Analysis of Sales Products Based on Data Mining

Zhuofan Zhong[1], Sihan Wu[2], Runkang Yang[1], Yiming Qian[1]
1. Hangzhou Normal University, Zhejiang 310000, China.
2. Zhejiang A&F University, Zhejiang 310000, China.

Abstract: With the popularization and development of the Internet, online sales are gradually replacing offline sales and occupy a major position in the sales industry. The transformation of sales mode will also be a huge challenge for commodity companies. Analyzing and mining market information and consumer feedback has become the key to network marketing. Based on the data of mathematical modeling competition for American college students in 2020, this paper mainly uses mathematical methods such as correlation analysis, multiple Logistic regression and natural language processing to formulate sales strategies for microwave ovens, hair dryers and baby nipples launched and sold by Sunshine Company in the online market, and analyze customer feedback data.

Keywords: Correlation Analysis; Multinomial Logistic Regression; EWM.

1. Restatement of the Problem

Sunshine Company plans to launch and sell microwave ovens, hair dryers and baby pacifiers in the online marketplace. To understand these three commodity markets and develop sales strategies, it is necessary to analyze the customer feedback data. We will accomplish the following tasks according to the given data:

(1) Develop sales strategy for sunshine company.

(2) Identify potential important design features that would enhance product desirability.

To accomplish the above two tasks, our specific work is as follows:

- Analyze the relationship between star rating, helpful vote and review.
- Make in-depth analysis of reviews and ratings, and distinguish the advantages and disadvantages of products, to make suggestions for product improvement.
- Establish a product rating system based on the analysis of reviews and ratings and select high-quality brand products to recommend to Sunshine Company for sales.

2. Model I: Analysis of Star rating， Helpfulness rating and Review

2.1 Word Segmentation and Sentiment Analysis Based on VADER

In this literature, we construct a novel model for sentiment analysis based on the review text. We put 80% of the data as the training set and all the rest 20% as the testing set of evaluations. Sentences in the review body from the training set are broken down into separated words, among which are statistically calculated their frequency. The high-frequency emotion words are picked out as seed words and manually annotated by us, while the low-frequency ones are discarded. The annotator (one of our group members) will incorporate his expertise natural-language processing knowledge for the classifying all the selected emotion words into five groups i.e., strong positive, weak positive, moderate, weak negative and strong negative.[1][2]

2.2 Correlation Analysis

We apply the methods of correlation analysis on the percentage of helpful votes (Hr), star ratings, the number of words in each review (count) and emotional score (compound), which is measured by Pearson correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})\,(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \tag{1}$$

X_i represents the star rating, review length, emotional score, Y_i represents the helpfulness rating, $\overline{X}, \overline{Y}$ represents the average number of these indexes. The r is the correlation coefficient between H_r and star rating, count, compound,

The results of the correlation analysis are

Table 1: Data Statistic Correlation Coefficient

Index	Microwave	Hair dryer	Pacifier
Star rating	-0.161312	-0.149756	-0.140086
Review words	0.244273	0.287867	0.298780
Emotional score	-0.211962	-0.178654	-0.131017

2.3 Multinomial Logistic Regression Model

Multinomial logistic regression model is adopted to further analyze the impact of star rating, review length and emotional score on the helpfulness rating. In fact, the helpfulness rating is the cumulative result of each consumer's voting (yes or no). Therefore, the number of helpful votes is subject to binomial distribution, and logistic model is suitable for empirical analysis of such kind of data.[3]

Multivariate logistic regression can determine the role and intensity of the explanatory variable X_n in predicting the probability of occurrence of strain Y. Suppose X is the response variable and P is the response probability of the model, and the corresponding regression model is as follows:

$$ln\left(\frac{p_1}{1-p_1}\right) = \alpha + \sum_{k=1}^{k} \beta_k x_{ki}$$

Then the probability of an event happening of an event is a non-linear function composed of the explanatory variable X_i . Here is the expression:

$$p = \frac{exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}{1 - exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}$$

Meanwhile, considering the impact of whether the customers are vine members and whether the purchase is verified on the authenticity and validity of reviews, we take these two indicators as independent variables.

