Evaluation and Prediction of Cell Phone Sales Based on Various Techniques

Abstract

With the progress of our society as well as the technology, online shopping gradually becomes a trend increasingly preferred by young people. This work mainly speculates on the sales of cell phones as a representative, aiming to construct a model capable of analyzing which are the most crucial factors and traits promoting the success of certain types of cell phones.

To begin with, we use Information Entropy to extract the most crucial factors: Comment Count, Good Comment Count and Search Count. We also employ Principal Component Analysis to complete the same goal. The top significant factors are Display Resolution, Recording Definition, RAM and ROM. Next, we apply the results above to Linear Regression and a Weight determination Technique for the modeling, in pursuit of further detailed conclusion. The method of Weight determination Technique yields straightforward graphs by using qualitative analysis, providing further insight to which specific traits contribute more to the success of the sales volume of that certain type of cell phone.

Furthermore, we optimize all these models with three different methods and employing BP Neural network, Principal Component Regression and Bayes Distinction respectively for quantitative analysis, also concerning which specific traits are more crucial to the sales volume. For the last step of optimization, XG BOOSTING algorithm is applied to produce more reliable and stable results. The feasibility and sensibility of the model are finally tested using the data in the testing set, establishing the application value of the model.

In a word, the model constructed not only yields the ranking of the significance of individual variables related to sales of phones volume but also gives insight about which particular traits contribute more to sales volume. It also enables the manufactures to predict sales volume, given its related features, and they can be more informed of the needs of customers and thus maximizing their profits. The testing of the model proves its stability as well as reliability, making it accessible and valuable for the further application in real life. Besides the practical application, the mathematics methods applied to the model are also better than the previous researches, which yield inconclusive and vague results. Therefore, we believe that the optimized model proposed is a huge improvement both in application and methodology, which fills in the vacancy in a nowadays major economic domain and will yield significant social value. **Key Words**: Information Entropy, Principle Component Regression, Bayes Distinction, BP Neural Network Fitting, XG Boosting algorithm

Contents

1 Background

1.1 Research Background

With technological advancement and social development, the use of the Internet has gradually become widespread around the world. The Internet now has developed to provide a platform for uses ranging from completing daily demands to conducting research. With respect to completing daily demands, the Internet has provided a possibility for online shopping. Given the fact that people nowadays have overwhelmed schedules and heavy workloads due to the fast pace of our society, more and more people prefer to shop online instead of going to department stores and supermarkets in person. However, online shopping possesses deficiencies and inconvenience despite its advantages. Shortcomings like being unable to see the products in person have become the greatest worry among customers as they may risk purchasing low-quality products due to lack of key information presented online. On the other hand, producers also suffer from the worry of selling their products. As a result, determining what characteristics of products are crucial to sales volume is the main challenge for online companies. To solve this problem, we choose a specific kind of productcell phoneto analyze what kinds of cell phones have the highest sale volume.

1.2 Current Research Status

Current Research mainly focuses on several key factors which are considered to influence sales volume. Abroad, Judith Chevalier et al [1] discovers that positive comments are crucial to the purchase choices of customers by examining online comments on Amazon. Christy M.K. Cheung, based on the dual process theory, constructs the model of receiving information to study the factors that influence the online consumer information receiving and finds that comprehensiveness and correlation are the most important factors. Kelly o. Cowart conducts a questionnaire survey of 357 sample of university students in the United States through consumer decision-making form. He finds that in online purchase of clothing, quality consciousness, brand consciousness, fashion consciousness, hedonism, impulsivity, and brand loyalty are positively correlated to consumer buying behavior, while price sensitivity is a negative correlation. Michael d. Smith et al [2] by comparing the shopping network of 20268 valid samples for empirical research, finds that goods brand is one of the most important determinants of consumer decision-making. At the same time, if the package goods and services cannot be apart, brands are considered as the credit guarantee of retailers.

Domestically, Jie Zhang and Jianan Zhong [3] conducted research to analyze how sale promotion influences the minds of customers and predict the purchase choices of customers. Gang Du and Zhenyu Huang [4] employed the Teradata platform to build decision-making tree model to predict purchasing behaviors of customers , further improving the efficiency and accuracy of prediction. Zhanbo Zhao, Luping Sun, and Meng Sun [5] discovered that factors influencing page view and sales volume are substantially different. To be more specific, price, scale, reputation, and insurance have a significant influence on page view and sales volume. Zhihai Hu, Dandan Zhao and Yi Zhang [6] employed sales of skin care products on Taobao as an example to analyze the influence

of online comments on sales volume. The aforementioned researches mainly explored certain factors influencing sales volume but lacked generality. Therefore, online sellers were unable to determine the influential order of all these factors.

