Analyzing Factors of Users Click Behavior on Ads Based on Logistic Regression and Machine Learning

Sitong Zhou(✉)

College of Urban Transportation and Logistics, Shenzhen Technology University, Shenzhen 518118, China
202004010102@stumail.sztu.edu.cn

Abstract. With the development of digital marketing, e-commerce industry is gradually increasing the market share it occupies, so one of the most concerned things for e-commerce platform is the users' behavior of clicking online advertising. High clicks indicates that users have greater possibilities to buy products. There are several methods for analyzing click behavior, while less of them value e-commerce and specific features of users and ads. Among these methods, Logistics Regression (LR) was adopted most before, but it can only analyze linear relationship. Based on the Taobao users data of AliCloud Tianchi, Decision Tree Model of machine learning is proposed innovatively in theis paper to explore the impact of different users characteristics on clicking behavior and compares Decision Tree Model with LR. The results present that gender, age, consumption level and brand are filtered out by both two methods to build model because of these variables' significance. Differently, Decision Tree Model analyze users characteristics precisely and in more detail, and it performs better in handling nonlinear and complex relationships. Additionally, from time's perspective this paper find that click-through rate (CTR) may be higher when people are spiritually active and may be in connection with users' life-shopping cycles instead of weekends. This article can provide guidance to e-commerce platforms' personalized and high-efficient online ad placement strategies to improve the competitiveness of platforms and users conversion rates and maximize e-stores' benefits.

Keywords: Online advertising · Click behavior · Logistic Regression · Machine Learning

1 Introduction

1.1 Background

In the recent years, the e-commerce industry is becoming active increasingly. With the outbreak of epidemic, offline business activities are affected to varying degrees. By December 2021, China's online shopping usage rate reached 81.6% and online retail sales of physical goods accounted for 24.5% of total retail sales of goods in 2021.

G. Ali et al. (Eds.): ISEMSS 2022, ASSEHR 687, pp. 2538–2549, 2022.
https://doi.org/10.2991/978-2-494069-31-2_299

Therefore, it is obvious that e-commerce is becoming more and more important and the competition among e-commerce platforms is getting tougher.

To attract customers and boost click-through rate (CTR) quality and appropriate online ads are necessary. And more clicks and higher CTR considerably impact on users conversion rate, which refers to more people tend to spend money on e-stores. Simultaneously, there will be a remarkable enhancement in the product sales volume of e-commerce platforms. Besides, user-targeted personalized advertisements can not only upgrade the competitiveness of e-stores, but also make it more user-friendly for users to spend less time and discover the products they need. That is precisely why it is necessary to study users click behavior on ads.

1.2 Related Research

With the development of the Internet, online shopping has become an important part of people's lives. Thus, online advertising is an essential part of marketing, and the study of clicking and CTR has become a popular research direction.

First, many previous studies revealed that the rich and appropriate contents of advertisements have a great influence on clicking advertising behavior and this conclusion indicates the importance of personalization. In 2015, Liu et al. explored the factors influencing the CTR of products showed on the shopping platforms by empirically analyzing the data based on the formula of CTR, and they found that product's name and product's discount-level showed on web page impact CTR significantly. They also believed that people are more likely to click ads because of the certain information of ads [1]. To study the feature of ads Kim et al. applied LR Model in 2016 and they detected that people tend to click on ads of Facebook which they perceive as informative [2]. Moreover in 2016, Zanjani and Khadivi proposed a model based on the search engine to extract features of new ads and they observed that users' clicking behavior is related to the keywords and brands of ads [3]. And in 2019, based on the information processing theory, Mattke studied several concrete features including informativeness, personalization and so forth and discovered these characteristics affect user's clicking behavior remarkably [4].

