Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

# Research on telecommunication subscriber churn prediction model based on quadratic classification method

Chunyu Zhang[1, ✉]

[1] *Management School of Northwestern Polytechnical University, Xi'an, Shaanxi, 710129, China*

ABSTRACT

The development of communication technology and the rapid growth of the number of mobile network service users have made the competitive situation in the market of communication service increasingly fierce, and maintaining the stock of users is of great significance to the sustainable development of telecommunication enterprises. In this paper, we collect relevant data features of telecommunication users, and after pre-processing the features with RFM model, we use XGBoost model to analyze the importance of each user's feature value. Then we use the secondary classification Stacking integration model that combines the base learner and the meta-learner to predict the telecom subscriber churn. Comparative validation reveals that the prediction model in this paper shows excellent prediction performance in all four datasets. Practical application results show that the effectiveness of churn maintenance efforts by telecom companies is improved after applying the model, and the average maintenance response rate reaches 50.63% in the first quarter of 2024. The prediction model proposed in this paper based on the binary classification method can assist telecommunication companies to manage the stock of subscribers, optimize the maintenance work plan, and reduce the subscriber churn rate in the telecommunication work period.

*Keywords:* RFM model, xgboost, eigenvalue, stacking algorithm, subscriber churn prediction

## 1. Introduction

With the application of big data, Internet of Things, cloud computing, 5G, Web3.0 and other technologies in reality, a variety of emerging advanced technologies are silently promoting an informatization change [1-2]. In the era of intelligence and informatization that can be seen everywhere, in order to enjoy the convenience brought by intelligence, one phone per hand makes the telecommunication users show a geometric growth, and the competition in the telecommunication industry becomes more and more intense, and the market share is higher when there are more users [3-5]. However, with the introduction of a series of policies such as number portability and speed

✉ Corresponding author.
*E-mail address:* cyz_001@mail.nwpu.edu.cn (C. Zhang).
Received 20 February 2024; Revised 15 May 2024; Accepted 20 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-438

reduction, the rate of subscribers leaving the network is getting higher and higher, and subscriber churn not only brings about a decrease in revenue, but also affects the reputation of operators [6-8]. Therefore, it is very important and urgent for operators to predict subscriber churn in a timely and accurate manner and take appropriate measures to retain them [9-10].

Understanding the causes of subscriber churn is essential to predict and prevent churn. Generally, the causes of subscriber churn can be categorized into internal and external factors. Intrinsic factors include subscriber satisfaction, service quality, product price, and competitors' advantages [11-14]. Extrinsic factors, on the other hand, include user's life changes, relocation, job changes, etc. These factors can directly or indirectly affect users' demand and choice of telecommunication services. Understanding the causes of subscriber churn can be targeted to develop preventive measures to effectively reduce subscriber churn [15-18]. The commonly used telecom user churn methods include data analysis, machine learning, text mining, secondary classification, etc. The essence is to use artificial intelligence algorithms and data mining techniques to mine the personal information and business information of past telecom users to extract information that can directly or indirectly reflect the user's status, such as package costs, consumption records, whether churn and other related data [19-22]. AI algorithms are used to train a classification prediction model, and then the trained model is used to predict the probability of user churn. The users with high churn probability are visited in advance to retain them, which ensures the business revenue of telecommunication companies and helps the stock business operation [23-26].

The telecom industry is a highly competitive and challenging industry, and subscriber churn has always been a major concern for organizations. Accurate prediction of subscriber churn can help telecommunication companies take timely measures to improve subscriber retention and reduce business costs. Literature [27] describes the huge loss caused by subscriber churn and applies exploratory data analysis to predict subscriber churn, which realizes the prediction of potential subscriber churn through the use of machine learning models and effectively reduces the loss of enterprises. Literature [28] explored the Net Promoter Score (NPS) affecting telecom subscriber churn with the aim of identifying the factors of subscriber churn, proposing a prediction framework, and developed the propensity to churn through the use of logistic regression, classification, and regression tree (CART), which revealed that CART is the most accurate in predicting subscriber churn. Literature [29] developed an analysis of subscriber churn behavior with the aim of testing the efficiency and performance of commonly used data mining techniques to predict subscriber churn behavior, table by understanding the severity of churn, telecommunication companies can take steps to increase subscriber retention. Literature [30] mentions a rule-based intelligent decision making technique based on rough set theory, which can effectively categorize churned and non-churned subscribers and predict subscribers with churn intent, providing optimal solutions for telecommunication sectors. Literature [31] aims to establish a churn prediction model to predict the churn of telecommunication subscribers through subscriber segmentation, which is helpful to help telecommunication companies to effectively predict the churn rate of subscribers, so as to take targeted measures to avoid churn and increase profits. Literature [32] aims to examine the machine learning algorithms required to construct a subscriber churn prediction model and determine the causes of churn, and the results of the study effectively enhance subscriber churn prediction, and the provided prediction model improves the retention rate of existing subscribers and effectively prevents subscriber churn. Literature [33] builds a data processing method for churn prediction in telecommunication services, the model is able to effectively predict the pattern of churn and shows certain advantages, and predicts churn by proposing a hybrid fuzzy disordered rule induction algorithm and fuzzy C-mean clustering. Literature [34] investigated the role of Markov chains in subscriber churn and retention in the mobile telecom industry, and by analyzing the data generated it was shown that MTNs have the highest subscriber retention in the study area, which is instructive for telecom companies. Literature [35] describes the problem exhibit related to subscriber churn and analyzes the adoption of machine

learning techniques to predict subscriber churn through models such as XGBoost, GradientBoost, AdaBoost and compares the effectiveness of the models in terms of accuracy. Literature [36] identified important predictors of subscriber churn and strategies to improve subscriber retention in telecom industry by applying dataset and data analysis, machine learning techniques. Literature [37] outlined the data mining techniques used for subscriber churn based models based on literature review and summarized the existing literature on telecommunication and provided subscriber churn prediction techniques such as decision trees, regression analysis, etc., which is informative for other studies. Literature [38] illustrates the results of different classification models for a telecommunication company as an example and verifies that the use of models in further classification of subscribers to predict the loyalty of subscribers to the company helps to reduce subscriber churn.

In this paper, daily and monthly data of users are collected from relevant departments and platforms of telecommunication companies, such as users' call, SMS and Internet records, etc. The collected telecommunication user data are then processed and organized to obtain user dynamic features used as inputs for prediction models. The RFM model is used to simplify the user dynamic features and reduce the feature dimensions in order to improve the prediction ability of the model. Subsequently, the XGBoost model is used to calculate the importance and classify the user data features processed by the RFM model, and then the Stacking integration algorithm improved by Bootstrap self-sampling method and precision weighting method is used for user churn prediction, and the secondary classifier is established to realize the accurate prediction of telecom user churn. The performance of the prediction model is evaluated through comparative experiments, and then the model is applied in J Telecom Company to explore the effectiveness of the model in terms of the response rate of churn subscriber maintenance.

