**第一页为封面页**

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：

**Summary**

blablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablablabla

**Key Words:** Grey Relational Analysis, Entropy of Information, Analytic Hierarchy Process, KNN algorithm, Principle Component Regression, Bayes Distinction, BP Neural Network Fitting, K-Ford algorithm, BOOST algorithm

**Contents**

1. **Background**
   1. **Current Research Status**
   2. **Literature Review**
   3. **Restatement of the Problem**
2. **Assumptions**
   1. **Assumptions**
   2. **Definitions**
3. **Data Procurement and Process**
   1. **Data extraction**
   2. **Grey Relational Analysis**
   3. **Entropy of Information**
   4. **Principal Component Analysis**
4. **Modeling**
   1. **Basic Statistics**
   2. **Analytic Hierarchy Process**
   3. **Linear Regression**
   4. **KNN Algorithm**
5. **Optimization**
   1. **Principal Component Regression**
   2. **Bayes Distinction**
   3. **BP Neural Network Fitting**
   4. **BOOST Algorithm**
6. **Application**
7. **Conclusion**
   1. **Sensitivity Analysis**
   2. **Strength and Weakness**
   3. **Conclusion**
8. **Reference**
9. **Acknowledgement**
10. **Appendix**
11. **Background**
    1. **Current Research Status**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

* 1. **Literature Review**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

* 1. **Restatement of the Problem**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

1. **Assumptions**
   1. **Assumptions**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

* 1. **Definitions**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

1. **Data Procurement and Process**
   1. **Data extraction**

We obtained a partial of the information of the goods selling in AliExpress which is under control of Alibaba, which 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.

We utilizes PYTHON to extract the parameter cells, which contains several standardized description of the phones. With the help of XLRD module and XLWR module, we search for the 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 either detachable, not detachable, or unknown, we first split the cell by “<br>” which stands for break to obtain strings that merely possesses 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”, which stands 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.

In the second stage, there are some special case for us to pay attention to. When we extracting the color parameters, we search the name of the colors individually, for the reason that a page may contains 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.

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, which triggers the inconsistency and inconformity of the units. To solve the 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 are 36.86334 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.

* 1. **Grey Relational Analysis**

In the real world, it is commonly seen that what influence a system tends to be multi-factors instead of a single counterpart, while the relationship between the factors is complex, which give 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.

The grey system theory holds that, although the appearance of objective system seems to be complicated and the data is irrelevant, it always has the function of the whole, so it must also contain some inherent and can excavate law, and the key is how to choose the proper way to excavate the law of the data and utilize it.

The gray correlation degree is calculated as follows 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 denotes the click rate sequence

And we let the following sequence as the convert rate sequence

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

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

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

We finally calculate correlation coefficient and the calculation correlation degree, of which the formula is as follows.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 dependent variables which have less value in it are apt to have higher correlation values which symbolize that they have a closer connection with the independent variables. Moreover, the independent variables which have the same number of value possess identical correlation degree, which renders 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, which indicates that it is not an ideal technique for us to determine how tight the relationship is under this situation.

* 1. **Entropy of Information**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

* 1. **Principal Component Analysis**

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.

We utilizes the 26 original variables mentioned in 3.4 as the original data. We still use to denote independent variables matrixes, . The original variables are ; the new variables are . We use to denote the number of samples; we use to denote the number of variables in each samples. Thus the data matrixes is

Since the data varies in dimensions and ranges, we need to standardize the data. We adopt variance standardized 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 18-20

(18-20)

denotes the standardized data at row and column ; denotes the data at row and column before standardization. denotes total column number and denotes total row number.

Then we establish the correlation coefficient matrix , of which the formula are as the following formula 24-25.



(24)

(25)

Then we obtain the characteristic roots which satisfies for and characteristic vectors to determine the load on each new principal component variables of the original variables , which is equal to the larger characteristic values of the correlation matrix corresponding to the characteristic vectors. is the th value of the characteristic vectors. The formula is as formula 26

(26)

In the formula, denotes each characteristic vectors, denotes each characteristic value. The characteristic roots are in the following table 25. Characteristic vector matrix is in the appendix.

Table 25: 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 |  |

Since the data is standardized before the analysis, each coefficients are equally likely. We can use the independent variables of which the principal component coefficients are relatively larger in the first several 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.

1. **Modeling**
   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 chart to reveal the proportions of the phones with each characteristics over the ensemble, as shown in figure 5-6. 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 20 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Clickrate Catagory  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 |

* 1. **Analytic Hierarchy Process**

In order to choose in diverse factors and judge the sales of certain types of phones, we utilizes Analytic Hierarchy Process(AHP) to achieve the goal which is to determine the weight of each options 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 AHP 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 is 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 is 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 entropy of information part. Figure 7 shows the diagram from RAM to click rate.



