Dean&Francis
ISSN 2959-6157

Research on the Accurate Prediction and Optimization Strategy of "10 billion Subsidies" Activity

Yuxuan Chen

Department of Big Data, Xi'an Jiaotong-Liverpool University, Suzhou, China

*Corresponding author: Yuxuan.Chen21@student.xjtlu.edu.cn

Abstract:

The central objective of this essay is to meticulously investigate the "10 billion subsidies" campaign implemented by the Pinduoduo platform, with a specific focus on understanding the multifaceted impact that these subsidies exert on consumer behavior patterns and the overall financial performance, as measured by platform turnover. To delve deeper into this analysis, the essay endeavors to construct a sophisticated intelligent prediction system. This system is designed to harness the power of users' historical consumption behavior and their usage patterns of coupons, aiming to predict with a high degree of accuracy the likelihood of users availing themselves of the subsidies offer. Through a rigorous process of data cleaning and processing, the essay will identify and extract the most relevant features that influence subsidy usage. By integrating these features with advanced machine learning algorithms, the essay seeks to generate precise predictive models. Ultimately, the findings of this report are intended to serve as a critical decision support tool for the Pinduoduo platform, enabling it to refine and enhance the strategic effectiveness of the "10 billion subsidies" campaign, thereby maximizing its return on investment and solidifying its market position in the competitive e-commerce landscape.

Keywords: optimization strategy, Pinduoduo, machine learning, accurate prediction

1. Introduction

In recent years, Pinduoduo platform has risen rapidly in the highly competitive e-commerce industry by virtue of its unique "preferential low price" strategy. As one of the platform's core competitive advantages, the "10 billion subsidies" activity has successfully attracted the widespread attention and participation of many users. In the field of e-commerce, the implementation of subsidies policy is not only an important means for platforms to attract and retain users, but also a key strategy to increase platform turnover

ISSN 2959-6157

and market share. Therefore, how to achieve accurate positioning and personalized distribution of user subsidies through data analysis and user behavior research has become the core element of the successful implementation of the "10 billion subsidies" campaign.

The objective of this essay is to construct an intelligent forecasting system to accurately predict whether a user may use the subsidies or not by using the historical consumption behavior and coupon use data of Pinduoduo users. Through in-depth study of user behavior patterns, a more effective strategy is provided for the "10 billion subsidies" campaign, improving user participation and promoting the growth of platform turnover.

2. Related Work

In the realm of e-commerce, personalized recommendations and precision marketing have proven to be invaluable tools for increasing user engagement and boosting sales. Extensive research, including the work by Shreyas Mohapatra and S. P. Mishra [1], has been dedicated to predicting users' purchase intentions through the meticulous analysis of their behavioral data, with the ultimate goal of formulating effective recommendation and marketing strategies based on these insights. Machine learning algorithms, such as those utilized by Qianchen Xia et al. [2], have become the cornerstone of achieving accurate predictions of user behavior.

In the context of the "10 billion subsidies" campaign on the Pinduoduo platform, the literature has seen a surge in studies focusing on data mining and user behavior analysis. For example, Wang Xingfen et al. [3] and Peiyi Song and Yutong Liu [4] have contributed to the understanding of users' purchasing decisions by mining their shopping habits and browsing history. Additionally, D. Alghazzawi et al. [5] have explored the role of coupon usage in decision-making. Despite this, the area of subsidy usage behavior remains relatively underexplored, with the current project aiming to address this void by employing an intelligent prediction system, inspired by the work of Bichen Zheng and Bingwei Liu [6], to deeply understand users' subsidy consumption patterns.

The project's approach is multifaceted, involving the cleaning and processing of historical data, as well as the application of sophisticated machine learning algorithms, similar to those proposed by Wenle Wang et al. [7] and Lin Gan [8]. By integrating these methods, the project seeks to enhance the user experience and deliver substantial commercial value to the platform, akin to the sales conversion optimization analysis conducted by Kristiawan Nugroho et al. [9].

