ISSN 2616-5775 Vol. 6, Issue 8: 165-170, DOI: 10.25236/AJCIS.2023.060820

Forecast of O2O Coupon Consumption Based on XGBoost Model

Weihan Yang[1,a,#], Zijie Zhang[1,b,#], Runze Liu[1,c,#], Jinjiang Liu[1,d,#]

[1]Beijing 21st Century International School, Beijing, China
[a]yangweihan211410@163.com, [b]zhangzj6621@163.com, [c]1668963885@qq.com,
[d]ljj20060824@163.com
([#]Co-first author)These authors contributed equally to this work.

Abstract: *A precise delivery of coupon is an important way to engage existing customers or attract new ones to physical stores in O2O marketing approach. And a suitable strategy of coupon distribution can significantly heighten the user experience and facilitate coupon re-consumption. In this paper, we design a prediction model of O2O coupon usage based on XGBoost and compare the performance of XGBoost with another model based on the average AUC value. By contrast, the XGBoost performs better than the other model with 0.9584 average AUC value. So this model can help merchants to locate the target accurately.*

Keywords: *O2O, Coupons, Machine learning, Prediction, XGBoost*

1. Introduction

With the improvement and the widespread use of mobile devices, the mobile internet has entered a high-speed development stage across various industries, with Online to Offline (O2O) consumption attracting the most attention. O2O commerce is a well-known business model that links offline business activities with online channels [1]. Utilizing coupons to reactivate existing customers or attract new customers to make in-store purchases is an important marketing method in the O2O industry. However, randomly distributing coupons creates meaningless disturbances for most users [2]. Hence personalized delivery of coupons is a crucial technology for improving coupon redemption rates, as it enables consumers who have specific preferences to receive some benefits. And in terms of the benefits that merchants receive is to improve coupon verification rate, giving merchants stronger marketing ability, helping merchants' precision marketing, increase user stickiness by locating costumers precisely, and helping merchants to have reasonable and rational coupon distribution strategies [3].

In this paper, we mainly discuss the question about predicting if the user will consume corresponding coupons within about 15 days by accurately analyzing the model and the given data under O2O circumstance. For this issue, the XGBoost model is chose to analyze this problem in 3 aspects: the user, coupon, and merchant.

This paper will introduce the designed model in several sections: Part 2 will talk about the work from others that related to our research. Part 3 is going to describe the data that computer studied from, feature engineering and the overview of algorithm; Part 3 will evaluate the result based on a specific standard. The paper will end up in Part 4 which will be the part for conclusions.

2. Related work

As for the coupons delivery, there are some researchers have already explored the system of coupon delivery. For example, the PiCoDa, it is special because it verifies a user's eligibility for a coupon, and it protects the vendor's privacy by not revealing the targeting strategy when compared with the conventional behavioral targeted advertising [4]. Paper[5] have established a new secure and targeted mobile coupon delivery scheme based on blockchain to enable the secure delivery of targeted coupons. Paper[6] shows that by contrast, men use fewer coupons than women. Pape r[7] make a comparison between online coupons and print coupons then found that E-coupons did not perform significantly better than off-line coupons. Paper[8] builds a coupon recommendation system to attract car users by using machine learning algorithms. Apart from coupon delivery, machine learning can also be used in many fields. Paper

[9] introduces the application of machine learning in biomedicine, to be specific, increase the accuracy of disease perception and diagnosis and promoting objectivity in the decision-making process. Paper[10] discusses the machine learning models in credit card fraud detection according to a specific dataset. Paper[11] uses machine learning algorithms to establish an advanced climate model to verify the global warming and determine the factors leading to global warming.

3. Data and algorithm

The second part mainly talk about the description of collected data, the features, and the algorithm of the model.

3.1 Data

The data set contains users' offline and online behavior from January 1 to June 30 in 2016, for 1754884 rows × 7 columns and 11429826 rows × 7 columns respectively. The data is used for XGBoost model to study and forecast whether a specific user will consume a coupon for 15 days period since the user received the coupon and setting up a probability model to determine the probability based on predicting positive sample(consuming the coupon) and the negative sample(not consuming the coupon). For this model, we mainly focus on the offline data and the attributes of data are explained in detail in Table 1.

