Research on the Influencing Factors of E-commerce Conversion Rate Based on Regression Analysis and CART Decision Tree Algorithm in the Context of Big Data

Xuanyi Zhu*

Business School, University of Shanghai for Science & Technology, Shanghai, China, 200093

*Corresponding author: atxuanyizhu@163.com

Abstract. The issue of enhancing the platform conversion rate has emerged as a primary challenge hindering the advancement of e-commerce platforms. This study undertakes a comprehensive analysis of the interplay between the platform conversion rate and its influencing factors. To this end, a two-pronged approach is adopted, employing both a polynomial fitting model and a decision tree model to elucidate the underlying mechanisms. Initially, a thorough data cleansing and descriptive analysis is conducted on the voluminous data set provided by Alibaba Tianchi Lab. Subsequently, a polynomial fitting model is formulated. The random forest feature importance analysis revealed that the total number of user visits is the primary factor influencing the platform conversion rate. This variable was then employed to construct a polynomial fitting model based on the least squares method, which was found to have a strong positive correlation with the platform conversion rate. Finally, the decision tree model was established. The final decision tree model is obtained by using CCP pruning after establishing the decision tree model with depth restriction. The study shows that the conversion rate of the e-commerce platform is mainly related to user browsing, user age, and user attributes. The results of this research can provide a scientific basis for e-commerce enterprises to accurately formulate marketing strategies and optimize the operation process, which can strongly promote the efficient development of the e-commerce industry.

Keywords: Big Data, E-commerce, Conversion Rate, Polynomial Fitting, Decision Tree.

1. Introduction

In recent years, with the rapid development of Internet technology, e-commerce platforms have become an indispensable part of modern business. However, the question of how to improve the platform conversion rate has become a major factor plaguing the development of e-commerce platforms. A considerable body of research has been conducted in this field by numerous scholars. Wang et al. (2022) developed a traffic game model from the perspective of the platform economy and investigated the optimal pricing problem of traffic data. They identified the conversion rate of sales volume and the conversion rate of private domain traffic as the pivotal factors in the pricing problem, yet it was limited by its focus on the platform economy perspective, which overlooked the impact of other factors such as customer behavior on conversion rates [1]. Li et al. (2022) demonstrated live streaming marketing impacts through apparel sector comparisons, but their industry-specific focus and reliance on projected sales data limit cross-sector generalizability [2]. While Zhang et al. (2021) emphasized the impact of product images and pricing through cross - border platform analysis, they did not construct a specific mathematical model, focusing more on theoretical exploration. The case study method they adopted can deeply analyze the changes in the conversion rate of specific products. However, due to the limited number of cases and lack of large - sample verification, the universality of the research conclusions is not satisfactory [3]. These studies provide a theoretical foundation for examining the conversion rate of e-commerce platforms from diverse perspectives. However, there are still many crucial aspects that require in-depth study on the key issue of e-commerce conversion rate influencing factors. In the context of the big data era, data mining and analysis techniques offer a novel opportunity to address these problems. In contrast to previous studies constrained by single-industry perspectives or limited sample sizes, this study developed a dual-model analytical framework integrating polynomial fitting and CART decision tree algorithms to dissect multidimensional

influencing factors of the conversion rate of e-commerce. Leveraging large-scale real-world transaction data from Alibaba Tianchi Lab, this study achieves an integration of parametric and non-parametric modeling to capture both linear relationships and complex decision boundaries in conversion rate mechanisms and a development of an interpretable prediction framework balancing algorithmic accuracy with operational implementability. These methodological breakthroughs not only address the generalization limitations of prior research but also establish a new paradigm for conversion rate optimization through data-driven feature engineering and model fusion techniques. This study hopes to fill the gaps in the existing research, provide strong theoretical support and practical guidance for the e-commerce industry to optimize the operation strategy and enhance market competitiveness, and promote the sustainable development of the e-commerce industry.

2. Data acquisition and pre-processing

2.1. Data source

The research data presented in this paper is courtesy of Alibaba Tianchi Lab. This data set encompasses user information and webpage access data pertaining to a specific product over a designated period on a particular platform.

