Scientific and Social Research, 2025, Volume 7, Issue 5
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Comparative and Innovative Application Research of Machine Learning Algorithms in User Churn Prediction

Qiuyue Chen*

Guangzhou Huanan Business College, Guangzhou 510550, Guangdong, China

**Author to whom correspondence should be addressed.*

Abstract: Customer churn has a significant impact on enterprises as it directly affects their revenue and profit. This paper focuses on analyzing user behavior data based on a machine learning-based user churn prediction model, constructs a prediction model, and verifies its effectiveness. This aims to help enterprises better understand user behavior and prevent user churn in advance. Meanwhile, by collecting and analyzing user data, multiple machine learning algorithms are used for modeling and evaluation to determine the optimal prediction model, which is then applied to practical business scenarios. The experimental results show that the model can accurately predict user churn, providing strong support for enterprises to develop personalized marketing strategies and enhancing user retention rates and enterprise competitiveness.

Keywords: Machine learning; User churn; Prediction model; Behavior analysis; Random forest

Online publication: June 6, 2025

1. Introduction

With the transformation of the industrial structure, the market environment is becoming increasingly severe, forcing enterprises to face the serious problem of user churn. User churn not only affects the economic benefits of enterprises but may also lead to negative word-of-mouth, influencing the enterprise's brand image and market competitiveness. Traditional user churn prediction methods have problems such as low prediction accuracy and inability to handle large-scale data [1]. However, the development of machine learning technology provides new ideas and methods to solve this problem. With the development of information technology, the application of machine learning in user churn prediction models is becoming more and more widespread. Machine learning algorithms can efficiently process large-scale datasets, extract valuable patterns and rules from them. At the same time, through training and optimizing the model, machine learning algorithms can more accurately predict the risk of user churn, providing a more reliable decision-making basis for enterprises [2].

2. Relevant theories and technical foundations

2.1. Definition and influencing factors of user churn

User churn refers to the behavior of users stopping using a product or service and no longer interacting with or making purchases. There are many reasons for user churn, including poor product experience, inadequate service, fierce competition, lack of personalized service, and individual customer factors. Product quality is the basis for customer selection and retention. If a product has frequent problems, customers' trust will quickly collapse, and they will share negative experiences through various channels, affecting the purchase decisions of other potential customers. Analyzing the reasons can help effectively understand the key factors of user churn and provide a basis for subsequent model construction [3].

2.2. Overview of machine learning algorithms

Logistic regression: Logistic regression is a statistical method used to model the relationship between independent variables and a binary-classified dependent variable. Its core is to estimate parameters through maximum likelihood and output the probability value of an event occurring [4]. Logistic regression has a certain degree of interpretability, and the direction and intensity of the learning algorithm can be explained through specific symbols and data.

Decision tree: A decision tree is a classification algorithm based on a tree structure. It constructs a decision tree by dividing the dataset. The decision-tree model is easy to understand and interpret, and can handle non-linear data. However, it is sensitive to data changes and is prone to overfitting [5].

Random forest: As a machine learning algorithm, the core concept of the random forest lies in "collective wisdom", that is, combining the prediction results of multiple individual decision-tree models to achieve more accurate and robust classification or regression tasks. The random forest performs well in handling high-dimensional and non-linear data and has a high ability to resist overfitting [6].

Gradient boosting decision tree (GBDT): It is mainly used to solve regression and classification problems. Its core idea is to iteratively improve the model step by step. Each iteration corrects the errors of the previous-round model, gradually reducing the prediction error. Each weak classifier fits the residuals of the previous round, gradually reducing the value of the loss function. GBDT can automatically handle the non-linear relationships between features and has strong prediction ability and generalization performance.

2.3. Evaluation metrics

Accuracy: It is a concept widely used in scientific experiments, data analysis, and machine learning. Its core meaning is to measure correctness. The formula is: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Precision: Precision is one of the important indicators for evaluating the performance of a classification model. It represents the proportion of true positives in the prediction results of the classification model. The formula is: $precision = \frac{TP}{TP+FP}$

Recall: Recall is an important indicator for evaluating the performance of a model or system. It is usually used to measure the proportion of samples that are actually positive and are correctly identified as positive. The

formula is: $F1 = \frac{2 \times \mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$

F1-Score: The F1-score is the harmonic mean of precision and recall. The formula is: $F1 = \frac{2 \times \mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$

The F1-score aims to balance the relationship between precision and recall [7].