Second, some researches focused on the click behavior influenced by users' characteristics. Through field study, Haans et al. investigated CTR of search engine. They discovered that the high involved Internet users are more inclined to click on ads [5]. Similarly, in 2016, through LR model Kim also found that users with a high level of involvement tend to click ads in Facebook [2]. In addition, a multivariate testing method was proposed by Higgins et al. to explore the effect of users' age and gender congruity on click behavior in 2018. They showed that a combined age and gender consistency has more significant CTR than inconsistency online advertisements [6]. In 2020, Lu and Chen introduced browsing situation routineness to explore the impact of browsing situations on users' advertisement clicks, and they detected that users have higher clicks during unconventional time [7]. Also in 2020, through the statistics, Asogwa et al. discovered that gender not only influences the duration of engagement in social platforms, but also be affected the popularity of products in online ads [8].

Third, many researches were based on the Logistic Regression Model to study the impact of users' clicking behavior on the online ads. For example, Wang et al. presented a Multiple Criteria Linear Programming Regression (MCLPR) prediction model in 2013

and they found that MCLPR is a promising and accurate model in behavioral targeting tasks [9]. In 2015, Kumar et al. applied LR to predict CTR on the search engine and their research indicated that LR can achieve about 90% precision on CTR estimation [10].

To sum up, most researches were focus on the search engines and social platforms, and ignored the importance of e-commerce industry. Especially, very few papers studied clicks behaviors from users' perspective. Even if some papers examined this field, they just only explored single features instead of several specific features. Besides, LR can only deal with linear relationship and it is unsuitable to handle complex relationships.

1.3 Objective

LR was used frequently to study CTR prediction and most of the previous research studied users' click behavior on ads in terms of search engines, social platforms and the characteristics of ad itself. The main purpose of this study is to explore the influencing factors of users click pattern from both users' perspective. To analyze the relevant data of Taobao LR and Decision Tree Model in machine learning are adopted and compared in this article. Then the coefficients, significance and the decision tree node of each factor can be obtained, respectively. Finally, to maximum the profit of e-stores and guarantee the effectiveness of ad placement, the study hopes to provide e-commerce platforms with relatively high reliability of individualized operation strategies according to the results.

2 Method

2.1 Data Source and Processing Tools

This study adopts datasets about advertisement and users' profiles of Taobao provided by Aliyun Tianchi to do the test and all the data are desensitized. As the most dominant shopping platform, Taobao, one of the largest C2C e-commerce platforms in China, has rich data of users and ads. Consequently, the data of Taobao can reflect users' behavior more truly. This article combines three excel files of Taobao datasets which include raw sample dataset, ad feature dataset and users profile dataset and deletes meaningless null values. SPSS is used mainly to figure datasets and build models based on the LR and Decision Tree Model.

2.2 Models

This study is divided into three parts to analyse why consumers click ads and what are the relevant factors. To gain the CTR, the equation is defined as follow:

$$\mathrm{CTR} = \frac{\text{clicks}}{\text{display}} \tag{1}$$

2.2.1 Users Characteristics

In this study there are several variables which may have a combined effect on click behavior and the function may be nonlinear, so the decision tree may be more accurate to estimate what parameter plays a significant role. And comparing Decision Tree Model and Logistic Regression Model is meaningful.

This paper formally testes impact of six users characteristics including gender, age, consumption level, shopping depth, occupation and brand as independent variables. These factors are general features of people in their lives, so the true reflection will be obtained. Independent variables and dependent variable are identified by numbers in the Table 1. Specially, as shown in Table 1, age independent variable has been divided into seven grades with specific age number in the raw dataset. Brand refers that if people tend to click an branded advertisement. Branded ad and non-branded ad equals 1 and 0, respectively.

In the Logistic Regression Model:

$$\begin{aligned} \text{Click} = {} & \alpha + \beta_1 \text{Gender} + \beta_2 \text{Age} \\ & + \beta_3 \text{Consumption Level} + \beta_4 \text{Shopping Depth} \\ & + \beta_5 \text{Occupation} + \beta_6 \text{Brand Tendency} \end{aligned} \tag{2}$$

where Shopping Depth (new, general, regular) means the length of time customers spend in an e-store.

