Research on the awakening Algorithm of sleeping members in drugstore

HaoTian Zheng, Tao Li*, Yan Song
College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China

*Corresponding author: litao@wust.edu.cn

Abstract. The Awakening of sleeping members is the focus of smart marketing. The cost of losing a sleeping member in existing pharmacies is about 1/4 of the cost of adding a new member. This paper abstractly classifies the problem of whether the sleeping members are easy to be woken up in drugstore into a dichotic problem, and solves the problem that the existing sleeping members' wake-up model is applied to the users of the sleeping members in drugstore and the prediction accuracy of the users' consumption in the store is not high. According to sleep members have rich behavior characteristics, behavior, medical conditions, gender, age, payment, the membership card integral, active level characteristics such as properties, through the analysis of these sleep member attributes in order to distinguish between normal users and sleep, characteristics and the actual scene at the same time, using the method of feature selection to remove redundant features to form a new expression model. On the basis of multiple perspectives, multiple algorithm models were integrated to wake up the sleeping members. The actual store-to-store consumption data of the sleeping member users were used to verify and compare the effectiveness of the algorithm model, and the characteristic expression model with the best discrimination and accuracy was selected. Through the experimental study of the abstract dichotomy algorithm, it is found that the sleeping members with what characteristics are easy to be woken up. The experiment proves that the feature expression model has a high degree of differentiation for the sleeping users, and has a high accuracy in judging the users who are likely to be woken up among the sleeping members of the pharmacy.

1. Introduction
With the rapid development of data informatization, the amount of data increases explosively. Coupon promotion is an important marketing method for merchants, but the problem of poor choice of coupon target and low redemption rate has been plaguing merchants. Under the background of the integration of the Internet and the real economy, more and more drugstores have begun to use online to offline marketing to expand their business, and to urge sleeping members of drugstores to make offline consumption in physical stores by means of coupons and other discounts. To explore the potential commercial value of users for pharmaceutical companies, so as to improve the benefits of pharmaceutical companies and reduce costs. However, directly issuing coupons to all sleeping members...
will lead to the limitation of too high efficiency and too low cost. Therefore, how to identify these members has become the focus of this paper. Start from the following two aspects:

From which aspects to model: what kind of feature vector to identify feature members, from which aspects to establish a special model is particularly important, so we need to find what kind of feature vector will have an impact on whether members come to consume, accurate feature vector will make the recognition effect better.

How to identify the characteristics of the member: looking for a good model, feature vector to establish good will have to consider what kind of algorithm to identify characteristics of the members, the traditional identification methods efficiency will be lower, so the members whether to come to consumption abstract for binary classification problems, choose two classification algorithm to sleep member model and compares relatively more efficient.

There are many algorithms that can be used for dichotomy, including support vector machine (SVM), decision tree algorithm, random forest algorithm, naive Bayes algorithm, logistic regression algorithm, etc. The essence of the classification algorithm is that given an object M, the object is divided into a predefined category Ni algorithm.

Nowadays, drugstores are homogeneous, transparent in price and have few advantages compared with other drugstores. When there is little difference between drugstores, members may consider buying drugs from nearby drugstores, and almost all major drugstores have serious membership loss. Adding new members is extremely difficult, and retaining a member would be much cheaper. This also depends on member marketing. From the perspective of members, marketing to a large number of members can not achieve the expected effect of merchants, but also may cause the disgust of members. From the perspective of merchants, blindly issuing coupons to most members will result in high marketing cost and low efficiency. In this case, drug screening target crowd, tailor marketing solutions, lower marketing costs and improve the efficiency of marketing, might need to send tens of thousands of text messages to a member of the operating activities, through member will sleep easily awakened abstraction algorithm for binary classification, the classification of the machine learning algorithms, probably thousands of text messages can be easily done, more accurate target.

For pharmacy, want to get a better development, you need to attract more new members, and safeguard the existing old members, reducing the cost of membership, and the development of new members is far higher than the cost of maintaining the existing cost, in the development of new users is relatively difficult at this stage, this paper mainly study sleep members forecast model (pharmacy members not consumption) 90 days), and in pharmacy sleep member identification technology, using the typical characteristic factor can effectively improve the efficiency of the classification of the model, to save health pharmacy data as an example, through investigating the member have to sleep the original data cleaning and visual analysis, Unearth hidden behind the data characteristics, whether can affect the member to the store consumption characteristics were extracted, and based on XGBoost, decision tree, random forest, gaussian, Bernoulli model mining user attribute data of sleep, the sleep user behavior information, commodity information, purchase information such as history, to estimate whether the user can be awakened. In order to further improve the prediction effect, the paper compares the prediction results of healthy sleep members’ data through the above five models, selects the model with the best prediction accuracy and recall rate, and adjusts the parameters with grid search algorithm to improve the prediction model.

