information

MDPI

Article

Machine-Learning-Based User Position Prediction and Behavior Analysis for Location Services

Haiyang Jiang [1], Mingshu He [2,*], Yuanyuan Xi [3] and Jianqiu Zeng [1]

[1] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China; jhy@bupt.edu.cn (H.J.); zengjianqiu@bupt.edu.cn (J.Z.)
[2] School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-10044 Stockholm, Sweden; yuaxi@kth.se
* Correspondence: hemingshu@bupt.edu.cn; Tel.: +86-130-0123-0413

Abstract: Machine learning (ML)-based methods are increasingly used in different fields of business to improve the quality and efficiency of services. The increasing amount of data and the development of artificial intelligence algorithms have improved the services provided to customers in shopping malls. Most new services are based on customers' precise positioning in shopping malls, especially customer positioning within shops. We propose a novel method to accurately predict the specific shops in which customers are located in shopping malls. We use global positioning system (GPS) information provided by customers' mobile terminals and WiFi information that completely covers the shopping mall. According to the prediction results, we learn some of the behavior preferences of users. We use these predicted customer locations to provide customers with more accurate services. Our training dataset is built using feature extraction and screening from some real customers' transaction records in shopping malls. In order to prove the validity of the model, we also cross-check our algorithm with a variety of machine learning algorithms. Our method achieves the best speed–accuracy trade-off and can accurately locate the shops in which customers are located in shopping malls in real time. Compared to other algorithms, the proposed model is more accurate. User preference behaviors can be used in applications to efficiently provide more tailored services.

Keywords: location-based service; machine learning; XGBoost; behavior analysis

check for updates

Citation: Jiang, H.; He, M.; Xi, Y.; Zeng, J. Machine-Learning-Based User Position Prediction and Behavior Analysis for Location Services. *Information* **2021**, *12*, 180. https://doi.org/10.3390/info12050180

Academic Editor: Hak-Seon Kim

Received: 12 April 2021
Accepted: 21 April 2021
Published: 22 April 2021

1. Introduction

With the development of Internet technology, people's expectations of services are becoming increasingly demanding. Many machine learning (ML)-based methods are used in user behavior analysis, hobby analysis, etc. [1]. People are trying to obtain more information on the characteristics of users from historical data and to build models to analyze these data [2,3]. In this study, we analyzed users' behavior trends using the user's transaction and related location information. At present, indoor positioning technology is mature and can be used to accurately locate users in shopping malls [4]. However, if only limited information is provided, such as WiFi and payment information, it will be difficult for shopping malls to obtain the users' locations. In this work, we used efficient machine learning algorithms to obtain valuable information from customers' transaction data. The machine learning generic algorithm can derive different classification information and algorithms according to different scenarios, increasing the accuracy and efficiency of the analysis of customers' transaction information. The analysis results provide the basis for the future advances in user behavior analysis.

For outdoor positioning, global position system (GPS) is the most widely used satellite navigation and positioning technology in the world. Its popularity and development have enabled people to obtain accurate and reliable geographical location information in outdoor environments. However, in indoor environments, buildings hinder the satellite signals, which leads to poor indoor positioning quality. To provide accurate indoor location services, there are many indoor wireless positioning technologies currently available, such as infrared positioning [5–7], radio frequency identification (RFID) positioning [8–11], ultrasonic positioning [12–14], WiFi positioning [15–18], Bluetooth low energy (BLE) [19,20], magnetic fields [21], etc. WiFi-based indoor positioning is currently the mainstream method because, with the rapid development of mobile devices, WiFi chips are widely used in various types of consumer terminals and, due to the faster network speed, wireless local area networks cover many cities. These advantages make WiFi-based indoor positioning one of the most effective methods currently. WiFi-based positioning includes the following methods: time of arrival/time difference of arrival (TOA/TDOA) [22], angle of arrival (AOA) [23], and received signal strength indication (RSSI) [24]. Due to the short propagation distance and fast attenuation of WiFi signals and due to shops in shopping malls being too closely distributed to use the TOA/TDOA and the AOA methods for positioning, based on the original data in this work as well as on actual environmental considerations of shopping malls, we show that indoor positioning using RSSI technology is more suitable. RSSI technology mainly compares various SSID data (including field strength and signal-to-noise ratio) received by the wireless network card with SSID data collected in advance at various locations indoors and then approximates the location of customers. However, in this work, the managers of the shopping malls did not input the cost to pre-collect the SSID data of each location in every shop; therefore, determining how to use the large amount of customer transaction record data containing location information to replace the pre-acquired data is another challenge in this study.

In addition to positioning problems, machine learning algorithms are widely used in WiFi-based positioning. K-nearest-neighbor (KNN) [25], a classification algorithm that is mature in theory, is one of the simplest machine learning algorithms. It has high classification accuracy. However, in the case of sample imbalance, KNN is less effective. The boost algorithm is one of the hottest branches in machine learning. Ensemble learning [26] is a method used to integrate multiple learners to solve a class of problems [27]. Over the years, researchers have performed many optimizations for ensemble learning algorithms, from the initial boost algorithm [28], which has complex conditions, to AdaBoost [29], which has simple conditions and good classification performance. GRNN-SGTM [30] is another ensemble model for prediction. The XGBoost algorithm [31] is one of the best improvements to ensemble learning. For ensemble learning algorithms, parameter adjustment has a considerable impact on the results. Therefore, parameter adjustment and algorithm usage were all considered in this work. There are also some deep learning (DL)-based methods used in user behavior analysis [32,33]. However, to better explain the business significance, we did not apply them.

In this paper, we propose a novel improvement to the XGBoost algorithm that is suitable for the conditions in this work to accurately locate consumers, showing in which shop they are located, so that we can provide more precise services. The contributions of this work can be summarized as follows:

- We propose a two-layer XGBoost algorithm, stacking multi-class XGBoost algorithm, and two-class XGBoost algorithm that can produce more accurate predictions.
- We data mine real consumers' transaction records, process and build feature engineering, and improve the prediction results.
- According to the model results, we propose some user behavior analysis methods based on location services.

The remainder of this paper is arranged as follows: The data are described and processed in Section 2. In Section 3, we introduce the detailed structure of the model and the the overall proposed study. Section 4 describes the experiments and an analysis of the results produced by the proposed model. Section 5 introduce some service and behavior analysis methods. Section 6 summarizes the study and provides our conclusions.

