Prediction of O2O Coupon Usage Based on XGBoost Model

Shi Qiongyu
Beijing Jiaotong University, Beijing, China
18120618@bjtu.edu.cn

ABSTRACT

Precise and personalized coupon distribution is an important marketing method for merchants. A proper coupon distribution strategy can improve user experience and promote re-consumption. In this paper, an O2O coupon usage prediction model based on XGBoost is proposed. The experiment shows that the AUC value of XGBoost model is 0.82, which is better than random forest and logistic regression. The model can help merchants to develop the coupon distribution strategy purposefully and accurately locate the target population.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning algorithms; Ensemble methods; Boosting..

KEYWORDS

XGBoost, O2O coupons, Machine Learning, Prediction

ACM Reference Format:
Shi Qiongyu. 2020. Prediction of O2O Coupon Usage Based on XGBoost Model. In *2020 The 11th International Conference on E-business, Management and Economics (ICEME '20), July 15–17, 2020, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3414752.3414775

1 INTRODUCTION

With the improvement and development of Internet technology, mobile Internet plus all walks of life have entered the stage of rapid development, among which O2O (Online to Offline) is also an important trend, and has effectively influenced our daily life. O2O is an e-commerce development mode combining offline business opportunities with online platforms [1.]. Over the years, O2O has become one of the most popular e-business models [2.]. Nowadays, various APPs record massive amounts of user behaviors and location records. These precious data resources make big data more accurate in marketing. Using coupons to revitalize old users or attract new customers to consume is an important marketing method for O2O. However, the random placement of coupons causes meaningless disruption to most users; for merchants, indiscriminate coupons may reduce brand reputation, and it is difficult to estimate marketing costs. Technical assistance helps enhance the customer experience [3.]. Personalized coupon delivery is an important technology to improve coupon verification rate. It allows consumers with certain preferences to receive real benefits and gives merchants stronger marketing ability. Predicting the usage of coupons can help merchants' precision marketing, target groups precisely to increase user stickiness, and help merchants formulate reasonable coupon distribution strategies, which is of great significance to merchants.

Many scholars currently study the delivery of coupons. Paper [4.] pointed out that coupons can play a role in establishing customer loyalty or maintaining customer relationships, and can stimulate customers to purchase again. Paper [5.] empirically concludes that online coupons have a positive effect on promoting consumption. Paper [6.] proposed a model based on location as a coupon channel to discuss how location affects customer purchase behavior. The consumption of coupons has also been modeled through machine learning models. Paper [7.] analyzed the user types of coupons by clustering and tried to find out the distribution strategy. Paper [8.] uses gradient boosting trees to predict coupon usage behavior. At the same time, XGBoost is widely used in data competitions and predictions. The paper [9.] establishes a portfolio model based on Xgboost for predicting retail sales in Germany, and also applies this method to the domestic retail brick-and-mortar industry and even e-commerce platforms for sales prediction, which is of great significance for improving store operating models, product prices, distribution methods, and targeted precision sales.

In this paper, the question of whether a user can spend within 15 days after receiving a coupon is considered as a binary classification problem. The XGBoost model is developed to study this classification problem in three dimensions: user, store, and coupon.

The paper is organized as follows: Part 2 introduces data, data preprocessing, feature engineering, and algorithms; Part 3 introduces experimental results and comparative analysis with other models; Part 4 is conclusions, and the final part is the conclusion.

2 DATA AND MODEL

2.1 Data

The data comes from the Aliyun Tianchi Competition. The data set contains a training set of users' online and offline behaviors and a test set. The goal of the competition is to predict the probability that the user will consume within 15 days after receiving the coupon. This can be regarded as a binary classification problem, that is, predicting whether a user will consume within 15 days after receiving a coupon, and setting a probability threshold, beyond which a positive sample (consumption) is predicted, and a negative sample (no consumption) is predicted otherwise. This article focuses on the training set with the user's offline consumption date. The original fields of the training set are shown in Table 1.

The dataset has a total of 1754884 data, has 539438 users who have acted on the coupon, from January 1 to June 30, 2016 consumption data.