With respect to the research methods, current researches mainly employed three methods: Grey Relational Analysis, C2C Model, and BP Neural Network Fitting. As for Grey Relational Analysis, Fatao Wang employed Grey Relational Analysis to determine the main factors for the development of online shopping. Naicong Hou, Xu Zhang, Enjun Zhang [7] presented reputation as the most influential factor of purchase. Xiao Shi [8] conducted a quantitative research of the interrelation of sales and price, comment rate, popularity with the utilization of Grey Relational Analysis. As for the C2C model, Youzhi Xue and Yongfeng Guo [9] employed a Tobit model to discover that customers valued more on price and delivery fee. Jingsha Fu [10] created a quantify model of influential factors. As for BP Neural Network Fitting, Yanli Ma built an evaluating system including refund rate, descriptions and online comments. All these aforementioned methods are theoretically capable of analyzing the influence of certain factors on sales volume but are lack of practicality. In conclusion, current researches have failed to analyze influential factors in a systematic and comprehensive way, and they have failed to reveal specific characteristics that achieve higher sales volume. Therefore, our research results improve the current research methods by offering a clear view into the characteristics that cellphones with high sales volume have and applying our results to predicting sales volume.

1.3 Research purpose and significance

Since online sellers constantly worry about ways to promote sales volume, we conduct research in the hope of offering a practical solution by determining which characteristics contribute to improving sales volume. Our research purposes can be summarized as below:

i. To conduct qualitative research to have a general understanding of the characteristics that contribute to high sales volume.

ii. To conduct quantitative research to rank factors that are considered to have an influence on sales volume.

iii. To determine specific characteristics within each factor that contribute to the highest sales volume.

iv. To predict the sales volume of cellphones with a given characteristic.

Our research results will be of great reference and help to online cellphone sellers by offering a clear explanation of what kinds of cell phones have the highest sales volume. Online cellphones sellers can consequently adjust their products according to our research results to achieve higher sales volume.

1.4 Research method and train of thinking

Figure 1 below presents the whole modeling process. After gathering data of information about product selling in AliExpress, we extract useful and relevant data concerning different influential factors and conduct basic statistics for further research. Next, we come to the data procurement to reduce the number of independent factors. In this process,