2. Research Background

Currently in electronic coupon marketing have more research, and part of the pharmacy marketing is belongs to the electronic coupon, a general of the traditional marketing model, this article on the traditional marketing model, based on user electronic coupons sleep later consumption and future user survey, abstract for binary classification problems research member easily awakened, the characteristic of what is converted into feature model, including XGBoost, decision tree classification algorithm, random forest, gaussian, Bernoulli algorithm.
In Reference 1, through exploratory analysis of data and based on relevant background business knowledge system, through visual analysis, the features hidden in the data are extracted, and the XGBoost method with better performance is used for rule mining, achieving good results. In order to further improve the prediction accuracy and generalization performance of XGBoost method, combined with the characteristics of engineering, the integrated method of study, using the GLMNET and XGBoost model fitting residual error, combined with LM, TSLM in trends and advantages of seasonal forecast in this paper, a optimization based on XGBoost combination model to forecast industry data, the composite model is verified by experiment has better accuracy and generalization ability. Literature 2 constructs a theoretical model from the perspective of personal characteristics and motivation to explore which factors have an impact on consumers’ use of coupons, including perceived entertainment, perceived convenience and perceived economic interests. Among them, perceived entertainment has a relatively large impact, followed by perceived convenience. Personal characteristics factors that influence the use intention -- personal innovation and personal coupon tendency have both direct influence on the use intention and indirect influence on the use intention through motivation factors. Literature 3 In the O2O marketing process, in order to improve the coupon utilization rate, three decision ideas are introduced into the coupon usage prediction problem, and XGBoost, an integrated algorithm in the machine learning algorithm, is used to build the model of coupon usage. Then, misclassification cost and learning cost are taken into account in the three decision-making processes, which makes the classification process closer to reality. Finally, the real consumption data of user coupons provided by Alibaba on Tianchi platform is analyzed experimentally. The results show that the classification accuracy can be improved effectively by using XGBoost-based binary classification algorithm. Merchants can not only maintain existing customers, but also identify potential new customers, thus reducing the marketing costs of merchants.

Literature 4 in order to achieve the new orientation of retail coupons, coupons to users using behavior prediction model is put forward, using the algorithm of XGBoost, broke through to TAM (technology acceptance model) model on the basis of explanation of the will of individual coupons use traditional method, and based on word-of-mouth network real transaction data of feature extraction and modeling user behavior. After the K-fold cross verification, the feature with a high contribution to consumers' decision use was determined through the variable importance score, and the AUC (Area Under Curve) accuracy was compared with random forest and GBDT (Gradient enhancement decision tree) algorithm. This study demonstrates the effectiveness of XGBoost-based integrated learning algorithm in the prediction of coupon usage behavior and has important practical significance for new retail precision marketing. Literature 5 systematically summarized the types and characteristics of electronic coupons at the present stage, deeply analyzed the specific marketing applications and advantages of different types of electronic coupons, such as network marketing, database marketing and offline marketing, and finally elaborated the development trend of the parties involved in electronic coupons in the future. Literature 5 illustrates that under the background of the integrated development of Internet and real economy, online coupons often assume the important experience of improving user experience and promoting re-consumption. Models such as gradient promotion tree and random forest were constructed to predict online coupon usage behavior, and the importance of influencing factors was ranked. After the experiment, the results showed that the average test accuracy and the area under the curve were 0.804 and 0.886, which were higher than the random forest and single decision tree algorithms. Coupon discount rate has a decisive influence on users' behavior of using coupons. Features such as the distance from the nearest store where users often operate and the time they receive coupons have a significant influence on users' behavior of using coupons. In reference 6, a new prediction model based on data integration of random forest is proposed by using the integration idea of random forest and the random segmentation and recombination of training data set. By random recombination, the original one-dimensional predictive variables are reorganized into high-dimensional vectors, and the final predicted value is the sum of the output of the model. The proposed model not only retains the convergence characteristics of the random forest, but also inherits its advantages, and its accuracy and robustness are greatly improved. Compared with the traditional method, the stochastic forest with data integration has
achieved remarkable improvement in the actual prediction effect. In addition, the idea of data integration is not limited to a particular model and can be ported to other algorithms. Literature 7 FORESTVis visual analysis system was designed and implemented, the system includes tree view, part of the dependency graph, t-SNE projection, characteristic view multiple interactive visual component, such as using the system, the relevant researchers and practitioners can intuitive understanding of the basic structure and working mechanism of random forests, and to assist the user to evaluate performance of the model. Finally, a case study on Kaggle’s open data set was carried out to verify the feasibility and effectiveness of the method.