ICEME '20, July 15–17, 2020, Beijing, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8801-6/20/07…$15.00
https://doi.org/10.1145/3414752.3414775

Table 1: Field description

Field	Description
User_id	User ID
Merchant_id	Merchant ID
Coupon_id	Coupon ID: null means no coupon consumption, the subsequent Date_received and Date are meaningless
Date_received	Date of receipt of coupon
Discount_rate	Discount rate: x in [0,1] represents the discount rate; x:y means full x minus y. The unit is Yuan.
Distance	The distance between the user's most active location and the nearest store in the store: 1) 10: more than 5 kilometers 2) x: x * 500 meters, x in [1,10] 3) 0: less than 500 meters 4) null means no such information
Date	The date of consumption: if Date=null & coupon_id ! =null:negative sample, coupon received but not used; if Date! = Null & coupon_id = null: ordinary consumption, not used for training; if the Date!= null & coupon_id!=null: positive sample, date of purchase with coupon

2.2 Feature Engineering

Feature engineering is a process that transforms the original data into training data. Its purpose is to obtain better training features and make machine learning approach the upper limit of the model [10.]. Feature engineering plays a very important role in the model and has a crucial influence on the prediction results of the model. Feature engineering determines the upper limit of the model, and the adjustment parameters only help the model get closer to this upper limit.

Before the feature engineering, the data should be pre-processed. Data pre-processing usually includes dumb coding of qualitative data, missing value processing, dimensionality reduction, etc. There are several ways to deal with missing values. If the missing samples account for a very high proportion of the total, you can discard the column, otherwise, it will affect the results. If the missing value is not very large, it can also be supplemented by fitting the data based on the existing value or supplemented with an average, median, etc. This article fills in missing values with -1.

For the filed Distance, convert the distance to a number between 0 and 10. X in (0,10) means the distance is 500 * x meters, 0 means less than 500 meters, and 10 means more than 5 kilometers.

For the field discount rate, there are two types of discounts, and the two types need to be unified. It is better to convert the full reduction into a discount rate, and finally get the three variables of discount_rate, discount_man, discount_jian.

Finally, add a label column to the model. According to the rules, When Date = null and coupon_id != null, it is a negative sample and is recorded as 0; if Date != null and coupon_id = null, not used for training, it is recorded as -1; if Date! = null and coupon_id != null, which is a positive sample, denoted as 1.

According to the existing features, four categories of features are extracted, as shown in Table 2:

2.3 Algorithm

The algorithm used in this article is XGBoost. XGBoost (eXtreme Gradient Boost) is an open-source machine learning project developed by Chen Tianqi et al. XGBoost effectively implements GBDT algorithm and makes many algorithmic and engineering improvements [11.]. XGBoost is a typical Boosting algorithm in ensemble learning, which is also characterized by ensemble learning naturally. It unifies the results of multiple base classifiers into a final decision, which greatly improves the decision power of a single model.

The objective function of XGBoost is defined as:

$$obj\left(\theta\right) = L\left(\theta\right) + \Omega\left(\theta\right) \tag{1}$$

Among them, $L(\theta)$ is the loss function, which can be customized; $\Omega(\theta)$ is the regular term, which can control the complexity of the model and effectively alleviate overfitting.

Given a dataset $\{(X_i, y_i)\}, i = 1, 2, \cdots, N$, where $X_i$is a feature set and $y_i$is a label. To find the optimal tree for each iteration, you only need to optimize the objective function, that is, the following function:

$$obj^{(t)} = \sum\nolimits_i l\left(y_i, \hat{y}_l^{(t-1)} + f_k\left(x_i\right)\right) + \sum\nolimits_{k=1}^{t} \Omega\left(f_k\right) \tag{2}$$

Where, f_k represents the k-th tree and $\hat{y}_i^{(t-1)}$ is the optimal solution of the existing t-1 tree. The regular term is defined as:

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda \sum\nolimits_{j=1}^{T} \omega_j^2 \tag{3}$$

Where, T is the number of leaf nodes, and ω_j represents the predicted value of the j-th leaf. The second-order Taylor expansion of the loss function at $\hat{y}_i^{(t-1)}$ can be deduced as follows:

$$obj^{(t)} \approx \sum\nolimits_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2\right] + \gamma T \tag{4}$$