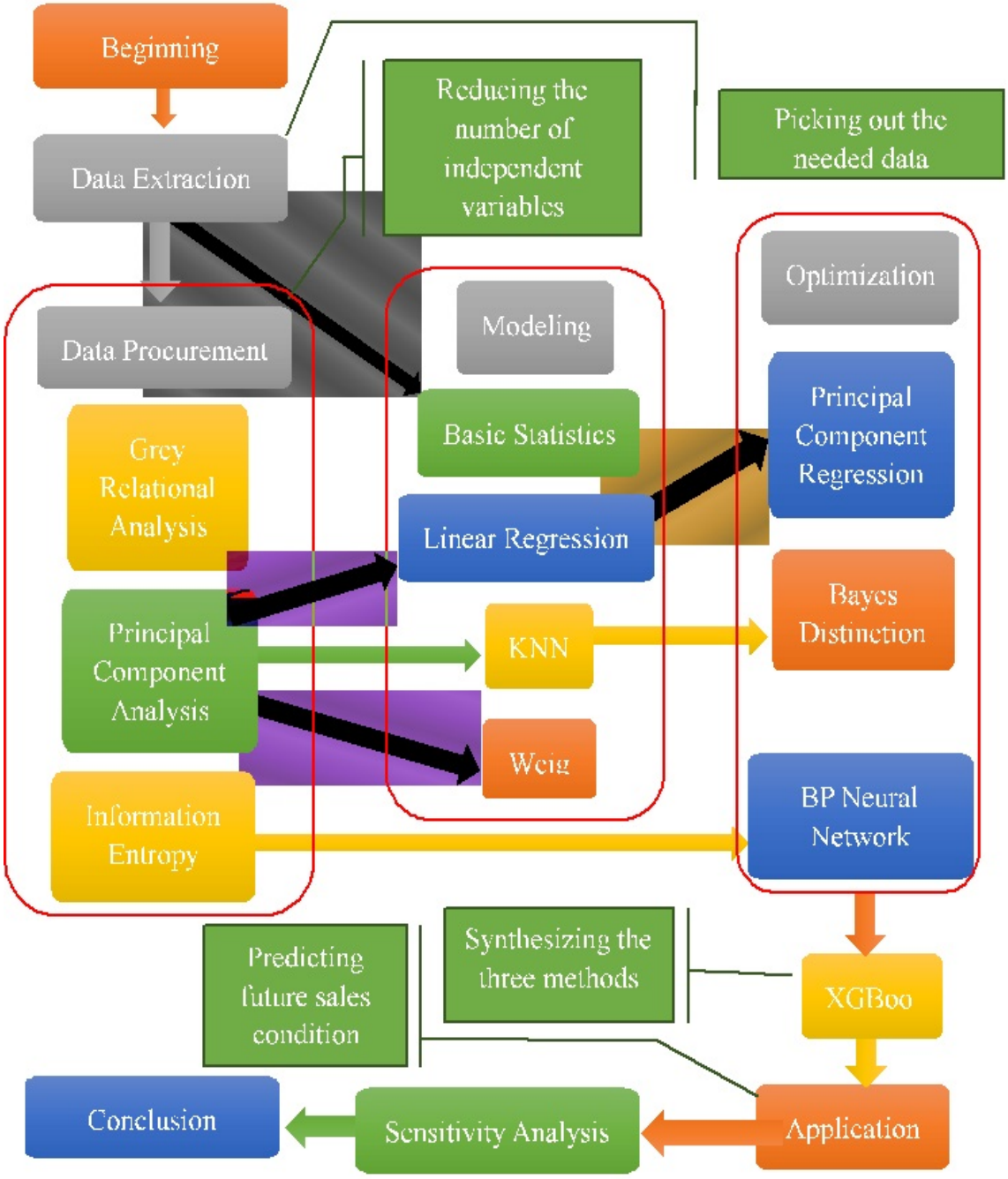

Figure 1: The flow chart of the whole modeling process

we apply three different methodGrey Relational Analysis, Principal Component Analysis, and Information Entropy. The Grey Relational Analysis fails to reduce the number of influential factors, while the other two methods effectively complete the goal. Then we apply the results of data procurement for modeling. In the modeling process, we apply results from Principal Component Analysis to Weight Determination Technique, KNN, and Linear Regression. At this point, we have reached the conclusion of the rank of different independent factors. Furthermore, we conduct optimization to each model. We optimize KNN by Bayes Distinction and Linear Regression by Principal Component Regression, while we optimize Entropy of Information to BP Neural Network Fitting. Afterward, we employ the XG Boosting algorithm to synthesize the three methods and reach the conclusion that which characteristics contribute to the highest sale volume. Finally, we practice the application of our research results by predicting future sales conditions.

2 Assumptions, Justifications, and Definitions

2.1 Assumptions and Justifications

We make the following assumptions in order to simplify the model without much loss of the core of the problem. We also include a justification part to show that our assumptions are reasonable.

- Assumption 1: We assume that considering these two sets of data as the bases for the Information Gain provide authentic information and reflect the ratio of people who are interested in and actually buy the cell phone.

  Justification: the data are also in a more consistent and standardized form which is convenient for later grouping and processing.

- Assumption 2: All the data which pass the data process and procurement part are credible and reliable, meaning that they have no error.

  Justification: In light of the fact that the data are provided by AliExpress, which comes from reliable sources, the data ought to be without fabrication.

- Assumption 3: Except the parameters given in the data, all other factors, including but not limited to the rest properties of the cell phones, such as the advertisement of the cell phones propagated by the manufacturer, are exactly the same, which indicates it has no impact of difference on the click rate and the convert rate between each phone.

  Justification: We make the assumption so as to simplify the problem, while we are unable to find and evaluate the data of the cell phones other than the given ones.

- Assumption 4: All the cell phones are suitable for customers to use, which means all the coasters have no potential safety hazards, and the coasters will not physically and/or mentally harm the users. For instance, the light generated by the screens will not harm the eyes of the users

Justification: In accordance with the local policies and the regulations, all the selling or upcoming cell phones ought to have passed the mandatory security test given by local authorities, which institute the rules to eradicate safety concerns.

2.2 Definitions

Here we clarify the definitions of the key terms we are going to use and the notations we will employ to expand the mathematical derivation.

- **Convert Rate** represents the ratio of the number of people who buy that certain type of cell phone to the number of people who click on the picture online for more detail.
- **Click Rate** represents the ratio of the number of people who click on the picture for more detail to the number of people who browse the internet and see the picture of that certain type of cell phone.

The following table 1 shows the definitions in the paper.