3. Model Overview

In this paper, a characteristic model was constructed based on the attributes of the sleeping members of the chain 500 pharmacies who recently issued coupons and the recent consumption data of all the sleeping members in the activities of the pharmacy. First member or coupon sleep data preprocessing new attributes arrive with 0 and 1 to indicate whether the member is awakened to consumption, abstract for binary classification problems, combined with a variety of binary classification learning algorithm to construct predictive model is applied to sleep members, according to the divide and the proportion of the training set and test set is used to model of training, will wake up after the training effect is the best model using a grid search algorithm which can adjust the parameter selection results of optimal parameters of the model. The flow chart is as follows:

![Flow chart](image-url)

Fig 1. Flow chart
4. Analysis of Model

4.1. The experimental data

Referring to the characteristics adopted in the prediction of the consumption behavior of sleeping members, the characteristics of the sleeping members, the behavior of sleeping members and the attributes of drugs are calculated. The characteristics of chain drugstores and those that may affect whether the sleeping members are awakened to consume are shown in Table 1:

| perspective | Character | Character description |
|-------------|-----------|-----------------------|
| The member properties | active (level) | Member's consumption activity in the store |
| | gender (sex) | Staff gender |
| | age (age) | male or female |
| | member of year (createdate) | Age of members |
| | distance (distance) | Membership card registration time so far |
| | integral (point) | The distance between the store and members |
| | Average consumption in the last 6 months (ma6) | Members earn a total of points by spending |
| | Average consumption in the last 6 months (mt6) | Average amount spent by members in the store in the last 6 months (total amount/total number of times) |
| | Categories of consumption in the last 6 months (num6) | The number of times members spent in the store in the last six months |
| | Average consumption in recent 1 year (ma1) | The total number of drugs purchased by members in the last six months |
| | Consumption times in recent 1 year (mt1) | Average amount spent by members in the store in the last year (total amount/total number of times) |
| | Categories of consumption in recent 1 year (num1) | The number of purchases members made in the store in the last year |
| | Average consumption in the past three months (ma3) | The total number of drugs purchased by members in the past year |
| | Consumption times in recent 3 months (mt3) | Average amount spent by members in the store in the last three months (total amount/total number of times) |
| | Categories of consumption in the last 3 months (num3) | The number of times members spent in the store in the last three months |
| | Number of times you use coupons (total) | The total number of drugs purchased by the members in the past three months |
| | Activity intimacy (intimacy) | The total number of times a member uses the discount to participate in an event within one year |
| | Chronic disease (is_suff) | The more times I participated in the activity and consumed in the store, the higher the intimacy was (the number of activities I participated in/total number) |

Whether the purchase of drugs for chronic diseases
4.2. Construction of feature model

The problem of whether the sleeping members will be awakened in the pharmacy can be regarded as a dichotomy problem. According to a large amount of historical information, it is found out what characteristics the sleeping members are likely to be awakened after receiving coupons and respond to consumption. Therefore, it is necessary to choose a model with high classification efficiency and fast recognition speed. XGBoost is a machine learning algorithm focusing on gradient enhancement. It combines the weak learner model into a strong learner in an iterative way. As an efficient implementation of GBDT, XGBoost is an algorithm with a particularly high upper limit and is a commonly used supervised integrated learning algorithm. Gradient Boosting algorithm, which is strong and convenient and can be used to build the model in parallel. A large amount of training data is used to predict the target variable. Figure 1 is the algorithm flow chart of XGBoost. Decision tree as XGBoost weak learning, every time training if individual learning, XGBoost will raise the weight the last data points wrong class some and then to the current weak learning machine learning, by adding new weak learning correct before learning the possible existence of residual, will eventually all weak learning weighted summation and final prediction.