Table 2: Feature description

Categories	Feature
User	the number of coupons received by the user
	the number of times a user has used a coupon
	users' consumption times
	coupon verification rate per user
	the proportion of consumption with coupons per user
Merchant	the number of coupons merchant released
	the number of deals the merchant made with coupons
	the number of deals of the merchant
	the verification rate of coupon merchant released
	the proportion of deals made with coupon per merchant
User and Merchant	the number of coupons received by the user at the merchant
	the times of deal made with a coupon of user at the merchant
	the verification rate of the coupon of the user at the merchant
Coupon	type of coupons
	discount rate
	historical verification rate per type of coupon

$$G_j = \sum\nolimits_{i \in I_j} \nabla_{\hat{y}_l^{(t-1)}} l\left(y_i, \hat{y}_l^{(t-1)}\right) \tag{5}$$

$$H_j = \sum\nolimits_{i \in I_j} \nabla^2_{\hat{y}_l^{(t-1)}} l\left(y_i, \hat{y}_l^{(t-1)}\right) \tag{6}$$

Among them, I_j is all sample sets belonging to the leaf node$j$. Assuming that the structure of the decision tree is known, the optimal solution of formula (3) is the predicted value of each leaf node:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{7}$$

Substituting ω_j^* into formula (4), the minimum value of the objective function can be obtained:

$$obj* = -\frac{1}{2}\sum\nolimits_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{8}$$

It is easy to calculate at this time, the difference between the objective function before and after the split is:

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma \tag{9}$$

XGBoost adopts maximizing gain to construct the decision tree, traversing all the values of all features. It also played a certain pruning effect. XGBoost also explicitly adds regular terms to control the complexity of the model, which is helpful to prevent overfitting, thereby improving the generalization ability of the model.

3 RESULT AND DISCUSSION

3.1 Evaluation Index

Appropriate and correct model evaluation indicators are very important for a model. The parameter adjustment of the model is based on the evaluation index of the model. Selecting the appropriate index will improve the interpretability of the model and determine a correct direction for the parameter adjustment.

Table 3: Confusion matrix

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

There are many model evaluation indicators for the two classifications, such as accuracy rate, recall rate, ROC curve, and AUC value.

The accuracy rate is the proportion of correctly classified samples to the total samples. The recall rate is the proportion of correctly classified samples to true positive samples. The calculation formula is as follows.

$$Accuracy = \frac{TP}{TP + FP} \tag{10}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{11}$$

Roc curve is to generate a set of key points on the curve by dynamically moving the cut-off point of the model (that is, the threshold to distinguish positive and negative prediction results). The abscissa of the ROC curve is the False Positive Rate (FPR), and the ordinate is the True Positive Rate (TPR), which can be calculated based on the confusion matrix (Table 3). The calculation formula is as follows.

$$FPR = \frac{FP}{P} \tag{12}$$

$$TPR = \frac{TP}{P} \tag{13}$$

AUC (Area Under Curve) [12.] is the area under the Receiver Operating Characteristic Curve (ROC). The normal value of AUC ranges from 0.5 to 1. The closer the value is to 1, the better the model performance.

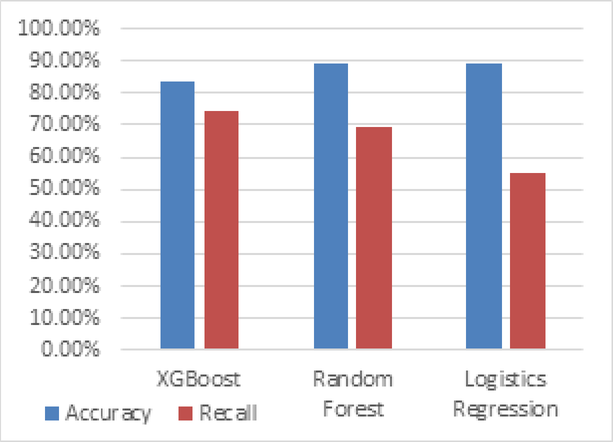

Figure 1: Accuracy rate and recall rate.

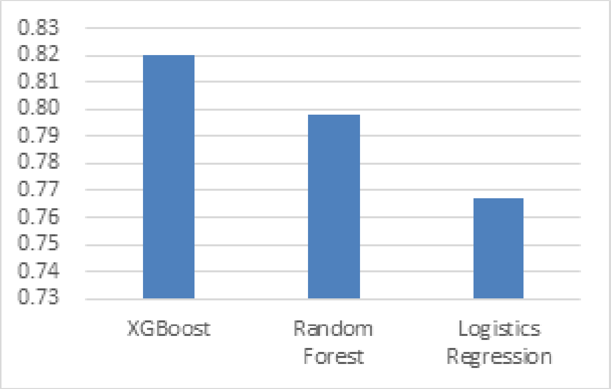

Figure 2: AUC of three models.

