Research on Customers' Purchasing Intention of Sports Shoes Based on Data Mining Application

Wanliu Che, Gang He, Haixia Gui, Changxia Shen and Fan Yang
School of economics and management, Anhui University of Science & Technology, Huainan, Anhui, 232001

Abstract. The rapid development of e-commerce makes online shopping the main means of social consumption. How to extract useful information from massive semi-structured and unstructured customer data has become a hot issue in e-commerce industry research. This article uses the Octopus data crawler software to capture the online purchase information of Jingdong Mall sports shoes customers, and classify and analyze the information, and analyze the price, color, size, brand, etc. of the most popular sports shoes purchased by consumers; On the other hand, GeNie is used to generate online reviews of the Bayesian network structure model for the platform, and the cross-validation method is used to test the accuracy of the model. Finally, the key factors affecting the satisfaction of the sneaker customers are analyzed, and the customer satisfaction is studied. There is a significant causal relationship between the target variable and other related variables, and there are different degrees of correlation between different variables. The most important factor for consumers to purchase sports shoes is the comfort, quality and brand of sports shoes.

1. Introduction
With the continuous development of e-commerce, online shopping has become the main means of social consumption. The speed and convenience of online shopping has led to an increase in the sales of footwear represented by sports shoes [1]. How to occupy a competitive advantage in many product types and brands has become the primary problem faced by major footwear suppliers and sellers, and the transparency of the network environment makes users purchase information becomes a key factor affecting product sales [2]. Therefore, researching the purchasing preferences of sports shoes consumers in the new era has important theoretical and practical significance for enterprises to grasp consumer preferences deeper and develop development strategies and marketing strategies.

2. Data collection and analysis

2.1. Data source acquisition and organization
With the core aim of “product, price and service” for many years, JD.com is one of the most popular and influential e-commerce websites in China's e-commerce field. The footwear brands are diverse and rich in categories, high quality, fast logistics, and praised by customers. Therefore, this article uses the Octopus crawler software to capture the corresponding customer purchase page, and transforms the semi-structured and even unstructured presentation modes such as browsing information and comment information generated between the human and the machine into effective visual information. The
static information purchased, such as price, product sales, store name, size of sports shoes, suitable for road surface, function, whether the product is discounted or closed, is classified and statistically processed, and dynamic information such as purchase evaluation is based on text analysis to extract comment keywords and synonyms. Replacing and other data processing methods to process the acquired sample data, so that the merchant has a new perspective on the customer, providing a reference for the product to achieve customer-centric sales.

(1) Data source acquisition

The data used in this paper is derived from the data retrieved by Jingdong Mall, using sports shoes as the key word. The collection time is July 10, 2019. After the web crawler software "Octopus" collects 5150 pieces of data on the corresponding product page. After data screening and cleaning, 5119 valid data were obtained. After preliminary sorting, the data warehouse was integrated. Then the pre-processed data was sorted by excel software, and the statistical information of static purchase information was constructed. The first five digits are sorted, as shown in Table 1.

| Price range | Colour | Size | Sneakers | Price range | Colour | Size | Sneakers |
|-------------|--------|------|----------|-------------|--------|------|----------|
| [100, 200)  | black  | 1664 | (32.49%) | [1000g, 1100g) | black  | 1648 | (32.19%) |
| [0, 100)    | white  | 1146 | (22.38%) | [700g, 800g)  | white  | 840  | (16.41%) |
| [300, 400)  | gray   | 501  | (9.78%)  | [500g, 600g)  | gray   | 765  | (14.94%) |
| [200, 300)  | red    | 228  | (4.45%)  | [800g, 900g)  | red    | 700  | (13.67%) |
| [400, 500)  | blue   | 196  | (3.83%)  | [600g, 700g)  | blue   | 317  | (6.19%)  |