2.2. Data pre-processing

In this paper, the presence of missing values and outliers is identified through a thorough examination of the dataset's information. These values are then processed in a systematic manner. For the category feature variables, this paper employs the use of nominal variables to convert them into uniformly digitized variables.

2.3. Descriptive analysis based on visualization

This study utilizes curve graphs, bar charts, and pie charts for the purpose of conducting a statistical analysis, thereby reflecting the variable probability density and distribution. Consequently, this study undertakes the following visual analysis of several core variables.

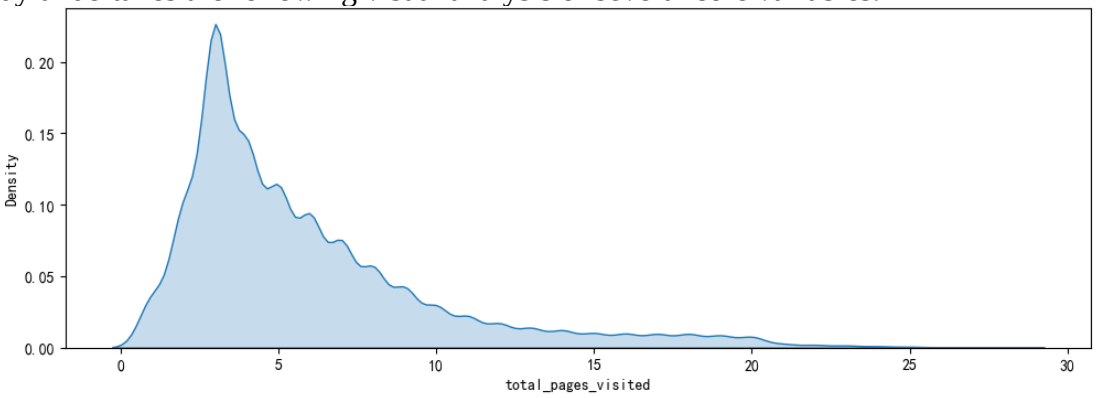

Figure 1. Density of user page visits distribution

As illustrated in Figure 1, the distribution of the total number of pages visited exhibits a tendency to approximate a normal distribution, with the majority of outliers occurring 13 times and beyond. The median of the data is approximately five times. This suggests that consumers who browse the product tend to reach a decision regarding its purchase within a specific number of visits.

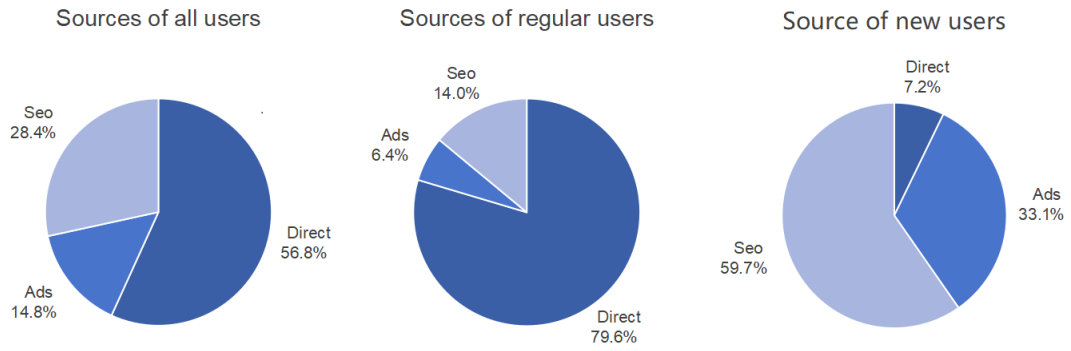

Figure 2. Source of users

As illustrated in Figure 2, when considering the aggregate of all users, it is evident that more than half of them access the merchandise interface directly. Older users exhibit a higher propensity to enter the product interface directly (Direct). In contrast, new users are more likely to access the product interface through search engine optimization (Seo) and advertising methods (Ads).