3.2 Analysis of Results

The data from January to May 2016 is selected for feature extraction, and the last month of data is used for modeling experiments. The experiment was done on the Jupyter Notebook of Anaconda, running on a PC with intel core i5 CPU and 8G RAM. In the experiment, accuracy is 83.37%, recall is 74.32% and AUC is 0.82. XGBoost has the advantage of high accuracy naturally.

In order to compare the effects of the models, this paper also constructed logistic regression and random forest models. Logistic regression is a classic classification model and a generalized linear regression model. Random forest belongs to the bagging method of integrated learning. It is a classifier that uses multiple independent CART trees to train and predict samples. Random forest can handle continuous and discrete variables and is suitable for multi-classification problems. Based on the decision tree-based base learner to build Bagging ensemble, it introduces random attribute selection in the training process of the decision tree [13.].

The accuracy, recall, and AUC values of the three models are shown in Figure 1 and Figure 2

Although the accuracy of the other two models is better than XGBoost, the recall rate does not perform well. This is because the proportion of positive and negative samples is unbalanced, and the accuracy rate only values the proportion of correctly classified samples, which is susceptible to the imbalance of positive and negative samples. But the AUC value of XGBoost is still the best, which also shows that the ensemble algorithm works better than a single model.

4 CONCLUSION

This paper builds the XGBoost model to predict O2O coupon usage problems and compares it with other models. A simple ensemble learning model is used to predict user usage within 15 days after receiving the coupon. The experiment shows that the performance of XGBoost is better than the traditional algorithms such as logistic regression and random forest.

In the future, the effect of the model can be improved in terms of feature engineering and model fusion, and the parameters can be adjusted by other methods to determine the model parameters accurately and quickly.

REFERENCES

[1.] Zhu ni. Research on the innovation of O2O marketing model in the era of big data [J]. E-commerce, 2019 (03) : 42-43.

[2.] Gaoyi Wu. 2018. On Theory and Application Studies of Online to Offline Platform Business Model in Upgrading and Transforming Traditional Industries. In *Proceedings of the 2018 9th International Conference on E-business, Management and Economics (ICEME 2018)*. Association for Computing Machinery, New York, NY, USA, 1–6. DOI:https://doi.org/10.1145/3271972.3271991

[3.] Guanzhen Wu, Li Cheng, and Liu Dong, "The Impact of E-Commerce on Customers' Purchasing Patterns in the Era of Big Data," Journal of Advances in Information Technology, Vol. 10, No. 3, pp. 109-113, August 2019. doi: 10.12720/jait.10.3.109-113

[4.] Clark R A, Zboja J J, Goldsmith R E. Antecedents of coupon proneness: a key mediator of coupon redemption[J].Journal of Promotion Management, 2013, 19(2) : 188-210.

[5.] Zhang jiantong, Fang chencheng. The influence of customers' historical behaviors and coupons on their purchasing decisions – based on an experimental study [J]. Soft science, 2017,31 (02) : 109-112.

[6.] ZOU Xiao, HUANG Kewei. Leveraging location-based services for couponing and infomediation[J]. Decision Support Systems,2015,78:93-103.

[7.] Huang Zheng, Sun Jingyang, Liu Danyang, Lu Hongyu, Hu Hongwei. Research on Coupon Issuance Based on Cluster Analysis on o2o Platform [J]. Modern Business, 2018 (12): 32-34.

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

[9.] Ye Qianyi. Research on physical retail sales forecast based on Xgboost method [D]. Nanchang University, 2016.

[10.] 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.

[11.] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*.Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785

[12.] Lei Yiming, Zhao Ximei, Wang Guodong, Yu Kexin. Cirrhosis recognition based on an improved LBP algorithm and over-limit learning machine [J]. Computer Science, 2017, 44 (10): 45-50.

[13.] Breiman L. Random Forests. Machine Learning, 2001, 45(1):5-32.