参赛队员姓名：曹凌微 钱成 田肇阳

中学：清华大学附属中学

省份：北京

国家/地区：中国

指导教师姓名：

论文题目：

Abstract

With the progress of our society as well as the technology, online shopping gradually becomes a trend increasingly preferred by young people. This paper 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.

As for the beginning, we gather the data from AliExpress, making them as the base for our further quantitative analysis. Methods including Gray Rational Analysis, Principal Component Analysis and Information Entropy are then used to process the data and extract the most crucial independent variables among the potential 26 factors. The Information Gain of these factors yield results that Comment Count, Good Comment Count and Search Count are at the top of the ranking, and Principal Component Analysis indicates Display Resolution, Recording Definition, RAM and ROM are also significant factors.

Next, we apply the results above to the further modeling process. The result from Principal Component Analysis are employed in Linear Regression, KNN algorithm and Analytic Hierarchy Process in pursuit of further detailed conclusion. For example, the method of AHP yields straightforward graphs by using qualitative analysis, and provides further insight to which specific traits in one individual variable contributes 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. The results from Information Entropy are applied to the BP Neural network for quantitative analysis, and the results from linear regression and KNN algorithm are employed to the Principal Component Regression and Bayes Distinction respectively. Instead of Analytic Hierarchy Process which only yields qualitative analysis, the three optimized methods above produce quantitative results concerning which specific traits are more crucial to the sales volume. For instances, the results indicates that cell phones with gold color, higher camera resolution and spacious ROM reveal more satisfactory sales condition. For the last step of optimization, BOOST algorithm is applied to produce more reliable and stable results. We also test the model's feasibility by using the data in the testing set, and the model can successfully predict the volume of sales. Finally, the sensitivity analysis validates the model's stability, thus making its use in real life viable and reliable.

In a word, the model constructed not only yields the ranking of individual variables' significance related to the phones' sales 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 customers' needs 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 the real life. Therefore, we believe that the optimized model proposed will yield significant social and economic value.

Key Words: Information Entropy, Principle Component Regression, Bayes Distinction, BP Neural Network Fitting, BOOST 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 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 product——cellphone——to analyze what kinds of cellphones 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 customers' purchase choices 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 customers' minds 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 customers' purchasing behaviors, 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 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 to 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 C2C model, Youzhi Xue and Yongfeng Guo [9] employed 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 influence on sales volume.

iii. To determine specific characteristics within each factor that contribute to 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 clear explanation of what kinds of cellphones 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

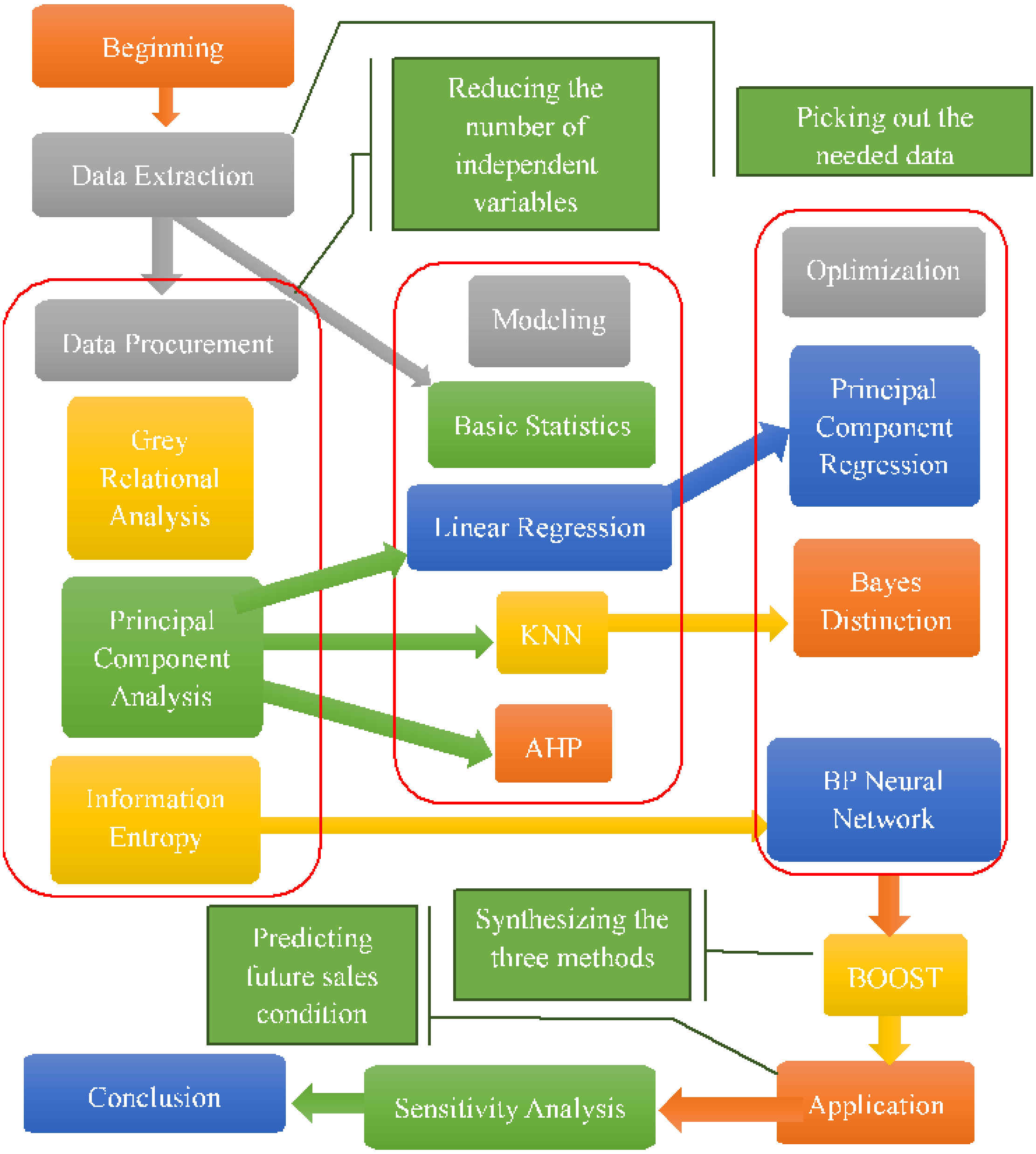


Figure 1: The flow chart of the whole modeling process

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, we apply three different method——Grey 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 from data procurement for modeling. In the modeling process, we apply results from Principal Component Analysis to Analytic Hierarchy Process, 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.

Afterwards, we employ the BOOST algorithm to synthesize the three methods and reach the conclusion that which characteristic contribute to the highest sale volume. Finally, we practice the application of our research results by predicting future sales conditions.

2 Assumptions

2.1 Assumptions

- Category Click 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.
- Category Convert 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.
- 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. In this way, the data are also in a more consistent and standardized form which is convenient for later grouping and processing.

2.2 Definitions

Table 1: the definition of notations

Notation	Definition
$\boldsymbol{A_{ij}}$	The element in i^{th} Row and j^{th} Column in matrix A
$\boldsymbol{X}$	The independent variables matrix
$\boldsymbol{x}$	Row vector of independent variables
$\boldsymbol{y}$	Row vector of dependent variables
$\boldsymbol{\overline{x}}$	The algebra average of several data
$\boldsymbol{Y_1}$	Click rate sequence in Grey Relational Analysis or click rate matrix in other parts
$\boldsymbol{Y_2}$	Convert rate sequence in Grey Relational Analysis or convert rate matrix in other parts
$\boldsymbol{X_k}$	The k^{th} independent variable sequence in Grey Relational Analysis
$\boldsymbol{Y_k^n}$	The n^{th} number in the k^{th} dependent variable sequence in Grey Relational Analysis

$\boldsymbol{X_k^n}$	The n^{th} number in the k^{th} independent variable sequence in Grey Relational Analysis
$\boldsymbol{\Delta X_k^n}$	The difference between each two adjacent terms in independent variable sequences in Grey Relational Analysis
$\boldsymbol{\Delta Y_k^n}$	The difference between each two adjacent terms in dependent variable sequences in Grey Relational Analysis
$\boldsymbol{CC(Y_k)}$	Correlation coefficient of k^{th} dependent variable sequence
$\boldsymbol{CC(Y_k, X_l)}$	Correlation coefficient between k^{th} dependent variable sequence and l^{th} independent variable sequence
$\boldsymbol{\gamma(Y_k, X_l)}$	The correlation degree between k^{th} dependent variable sequence and l^{th} independent variable sequence
$\boldsymbol{E(X)}$	The information entropy regarding the set of incidence X
$\boldsymbol{P_i}$	The possibility that incident numbered i will happen in the set X
$\boldsymbol{E(global)}$	The information entropy of Category Click and Convert Rate
$\boldsymbol{IGain}$	The information gain of individual variables related to the Category Click and Convert Rate.
$\boldsymbol{A_{i,j}}$	The data in the i^{th} line and j^{th}column in the table of data processing concerning information entropy
$\boldsymbol{x_p}$	The p^{th} original variable
$\boldsymbol{z_q}$	The q^{th} New variable
$\boldsymbol{m}$	The number of samples
$\boldsymbol{l}$	The number of variables in each sample
$\boldsymbol{x_{ij}}^*$	The standardized data at row i and column j
$\boldsymbol{x_{ij}}$	The data at row i and column j before standardization
$\boldsymbol{R}$	The correlation coefficient matrix in principal component analysis
$\boldsymbol{\lambda_q}$	The q^{th} characteristic roots or eigenvalues in AHP
$\boldsymbol{a_q(A_q)}$	The q^{th} characteristic vectors
$\boldsymbol{a_{pq}}$	The pth value of the q^{th} characteristic vectors
$(\boldsymbol{w_1} \quad \dots \quad \boldsymbol{w_n})$	Weight vector in Weight Determination Method
$\boldsymbol{n}$	The number of choices of target layer in Weight Determination Method
$\boldsymbol{w}$	The eigenvector in Weight Determination Method
$\boldsymbol{\beta}$	Coefficient matrixes of the original data
$\boldsymbol{\beta'}$	Coefficient matrixes of Principal Component Regression