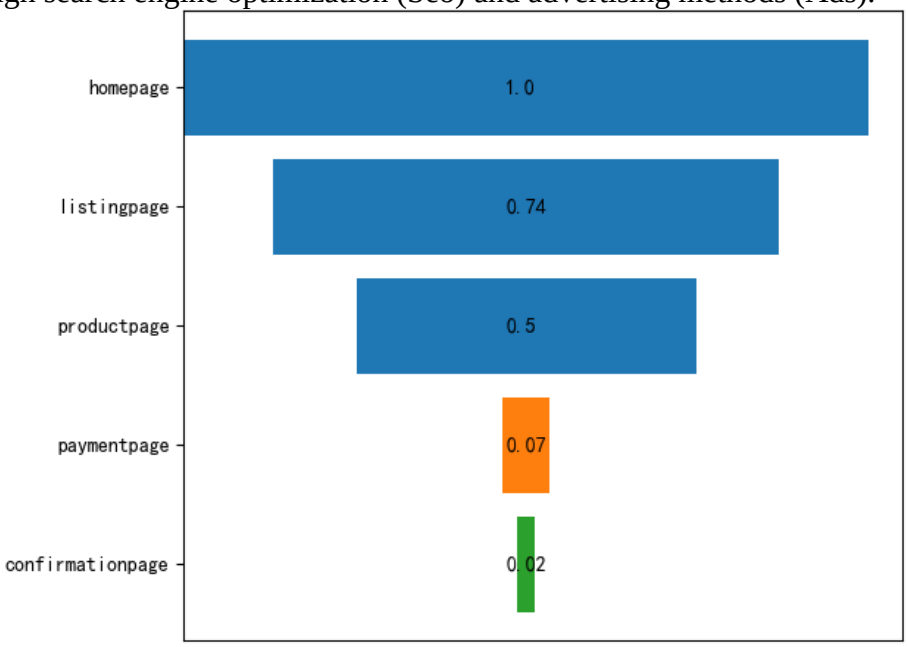

Figure 3. Conversion rate from the home page to each stage page

As illustrated in Figure 3, the conversion rate exhibited a decline in the initial three stages. However, it was relatively sustained at a certain level until the fourth stage (payment page), where a substantial decrease in the conversion rate was observed, from 0.5 to 0.07. In the final confirmation process, the conversion rate further decreased to 0.02. This decline underscores the challenges inherent in the pivotal process spanning from the product page to the payment and confirmation page.

3. Results and analysis

3.1. The single factor study based on polynomial fitting

This study applies a polynomial fitting algorithm based on the least squares method to investigate the effect of a single variable on the conversion rate of the longest process from home page to confirmation page.

(1) The principle of polynomial fitting

A polynomial fitting algorithm is a learning algorithm that uses known data to predict or estimate unknown data. This study chooses a polynomial fitting algorithm based on the least squares method to approximate the distribution of known data by constructing a polynomial function. The parameters are then adjusted to minimize the residual sum of squares between the points on the curve of this

function and the actual values, thereby achieving optimal fitting. The function curve is subsequently employed to analyze the user and purchase data of the e-commerce shopping platform [4].

The fundamental principle of the polynomial fitting algorithm based on the least squares method can be outlined as follows:

Initially, a viable form of $P(x)$ is ascertained based on the approximate trend of the original data to be fitted, as demonstrated in the polynomial form Equation (1):

$$P_n(x) = a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n = \sum_{i=0}^{n} a_i x^i \tag{1}$$

Where: $P_n(x)$ is defined as an n th-degree polynomial in x.