Table 1: Attributes description

Attributes	Description
User_id	The id for each user, only containing the numbers, length of id is not fixed
Merchant_id	The id for each merchant, only containing the numbers, length of id is not fixed
Coupon_id	The id for each coupon, same id means the coupons are from same merchant, length is not fixed, null means purchase without coupon
Discount_rate	the discount, x within the range of 0 to 1 means the rate of discount; x:y means when the price reaches x then the price will minus y; the unit is Yuan.
Distance	The smallest distance between user's location and the store: a. 10 means more than 5 km b. x: x*500 meters in the range of 1 to 10(both inclusive) c. 0 means the distance is less than 500 meters d. null means the distance is not recorded
Date_received	date of coupon receiving, in the format of Y/m/d, null means not receiving the coupon
Date	The buying date: if coupon_id !=null and Date=null: it is a negative sample, user receives coupon but not use. if coupon_id = null and Date != Null: user buy the product without coupon, useless for the training. if the coupon_id!=null and Date!= null: positive sample, coupon is used when purchase.

3.2 Feature Engineering

Feature engineering to change the original data into training data in order to help the model to have better performance and approaching the best performance of the model[12]. Feature engineering plays an irreplaceable role in the model, and it is decisive to the final performance of model. It defines the upper limit of the model[3].

In order to obtain effective features, the data that are collected need to be processed before the feature engineering. The processing of data usually involves dumb coding of qualitative data, missing value processing, dimensionality reduction, etc. For the missing values, there are several ways to deal with. When the missing value occupies a considerable proportion of the total data, the columns or rows of the missing data can be abandoned, otherwise discarding the missing value will have a significant impact on the final result. When the missing value does not take up a large proportion, we can handle the missing value by filling data that related to the existing data. For instance, median, average number, or mode[3]. In this paper, we replace the missing data by ‘null’.

After handling the missing data, some features are generated based on the data.

Firstly, for the attribute Distance, converting each distance into a number between 0 to 10(both inclusive) without decimals, 10 means more than 5 kilometers; 0 means less than 500 meters; for the number between 0 to 10, it means the distance is 500*x meters; as for null value, converting it into -1.

Secondly, for the attribute Discount_rate, there are 2 sorts of discount, discount x within the range of 0 to 1 means the rate of discount, label 0 in feature discount_type; for the format of x:y, it means full x minus y when purchase, label 1 in feature discount_type. And labelling -1 for null discount rate. Then, converting the latter discount format into the former by 1 minus the quotient of second number and first number, assigning each rate into feature Discount_rate. After that assign the first and second number of second format discount into feature discount_man and discount_jian respectively.

Finally, adding a column to the model based on specific rules:

1) When Date equals null and coupon_id does not equal null, it is a negative sample and is assigned 0 to the label.

2) When Date does not equal null and coupon_id equals null, it is useless for machine learning, so it is denoted as -1.

3) When Date does not equal null and coupon_id does not equal null, it is recorded as 1.

After that, according to features above, some new features are formed and they are divided into 2 categories, which are described in Table 2.

Table 2: Feature description

Categories	**Feature**
Coupon	discount rate
	the requirement for the minimum expenditure available for discount
	the amount of money subtracted when meet the requirement
Other	whether the date of purchase is on weekend
	minimum distance between merchant to user
	Distinguish the given date is which day of a week
	the time interval from user’s last purchase to the newest purchase

3.3 Algorithm

The algorithm used in this paper is eXtreme Gradient Boosting package (XGBoost). Efficient linear model solver and tree learning algorithm are included in the package. And it is available for several objective functions, comprising classification, ranking and regression. Also the package is extendible which means users can easily define their own objectives [12].

The traditional Boosting Tree models uses only the first derivative information, so it is difficult to implement distributed training when training the nth tree in the light of the residual of the former n-1 trees are already used, while XGBoost performs a second-order Taylor expansion on the loss function and can use the multithreading of the CPU for parallel computing automatically. Next, we are going to briefly illustrate the algorithm and the parameters of XGBoost model [13].

