved the highest values in terms of *Recall*, *Precision*, *F*1 score, *G-mean*, *Kappa*, and *AUPRC* value. In the dataset *Churn*, after IADASYN sampling, the *Recall*, *F*1 score, *G-mean* and *AUPRC* are increased by 26.51 %, 8.49 %, 12.77 % and 8.24 %,

[2] https://www.kaggle.com/mathchi/churn-for-bank-customers.
[3] https://https://www.kaggle.com/barun2104/telecom-churn.

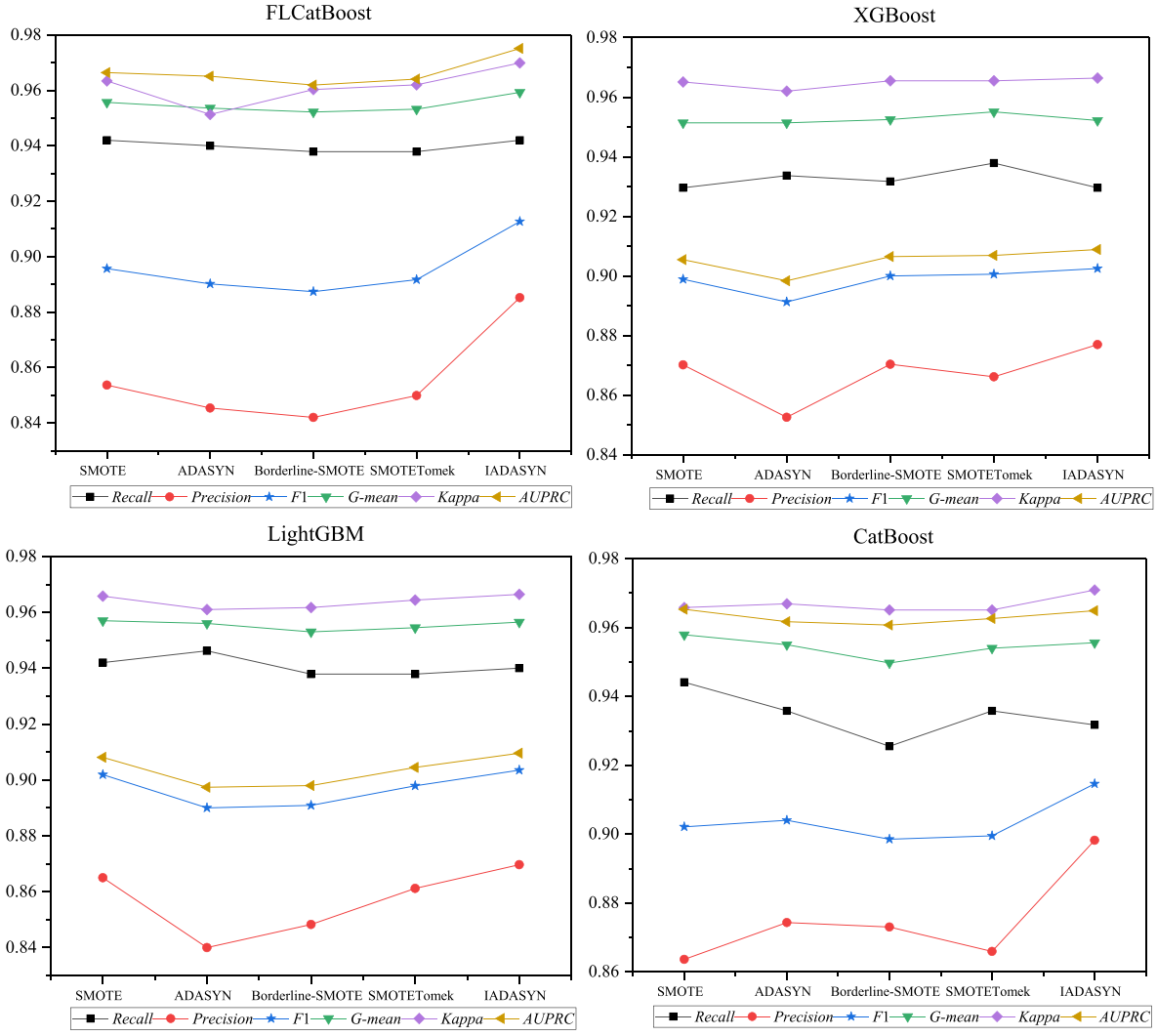


Fig. 9. Comparison of different oversampling methods.

