Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/71/20241622

Research on classified prediction algorithm of platform customer value based on "tree" concept

Zening Cheng[1,*], Shuangbin Tao[2], Meiling Ge[3], Weiwei Xing[3], Tianyu Li[1]

[1]Baoding University of Technology, Baoding, China
[2]Shenyang Normal University,China
[3]Langfang Yanjing Vocational Technical College

[*]chengzening@sina.com

Abstract. The concept of "trees," representing decision trees, random forests, and XGBoost algorithms, has gained increasing attention in the field of classification prediction in recent years. These tree-based machine learning algorithms have been widely applied in the e-commerce sector. This paper discusses the decision tree, random forest, and XGBoost algorithms individually, and extracts useful classification rules from large volumes of online customer data to provide intelligent decision support for computer network customer management. The study finds that algorithms related to the "tree" concept can integrate a customer's current value (such as purchasing behavior) with potential value (such as customer interest inferred from online reviews) through detailed classification criteria, thereby constructing a customer value measurement model. The evolution from decision trees to random forest and XGBoost algorithms has effectively promoted research on customer value hierarchy prediction, improving the efficiency of data mining in computer networks.

Keywords: decision tree, random forest, XGBoost, machine learning, computer networks, hierarchical prediction.

1. Introduction

With the development of intelligent data processing methods in computer networks, the use of machine learning algorithms in computer network data management has become increasingly necessary. In the era of big data, leveraging the secondary value of network data has become key to the growth of e-commerce enterprises. For example, enhancing network data management can guide businesses to improve products, services, and communication techniques, thereby enhancing user experience [1]. However, there are several challenges in improving the intelligence of computer network data management through machine learning algorithms. First, the issue of data authenticity always affects the predictive accuracy of machine learning models, preventing the algorithms from realizing their full potential. Second, the complex and variable nature of data attributes in computer networks makes it difficult to establish a uniform "processing" standard, making data preprocessing a challenging aspect of developing intelligent data production. Third, the immaturity of related machine learning classification algorithms is another issue. Data scientists currently focus on basic mathematical principles, but research on how to process and apply these principles is still insufficient. In fact, some social science research methods offer ideas for systematically collecting and organizing computer

DOI: 10.54254/2755-2721/71/20241622

network data. This paper aims to predict customer value hierarchies using three major tree-based machine learning algorithms: decision trees, random forests, and XGBoost. The paper also discusses classification prediction issues that arise during the evolution from decision trees to XGBoost algorithms.

2. Research Methods

2.1. Decision Trees and Random Forests

Decision trees are the earliest classification prediction algorithms based on the "tree" concept. The random forest algorithm, an ensemble learning method based on decision trees [2], uses the Bagging technique [3] to construct multiple "good and different" [4] decision tree classifiers in parallel. Random forests are mainly used to address classification and regression problems [5] and are particularly effective in handling large-dimensional customer behavior data. As a classifier composed of multiple decision trees, the output class is determined by the mode of the classes output by the individual trees. Random forests can handle classification and regression problems as well as dimensionality reduction, and they offer better predictive and classification performance compared to decision trees [6].

2.2. XGBoost

XGBoost is an optimized distributed gradient boosting library designed for efficiency, flexibility, and portability. It implements machine learning algorithms under the Gradient Boosting framework, offering parallel tree boosting (also known as GBDT or GBM) that can quickly and accurately solve many data science problems. The same code runs in major distributed environments (Hadoop, SGE, MPI) and can address problems beyond billions of examples [7]. XGBoost [8] is an improvement on the gradient boosting algorithm, using Newton's method to solve the extremum of the loss function and expanding the loss function to the second order using Taylor series. Additionally, a regularization term is added to the loss function. Compared to traditional GBDT algorithms, XGBoost can perform second-order Taylor expansions, supports custom loss functions [9], and adds penalty terms to reduce model variance, effectively preventing overfitting [10]. Moreover, XGBoost supports data parallelization: each feature is pre-sorted by feature value and stored in a block structure, allowing multi-threaded parallel search for the best split point for each feature during node splitting, significantly improving training speed [11].