3.3.1 Overview of algorithm

Assuming there are K trees in total, F is used to represent the basic tree model, integrating the tree model with addition method:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \tag{1}$$

The objective function is:

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

Where l is the loss function, representing the error between the predictive value and the true value.

After the second-order Taylor expansion of the objective function and some other calculations, the information gain of the objective function after each split is:

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{3}$$

As shown in (3), a splitting threshold γ is added to control tree’s growth and averting the overfitting. And the leaf node is only allowed to split when the information gain is greater than γ [13].

3.3.2 Parameter adjustment

Also, when using the XGBoost model, we adjust the parameters in order to help the model having better performance:

A. learning_rate

Affecting the performance of model heavily, in this model, we set the learning rate to 0.03.

B. n_estimators

The number of iterations when training, when this parameter is too small, it will cause underfitting. But when the parameter is too big, also leading to overfitting. We set this parameter to 1000.

C. max_depth

The maximum depth of the tree. Increase the maximum depth will make the tree model become more complicated. Although greater maximum depth makes the model becomes stronger, it is much easier to overfit. We set this parameter to 5.

D. min_child_weight

The sum of sample weight for the smallest leaf nodes, this parameter is designed to prevent overfitting. We set this parameter to 1.0.

E. subsample

The sampling rate of training samples. We set this parameter to 0.85.

F. colsample_bytree

The feature sampling rate when constructing each tree. We set this parameter to 0.8.

3.4 Result Evaluation

3.4.1 Evaluation method

A suitable evaluation method is very important for a model. For two classifications we can use accuracy rate, ROC curve, recall rate and AUC value to examine the quality of model[3]. In the paper, we use the average AUC values as the model evaluation indicator.

Table 3: Confusion matrix

		Forecast result	
		1	0
Actual result	1	TP	FN
	0	FP	TN

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

The ROC curve is to form a set of key points on the curve by continuously moving the cut-off point of the model in order to distinguish the positive and negative prediction results[3]. The curve is produced by plotting sensitivity(TPR) on the y-axis against 1-specificity(FPR) on the x-axis, which can be calculated from the confusion matrix (Table 3) and the calculation formulas are shown above [14]. In this paper, we choose ROC curve as the evaluation index because ROC curve is not affected by the distribution of positive and negative samples and can describe the performance of the model well.

AUC (Area Under Curve) is the area enclosed by Receiver Operating Characteristic Curve (ROC), the x-axis(FPR) and the y-axis(TPR). The normal value of AUC is between 0.5 to 1(both inclusive). When the AUC value equals 0.5, it means that the model is meaningless; When the AUC value equals the model is perfect. So the higher AUC value, the more accurate coupon consumption prediction that the model does (Figure 1).

3.4.2 Analysis of the result

In this experiment, the highest AUC is 0.9584. So the XGBoost model has the merit of high accuracy.

We also compare the effects and performances of model GBDT with XGBoost. GBDT is extensively used in machine learning, and it is a powerful tool in many sorts of applications, including flocculation process modeling, learning to rank, learning to transfer, multiclass classification and click prediction. Also it produces abundant cutting-edge results for data mining competitions. Decision trees are used as the fundamental learner and GBDT adds the predictions of a series of trees together. This model is famous for its efficiency, accuracy, and interpretability [15].

As shown in the Figure 1, the XGBoost model with AUC value 0.9584 is higher than adopting GBDT model(AUC value of 0.9580), hence the XGBoost produces a more accurate result compared to the GBDT model.