Table 7
Comparison results with other imbalanced classification algorithms.

Algorithms	*Recall*	*Precision*	*F1*	*G-mean*	*Kappa*	*AUPRC*
SMOTEBoost	0.9337	0.7580	0.8367	0.9387	0.9334	0.9285
RUSBoost	0.9400	0.6953	0.7993	0.9310	0.9140	0.9295
SMOTEBagging	0.8861	0.8327	0.8586	0.9254	0.9465	0.9198
OverBoost	0.9545	0.7317	0.8284	0.9441	0.9279	0.9428
SPE	0.8923	0.8434	0.8672	0.9297	0.9499	0.9392
IADASYN-FLCatBoost	0.9420	0.8852	0.9127	0.9593	0.9699	0.9751

respectively, and *Precision* and *Kappa* are decreased by 10.85 % and 0.65 %, respectively. but by improving the loss function, the IADASYN-FLCatBoost model significantly improves the *Precision* and *Kappa* coefficient, which are significantly higher than the original CatBoost model, compensating for the decrease in these two evaluation indexes after sampling, and the six indexes improve by 21.48 %, 14.79 %, 17.57 %, 13.08 %, 9.69 %, and 21.28 %, respectively. In the dataset *TeleChurn*, the stage experimental test results also performed better. With the improvement of the data level, *Recall*, *F1* score, *G-mean* and *AUPRC* value of the IADASYN-FLCatBoost model increased significantly by 25.79 %, 14.43 %, 13.96 % and 15.00 %, while the *Precision* and *Kappa* increased slightly by 1.04 % and 4.31 % respectively, and the subsequent algorithm level is further improved. Figs. 12 and 13 also clearly show the comparison results between different models. Therefore, the IADASYN-FLCatBoost model proposed in this paper is also very effective for customer churn data sets of other banks and other industries.

6. Conclusion

In the era of Bank 4.0, it is imperative for banks to effectively use the massive amount of stored structured, semi-structured and unstructured data to construct customer churn prediction model. As customer churn prediction is a high-dimensional imbalanced classification problem, the operation efficiency of the model will be affected if the dimension is too high, and the data imbalance will affect the classification effect of the model, resulting in the misjudgment of churn customers. In addition, the key concern of bank managers is whether the churn customers can be correctly predicted. Therefore, this paper proposes a multi-strategy collaborative processing method to solve this problem at dual perspectives of data and algorithm. The EN and XGBoost algorithm, which are widely used in high dimensional feature space problems, are used to select the key factors affecting credit card customer churn and get the optimal feature subset. Then, at the data level, IADASYN algorithm is used to generate minority samples to balance the data distribution. At the algorithm level, Focal Loss is introduced to embed CatBoost

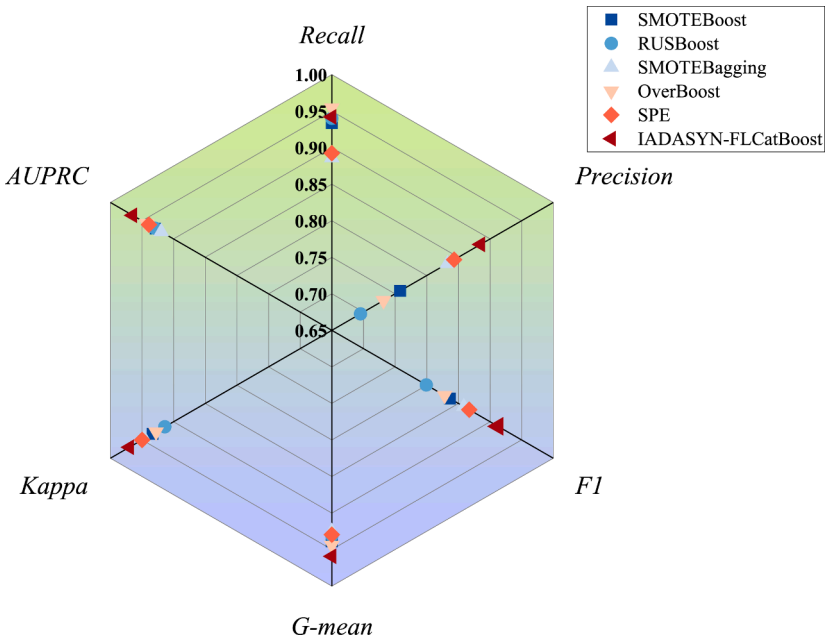


Fig. 10. Comparison of imbalanced algorithms.