3. Experimental Design

This paper first collects and preprocesses customer value-related information from online platforms and simulates platform owner requirements to capture user behavior data using web scraping technology. Decision tree, random forest, and XGBoost models are then constructed based on the following equation (1). General data includes gender, response to feedback, age group, promotions such as red envelopes and coupons, patient customer service, lotteries, and small gifts—indicators related to predicting user value. Additionally, based on the collected data, this study categorizes customer value as high or low, assigning a virtual classification label of 0 or 1. This process is essential for training the machine learning models.

$$Gini(D) = 1 - \sum_{k=1}^{100} (P_k)^2 \tag{1}$$

In this study, K represents the 100 data groups in the experimental dataset D, and *Pk* denotes the proportion of samples in D that belong to the kth class out of the total samples.

Regarding the specific training process of the random forest model, it begins by evaluating each feature using the information gain formula to select the optimal feature for constructing the decision tree. Then, based on the chosen feature, the dataset is divided into multiple subsets, with each subset corresponding to a node in the decision tree. The above steps are then recursively applied to each subset

DOI: 10.54254/2755-2721/71/20241622

to construct a complete decision tree. Finally, by integrating multiple decision trees, the random forest's generation formula is used to predict or classify new samples. It is noteworthy that in a random forest, the selection of features is done randomly, with each feature's selection probability calculated based on its importance.

Additionally, XGBoost was used in this study to compare the prediction results with those of the random forest model.

4. Experimental Results

Figure 1 shows the structure of the decision tree using the experimental data, where the internal nodes present the specific split conditions of the branching features, that is, the data is divided based on a particular split value of a feature.

Figure 1. Decision Tree Structure for Online Customer Value Hierarchy

DOI: 10.54254/2755-2721/71/20241622

As shown in Figure 1, the Gini information entropy is used to determine which feature to split. The sample category distribution represents the number of samples belonging to each classification group within that node.

Table 1. Random Forest Model Parameters

Parameter name	Parameter value
Training time	0.092s
Data segmentation	0.7
Data shuffle	No
Cross validation	No
Node evaluation criterion	gini
Number of decision trees	100
There is back sampling	true
Out-of-pocket data test	false
Maximum Feature Proportion Considered in Partition	auto
Minimum number of samples of internal nodes	2
Minimum sample number of leaf nodes	1
Minimum Weight of Samples in Leaf Nodes	0
Maximum depth of a tree	10
Maximum number of leaf nodes	50
Threshold of impurity of node division	0

According to formula (2), Figure 2 illustrates the importance of input features in predicting customer value using the random forest algorithm.

$$S = \sum_{i=1}^{s} importance(i) \tag{2}$$

where SSS represents the sum of the selection probabilities of all features in the feature set, and importance(i) denotes the selection probability of feature i. Feature weights indicate the significance of each feature's contribution to the model, with their sum equaling 1.