## 2. Method

### *2.1. Telecommunications subscriber data collection*

2.1.1. User day data collection. The daily data of the users of telecommunication companies that need to be collected are the daily behavioral data of the users, including the users' call records, SMS records, and Internet records. These data can reflect the user's daily behavior and are an important basis for predicting user churn in this paper.

1) User call information. Detailed information about the user's calls on the same day, including the length of each call, whether the caller or the called, whether the call was local or roaming, whether it was a long-distance call or not, and the billing length of the call.

2) User Internet behavior data. Detailed information about the user's traffic on the Internet on the same day, such as how much upstream traffic and how much downstream traffic each time, the length of Internet access, whether the Internet access is in the local area or roaming, and so on.

3) User SMS information. User's SMS usage on the same day, such as when the user sends SMS, whether the SMS is charged or not, and so on.

4) Value-added service data. Information on value-added services subscribed by the user on the same day, such as video ringtone, cloud disk, cloud business card, upgrade package, gasoline package, super member and so on.

5) Carry forward related information. Information related to the user's porting service on the same day, such as the user's query for porting code, the user's query for authorization code, the query for the porting code verification result (whether it is allowed to be ported out or not), and so on.

2.1.2. Monthly user data collection. In user churn prediction research, the collection of monthly data can reflect the overall situation of users more comprehensively and is very important for constructing prediction models. The monthly data collected in this study can be categorized as follows.

1) Basic user information. The user's basic business information, such as the user's network entry time, the user's package type, the user's last switching time, the user's terminal information, and the user's contract expiration time.

2) User traffic information. Detailed list of the user's traffic usage in the month, including the total traffic of the user's package, the user's 2/3G traffic used in the month, the user's 4G traffic used in the month, the user's 5G traffic used in the month, and information on apps accessed by the user in the month.

3) User call information. Detailed list of the user's call information for the month, including call duration, number of the other party, calling or called, local or roaming, long distance or not, and so on.

4) User Complaint Information. User's complaint record in the month, including the time of the complaint.

5) User SMS information. User's SMS details for the month, including the time of SMS sending, the other party's number, etc.

*2.2. User feature construction and processing*

2.2.1. Dynamic feature construction. Dynamic feature construction is to use the collected five categories of user day data for further data processing, and then get the dynamic features that can be used as input to the model, the specific features of the construction process in this paper according to the collected five categories of data are introduced.

1) User call features. According to the detailed records of the user's calls, whether the user has made calls on that day, the total duration of calls and the number of calls are counted. According to the caller and callee information of the user's calls, it can be recognized whether the user has active calls or not, and if there is only a caller call but no callee call on the same day, it can be recognized that the user is an active communication user. According to whether the user's call is local or not, it can be determined whether the user has roaming calls on that day. According to the attribution information of the user's caller's number, it can be determined whether the user has long-distance calls on the same day.

2) Characteristics of User Internet Behavior. According to the user's online traffic records on the same day, it can be counted whether the user went online, upstream traffic, downstream traffic, the length of online traffic, and the number of times the user went online on the same day.

3) Characteristics of user's SMS. According to the user's SMS records on the same day, we can count whether the user sends SMS and the number of SMS on the same day.

4) Value-added service features. According to the information of value-added services used by the user, we can determine whether the user uses value-added services on the same day.

5) Carry forward related features. According to the list of the user's carry code inquiries on the same day, we can count the number of carry eligibility inquiries and whether the user is allowed to carry out on the same day.

6) Crossover features. Cross-analysis of the above features can lead to further user behavioral features, for example, if the user has calls or traffic or SMS, it can be judged whether the user is active or not, and it can be judged whether there is only voice, whether there is only SMS, whether there is only value-added services based on calls, SMS, and value-added services.

The final dynamic characteristics derived after analyzing the user's characteristics are shown in Table 1, with a total of 21 telecom user characteristic variables.

**Table 1.** Dynamic characteristic variable

| Serial number | Characteristic variable | For short | Serial number | Characteristic variable | For short |
|---|---|---|---|---|---|
| 1 | Date | STAT_DATE | 12 | Internet number | FLUX_TIMES |
| 2 | Whether to call | IS_CALL | 13 | Text message | IS_SMS |
| 3 | Length of call | CALL_DURATION | 14 | Text strip | SMS_NUM |
| 4 | Number of calls | CALL_TIMES | 15 | Appreciation | IS_INCR |
| 5 | Whether to be proactive | IS_INITV | 16 | Whether to allow | OUT_TAG |
| 6 | Long distance call | IS_LONG_CALL | 17 | Number of eligibility queries | QUERY_TIMES |
| 7 | Roaming calls | IS_ROAM_CALL | 18 | Be active | IS_ACTIVE |
| 8 | Internet access | IS_GPRS | 19 | Is it just voice | IS_ONLY_CALL |
| 9 | Upward flow | FLUX_UP | 20 | Is it just only have text messages | IS_ONLY_SMS |
| 10 | Downstream flow | FLUX_DOWN | 21 | Whether it's just appreciation | IS_ONLY_INCR |
| 11 | Internet length | FLUX_DURA | —— | —— | —— |

2.2.2. Dynamic Feature RFM Processing. Due to the relatively large number of dynamic features of the user and the fact that traditional machine learning algorithms such as the XGBoost algorithm used in this paper cannot directly use time series data, and if the user's daily features are expanded into a row vector and calculated according to the time step of 30 days, there will be 600 features as inputs to the model. In order to improve the efficiency of the model and reduce the number of feature dimensions, this paper first uses the RFM model [39] for feature processing, which not only simplifies the process of feature processing, but also provides more information about the user's behavior of the feature.

By subjecting the multi-day features to RFM, it is possible to transform them into three separate features, i.e., the most recent time of use of this feature, the frequency of use, and the total number of times it has been used. For this purpose, this paper defines the $R$, $F$, and $M$ values for each time-behavioral feature $F_i$.

In this paper, the Recency of a feature $Fi$ within a specific time window $[t_1,t_2](t_2-t_1\le 30)$ is defined as, the number of time steps (days) from the last $F_i(t_s)$ non-zero value to the $t_2$ time step, as shown in the following equation:

$$R(F_i,t_1,t_2)=\sum_{t_s}^{t_2}(1)\ if\ t_s<t_2 els 0 \tag{1}$$

In this paper, we define the Frequency of $F_i$ as the number of time steps (in days) that are non-zero values of $[t_1,t_2](t_2-t_1\le n)$ within a specific time window, as shown in the following equation:

$$F(F_i,t_1,t_2)=\sum_{t=t_1}^{t_2}(f_i(t)>0) \tag{2}$$

In this paper, Monetary of Definition $F_i$ is defined as the sum of the values of $[t_1,t_2](t_2-t_1\le n)$, $F_i$ in a specific time window as shown in the following equation:

$$M\left(F_i,t_1,t_2\right)=\sum_{t=t_1}^{t_2}f_i(t) \tag{3}$$

In this way, in this paper, according to the definitions of $R$, $F$ and $M$ above, the features are processed by RFM respectively, and the row vector features of each user can be obtained.