$\boldsymbol{P\{X\}}$	The probability that satisfies condition X
$\boldsymbol{\alpha}$	Reliability in Regression
$\boldsymbol{\theta}$	Parameters to be estimated of the ensemble in Regression
$\boldsymbol{\hat{\theta}_1}$	The confidence upper limit in Regression
$\boldsymbol{\hat{\theta}_2}$	The confidence lower limit in Regression
$\boldsymbol{d(X,Y)}$	The Mahalanobis distance of the data
$\boldsymbol{\Sigma}$	The covariance matrix
$\boldsymbol{P(B_i\|A)}$	Posteriori probability in Bayes Distinction
$\boldsymbol{P(A\|B_i)}$	Priori probability in Bayes Distinction
$\boldsymbol{P(B_i)}$	The frequency at which the sample appears in Bayes Distinction
$\boldsymbol{G_i}$	The ensemble in Bayes Distinction
$\boldsymbol{f(x)}$	Probability density function of G_i in Bayes Distinction
$\boldsymbol{p_i}$	The priori probability of G_i In Bayes Distinction
$\boldsymbol{k}$	The number of G_i in Bayes Distinction
$\boldsymbol{P\left({}^{j}/_{i}\right)}$	The condition probability of wrongly categorizing the sample of G_i to the ensemble G_j
$\boldsymbol{C\left({}^{j}/_{i}\right)}$	The loss caused by the wrong categorization
$\boldsymbol{D_k}$	A division of a set of distinction samples
$\boldsymbol{ECM}$	The average wrong distinction loss

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 “
” 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 key words for the first stage; we set “yes” and “no” as the key words for the second stage.

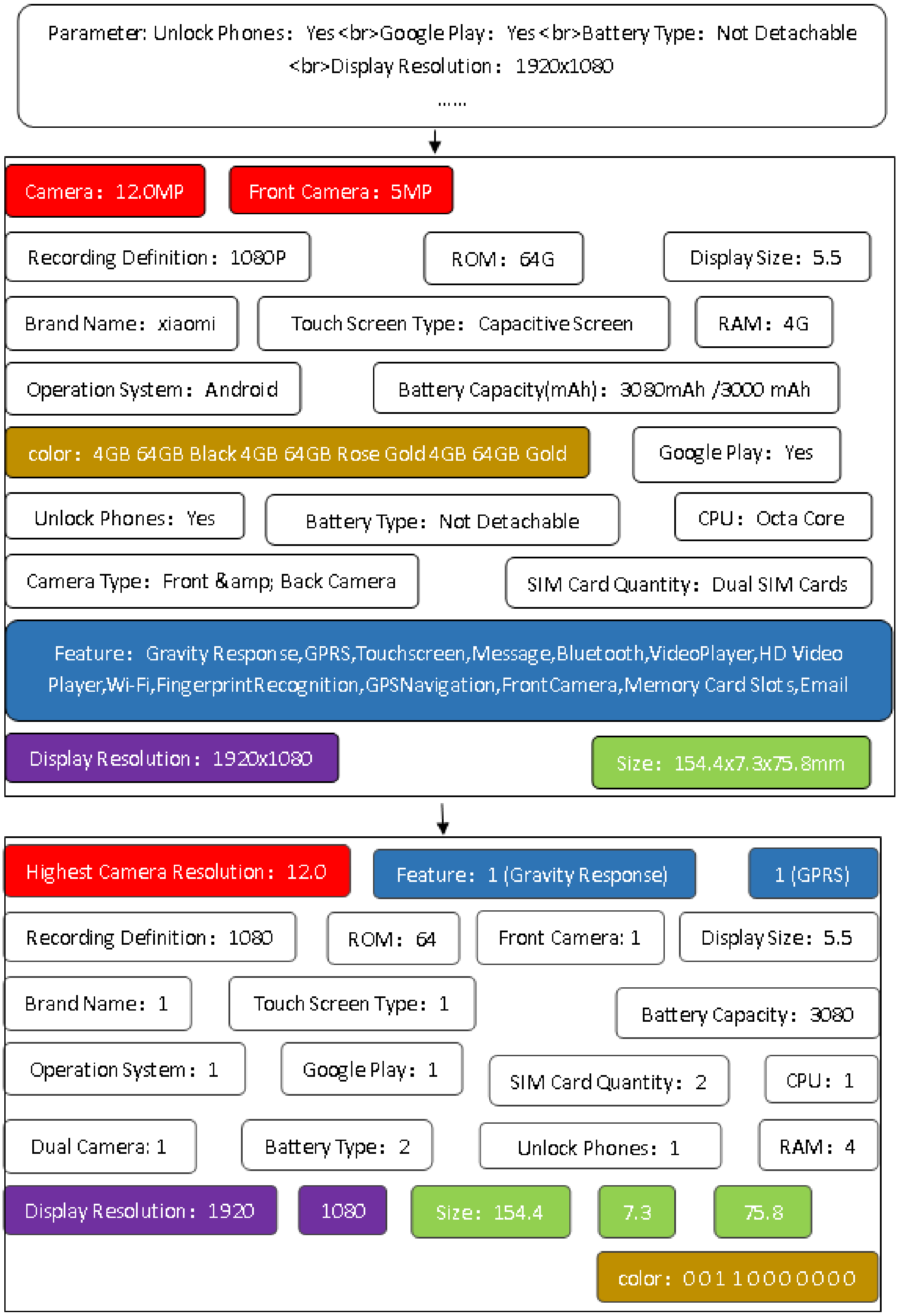


Figure 2: Data Extraction diagram

In the second stage, there are some special 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 key words, which allows us to obtain eleven-dimensional binary array to demonstrate the colors. The following figure 2 demonstrates 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 the rest.

As for extracting the size, we come up with a problem that some of the sizes are expressed in inches, while others are in centimeters or millimeters, triggering inconsistency and inconformity in units. To solve this issue, we first use “x” or “*” to split the value of three dimensions, before we multiply the 3 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, in millimeters and in centimeters. We then obtain the square roots of the products of the volume in inches and in centimeters, as well as in centimeters and in millimeters, which are regarded as the critical value. We obtain the critical volume value, which are36.86334 and 4712.451. If the product is less than 36.86334, we regard 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 Grey 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 complex, 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 concept. [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 laws 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 difference of the sequence of difference; we calculate the correlation coefficient and the calculation correlation degree.

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

$$Y_1 = Y_1^1, Y_1^2, Y_1^3, Y_1^4, \dots, Y_1^{1324} \tag{1}$$

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

$$Y_2 = Y_2^1, Y_2^2, Y_2^3, Y_2^4, \dots, Y_2^{1324} \tag{2}$$

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

$$X_1 = X_1^1, X_1^2, X_1^3, X_1^4, \dots, X_1^{1324} \tag{3}$$

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

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

Then we standardize the data, making the variance of each sequences change into 1 and the mean into 0. We compute the difference between each two adjacent terms, which can be shown as following formula 5-6:

$$\Delta X_k^n = X_k^{n+1} - X_k^n (k\epsilon\{k\epsilon\mathbb{N}^*|k \le 26\}, n\epsilon\{n\epsilon\mathbb{N}^*|n \le 1323\}) \tag{5}$$

$$\Delta Y_k^n = Y_k^{n+1} - Y_k^n (k\epsilon\{k\epsilon\mathbb{N}^*|k \le 2\}, n\epsilon\{n\epsilon\mathbb{N}^*|n \le 1323\}) \tag{6}$$

We finally calculate correlation coefficients and the correlation degree, of which the formula is as following formula 7-9.

$$CC(Y_k) = \left|\sum_{i=1}^{1323} \frac{\Delta Y_k^i}{n}\right| (k\epsilon\{k\epsilon\mathbb{N}^*|k \le 2\}) \tag{7}$$

$$CC(Y_k, X_l) = \left|\sum_{i=1}^{1323} \frac{\Delta Y_k^i - \Delta X_l^i}{n}\right| (k\epsilon\{k\epsilon\mathbb{N}^*|k \le 2\}, l\epsilon\{l\epsilon\mathbb{N}^*|l \le 26\}) \tag{8}$$

$$\gamma(Y_k, X_l) = \frac{1 + CC(Y_k)}{1 + CC(Y_k) + CC(Y_k, X_l)} (k\epsilon\{k\epsilon\mathbb{N}^*|k \le 2\}, l\epsilon\{l\epsilon\mathbb{N}^*|l \le 26\}) \tag{9}$$

Table 2: Grey Relational Analysis Result

	Google Play	Battery Type	Battery Capacity(mAh)	Display Resolution	Operation System	SIM Card Quantity
Click Rate	0.958033	0.958033	0.54663	0.958033	0.958033	0.958033
Convert Rate	0.989033	0.989033	0.563529	0.989033	0.989033	0.989033

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

From the obtained correlation degree, we find that the independent variables which have less value in them are apt to have higher correlation values, symbolizing that a closer connection with the dependent variables. Moreover, the independent variables which have the same number of value possess identical correlation degree, rendering it impossible for us to distinguish how close the connections are between these independent variables and the target dependent variables. We can conclude that the Grey Relational Analysis suits for continuous variables rather than discrete variables, indicating that it is not an ideal technique for us to determine how tight the relationship is under this situation.