Moreover, the project draws on the broader application of machine learning in digital marketing, as discussed by Mithun S. Ullal et al. [10], and its specific application in e-commerce by Nikhil Kandekar [11]. Adarsh Srivastava et al. [12] have also highlighted the importance of machine learning algorithms in online marketing strategies, which resonates with the current project's objectives. By leveraging these algorithms, the project aims to provide a more nuanced "10 billion subsidies" distribution strategy for the Pinduoduo platform, aligning with the principles of causal machine learning for marketing strategies as proposed by Henrika Langen and M. Huber [13].

3. Method

The aim of this thesis is to accurately predict whether a user will use subsidies or not by constructing an intelligent prediction system that utilizes data such as historical consumption behavior and coupon usage of Pinduoduo users. To achieve this goal, the following methods are adopted:

3.1 Data Preprocessing

Data preprocessing is an essential phase in data analysis that has a direct impact on the efficiency and precision of later model training. For this research, the data preprocessing entails the following procedures:

First, the data is read by Panda's library and some unwanted columns are removed. Then, the data is processed for missing values. In this process, forward padding method is used to fill the missing values to ensure the integrity of the data.

Feature engineering is an important tool to enhance the performance of the model. Features are chosen, created, altered, and integrated to better accommodate the training and predictive needs of machine learning models. In this task, some unwanted columns such as "redemption_status", "start_date", "end_date ", "id", etc. are removed, and then training set and test set are detached from the data set, in which the proportion of the test set is 10%.

Data exploration is an important stage of data analysis, where a basis is provided for subsequent analysis and modeling by calculating the statistical values of the data (e.g., total number of rows, total number of columns, unique values, mean, media, plurality, and standard deviation, etc.) and evaluating the data for the presence of missing values.

3.2 Feature Selection

Feature selection is an important part of machine learning, which can help us filter out features from a large number of features that have a considerable influence on the pre-

dictive performance of the model. In this study, the following feature selection methods are adopted:

Statistical tests are used to evaluate the relationship between each feature and the target variable and select those features that are statistically significant.

The built-in feature importance assessment function of some machine learning models (e.g., Random Forest) is used to select features that have a significant influence on the predictive performance of the model.

3.3 Model Construction

In this study, two machine learning models are constructed to predict whether users will use subsidies:

Random forest is an integrated learning method that makes predictions by constructing multiple decision trees and averaging their results (for regression problems) or majority voting (for classification problems). A key benefit of Random Forest is its capability to effectively manage a high volume of input variables, obviating the need for feature selection. In addition, it can evaluate the importance of individual features for classification problems.

Random forest algorithm specific construction process is as follows:

(1) Using the self-help method's approach to random sampling, the decision tree selects the most suitable feature from a pool of m candidate features at a given node to split into its left and right child nodes. In the case of Random Forest (RF), the method is refined by randomly picking a smaller group of features, referred to as m', which is less than the total number m. The best feature for the node's subtree division is then determined from within this randomly chosen subset of m' features. This strategy effectively boosts the model's ability to generalize.

(2) According to the method described in (1), k decision tree models are trained and then combined to form a random forest model. The test set samples are input into this model to obtain the corresponding classification results. The random forest algorithm predicts the final result by aggregating the outcomes of each decision tree: in classification tasks, it adopts a straightforward majority rule voting mechanism, with the ultimate classification being based on the most frequent outcome across the ensemble of decision trees; in regression tasks, it utilizes straightforward averaging, with the final prediction being the average of the results from the ensemble of decision trees.

Logistic regression is a statistical learning method that is mainly used to deal with binary classification problems. The method transforms the output from linear regression into a probability within the range of 0 to 1 by employing a logistic function, and subsequently translates this probability into a categorical prediction using a predefined threshold. The main advantage of logistic regression is that it is simple and efficient, does not require much computational resources or scaling of input features, and is suitable for problems that provide probabilistic predictions. Logistic regression achieves categorization by mapping the basic linear regression model through a certain function. The linear function space is shown below:

Sigmoid function is introduced to map the linear function into probability. In Fig. 1 can be seen, the range of values of its x is the natural number R, y's range of values is [0,1]. The function itself is for the S-shape, f (0) = 0.5 and when the value of x is away from 0, the value of y will be very close to 0 or 1, so the value of y can be treated as the probability of certain extent, which is categorized by determining whether it is greater than 0.5.