2.2.3. Characteristic importance analysis model. In this paper, we use XGBoost algorithm model [40] to analyze the importance of telecommunication user characteristics, XGBoost has been improved compared with the traditional GBDT to deal with missing values and prevent overfitting. When there are missing values in the samples, XGBoost is able to automatically divide the samples with missing values, GBDT does not automatically handle missing values [41].

XGBoost model representation:

$$\hat{y}=\sum_{k=1}^{k}f_k\left(x_i\right),f_k\in F \tag{4}$$

Here $K$ is the number of trees, $F$ denotes all possible CART trees, $f$ denotes a specific CART tree, and this model consists of $K$ CART trees.

The prediction function of XGBoost is:

$$w_j^*=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda} \tag{5}$$

The objective function of XGBoost is:

$$obj\left(\theta\right)=\sum_{j}^{n}l\left(y_i,\hat{y}_i\right)+\sum_{k=1}^{k}\Omega\left(f_k\right) \tag{6}$$

The objective function is the superiority of the tree structure, the smaller the value, the better the structure, i.e., the value is a measure of how good the structure of the tth CART tree is. The objective function also contains two parts, the first part is the loss function, the second part is the regularization term, the sum of the regularization terms of $K$ tree is the regularization term here.

The regularization term of XGBoost:

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2\left(1-3\right) \tag{7}$$

A tree has $T$ leaf node, and the values of these $T$ leaf nodes form a $T$-dimensional vector $w$, $\gamma$ and $\lambda$, which is customized by XGBoost, and whose value can be set at the time of use, with the value of $\lambda$ getting bigger with the value of $\gamma$, indicating that the more you want to get a tree with a simple structure.

XGBoost's loss function, for classification problems, the commonly used loss function is the logarithmic loss function:

$$L(\theta)=\sum_{I}\left[y_i\ln\left(1+e^{-\hat{y}_i}\right)++_{-}\left(1-y_i\right)\ln\left(1+e^{\hat{y}_i}\right)\right] \tag{8}$$

### *2.3. Telecom Churn Prediction Model*

2.3.1. Improved Stacking Integration Algorithm. In this paper, we propose the Stacking integration algorithm [42] with self-sampling and accuracy weighting [43] to achieve the purpose of improving the classification effect of each base learner and providing the credibility of the base learners to the

meta-learners, so as to improve the generalization performance of the whole classification prediction system.

1) Based on the original dataset using Bootstrap self-sampling method and randomly selecting attributes according to the importance probability to get $m$ data subset, the $T$ base learners BM in the classical Stacking algorithm are replicated into $m$ copies, i.e.:

$$\begin{aligned} BM_1 &= \left(BM_{11}, BM_{12}, \ldots, BM_{lm}\right), \ldots, BM_T \\ &= \left(BM_{T1}, BM_{T2}, \ldots, BM_{Tm}\right) \end{aligned} \tag{9}$$

A subset of $m$ data is input into the base learner for $m$ training times respectively to obtain $m$ classification results for that base learner.

2) Calculate the out-of-bag error rate OOBError for each training of the base learner based on the OOB samples produced by each base learner for each sampling, i.e:

$$err_{ij} = \frac{n_{ij}}{N_{ij}} \tag{10}$$

where $i = 1, 2, \ldots, T$ ; $j = 1, 2, \ldots, m$ , $n_{ij}$ denote the number of misclassified OOB samples for the $j$ th training of the $i$ th base classifier, and $N_{ij}$ denotes the total number of OOB samples for the $j$ th training of the $i$ th base classifier.

In order to assign greater weights to the base classifiers with high prediction accuracy, the weights of each base classifier for each training are set:

$$\beta_{ij} = \frac{1 - err_{ij}}{err_{ij}} \tag{11}$$

Where $\beta_{ij}$ is the weight of the $j$ rd training of the $i$ nd base learner, it can be seen that as the error rate $err_{ij}$ gets larger, the weight of the classifier $\beta_{ij}$ gets smaller. The final prediction of the $i$ th base learner is obtained by weighted average of the predictions of the $m$ th training, i.e.:

$$\sum \beta_{ij} \cdot y_{ij}(t) \left(i = 1, 2, \ldots T; j = 1, 2, \ldots m; t = 1, 2, \ldots n\right) \tag{12}$$

The classification result $\sum \beta_{ij} y_{ij}(t) \left(i = 1, ..T; j = 1, ..m; t = 1, \ldots n\right)$ of the $T$ base learner is treated as the input to the second layer of the meta-learner (MM).

2.3.2. Predictive modeling. In this paper, the improved Stacking integration algorithm with two-layer structure of self-sampling and accuracy weighting is used to establish the telecom subscriber churn prediction model. For the Stacking integration algorithm, the selection and combination of the base learners in the first layer is crucial, which not only requires high prediction accuracy of each base learner individually, but also needs to consider the prediction effect after the combination of different base learners. Since a base learner with good prediction performance can improve the overall effect of the prediction model, the first layer of base learners should first consider choosing a model with high pre-precision, and at the same time, it is also necessary to ensure that there is differentiation between the models of each base learner, mainly because different prediction models use different principles, algorithms, data, etc., to construct models with different properties for the same data for the same learning task, and then combining them together can uncover information in all directions. Therefore, the selection of base learner models should pursue the principle of “quasi but different”, and the selection of differentiated base learners with strong prediction performance can gather the advantages of different models, so that the base learner models can complement each other's strengths and weaknesses, and ultimately improve the overall prediction performance of Stacking models.

In this paper, we study the problem of predicting the churn of telecom subscribers, the purpose is to predict whether the telecom subscribers will continue to use the service, which is a binary classification problem, and the framework of the final Stacking integration model is shown in Fig. 1. The selection of the first layer of base learners is based on the principle of "quasi but different", and the prediction performance of the machine learning model is directly related to the hyper-parameter settings of the model. Nowadays, the more widely used tuning methods include manual tuning, grid tuning, random search and Bayesian search. In this paper, the grid tuning method is chosen to determine the parameters of each base learner, the grid tuning method is similar to manual tuning, by setting the grid of possible configurations of the specified parameters in the parameter space, and establishing the prediction model for all the permutations of the parameter values in the grid, so as to evaluate and select the model parameters with the best prediction performance. According to the principle of Stacking algorithm, the input of the second layer meta-learner is generated from the prediction results of the first layer base-learner, and since the first layer has already used complex nonlinear transformation models, including decision tree, LightGBM, and MLP, three differentiated machine learning models, in order to reduce the risk of overfitting, the meta-learner of the second layer does not need to choose overly complex classifiers. Generally, simple and stable models are chosen, and in this paper, logistic regression is chosen as the meta-learner of the second layer.