2. Data Analysis and Processing

Nowadays, mobile payment technology is becoming increasingly mature. Most customers use mobile payments instead of cash. When customers use mobile payments in a shopping mall, some manufacturers capture the transaction record of customers, which contains their mobile device information and current environmental information surrounding the user at each payment. The dataset used in this study contained this consumer information.

2.1. Data Description

In this work, we collected 1,138,015 transaction records from different customers at different shops, which were divided into a training dataset and a test dataset (the dataset is available at [34]). The training dataset contained 849,505 transaction records for the first three weeks of a month, and the test dataset contained the 288,510 transaction records of the last week of the month. Note that all data were anonymized, and necessary desensitization measures such as biased sampling and filtering were applied to ensure the privacy of the customers and sellers.

In Table 1, shop_id and mall_id represent the unique identifier of a shop and a shopping mall, respectively; category_id is the shop type; longitude and latitude indicate the shop location by GPS positioning, which is not very accurate; and price is transaction per person in this shop. In Table 2, time_stamp is the mobile payment moment and wifi_infos is the unique identifier of the bssid of the WiFi that can be searched by this device, the signal strength of this WiFi, and a binary number representing whether this device is connected to the access point (AP) of a WiFi connection.

Table 1. Part of the shops' information.

	Shop_id	Category_id	Longitude	Latitude	Price	Mall_id
0	s_26	c_4	122.346736	31.833507	57	m_690
1	s_133	c_6	121.134362	31.197511	58	m_6587
2	s_251	c_38	121.000505	30.907667	34	m_5892
3	s_372	c_30	119.864982	26.659876	44	m_625
4	s_456	c_26	122.594243	31.581499	44	m_3839

Table 2. Part of the transaction records.

	User_id	Shop_id	Time_Stamp	Longitude	Latitude	Wifi_infos
0	u_376	s_2871718	2017-08-06 21:20	122.308219	32.088040	b_6398480 | -67...
1	u_376	s_2871718	2017-08-06 21:20	122.308162	32.087970	b_6396480 | -67...
2	u_1041	s_181637	2017-08-02 13:10	117.365255	40.638214	b_8006367 | -78...
3	u_1158	s_609470	2017-08-13 12:20	121.134451	31.197416	b_26250579 | -73...
4	u_1654	s_3816766	2017-08-25 10:50	122.255867	31.351320	b_39004150 | -66...

Segmenting non-orthogonal data not only improves the accuracy of the prediction results but also significantly shortens the runtime of the algorithm. In order for our algorithm to build a more accurate predictive model, we simply classified the data based on the raw dataset. Through experimental observation, we found a clear distinction between shop locations and customer locations between different shopping malls. Per our statistics, there are a total of 97 shopping malls in the raw dataset, which means we manually classified the data using these 97 locations first. In this way, all data analysis and machine learning algorithms were based on the 97 location classification.

2.2. Feature Extraction

In Section 2.1, we introduced the contents of the raw dataset, from which we extracted orthogonally uncorrelated features for supervised learning algorithms. The feature extraction largely determines the accuracy of the final prediction to improve the prediction accuracy; in addition to the direct features that can be found in the raw dataset, we built many statistical features [35]. For direct features, we took the longitude and latitude of the shops provided by a GPS, the shop types, and the transaction per person in shops from the shop information. We used the unique identifier of customers and shops, the longitude and latitude of customers provided by GPS, and the WiFi information from the transaction records. In addition, we built many statistical features using some direct features as follows:

WiFi signal strength: for each transaction record, the WiFi information received by the consumer at the payment moment is recorded. The WiFi information corresponds to the 'wifi_infos' column in Table 2 and includes the WiFi signal strength. We extracted the WiFi signal strength from the 'wifi_infos' column in Table 2 as multi-dimensional features. Because the WiFi received by the consumers at the payment moment is always more than one and not equal between different transaction records, if the number of WiFi connections in one transaction record was less than another transaction record, we make up the features with NULL for the same number of feature dimensions between different transaction records.

The average WiFi signal strength from each transaction record was counted as a one-dimensional feature.

The standard deviation of WiFi signal strength from each transaction record was counted as a one-dimensional feature.

The maximum WiFi signal strength from each transaction record was counted as a one-dimensional feature.

The minimum WiFi signal strength from each transaction record was counted as a one-dimensional feature.

The quartile of WiFi signal strength: the first, second, and third quartiles of the WiFi signal strength from each transaction record were counted as three-dimensional features.

The number of WiFi connections from each transaction record was counted as a one-dimensional feature.

The number of shops around the WiFi connection: according to shop_id and wifi_infos from each transaction record, the number of shops where the WiFi had the strongest signal strength in each record was counted as a one-dimensional feature.

The number of WiFi connections around the shops: according to "shop_id" and "wifi_infos" from each transaction record, we picked the strongest WiFi signal from a record and counted the number of shops where this WiFi appears as a one-dimensional feature.

The transaction per person of WiFi: according to the statistics of the number of shops around the WiFi, we selected the shop with the most occurrences of WiFi connection and counted the shops with the most occurrences of WiFi connection corresponding to each record and used the corresponding transaction per person. The average value of these transactions per person was counted as a one-dimensional feature.

The count of transaction: according to user_id and shop_id from each transaction record, the number of times the consumer consumes a product in the shop was counted as a one-dimensional feature.

As can be seen from the above statistical characteristics, we counted a large number of WiFi-related features. GPS positioning often has large errors in precise indoor positioning. In actual environments, WiFi features are especially important compared with the latitude and longitude features, not only because of the large number WiFi connections in shopping malls but also the different signal strengths of WiFi for different customer locations, which is more effective than latitude and longitude features in location prediction.

2.3. Feature Screening

After feature extraction, we found that many features were relatively independent, which were not universal and could not help with our predictions, and that these redundant features would affect the runtime of the algorithms. Therefore, it was necessary to screen out redundant features. Although accurate prediction cannot be achieved by GPS data alone, GPS data can roughly represent the positional relationship between shops and customers. To verify this conclusion, we constructed a relationship diagram between the latitude and longitude of a shop in a shopping mall and the latitude and longitude of customers with the transaction records in this shop, as shown in Figure 1, where the crosses indicate the latitude and longitude of the shop and the circles denote the latitude and longitude of the customer in each transaction record. The latitude and longitude of most customers are concentrated together near the latitude and longitude of the shop. This shows that the latitude and longitude feature roughly indicates the positional relationship between shops and customers.