3.3 Information Entropy

Information entropy is used to reflect the complexity of the information being processed. Higher information entropy value indicates higher degree of information complexity. Thus, information entropy can be applied to analyze the information in a quantitative way. Information entropy is defined by the formula 10 below:

$$E(X) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{10}$$

where $E(X)$ represents the Information Entropy of X, the set of incidents taken into consideration (in the formula 10 the total number of incidents is n) and $p_i$represents the possibility that the incident numbered i will happen in the set X. The information entropy is calculated in the form of sum of each individual incident.

However, in determining which factor is more important for us to take into consideration among 26 individual variables related to the cell phone as extracted, what is needed should be the amount of information that can be acquired from analyzing on factor instead of its complexity as reflected by the information entropy. Therefore, we utilize information gain to consider which factors are the top ones that should be taken into account as the most crucial. In other words, what kinds of factors contribute more or promote the sale of the smart phones in general. The calculation of information gain of each factor involves its information entropy and is a deliberate and complex process. In the next part of this section, we will mainly discuss the data processing related to the information gain.

First, we identify 26 individual variables as the potential influential factors for sale volume of cellphones including Google play, battery type, brand, RAM, ROM, dual camera, front camera, display size, etc. Then, types of data representing the actual sale volume of cellphones are regarded as the bases for calculating the information gain. Instead of choosing the actual sales volume, we consider the Category Click Rate and Category Convert Rate. Reasons are illustrated in the assumption.

We then divide the Category Click Rate and Category Convert Rate into five groups respectively and reasonably, according to the individual value of the data, from high to low, categorized from 1 to 5. After categorizing the data related to Category Click and Convert Rate, we use formula 10 to calculate the global information entropy of those two sets respectively. As applying the formula to the Category Click Rate, $E(X)$ now represents the information entropy of the Category Click Rate, and p_i represents the possibility of category numbered i will happen. Specially, since there are 5 categories, the number n equals to 5. The same can be applied to the Category Convert Rate, and the final results are shown in table 3 below:

Table 3: The Global information entropy

	Category Click Rate	Category Convert Rate
Global information entropy $E(global)$	2.200779	2.081891

The information we can get from each individual variable is calculated respectively, and the individual variables can be generally classified into two groups: group one with relevant data presenting in inconsistent ways, including factors like Is Gallery Featured and Dual Camera, in which the data only consist of 1, 0, or -1(in other words, the data are expressed in simple forms and can be calculated artificially); group two with relevant data presenting in consistent forms, including factors like Display Size and Display Resolution, in which the data are in various forms and need grouping for further calculation.

As for group one, we take ROM as an example to illustrate how the information entropy is calculated based on the grouping of Category Click Rate. First, we do the grouping and data processing. The data of ROM are presented as discrete variables, including 2, 4, 8, 16, 32, 64, 128, and 256. The grouping of data in ROM should also related to the grouping of Category Click Rate, so accordingly, there are in total 40 groups, which are presented in table 4 below:

Table 4:The grouping

ROM	Group number in Category Click Rate	1	2	3	4	5
	2	0	3	2	1	0
	4	5	6	13	7	0
	8	64	38	63	54	8
	16	110	89	132	143	36
	32	64	44	82	63	29
	64	83	38	62	49	16
	128	3	4	5	7	0
	256	0	0	1	0	0

Let i represents the i^{th} line in the table of the forty groups(as distinguished by double cross lines),j represents the j^{th} column in the table, and A_{ij}represents the number in the unit of the i^{th} line and j^{th} column. Thus, in the unit $A_{4,1}$, the number 110 represents that there are in total 110 data in ROM that are 16 and also in the group 1 as categorized according to the Category Click Rate. Notice that the sum of all the forty groups should equal to the total number of data(and in our data processing, the total number of data available is 1324).

After the grouping of ROM data related to the Category Click Rate, we further calculate the information entropy of the data in each line using formula 10. The information entropy of ROM in each line is shown in table 5:

Table 5: Information entropy of ROM

					Information entropy $E(ROM)$
0	3	2	1	0	1.459148
5	6	13	7	0	1.89366
64	38	63	54	8	2.122787
110	89	132	143	36	2.205866
64	44	82	63	29	2.242444
83	38	62	49	16	2.160525
3	4	5	7	0	1.931295
0	0	1	0	0	0

In order to acquire the total amount of information we can gain from the independent variable ROM, we need to further calculate the possibility that each line will happen. As for the first line$A_{1,j}$, we calculate the times data 2 appears and then divide the total number of data, 1324. Then, we multiply the possibility to the information entropy of each line, the results are shown in table 6:

Table 6: The Information entropy, possibility and their products of ROM

Information entropy	Possibility	Product

1.459148	0.004531722	0.006612
1.89366	0.023413897	0.044338
2.122787	0.171450151	0.363952
2.205866	0.385196375	0.849692
2.242444	0.212990937	0.47762
2.160525	0.187311178	0.40469
1.931295	0.014350453	0.027715
0	0.000755287	0

The sum of all 8 products is the total information entropy we can get from the individual variable ROM. However, for the information gain as related to the Category Click Rate, we need to use the global information entropy of Category Click Rate to subtract the sum of the product above, as the following formula 11 presents:

$$IGain(Category\ Click\ Rate, Rom) = E(global) - \sum Information\ entropy \times Possibility \quad (11)$$

where $E(global)$here represents the global information entropy of the Category Click Rate, since the gain is related to the Category Click Rate. The final gain is presented in table 7 below:

Table 7: The final information gain of ROM

	Sum of the products	$IGain$
ROM	2.174619842	0.026159369

Similarly, the information gain of ROM related to the Category Convert Rate can also be calculated using the method above, and the only difference will be the data in the 40 groups and in the final formula, $E(global)$ should represent the global information entropy of the Category Convert Rate.

As for the group two, we consider the Search Count(the number of time that a certain type of phone is exposed to the customer) as related to the Category Click Rate in order to illustrate the difference of data processing from group one. From the data we have extracted, it is obvious that the data in Search Count are not discrete and the majority of the data of this independent variable are different. However, for the calculation of the information entropy, the number of data in the group should reach a substantial amount, or the final result will be meaningless. Thus, we divide the 1324 data into 5 groups reasonably in order to ensure the number of data in each group for an effective final result.

We divide the data into 5 groups, which are: [0,3000], [3000,30000], [30000,100000], [100000,500000] and [500000, max value]. The later data processing parts are similar as that for the independent variables in group one. The following table 8 presents the grouping of Search Count after the data division:

Table 8: Grouping of Search Count

Search Cnt	Group number in Category Click Rate	1	2	3	4	5

	[0,3000)	220	2	17	13	2
	[3000,30000)	31	106	107	94	19
	[30000,100000)	34	43	77	66	8
	[100000,500000)	29	61	96	81	8
	[500000, max value)	15	10	63	70	52

The data can later be processed as the same way above, and the final information gain of the Search Count related to the Category Click Rate is as shown in table 9:

Table 9: The final information gain of Search Count

	Sum of the products	$IGain$
Search Count	1.808392333	0.392386753

As for other individual variables whose data are not discrete numbers, the same data processing method can be applied. Thus, the information gain of each 26 individual variables as related to the Category Click Rate and Category Covert Rate can thus be calculated. The final information gain is presented in table10 and 11:

Table10: Information gain of each individual variables related to the Category Click Rate

Comment Count	0.732792417
GoodCommentCount	0.680453664
Search Count	0.392386753
Score	0.173242295
Brand	0.124112475
Is Gallery Featured	0.060358001
Battery Capacity(mAh)	0.050232189
RAM(G)	0.031072544
Size	0.028079662
Highest camera resolution	0.026589357
ROM(G)	0.026159369
Price	0.02284587
Color	0.021268354
Display Resolution	0.019607145
Feature(gravity and GPRS)	0.016959237
Is High Quality	0.016942427
CPU	0.016022113
Recording Definition (P)	0.012770426
Display Size	0.011465334
Battery Type	0.010742922
Touch Screen Type	0.008087216
Operation System	0.006372009
SIM Card Quantity	0.006106585
Front Camera	0.001569914
Dual Camera	0.00153786

Google Play	0.000108253

Table 11: Information Gain of each individual variables related to the Category Convert Rate

Comment Count	0.950131659
GoodCommentCount	0.910616696
Search Count	0.631528548
Score	0.288394004
Brand	0.261397755
Is Gallery Featured	0.220147548
Battery Capacity(mAh)	0.102065066
Highest camera resolution	0.067310002
Color	0.052728838
Size	0.052070786
Price	0.040606491
RAM(G)	0.040286271
ROM(G)	0.039561052
Recording Definition (P)	0.038203566
CPU	0.037314954
Display Resolution	0.032897632
Battery Type	0.0300117
Feature(gravity and GPRS)	0.024420792
Display Size	0.020633742
Is High Quality	0.014778733
SIM Card Quantity	0.010565193
Operation System	0.008369328
Front Camera	0.006603513
Touch Screen Type	0.005934561
Google Play	0.005319893
Dual Camera	0.002904891

Higher information gain of the individual variable indicates greater importance of that factor contributing to the sale volume of the product. From tables 10 and 11 above, we can conclude that Comment Count, Good Comment and Search Count contribute more to the sales volume of the products as a whole, while Score, Brand and Is gallery Featured are also significant factors promoting the phone's sales. One thing particularly noticeable is that Category Click Rate and Category Convert Rate are considered separately in the data processing, but they yield similar final result, the fact of which lend the method credibility. Thus, when deciding which variables are more crucial and can be taken into account for the modeling further, the method of information entropy is a relatively clear and reliable way.