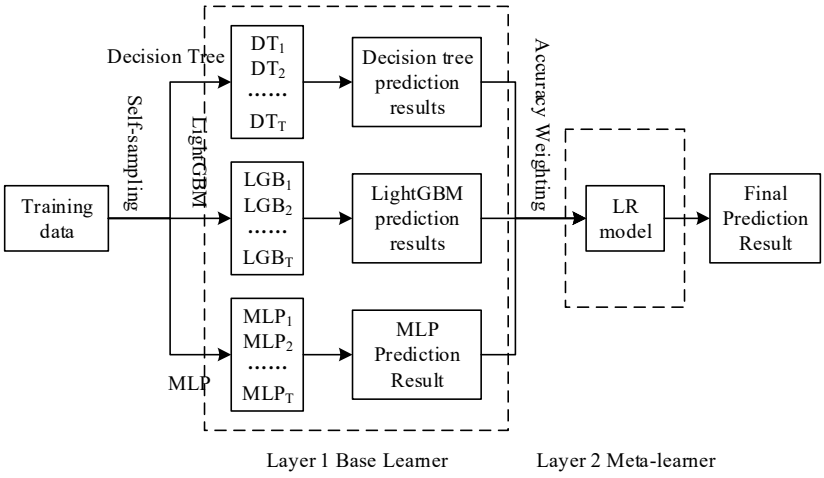


**Fig. 1.** The telecom user drain prediction model

## 3. Results and discussion

### *3.1. Results of Telecommunications Subscriber Characterization*

In this paper, two feature selection methods, namely feature importance ranking and feature relevance, are used to select feature variables from the model and statistical perspectives in a comprehensive manner. In constructing the telecom user churn prediction model, the importance of user data features was first calculated using the XGBoost model. The calculated results of the importance ranking analysis of user features are shown in Table 2. The top 5 ranked user features are whether to call (IS_CALL), whether to access the Internet (IS_GPRS), the length of Internet access (FLUX_DURA), the number of times to access the Internet (FLUX_TIMES), and whether to text message (IS_SMS), with eigenvalues of 0.182, 0.181, 0.176, 0.156, and 0.149, respectively.

**Table 2.** Analysis of characteristics

| Characteristic variable | Importance | Ranking | Characteristic variable | Importance | Ranking |
|---|---|---|---|---|---|
| IS_CALL | 0.182 | 1 | CALL_TIMES | 0.084 | 12 |

| | | | | | |
|---|---|---|---|---|---|
| IS_GPRS | 0.181 | 2 | IS_ONLY_SMS | 0.07 | 13 |
| FLUX_DURA | 0.176 | 3 | STAT_DATE | 0.067 | 14 |
| FLUX_TIMES | 0.156 | 4 | IS_LONG_CALL | 0.066 | 15 |
| IS_SMS | 0.149 | 5 | SMS_NUM | 0.057 | 16 |
| IS_ONLY_CALL | 0.135 | 6 | OUT_TAG | 0.049 | 17 |
| CALL_DURATION | 0.113 | 7 | FLUX_DOWN | 0.049 | 18 |
| IS_INITV | 0.107 | 8 | IS_ACTIVE | 0.042 | 19 |
| IS_ROAM_CALL | 0.101 | 9 | IS_INCR | 0.041 | 20 |
| FLUX_UP | 0.088 | 10 | IS_ONLY_INCR | 0.036 | 21 |
| QUERY_TIMES | 0.085 | 11 | —— | —— | —— |

After that, the correlation coefficients between these 21 features are calculated, and by evaluating the correlation degree of each feature, it can help to improve the prediction performance of the user churn prediction model. The results of the correlation analysis between each user characteristic variable are shown in Figure 2, and the names of the characteristic variables corresponding to numbers 1-21 in the figure are shown in Table 1. Among them, the correlation between the characteristic variables "whether to call", "whether to surf the Internet" and "whether to be active" were 0.096 and 0.291, respectively, and the correlation coefficients between "whether to surf the Internet" and "Internet access" and "Internet access" were 0.231 and 0.204.

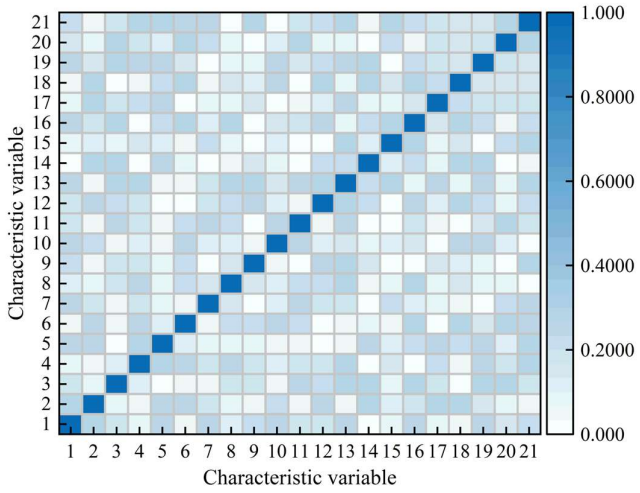


**Fig. 2.** The correlation between characteristics and target variables

### *3.2. Predictive Model Evaluation Experiments and Analysis of Results*

3.2.1. Experimental data set. In this paper, four publicly available telecom churn datasets are used to evaluate the effectiveness of the prediction methods, and the basic information of the datasets is shown in Table 3.The Cell2cell dataset is an open-source telecom churn dataset from Cell2cell, Inc. in the United States. This dataset has 52634 samples, of which there are 14885 churned subscribers. The total number of samples in this dataset is large and the number of churned subscribers is much smaller than the number of non-churned subscribers, making it a large sample dataset with an unbalanced distribution.The WAFN dataset is derived from Kaggle, a data mining competition platform, and is provided by IBM, and contains subscriber information for a fictitious telecom company that provides home phone and Internet services to customers in California in the third quarter. The dataset has 7,542 samples with 1,954 and 5,588 churned and non-churned subscribers, and the percentage of churned subscribers is 25.91%.The Duke customer dataset has a higher total number of samples and a similar number of churned and non-churned subscribers, and it is a large-sample dataset with a balanced

distribution.The DF churn dataset has a lower total number of samples and the number of churned subscribers is much The DF churn dataset has fewer samples and the number of churned users is much smaller than the number of non-churned users, which is a small sample dataset with unbalanced distribution.

**Table 3.** Information about public data sets

| Data set | Cell2cell | WAFN | Duke customer | DF churn |
|---|---|---|---|---|
| Sample size | 52634 | 7542 | 120000 | 11254 |
| Drain user | 14885 | 1954 | 59635 | 1152 |
| Non-lost users | 37749 | 5588 | 60365 | 10102 |
| Total characteristic number | 55 | 21 | 87 | 62 |
| Category number | 25 | 6 | 72 | 48 |
| Numerical characteristic number | 30 | 15 | 15 | 14 |
| Classification number | 2 | 2 | 2 | 2 |
| Loss of user ratio | 28.28% | 25.91% | 49.70% | 10.24% |

3.2.2. Assessment of indicators. Since there are experiments on datasets with unbalanced distributions, general binary classification metrics such as accuracy will be affected by negative class samples and cannot truly reflect the performance of the model, so AUC, AP and F1_score are used as the main metrics to judge the model in this experiment.