A growth method of decision tree, named Chi-squared Automatic Interaction Detector (CHAID) is daopted because it is suitable to classify variables and handle nonlinear problems. The principle of this model is that overall data is classified by the algorithm and decision nodes generates until the data can not be categorized. This paper sets the proportion of learning sample and test sample was 70% and 30%, respectively.

In the CHAID of Decision Tree Model:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \tag{3}$$

where n, f_i, p_i and np_i refers to sample size, frequency of the sample occurrence, probability of a sample occurrence and theoretical frequency number of the sample, respectively. This equation describes the classification criteria algorithm of constructing decision tree model.

2.2.2 Time

To explore CTR at different time of a day and different days of a week, this article extracts needed data and recombines data. Then time range and day are set as independent variables and CTR is set as dependent variable. Each independent variable has its own click and CTR. Besides, time of a day is divided into twelve parts averagely and each parts has two hours.

2.2.3 Ads Characteristics

Logistic Regression Model is built as follow:

$$\mathrm{CTR} = \alpha\mathrm{Price} + \beta\mathrm{Brand} \tag{4}$$

This study identifies the price: range from 0 to 150 equals 1, range from 150 to 300 equals 2, range from 300 to 450 equals 3, range from 450 to 600 equals 4, range from 600 to 1000 equals 5 and price range which is greater than 1000 equals 6. Identification of brand shows in Table 1.

3 Results

3.1 Users Characteristics

3.1.1 Results of Logistics Regression Model

Table 2 demonstrates that the p-value of the whole fitted model is less than 0.01, which indicates this model has a very significant statistical meaning. And because p-values of gender, age, consumption level and brand are all less than 0.01, these several independent variables contribute remarkably to this model.

As displayed in Table 3, based on the last classification of every independent variable, the positive and negative of coefficient indicate the positive and negative correlation between different variables and click behavior. Standard error represents the mean error of the estimate value. Odds ratio (OR) refers that each classification compares with the reference quantity in an unchanged condition and each variable takes its last classification as a baseline value which equals 1. Since the click scale which between men and women is about 0.941:1 in the OR column, women prefer to click online ads. And as age increase from grade 0 to grade 6, Fig. 1 displays that the probability of clicking drops first and then upgrades. This phenomena indicate that people of the middle-age group are busier and have not much time to entertain. Because the probability of Consumption Level decrease gradually from 1 to 3, it indicates that people buying low and medium price products tend to click ads online. And it is possible that people who have high consumption level think propaganda of goods is not very worthy to believe online. Through OR this paper also detects that customers are likely to click non-branded ads. It can be predicted that the appearance of this situation is due to non-branded products with lower price usually considered cheaper.

Given these results, formulating operation strategies and personalized recommendation in terms of different gender, age, consumption level and brand is very useful for e-stores to ensure the investment on advertisements highly targeted. For instance, to ensure accurately delivery ads analyzing preference of females for products is the focus because ads are more attractive to femals. And for users' age, increasing further through matching preference of grade six users is potential because this type of users has the most click behavior. While these methods need more detailed market researches. And from shopping depth perspective, holding more promotion activities to grasp customers and stimulate customers' desire to browse and buy would be helpful. In short, for e-commerce platforms, it is worth formulating and adjusting advertisements delivery strategies from users' perspective.

Table 1. Parameter settings

Item	Digital Conversion		
Click	Not click = 0	click = 1	
Gender	Man = 1	Woman = 2	
Consumption Level	Low = 1	Medium = 2	High = 3
Shopping Depth	New = 1	General = 2	Regualr = 3
Occupation	Undergraduate = 1	Non-graduate = 0	
Brand	Non-brand = 0	Brand = 1	

Table 2. Model Fitting Result

Item	P-value
Logistic Regression Model	0.000
Gender	0.000
Age	0.000
Consumption Level	0.000
Shopping Depth	0.322
Occupation	0.491
Brand	0.000