Table 1 shows the price range, color, size and weight of the goods. The left column of the product feature represents the feature attribute value, and the right column represents the number of people purchased in the statistics, and the proportion in parentheses. It can be clearly seen that the number of purchases is the highest when the price range is [100, 200), reaching 2,592, accounting for 50.63% of the overall sample. This shows that more than half of the Chinese sports shoes consumers think that the reasonable price of sports shoes is between 100 and 200 yuan. In the first four digits of the price range, [0, 400) yuan, the total number of purchases was 4222, accounting for 82.48% of the total sample, indicating that the majority of sports shoe buyers can receive between 0 and 400 yuan. For the color attribute, the author classifies and counts according to the standard color of the Jingdong shopping platform. It is obvious from the color attributes that black, white, gray, red, and blue are the most frequently purchased colors by the purchaser. The purchase amount of the five color attributes reaches 3,735 pairs, accounting for 72.96% of the total, of which black is the most common purchaser. The number of purchases was 1,664, which is close to 1/3 of the total sample. For the size of sports shoes, the author's statistical interval is [35, 46], with an increase of 0.5 yards. Among them, 27.75% of the sample buyers choose 42 yards. The gross weight of sports shoes is also an important factor in sports shoes. Through statistical data, sports shoes with weights ranging from 1000g to 1100g are the most purchased in the sample, and the weight of products that most consumers can accept. The sales volume reaches 1648 pairs, accounting for the total 32.19%. At the same time, the weight of sports shoes in the three ranges of 700–800g, 500–600g and 800–900g also account for a large proportion. Most people don't care about the weight of shoes, and Yan Pei's research results show that the weight of sports shoes has an important impact on comfort, and the subjective comfort of the human body will increase with the weight reduction, so the weight of the sports shoes should also be a factor considered by the seller.
At present, there are many sports shoe researches at home and abroad that focus on the material of the upper, the material of the sole, the suitable road surface and the way of closing. This article will show the relevant sports shoe feature results displayed in Table 2.

### Table 2. Sports shoes purchase information statistics classification.

| Upper material | Sole material | Suitable for pavement | Closed way |
|----------------|---------------|-----------------------|------------|
| 1 net          | rubber        | Runway, road, trail   | Lace       |
| (35.04%)       | (58.84%)      | (25.34%)              | (94.66%)   |
| 2 Fabric       | PU            | Floor, runway, highway| Set of     |
| (33.87%)       | (8.55%)       | (24.58%)              | (2.48%)    |
| 3 PU           | EVA           | track                 | Elastic    |
| (5.18%)        | (6.52%)       | (15.43%)              | (1.15%)    |
| 4 cloth        | PHYLO         | Floor, runway, road, trail| Not tied |
| (3.98%)        | (1.88%)       | (6.16%)               | (0.69%)    |
| 5 Mixed Material| Mixed Material| Runway, highway      | Buckle     |
| (2.98%)        | (1.84%)       | (5.32%)               | (0.46%)    |

It can be seen that the upper composed of mesh cloth and fabric and the sole composed of rubber are the most sold, among which the mesh shoes are most loved by the people, and the mesh shoes can constitute the microclimate in the shoes, and the comfort of the sports shoes is improved inseparable [3]. The upper made of fabric can better meet the requirements of fashion and performance design [4]. The top three sales rankings for sports shoes are “runway, road, trail”, “floor, runway, highway” and “runway”. The total sales of the three shoes reached 3,345 pairs, accounting for 63.35% of the total sample. In the overall sample, 94.66% of the sold sports shoes are tied. In the mall, we can see that the sports shoes are mostly closed with laces, but this is often ignored by buyers and sellers. Studies have shown that the closure of the strap not only enhances the fit of the shoe, enhances its function and performance, but also helps to reduce the pain of the foot [5].

At present, there are various sports shoes brands distributed in Jingdong Mall. The sales and promotion of each brand are generally completed by brand flagship stores, specialty stores and Jingdong self-operated stores. In this paper, while crawling the product data, we also grabbed the sales of the brand of the sports shoes, the brand operation store and the sales of the sports shoes. A total of 138 domestic and foreign sports shoes brands were collected, the current sales of effective sports shoes in the store is 66,831,358 pairs. Secondly, the classification of brand shoes with more than 5% of the total number of stores will be counted, the number of sports shoes sales accounted for the top six in the total number of sales statistics (unit: million pairs).

It can be seen that the most popular sports shoes for sneaker buyers are Xtep and Anta, and the sales volume is more than 8 million pairs. At present, domestic sports shoes brands occupy more than half of the Chinese sports shoes market regardless of the number of stores or product sales. This shows that the Chinese sports shoe industry is gradually moving closer to international sports shoes. The Adidas sales store is the largest of all the sample stores, more than twice as many as Nike stores, but its total sales volume is nearly half that of Nike. According to the financial report data in recent years, Adidas has some problems, but it is still growing steadily in China [6]. In addition, in the Jingdong Mall, whether it is Nike, Adidas, Jordan and other foreign brands or Xtep, Anta, Li Ning and other domestic brands are mostly based on Jingdong self-operated sales, the sports shoes sales shop can choose the domestic sports brands such as Xtep and Anta in the selection of marketing shoes. It is best to choose to enter the Jingdong self-operated market in Jingdong Mall.