The AUC indicator can reflect the fit between the model output results and the actual results, the larger the AUC indicates that the model performs better and the output results are closer to the real results. The ROC curve has TPR as the vertical coordinate and FPR as the horizontal coordinate, and the size of the area enclosed by the ROC curve with the x-axis and y-axis is the AUC (TPR is the ratio of the number of correctly predicted positive classes to the total number of samples, and FPR is the ratio of the number of incorrectly predicted positive classes to the total number of samples). The computation of the AUC, TPR, and FPR is defined as follows:

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

$$FPR = \frac{FP}{FP + TN} \tag{14}$$

$$AUC = \frac{\sum_{i \in PositiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{15}$$

where $rank_i$ is the serial number of the ith sample, M is the number of positive samples, N is the number of negative samples, TP represents the number of samples correctly classified as positive classes, TN represents the number of samples correctly classified as negative classes, FP represents the number of samples incorrectly classified as positive classes, and FN is the number of samples incorrectly classified as negative classes.

The geometric meaning of the AP metric is the area enclosed by the PR curve with recall as the x-axis and checking accuracy as the y-axis and the coordinate axis, reflecting the ability of the classifier to correctly classify the positive classes. Precision is defined as a function of Precision for Recall, and the specific calculation of AP is defined as follows:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{16}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{17}$$

$$AP = \int_0^1 \mathrm{Pr}\,ecision(\mathrm{Re}\,call)\,dRecall \tag{18}$$

The F1 value is a new metric defined by considering both precision and recall, based on the balance point between them.The F1 value will maximize both at the same time, striking a balance.The formula for the F1 value is as follows:

$$F1 = \frac{2 \times \mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{19}$$

3.2.3. Results of model evaluation. The difference in performance between the model and the comparison model is verified through comparison experiments, and the comparison models are Support Vector Machine (SVM), CART Decision Tree, Logistic Regression (LR), Random Forest (RF), Tabnet, Deep Forest (Gcforest), and LightGBM.The experimental results of the model evaluation metrics are shown in Table 4, and the corresponding ROC curves are shown in Fig. 3, and the ( a)-(d) are the analysis results of Cell2cell dataset, WAFN dataset, Duke customer dataset and DF churn dataset, respectively. The integrated prediction model proposed in this paper achieves superior performance on all four publicly available datasets. The model's classification prediction performance ranks high in both large and small sample datasets. This is due to the multilayer perceptron's practice of iteratively updating the loss function using the stochastic gradient descent algorithm, which enables it to accurately fit the sample features. To summarize the classification prediction performance of the model in this paper, the AUC metric is second only to LightGBM (0.8416) on the WAFN dataset (0.8390), and achieves the best results on three datasets, namely, Cell2cell (0.6794), Duke customer (0.6873), and DF churn (0.6490).F1 score metrics are optimal on the Cell2cell dataset (0.1599) and the WAFN dataset (0.5933), and better than Tabnet (0.6295 and 0.1037) on the Duke customer dataset (0.6251) and the DF churn dataset (0.1024). Overall the model performance in this paper is excellent and outperforms the existing non-integrated machine learning algorithms in all three evaluation metrics, thanks to the fact that the model in this paper fully exploits the feature information in the samples, and reduces the bias by mining the rich feature information, although the model does not perform as well as the optimized gradient boosting tree model in some aspects, it is sufficient to show that the model has sufficiently good prediction performance.

**Table 4.** Model evaluation index comparison analysis results

| Data set | Model | SVM | CART | LR | Tabnet | Gcforest | RF | Light GBM | This model |
|---|---|---|---|---|---|---|---|---|---|
| Cell2cell | AUC | 0.6348 | 0.6356 | 0.6364 | 0.6411 | 0.6536 | 0.6679 | 0.6696 | **0.6794** |
| | AP | 0.3564 | 0.3929 | 0.4196 | 0.4230 | 0.4300 | 0.4312 | **0.4406** | 0.4348 |
| | F1 score | 0.0593 | 0.0608 | 0.0712 | 0.0738 | 0.0782 | 0.1111 | 0.1594 | **0.1599** |
| WAFN | AUC | 0.7825 | 0.7841 | 0.7850 | 0.8030 | 0.8095 | 0.8404 | **0.8416** | 0.8390 |
| | AP | 0.5902 | 0.6029 | 0.6097 | 0.6149 | 0.6275 | 0.6400 | 0.6405 | **0.6486** |
| | F1 score | 0.5194 | 0.5231 | 0.5316 | 0.5342 | 0.5541 | 0.5709 | 0.5738 | **0.5933** |
| Duke customer | AUC | 0.6538 | 0.6551 | 0.6587 | 0.6610 | 0.6642 | 0.6701 | 0.6755 | **0.6873** |
| | AP | 0.6019 | 0.6079 | 0.6097 | 0.6329 | 0.6329 | 0.6339 | 0.6361 | **0.6376** |
| | F1 score | 0.5873 | 0.5949 | 0.6012 | **0.6295** | 0.6096 | 0.6200 | 0.6215 | 0.6251 |
| DF churn | AUC | 0.5514 | 0.5516 | 0.5699 | 0.6249 | 0.6413 | 0.6449 | 0.6481 | **0.6490** |
| | AP | 0.2626 | 0.2797 | 0.2870 | 0.3039 | 0.3055 | 0.3162 | 0.3342 | **0.3420** |
| | F1 score | 0.0548 | 0.0579 | 0.0772 | **0.1037** | 0.0874 | 0.0989 | 0.0870 | 0.1024 |