3.4 Principal Component Analysis

We regard the sales condition as dependent variables and the properties of phones as independent variables. We try to reduce the dimensionality, diminishing the vast amount of the original data and variables into less data and variables, while the new variables can retain the information in the original data by and large. [13]

We utilize the 26 original variables mentioned in 3.4 as the original data. We still use X to denote independent variables matrixes, $Y_n(n \in \{n \in N^*|n \leq 2\})$. The original variables are $x_p(p \in \{p \in N^*|p \leq l\})$; the new variables are $z_q(q \in \{q \in N^*|q \leq p\})$. We use m to denote the number of samples; we use l to denote the number of variables in each sample. Thus, the data matrix is as matrix 12

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1l} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{ml} \end{bmatrix} \tag{12}$$

Since the data vary in dimensions and ranges, we need to standardize the data. We adopt variance standardization technique to operate the data so that the variance of the standardized data is 1, while we conduct the central translation so that the mean of the data is 0. The formula is as formula 13

$$\overline{x_j} = \sum_{t=1}^{i} \frac{x_{tj}}{i}, \sigma_j = \sqrt{\sum_{i=1}^{n} \frac{\left(\overline{x_j} - x_{ij}\right)^2}{n-1}}, {x_{ij}}^* = \frac{x_{ij} - \overline{x_j}}{\sigma_j} \tag{13}$$

${x_{ij}}^*$denotes the standardized data at row i and column j; x_{ij}denotes the data at row $i$and column j before standardization. i denotes total column number and j denotes total row number.

Then we establish the correlation coefficient matrix R. The formulas are shown in formula 14-15.

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}} \tag{14}$$

$$R = \left(r_{ij}\right)_{l \times l} \tag{15}$$

Then we obtain the characteristic roots $\lambda_q(q \in \{q \in N^*|q \leq l\})$ which satisfy$\lambda_x > \lambda_y$ for $\forall 1 \leq x < y \leq q$ and characteristic vectors $a_q(q \in \{q \in N^*|q \leq l\})$to determine the load a_{pq}on each new principal component variables $z_q$of the original variables x_p, which are equal to the q^{th}larger characteristic values of the correlation matrix corresponding to the characteristic vectors. $a_{pq}$is the p^{th} value of the q^{th}characteristic vectors. The formula is as formula 16

$$RA = \lambda A \tag{16}$$

In the formula, A denotes each characteristic vector, λ denotes each characteristic value. The characteristic roots are shown in table 12. Characteristic vector matrix is in the appendix.

Table 12: Principal Component Regression Characteristic Value

5.830569	2.108099	2.023791	1.548141	1.270862	1.142889	1.079039
0.984911	0.973456	0.896728	0.878234	0.848337	0.823186	

0.770015	0.68834	0.646611	0.571062	0.49684	0.465049	0.413485
0.364857	0.324523	0.276729	0.245753	0.200672	0.127823	

The contribution rate formula and the total contribution rate formula is as 17-18.

$$\frac{\lambda_i}{\sum_{k=1}^{q}\lambda_k}\ (i=1,2,\ldots,p) \quad \text{and} \quad \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{q}\lambda_k}\ (i=1,2,\ldots,p) \tag{17-18}$$

We obtain the total contribution rate until the fourteenth principal component is 81.45%, which is larger than 80%. Therefore, we take the first fourteenth eigenvalue as the principal component. Suppose the principal component is formula set 19

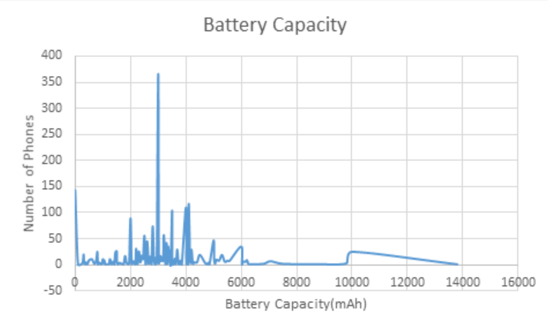
$$z_1 = a_{11}x_1 + a_{21}x_2 + a_{31}x_3 + a_{41}x_4 + a_{51}x_5 + \cdots + a_{26\,1}x_{26}$$
$$z_2 = a_{12}x_1 + a_{22}x_2 + a_{32}x_3 + a_{42}x_4 + a_{52}x_5 + \cdots + a_{26\,2}x_{26} \tag{19}$$
$$\cdots$$
$$z_{14} = a_{1\,14}x_1 + a_{2\,14}x_2 + a_{3\,14}x_3 + a_{4\,14}x_4 + a_{5\,14}x_5 + \cdots + a_{26\,14}x_{26}$$

Since the data are standardized before the analysis, each coefficient is equally likely. We can use the independent variables of which the principal component coefficients are relatively larger in the first several principal components. For instance, there are several original variables in the first principal components, of which the coefficients are relatively larger among all the coefficients of the original variables in the first principal components. We choose two deputies of them to denote the resembling original variables and repeat the process for the second and third principal components. Thus we obtain 6 properties of the phones for further analysis, which are Display Resolution, Recording Definition, RAM, ROM, CPU, Highest camera resolution, and Price.

4 Modeling

4.1 Basic Statistics

After obtaining the original data, we do the basic statistics process. We set the click rate and the convert rate as the dependent variables, while other variables as independent variables. On the one hand, we make pie charts as well as line charts reveal the proportions of the phones with each characteristics over the ensemble, as shown in figure 3-4. On the other hand, to show the cross relationship between the independent variables and dependent variables, we draw the bivariate tables to reveal the proportions of the phones with each characteristics over a certain type of phones. We first categorize the continuous variables into several ranges, in order to discretize the variables. Table 13 is the statistic table of Battery Capacity. We divide the click rate into 5 categories, which are 0-0.1, 0.1-0.2, 0.2-0.225, 0.225-0.3, 0.3-0.464. We divide the convert rate into 5 categories, which are 0-0.1, 0.1-0.2, 0.20-0.22, 0.22-0.23, 0.23-0.468.



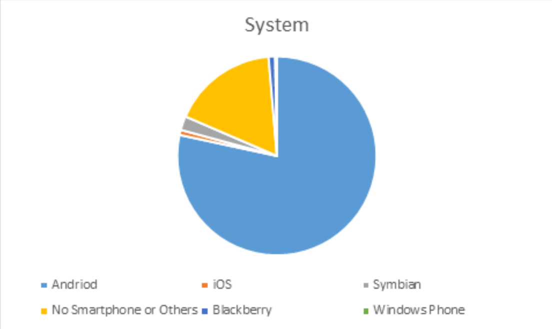


Figure 3: Line Chart of Battery Capacity Figure 4: Pie Chart of System

Table 13: Statistics from Battery Capacity to Click Rate Category

ClickrateCatagory / mAh	1	2	3	4	5
Lower than 3000	93	67	92	88	14
3000	59	40	66	60	8
More than 3000 but less than 4000	80	57	85	79	16
4000mAh to 4100	24	19	53	60	42
More than 4100	73	39	64	37	9

The previous charts demonstrate, for instance: most of the phones possess 3000 mAh, 4000 mAh, or 4100 mAh battery. The phones with Android systems lead the ranking or systems, while Apple system is the second one. For the phones which achieve higher click rate category, they are more likely to have the upper-middle battery capacity.

4.2 Weight Determination Technique

In order to choose by diverse factors and judge the sales of certain types of phones, we create a new Weight Determination Technique, which imitates the Analytic Hierarchy Process (AHP), to achieve the goal which is to determine the weight of each option in complicated and uncertain problems. We define the properties of the phones, which are display resolution, recording definition, RAM, ROM, CPU core, highest camera resolution, and price, obtained from the Principal Component Analysis, as the scheme layer, while defining the click rate and the convert rate as the target layer, to build up the weight determining model with one mere layer but several groups. We divide RAM in to 3 groups, less than 1 GB, no less than 1 GB but less than 4 GB, and more than 4 GB, of which are groups 1, 2, and 3 respectively. We divide ROM in to 3 groups, less than 8 GB, no less than 8 GB but less than 64 GB, and more than 64 GB, of which are groups 1, 2, and 3 respectively. We also divide display resolution, recording definition, highest camera resolution, and price into several categories, of which the standard is the same as what we do in the Information Entropy part. Figure 5 shows the diagram from RAM to click rate. [14]