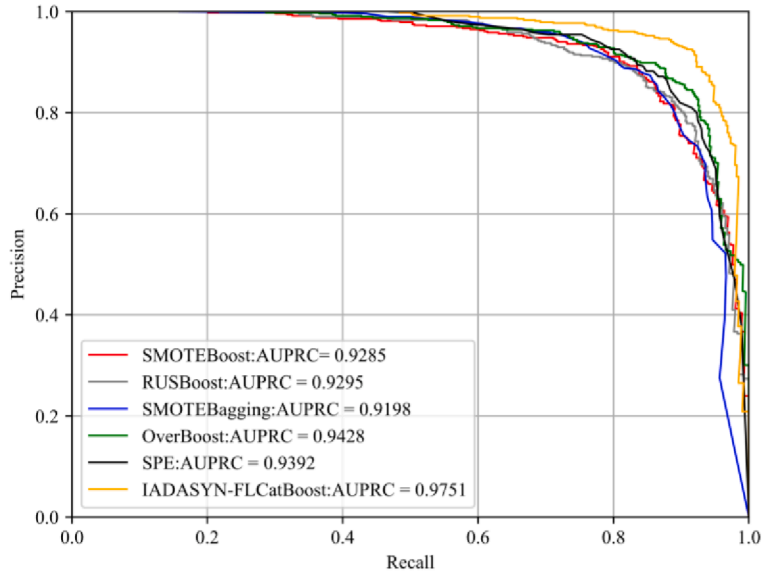


Fig. 11. Comparison of PR curves.

integrated learning framework for imbalanced data classification, and a customer churn warning model named IADASYN-FLCatBoost based on the multi-strategy collaborative processing method is proposed. The feasibility of the model is proved by the comparison of the staged improvement and other methods, and the prediction effect proposed in this paper is better in the field of credit card customer churn prediction. And the model is further verified by another bank customer churn dataset *Churn* and the telecom user churn dataset *Telechurn* on the Kaggle platform. The results show that the IADASYN-FLCatBoost model proposed in this paper has certain universality in the field of customer churn prediction.

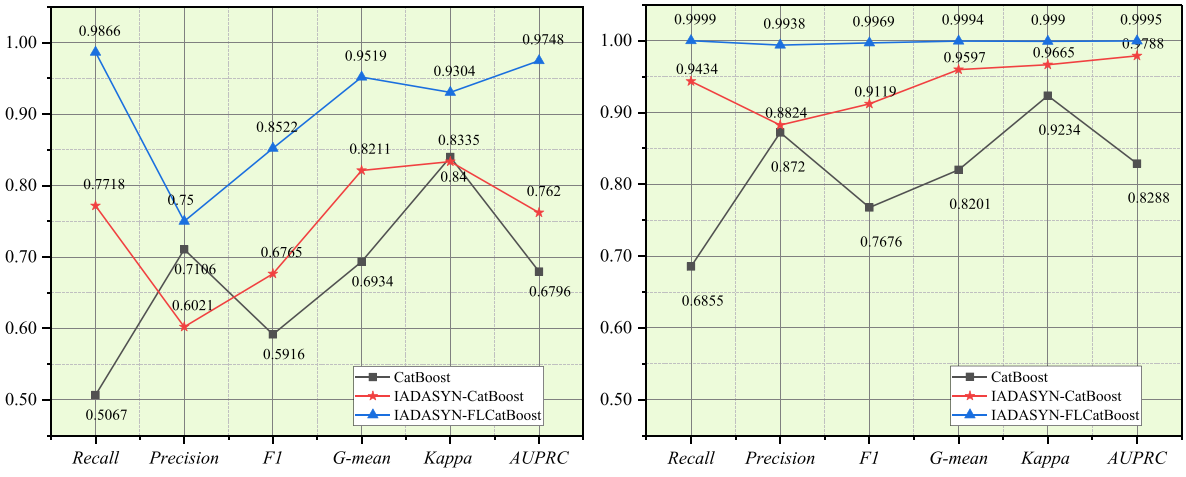
Although FLCatBoost ensemble algorithm improves the classification ability by fusing multiple base classifiers, the classification accuracy will be affected by its internal hyper-parameters to a certain extent. Manual parameter adjustment is not only a hit-or-miss endeavor but also entails an intensive workload, so it is difficult to find the optimal parameter combination, thus affecting the prediction accuracy of the model. To address this issue, while some researchers have employed swarm intelligence optimization algorithms for parameter optimization, the classical optimization algorithm may carry intrinsic shortcomings and cannot achieve effective optimization. From this perspective, in our