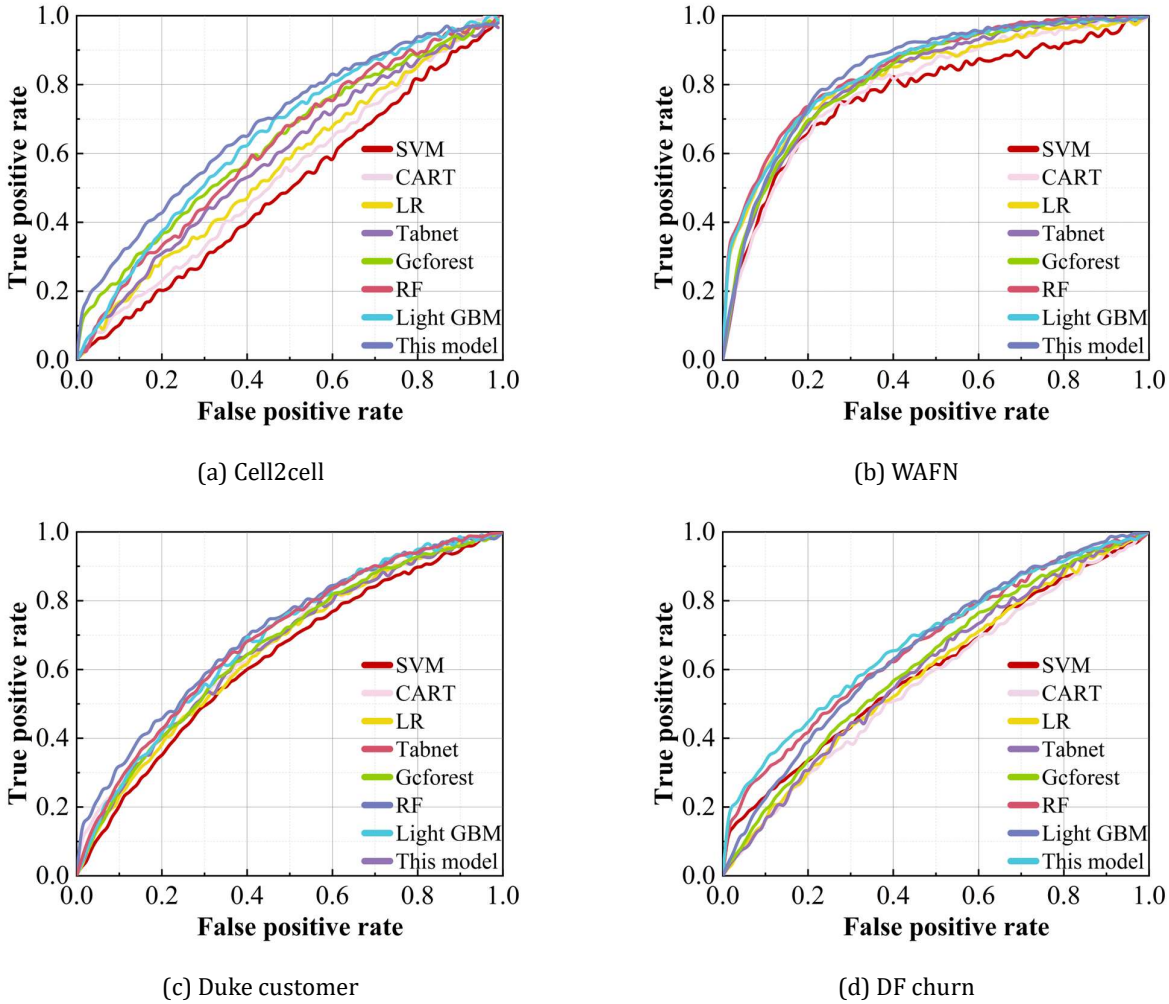


**Fig. 3.** ROC curve analysis results of the model

3.2.4. Model robustness tests. In practical applications, the optimal model not only requires that it has good generalization ability, but also requires that the model has strong robustness. Since modeling with different proportions of training and test sets may make differences in the final modeling results, this section will examine the robustness of the Stacking-based fusion prediction model in terms of adjusting the proportion of data. In this section, 90%, 80%, 70%, 60%, 50% and 40% of the sample data from the Cell2cell dataset will be selected as the training set, and the remaining samples will be used as the test set, comparing the ACC, F1 value and AUC of this fusion model under the different proportions of the training set, and the specific results are shown in Table 5. As the proportion of the training set decreases, the ACC, F1 value and AUC of the predictive model decrease, but the magnitude is small. When the training set is reduced to 40%, the AUC, ACC and F1 values of the prediction model are 0.6559, 0.4033 and 0.1349, respectively.In summary, it can be concluded that the telecom subscriber churn prediction model proposed in this paper has strong robustness.

**Table 5.** Model evaluation index of different proportional training set

| Training set | Evaluation index | | |
|---|---|---|---|
| | AUC | ACC | F1 score |
| 90% | 0.6691 | 0.4184 | 0.1481 |
| 80% | 0.6691 | 0.4137 | 0.1427 |
| 70% | 0.6684 | 0.4095 | 0.142 |

| | | | |
|---|---|---|---|
| 60% | 0.6648 | 0.4068 | 0.1399 |
| 50% | 0.6569 | 0.4065 | 0.138 |
| 40% | 0.6559 | 0.4033 | 0.1349 |

### *3.3. Practical application case results and analysis*

3.3.1. Data sources for telecommunication operators. In this paper, a randomly sampled subset of the database of telecom operator subscribers in the city of J for the year 2023 was used for analysis. The dataset contains 5500 samples, which include information about the personal information of telecom users, records of their consumption behavior, and the grade of each indicator. The data of some subscriber samples before cleaning is shown in Table 6, almost all subscribers have online behavior on the day of data collection, and the average length of online time is about 2.9h.

**Table 6.** The first part of the cleaning is according to the sample

| Serial number | Gender | Age | Whether to call | Number of calls (min) | Internet access | Internet length (h) | Text message |
|---|---|---|---|---|---|---|---|
| 1 | Female | 33 | 0 | 0 | 1 | 2.25 | 0 |
| 2 | Male | 38 | 1 | 6.5 | 1 | 1.86 | 1 |
| 3 | Male | 47 | 0 | 0 | 1 | 4.15 | 1 |
| 4 | Male | 19 | 0 | 3 | 1 | 1.93 | 0 |
| 5 | Male | 22 | 0 | 0 | 1 | 4.35 | 0 |
| 6 | Male | 31 | 0 | 0 | 1 | 2.43 | 1 |
| 7 | Female | 36 | 0 | 9 | 1 | 4.42 | 0 |
| 8 | Female | 56 | 1 | 0 | 1 | 0.37 | 1 |
| 9 | Female | 31 | 0 | 2.9 | 1 | 2.24 | 0 |
| 10 | Male | 46 | 0 | 0 | 1 | 4.11 | 0 |
| 11 | Male | 55 | 0 | 8.7 | 1 | 0.31 | 1 |
| 12 | Female | 31 | 1 | 6.7 | 1 | 2.66 | 1 |
| 13 | Male | 22 | 0 | 0 | 1 | 4.79 | 1 |
| 14 | Female | 35 | 1 | 7.1 | 1 | 4.53 | 0 |
| 15 | Female | 22 | 0 | 5.6 | 1 | 2.86 | 0 |
| 16 | Male | 35 | 1 | 0 | 1 | 4.5 | 1 |
| 17 | Female | 21 | 1 | 3.8 | 1 | 1.64 | 0 |
| 18 | Male | 19 | 0 | 9.9 | 1 | 3.67 | 1 |
| 19 | Female | 56 | 1 | 0 | 1 | 4.25 | 0 |
| …… | …… | …… | …… | …… | …… | …… | …… |
| 5500 | Female | 58 | 0 | 0 | 1 | 2.52 | 1 |

3.3.2. Effectiveness of model application. In order to verify whether the telecom subscriber churn prediction and analysis model established in this study is of practical help in solving the problem of subscriber churn in Telecom Company of City J, the next step is to carry out the application feasibility experiment by combining the model and the actual data of the company. After communicating with the business department marketing department and customer service department of this telecommunication company about the application method of the model and the experimental plan, in early November 2023, the user database information was imported into the model in bulk, and the calculation results predicted that there were 37,714 users in the city in December with a high probability of churning possibility. Therefore the customer service staff's maintenance and retention efforts in December are focused on this list of 37,714 subscribers. The results of the retention efforts for the list of high probability churn subscribers identified using the model are shown in Figure 4. The

success rate of the telco's retention response in December 2023 for prospective churn users, high probability churn users, low balance users, and low usage users is 47.85%, 51.26%, 53.24%, and 56.94%, respectively. Compared with the original method, identifying the user groups with high churn tendency with the help of prediction model and then implementing the maintenance work category classification has reduced the workload of the customer service department in screening the pre-churned users in the city's mobile network users in batch, and the customer service department can more accurately follow up the service of these users identified as having a high churn tendency to improve the maintenance efficiency and the success rate of retaining them.