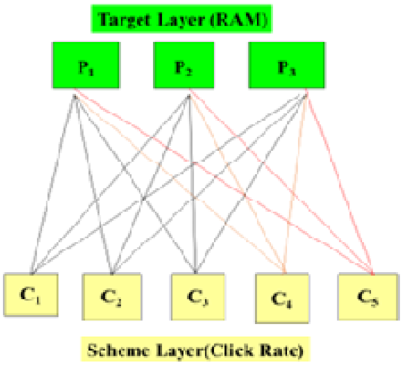


Figure 5: Structure diagram

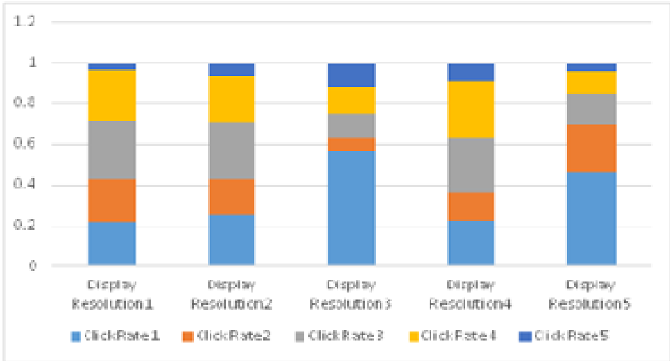


Figure 6: Result analysis

First, we define the amounts of phones that possess certain properties under certain type of sales conditions, which refers to the amount of a certain target choice under a scheme layer condition, as w. In accordance with the target choice, we obtain a weight vector $(w_1 \quad \cdots \quad w_n)$ (n stands for the number of choices of target layer). We compute the ratio between the number, $w_i (1 \leq i \leq n)$, of each scheme layer choice under a common target layer choice and regard it as the weight of paired comparison matrix. As they are consistent matrixes, we do not need to apply consistency tests to the matrixes, for they are automatically consistent, which means that the eigenvalues are all identical. With the help of the formula of the eigenvalue and eigenvectors shown in formula 20,

$$Aw = \lambda w \tag{20}$$

we can obtain the eigenvectors, w. Composing the eigenvalues of each scheme layer, we obtain the eigenvector matrixes as well as weight vector matrixes from target layer to scheme layer.

Then we repeat the process from each scheme layer, which is the sales condition, to each target layer, which is the properties of the phones, to achieve the goal that for each scheme the sum of the weight vector is 1 to transversely compare which option is more welcomed under the same sales condition. Comparing the weight of each scheme to one single target vertically, we obtain which kinds of phones are more welcomed under the same standard.

Finally, we draw the statistical chart with each weight vector, such as stacked column charts, to clearly express the interference of the properties of the phones to the result. The charts are shown in the appendix, one of which is shown as figure 6.

We can clearly see that phones with middle display resolution tend to attract more customer to click in and purchase. Phones with lower and higher recording definition are more welcomed, while phones with medium counterpart are less intriguing. Phones with lower RAM, ROM, and CPU involve in more click rate, whereas phones with higher equivalents involve in more convert rate. For both highest camera resolution and price, the medium ones are both attractive.

4.3 Linear Regression

The third modeling method we use is linear regression. We can regard the properties of phones as independent variables, and the sales as dependent variables. Based on the samples, each data can be viewed as a mapping from the independent variables, which are the properties, to the dependent variables, which are sales. As each data is expressed numerical, we can find the function from the independent variables to the dependent variables through linear regression from the data. [15]

Let x_1 to x_7 respectively denote display resolution, recording definition, RAM, ROM, CPU core, highest camera resolution, and price. Let $y_1$denotes click rate and $y_2$denote convert rate. The value of the independent variables and dependent variables is the numbers of each option.

We utilize regression formula 21.

$$y_n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 (n \in \{1,2\}) \qquad (21)$$

Let X denotes the independent variables matrix; $Y_n(n \in \{n \in N^* | n \leq 2\})$ denote dependent variables matrixes;β denotes coefficient matrixes. We apply Least Square Regression Method to the issue, of which the formula is shown in formula 22:

$$\beta' = (X^T X)^{-1} X^T Y = \left(\sum x_i \, x_i{}^T\right)^{-1} \left(\sum x_i \, y_i\right) (i \in \{i \in N^* | i \leq n\}) \qquad (22)$$

The formula is set to solve out the value of the coefficient matrixes of point estimation. With MATLAB giving solution, we obtain the coefficient matrixes which are presented in table 14:

Table 14: Linear Regression Coefficient

	Click Rate	Convert Rate
β_0	0.017671	0.018321
β_1	7.71E-10	6.24E-10
β_2	1.83E-06	1.68E-06
β_3	-0.00055	-0.00053
β_4	8.85E-06	8.64E-06
β_5	-0.00062	-0.00081
β_6	2.61E-06	-9.50E-06

Point estimation possesses a drawback that it cannot express the accuracy of the data obtained, thus we utilize interval estimation to reuse the Least Square Regression Method, the formula as in formula 23:

$$P\{\hat{\theta}_1 < \theta < \hat{\theta}_2\} = 1 - \alpha \qquad (23)$$

θ denotes the parameters to be estimated of the ensemble; P denotes probability; $\hat{\theta}_1$ denotes Confidence upper limit; $\hat{\theta}_2$ denotes Confidence lower limit; α denotes reliability which satisfies $0 < \alpha < 1$. In this way, we obtain formula 24

$$P\{\hat{\beta}_{n\,1} < \beta < \hat{\beta}_{n\,2}\} = 1 - \alpha \ (n \in \{n \in N^* | n \leq 2\}) \qquad (24)$$

With MATLAB program, we set α as 0.95, under which the regression coefficient bound is shown in table 15:

Table 15: Linear Regression Coefficient Bound

	Click Rate Lower Bound	Convert Rate Lower Bound	Click Rate Upper Bound	Convert Rate Upper Bound
β_0	0.013262	0.013966	0.02208	0.022677
β_1	-5.27E-10	-6.58E-10	2.07E-09	1.91E-09
β_2	-2.08E-06	-2.20E-06	5.75E-06	5.55E-06
β_3	-0.0014	-0.00136	0.0003	0.000311
β_4	-3.08E-05	-3.06E-05	4.85E-05	4.78E-05
β_5	-0.00195	-0.00212	0.000702	0.000499
β_6	-0.00019	-0.00019	0.00019	0.000185

Residual graphs are in the appendix, one of which is shown in figure 7. When examining correlation coefficients, we find the correlation coefficients are as presented in table 16:

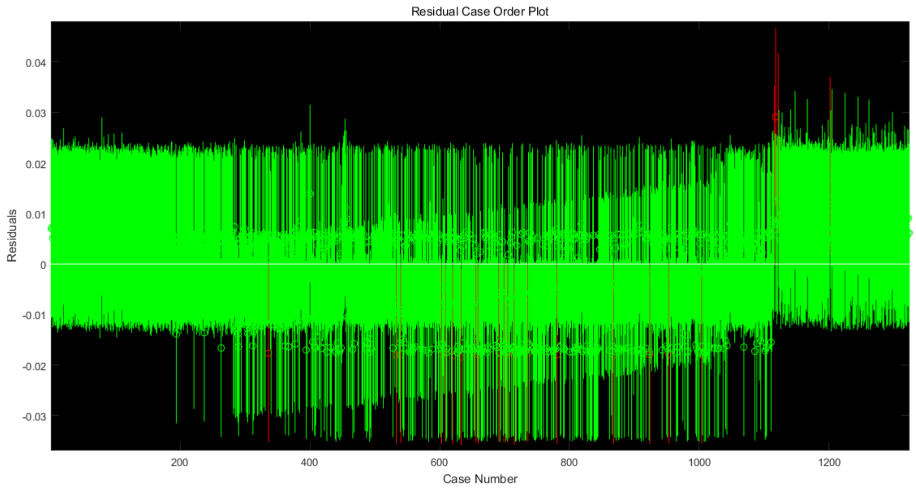


Figure 7: Residual Case Order Plot of Linear Regression

The comparison between the distinction result and the real result is shown in table 17, which shows the accuracy of the distinction. In light of the fact that the accuracy is relative low, which is insufficient to reveal the features of each variables precisely, we consider to take the advantage of other methods.

Table 17: Linear Regression Correlation Coefficient

Click Rate	Convert Rate
0.115124	0.162528

4.4 KNN Algorithm

In accordance with the given data, we try to randomly sample two-thirds of the data as learning samples and one-third of the data as the test data to highly merge the vast amount of the data and find the shared features and characteristics of each sample to obtain the common properties of the phones under similar sales condition to determine the relationship. [16]

We utilize Mahalanobis distance distinction to operate these data, which is processed after principal component analysis and features eradicating the dimension of each independent variables. The formula is as the following formula 25.

$$d(x,y)=\sqrt{(x-y)\Sigma^{-1}(x-y)^{T}} \tag{25}$$

Among the formula, x and y denote two row vectors; Σ denotes the covariance matrix; $d(x,y)$ denotes the obtained Mahalanobis distance of the data.

For the click rate, we correctly categorized 51 samples out of 444, achieving an accuracy of 11%; for the convert rate, we correctly categorized 72 samples out of 444, achieving an accuracy of 16%, which is too low for further application. Thus, we made an optimization in 5.2.

5 Optimization

5.1 Principal Component Regression

Principal Component Regression is specifically suits for the problems that have vast amount of independent data types, not all of which are tightly connected to the dependent data, which means some of the data are loosely related to the data. In view of considering that our problem has 26 independent variables, the method is highly compatible with our research.

We can still do as part 4.3, regarding the sales condition as dependent variables and the properties of phones as independent variables. We try to reduce the dimensionality, diminishing the vast amount of the original data and variables into less data and variables, while the new variables can retain the information in the original data by and large. [17]

We utilize the 26 original variables mentioned in 3.4 as the original data. We still use X to denote independent variables matrixes, $Y_n(n \in \{n \in N^*|n \le 2\})$. The original variables are $x_p(p \in \{p \in N^*|p \le l\})$; the new variables are $z_q(q \in \{q \in N^*|q \le p\})$. We use m to denote the number of samples and use l to denote the number of variables in each sample.