Table 8
Comparison results for dataset *Churn*.

Model	*Recall*	*Precision*	*F1*	*G-mean*	*Kappa*	*AUPRC*
CatBoost	0.5067	0.7106	0.5916	0.6934	0.8400	0.6796
IADASYN-CatBoost	0.7718	0.6021	0.6765	0.8211	0.8335	0.7620
IADASYN-FLCatBoost	0.9866	0.7500	0.8522	0.9519	0.9304	0.9748

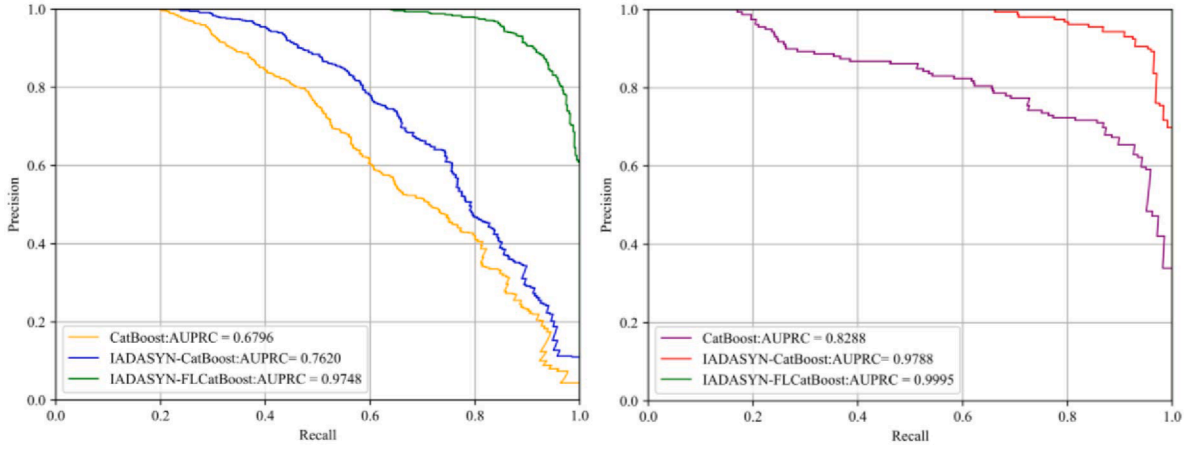
Table 9
Comparison results for dataset *Telechurn*.

Model	*Recall*	*Precision*	*F1*	*G-mean*	*Kappa*	*AUPRC*
CatBoost	0.6855	0.8720	0.7676	0.8201	0.9234	0.8288
IADASYN-CatBoost	0.9434	0.8824	0.9119	0.9597	0.9665	0.9788
IADASYN-FLCatBoost	0.9999	0.9938	0.9969	0.9994	0.9990	0.9995



(a) *Churn* (b) *Telechurn*

Fig. 12. Comparison of evaluation indicators.



(a) *Churn* (b) *Telechurn*

Fig. 13. Comparison of PR curves.

future work, we will propose an improved algorithm to optimize the important hyper-parameters of FLCatBoost algorithm for the classical intelligent optimization algorithms, and develop a hybrid model based on machine learning and swarm intelligence optimization algorithm.