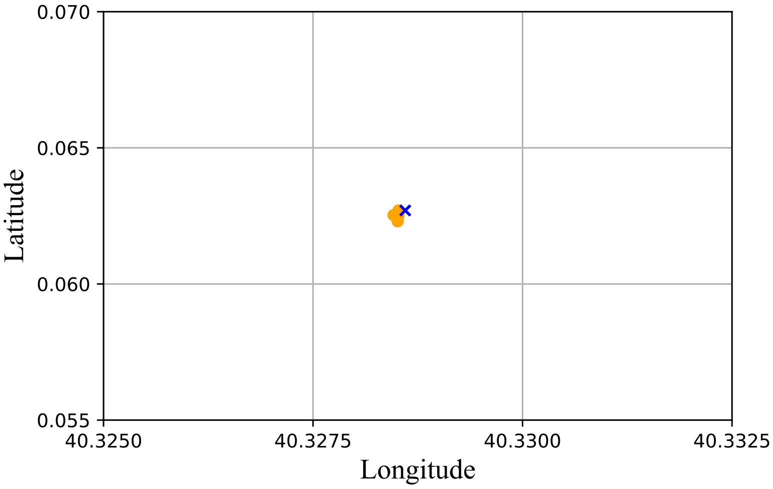

Figure 1. Diagram of the relationship between a shop and customer latitude and longitude.

However, the situation described in Figure 2 appears in this shopping mall. In Figure 2, the latitude and longitude of a consumer is far from the latitude and longitude of the shop and most latitudes and longitudes of other customers. This circle in Figure 2 is an outlier. According to our statistics, almost every shopping mall has such a relationship with the latitude and longitude of a shop and its customers. The reason for the appearance of this outlier is that the GPS information acquired by a mobile device when the customers are indoors is inaccurate. Such outliers will likely cause errors in the shop prediction. Therefore, we deleted the transaction record if the Euclidean distance of the latitude and longitude of a customer and the latitude and longitude of the shop exceeded a certain threshold.

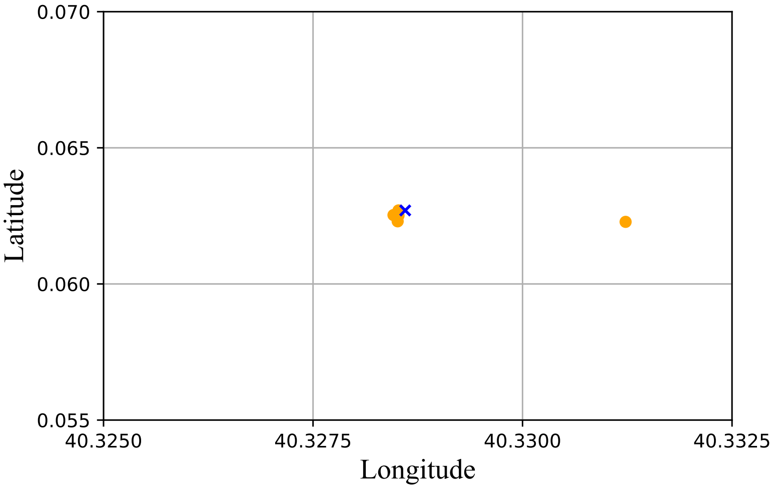

Figure 2. Diagram of the relationship between a shop and customer latitude and longitudes including outliers.

The statistics showed that a large amount of WiFi connections in the raw dataset, which contains more than one million consumer records, only appeared several times in a whole month. These WiFi connections are obviously not WiFi connections inside the shop or inside the shopping mall. We defined this type of WiFi with less than 20 transaction records in a month as a personal hotspot that travels with individual customers. Since the signal strength of a personal hotspot is based on the customer's mobile device and not the shop, if we used these personal hotspots as features, the prediction results were likely to be wrong. The number of personal hotspots was very large, which slowed the runtime of the algorithms. Therefore, we removed WiFi connections with less than 20 transaction records in a month to improve both the run speed of the algorithms and the accuracy of the prediction.

3. Methodology

In this section, we introduce the XGBoost architecture, discuss our improvements to it, and provide a comparison of its performance with that of other algorithms.

3.1. XGBoost Model

The XGBoost model [36] is composed of a number of classification and regression trees (CARTs) [37]. For a single CART, its parameters include (1) the structure of the CART and (2) the leaf node value. Therefore, the prediction of XGBoost can be expressed as follows:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \tag{1}$$

where $f_k \in F$ and K are the number of CARTs in an XGBoost model and $f_k(x_i)$ is the leaf node value of sample x_i on the kth-specific CART. Each sample corresponds to a leaf node on each CART. The prediction of this sample is equal to the sum of the corresponding leaf node values of the sample on all CARTs in the model. The objective function of XGBoost is defined as follows:

$$obj(\theta) = \sum_{i}^{n} l(y_i\, , \widehat{y_i}) + \sum_{k=i}^{K} \Omega(f_k). \tag{2}$$

The first term is the loss function of the model, which can be customized. The second term is a regular term.