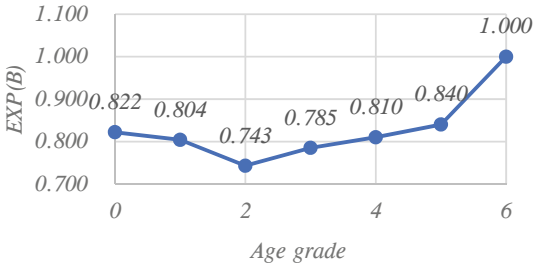

Fig. 1. The Probability of Clicking Ads

3.1.2 Results of Decision Tree Model

Concrete information of this model is displayed in Table 4, which consists of specification and results. To obtain an accurate classification and results, the study sets minimum cases of parent node and child node are 600 and 300, respectively. Because data and feature values are relatively plenty, maximum tree depth is needed to set to handle core module of the prepruning which can improve performance of the decision tree. The results demonstrate that brand, gender, age and consumption Level eventually are filtered. This phenomena confirms that these indicators play an important role on click behavior.

Table 3. Descriptive Statistical Data of Different Independent Variables

Click	Coefficient	Standard Error	P-value	OR
Gender=1	-0.067	0.016	0.000	0.935
Gender=2	---	---	---	1
Age=0	− 0.196	0.461	0.671	0.822
Age=1	− 0.219	0.065	0.001	0.804
Age=2	− 0.297	0.051	0.000	0.743
Age=3	− 0.242	0.048	0.000	0.785
Age=4	− 0.211	0.049	0.000	0.810
Age=5	− 0.174	0.049	0.000	0.840
Age=6	---	---	---	1
Consumption Level=1	0.107	0.029	0.000	1.112
Consumption Level=2	0.101	0.027	0.000	1.106
Consumption Level =3	---	---	---	1
Shopping Depth=1	− 0.090	0.148	0.544	0.914
Shopping Depth=2	0.049	0.036	0.169	1.105
Shopping Depth=3	---	---	---	1
Occupation=0	0.028	0.040	0.492	1.028
Occupation=1	---	---	---	1
Brand=0	0.112	0.019	0.000	1.119
Brand=1	---	---	---	1

As shown in Fig. 2, Decision Tree Model of training sample indicates the visualized interactions and nonlinear relationships among the variables represented by each root note. It can be seen that the probability of the users' click behavior in the root node is 4.8%. The first layer is classified in terms of the brand of 1 and 0. Branded and non-branded ads account for 82.4% and 17.6%, respectively. On the contrary, Users' clicks on the non-branded ads with 5.3% slightly higher than branded ads with 4.7%. In the second layer, node 1 is classified in terms of gender of 1 and 2 because gender plays a decisive classification role on branded ads. The proportion of females with 52.2% is higher than males with 30.2%, and females also click more. Node 2 is categorized according to the consumption level which is divided into two parts: one is grade 1 and 2, another is grade 3. It suggests that low and middle consumption level users have higher click probability. Then the third layer is split down in node 4 and node 6. Node 4 is categorized according to the age which is divided into three parts: the first is grade 1 and 5, the second is grade 6, and the other age grade belongs to the third parts. The results discover that male users of grade 6 have the highest clicking proportion with 6.4% and users of grade 1 and 5 have the lowest clicking percentage with 5.2%. Node

Table 4. Model Summary

Specification	Detailed information
Growing Method	CHAID
Dependent Variable	click
Independent Variables	Gender, Age, Consumption Level, Shopping Depth, Occupation, Brand
Validation	Split Sample
Maximum Tree Depth	3
Minimum Cases in Parent Node	600
Minimum Cases in Child Node	300
Results	Detailed information
Independent Variables Included	Brand, Gender, Age, Consumption Level
Number of Nodes	12
Number of Terminal Nodes	7
Depth	3

Table 5. Classification

Sample	Predicted	Risk	
		Estimate	Standard Error
Training	95.2%	4.8%	0
Test	95.0%	5.0%	0.1%

6 is categorized in terms of gender of 1 and 2. Thus, click behavior of males with 5.4% is significantly more than females with 3.6% for high consumption level users.