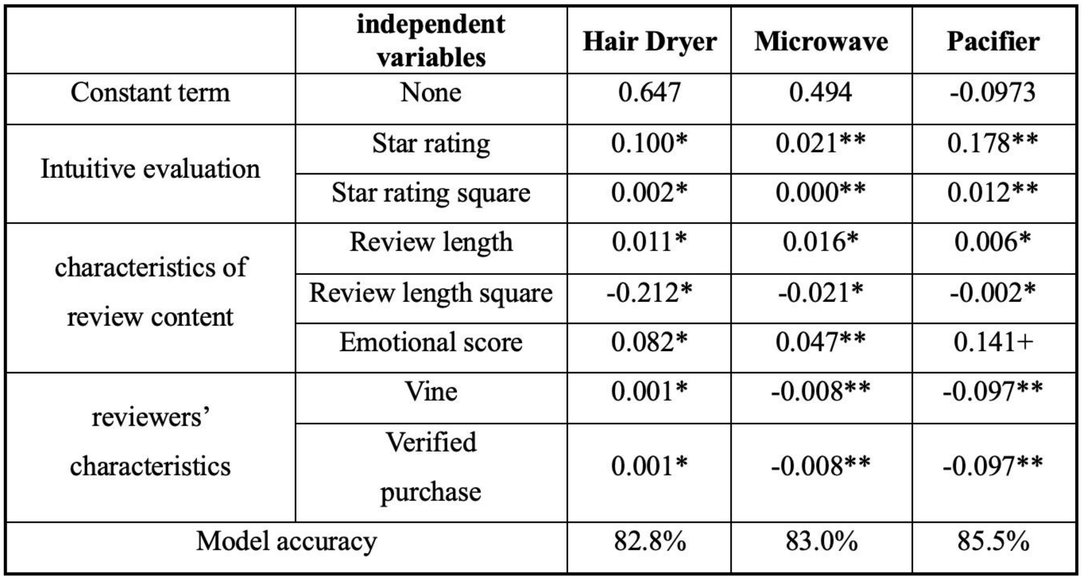
Independent variables are divided into three levels, which represent the intuitive evaluation (star rating, star rating square) respectively, characteristics of review content (review length, review length square emotional score), and reviewers' characteristics (vine, VP), to explore the various reasons that affect the helpfulness ratings in detail.[4]

As for dependent variables, we divide them into two categories according to their numerical values, namely:

$$helpfulness\ ratings = \begin{cases} 0 & if\ the\ p \leqslant 0.5 \\ 1 & else \end{cases}$$

Note: 0 indicates that the comment is not credible, 1 indicates that the comment is credible

Table 2: Index and regression coefficient



	independent variables	Hair Dryer	Microwave	Pacifier
Constant term	None	0.647	0.494	-0.0973
Intuitive evaluation	Star rating	0.100*	0.021**	0.178**
	Star rating square	0.002*	0.000**	0.012**
characteristics of review content	Review length	0.011*	0.016*	0.006*
	Review length square	-0.212*	-0.021*	-0.002*
	Emotional score	0.082*	0.047**	0.141+
reviewers' characteristics	Vine	0.001*	-0.008**	-0.097**
	Verified purchase	0.001*	-0.008**	-0.097**
Model accuracy		82.8%	83.0%	85.5%

Note: *:p<0.05;**:p<0.01;+:p<0.1

It is clear from the above table, each variable passed the significance test. The regression coefficients of star ratings of microwave oven, hair dryer and pacifier are all positive, and the regression coefficients of star rating square terms are also positive. It indicates that there is a "U-type" relationship between star ratings and the helpfulness ratings of reviews. Our analysis suggests that the reviewers who score higher or lower star ratings may be more willing to express a clear attitude towards the products, so as to provide more valuable reviews; while the reviewers who give intermediate star ratings may lack reference value due to their less distinctive attitude.

From the data in the table, the regression coefficient of the review length term is positive, while one of the square terms of the review length is negative, which indicates that the review length and helpfulness rating are of an inverted-U-type. According to our analysis, reviews that are too short are often limited in content and cannot provide sufficient useful information. However, overlong reviews may provide a lot of valuable information though, other customers may not be patient to spend time and energy reading them. Therefore, reviews with high helpfulness ratings should be of moderate length.