Training XGBoost uses the method of addition training, which means optimizing a CART first and then optimizing the other. The training process can be expressed as follows:

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) \ + \ f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\cdots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) \ = \ \hat{y}_i^{(t-1)} + f_t(x_i).
\end{aligned} \tag{3}$$

Therefore, the objective function can be transformed as follows:

$$obj^{(t)} = \sum_{i=1}^{n} l\Big(y_i\, , \widehat{y_i}^{(t-1)} + f_t(x_i)\Big) + \Omega(f_t) + const, \tag{4}$$

when using addition training. Then, Taylor series expansion is applied for Equation (4). The objective function can be described as follows:

$$obj^{(t)} = \sum_{i=1}^{n}\left[l\left(y_i\,,\,\hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) + const, \tag{5}$$

where

$$g_i = \partial_{\hat{y}_i^{(t-1)}}\, l\left(y_i\,,\,\hat{y}_i^{(t-1)}\right) \tag{6}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i\,,\,\hat{y}_i^{(t-1)}\right) \tag{7}$$

Additionally, $obj^{(t)}$ can be minimized to the following:

$$min\sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t). \tag{8}$$

Then, $f_t(x)$ and the regular term in (8) can be defined as follows:

$$f_t(x) = w_{q(x)} \tag{9}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2, \tag{10}$$

where $q: \;\; R^d \rightarrow 1,2,\ \ldots, T$ represents the mapping of samples to leaf nodes and T is the number of leaf nodes on a CART. Substituting (9) and (10) into (8), the optimization task becomes the following:

$$min\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T, \tag{11}$$

where I_j is a set composed of all samples mapping the jth leaf node. For the result of (10), let $obj'^{(t)} = 0$. The solution of (10) is as follows:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{12}$$

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{13}$$

where $G_j = \sum_{i\in I_j} g_i, H_j = \sum_{i\in I_j} h_i.obj^*$ is the minimum of the objective function, which can be used to measure the quality of the structure of CARTs. The smaller the obj^*, the better the structure of the CART. Because G_j and H_j are only determined by the structure of CART, as long as the structure of the CART is determined, the optimization parameters can be determined. According to the above introduction, we observed that XGBoost provides two advantages over other algorithms: (1) the run time is fast because of its parallel calculated parameters and (2) the loss function can be customized to only need to satisfy the second differentiable condition.

3.2. Two-Layer XGBoost Algorithm

In general, as one of the supervised learning algorithms, XGBoost is tasked with extracting useful features and then building a mapping from features to labels by learning the relationship between the extracted features and the labels in the training set. For the test set, the labels are predicted using features in the test set through the trained model. That is, the features are the condition for making the prediction,and the labels are the prediction result. Considering the specific conditions of this work, we designed a two-layer XGBoost algorithm that can accurately predict the shop in which the customer is located. The structure is described in Figure 3. The first layer of our model is a multi-class XGBoost, which predicts the accuracy of all shops. Then, we input the top N predicted shops into the second layer of our model, which is a two-class XGBoost, predicting whether the shop is the ground-truth shop.

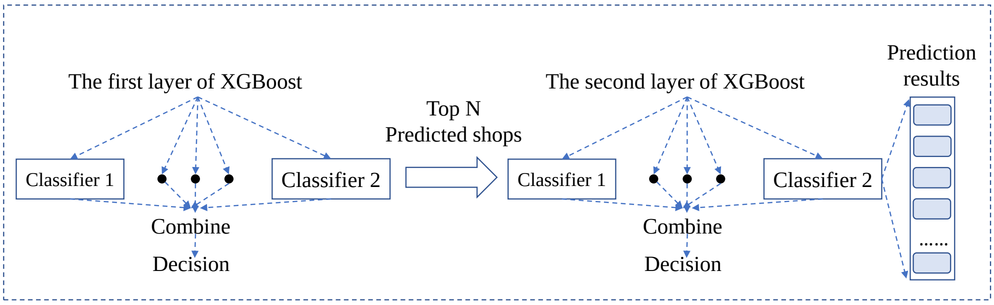

Figure 3. Network architecture of the two-layer XGBoost algorithm.

3.2.1. Multi-Class Problem for the First Layer

A multi-class problem is one in which there are more than two types of predictive labels. Because there are obviously different classifications between shops in different shopping malls, our algorithm differentiates by shopping malls first. In the first layer of our algorithm, we aimed to predict the shop ID of one transaction record that contains many features, which means that the number of predicted labels categories is the number of shops in a shopping mall. Therefore, we defined the first layer of our algorithm to solve the multi-class problem: the number of predicted labels categories is the number of shops in a shopping mall.

Since the test dataset did not contain any shop information, it was not helpful for predicting shop-related features such as shop latitude and longitude, and shop type in training. Therefore, in training, the features we used were the WiFi signal strength, the dimension of which is the number of WiFi shared by the training set and the test set, and the latitude and longitude of customers, with two dimensions. In addition, we used the statistical features containing the maximum, minimum, average, standard deviation, and the four quartiles of the WiFi signal strength of each transaction record, which were introduced in Section 2. The statistical features had 10 dimensions.

After the features were determined, we defined different shops in a shopping mall as different categories of labels, so that the number of labels was the same as the total number of shops in a shopping mall. Then, we sent the features and labels as the input to the XGBoost algorithm for training. Using the built XGBoost algorithm model, the shop IDs of the training dataset and the test dataset were predicted. The prediction result for each transaction in the test dataset is the probabilities of all labels, which means the probability of the customer in each shop is predicted.

We chose the shop with the highest probability of prediction in each transaction record in the test dataset as our prediction result of the first layer of our algorithm.

3.2.2. Two-Class Prediction for Second Layer

As mentioned above, the shop-related features were not used in the first layer of our algorithm, which are helpful for prediction. Therefore, after obtaining the probability of shops predicted by the first layer of our algorithm, we added shop-related features into the second layer our algorithm to obtain a more accurate prediction result. The specific method is as follows:

We defined the shops with the top five probabilities after each transaction record was predicted by the first layer of our algorithm as an alternative set. Then, we expanded each of the transaction records in the training dataset and the test dataset into five, adding the top five probability shop IDs and the corresponding probabilities. In this way, shop information was contained in the test dataset in each transaction record, which means that we could use the shop-related features in training.

The new features included in the second layer of our algorithm were the latitude and longitude of shops, the transaction per person in shops, the number of WiFi connections around the shops, the transaction per person corresponding to WiFi connections, the number of transactions in shops, and the number of shops in which a transactions occurred by a total of 7-dimensional features, which are introduced in Section 2.2.

Then, we labeled the shops that were the same as the ground-truth shop with a positive label of 1, and the shops that were not the same as the ground-truth shop with a negative label of 0 for each transaction record in the training dataset. This 0/1 label is the prediction result of the second layer of our algorithm in the test dataset, which means that the second layer of our algorithm provides a two-class prediction. In the two-class prediction, the output is the probability of the negative label 0 and the positive label 1. For the five transaction records expanded from one transaction record, we selected the shop with the highest probability of the positive label 1 among the five transaction records as the final predicted shop.