Applying Least squares regression, point estimation and interval estimation method which has previously been mentioned, we obtain the principal coefficient matrix β' as shown in table 18 with formula 25.

$$y_n{}^* = \beta_1{}'z_1 + \beta_2{}'z_2 + \beta_3{}'z_3 + \cdots + \beta_{14}{}'z_{14}(n \in \{n \in N^*|n \le 2\}) \tag{26}$$

Table 18: Coefficient Matrix of principal component

Click rate	Convert rate	Click rate bond		Convert rate bond	
0.006462	0.0062	0.004059	0.008866	0.003961	0.008439
-0.00041	-0.00037	-0.00051	-0.00032	-0.00046	-0.00028
0.0002	2.03E-05	-0.00024	0.000636	-0.00039	0.000427
0.002007	0.001957	0.001797	0.002217	0.001762	0.002153
0.000336	0.000307	0.000162	0.00051	0.000145	0.000468
0.000999	0.0011	0.000672	0.001325	0.000796	0.001405
-0.00419	-0.00407	-0.00458	-0.00379	-0.00444	-0.0037

0.001884	0.001854	0.001271	0.002498	0.001283	0.002426
-0.00422	-0.0042	-0.00502	-0.00343	-0.00494	-0.00347
-0.0012	-0.00111	-0.00209	-0.00031	-0.00194	-0.00029
0.000897	0.000965	0.000346	0.001447	0.000452	0.001478
-0.00077	-0.00071	-0.00107	-0.00047	-0.00099	-0.00043
-0.00069	-0.00067	-0.00091	-0.00048	-0.00087	-0.00047
-0.00056	-0.00059	-0.00098	-0.00014	-0.00098	-0.0002
-0.00091	-0.0009	-0.00106	-0.00076	-0.00104	-0.00076

The correlation coefficients of this method are 0.805032 and 0.826614, which are satisfactory for further calculation. Ultimately, we conduct the inverse standardization process and obtain the equation interpreted in the original data, which is formula 27, and the final coefficient matrix, as shown in table 19.

$$y_n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{26} x_{26} (n \in \{1,2\}) \tag{27}$$

Table 19: Final Coefficient Matrix of original variables

Click rate	Convert rate	Click rate Bound		Convert Rate Bound	
0.0064624	0.0062001	0.0040588	0.008866	0.0039614	0.0084389
-0.000786	-0.000838	-0.001242	-0.00033	-0.001262	-0.000413
-0.000207	-0.000214	-0.000476	0.000063	-0.000465	0.0000374
1.42E-08	4.98E-08	-0.000465	0.0004646	-0.000433	0.0004328
4.31E-10	-3.61E-09	-0.000131	0.0001314	-0.000122	0.0001224
0.0006882	0.0005926	0.0003245	0.0010518	0.0002539	0.0009314
0.000785	0.0007886	0.0012951	0.0002749	0.0012637	0.0003134
1.08E-07	-1.16E-07	-0.000158	0.0001582	-0.000147	0.0001471
0.0009	0.0008054	0.0007563	0.0010438	0.0006715	0.0009393
0.0000249	0.0000324	0.0000949	-0.000045	0.0000976	-3.27E-05
0.0000167	0.0000174	0.0001232	-8.98E-05	0.0001166	-8.17E-05
-8.94E-05	-9.18E-05	-0.000135	-4.39E-05	-0.000134	-4.94E-05
-0.000236	-0.000209	-0.000139	-0.000334	-0.000118	-0.0003
-5.25E-05	-4.31E-05	-0.000661	0.0005556	-0.00061	0.0005232
-4.03E-05	-3.61E-05	-9.11E-05	0.0000104	-8.34E-05	0.0000112
-0.004318	-0.004261	-0.004665	-0.003971	-0.004584	-0.003938
-0.00228	-0.002102	-0.002718	-0.001841	-0.00251	-0.001693
-9.58E-06	-3.22E-06	-0.000632	0.0006129	-0.000583	0.0005766
-0.000208	-0.000244	-0.0003	-0.000117	-0.000329	-0.000158
0.0009217	0.001011	0.0011259	0.0007175	0.0012013	0.0008208
-4.81E-06	-4.41E-06	-3.56E-06	-6.05E-06	-3.24E-06	-5.56E-06
2.92E-10	9.89E-09	-6.03E-05	0.0000603	-5.62E-05	0.0000562
2.09E-06	1.99E-06	-8.34E-05	0.0000876	-7.77E-05	0.0000816
0.0033081	0.0033025	0.0036388	0.0029774	0.0036105	0.0029945
0.0009639	0.0010324	0.0003938	0.001534	0.0005014	0.0015634
-0.002635	-0.002544	-0.003126	-0.002145	-0.003001	-0.002087

0.0016046	0.0016475	0.0011791	0.00203	0.0012513	0.0020438

5.2 Bayes Distinction

Bayes Distinction ideally satisfies the requirements of such issue that each individuals of the ensemble exists at different frequencies, which indicates that we need to take into consideration that the different possibilities that each individuals exists. As for our research, each phones is obviously impossible to appear at identical frequencies, so we apply Bayes Distinction to our study.

In the distance distinction method above, it does not take into account the frequency of each sample in the whole, and does not take into account the loss caused by the wrong distinction. The Bayes distinction method modifies on the basis of distance distinction, and the formula is defined as in formula28: [18]

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\Sigma P(A \mid B_i)P(B_i)} \tag{28}$$

Among which $P(B_i|A)$represents a posteriori probability; $P(A|B_i)$represents a prior probability;$P(B_i)$represents the frequency at which the sample appears;Σ represents the total covariance matrixes. The distinction rule is that the posterior probability is the highest and the average wrong distinction loss is the lowest, which brings out the rule is as follows: If the condition meets the following formula 29:

$$P(G_l \mid x_0) = \frac{p_l f_l(x_0)}{\Sigma p_j f_j(x_0)} = \max_{1 \le i \le k} \frac{p_i f_i(x_0)}{\Sigma p_j f_j(x_0)} \tag{29}$$

Then we categorize$x_0$into G_l,among which G_i is the ensemble, $f(x)$ is the probability density function of G_i, p_i is priori probability of G_i, which is the probability that it belongs a certain category when sample $x_0$occurs, and k is the number of G_i. The solution formula for distinction analysis is as the following formulas 30-31:

$$ECM = \sum_{i=1}^{k} p_i \sum_{j \ne i} C(j/i)P(j/i) \tag{30}$$

$$p(j/i) = P(X \in D_j / G_i) = \int_{D_j} f_i(x)dx \qquad i \ne j \tag{31}$$

In this case, $P\left({}^{j}/_{i}\right)$ represents the conditional probability of wrongly categorizing the sample of $G_i$to the ensembleG_j. $C\left({}^{j}/_{i}\right)$is the loss caused by this categorization.$D_k$is a division of a set of distinction samples. ECM isthe average wrong distinction loss. The solution to a Bayes distinction analysis is to make the smallest set of solutions.

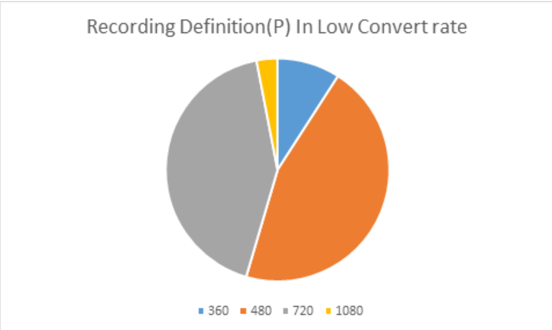
Using the MATLAB program, we still randomly sample $\frac{2}{3}$ of the ensemble as a learning sample and $\frac{1}{3}$ as a test set to carry out Bayes distinction solution. We utilize the data after principal component analysis to study the condition of the distinction.

The result is shown in the appendix, part of which is as following figure 8-9 and table 20. For instance, the number “91” shows that there are 91 samples with 2G RAM are judged as click rate category 1.

For the click rate, we correctly categorized 89 samples out of 444, achieving an accuracy of 20%; for the convert rate, we correctly categorized 176 samples out of 444, achieving an accuracy of 40%, which is relatively higher than the accuracy obtained from KNN algorithm.

Table 20: RAM result in click rate

	Category 1	Category 2	Category 3	Category 4
0.125	2	0	0	0
0.5	7	0	0	1
1	49	0	1	7
1.5	1	0	0	0
2	91	0	15	25
3	41	1	64	19
4	1	13	68	2
6	0	11	22	0
8	0	0	1	0



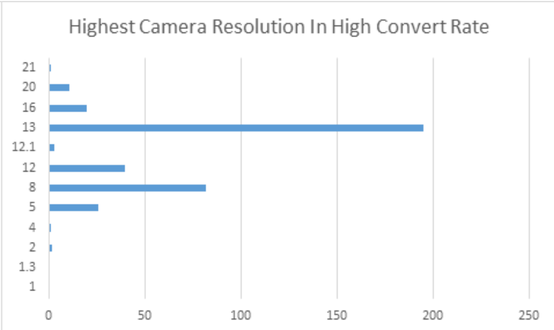


Figure 8: Bayes Result of Recording Definition in Low Convert Rate

Figure 9: Bayes Result of Highest Camera Resolution in High Convert Rate

From the results given, we can clearly figure out the trend that the higher the mobile configuration is, the high click rate and convert rate the sample has. To be specific, phones with higher display resolution, higher recording definition, higher camera resolution, more CPU cores, larger RAM, and more spacious ROM are apt to reveal more satisfactory sales condition. The phones that display weaker sales performance tend to possess lower counterparts of the features listed above.