Table 5 reveals the comparison between training sample and test sample, in which the model accuracy rate is 95.2% and 95.0%, respectively. Risk column shows that the risk estimation of training sample and test sample is 4.8% and 5.0%. This result verify the high precision of Decision Tree Model.

This model picks out independent variables to put into analyzing through the size of contribution and it details the users features, which is more intuitive, more personalized and more effectively digital-driven. Furthermore, it is a good way to handle complex links by this model and it has relatively high credibility in analyzing influencing factors. Based on the results, this paper proposes that formulating detailed operation strategies according to different characteristics of users and characteristics subdivided in each feature will help platforms to achieve maximum personalized service and increase the probability of clicking behavior.

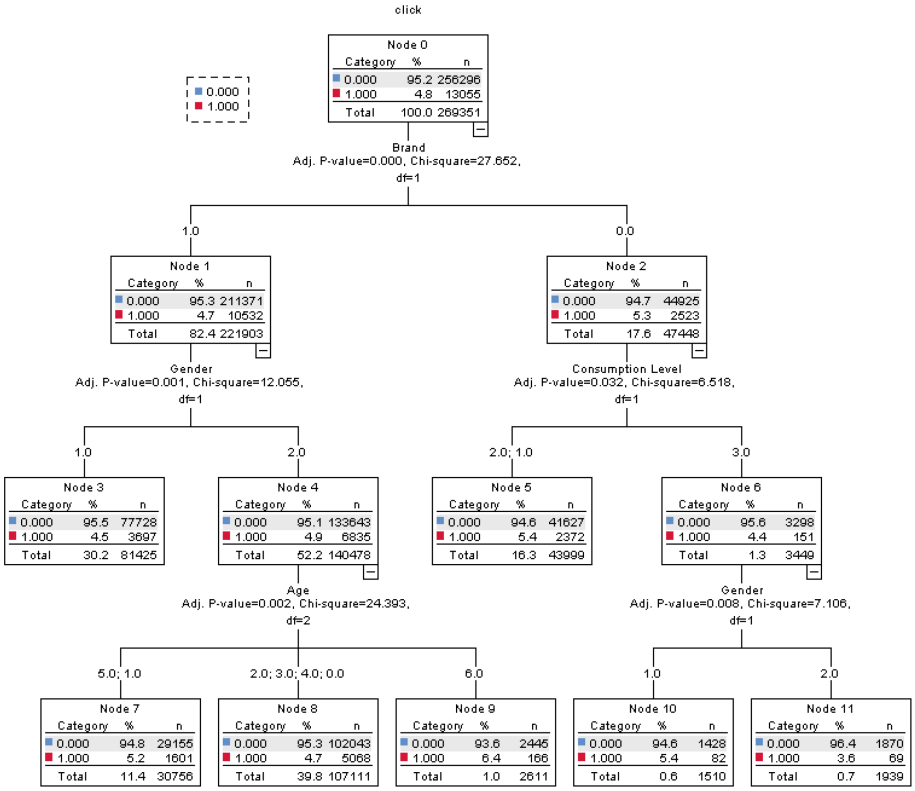

Fig. 2. Training Model

3.2 Time of Users

As illustrated in Fig. 3, the number of clicks from 8 p.m. to 10 p.m., from 10 a.m. to 16 p.m. and from 10 p.m. to 12 p.m. are relatively high. This phenomenon reflects that clicks of ads may be related to users' timetable and people tend to browse ads during lunch and night break. The results detect the clicks of period range from 2 p.m. to 4 p.m. is higher than time range from 8 a.m. to 10 a.m., which may indicates people are busier in the morning and have less time to browse ads. Specially, the CTR of morning is much higher than evening, which is opposite with clicks. From 10 p.m. to 12 p.m. CTR is also opposite with clicks. These phenomenons demonstrate that people may be sleepy in the afternoon and late night, in which they have no big desire to click, though they browse many ads. While in the morning, lunch and evening rest, people are mentally excited and the possibility of clicking is greatly increased. In light of this results, putting the right number of online ads at a right time is helpful to platforms. This method can match better with users' time and increase the amount of CTR. High CTR means more users click ads under the same amount of ads placement and users are attracted easily.