4. Experiments and Results

We grouped the features by relevance to verify the impact of performance using different features. We randomly selected eight shopping malls as our verification dataset. The result is shown in Figure 4. In the figure, the numbers one to six on the horizontal axis represent the WiFi information features, shop and user transaction record features, the WIFI statistical features, other features, the latitude and longitude features, and the time features, respectively. We concluded that the latitude and longitude features and the time features considerably improved the accuracy of the result.

The hyperparameters were as follows: the *k*-value of the KNN was 3 and the weight was 1, $\frac{1}{3}$, and $\frac{1}{3}$. For the XGBoost model, we used gbtree as the booster. The evaluation index was accuracy, and the minimum loss of leaf nodes was 0.1. In the experiments, the length of the training process of KNN was about 2 h, and the proposed model needed to be trained for about 4 h.

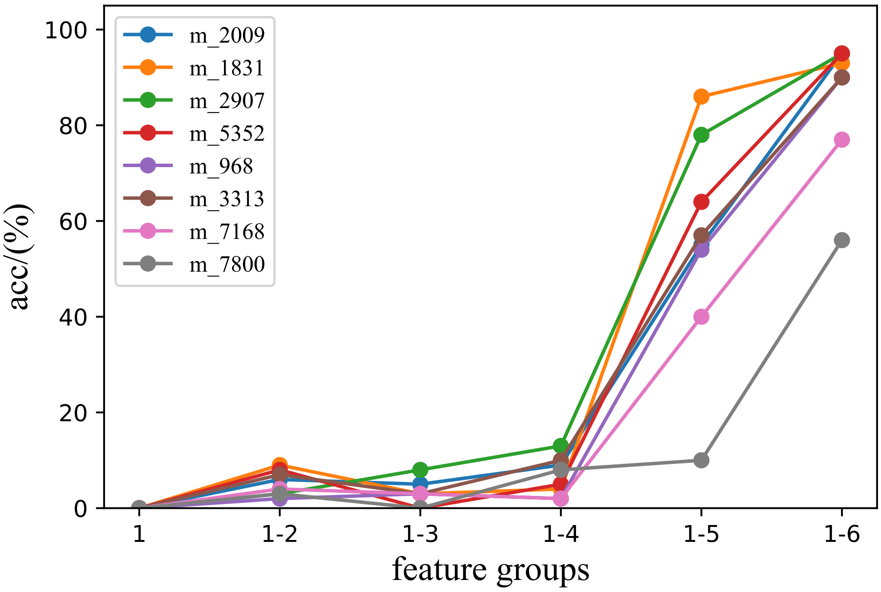

Figure 4. Performance comparison using different features.

Using multidimensional features, the accuracy of the first layer of our algorithm was high, as shown in Figure 5. The accuracy of the second layer of our algorithm is higher compared with the first layer because of the shop-related features that were not used in the first layer. The results of our two-layer XGBoost algorithm are shown in Table 3. Due to the dataset used in this paper not being a public dataset for special research, few previous papers were based on it. Some researchers only shared their methods and results in blogs [38,39], with precision being 88%. Moreover, to indicate our improvements in detail, we compared the performance between the proposed model and some common algorithms in Table 3. The differences between these structures are also clarified in Section 4.1.

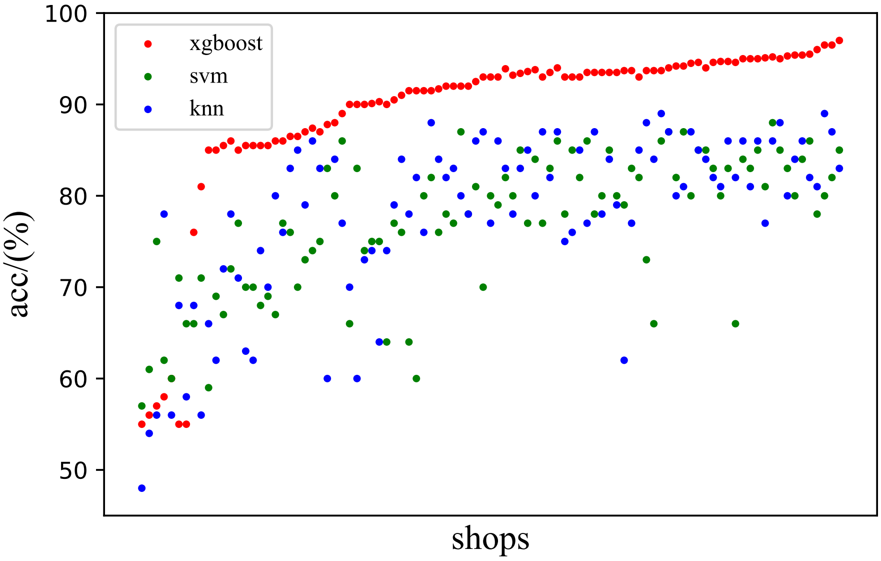

Figure 5. The prediction accuracy of different algorithms for different shopping malls. The red dots represent our two-layer XGBoost algorithm, the blue dots represent the statistical rule algorithm, and the green dots represent the K-nearest neighbor (KNN) classification algorithm.

Table 3. The results of different algorithms.

Method	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Statistical rule algorithm	68.80	72.40	70.72	69.03
KNN algorithm	73.41	78.60	75.82	76.59
First layer of our algorithm	88.52	91.30	89.82	90.20
Two-layer XGBoost algorithm	91.43	94.35	91.89	92.40

4.1. Comparison with the Statistical Rule Algorithm

We designed a statistical rule algorithm because most of the features we used were WiFi signal strength features, which only consider the WiFi signal strength in each transaction record. The main idea of the algorithm is to count the number of times that each WiFi appears in the transaction records of all shops in the training dataset. Then, for a certain transaction record in the test dataset, we selected the WiFi connection with the strongest signal strength from this transaction record and defined the shop with the most occurrences of this WiFi connection as the prediction result. The algorithm steps are shown in Table 4.

Table 4. Flow chart of the statistical rule algorithm.