3 Data Procurement and Process

3.1 Data extraction

We have obtained information about sale records on AliExpress, which is under the control of Alibaba. The original data is in the appendix. With the algorithm and formula given by AliExpress, we convert the original data into the readable and understandable data, which can also be seen in the appendix. [11][12]

We utilize PYTHON to extract the parameter cells, which contain several standardized descriptions of the phones. With the help of XLRD module and XLWR module, we search for cells with the assigned field one after another. We divide the searching process into two stages. The first stage is to separate the entire parameters into several fields that contain only one property each; The second stage is to check what each field denotes and use numerical data to characterize the words. For instance, when we search for the battery property, which is detachable, not detachable, or unknown, we first split the cell by ¡br¿ which stands for breaks to obtain strings that merely possess one property in lieu of many.

Then we use the if function to determine whether the obtained string includes target string, which is yes or no standing for detachable or not detachable. If it includes the prior one, we define the corresponding value in the new Excel table as 1. If it includes the latter one, we define the corresponding value in the new Excel table as 2. If it includes neither one, we define the corresponding value in the new Excel table as 0, which stands for unknown.

We set Unlock Phones, Google Play, Battery Type, Display Resolution, Operation System, Gravity Response, GPRS, SIM Card Quantity, Size, Battery Capacity, Camera, Recording Definition, Display Size, Brand Name, CPU, Touch Screen Type, RAM, and

ROM as the keywords for the first stage; we set yes and no as the keywords for the second stage.

In the second stage, there are some individual cases for us to pay attention to. When we extract the color parameters, we search the name of the colors individually, for the reason that a page may contain phones with various colors. We use the binary combinations to express the colors of the phones. We set White, Blue, Rose, Gold, Silver, Grey, Pink, Brown, Orange, Yellow, and Red as the detection keywords, which allows us to obtain eleven-dimensional binary array to demonstrate the colors. The following figure 2 illustrates the process.

When we are extracting the highest camera resolution fields, we search all the fields with camera: and comparing the numerical part of all the fields featuring above, retaining the largest one and disposing of the rest.

As for extracting the size, we come up with a problem that some of the dimensions are expressed in inches, while others are in centimeters or millimeters, triggering inconsistency in units. To solve this issue, we first use x or * to split the value of three dimensions, before we multiply the three parameters, get the volume of the phones, and use a method to determine the critical value that decides the unit of the phones. We select a phone that we regard as normal, calculating the volume among in inches, millimeters, and centimeters. We then obtain the square roots of the products of the size in inches and centimeters, as well as in centimeters and in millimeters, which are regarded as the critical value. We obtain the essential values of volume, which are 36.86334 and 4712.451. If the product is less than 36.86334, we appreciate the unit as inches. Then we multiply the length, width, and height of the phone with 25.4 to obtain the corresponding value in millimeters. If the product is more than 36.86334 but less than 4712.451, we regard the unit as centimeters. Then we multiply the length, width, and height of the phone with 10 to obtain the corresponding value in millimeters. If the product is more than 4712.451, we regard the unit as millimeters. Then we straight write the length, width, and height of the phone into the tables.

Finally, we write the value into the Excel table and obtain the data that we use, which can be seen in the appendix.

3.2 Data extractionGrey Relational Analysis

In the real world, it is commonly seen that what influences a system tends to be multi-factors instead of a single counterpart, while the relationship between the factors is complicated, which gives rise to the fact that it is easy to cover up its essence with mere regards of its appearance, which makes it difficult to get accurate information and distinguish the primary and secondary factors. The grey system analysis method is essentially an analytic method that replaces discrete data with linked concepts. [8]

The grey system theory holds that, although the appearance of the objective system seems to be complicated, and the data is irrelevant, it always functions as a whole, which means it is not random but proves to contain some inherent laws that can be discovered and explored, and the key is how to choose the proper way to figure out the rules of the data and utilize them.

The gray correlation degree is calculated as following in general: First, we standardize the collected evaluation data to ensure that it is treated without dimension; we obtain the

sequence of difference and compute the maximum and minimum variance of the series of difference; we calculate the correlation coefficient and the calculation correlation degree.