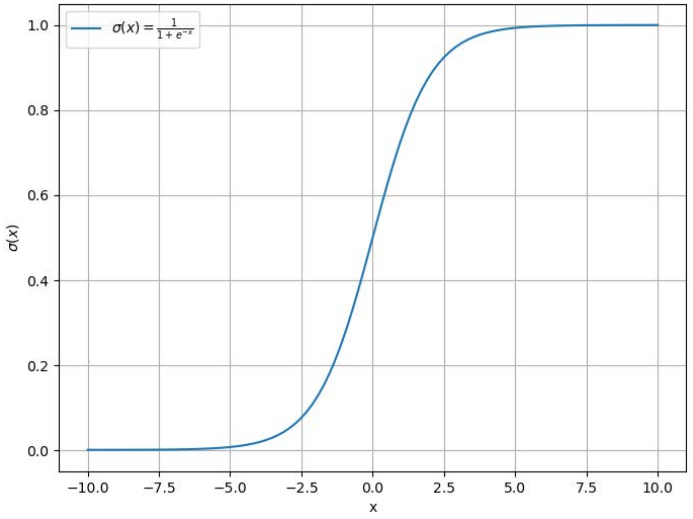

Fig. 1 Image of Sigmoid function

ISSN 2959-6157

3.4 Model Evaluation

To assess the efficacy of the model, the following metrics for evaluation are applied:

Area under the Receiver Operating Characteristic (ROC) Curve (AUC) is a commonly applied assessment metric for binary classification algorithms. "True positive rate" represents the proportion of actual positive instances correctly classified among all positive instances, while "false positive rate" indicates the proportion of actual negative instances incorrectly identified as positive within the total negative instances. The ROC curve is visualized by plotting "true positive rate" on the vertical axis and "false positive rate" on the horizontal axis. The AUC is the measure of the space under this curve. An AUC value of 0.5 suggests that the model's efficacy is equivalent to random guessing. Therefore, the AUC value should generally be greater than 0.5 but less than 1. In such cases, a higher AUC value signifies a better-performing model.

Accuracy (ACC) is the most commonly used performance metric for classification tasks, representing the ratio of correctly identified samples to the overall sample count. Accuracy is one of the most commonly used metrics for evaluating classification models, and it measures the overall accuracy of the model, which means the proportion of correct classifications made by the model on all categories.

4. Experimental Results

4.1 Description of the experimental dataset

The dataset size of the original data is 7836946. 'Redemption_status' is whether the user subsidies are being used or not, which is also the target variable of this project. The other features are X, such as 'Total Discount Amount', 'activity_type', 'start_date', 'end_date', 'activity_duration', etc. The data set is shown in Table 1:

Table 1. Data set

Id	Campaign_id	Customer_id	Redemption_status	brand	…
1	13	1053	0	1105	
2	13	48	0	56	
3	9	205	0	560	
4	13	1050	0	611	
5	8	1489	0	1558	
6	11	793	0	948	
7	9	590	0	133	
8	29	368	0	278	
…					

4.2 Data Preprocessing

Data preprocessing for this task consists of three parts: data cleaning, feature engineering and data exploration.

First, the data is read using Pandas library and some unwanted columns are removed. Then the data is processed for missing values. In this process, forward padding method is used to fill the missing values. This method will use the previous non-missing values to fill the missing values. Next, some unneeded columns are removed, such as "redemption_status", "start_date", "end_date ", "end_date", "id" and so on, and then the dataset is separated into training set and test set, in which the proportion of the test set is 10% (Table 2).

Table 2. Scale division of data sets

Training set	Test set
90%	10%

In the data exploration phase of this task, statistical values of the data such as total number of rows, total number of columns, unique values, mean, median, plurality and standard deviation are calculated, which are shown in Table 3. In addition, the data is checked for missing value cases (a total of 34,708 samples had missing value cases, representing 0.44 of the total sample size).

Table 3. Sample of individual missing values in the data

Feature	Amount
Id	7
Campaign_id	13
Coupon_id	644
Customer_id	1050
Redemption_status	0
Brand	611
Brand_type	0
Category	6
Cd_sum	17.81

4.3 Feature selection

Choosing the appropriate feature selection method has an important impact on the model prediction results. This task compares the performance of the model on the test set by different feature selection methods (using all features, feature selection, principal component analysis (PCA) dimensionality reduction) and evaluates the impact of different feature selection methods on the performance of the classification model in terms of ROC curves and AUC values. The comparison results are shown in Table 4, which shows that the feature filtered out of the individual features method yields the best AUC values.