Subsequently, the least squares method is employed to solve the Equation (1). The m data points (x_i, y_i) are utilized to ascertain the coefficient a_i in Equation (1). Let δ_i denote the deviation of each point in the equation from the actual data points, then the sum of the squares of the deviations is shown in Equation (2):

$$\sum_{i=1}^{m} \delta^2 = \sum_{i=1}^{m} \left[\sum_{j=0}^{n} a_j x_i^j - y_i \right]^2 \tag{2}$$

The method of least squares is defined as the method that minimizes the value of the sum of squares of the deviations of the analyzed data set and in this way determines the coefficients of the polynomial equations set [4]. The m data points (x_i, y_i) mentioned above are brought into Equation (2) to obtain the system of equations as shown in Equation (3):

$$\begin{cases} a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n = y_1 \\ \cdots \\ a_0 + a_1 x_m + a_2 x_m^2 + \cdots + a_n x_m^n = y_m \end{cases} \tag{3}$$

This system of equations consists of m equations, among which the number of unknowns a_j is $n+1$. The specific form is shown in Equation (4):

$$a = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix}, c = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix} \tag{4}$$

Denote the Equation (4) by $Ca = \gamma$, and construct a normal linear system of equations $C^T Ca = C^T \gamma$ containing $n+1$ unknowns in formula (3). When $|C^T C| \neq 0$, that is, its coefficient determinant is not equal to zero, then formula (4) has a unique set of optimal approximate solutions that minimize the sum of squared deviations of formula (2). Thus, the polynomial obtained by the least squares fitting can be found. Since the $n+1$ column vectors of matrix C are linearly independent, that is, $x_1, x_2, ..., x_m$ are distinct, the rank of matrix C, $R(C) = n+1$. That is, $C^T C$ is non-singular, so the solution of formula (3) exists and is unique [4].

(2) Variable selection

Through simple mathematical analysis of some of the acquired data, it was found that four factors, namely user gender (sex), user age (age), user attribute (new_user), and total number of pages visited by the user (total_pages_visited), have relatively significant impacts on the user's purchase outcome.

Therefore, the feature importance of the random forest was utilized to evaluate these four variables, exploring the extent of their influence on conversion.

Table.1. Random Forest Feature Importance

Feature	Importance
total_pages_visited	0.665
age	0.246
new_user	0.067
sex	0.021

As demonstrated in Table 1, the four factors of user gender (sex), user age (age), user attribute (new_user), and total number of pages visited by the user (total_pages_visited) all have varying degrees of influence on the user's purchase outcome. Among them, the feature importance of the total number of pages visited by the user (total_pages_visited) on conversion reaches 0.665. Consequently, it is identified as the focal point of subsequent research endeavors.

(3) Model construction

The mean conversion rate corresponding to varying numbers of user clicks is considered the sample point. In order to reduce over-fitting, the order of the polynomial is set to 4 in this study, without sacrificing the accuracy of this model.

The least-squares method is used to perform polynomial fitting, and a single-variable polynomial model is established, as shown in Equation (5):

$$Y=-0.00001541x^4+0.0008553x^3-0.01326x^2+0.07324x-0.1011 \quad (5)$$

Where: Y is the conversion rate; x is the total number of visits.

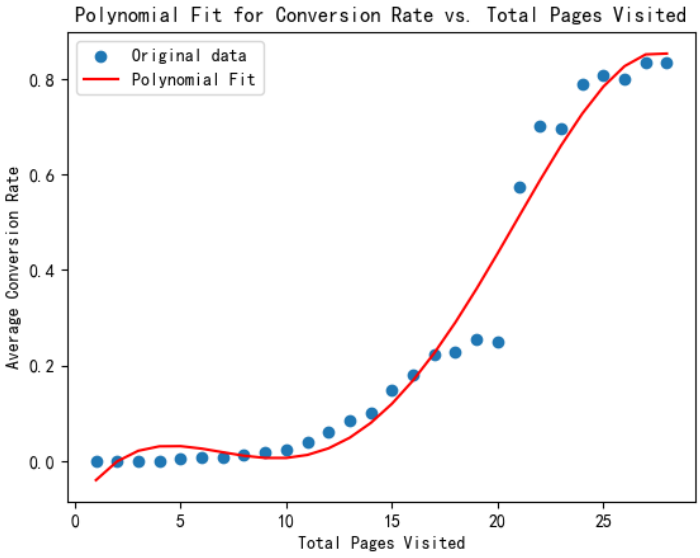

Figure 4. Polynomial fitting curve

As illustrated in Figure 4, this study offers preliminary conclusions. When the total number of user visits is less than 10, the conversion rate of consumers predicted by the model approaches zero. Overall, the model demonstrates a strong positive correlation between conversion rate and total number of user visits.