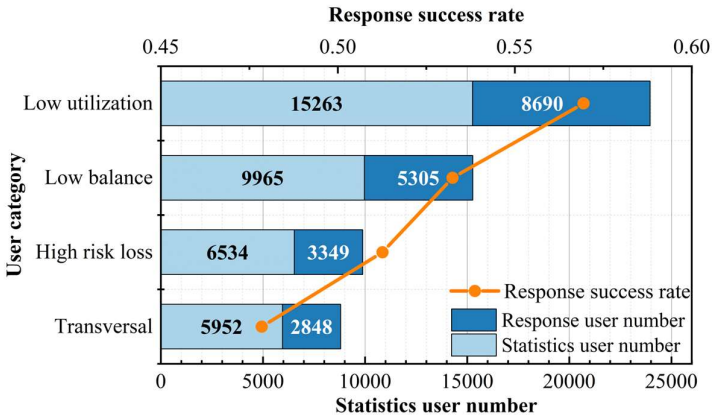


**Fig. 4.** The results were maintained in November 2023

Since the churn rate of communication subscribers is affected by external factors such as phased policies (e.g., door-opening campaigns, holiday promotions) and holiday personnel turnover, this study compares the year-on-year growth of the telecom subscriber maintenance results of the first quarter of 2023 with the results of the telecom subscriber maintenance results of the first quarter of 2024.The results of the comparison of the pre-churned subscriber maintenance of the first quarter of 2023 and 2024 are shown in Table 7 The number of subscribers counted in 2023 is the number of pre-churned subscribers selected by using the report screening method for the whole network, the number of subscribers counted in 2024 is the number of subscribers with high probability of churn calculated by the churn prediction model, and the number of respondents is the number of subscribers in the screened stock who accepted the promotional activities or policy measures offered by customer service personnel or business personnel.The results of the city telecommunication company's maintenance work for each type of churn subscriber in the first quarter of 2023 and 2024 were compared year-on-year. The average success rate of the city's telecommunications companies in maintaining all types of churned subscribers in the middle of the year was 15.43% and 50.63%, respectively. In the case of quasi-churned telecom subscribers, the success rate of response in the first quarter of 2024 ranged from 46.35% to 58.55%, while the average response rate in the first quarter of 2023 was only 15.80%. It shows that the customer service department of J city telecom company has some effect on subscriber management with the help of churn prediction model.

**Table 7.** The pre-loss user maintenance results were compared

| Statistical category | Time | Statistics user number | Response user number | Response success rate |
|---|---|---|---|---|
| Transversal user | 2023.1 | 9526 | 1771 | 18.59% |
| | 2024.1 | 6253 | 3661 | 58.55% |
| | 2023.2 | 11254 | 1950 | 17.33% |
| | 2024.2 | 7596 | 4205 | 55.36% |
| | 2023.3 | 9562 | 1097 | 11.47% |
| | 2024.3 | 5586 | 2589 | 46.35% |
| High risk loss users | 2023.1 | 19526 | 3542 | 18.14% |
| | 2024.1 | 6984 | 3965 | 56.77% |
| | 2023.2 | 23541 | 3571 | 15.17% |
| | 2024.2 | 9526 | 5069 | 53.21% |
| | 2023.3 | 16524 | 2229 | 13.49% |
| | 2024.3 | 8745 | 3945 | 45.11% |
| Low balance user | 2023.1 | 29563 | 4118 | 13.93% |
| | 2024.1 | 12542 | 6358 | 50.69% |
| | 2023.2 | 26354 | 4744 | 18.00% |
| | 2024.2 | 11254 | 5691 | 50.57% |
| | 2023.3 | 23652 | 2755 | 11.65% |
| | 2024.3 | 9856 | 4547 | 46.13% |
| Low utilization user | 2023.1 | 35264 | 6390 | 18.12% |
| | 2024.1 | 12546 | 5759 | 45.90% |
| | 2023.2 | 26595 | 2867 | 10.78% |
| | 2024.2 | 13254 | 6235 | 47.04% |
| | 2023.3 | 27754 | 5126 | 18.47% |
| | 2024.3 | 13526 | 7012 | 51.84% |

## 4. Conclusion

In this study, the XGBoost model is used to analyze the importance of telecom subscriber features, and then the construction of telecom subscriber churn prediction model is carried out based on the improved binary integration prediction algorithm Stacking. The performance of this paper's model is verified through model comparison experiments and evaluation indexes, and it is found that the F1 score of this paper's model is optimal on both Cell2cell dataset (0.1599) and WAFN dataset (0.5933), and better than the existing non-integrated machine learning algorithms on the remaining two evaluation indexes, and the model still has strong robustness in the case of the reduction of the training samples. The model is still robust when the training samples are reduced. The application of the model in J city telecom company found that the success rate of maintenance response in the first quarter of 2024 is between 46.35% and 58.55%, and the success rate of churn maintenance response is much higher than the original method, which indicates that the model is effective in assisting the telecom company in churn maintenance work. In the future, the model can be dynamically adjusted and optimized according to the changing data in each month or quarter, so that the prediction effect of subscriber churn can be further improved.

## References


[1] Kravchuk, O., Shoturma, N., Grabina, G., Myloserdna, I., Vedenieiev, V., & Shtelmashenko, A. (2022). The Information Revolution in the Post-Industrial Society: Dangers in Political Processes. Postmodern Openings, 13(4), 113-126.

[2] Kim, S. W., & Lee, M. S. (2020). Understand the current status of teaching and learning informatization and develop indicators in the 4th industrial revolution. Journal of Digital Convergence, 18(4), 67-74.

[3] Liu, C., & Li, L. (2021). Place-based techno-industrial policy and innovation: Government responses to the information revolution in China. China Economic Review, 66, 101600.

[4] Faccio, M., & Zingales, L. (2022). Political determinants of competition in the mobile telecommunication industry. The Review of Financial Studies, 35(4), 1983-2018.

[5] Meena, M. E., & Geng, J. (2022). Dynamic competition in telecommunications: A systematic literature review. Sage Open, 12(2), 21582440221094609.

[6] Alizadeh, M., Beheshti, M. T., Ramezani, A., & Bolouki, S. (2023). An optimized hybrid methodology for short-term traffic forecasting in telecommunication networks. Transactions on Emerging Telecommunications Technologies, 34(12), e4860.

[7] Kwon, H., & Kim, D. (2016). A method for churn analysis of new users of mobile games using process mining. animation, 14, 15.

[8] García, D. L., Nebot, À., & Vellido, A. (2017). Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems, 51(3), 719-774.

[9] Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. IEEE access.