In addition, according to the above classified statistical data and the research conclusions of the literature, the best positioning price of sports shoes in China is not 50-100 yuan. The MBA think tank 2017 China consumption upgrade research report points out that the rise of China's high-income groups in recent years has emerged. Consumer demand has exploded, the consumption structure has

---

3
changed, and the purchasing power of mass consumers is also increasing [7]. China is experiencing a new round of consumption upgrades, grasping the big waves of consumption upgrades, and always insisting on customer-centered. From a perspective, China will certainly manufacture more and more international sports shoes brands.

3. Bayesian network comment model

The Bayesian network, also known as the Bayesian belief network, was first proposed by Laurie Cinn and Spitzer Hotter, and consists mainly of a directed acyclic graph (dag) and a conditional probability table [8]. The framework proposed by the network is suitable for customer satisfaction modeling and sentiment analysis. The accuracy of seeking research focus is 88.3% [9]. Therefore, this paper uses the purchase comments of e-commerce customers to construct Bayesian network for e-commerce customer movement. The shoe purchase evaluation is analyzed and researched to find out the key factors affecting the customer's satisfaction with the type of sports shoes, and provide objective theoretical reference for the e-commerce platform and suppliers.

3.1. Comment information element extraction

Firstly, the spam comments and invalid comments are removed by pre-processing, the comment keywords are extracted and sorted, and the frequency of each keyword is counted, the high-frequency keywords are extracted, and they are classified into feature factor variables, and the unstructured data is processed and converted into structured data. Secondly, the words with similar meanings in the high-frequency keywords and representing the same product features are classified and summarized, which is used as the evaluation index variable of the Bayesian network modeling. From the perspective of customer reviews, almost all buyers are satisfied with their products. We also classify satisfaction (δ) into three levels according to the content of the comments, namely “satisfaction”, “general” and “unsatisfactory”. From the analysis of the characteristics of sports shoes, the highest level of attention paid by the purchaser to the characteristics of sports shoes is “comfort”, followed by “quality”. In addition, consumers pay more attention to the characteristics of “sports”, “appearance”, “price”, “size” and “function” of sports shoes. In contrast, “professional”, “color”, features such as "physical attributes" are relatively low in focus.

3.2. Bayesian network principle

The Bayesian network modeling core is the Bayesian formula, and is defined by the joint probability distribution of all nodes in the network [10]. For the event δ that affects the characteristics of the sneaker customer satisfaction, set the event set that affects its occurrence as \( A_i \ (i = 1 \sim 10) \), then the relevant Bayesian formula is:

\[
P(A_i|\delta) = \frac{P(\delta|A_i)P(A_i)}{\sum_{j=1}^{10} P(\delta|A_j)P(A_j)}
\]

among them, \( P(A_i|\delta) \) for the posterior probability, \( P(A_i) \) for prior probability, \( P(\delta|A_i) \) for the event \( A_i \) under the condition U.

For the Bayesian network model that has been constructed, the joint probability density of each feature variable node of the sneaker in the network can be expressed as:

\[
P(A) = \prod_{i=1}^{10} P(A_i|A_{\text{pa}(A_i)})
\]

among them, \( A_{\text{pa}(A_i)} \) representing feature variable nodes \( A_i \) all parent nodes.
3.3. Construction of Bayesian network model

After converting the pre-processed semi-structured data into structured data, it is imported into the Bayesian network software GeNie, and after reading, filtering, it enters the modeling process of the Bayesian network to construct the network structure and calculate the network. The conditional probability of the node and the importance of each factor node determine the degree of customer satisfaction with the purchased product.

In the Bayesian network structure diagram, user satisfaction is taken as the target variable to be examined. It has obvious causal logic relationship with many factors such as appearance, physical characteristics, color, quality and professionalism of sports shoes. In addition, there is a certain correlation between the input variables. Among them, the function, price, professionalism and comfort of sports shoes are related to the brand of sports shoes; the quality and price of sports shoes are related to the professionalism of sports shoes; the appearance and physical characteristics of sports shoes also have certain influence on their quality.