Figure 4 illustrates that CTR of Tuesday, Thursday and Saturday is higher.. The results show clicks of Tuesday is less than its CTR, so ads placement is the most effective

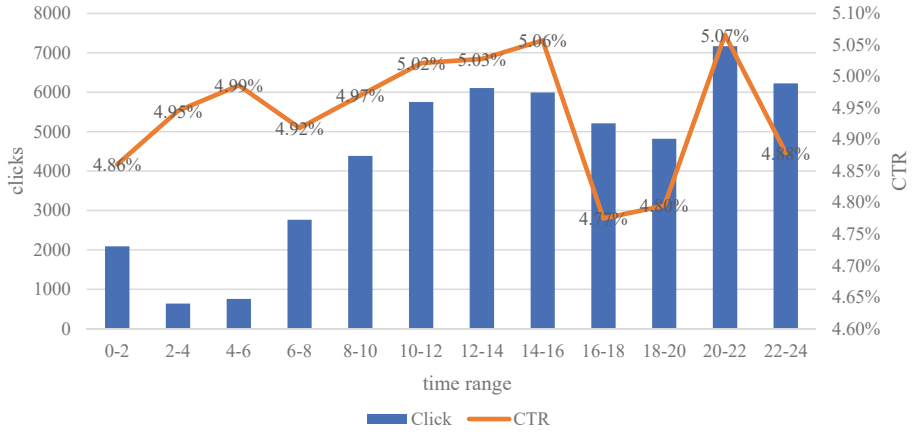

Fig. 3. CTR of Different Time in A Day

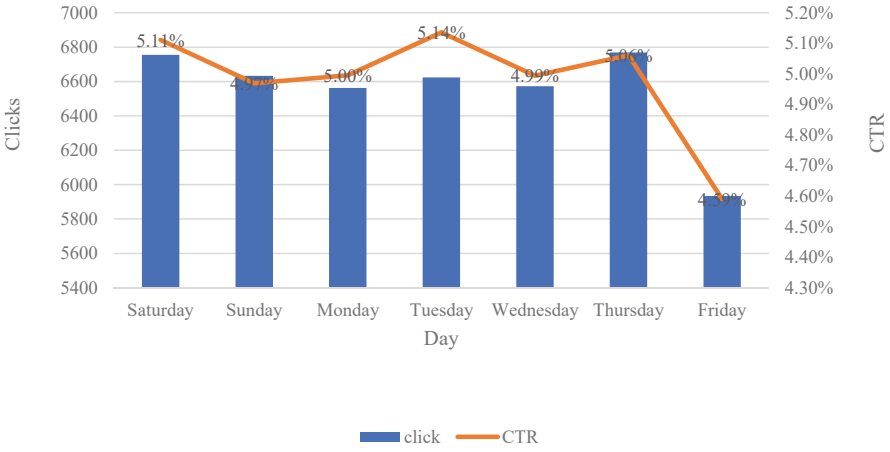

Fig. 4. CTR of Different Days in A Week

on Tuesday. And though Sunday is weekend, clicks and CTR are both the lowest in a week, so the study speculate that people are busy preparing working and studying plans before a new week on Sunday. While Saturday is the first day after a week, people tend to relax themselves. Besides, through the results, there is an interesting phenomenon that browsing ads and shopping on Tuesday, Thursday and Saturday may be a shopping cycle in the daily life of people. E-commerce platforms can propaganda more promotional activities and put more ads online in the peak period of clicking and CTR. This recommendation can not only improve users' liveness and increase the possibility of commodities sales, but also can decrease unnecessary capital investment. Furthermore, platforms can hold promotion activities on weekend to increase the profits on Saturday and encourage people to browse ads and buy goods on Sunday.