[10] Senthan, P., Rathnayaka, R. M. K. T., Kuhaneswaran, B., & Kumara, B. T. G. S. (2021, April). Development of churn prediction model using xgboost-telecommunication industry in sri lanka. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-7). IEEE.

[11] Bhattacharyya, J., & Dash, M. K. (2022). What do we know about customer churn behaviour in the telecommunication industry? A bibliometric analysis of research trends, 1985–2019. FIIB Business Review, 11(3), 280-302.

[12] De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. Journal of Discrete Mathematical Sciences and Cryptography, 25(7), 1965-1985.

[13] Ribeiro, H., Barbosa, B., Moreira, A. C., & Rodrigues, R. G. (2024). Determinants of churn in telecommunication services: a systematic literature review. Management Review Quarterly, 74(3), 1327-1364.

[14] Lappeman, J., Franco, M., Warner, V., & Sierra-Rubia, L. (2022). What social media sentiment tells us about why customers churn. Journal of Consumer Marketing, 39(5), 385-403.

[15] Mahajan, V., Misra, R., & Mahajan, R. (2017). Review on factors affecting customer churn in telecom sector. International Journal of Data Analysis Techniques and Strategies, 9(2), 122-144.

[16] Amin, A., Al-Obeidat, F., Shah, B., Tae, M. A., Khan, C., Durrani, H. U. R., & Anwar, S. (2020). Just-in-time customer churn prediction in the telecommunication sector. The Journal of Supercomputing, 76, 3924-3948.

[17] Gao, W., Wang, Y., Cheng, X., Xu, L., & Song, C. (2024, April). Churn Reason Prediction Based on Telecom Operator Data. In 2024 Sixth International Conference on Next Generation Data-driven Networks (NGDN) (pp. 158-163). IEEE.

[18] Wu, F., Lyu, F., Ren, J., Yang, P., Qian, K., Gao, S., & Zhang, Y. (2023). Characterizing internet card user portraits for efficient churn prediction model design. IEEE Transactions on Mobile Computing, 23(2), 1735-1752.

[19] Banda, P. K., & Tembo, S. (2017). Application of system dynamics to mobile telecommunication customer churn management. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(3), 101-104.

[20] Do, Q. H., Doan, T. T. H., Nguyen, T. V. A., Duong, N. T., & Linh, V. V. (2020). Prediction of data traffic in telecom networks based on deep neural networks. Journal of computer science, 16(9), 1268-1277.

[21] Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14, 100342.

[22] Senthilselvi, A., Kanishk, V., Vineesh, K., & Raj, A. P. (2024, May). A Novel Approach to Customer Churn Prediction in Telecom. In 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-7). IEEE.

[23] Shukla, J., Dhariwal, N., & Marken, G. (2024, September). Churn Prediction and Customer Behavior Analysis in Telecommunications. In 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET) (pp. 1-8). IEEE.

[24] Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. Machine Learning with Applications, 17, 100567.

[25] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE access, 7, 60134-60149.

[26] Zdziebko, T., Sulikowski, P., Sałabun, W., Przybyła-Kasperek, M., & Bąk, I. (2024). Optimizing customer retention in the telecom industry: A fuzzy-based churn modeling with usage data. Electronics, 13(3), 469.

[27] Pareek, A., Arora, P., Madan, S., & Gupta, N. (2024). Telecom customer churn prediction model: Analysis of machine learning techniques for churn prediction and factor identification in the telecom sector. Journal of Information and Optimization Sciences, 45(2), 613-630.

[28] Mustafa, N., Ling, L. S., & Razak, S. F. A. (2021). Customer churn prediction for telecommunication industry: A Malaysian Case Study. F1000Research, 10, 1274.

[29] Melian, D. M., Dumitrache, A., Stancu, S., & Nastu, A. (2022). Customer churn prediction in telecommunication industry. A data analysis techniques approach. Postmodern Openings, 13(1 Sup1), 78-104.

[30] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242-254.

[31] Zhang, T., Moro, S., & Ramos, R. F. (2022). A data-driven approach to improve customer churn prediction based on telecom customer segmentation. Future Internet, 14(3), 94.

[32] Krishna, R., Jayanthi, D., Sam, D. S., Kavitha, K., Maurya, N. K., & Benil, T. (2024). Application of machine learning techniques for churn prediction in the telecom business. Results in Engineering, 24, 103165.

[33] Choudhari, A. S., & Potey, M. (2018, August). Predictive to prescriptive analysis for customer churn in telecom industry using hybrid data mining techniques. In 2018 Fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-6). IEEE.

[34] Adebiyi, S. O., Oyatoye, E. O., & Mojekwu, J. N. (2015). Predicting customer churn and retention rates in Nigeria's mobile telecommunication industry using Markov chain modelling. Acta Universitatis Sapientiae, Economics and Business, 3(1), 67-80.

[35] Hota, L., & Dash, P. K. (2021). Prediction of customer churn in telecom industry: a machine learning perspective. Computational Intelligence and Machine Learning, 2(2), 1-9.

[36] Farman, H., Khan, A. W., Ahmed, S., Khan, D., Imran, M., & Bajaj, P. (2024). An analysis of supervised machine learning techniques for churn forecasting and component identification in the telecom sector. Journal of Computing & Biomedical Informatics, 7(01), 264-280.

[37] Mahajan, V., Misra, R., & Mahajan, R. (2015). Review of data mining techniques for churn prediction in telecom. Journal of Information and Organizational Sciences, 39(2), 183-197.

[38] Fareniuk, Y., Zatonatska, T., Dluhopolskyi, O., & Kovalenko, O. (2022). Customer churn prediction model: a case of the telecommunication market. ECONOMICS-INNOVATIVE AND ECONOMICS RESEARCH JOURNAL, 10(2).

[39] Xiaoyong Lv,Zhiwu Yu,Jian Yin & Peng Liu. (2025). Seismic behaviour and restoring force model of a novel prefabricated element precast concrete frame. Journal of Building Engineering112003-112003.

[40] Meysam Alizamir,Mo Wang,Rana Muhammad Adnan Ikram,Aliakbar Gholampour,Kaywan Othman Ahmed,Salim Heddam & Sungwon Kim. (2025). An interpretable XGBoost-SHAP machine learning model for reliable prediction of mechanical properties in waste foundry sand-based eco-friendly concrete. Results in Engineering104307-104307.

[41] Zhenwei Yang,Han Li,Xinyi Wang,Hongwei Meng,Tong Xi & Zhenhuan Hou. (2025). Source identification of mine water inrush based on GBDT-RS-SHAP. Environmental Earth Sciences(4),114-114.

[42] Yi Li,Benwen Zhang,Hongming Mo,Jiancheng Hu,Yuncheng Liu & Xingqiang Tan. (2024). Unsupervised attribute reduction based on variable precision weighted neighborhood dependency. iScience(12),111270-111270.

[43] Jingli Wang & Jingxiang Gao. (2025). Stacking ensemble learning algorithm based rapid inverse modelling of copper grade using imaging spectral data. Chemometrics and Intelligent Laboratory Systems105308-105308.