Table 4. AUC values of different feature selection methods

Full features	Individual features filtered out by features selection	Individual features downscaled by PCA
0.9212	0.9254	0.8796

4.4 Prediction and Optimization Models for Pinduoduo Platforms

The comparison of the outcomes of various machine learning models on this task is illustrated in Table 5, for logistic regression the original model is used for prediction, while for the random forest model Bayesian search is used to regulate the hyperparameters of both models.

The table indicates that the Random Forest's ACC on the test set is 0.0001 superior to logistic regressions, and its AUC on the test set is 0.05 greater, which shows that the Random Forest model is better from an overall point of view.

Table 5. Comparative analysis of different machine learning models

Model Category	Training set ACC	Test set ACC	Training set AUC	Test set AUC
Logistic regression	0.9908	0.9903	0.9905	0.9171
Random forest	0.9908	0.9904	0.9905	0.9221

5. Conclusion

The ultimate goal of this project is to build an intelligent prediction system to forecast whether users will use subsidies or not by analyzing the historical consumption behavior and coupon usage of Pinduoduo users. By cleaning and processing the historical data, effective features are extracted and combined with machine learning algorithms to make accurate predictions. The whole process includes preprocessing the data, dividing the dataset into a training set and a test set, and using machine learning models such as random forest and logistic regression for training. The ultimate findings reveal that the random forest algorithm outperforms the logistic regression model in terms of predictive efficacy.

ISSN 2959-6157

References

[1] Shreyas Mohapatra, S. P. Mishra. Building Predictive Models for Market Research Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 2022, 1017-1022.

[2] Qianchen Xia, J. Lv, Shilong Ma, Bocheng Gao, Zhenhua Wang. A new information-theoretic method for advertisement conversion rate prediction for large-scale sparse data based on deep learning. Entropy, 2020, 22(6): 643.

[3] Wang Xingfen, Yan Xiangbin, Ma Yangchun. Research on User Consumption Behavior Prediction Based on Improved XGBoost Algorithm. IEEE International Conference on Big Data, 2018, 4169-4175.

[4] Peiyi Song, Yutong Liu. An XGBoost Algorithm for Predicting Purchasing Behaviour on E-Commerce Platforms, 2020, 1467-1471.

[5] D. Alghazzawi, Anser Ghazzaal Ali Alquraishee, Sahar Badri, S. Hasan. ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review. Sustainability, 2023, 15(9):7076.

[6] Bichen Zheng, Bingwei Liu. A scalable purchase intention prediction system using extreme gradient boosting machines with browsing content entropy. IEEE International Conference on Consumer Electronics, 2018, 1-4.

[7] Wenle Wang, Wentao Xiong, Jing Wang, Lei Tao, Shan Li, Yugen Yi, Xiang Zou, Cui Li. A User Purchase Behavior Prediction Method Based on XGBoost. Electronics, 2023, 12(9):2047.

[8] Lin Gan. XGBoost-Based E-Commerce Customer Loss Prediction. Computational Intelligence and Neuroscience, 2022, 2022 (9):1-10.

[9] Kristiawan Nugroho, Th. Dwiati Wismarini, Hari Murti. Sales Conversion Optimization Analysis Using the Random Forest Method. SinkrOn, 2023, 8(4).

[10] Mithun S. Ullal, Iqbal Thonse Hawaldar, Rashmi Soni, Mohamad Nadeem. The Role of Machine Learning in Digital Marketing. SAGE Open, 2021, 11(4):1-12.

[11] Nikhil Kandekar. Ecommerce Assisted by Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 2022, 10(3):351-353.

[12] Adarsh Srivastava, Mr. Lokendra Singh Umrao, Mr. Dilip Kumar. Application of Machine Learning Algorithms in Online Marketing. International Journal for Research in Applied Science and Engineering Technology, 2023, 11(3):2306-2312.

[13] Henrika Langen, M. Huber. How causal machine learning can leverage marketing strategies: Assessing and improving the performance of a coupon campaign. PLoS ONE, 2022, 18(1):e0278937.