Specifically, we consider the dependent variables, which are click rate and conversion rate, as the reference sequence. As shown in the appendix, we let the following sequence 1 denotes the click rate sequence:

$$Y_1 = Y_1^1, Y_1^2, Y_1^3, ..., Y_1^{1324} \tag{1}$$

And we let the following sequence 2 as the convert rate sequence:

$$Y_2 = Y_2^1, Y_2^2, Y_2^3, ..., Y_2^{1324} \tag{2}$$

We consider the 26 series of independent variables as comparing sequence. As shown in the appendix, we let the following sequence 3 denotes the Google play sequence:

$$X_1 = X_1^1, X_1^2, X_1^3, ..., X_1^{1324} \tag{3}$$

And so on, we let the following sequence 4 as the can-design-product sequence:

$$X_{26} = X_{26}^1, X_{26}^2, X_{26}^3, ..., X_{26}^{1324} \tag{4}$$

	Google Play	Battery Type	Battery Capacity(mAh)	
Click Rate	0.958033	0.958033	0.54663	
Convert Rate	0.989033	0.989033	0.563529	
	Display Resolution	Operation System	SIM Card Quantity	
Click Rate	0.958033	0.958033	0.958033	
Convert Rate	0.989033	0.989033	0.989033	
	Recording Definition (P)	Touch Screen Type	RAM(G)	
Click Rate	0.958033	0.958033	0.588991	
Convert Rate	0.989033	0.989033	0.570534	
	ROM(G)	CPU	Display Size (inches)	Size
Click Rate	0.337011	0.958033	0.958033	0.771116
Convert Rate	0.330881	0.989033	0.989033	0.805193
	Brand	Color	Feature	Price
Click Rate	0.958033	0.486192	0.958033	0.555845
Convert Rate	0.989033	0.499513	0.989033	0.539377
	Highest camera resolution(MB)	Dual Camera	Front Camera	
Click Rate	0.771525	0.958033	0.958033	
Convert Rate	0.805639	0.989033	0.989033	
	SearchCnt	GoodCommentCount	Score	
Click Rate	0.752451	0.550346	0.958033	
Convert Rate	0.722596	0.56748	0.989033	
	IsGalleryFeatured	IsHighQuality	CanDesignProduct	
Click Rate	0.958033	0.958033	0.958033	
Convert Rate	0.989033	0.989033	0.989033	

Table 1: Grey Relational Analysis Result

[1] J. Chevalier, D. Mayzlin. The Effect of Word Of Mouth on Sales: Online Book Review [J]. Working Paper, 2003.12.

[2] Michael D. Smith, Erik Brynjolfsson. Consumer Decision-Making at an Internet Shop bot: Brand Still Matters [J]. The Journal of Industrial Economics, 2001.12(4):541-558.

[3] Jie Z., Jianan Z. Research of promotions influence to customers purchasing behaviors [A]. In The 11th National Conference on Psychology [C]. Kaifeng, China, 2007: 278.

[4] Gang D., Zhenyu H. Prediction of customers purchasing behaviors in the Big Data environment [J]. Modernization of Management, 2015, 1(14): 40-42.

[5] Zhanbo Z., Luping S. and Meng S., Research of comparison between factors in C2C influencing page view and sales volume[J].Journal of Management Science,2013,26(1): 58-67.

[6] Zhihai H., Dandan Z. and Yi Z. An Empirical Study on the Effect of Online Reviews on Product Sales [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2015, 12(11): 52-55.

[7] Naicong H., Xu Z., Enjun Z. Grey Relational Analysis of online reputation and sales volumemovie data as an example [J], Modernization of Management, 2015, 2(10):28-30.

[8] Xiao S. Research of influential factors of online sales based on Grey Relational Analysis [D], Yunnan University of Finance and Economics, Yunnan, 2017.

[9] Youzhi X., Yongfeng G. Competitive Strategy of E-business Sellers on Consumer-to-Consumer Platform: Based on Data from Taobao.com [J], Nankai Business Review, 2012, 15(1): 129-140.

[10] Jingsha F. The Study of Influencing Factors and Index System of C2C Online Shop Sales Volume Based on the Soft Set Theory [D], Chongqing Jiaotong University, Chongqing, 2016.

[11] Wenxuan H. Study on the factors influencing the purchase behavior of Wechat business customers [D], Nanchang University, Nanchang, 2016.

[12] Jiao L. Research on customer purchase behavior analysis system based on Data Mining [J], Time Finance, 2015, 2(2): 320-321.

[14] Mingbei C., Gang H., Guoufu Z. Comprehensive evaluation of takeaway website based on AHP methodEleme website as an example [J]. Modern Business, 2015, 12: 57-58.

[16] Jiang W. Comparative Study of Fisher Discriminant and Mahalanobis Distance Discriminant [J]. Journal of Ningbo Polytechnic, 2017, 21(5); 91-94.

[18] Haiwei W. Yu X., Yalin W.A bivariate hierarchical Bayesian approach to predicting customer purchase behavior [J], Journal of Harbin Engineering University, 2007, 28(8): 949-954.

[19] Wu P. Application of Cigarette Sales Forecasting Based on Neural Network [J], Computer Simulation, 2012, 29(3): 227-230.