CRediT authorship contribution statement

Congjun Rao: Conceptualization, Methodology, Data curation. **Yaling Xu:** Methodology, Software, Writing – original draft. **Xinping Xiao:** Supervision, Writing – review & editing. **Fuyan Hu:** Visualization, Investigation, Validation. **Mark Goh:** Formal analysis, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 72071150, 72371194).

References

Abraham, B., & Nair, M. S. (2018). Computer-aided diagnosis of clinically significant prostate cancer from MRI images using sparse autoencoder and random forest classifier. *Biocybernetics and Biomedical Engineering, 38*(3), 733–744.

Amini, F., & Hu, G. P. (2021). A two-layer feature selection method using genetic algorithm and elastic net. *Expert Systems with Applications, 166*, Article 114072.

Batista, G., Prati, R. C., & Monard, M. C. (2014). A study of the behavior of several methods for balancing machine learning training data. *Acm Sigkdd Explorations Newsletter, 6*(1), 20–29.

Bentejac, C., Csorgo, A., & Martinez-Munoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937–1967.

Bileki, G. A., Barboza, F., Silva, L. H. C., & Bonato, V. (2022). Order book mid-price movement inference by CatBoost classifier from convolutional feature maps. *Applied Soft Computing, 116*, Article 108274.

Cao, H. L., Xie, X. J., Shi, J. B., & Wang, Y. X. (2022). Evaluating the validity of class balancing algorithms-based machine learning models for geogenic contaminated groundwaters prediction. *Journal of Hydrology, 610*, Article 127933.

Chan, J. P., Papaioannou, I., & Straub, D. (2023). Bayesian improved cross entropy method for network reliability assessment. *Structural Safety, 103*, Article 102344.

Chan, K. Y., Kwong, C. K., & Jiang, H. M. (2021). Analyzing imbalanced online consumer review data in product design using geometric semantic genetic programming. *Engineering Applications of Artificial Intelligence, 105*, Article 104442.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*(1), 321–357.

Chen, C., Zhang, Q. M., Yu, B., Yu, Z. M., Lawrence, P. J., Ma, Q., & Zhang, Y. (2020). Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. *Computers in Biology and Medicine, 123*, Article 103899.

Chen, G. C., & Qin, H. B. (2021). Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. *Visual Computer, 38*(3), 1051–1063.

Chen, S. X., Wang, X. K., Zhang, H. Y., & Wang, J. Q. (2021). Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. *Expert Systems with Applications, 173*, Article 114756.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327.

Cui, L. X., Bai, L., Wang, Y. C., Jin, X., & Hancock, E. R. (2021). Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection. *Pattern Recognition, 114*, Article 107835.

Dai, Q., Liu, J. W., & Liu, Y. (2022). Multi-granularity relabeled under-sampling algorithm for imbalanced data. *Applied Soft Computing, 124*, Article 109083.

Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences, 548*, 497–515.

Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing, 19*, 31–40.

Fonseca, J., & Bacao, F. (2023). Geometric SMOTE for imbalanced datasets with nominal and continuous features. *Expert Systems with Applications, 234*, Article 121053.

Han, H., Wang, W., & Mao, B. (2005). Border-line-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the 2005 International Conference on Intelligent Computing* (pp. 878-887). Cham: Springer Nature Switzerland.

Hussain, S., Mustafa, M. W., Jumani, T. A., Baloch, S. K., Alotaibi, H., Khan, I., & Khan, A. (2021). A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. *Energy Reports, 7*, 4425–4436.

Jain, N., Tomar, A., & Jana, P. K. (2021). A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. *Journal of Intelligent Information Systems, 56*(2), 279–302.

Jelen, G., Podobnik, V., & Babic, J. (2021). Contextual prediction of parking spot availability: A step towards sustainable parking. *Journal of Cleaner Production, 312*, Article 127684.

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing, 24*, 994–1012.

Kumar, S., Biswas, S. K., & Devi, D. (2019). TLUSBoost algorithm: A boosting solution for class imbalance problem. *Soft Computing, 23*(21), 10755–10767.

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications, 77*(2), 277–285.

Li, J., Zhu, Q. S., Wu, Q. W., & Fan, Z. (2021). A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences, 565*, 438–455.