Step	Description
Input	training dataset, test dataset
1	all_wifi = [all WiFi_id in training dataset]
2	wifi_to_shop = WiFi_id: the number of times this WiFi appears in each shop
3	For each transaction record in the test set, find the WiFi with the strongest signal strength
4	Query the shop result_shop with the most occurrences of this WiFi in the dictionary wifi_to_shop
5	Return result_shop

The statistical rule algorithm only uses WiFi features, which means we could not combine other features to fit a more accurate generalization model. Compared to our algorithm, the statistical rule algorithm without using other features was significantly less accurate, which means that other features also affected the construction of the predictive model. The accuracy of the statistical rule algorithm is shown in Table 3.

4.2. Comparison with KNN Algorithm

The main idea of the KNN classification algorithm is to compare the features in the test dataset with the features in the training set and to find the top K data with the most similar data in the training dataset. Then, the class of each data in the test dataset is the one with the most occurrences among the K data. The algorithm steps are shown in Table 5.

Table 5. Flowchart of the KNN algorithm.

Step	Description
Input	training dataset, test dataset
1	Calculate the Euclidean distance between each testing datum and each training datum
2	Sort by the increasing Euclidean distance
3	Select the K points with the smallest Euclidean distance
4	Determine the frequency of occurrence of the category of the top K points
5	Return the category with the highest frequency among the top K points as the prediction classification of the testing data.

We varied from 1 to 25 to determine the value of K; the result is shown in Figure 6. We found that, when K = 1, the accuracy of the algorithm was the highest but was still lower than that of our algorithm. The accuracy of the KNN algorithm is shown in Table 3. This occurred because, for each datum in the test dataset, the KNN algorithm uses all data in the training dataset; therefore, the KNN algorithm does not first learn the training set and then predict the test dataset. The main disadvantage of the algorithm in classification is that, when the sample is unbalanced, when a new sample is input, the sample of the large-capacity class among the K neighbors of the sample will be dominant. The algorithm only calculates the nearest neighbor samples. Therefore, the KNN algorithm is not suitable for this task.

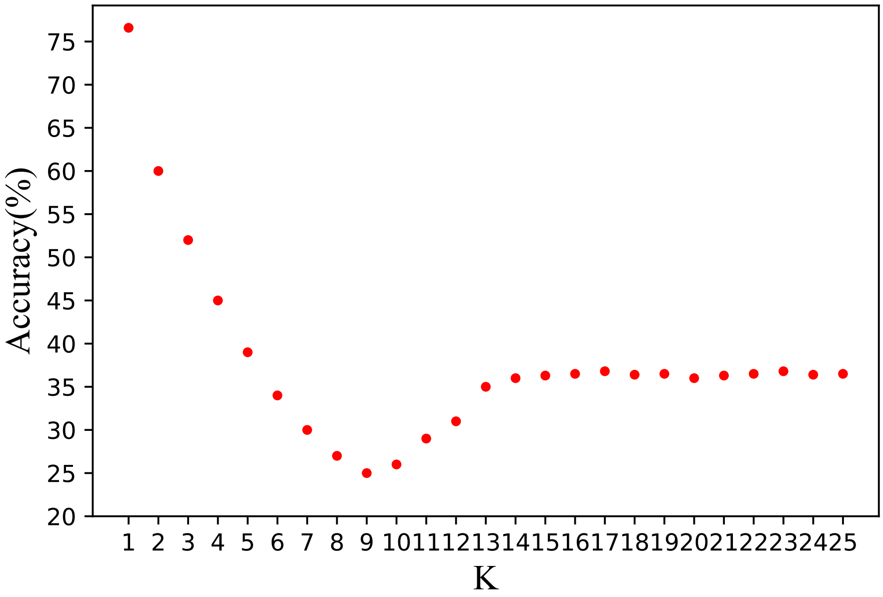

Figure 6. The accuracy of the KNN algorithm with K varying.

As shown in Figure 5, our two-layer XGBoost algorithm provided more accurate predictions in most shopping malls compared to the other two algorithms but the prediction accuracy of our algorithm was lower than the statistical rule algorithm or the KNN classification algorithm in individual shopping malls. This showed that our two-layer XGBoost algorithm has a strong predictive power but that underfitting still occurred for some specific models. The prediction results of the KNN classification algorithm were very unstable, and the generalization ability was low. For the statistical rule algorithm, since only statistical features were used, the model provided inaccurate predictions and the overall prediction result was not satisfactory. Therefore, based on a series of conditions such as algorithm effect, algorithm complexity, and generalization ability, our two-layer XGBoost algorithm is the better choice in this work.

5. Services and Behavior Analysis

We provide a series of shop-related services for the shops where the customers can be located, including comprehensive shop evaluation, shop product recommendation, shop discount summary, and shop peripheral services, as shown in Figure 7.

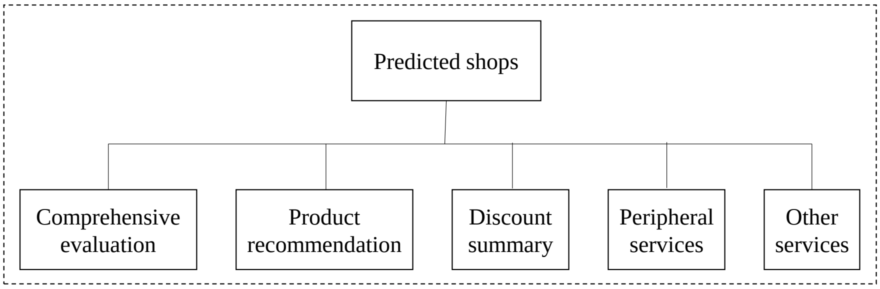

Figure 7. Services provided to customers using the predicted shops.

For comprehensive shop evaluation, by identifying the shops, we can introduce brand information to customers, provide relevant product photos, and allow customers to rate shops in combination with rating software. Additionally, we can obtain past customers' experience evaluation using third-party commenting, which is valuable to new customers. Customers can select a suitable shop according to their age, gender, preferences, and other characteristics through past customers' evaluations. In addition, we can provide information such as the per capita transaction status of the shops for the customers, so that customers can determine their intended level of interaction in the shop in advance. Such functions allow customers to obtain comprehensive and objective shop information during their first visit. With these functions, customers can quickly obtain a more comprehensive understanding and judgment of the shop while eliminating complicated and long screening time.