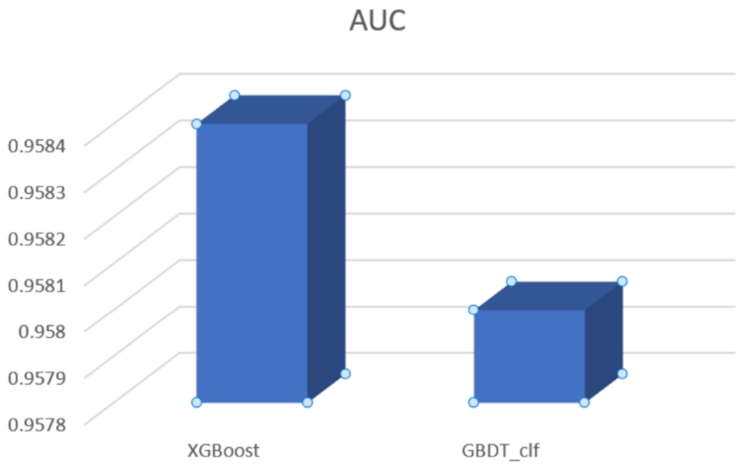

Figure 1: AUC of 2 models

4. Conclusion

This paper aims to predict the O2O coupon usage issues by building the XGBoost model and compare the performance of XGBoost model with the GBDT model. We use the XGBoost model to predict is the user will consume the coupon in 15 days after receiving the coupon. And the experiment shows that the XGBoost is superior to GBDT algorithm.

In the future, the performance of the model can be continually enhanced by changing the parameters of the model and feature engineering to make the prediction more effective and let the model run faster.

References

[1] Partridge, K., Pathak, M. A., Uzun, E., & Wang, C. (2012, July). Picoda: Privacy-preserving smart coupon delivery architecture. In Proc. HotPETs (pp. 94-108).
[2] Wu, J., Zhang, Y., & Wang, J. (2018). Research on Usage Prediction Methods for O2O Coupons. Neural Information Processing, 175–183. https://doi. org/10. 1007/978-3-030-04221-9_16
[3] Suo, Qiongyu. (2020). Prediction of O2O Coupon Usage Based on XGBoost Model. ICEME. https://doi. org/10. 1145/3414752. 3414775
[4] Alpar, P., & Winter, P. (2014). Comparison of redemption of print and electronic coupons. ACIS.
[5] Zhang Chunfu, Wang Song, Wu Yadong, Wang Yong, Zhang Hongying. Diabetes risk prediction based on GA_Xgboost model [J]. Computer Engineering, 2020, 46 (03): 315-320.
[6] Gu, Y., Gui, X., Xu, P., Gui, R., Zhao, Y., & Liu, W. (2018). A secure and targeted mobile coupon delivery scheme using blockchain. In Algorithms and Architectures for Parallel Processing: 18th International Conference, ICA3PP 2018, Guangzhou, China, November 15-17, 2018, Proceedings, Part IV 18 (pp. 538-548). Springer International Publishing.
[7] Harmon, S. K., & Jeanne Hill, C. (2003). Gender and coupon use. Journal of Product & Brand Management, 12(3), 166-179.
[8] Md. Abdul Hai, Rafsan Shartaj Uddin, Rahman, Y., & Raudatul Mahfuza. (2022). A Methodology for Recommending In-Vehicle Coupons Incorporating Machine Learning Algorithms for Efficient Financial Schemes. Proceedings of International Conference on Fourth Industrial Revolution and beyond 2021, 15–27. https://doi. org/10. 1007/978-981-19-2445-3_2
[9] Meherwar, Fatima, Maruf Pasha. (2017). Survey of Machine Learning Algorithms for Disease Diagnostic - Open Access Library. Www. oalib. com. https://www. oalib. com/paper/5282280
[10] Jiaxin, Gao, Zirui, Zhou, Jiangshan, Ai, Bingxin, Xia, Stephen, Coggeshall. (2019). Predicting Credit Card Transaction Fraud Using Machine Learning Algorithms - Open Access Library. Www. oalib. com. https://www. oalib. com/paper/5420837
[11] Harvey Zheng. (2018). Analysis of Global Warming Using Machine Learning - Open Access Library. Www. oalib. com. https://www. oalib. com/paper/5298942
[12] Chen, T., He, T., & Benesty, M. (2016). Xgboost: extreme gradient boosting.
[13] Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. Frontiers in Genetics, 10. https://doi. org/10. 3389/fgene. 2019. 01077
[14] Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? Emergency Medicine Journal, 34(6), 357-359.
[15] Zhang, Z., & Jung, C. (2020). GBDT-MO: Gradient-boosted decision trees for multiple outputs. IEEE transactions on neural networks and learning systems, 32(7), 3156-3167.