(4) Model evaluation

As demonstrated in the above results of the polynomial fitting, the coefficient of determination (R^2) is 0.97, and the mean squared error (MSE) is 0.003. This suggests that the model exhibits a high degree of fit.

Subsequent evaluation of the model construction revealed that the polynomial fitting model has certain inherent defects. First, it can only consider the impact of a single factor on the conversion rate. Conversely, although the total number of pages visited by users (total_pages_visited), which is of

paramount importance to the conversion rate, was selected, the model still exhibits over-fitting when the total number of pages visited by users is relatively low (less than or equal to 2), resulting in a negative conversion rate. It is important to note that the current model does not allow for simultaneous improvement in model accuracy and reduction in over-fitting through alterations to the order of the polynomial or other methods. In light of these challenges, the subsequent sections of this study will utilize alternative models to further investigate the factors influencing the conversion rate.

3.2. Research on conversion rate factors based on CART decision tree algorithm.

Given that the majority of contemporary research on conversion rate analysis is rooted in the domain of applied economics, there persists a dearth of a clearly delineated scope for the selected indicators. Concurrently, the issue of an overabundance of indicators and a plethora of classifications has come to the fore. In light of these observations, this study employs a methodology that utilizes the CART algorithm, grounded in the principles of polynomial fitting, to meticulously analyze the multifaceted factors influencing conversion rates. The decision tree algorithm has been shown to exhibit superior performance in terms of rapid execution, minimal computation, and high classification accuracy when compared to other classification algorithms [5]. This study proposes the establishment of a CART decision tree model, utilizing the Gini coefficient to analyze the factors influencing conversion rate with the help of big data.

(1) The principle of CART decision tree algorithm

The decision tree algorithm is a machine learning algorithm for classification and regression problems that predicts the input data by constructing a tree-like decision model [6]. Currently, the commonly used decision tree algorithms are ID3, C4.5, and Classification and Regression Tree (CART) algorithm [7].

The feature gain index of the ID3 algorithm is known to tend to discrete attributes with a large number of attributes, and can not deal with missing values and continuous attribute values; C4.5 can deal with continuous attributes and missing values, but is computationally complex and has low efficiency in generating multi-branch trees [8]. In light of these observations, this study proposes the adoption of an enhanced CART algorithm for e-commerce conversion rate analysis, a method frequently employed in addressing classification and regression challenges.

In the CART algorithm, the split point of each attribute is first calculated, and the attribute that increases the purity of the data set the most is selected as the best split attribute. For continuous attributes, the optimal split point should be determined to achieve the highest purity, and the decision tree nodes are constructed accordingly. Gini impurity or Entropy are commonly used to measure purity in classification problems [9]. The Gini impurity is shown in Equation (6):

$$Gini(S) = \sum_{i=1}^{n} p_i(1-p_i) = 1-\sum_{i=1}^{n} p_i^2 \tag{6}$$

Where: p_i is the sample set; S is the probability of category C_i; n is the total number of categories.

If a certain eigenvalue a of attribute a in dataset S divides S into two sample subsets S_1 and S_2, the Gini coefficients at this point are transformed into Equation (7):

$$Gini(S,A) = \frac{|S_1|}{|S|}Gini(S_1) + \frac{|S_2|}{|S|}Gini(S_2) \tag{7}$$

The CART algorithm repeats the above steps for each subset after splitting until it meets the stopping conditions, such as reaching the preset maximum tree depth, the number of samples in the node is less than a certain threshold, or it cannot be split further, and in turn recursively forms a binary tree. Eventually, pruning is performed to optimize the model to prevent overfitting and improve the model accuracy.

(2) Model construction

Under the premise of limiting the accuracy without losing precision, after considering the relationship between the depth of the decision tree and the classification accuracy, this study limits the depth of the decision tree to 4 and constructs a decision tree model based on the CART algorithm.

Based on this model, the characteristic importance of each factor is derived as shown in Table 2.