3.4. Node correlation analysis

The data is imported into SPSS25.0, and the processed data is analyzed for correlation size. Since the target node satisfaction and other related nodes are continuous variables, the correlation analysis selects Pearson correlation, and the correlation of nodes Sex analysis, screening out the main related variables, sees the results in Table 3.

| node            | Related variable | Relevance |
|-----------------|------------------|-----------|
| Satisfaction    | Comfort          | 0.356**   |
|                 | quality          | 0.478**   |
|                 | Brand            | 0.258**   |
|                 | Features         | 0.248**   |
|                 | Physical         | 0.244**   |
|                 | characteristics  |           |
| Features        | colour           | 0.470*    |
|                 | professional     | 0.464**   |
|                 | Physical         | 0.453***  |
|                 | characteristics  |           |
| Exterior        | professional     | 0.380**   |
|                 | price            | 0.357**   |
| colour          |                  |           |
|                 | Physical         | 0.714**   |
|                 | characteristics  |           |
|                 | professional     | 0.482**   |
|                 | Comfort          | 0.444**   |

Note: ** At the 0.01 level (two-tailed), the correlation is significant.

It can be seen from Table 3 that the degree of correlation between different nodes is different, and there is a positive significant correlation between all nodes. Among them, the correlation between color and physical characteristics is the strongest, up to 0.714 on the 0.01 level, followed by brand and comfort, and the correlation is at 0.01 level, reaching 0.643.

3.5. Model validation

In order to verify the prediction performance of the Bayesian network model, the validity of the model is tested by the Leave-One-Out-Cross-Validation (LOO-CV) test method provided by GeNie software. First, select the "Validate" button; secondly, click "Leave one out" as the verification method to calculate the prediction accuracy of each node in the model. The results are shown in Table 4.
Table 4. LOO-CV cross-validation results.

| node          | customer satisfaction | colour | price | Comfort | Physical characteristics | quality |
|---------------|-----------------------|--------|-------|---------|--------------------------|---------|
| Accuracy node | 0.985                 | 0.931  | 0.824 | 0.983   | 0.794                    | 0.865   |
| Accuracy      | 0.902                 | 0.783  | 0.851 | 0.826   | 0.909                    |         |

It can be clearly seen that in the predictive accuracy of many variables, the target node variable "user satisfaction" has the highest prediction accuracy of 0.985; the second is the prediction accuracy of the node "comfort", reaching 0.983; the node "color" prediction accuracy for the whole 0.931, the prediction accuracy of most nodes is above 0.8, and the overall prediction accuracy of the node is 0.813, which indicates that the constructed Bayesian network model has higher prediction accuracy and can be used for causality analysis and model reasoning.

3.6. Node posterior probability analysis

The posterior probability analysis is the most widely used in Bayesian network inference. By knowing the probability of occurrence of the target node, the reason that is most likely to cause the result is found, that is, the importance degree of each characteristic variable. Based on the constructed Bayesian network model structure, the trained user comment data is imported into Bayesian software, and the probability of selecting user satisfaction as "satisfactory" is 100%, and the degree of importance of the remaining feature variables is output. The result is shown in Figure 1.

![Figure 1. Characteristic variable posterior importance.](image)

It can be clearly seen that under the condition that the satisfaction of sports shoes customers is high, the comfort, quality and brand of sports shoes are the three factors that consumers pay most attention to; the sum of the three places reaches 0.579, accounting for the whole. More than half of the importance of 1, the remaining seven indicators are divided into two categories with the importance of 0.1 as the dividing line: ① More important indicators, including the functional and physical characteristics of sports shoes; ② Less important indicators, including the appearance, professionalism, price, size and color of the sneakers. Therefore, sports shoes sales shops and manufacturers must choose the design of sports shoes with the core aim of the comfort and quality of sports shoes, and strive to create a sports shoes brand that consumers trust. Secondly, the remaining seven types of variables are taken into account, achieving a customer-centric purpose.
4. Conclusion

This article takes Jingdong Mall as the representative of the e-commerce platform, and uses the Octopus data acquisition software to capture the purchase information of some sports shoes users, and transforms the collected unstructured data into corresponding structured by relevant data preprocessing methods. The professional Bayesian analysis software GeNie is used to construct the Bayesian network structure model of sneaker customer satisfaction, so as to find out the most concerned factors when consumers buy sports shoes, and then provide reference for merchant marketing. The research conclusions are as follows:

(1) According to the static information collected by the sneaker buyers, the price of 100–200 yuan in the Chinese sports shoe sales market is the most frequently purchased price range for consumers. The black and white colors are the most common choices for consumers; 42, 41, 40 are three sizes people often to buy; 1000g ~ 1100g sports shoes gross weight is the most commonly purchased weight.