Li, Y., Jia, X. Y., Wang, R. L., Qi, J. F., Jin, H. B., Chu, X. Q., & Mu, W. S. (2022). A new oversampling method and improved radial basis function classifier for customer consumption behavior prediction. *Expert Systems with Applications, 199*, Article 116982.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 318–327.

Liu, W. A., Fan, H., Xia, M., & Xia, M. (2022). A focal-aware cost-sensitive boosted tree for imbalanced credit scoring. *Expert Systems with Applications, 208*, Article 118158.

Liu, Y. N., Yu, Z. M., Chen, C., Han, Y., & Yu, B. (2020). Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net. *Analytical Biochemistry, 609*, Article 113903.

Lui, T. C. C., Gregory, D. D., Anderson, M., Lee, W., & Cowling, S. A. (2022). Applying machine learning methods to predict geology using soil sample geochemistry. *Applied Computing and Geosciences, 16*, Article 100094.

Meng, D. X., & Li, Y. J. (2022). An imbalanced learning method by combining SMOTE with center offset factor. *Applied Soft Computing, 120*, Article 108618.

Mitra, R., Bajpai, A., & Biswas, K. (2023). ADASYN-assisted machine learning for phase prediction of high entropy carbides. *Expert Systems with Applications, 223*, Article 112142.

Mushava, J., & Murray, M. (2022). A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications, 202*, Article 117233.

Nie, G. L., Wei, R., Zhang, L. L., Tian, Y. J., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications, 38*(12), 15273–15285.

Pang, Y., Peng, L. Z., Chen, Z. X., Yang, B., & Zhang, H. L. (2019). Imbalanced learning based on adaptive weighting and Gaussian function synthesizing with an application on Android malware detection. *Information Sciences, 484*, 95–112.

Praveen, L., Manas, K. M., Jasroop, S. C., & Pratyush, S. (2022). Customer churn prediction system: A machine learning approach. *Computing, 104*, 271–294.

Pulicherla, P., Kumar, T., & Abbaraju, N. (2019). Job shifting prediction and analysis using machine learning. *Journal of Physics Conference Series, 1228*(1), Article 012056.

Pustokhina, I. V., Pustokhin, D. A., Nguyen, P. T., Elhoseny, M., & Shankar, K. (2023). Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. *Complex & Intelligent Systems, 9*, 3473–3485.

Qiu, W. (2020). *Research on the application of telecom customer churn prediction based on random forest*. China: South China University of Technology.

Rao, C. J., Liu, M., Goh, M. Y., & Wen, J. H. (2020). 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers. *Applied Soft Computing, 95*, Article 106570.

Rao, C. J., Huang, Q. F., Chen, L., Goh, M., & Hu, Z. (2023). Forecasting the carbon emissions in Hubei Province under the background of carbon neutrality: A novel STIRPAT extended model with ridge regression and scenario analysis. *Environmental Science and Pollution Research, 30*(20), 57460–57480.

Shen, F., Liu, Y. Y., Wang, R., & Zhou, W. (2020). A dynamic financial distress forecast model with multiple forecast results under imbalanced data environment. *Knowledge-Based Systems, 192*, Article 105365.

Srikanth, B., Papineni, S. L. V., Sridevi, G., Indira, D. N. V. S. L. S., Radhika, K. S. R., & Syed, K. (2022). Adaptive XGBoost hyper tuned meta classifier for prediction of churn customers. *Intelligent Automation and Soft Computing, 33*(1), 21–34.

Sundarkumar, G. G., & Ravi, V. (2015). A novel hybrid undersampling method for mining imbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence, 37*, 368–377.

Swetha, P., & Dayananda, R. B. (2020). Improvised – XGBoost Machine learning Algorithm for Customer Churn *Prediction*. *EAI Endorsed Transactions on Energy Web, 7*(30), 1–7.

Tang, L. F., Xie, H. P., Wang, Y., Zhu, H., & Bie, Z. H. (2022). Predicting typhoon-induced transmission line outages with coordination of static and dynamic data. *International Journal of Electrical Power & Energy Systems, 142*, Article 108296.