Table 6. Logistic Regression Results

Item	Coefficient	Standard Error	P-value
Brand	-0.086	0.019	0.000
Price	-0.042	0.005	0.000

3.3 Ads Characteristics

As shown in Table 6, p-values of brand and price are both less than 0.01, so these two variables are significant in statistics. And they are both negatively correlated with click behavior. Thus, reducing the price and display of brands in ads would be a good method for platforms. And it is useful to adjust the information of ads online as users demand and browsing behavior change to guarantee users would be interest in ads.

4 Conclusion

This paper utilizes Logistic Regression to investigate the influence of six users features and two ads characteristics on clicking advertising behavior of users. A new idea is proposed to explore the impact of concrete users characteristics on clicking pattern based on the Decision Tree Model in this paper. From users' perspective, the study finds that p-values of gender, age, consumption level and brand are all significant statisticsly in LR. Similarly, above the four variables are included in the analytical model of decision tree. By comparing two models, the results both indicates that these four factors are sensitive to clicking behaviors on online ads. While the difference is that Decision Tree Model provides a more refined analysis of individual variables and can analyzes complex relationships between multiple independent variables and dependent variable. Thus, decision tree is more accurate and better than LR.

The clicks and CTR at different time of a day and different days of a week are also studied by visualized bar-line graphs. And this article detects that users click online ads at a high rate in the morning, lunch and evening break. This may due to people active mentally in the above three time ranges. The study also discovers that CTR is higher on Tuesday, Thursdays and Saturday, which can be anticipated click behavior is related to shopping cycles of customers rather than whether it is a weekend.

Therefore, this paper recommends that e-commerce platforms should not only pay more attention to the above four users factors, but also consider decision tree and users' sensitivity to the price and brand of online ads based on the timetable of users to develop a more personalized online ad placement strategy. This recommendation can help e-stores to make precise investment and facilitate maximization of e-stores' benefits.

References

1. Y. Liu, P. Yuan, W. Liu, & X. Li (2015). What drives click-through rates of tourism product advertisements on group buying websites?. Procedia Computer Science, 55, 221-230.
2. Y. Kim, M. Kang, S. M. Choi, & Y. Sung (2016). To click or not to click? Investigating antecedents of advertisement clicking on Facebook. Social Behavior and Personality: an international journal, 44(4), 657-667.
3. M. Daryaie Zanjani, & S. Khadivi (2015). Predicting user click behaviour in search engine advertisements. New Review of Hypermedia and Multimedia, 21(3-4), 301-319.
4. J. Mattke, C. Maier, L. Reis, & T. Weitzel (2021). In-app advertising: a two-step qualitative comparative analysis to explain clicking behavior. European Journal of Marketing.
5. H. Haans, N. Raassens, & R. van Hout, (2013). Search engine advertisements: The impact of advertising statements on click-through and conversion rates. Marketing Letters, 24(2), 151-163.
6. S. F. Higgins, M. D. Mulvenna, R. B. Bond, A. McCartan, S. Gallagher, & D. Quinn, (2018). Multivariate testing confirms the effect of age–gender congruence on click-through rates from online social network digital advertisements. Cyber psychology, Behavior, and Social Networking, 21(10), 646-654.
7. X. Lu, & Y. Chen, (2020). Situations Matter: Understanding How Individual Browsing Situation Routineness Impacts Online Users' advertisement Clicks Behavior. Journal of Electronic Commerce Research, 21(2).
8. C. E. Asogwa, S. V. Okeke, V. C. Gever, & G. Ezeah, (2020). Gender disparities in the influence of social media advertisements on buying decision in Nigeria. Communication: South African Journal of Communication Theory and Research, 46(3), 87-105.
9. F. Wang, W. Suphamitmongkol, & B. Wang, (2013). Advertisement click-through rate prediction using multiple criteria linear programming regression model. Procedia Computer Science, 17, 803-811.
10. R. Kumar, S. M. Naik, V. D. Naik, S. Shiralli, V. G. Sunil, & M. Husain, (2015, June). Predicting clicks: CTR estimation of advertisements using logistic regression classifier. In 2015 IEEE international advance computing conference (IACC) (pp. 1134-1138). IEEE.