Table.2. CART decision tree model (before pruning) characteristic importance of factors

Variables	Column
new_user	0.12
sex	0.00
age	0.07
total_pages_visited	0.80

The feature importance of gender (sex) in this model is 0.00, which means that the factor of gender (sex) in this model has a negligible effect on the final results. Therefore, in the prediction of this model with the addition of big data, this study considers that gender (sex) will not affect the change of conversion rate of e-commerce.

In this model, when the total number of visits is lower than 14.5 times (i.e., lower than or equal to 14 times), the probability that the consumer of this product realizes conversion under the prediction of this model is zero, and when the number of visits of the product interface is greater than 14.5 times (i.e., greater than or equal to 15 times), the conversion rate of the product is ushered in to increase. At this point, the split attribute points are constantly undergoing changes. When the number of visits on the product page is in the range of 14.5 to 20.5, the split attribute point is age. In this condition, when age is greater than 19.5 years old, the model predicts that the probability of the consumer of the product achieving conversion is zero. When the age is less than 19.5 (i.e., less than or equal to 19.5), whether the user is a new user becomes the split attribute point. In summary, the model predicts that if the consumer is a new user under the age of 19.5 and has visited the item more than 14.5 times, the consumer will convert the browsing behavior into a consumption behavior.

When a consumer visits the product more than a certain number of times (the model predicts more than 20.5 times), the consumer will purchase the item. Under this condition the split analysis is performed again then the split attribute is whether the consumer is a new consumer. When the consumer is a new consumer and is less than 39.5 years old, the model determines that the consumer will purchase the item. When the consumer is not a new consumer and is less than 26.5 years old, the model determines that the consumer will purchase the good. Under other conditions, this model determines that the consumer will not purchase the item.

The performance of the model was evaluated using the confusion matrix. The confusion matrix data is shown in Table 3.

Table.3. Confusion matrix for CART decision tree model (before pruning)

	Forecast non-conversions	Forecast conversions
Actual non-conversions	19234	41
Actual conversions	361	116

As illustrated by the confusion matrix presented in Table 3, the model demonstrates an accuracy of 97.96% in the test data of the test set.

(3) Model pruning

In the subsequent optimization of this model, in order to prevent overfitting during the modeling step, a post-CCP pruning approach is used, where the tree is pruned and simplified after the construction of the decision tree is completed, resulting in the minimization of the loss function as in Equation (8) [9]:

$$L=\sum_{i=1}^{T}\frac{Ni}{N}Li+\alpha T \tag{8}$$

Where: T is the number of leaf nodes; N is the number of all samples; N_i is the number of samples on the i th leaf node; L_i is the loss function of the i th leaf node; α is a to-be-determined coefficient used to penalize the number of nodes and guide the model to use fewer nodes.

In this model, L_i is the Gini coefficient of the i th leaf node.

In accordance with this principle, the original CART decision tree model CCP paths are calculated to explore the relationship between α and tree impurity (the above loss function only considers the case of Gini weighted sums), as illustrated in Table 4.

Table.4. Relationship between α and tree impurity

α	Tree impurity
0.00E+00	0.03251
7.81E-06	0.032518
1.15E-05	0.032529
1.18E-05	0.032541
4.06E-05	0.032582
7.11E-05	0.032653
7.95E-05	0.032732
9.15E-05	0.032824
1.04E-04	0.032928
4.61E-04	0.033389
4.74E-04	0.033864
5.35E-04	0.034398
7.37E-04	0.035135
7.65E-04	0.035901
3.20E-03	0.039099
7.96E-03	0.047063

In this model, the acceptable tree impurity is determined to be 0.032928, which corresponds to an α value ranging from 1.04E-04 to 4.61E-04. Subsequent to establishing α at 4.0E-04, the model underwent retraining to generate a pruned decision tree.

The result of the original model after pruning is shown in Figure 5.