Tarkocin, C., & Donduran, M. (2024). Constructing early warning indicators for banks using machine learning models. *The North American Journal of Economics and Finance, 69*, Article 102018.

Tekouabou, S. C. K., Gherghina, S. C., Toulni, H., Mata, P. N., & Martins, J. M. (2022). Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. *Mathematics, 10*(14), 2379.

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences, 513*, 429–441.

Thammasiri, D., Hengpraprohm, S., Hengpraprohm, K., & Mukviboonchai, S. (2018). Imbalance classification model for churn prediction. *Advanced Science Letters, 24*(2), 1348–1351.

Truong, V., Tangaramvong, S., & Papazafeiropoulos, G. (2024). An efficient LightGBM-based differential evolution method for nonlinear inelastic truss optimization. *Expert Systems with Applications, 237*, Article 121530.

Wang, C., Deng, C. Y., & Wang, S. Z. (2020). Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters, 136*, 190–197.

Wang, G., & Chen, Y. F. (2021). Robust feature matching using guided local outlier factor. *Pattern Recognition, 117*, Article 107986.

Wang, H. M., Liu, Y. H., Yin, C. C., Ren, X. Y., Cao, J., Su, Y., & Xiong, B. (2021). Fast 3D time-domain airborne EM forward modeling using random under-sampling. *Journal of Applied Geophysics, 191*, Article 104357.

Wang, J., Rao, C. J., Goh, M., & Xiao, X. P. (2023). Risk assessment of coronary heart disease based on cloud-random forest. *Artificial Intelligence Review, 56*(1), 203–232.

Wang, W. D., & Ji, C. (2018). Key elements of retail banking operations and countermeasures in the Bank 4.0 era. *New Finance, 10*, 17–20.

Wang, X., Li, S. K., Tang, T., Wang, X. N., & Xun, J. (2019). Intelligent operation of heavy haul train with data imbalance: A machine learning method. *Knowledge-Based Systems, 163*, 36–50.

Wei, X., Rao, C. J., Xiao, X. P., Chen, L., & Goh, M. (2023). Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model. *Expert Systems with Applications, 219*, Article 119648.

Wen, X. H., Wang, Y. H., Ji, X. D., & Traore, M. K. (2022). Three-stage churn management framework based on DCN with asymmetric loss. *Expert Systems with Applications, 207*, Article 117998.

Wong, M. L., Seng, K., & Wong, P. K. (2020). Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. *Expert Systems with Applications, 141*, Article 112918.

Wu, X., Li, P., Zhao, M., Liu, Y., González Crespo, R., & Herrera-Viedma, E. (2022). Customer churn prediction for web browsers. *Expert Systems with Applications, 209*, Article 118177.

Wu, X. T., Zhao, X., Guo, Y. W., & Run, H. F. (2016). Prediction of Online Game User Turnover: Comparison and analysis of sampling methods based on unbalanced data. *Journal of Chinese Information, 30*(4), 213–222.

Xiao, J., Xie, L., He, C. Z., & Jiang, X. Y. (2012). Dynamic classifier ensemble model for customer classification with imbalanced class distribution. *Expert Systems with Applications, 39*(3), 3668–3675.

Xiao, Q. Z., Gao, M. Y., Chen, L., & Jiang, J. C. (2023). Dynamic multi-attribute evaluation of digital economy development in China: A perspective from interaction effect. *Technological and Economic Development of Economy*, 2023, in press, DOI: doi: 10.3846/tede.2023.20258.

Xie, Y. Y., Li, X., Ngai, E. W. T., & Ying, W. Y. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445–5449.

Zhang, T. Y., & Ding, L. X. (2023). Another use of SMOTE for interpretable data collaboration analysis. *Expert Systems with Applications, 228*, Article 120385.

Zhang, X. L., Rao, C. J., Xiao, X. P., Hu, F. Y., & Goh, M. (2024). Prediction of demand for staple food and feed grain by a novel hybrid fractional discrete multivariate grey model. *Applied Mathematical Modelling, 125*, 85–107.

Zhao, J. K., Jin, J., Chen, S., Zhang, R. F., Yu, B. L., & Liu, Q. F. (2020). A weighted hybrid ensemble method for classifying imbalanced data. *Knowledge-Based Systems, 203*, Article 106087.