(2) According to the characteristics of sports shoes, the upper materials of “mesh” and “fabric” accounted for 68.91% of the proportion of consumers purchased; the sports shoes with rubber as the sole material accounted for 58.84% of the total proportion; The shoes are suitable for “runway, road, small road”, “floor, runway, highway” and “runway” in the road, accounting for 65.35% of all sample sports shoes sales; sports shoes with “lacing” as the closed mode account for the purchase ratio 94.66%.

(3) By counting the number of sports shoes sold and the sales of sports shoes, Adidas sales stores are the most among all sample stores, more than twice as many as Nike stores, but their sample sales are nearly half that of Nike; sports shoes sales most of them are domestic sports shoes brands Xtep and Anta, followed by Nike and Hongxing Erke; in Jingdong Mall, whether it is Nike, Adidas, Jordan and other foreign brands or Xtep, Anta, Li Ning and other domestic brands are mostly sold by JD. Mainly.

(4) By constructing the user satisfaction Bayesian network model, it can be seen that user satisfaction as the target variable to be examined has obvious causal logic relationship with many factors such as appearance, physical characteristics, color, quality and professionality of sports shoes. In addition, the function, price, professionalism and comfort of sports shoes are related to the brand of sports shoes; the quality and price of sports shoes are related to the professionalism of sports shoes; the appearance and physical characteristics of sports shoes also have a certain influence on their quality.

(5) Using the constructed Bayesian network for post-node probabilistic analysis, it is found that the comfort, quality and brand of sports shoes are the three most important factors for consumers; the sum of the three has reached 0.579. Therefore, sports shoes sales shops and manufacturers must choose the design of sports shoes with the comfort and quality as the core tenet, and strive to build a brand of sports shoes that consumers trust, and take into account the remaining seven types of variables to achieve a customer-centric purpose.

References

[1] Deng Lingbin, Shen Hui. An empirical study on the influence of commodity recommendation information characteristics of e-commerce platform on consumers' willingness to buy [J]. Journal of Nanhua University (Social Science Edition), 2019, 20 (02): 60-65.

[2] Oliver Schaer, Nikolaos Kourentzes, Robert Fildes. Demand forecasting with user-generated online information [J]. International Journal of Forecasting, 2018(5): 98-102.

[3] Chen Jianye. The relationship model between online store credit, price, praise and store sales—based on the adjustment of service quality [A]. China Statistical Education Society, 2015.

[4] Meng Yuan, Wang Hongwei, Wang Wei. The impact of internet word of mouth on product sales: based on fine-grained sentiment analysis method [J]. Management Review, 2017, 29(01): 144-154.

[5] Xiao Jinghua, Xie Kang, Wu Yao, Liao Xuehua. The transformation from supply chain to consumer-oriented supply chain--a case study of e-commerce enterprise supply chain [J]. Management World, 2015(04): 137-154+188.
[6] Zhu Liye, Yuan Denghua, Zhang Hongming. How do customers participate in brand co-creation to enhance brand loyalty?——Research on the influence and mechanism of co-creation behavior type on brand loyalty [J]. Foreign Economics and Management, 2018, 40(05): 84-98.

[7] Shi Wenhua, Wang Wei, Sheng Na, Cai Jialong. A comparative study of the impact of online initial comments and online additional comments on product sales [J]. Management Review, 2018, 30 (01): 144-153.

[8] Xie Guangming, Jin Daxiang, Hu Pei. Analysis of the impact of network word-of-mouth disclosure on consumers' purchase behavior based on product sales [J]. Nankai Management Review, 2018, 21(6): 53-66.

[9] Lu Liying, Zhang Rong, Liu Bin. The impact of b2c retailer's distribution pricing strategy on consumers' purchase behavior [J]. Journal of Jiangsu University of Science and Technology (Natural Science Edition), 2018, 32(04): 569-575.

[10] Zhou Xiongwei, Ma Xinchao, Chen Xiaohong. The effect of discount ratio on the positive feedback behavior of consumers in price comparison advertisements [J]. Journal of Management, 2018, 15(03): 410-419.