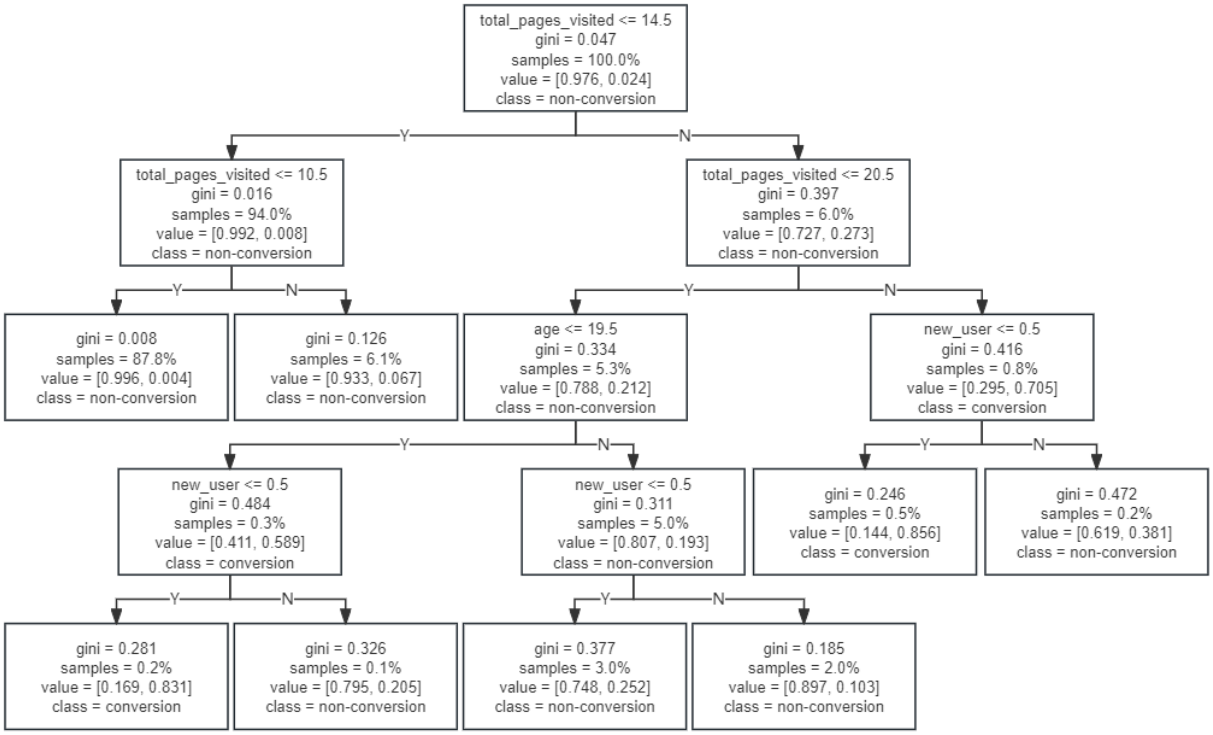

Figure 5. CART decision tree model (after pruning)

As illustrated in Figure 5, the pruned model has been simplified to enhance its accuracy. This has been achieved by simplifying the case of more than 20.5 visits and eliminating the redundant analysis for conditions involving less than 14.5 visits.

(4) Model evaluation

The model was subjected to a pruning operation where some of the redundant branches were cut off. The confusion matrix for evaluating the model after pruning is shown in Table 5.

Table.5. Confusion matrix for CART decision tree model (after pruning)

	Forecast non-conversions	Forecast conversions
Actual non-conversions	19248	27
Actual conversions	365	112

As illustrated by the confusion matrix in Table 5, the model demonstrates an accuracy of 98.02% in the test data of the test set. This indicates an enhancement in the model's precision.

In comparison to the model with unpruned leaves, the pruned model has demonstrated efficacy in preventing overfitting, enhancing the number of successful samples for prediction, and simplifying the generated decision tree, thereby ensuring more concise model results [10].

However, due to the inherent limitations of the model and within the context of applied economics, the criteria for the attributes remain ambiguous. In the construction of the decision tree model, there is a situation in which the model is quite ambiguous about the results, i.e., the Gini coefficient appears to be close to 0.5 in some branches. While the model has identified the factors that significantly impact the effectiveness of the largest share of the numerous factors influencing conversion rate, the presence of a high Gini coefficient indicates that the model lacks adequate differentiation in certain end nodes. This is a primary impediment to enhancing the model's accuracy.