The XGBoost base model is a decision tree, and you start by defining the output of a single decision tree as

\[ f(x) = \omega_{q(x)}, x \in \mathbb{R}^q, R^q \rightarrow \{1, 2, \ldots, T\} \]  

(1)

Where \( x \) is the input vector, \( Q \) represents the structure of the tree, the structural function \( Q(x) \) represents the index number of the input mapped to the leaves, \( W \) represents the fraction of the leaves corresponding to each index number, \( T \) is the number of middle nodes of the tree, and \( D \) is the characteristic dimension. The XGBoost algorithm can be viewed as a set of \( K \) decision trees, and the output of the set of \( K \) trees is

\[ y_i = \sum_{k=1}^{K} f_k(x_i) \]  

(2)
According to the decision tree model, the complexity calculation formula of a single decision tree is

$$\Omega(f_i) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2$$

(3)

Similarly, the complexity of the integration tree can be expressed as

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{i=1}^{T} \omega_i^2$$

(4)

In addition, the iteration of the target function of the XGBoost algorithm at step T can be expressed as

$$J^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{i=1}^{T} \Omega(f_i)$$

(5)

The formula consists of two parts: the first part represents the sum of the errors of the true value and the predicted value, and L is the error function. The second part represents the sum of the complexity of a single decision tree. Where, the objective function can be converted into

$$Obj^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{i=1}^{T} \Omega(f_i) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f(x_i)) + \sum_{i=1}^{T} \Omega(f_i)$$

(6)

In a word, XGBoost algorithm is an improvement of GBDT algorithm, which has some advantages in practical application. For example, regularization steps are introduced to prevent over-fitting; Parallel processing improves the operation speed; Allows users to customize optimization goals and evaluation criteria, adding soul; Contains rules for handling missing values; There are special pruning steps to control the complexity of the decision tree.

The prediction process of the drug store sleep member awakening model is as follows:

The sample data set was divided into training set and test set according to the ratio of 7:3.

For the training set, repeat steps A-C.

a. Starting from the root node, the split point is recursively found according to the formula until the condition is satisfied. At this point, all features are transformed into a node in a regression tree.

b. Cyclic execution of step A enables multiple regression trees established to maintain a downward trend on the gradient of loss function.

c. After the combination of multiple regression trees, a sleep member awakening model based on XGBoost algorithm was established. The grid search algorithm is used to realize the automatic parameter tuning of the model and get the optimal parameter set of the model. The optimal parameters are substituted into the XGBoost model to improve classification performance. When processing the predicted value generated by the model, set 0.5 as the threshold value, greater than 0.5 output 1, or output 0;

Test sets were used to validate and evaluate the sleep member awakening model.

4.3. Experiment and Analysis

In this paper, the data of sleeping members provided by a drugstore affiliated to a drugstore company in Wuhan was selected as the experimental subjects. During the cooperation with the company, 15,640 data were obtained, including 3,267 data from sleep members and 12,373 data from non-sleep members.
Characteristics of the data set processing: remove duplicate data, to get rid of the dirty data, computer can't identify new arrive field is used to represent the sleep members whether to shop after receiving text messages, if the shop is marked as 1, before the shop flag is 0, will be and whether all replacement for 1s and 0s, according to the age, date of birth so calculate the new field delete has no effect on the user consumption behavior of indicators, such as card number, etc.

4.4. Evaluation indicators
Members discussed in this article is to sleep in the question of whether after received coupons to the store consumption abstraction for machine learning classification problem, two for binary classification problems, this paper choose the confusion matrix to evaluate the forecasting results classified as real class (TP) : forecasts is awakened and sleep members to store consumption, actually negative class (TN) : forecasts not awakened and not to actually shop sleep member, false positive (FP) : predict awakened but actually not to shop of sleep member, false negative (FN) : forecasts not awakened but actually the shop member of sleep. The confusion matrix is shown in the table:

| Confusion matrix | The real value |
|------------------|---------------|
| Predictive value | Positive      | Negative     |
| Positive         | TP            | FP           |
| Negative         | FN            | TN           |

Due to the particularity of sleep members, namely the statistics after the consumption of sleep membership is much lower than before the shop membership of sleep, led to the positive and negative samples data imbalance, negative samples than positive number of samples, combined with the actual situation, this paper USES the precise rate (precision), the recall rate (recall) and AUC three values to assess model. Precision rate refers to the proportion of the correct predicted to be positive in all the predicted to be positive. The calculation formula is: Precision = TP / (TP + FP). Recall rate refers to the ratio of the correctly predicted positive to the actual positive. The calculation formula is: recall = TP / (TP + FN). AUC is defined as the area under the ROC curve, which will not be greater than 1. Moreover, since THE ROC curve is generally located above the line y=x, the value range of AUC is generally between 0.5 and 1. As a numerical value, the greater the CORRESPONDING AUC value, the better the model effect will be.