2.4. Data collection and preprocessing

2.4.1. Data source

This study uses the user data of an e-commerce platform, including users' basic information (such as age, gender, region, etc.), transaction records (such as purchase amount, purchase frequency, category of purchased goods, etc.), browsing behavior (such as number of views, browsing duration, browsed pages, etc.), and user feedback (such as evaluation scores, number of complaints, etc.) [8]. The data covers the user activity information over a certain period in the past, and a total of 10,000 pieces of user data were collected [9].

2.4.2. Data cleaning

Removing duplicate data: By checking for duplicate records in the dataset, duplicate user data is deleted to ensure the uniqueness of each piece of data.

Handling missing values: For features with missing values, an appropriate filling method is selected according to their data type and distribution. For categorical features, the mode can be used for filling. For example, for the age feature of users, if there are missing values, the mean value of that age group can be used for filling [10].

Handling outliers: Identify and handle outliers in the data to avoid their negative impact on model training. For example, for the purchase-amount feature of users, if a user's purchase amount is much higher than that of other users, it may be abnormal data and needs further verification and processing [11].

2.4.3. Data preprocessing

Raw data usually contains noise, missing values, etc., and needs to be preprocessed [12].

2.4.4. Feature engineering

Feature engineering is a key step in machine learning, aiming to transform raw data into features more suitable for model use. By analyzing user behavior data, features related to user churn, such as activity level, consumption amount, and login frequency, are extracted [13].

Feature selection: Feature selection methods, such as correlation analysis and chi-square test, are used to screen out feature variables related to user churn. For example, calculate the chi-square value or correlation coefficient between each feature and the user-churn label, and select features with larger chi-square values or correlation coefficients as important features.

Feature transformation: Some features are transformed to meet the input requirements of the model or improve the model's performance. For example, numerical features are standardized so that their mean is 0 and their standard deviation is 1 [14].

3. Model construction and training

3.1. Model selection

This paper selects the random forest as the basic model because of its good classification performance and anti-overfitting ability.

3.2. Model formula

Let X = (1, x2,…,) be the feature vector, and y be the target variable (churn or not, churn is 1, no churn is 0). Then the model can be expressed as: y=f(X)=RandomForest(X), where RandomForest (X) represents the prediction result of the random-forest model for the feature vector X.

The random forest is composed of multiple decision trees, and the final prediction result is determined by voting on the prediction results of each decision tree. The prediction formula for a single decision tree is: yi=g(mi;0i) where yi represents the prediction result of the i-th decision tree, g represents the decision-tree model, and 0i represents the parameters of the i-th decision tree.

The final prediction result of the random forest is: $\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ where N represents the number of decision trees in the forest.

3.3. Model training and validation

The training dataset is used to train the random-forest model, and the model performance is optimized by adjusting model parameters (such as the number of decision trees, tree depth, etc.) [15].

3.4. Logistic regression model

3.4.1. Model building

Let x be the feature vector of the user, and y be the user-churn label (y=1) indicates user churn, (y = 0) indicates no user churn. The form of the logistic regression model is:

$$p(y=1|X)=\frac{1}{1+e^{-(\omega^T X+\beta)}}$$

where *w* is the weight vector and *b* is the bias term. The parameters *w* and *b* of the model are solved by maximizing the likelihood estimation (MLE).

3.4.2. Model training

The preprocessed data is divided into a training set and a test set with a ratio of 8:2. The training set is used to train the logistic regression model, and the gradient-descent method is used to optimize the model parameters. During the training process, the value of the loss function (such as the log-likelihood loss function) of the model is recorded. When the loss-function value converges or reaches the preset number of training rounds, the training stops.