For shop product recommendations, when the customers are satisfied with information such as the comprehensive score and transaction level of the shop, we can provide customers with more detailed shop goods or service information. For instance, we can recommend the current hot-selling products or services at the shop using social software and shop data. Based on this, the customers can further determine whether the shop meets their needs. Network celebrities' recommendations and the shop catalogue can be provided to customers as a reference. Through other people's recommendations, customers can identify the best products and services that meet their needs. Additionally, according to the customer′s past browsing and transaction records, the customer′s preferences can be estimated. Different products are recommended to different customer groups. We can provide personalized, differentiated, and high-level services for different customers. Targeted referral services can more suitably match products with customers and can reduce customers' search time, which means that we can provide consumers with precise and rapid services.

For shop discounts, we can provide online and offline price comparisons through an e-commerce platform. The price comparison interface can provide the prices of major e-commerce suppliers and offline shops. Product information and customer evaluation can be found too. A comprehensive price comparison service allows customers to find affordable products from various shops' information. This service can reduce the customer′s search time. Additionally, online purchases and offline transaction are feasible for customers. When the online price is better or customers want to re-buy a product after seeing the products in a shop, customers can purchase the products and services online. A summary of shops' preferential activities, such as coupons, membership cards, and other activities can be easily obtained. We can also provide a comparison of similar products within the current shop and at other shops in order to minimize expenditure to meet customer transaction needs and to improve customer effectiveness.

In terms of shop peripheral services, when customers are faced with a variety of shops and products, customers often cannot find a suitable store. Perhaps they will miss suitable products, which then reduces satisfaction. According to the map of the mall and the current location of the customer, we can provide customers with more intuitive map navigation for shopping malls, elevators, toilets, service desks, etc. Clear and intuitive floor and route navigation can help customers overcome shopping difficulties, can relieve the psychological pressure on customers, can stimulate customers' desire to purchase, and can provide a comfortable and enjoyable shopping experience. We can indicate the standard parking lot charge related to the current shop's transaction coupon. We not only meet the needs of consumers when they are shopping but also meet pre- and post-consumer needs. We are committed to providing customers with comprehensive, full-time, overall services and to creating a comfortable and satisfying consumer environment for customers.

In terms of other services, for specific types of shops, personalized services can be provided according to customer needs. For example, food shops can check the number of people in the current shop. If the customers need to be somewhere else, one-click access services can be provided. The customer can queue remotely and intelligently, which means that they can arrange their wait time reasonably. The service can remind the customer of meal times according to the service speed of the restaurant. This service can solve the long wait-times problem during peak meal times. For specific types of customers, personalized services can also be provided, such as inquiries for an infant room or a baby dining chair. We are devoted to providing a full range of services to proactively meet the individual needs of customers.

6. Conclusions

In this paper, we proposed an accurate method of shopping prediction. We used the XGBoost algorithm, which is a machine learning algorithm, to predict the shops where customers are currently located. We evaluated our method on a dataset composed of real customers' transaction records in shopping malls, and the results showed that our method is useful in this task. The accuracy reached 92.40%. With the predicted shops, we can provide customers with more shop-related services, which can be easily integrated with current mobile apps to provide more accurate user services. In our future work, we plan to discuss in detail the services after obtaining accurate customer positioning, which was only briefly mentioned in this paper. Additionally, we want to verify the generalization of our method on different datasets.

Author Contributions: Conceptualization, J.Z.; Investigation, Y.X.; Methodology, H.J.; Writing—original draft, M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study did not involve humans or animals.

Informed Consent Statement: The study did not involve humans.

Data Availability Statement: The data used in this study is available at https://tianchi.aliyun.com/competition/entrance/231620/information accessed on 21 April 2021.

Acknowledgments: Thanks for the experimental environment provided by the School of Economics and Management, Beijing University of Posts and Telecommunications.

Conflicts of Interest: We declare that we do not have any commercial or associative interests that represent conflicts of interest in connection with the work submitted.

References

1. Sarker, I.H.; Kayes, A. ABC-RuleMiner: User behavioral rule-based machine learning method for context-aware intelligent services. *J. Netw. Comput. Appl.* **2020**, *168*, 102762. [CrossRef]
2. Sarker, I.H.; Colman, A.; Han, J.; Kayes, A.; Watters, P. CalBehav: A Machine Learning-Based Personalized Calendar Behavioral Model Using Time-Series Smartphone Data. *Comput. J.* **2020**, *63*, 1109–1123. [CrossRef]
3. Purba, K.R.; Asirvatham, D.; Murugesan, R.K. Classification of instagram fake users using supervised machine learning algorithms. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 2763. [CrossRef]
4. Huang, H.; Liu, L. Analysis of the Location of Nanning Large-Scale Mall Based on BP Neural Network. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *42*, 975–979. [CrossRef]
5. Arai, T.; Yoshizawa, T.; Aoki, T.; Zempo, K.; Okada, Y. Evaluation of indoor positioning system based on attachable infrared beacons in metal shelf environment. In Proceedings of the 2019 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–13 January 2019; pp. 1–4.
6. Santo, H.; Maekawa, T.; Matsushita, Y. Device-free and privacy preserving indoor positioning using infrared retro-reflection imaging. In Proceedings of the 2017 IEEE International Conference on Pervasive Computing and Communications (PerCom), Kona, HI, USA, 13–17 March 2017; pp. 141–152.
7. Martín-Gorostiza, E.; García-Garrido, M.A.; Pizarro, D.; Salido-Monzú, D.; Torres, P. An Indoor Positioning Approach Based on Fusion of Cameras and Infrared Sensors. *Sensors* **2019**, *19*, 2519. [CrossRef] [PubMed]
8. Saab, S.S.; Nakad, Z.S. A standalone RFID indoor positioning system using passive tags. *IEEE Trans. Ind. Electron.* **2010**, *58*, 1961–1970. [CrossRef]
9. Huang, C.H.; Lee, L.H.; Ho, C.C.; Wu, L.L.; Lai, Z.H. Real-time RFID indoor positioning system based on Kalman-filter drift removal and Heron-bilateration location estimation. *IEEE Trans. Instrum. Meas.* **2014**, *64*, 728–739. [CrossRef]
10. Xu, H.; Ding, Y.; Li, P.; Wang, R.; Li, Y. An RFID indoor positioning algorithm based on Bayesian probability and K-nearest neighbor. *Sensors* **2017**, *17*, 1806. [CrossRef] [PubMed]
11. Wu, C.; Wang, X.; Chen, M.; Kim, M.J. Differential received signal strength based RFID positioning for construction equipment tracking. *Adv. Eng. Inform.* **2019**, *42*, 100960. [CrossRef]
12. De Angelis, A.; Moschitta, A.; Carbone, P.; Calderini, M.; Neri, S.; Borgna, R.; Peppucci, M. Design and characterization of a portable ultrasonic indoor 3-D positioning system. *IEEE Trans. Instrum. Meas.* **2015**, *64*, 2616–2625. [CrossRef]
13. Yayan, U.; Yucel, H.; Yazıcı, A. A low cost ultrasonic based positioning system for the indoor navigation of mobile robots. *J. Intell. Robot. Syst.* **2015**, *78*, 541–552. [CrossRef]
14. Lin, Q.; An, Z.; Yang, L. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In Proceedings of the 25th Annual International Conference on Mobile Computing and Networking, Los Cabos, Mexico, 21–25 October 2019; pp. 1–16.