4. Conclusions

This study offers technical support and guidance for enhancing the conversion rate of the e-commerce platform. To this end, a polynomial fitting model and a decision tree model have been constructed, yielding the following conclusions:

Utilizing the polynomial fitting model, this study ascertains a robust positive correlation between the number of user visits and the conversion rate. This finding can assist e-commerce companies in determining the optimal range of the number of user visits, thereby averting the squandering of resources engendered by over-promotion, and furnishing a foundation for the selection of attraction mode. Nevertheless, the model is constrained in its capacity to consider the influence of a specific factor on the conversion rate in isolation. Furthermore, in the context of overfitting while maintaining model accuracy, it is not feasible to continue reducing the degree of overfitting by modifying the polynomial order and other related factors.

This study utilizes the CART decision tree model to elucidate the comprehensive impact of each factor on the conversion rate, particularly the interaction of user visits, user age, and user attributes. This model offers a robust framework for e-commerce accurate publicity, enhancing the efficacy of publicity and attracting traffic. However, the model exhibits suboptimal differentiation at certain end nodes when confronted with large-scale data, hindering further enhancement of accuracy.

This study provides valuable insights and practical tools for e-commerce platforms to optimize their conversion rates through data-driven decision-making, offering both predictive and interpretative models to enhance marketing strategies and resource allocation. Future research should further explore adaptive modeling techniques that integrate real-time user behavior data and dynamic market factors, potentially combining ensemble learning with deep neural networks to address current limitations in feature interaction analysis and overfitting control, while incorporating multi-source heterogeneous data such as social media traces and geolocation patterns to achieve more granular conversion rate predictions and personalized recommendation strategies.

References

[1] Wang Yong, Liu Leyi, Chi Xi, et al. Traffic Game and Optimal Pricing of Traffic Data: From the Perspective of E-commerce Platforms[J]. Management World, 2022, 38(08): 116-132.

[2] Li Yumin, Miao Ruijing, Mao Yueyu, et al. Clothing e-commerce design-making analysis under different webcast marketing modes from the perspective of game theory[J]. Journal of Silk, 2022, 59(02): 68-76.

[3] Zhang Xiaoli, Han Xiaoxiao, Xu Yue. Research on the Key Influencing Factors of Product Conversion Rate in Cross-border E-commerce[J]. China Economic & Trade Herald (Middle Edition), 2021,(04):132-134.

[4] Hu Jie, Su Jianhui, Du Yan, et al. Parameter Identification of Pemfc Stack Model Based on Least Square Method[J]. Acta Energiae Solaris Sinica, 2021, 42(06): 1-4.

[5] Zhang Min, Peng Hongwei, Yan Xiaoling, et al. Improved Algorithm of Fuzzy Decision Tree Based on Neural Network[J]. Computer Engineering and Applications, 2021, 57(21): 174-179.

[6] Wang Jingxiang. Research on the Principle and Practical Application of Decision Tree Algorithm[J]. Computer Programming Skills & Maintenance, 2022,(08):54-56, 72.

[7] Zhang Maojun, Rao Huacheng, Nan Jiangxia, et al. Timing Selection of Quantitative Trading Based on Decision Tree[J]. Systems Engineering, 2022, 40(02): 118-130.

[8] Liu Yafen. Improved CART Decision Tree Algorithm Based on GA and Its Application[D]. Guangzhou University, 2020.

[9] Yao Yuesong, Zhang Xianyong, Chen Shuai, et al. Decision-tree induction algorithm based on attribute purity degree[J]. Computer Engineering and Design, 2021, 42(01): 142-149.

[10] Zhang Chi, Wang Dong, Zhao Shuhan, et al. Automatic Fault Diagnosis System of Energy Metering Device Based on Improved Decision Tree Algorithm[J]. Automation & Instrumentation, 2024, 39(02): 1-4, 10.