3. Model II: Weight Distribution Model based on EWM Model

3.1 Basic introduction of EWM

The basic idea of entropy method is to determine the objective weight according to the variability of the index. If the information entropy of an index is smaller, it indicates that the greater the degree of variation of the index, the more information it provides, the greater the role it can play in the comprehensive evaluation, and the greater its weight. On the contrary, the greater the information entropy of an index, the less the degree of variation of the index, the less the amount of information provided, and the smaller the role played in the comprehensive evaluation, the smaller its weight.[5]

3.2 General steps for EWM

Considering that the emotional score is [-1,1], to ensure that the matrix is all positive, we re-standardize the data to form a positive matrix to pave the way for later probability calculation.

First of all, our two total evaluation indicators are standardized to form a positive matrix：

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}$$

Then the standardized matrix is extremely Z. Every element in Z is:

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^{2}}}$$

Since there is still a negative number after passing the above standardized formula, we use another standardized formula:

$$z_{ij} = \frac{x_{ij} - min\{x_{1j}, x_{2j}, \cdots x_{nj}\}}{max\{x_{1j}, x_{2j}, \cdots x_{nj}\} - min\{x_{1j}, x_{2j}, \cdots x_{nj}\}}$$

After standardization, the non-negative matrix is

$$\tilde{Z} = \begin{bmatrix} \tilde{z}_{11} & \tilde{z}_{12} \\ \tilde{z}_{21} & \tilde{z}_{22} \\ \vdots & \vdots \\ \tilde{z}_{n1} & \tilde{z}_{n2} \end{bmatrix}$$

Then we calculate the probability matrix *P*, in which each element P_{ij} in *P* is calculated as follows:

$$p_{ij} = \frac{\widetilde{z_{ij}}}{\sum_{i=1}^{n} \widetilde{z_{ij}}}$$

It is easy to verify the following conclusions

$$\sum_{i=1}^{n} p_{ij} = 1$$

This ensures that the probability sum corresponding to each index is 1.

Next, we calculate the information entropy of each index, and calculate the information utility value, from which we get the entropy weight of each index. For the j index, the formula for calculating its information entropy is[6]

$$e_j = -\frac{1}{ln\, n} \sum_{i=1}^{n} p_{ij}\, ln(p_{ij}) (j = 1, 2, \cdots, m)$$

Among them, the larger the e_j, the greater the information entropy of the j index, indicating that the information of the j index is less. For the information utility value, we define $d_j = 1 - e_j$,so the larger the information utility value is, the more information it corresponds to. By normalizing the information utility value, we can get the entropy weight of each index:

$$W_j = \frac{d_j}{\sum_{j=1}^{m} d_j} (j = 1, 2, \cdots, m)$$

3.3 Running results of EWM

After running, the entropy weight of each product is shown in the table:

Table 3: Entropy weight table of each commodity

	weight of SR	weight of Review
Hair dryer	0.49987	0.50012
Microwave	0.49989	0.50010
Pacifier	0.49996	0.50003

According to the EWM model, consumers are equally interested in star scores and reviews when buying goods, so in the prediction of subsequent comprehensive scores, we can give 50% weight to the average emotional score and the average star score respectively, and their sum is the comprehensive score of the product reputation. This lays a solid foundation for us to establish the prediction model of product reputation after that.

References

[1] HEAD Acoustic Loudness and Sharpness calculation with Artem[A]. HEAD Application note 2006.

[2] Zwicker E, Fast1 H. psychoacoustics-Facts and models[M]. 2nd edition Spring -Verlag Berlin,1999.

[3] Fast1 H. Psychoacoustics and Sound Quality[M]. Springer Berlin Heidelberg, 2005.

[4] Karst collapse risk assessment based on AHP-fuzzy comprehensive evaluation [J]. Groundwater, 2016. 38 (4): 114-116.

[5] Qiu, D, Systematic analysis of multi-index comprehensive evaluation method [M]. Beijing: China Statistics Press, 1991.

[6] Wang, G. Quantitative analysis and evaluation method [M]. Shanghai: East China normal University Press, 2003.