3.4.3. Model evaluation

The trained model is evaluated using the test set, and indicators such as accuracy, precision, recall, and F1-score are calculated. The evaluation results of the logistic regression model on the test set are shown in **Table 1**.

Table 1. The evaluation results of the logistic regression model on the test set

Index	Value
Accuracy	0.75
Precision	0.68
Recall	0.72
F1-Score	0.70

3.5. Decision-tree model

3.5.1. Model building

The decision-tree model selects the best splitting feature and splitting point by calculating the information gain or Gini index of each feature and constructs a binary tree. Common decision-tree algorithms include ID3, C4.5, and CART. In this study, the CART algorithm is used to construct the decision-tree model.

3.5.2. Model training

Similarly, the data is divided into a training set and a test set, and the training set is used to train the decision-tree model. During the training process, parameters such as the maximum depth of the decision tree and the minimum number of samples for splitting are set to prevent the decision tree from growing excessively.

3.5.3. Model evaluation

The model evaluation is shown in **Table 2**.

Table 2. Model evaluation

Index	Value
Accuracy	0.78
Precision	0.72
Recall	0.75
F1-Score	0.73

3.6. Random-forest model

3.6.1. Model building

The random-forest model is an ensemble-learning model composed of multiple decision trees. When constructing the random-forest model, several sub-sample sets are first randomly drawn from the training set, and a decision tree is constructed for each sub-sample set.

3.6.2. Model training

In machine learning, dividing data into a training set and a test set is a standard practice, which helps to evaluate the generalization ability of the model. As an ensemble-learning method, the random forest improves the accuracy and stability of the model by constructing multiple decision trees. Its core idea is to introduce the randomness of samples and features to avoid overfitting and improve the generalization ability of the model.

3.6.3. Model evaluation

The model evaluation is shown in **Table 3**.

Table 3. Model evaluation

Index	Value
Accuracy	0.82
Precision	0.76
Recall	0.79
F1-Score	0.77

3.7. Gradient-boosting decision-tree model

3.7.1. Model building

The gradient-boosting decision-tree model iteratively constructs multiple weak classifiers (decision trees), and each weak classifier fits the residuals of the previous round. In this study, the mean-squared-error loss function is used to construct the gradient-boosting decision-tree model.

3.7.2. Model training

The data is divided into two different samples. To ensure that the test set can objectively reflect the performance of the model in practical applications and avoid overfitting problems caused by data overlap. Parameters of the gradient-boosting decision-tree, such as the learning rate, maximum depth, and number of iterations, are set. During the training process, the loss-function value and validation error of the model are monitored to prevent overfitting.

3.7.3. Model evaluation

The model evaluation is shown in **Table 4**.

Table 4. Model evaluation

Index	Value
Accuracy	0.85
Precision	0.80
Recall	0.83
F1-Score	0.82

4. Experimental results and analysis

4.1. Experimental results

Dataset: An online-education-platform user-behavior dataset containing features such as user activity level, consumption amount, and login frequency is used for the experiment.

Experimental results: The accuracy of the model on the validation dataset.

4.2. Model performance comparison

By comparing the evaluation indicators of the four models of logistic regression, decision tree, random forest, and gradient-boosting decision tree on the test set, it can be seen that the gradient-boosting decision tree model performs well in terms of accuracy, precision, recall, and F1-score, outperforming the other three models. This indicates that the gradient-boosting decision-tree model has strong performance and generalization ability in

handling user-churn prediction problems.

4.3. Analysis of model advantages and disadvantages

Logistic regression model: The advantages are that the model is simple, easy to understand and interpret, and has a fast calculation speed. The disadvantage is that it has strict assumptions about the data, can only handle linearly separable problems, and has limited ability to handle complex non-linear relationships.

Decision-tree model: The advantages are that it can handle non-linear and high-dimensional data, and the model has strong interpretability. The disadvantages are that it is prone to overfitting, especially when the data dimension is high or the data volume is small. Also, the stability of the decision tree is poor, and its model structure may vary depending on different data-splitting methods.