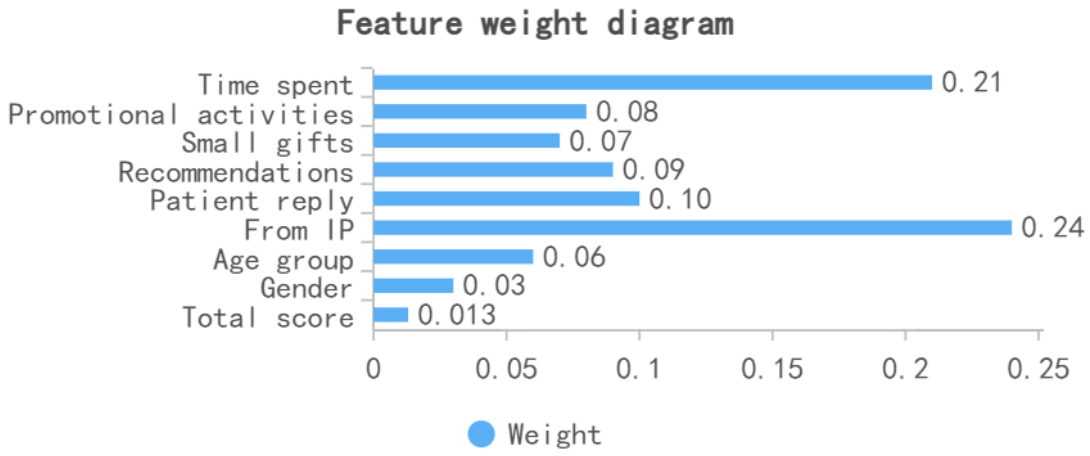

Figure 2. Feature Importance in Random Forest Classification Prediction

DOI: 10.54254/2755-2721/71/20241622

As shown in Figure 2, the proportion of the feature "IP" is 23.72%, making it the most influential feature with a critical role in model construction. The feature "time spent" has a proportion of 20.65%, indicating its significant importance in model building. The proportion for "total score" is 13.09%; "customer service patience" is 9.89%; and "respect for improvement suggestions" is 8.55%. Together, these five features account for 75.90% of the importance. The remaining features—"red envelopes," "coupons and promotions," "small gifts," "age group," and "gender"—have proportions of 7.87%, 7.40%, 6.30%, 2.53%, respectively.

Table 2. Evaluation Results of Random Forest Algorithm Classification Prediction

	Accuracy	Recall rate	Precision rate	F1
Training set	1	1	1	1
Test set	0.667	0.667	0.667	0.667

Table 2 presents the classification evaluation metrics for the training set and test set, which quantify the effectiveness of the Random Forest model's classification on the training and test data.

Accuracy: The proportion of correctly predicted samples out of the total samples. A higher accuracy indicates better performance.

Recall: The proportion of actual positive samples that were correctly predicted as positive. A higher recall is better.

Precision: The proportion of predicted positive samples that are actually positive. A higher precision is better.

F1: The harmonic mean of precision and recall. Precision and recall influence each other; while it is ideal to have both high, in practice, often one is high while the other is low. If both need to be considered, the F1 score is used.

Table 3 presents the various parameter configurations of the XGBoost algorithm and the model training duration.

Table 3. Parameter Settings of the XGBoost Algorithm

Parameter name	Parameter value
Training time	1.26s
Data segmentation	0.7
Data shuffle	No
Cross validation	No
Base learner	GBTree
Number of base learners	100
Learning rate	0.1
L1 regular term	0
L2 regular term	1
Sample characteristic sampling rate	1
Tree feature sampling rate	1
Node characteristic sampling rate	1
Minimum Weight of Samples in Leaf Nodes	0
Maximum depth of a tree	10

Figure 3 illustrates the feature importance in the XGBoost classification model, which compares the average values of the dataset. The feature importance indicates the degree of contribution each feature makes to the model, with the sum of all feature importances equaling 1. The weights are as follows: time used accounts for 13.80%, IP source accounts for 13.69%, red envelopes and coupon promotions account for 12.77%, customer service patience accounts for 10.98%, lottery accounts for 10.21%, and

DOI: 10.54254/2755-2721/71/20241622

small gifts account for 9.67%. The combined weight of these six features totals 71.12%. The remaining four features, respect for feedback, total score, gender, and age group, account for 9.65%, 8.27%, 5.56%, and 5.40%, respectively.