The comparison between the distinction result and the real result is shown in table 17, which shows the accuracy of the regression. We can see that it is much higher than linear regression.

Table 17: Principal Regression Correlation Coefficient

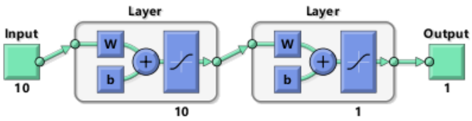
Click Rate	Convert Rate
0.805032	0.826614

5.3 BP Neural Network Fitting

BP Neural Network is a kind of multilayer feed-forward network, which highly fits for the problem that there are data with certain scale, the relationship between which is not too complicated to identify. When it comes to our target, we have a middle-sized database, while the process we want is fitting, which is not too intricate, which shows that the model can be applied to our goal.

We utilize BP neural network fitting as another method to promote the accuracy of the regression. BP neural network works to encode itself with its high-dimensional features and to carry out dimension reduction processing towards high-dimensional data. It is marked by feature extraction model with unsupervised learning, which can also combine few basic features to obtain higher-layer abstract features. [19]

We utilize Tangent Sigmoid function as the transfer function; we use Levenberg Marquardt algorithm (trainlm) as the training algorithm; we use the Gradient descent with momentum weight and bias learning function (learngdm) as the learning algorithm; we use the mean square error (MSE) method as the learning function. The structure of the network and the performance plot are shown in figure 10 and 11.



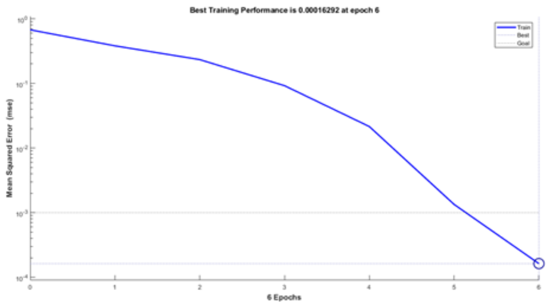


Figure 10: BP Neural Network Structure

Figure 11: the performance plot of BP Neural Network

We utilize the properties after the Information Entropy analysis to conduct the process. Using the MATLAB program, we still randomly samples $\frac{2}{3}$of the ensemble as a learning sample and$\frac{1}{3}$as a test set to carry out the BP neural network fitting. We divide the learning samples in to five groups, each time using four of the groups to carry out a model and then test the test set. Therefore we can obtain five identical models, and then we use the BOOST algorithm to get the means of the five result. The result is in the appendix, part of which is as the following table 21.

It can be seen that some of the predicted data run an accuracy that are higher than 99%. We also utilizes a formula to measure the error of our estimation, reaping an average score of 9.45 of click rate and 9.46 of convert rate out of 10, which shows tl model can successfully reflect the trend. The formula is as the following formul

$$S_k = max\left(0{,}10 - 10 \times \left|\frac{log_{10}\left|\frac{x_{predict}}{x_{real}}\right|}{5}\right|\right) \tag{32}$$

Table 21: BP Neural Network Result

O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AE
					ClickRate mean							ConvertRate	
0.02078	0.011384	0.008857	0.004261	0.021461	0.013348		0.014479	-0.00362	0.02533	0.022221	0.007262	0.013133	
0.020471	0.002053	0.005953	0.002532	-0.00187	0.005829		0.016294	0.001955	0.009609	0.013991	0.007917	0.009953	
0.021792	0.020835	0.013656	0.017043	0.010632	0.016791		0.016819	0.009979	0.020041	0.002321	0.021112	0.014054	
0.015907	0.013295	0.005368	0.002473	0.001955	0.0078		0.017374	0.007013	0.009318	0.010874	0.007137	0.010343	
0.012225	0.020101	0.01652	0.019162	0.021755	0.017953		0.016328	0.008709	0.013901	0.014945	0.017033	0.014183	
0.020453	0.010045	0.006676	0.002474	-0.00278	0.007373		0.013149	0.007797	0.007339	0.013637	0.008797	0.010144	
0.015864	0.010883	0.006219	0.002389	-0.00382	0.006307		0.015677	0.009632	0.007558	0.010519	0.007801	0.010238	
0.016053	0.003646	0.006422	0.0025	-0.00062	0.0056		0.016442	0.00814	0.008473	0.011153	0.007686	0.010379	
0.021603	0.021945	0.016347	0.019545	0.016197	0.019127		0.015357	0.019486	0.012833	0.01432	0.012956	0.01499	
0.018423	0.012065	0.007084	0.002474	-0.00151	0.007707		0.013997	0.007761	0.007401	0.013494	0.008178	0.010166	
0.01248	0.0216	0.012888	0.017519	0.017937	0.016485		0.015002	0.012329	0.016478	0.005335	0.012801	0.012389	
0.016136	0.010158	0.006237	0.002339	0.000183	0.007011		0.015353	0.009618	0.007716	0.011093	0.008007	0.010357	
0.015526	0.008238	0.006689	0.002239	0.002612	0.007061		0.014566	0.011249	0.006636	0.006922	0.004731	0.008821	
0.017688	0.010685	0.008449	0.005342	0.007715	0.009976		0.013769	0.008002	0.002354	0.021218	0.006657	0.0104	
0.024372	0.022045	0.016973	0.019874	0.018377	0.020328		0.014417	0.027674	0.015893	0.014049	0.012057	0.016818	
0.005814	0.020225	0.00924	0.00513	0.011588	0.010399		0.01433	0.010461	0.012536	0.009471	0.006633	0.010686	
0.023505	0.012874	0.016141	0.019691	0.01486	0.017414		0.01492	0.024921	0.017553	0.01721	0.012591	0.017439	
0.021031	0.003062	0.006779	0.002463	-0.00594	0.005479		0.016598	-0.00305	0.011454	0.017018	0.008545	0.010112	
0.007895	0.019178	0.005459	0.002725	0.006349	0.008321		0.014353	-0.00967	-0.00093	0.019134	0.006623	0.005903	
0.004709	0.011095	0.006211	0.002544	0.004508	0.005813		0.018337	0.007168	6.45E-05	0.010817	0.004856	0.008249	
0.020242	0.001131	0.006219	0.002481	-0.00251	0.005513		0.015004	0.007746	0.008241	0.012762	0.008476	0.010446	
0.015613	0.014638	0.006331	0.002231	0.00345	0.008452		0.014434	0.011731	0.006555	0.007063	0.005558	0.009068	
0.022037	0.009821	0.006919	0.001946	0.000596	0.008264		0.02166	0.003533	0.007823	0.010596	0.007032	0.010129	
0.016142	0.010266	0.017648	0.019628	0.022345	0.017206		0.015378	0.016143	0.022779	-0.00053	0.019856	0.014726	
-0.00863	0.017637	0.019993	0.017025	0.030315	0.015269		0.016931	0.016732	0.012842	-0.00257	0.009733	0.010734	

5.4 BOOST Algorithm

We utilize BOOST algorithm to obtain the average value of each methods of the samples. The formula is as the following formula 33

$$\bar{x} = \sum_{i=1}^{3} \frac{x_i}{3} \tag{33}$$

The theory of BOOST algorithm is as follows. For a complicated issue, it is a better judgement when synthesizing the judgment of each experts than that of a sole expert. For each step we generate a model accumulate each model to a whole model, which enables us to analyze the problems.

In the formula, x_1 denotes the original result of principle component analysis. x_2 denotes the result of Bayes distinction. x_3 denotes the original result of BP neural network fitting. For each categories in Bayes distinction, we utilizes the mid-value of each interval to numerate each category. We divide the click rate into 5 categories, which are 0-0.1, 0.1-0.2, 0.2-0.225, 0.225-0.3, 0.3-0.464, as well as the convert rate into 5 categories, which are 0-0.1, 0.1-0.2, 0.20-0.22, 0.22-0.23, 0.23-0.468. Therefore, we use 0.05, 0.15, 0.2125, 0.2625, and 0.382 to denote the 5 result of the categories. We use 0.05, 0.15, 0.21, 0.225, and 0.349 to denote the 5 result of the categories.

6 Application

We use the data which have exactly one zero of each data as the test sets and conduct the BP neural network illustrated in part 5 to obtain the final result to show that our models and methods can be applied to a broader range. We use the data after Principal Component Analysis for Principal Component Regression and Bayes Distinction and data after Information Entropy for BP neural network Fitting. The following table 22 is a part of the final result.