Random-forest model: The advantages are high accuracy and stability, the ability to handle high-dimensional and non-linear data, and strong anti-overfitting ability. The disadvantages are that the model training time is long, it requires a large amount of computing resources, and the interpretability of the model is relatively poor.

Gradient-boosting decision-tree model: The advantages are high prediction accuracy, the ability to automatically handle complex relationships between features, and strong generalization ability. The disadvantages are that the training process is relatively complex, it requires adjusting more parameters, and it is prone to overfitting.

5. Model application and practice

5.1. User-churn prediction

The trained gradient-boosting decision-tree model is applied to practical business scenarios to predict new user data. According to the prediction results of the model, the user groups that are likely to churn are identified, and corresponding marketing strategies are adopted to retain them. For example, for users predicted to be likely to churn, personalized coupons can be pushed to them, suitable products can be recommended, or better customer service can be provided.

5.2. Formulation of personalized marketing strategies

Based on the user-churn probability predicted by the model and the user's characteristic information, personalized marketing strategies are developed. For example, for young female users with a high churn probability, promotional activities for fashion and beauty products can be pushed according to their interests and consumption habits. For elderly male users with a low churn probability, preferential information for healthcare products can be recommended.

6. Conclusion

In summary, with the continuous development of big-data and artificial-intelligence technologies, user-churn prediction research will play an increasingly important role in enterprise precision marketing and customer-relationship management. By continuously optimizing the model and expanding application scenarios, it is expected to provide more effective decision-making support and solutions for enterprises.

Funding

Guangdong Province General University Youth Innovation Talent Project, Research on Machine Learning-Based

User Churn Prediction Model and Application (2021KQNCX252)

Disclosure statement

The author declares no conflict of interest.

References

[1] André B, Lin C, 2023, A Simple Model for Predicting User Churn. Journal of Economics and Management, 2(3): 111–134.

[2] Wang Y, 2023, Applied Research on User Churn Prediction Based on Machine Learning Models, thesis, Changchun University of Technology.

[3] Liu XQ, 2023, Research on Machine Learning in the User Churn Prediction Model of TV Shopping, thesis, China University of Geosciences (Beijing).

[4] Wang DC, 2023, Research on the Communication User Churn Prediction Model Integrating GAN, thesis, Jiangxi Normal University.

[5] Zheng GX, Xu K, 2022, User Churn Prediction Model for the Live-Streaming Industry Based on High-Dimensional Time-Series Feature Supplementation. Science and Technology & Innovation, 2022(23): 56–61.

[6] Huang ZY, 2022, Research and Application of the User Churn Prediction Model for Private-Equity-Fund Wealth-Management APPs, thesis, University of International Business and Economics.

[7] Hu YF, Xiong W, Gao W, 2022, A Method for Predicting the Churn of Online-Game Users Based on the Spark Platform. Computer Engineering and Science, 44(10): 1730–1737.

[8] Huang ZX, 2022, Research on Telecommunication User Churn Prediction Model Based on Attention Mechanism, thesis, Hangzhou Dianzi University.

[9] Zheng GX, 2022, Research and Application of User Churn Prediction Model for Live-streaming APPs Based on Data-Driven. South China University of Technology.

[10] Xu SD, 2021, Research and Implementation of User Churn Prediction Model Based on MLP. Journal of Guangdong Communication Polytechnic, 20(3): 35–39 + 47.

[11] Xie J, 2021, Research on Mobile Application User Churn Prediction Method Based on Hybrid Model, thesis, Wuhan University of Technology.

[12] Liao W, 2020, Research and Application of Internet Finance User Churn Prediction Model Based on Data Mining, thesis, South China University of Technology.

[13] He AS, Wang L, Huang S, 2020, Design and Implementation of User Churn Early-Warning System Based on Big Data. Radio & TV Information, 27(7): 93–94.

[14] Wen RJ, 2020, Research on the Application of Deep Learning Models in Bank User Churn Prediction, thesis, Chongqing University of Posts and Telecommunications.

[15] Wu HF, 2020, Research on User Portrait and Churn Prediction Model of UFIDA Based on Data Mining Methods, thesis, Dongbei University of Finance and Economics.