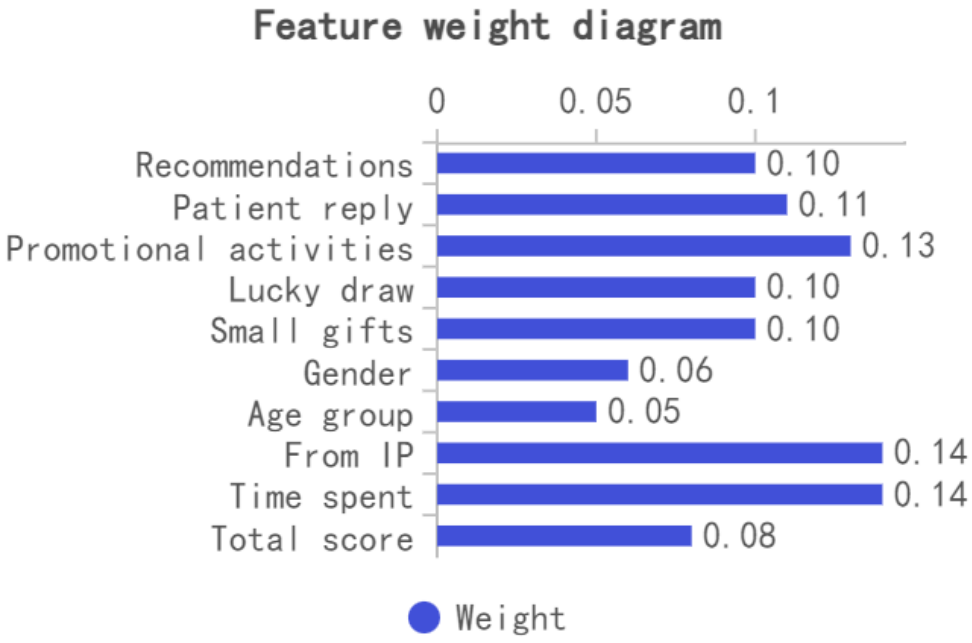

Figure 3. Feature Importance in XGBoost Classification Model

5. Conclusion

In the future, the study of platform customer value using Random Forest and XGBoost classification algorithms in computer networks will witness numerous innovations and opportunities. By collecting representative data from platform store customers and employing advanced feature selection using "tree" concept machine learning algorithms, we can gain deeper insights into customer value hierarchies. These "tree" concept machine learning models not only analyze customer value hierarchies but also perform classification and regression analysis on store operation data. Based on these analyses, businesses and store owners in e-commerce can develop marketing strategies and product recommendations that are better aligned with their specific conditions. This enables a precise match of marketing strategies with target audiences, adapting to the rapid changes in the market and user behavior.

References

[1] Maedeh Haghirad,Shahin Heidari,and Hojat Hosseini."Advancing personal thermal comfort prediction: A data-driven framework integrating environmental and occupant dynamics using machine learning."Building and Environment 262.(2024):111799-111799.

[2] Breiman L. Random forest[J]. Machine Learning,2001, 45: 5-32.

[3] Dudoit S,Fridlyand J.Bagging to improve the accuracy of a clustering procedure[J].Bioinf ormatics,2003,19(9) : 1090- 1099.

[4] Li G, Cherukuri K A. Embrace open-environment machine learning for robust AI.[J].Nati onal science review,2024,11(8):nwad300.

[5] Liaw A ,Wiener M. Classification and regression by Ran- domForest[J]. R News,2 002,2(3) : 18-22.

[6] Xu Y, Lin K, Hu C, et al.Interpretable machine learning on large samples for supporting runoff estimation in ungauged basins[J].Journal of Hydrology,2024,639131598-131598.

DOI: 10.54254/2755-2721/71/20241622

[7] Mouskeftara T ,Deda O ,Liapikos T , et al.Lipidomic-Based Algorithms Can Enhance Prediction of Obstructive Coronary Artery Disease. [J].Journal of proteome research,2024,

[8] Alzakari A S ,Menaem A A ,Omer N , et al.Enhanced heart disease prediction in remote healthcare monitoring using IoT-enabled cloud-based XGBoost and Bi-LSTM[J]. Alexandria Engineering Journal,2024,105280-291.

[9] Y GQiu, JZhou, Khandelwal M, et al. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost algorithms to predict blast-induced ground vibration[J]. Engineering with Computers, 2021, 38: 4145-4162.

[10] Torlay L, Perrone-BertolottiM, Thomas E, et al. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain informatics, 2017, 4(3): 159-169.

[11] Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree[J]. Information Sciences, 2021, 572: 522-542.