Table 22: Final Result

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	ID	Unlock Ph	Google Pl	Battery T	Battery C	Display R	Display R	Operation	SIM Card	Recording	Touch Scr	RAM(G)	ROM(G)	CPU	Display S	Size_X(mm	Size_Y(mm	Size_Z(mm	Hi
2																			
3	9083.0	1.0	1.0	2.0	4100.0	1280	720	1.0	2.0	720.0	1.0	2.0	0.0	1.0	5.0	139.24	69.96	8.65	13
4	9072.0	1.0	1.0	2.0	4100.0	1280	720	1.0	2.0	1080.0	1.0	3.0	0.0	1.0	5.0	139.24	69.96	8.65	13
5	8160.0	1.0	1.0	2.0	3120.0	1280	720	1.0	2.0	720.0	1.0	2.0	0.0	2.0	5.0	139.5	70.4	8.5	13
6	9273.0	1.0	1.0	2.0	4100.0	1920	1080	1.0	2.0	1080.0	1.0	4.0	64.0	0	5.5	151.0	76.0	8.45	13
7	8585.0	1.0	1.0	2.0	4100.0	1920	1080	1.0	2.0	1080.0	1.0	4.0	64.0	0	5.5	151.0	76.0	8.45	13
8	8574.0	1.0	1.0	2.0	4100.0	1280	720	1.0	2.0	720.0	1.0	3.0	0.0	1.0	5.0	139.24	69.96	8.65	13
9	9835.0	1.0	1.0	2.0	4100.0	1920	1080	1.0	2.0	1080.0	1.0	3.0	64.0	0	5.5	151.0	76.0	8.35	13
10	9078.0	1.0	1.0	2.0	4100.0	1280	720	1.0	2.0	720.0	1.0	2.0	0.0	1.0	5.0	139.24	69.96	8.65	13
11	9863.0	1.0	1.0	2.0	4100.0	1920	1080	1.0	2.0	1080.0	1.0	4.0	64.0	0	5.5	151.0	76.0	8.45	13
12	6724.0	1.0	1.0	2.0	4100.0	1920	1080	1.0	2.0	1080.0	1.0	3.0	32.0	0	5.5	151.0	8.35	76.0	13
13	9089.0	1.0	1.0	2.0	3060.0	1920	1080	1.0	2.0	1080.0	1.0	3.0	32.0	0	5.5	153.6	75.2	7.25	12
14	9856.0	1.0	1.0	2.0	4100.0	1920	1080	1.0	2.0	1080.0	1.0	4.0	64.0	0	5.5	151.0	76.0	8.45	13
15	9097.0	1.0	1.0	2.0	4100.0	1920	1080	1.0	2.0	1080.0	1.0	3.0	32.0	0	5.5	151.0	76.0	8.35	13
16	36.0	1.0	2.0	1.0	1500.0	128	160	4.0	2.0	360.0	3.0	0.125	2.0	0	1.77	115.0	49.0	13.5	1.
17	8592.0	1.0	1.0	2.0	3060.0	1920	1080	1.0	2.0	1080.0	1.0	3.0	32.0	0	5.5	153.6	75.2	7.25	12
18	8362.0	1.0	1.0	2.0	3060.0	1920	1080	1.0	2.0	1080.0	1.0	4.0	32.0	0	5.5	153.6	75.2	7.25	12
19	8579.0	1.0	1.0	2.0	3060.0	1920	1080	1.0	2.0	1080.0	1.0	4.0	32.0	0	5.5	153.6	75.2	7.25	12
20	8188.0	1.0	1.0	2.0	3000.0	1920	1080	1.0	2.0	1080.0	1.0	4.0	32.0	0	5.5	151.1	74.2	7.5	21
21	1799.0	1.0	2.0	1.0	3000.0	320	240	4.0	3.0	360.0	3.0	0.125	2.0	0	2.4	132.0	62.0	22.0	2.
22	503.0	1.0	2.0	1.0	1450.0	240	320	4.0	2.0	360.0	3.0	0.125	2.0	0	2.4	55.0	126.5	15.5	2.
23	7531.0	1.0	0	1.0	2600.0	1920	1080	1.0	1.0	1080.0	1.0	2.0	16.0	2.0	5.0	136.6	69.8	7.9	13
24	240.0	1.0	2.0	1.0	1500.0	320	240	4.0	2.0	360.0	3.0	0.125	2.0	0	1.8	100.0	43.0	18.0	0.
25	9612.0	1.0	1.0	2.0	3060.0	1920	1080	1.0	2.0	1080.0	1.0	3.0	32.0	0	5.5	153.6	75.2	7.25	12

7 Sensitivity Analysis

Sensitivity analysis is a method of studying and analyzing the sensitivity of the model to changes in system parameters or surrounding conditions. In our team's optimization methods, it can detect the stability of our model, especially when the given data is not accurate.

In this part we will mainly discuss the sensitivity of the application part. If we give the test set of the data an increase or an decrease of 1%, by changing the value of the original data matrix on the program, we discover that the output data of the principal component regression changes exactly 1%; almost all the results in the Bayes Distinction part have no change in categories; the majority of the output of BP neural network model fluctuates 1% approximately. The output after the change is small enough for us to make further adjustment. Therefore, it is acceptable in the modeling. This sensitive analysis also indicates that our model has universality and can be applied to more situations. For instance, if there is some error in the data, out final result does not vary rapidly correspondingly. Therefore, our model is relatively stable. The data of Sensitivity Analysis can be referred to the appendix part.

8 Conclusion

8.1 Strength and Weakness

The method we propose in the paper has effectively made up the vacancy and deficiency of the previous evaluating process regarding to the sale volume of cell phones, and several main advantages are as the following. For a start, it presents the ranking of the most important individual variables within the cell phone market, the results of which are seldom considered by manufacturers but actually of great significance. Manufacturers can take specific traits of cellphones into consideration, deciding which types or combinations of traits are more profitable to produce and fitting the need of their target customers. Furthermore, as the application section in the paper indicates, the process we propose can also be applied to pragmatic purposes. By using the method linked with BP neural network, the process can successfully predict the outcome of the sale volume of cellphones before they are released into the market, and the margin of error is within an acceptable level. Besides the application of the evaluating process in the real life, the method itself is also more advanced and comprehensive than that in the previous thesis. For the method of information entropy in data processing and BP neural network in the optimization, they not only fit in the exact needs of the data being processed and the expected outcome, but they are also more precise and reliable, ensuring the credibility of the model as a whole.

Admittedly, there are several shortcomings concerning the whole paper, like for some particular methods including Grey Relational Analysis, the results are not very desirable, and they are not quite useful for the later optimization. However, considering the methods being applied as a whole, the advantages obviously outweigh the deficiencies, thus making the modeling reliable for reference and have high practical value.

8.2 Conclusion

We discovered some intriguing and unexpected conclusion throughout the modeling process. Phones with upper-middle display resolution sell better, while phones with middle and the highest counterparts sell worse. Phones with the lowest and highest recording definition gain better sales. It is the most amusing that phones sold either the best or the worst focus onto the highest RAM, ROM and CPU. As for the Highest Camera Resolution and Price, the ones with middle and lower-middle condition sell well, and the ones with upper-middle or the highest equivalents sell experience a tough sell.

From the result above, we can explain that phones with excessive specs are apt to experience harsh sales, since the prices are too high for common users to purchase, while the users do not need such high specs on phones. The reason why phones are sold either the best or the worst focus onto the highest RAM, ROM, and CPU can be attributed to the fact that the highly-equipped phones have more risks in the cost performances, which means these phones are either very welcomed or cold-welcomed by customers, since the prices are relatively high for the highly-equipped phones. Thus the customers may be more cautious when buying a cost-consuming phones than buying a low-cost counterparts.

Despite the specific conclusion and some reasonable explanation, our research also yields significant results in the following four aspects. First, the method concerning AHP produces qualitative analysis of which specific traits in the individual variables promote the sale of the phones the best way. For example, regarding to the display resolution, target readers can clearly make out the third category as bringing more profit and contributing more to the sale. The results can be compiled into graphs and thus providing the whole picture in a straightforward way. Besides, AHP is an easily accessible method and produces relatively reliable result.

Second, the quantitative research can be used to rank the factors and determine which factors are the most crucial ones that the manufacturers should take into consideration. The results reflect the customers' tendency towards different types of cell phones, and their preference is carefully studied using information entropy in the data processing. The ranking of individual variables gives the target readers a broader view of which ones are the keys promoting the sale and lays the stepping stone for the further optimization relating to the different traits in individual variables.

Third, the optimization process allows the determination of specific characteristics that contribute to the highest sales volume. The optimization using Bayes distinction and BP neural network further explores the result in specific details. For example, as for the individual variable like color, it can be analyzed that gold contributes to the highest sale, the fact of which will definitely give the manufacturers more detailed references when making decisions about the production of certain cell phone. For other variables, the same method can be applied, either, yielding valuable insight to specific traits.

Last but not least, the sales volume of the cell phones can be successfully predicted applying the method in the optimization process, as mentioned in the application section. This enables the manufacturers to predict the sale with given characteristics and according to the sensitivity analysis and data testing, the model is reliable and can be applied for other practical uses.

9 Reference

[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 Shopbot: Brand Still Matters [J]. The Journal of Industrial Economics, 2001.12(4):541-558.

[3] Jie Z., Jianan Z. Research of promotion's 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 volume——movie 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.

[13] http://blog.csdn.net/MATLAB_matlab/article/details/59483185?locationNum=10&fps=1, MATLAB principal component analysis.

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

[15] https://wenku.baidu.com/view/99c8408e6529647d272852cd.html, Interval estimation and linear regression analysis with MATLAB.

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

[17] Yimeng F. Analysis of influencing factors of customer purchase behavior in E-commerce [J], Industrial & Science Tribune, 2014, 13(8): 138-139.

[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.

10 Acknowledgement

请说明每一个队员在研究报告撰写中承担的工作以及贡献；

并对他人协助完成的研究成果进行说明。

如果有必要，最后可以列出团队成员和指导老师的简历。

Blablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablabla

11 Declaration

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员：__________ 指导老师：__________

年 月 日

12 Appendix

Blablablablablablablablablablablablablablablablablablablablablablablablablablablabla blablabla