15. Sharp, I.; Yu, K. Indoor WiFi Positioning. In *Wireless Positioning: Principles and Practice*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 219–240.
16. Bisio, I.; Lavagetto, F.; Marchese, M.; Sciarrone, A. Smart probabilistic fingerprinting for WiFi-based indoor positioning with mobile devices. *Pervasive Mob. Comput.* **2016**, *31*, 107–123. [CrossRef]
17. Zou, H.; Jin, M.; Jiang, H.; Xie, L.; Spanos, C.J. WinIPS: WiFi-based non-intrusive indoor positioning system with online radio map construction and adaptation. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 8118–8130. [CrossRef]
18. Yang, C.; Shao, H.R. WiFi-based indoor positioning. *IEEE Commun. Mag.* **2015**, *53*, 150–157. [CrossRef]
19. Subedi S., Pyun, J.Y. Lightweight Workload Fingerprinting Localization Using Affinity Propagation Clustering and Gaussian Process Regression. *Sensors* **2018**, *18*, 4267. [CrossRef] [PubMed]
20. Spachos, P.; Plataniotis, K.N. BLE beacons for indoor positioning at an interactive IoT-based smart museum. *IEEE Syst. J.* **2020**, *14*, 3483–3493. [CrossRef]
21. Galván-Tejada, C.E.; Zanella-Calzada, L.A.; García-Domínguez, A.; Magallanes-Quintanar, R.; Luna-García, H.; Celaya-Padilla, J.M.; Galván-Tejada, J.I.; Vélez-Rodríguez, A.; Gamboa-Rosales, H. Estimation of Indoor Location Through Magnetic Field Data: An Approach Based on Convolutional Neural Networks. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 226. [CrossRef]
22. Fujiwara, R.; Mizugaki, K.; Nakagawa, T.; Maeda, D.; Miyazaki, M. TOA/TDOA hybrid relative positioning system based on UWB-IR technology. *IEICE Trans. Commun.* **2011**, *94*, 1016–1024. [CrossRef]
23. Hong, C.Y.; Wu, Y.C.; Liu, Y.; Chow, C.W.; Chen, Y.Y. Angle-of-Arrival (AOA) Visible Light Positioning (VLP) System Using Solar Cells with Third-Order Regression and Ridge Regression Algorithms. *IEEE Photonics J.* **2020**, *12*. [CrossRef]
24. Gansemer, S.; Großmann, U.; Hakobyan, S. Rssi-based euclidean distance algorithm for indoor positioning adapted for the use in dynamically changing wlan environments and multi-level buildings. In Proceedings of the 2010 International Conference on Indoor Positioning and Indoor Navigation, Zurich, Switzerland, 15–17 September 2010; pp. 1–6.
25. Keller, J.M.; Gray, M.R.; Givens, J.A. A fuzzy k-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybern.* **1985**, *1*, 580–585. [CrossRef]
26. Arbib, M.A. (Ed.) *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 2003.
27. Krawczyk, B.; Minku, L.L.; Gama, J.; Stefanowski, J.; Woźniak, M. Ensemble learning for data stream analysis: A survey. *Inf. Fusion* **2017**, *37*, 132–156. [CrossRef]
28. Lee, S.; Kim, J. Apparatus, Method, and Medium for Detecting Face in Image Using Boost Algorithm. U.S. Patent 7,835,541, 16 November 2010.
29. Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-class adaboost. *Stat. Its Interface* **2009**, *2*, 349–360. [CrossRef]
30. Izonin, I.; Tkachenko, R.; Vitynskyi, P.; Zub, K.; Dronyuk, I. Stacking-based GRNN-SGTM Ensemble Model for Prediction Tasks. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020.
31. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
32. Zeng, J.; Chen, Y.; Zhu, H.; Tian, F.; Miao, K.; Liu, Y.; Zheng, Q. User Sequential Behavior Classification for Click-Through Rate Prediction. In Proceedings of the International Conference on Database Systems for Advanced Applications, Jeju, Korea, 24–27 September 2020; Springer: Cham, Switzerland, 2020; pp. 267–280.
33. Sarker, I.H.; Colman, A.; Han, J.; Khan, A.I.; Abushark, Y.B.; Salah, K. Behavdt: A behavioral decision tree learning to build user-centric context-aware predictive model. *Mob. Netw. Appl.* **2020**, *25*, 1151–1161. [CrossRef]
34. Alibaba. Dataset. 2019. Available online: https://tianchi.aliyun.com/competition/entrance/231620/information (accessed on 22 April 2021).
35. Ramasubramanian, K.; Singh, A. *Machine Learning Using R*; Springer: Berlin/Heidelberg, Germany, 2017.
36. Nielsen, D. *Tree Boosting with XGBoost*; NTNU Norwegian University of Science and Technology: Trondheim, Norway, 2016.
37. Loh, W.Y. Classification and regression trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2011**, *1*, 14–23. [CrossRef]
38. CSDN. Comparison Results. 2018. Available online: https://blog.csdn.net/gentle_guan/article/details/78593865 (accessed on 22 April 2021).
39. CSDN. Comparison Methods. 2018. Available online: https://blog.csdn.net/gemnwing/article/details/81781971 (accessed on 22 April 2021).