4.5. Comparative experiment of sleep member model
In order to verify the effectiveness of the model selected in this paper on the awakening of sleeping members, a variety of binary classification algorithms and the optimal model proposed in this paper were selected to compare the prediction of the awakening of sleeping members under the same feature vector, and the XGBoost algorithm was compared with random forest, decision tree, Gaussian and Bernoulli algorithm in experiments. At the same time, in order to make the experimental results more accurate and reduce the impact caused by the imbalance of positive and negative samples, the method of stratified sampling was adopted to take the negative samples and the number of positive samples was 1:1. The ROC curve results of different models were as follows:
Fig 3. Algorithm flow chart

According to the figure above, compared with random forest, decision tree, Gauss, Bernoulli and other classification methods, XGBoost algorithm has significantly higher AUC value, better overall effect and higher accuracy rate. The experimental results show that the XGBoost algorithm also improves the accuracy, recall rate and harmonic average value compared with other classification algorithms. The main reason is that the XGBoost algorithm is the accumulation of all weak learner predictions and residuals correction and the weighted sum of multiple learners to make the final prediction. From the above results and analysis, it can be concluded that the XGBoost algorithm combined with the characteristics of the sleeping member proposed in this paper can effectively predict the awakening of the sleeping member.

5. Conclusion
Members mainly pharmacy in this paper, we study the sleep wake up problem, research purpose is based on sleep member has rich characteristics, establish training set test set, dig out what the characteristics of the members after received favourable activity information issued by the pharmacy of easily awakened to consumer, what kind of binary classification model can predict more effective member of sleep wake up. After the prediction, the coupon information will be distributed to the most likely sleep members. The cost of waking up a sleep member is about one quarter of the cost of adding a new member, so as to increase the number of pharmacy members and reduce the marketing cost as much as possible.

References
[1] Ye Qianyi, Rao Hong, Himeji book.Commercial sales forecast based on Xgboost [J]. Journal of nanchang university (science science edition),2017,41(03):275-281.
[2] Liu Fen, ZHAO Xuefeng, ZHANG Jinlong, Tang Qing.Research on consumers' willingness to use mobile coupons: From the perspective of personal characteristics and motivation [J].Management review,2016,28(02):93-102.
[3] Zhang Weiwei, LIU Dun, Jia Xiuyi.Journal of nanjing university of aeronautics and astronautics,2019,51(05):643-651.
[4] Xu Ning, La Lei. New retail coupon behavior prediction based on XGBoost [J]. Journal of Southwest Normal University (Natural Science Edition), 2019, 44(03): 101-105.

[5] Lu Ping, Chen Xiaotian. Prediction of network coupon usage based on gradient promotion tree model [J]. Science, Technology and Engineering, 2019, 19(18): 234-238.

[6] Xie Kun, Rong Yutian, Hu Fengping, Chen Huan, Yao Xiaolong. Based on the data integration of random forest algorithm [J/OL]. Computer Engineering: 1-15 [2020-11-15].

[7] Yang Ye-min, Zhang Huijun, Zhang Xiaolong. Study on interpretable visual analysis method of Random Forest [J/OL]. Computer Engineering and Application: 1-13 [2020-11-15].

[8] Zhao Y. Research on the development and Countermeasures of O2O commerce in traditional retail enterprises under the background of transformation [J]. IEEE Transactions on Engineering Management, 2017, 40(3), 90-92.

[9] AMROLJCHE N, YAN R L. A manufacturer distribution is sue: how to manage an online and a traditional retailer. Annals of Operations Research, 2016, 244(2): 257-294.

[10] LU Jia-ning. An Offline Transferable and Divisible Mobile Coupon Based on NFC [C]//Proceedings of International Conference on Computing, Mechanical and Electronics Engineering. Singapore: International Institute of Engineers, 2015.

[11] Song Guoqin, Liu Bin. Establishment and Application of Screen Skip Index based on XGBoost Feature Selection [J]. Journal of University of Electronic Science and Technology of China, 2018, 47(6): 921-926.

[12] Li Gen. Model of highway interleaved area convergence based on gradient enhancement decision tree [J]. Journal of Southeast University (Natural Science), 2018, 48(3): 563-567.

[13] Xu Shaocheng, Li Dongxi. [J]. Statistics and Decision-making, 2018, 48(18): 25-25. (in Chinese).

[14] Zhang G F. Membership management and health management [J]. China Pharmacy, 2007(09): 72-73.

[15] WU Jie, ZHANG Yulai, WANG Jianfen. Research on usage prediction methods for O2O coupons Proceeding of the 25th International Conference on Neural Information Processing. Cambodia: [s.n.] 2018: 175-183.