First, we define the amount of phones that possess certain properties under a certain type of sales conditions, which refer to the amount of a certain target choice under a certain scheme layer condition, as . In accordance with the target choice, we obtain a weight vector ( stands for the number of choices of target layer). We compute the ratio between the number, , of each scheme layer choices 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 the eigenvalues are all identical. With the help of the formula of the eigenvalue and eigenvectors, , we can obtain the eigenvectors, . 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 schemes 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 schemes 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 chart, 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 8.

From the figure in the appendix, 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.

* 1. **Linear Regression**

The third 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 from the given data to find the function from the independent variables to the dependent variables through linear regression.

Let to respectively denote display resolution, recording definition, RAM, ROM, CPU core, highest camera resolution, and price. Let denotes click rate and denote convert rate. The value of the independent variables and dependent variables are the numbers of each options.

We utilizes regression formula 14,

Let denotes the independent variables matrix, denote dependent variables matrixes, denote coefficient matrixes. We apply Least Square Regression Method to the issue, of which the formula is shown as the following formula 15

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 as following table 21

Table 21: Linear Regression Coefficient

|  |  |  |
| --- | --- | --- |
|  | Click Rate | Convert Rate |
|  | 0.017671 | 0.018321 |
|  | 7.71E-10 | 6.24E-10 |
|  | 1.83E-06 | 1.68E-06 |
|  | -0.00055 | -0.00053 |
|  | 8.85E-06 | 8.64E-06 |
|  | -0.00062 | -0.00081 |
|  | 2.61E-06 | -9.50E-06 |

Point estimation possesses a drawback that it cannot express the accuracy of the data obtained, thus we utilizes interval estimation to reuse the Least Square Regression Method, the formula as following formula 16

denotes the parameters to be estimated of the ensemble; denotes probability; denotes Confidence upper limit; denotes Confidence lower limit; denotes reliability which satisfies . In this way, we obtained formula 17

With MATLAB program, we set as 0.95, under which the regression coefficient bound is as following table 22

Table 22: Linear Regression Coefficient Bound

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

Residual graphs are in the appendix, one of which shown as following. When examining correlation coefficient, we find the correlation coefficient is as the following table 23.



Table 23: Linear Regression Correlation Coefficient

|  |  |
| --- | --- |
| Click Rate | Convert Rate |
| 0.011722 | 0.011268 |

* 1. **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 samples to obtain the common properties of the phones under similar sales condition to determine the relationship.

We utilizes 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 34.



Among the formula, and denote two row vectors; denotes the covariance matrix; denotes the obtained Mahalanobis distance of the data.

The comparison between the distinction result and the real result is shown in table 29, 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 29: Linear Regression Correlation Coefficient

|  |  |
| --- | --- |
| Click Rate | Convert Rate |
| 0.115124 | 0.162528 |

1. **Optimization**
   1. **Principal Component Regression**

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.

We utilizes the 26 original variables mentioned in 3.4 as the original data. We still use to denote independent variables matrixes, . The original variables are ; the new variables are . We use to denote the number of samples; we use to denote the number of variables in each samples.

After the work in part 3.4 ,we apply the contribution rate formula and the total contribution rate formula 27-28

and 

(27-28)

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 29

(29)

Applying Least squares regression, point estimation and interval estimation method which has previously been mentioned, with formula 32 we obtain the principal coefficient matrix as shown in table 25

(32)

Table 25: Final Coefficient Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 is satisfactory for further calculation. Ultimately, we conduct the inverse standardization process and obtained the equation interpreted in the original data, which is formula 33, and the final coefficient matrix, as shown in table 26.

(33)

Table 26: Final Coefficient Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Click rate | Convert rate | Click rate bond | | Convert rate bond | |
| 0.0064624 | 0.00620011 | 0.00405878 | 0.00886602 | 0.00396136 | 0.00843887 |
| -0.0007861 | -0.0008376 | -0.001242 | -0.0003302 | -0.0012622 | -0.0004129 |
| -0.0002066 | -2.14E-04 | -0.0004761 | 6.30E-05 | -0.0004647 | 3.74E-05 |
| 1.42E-08 | 4.98E-08 | -0.0004646 | 0.00046463 | -0.0004327 | 0.0004328 |
| 4.31E-10 | -3.61E-09 | -0.0001314 | 0.00013144 | -0.0001224 | 0.00012242 |
| 0.00068817 | 0.00059262 | 0.00032451 | 0.00105183 | 0.00025392 | 0.00093136 |
| 0.00078499 | 0.00078859 | 0.00129513 | 0.00027486 | 0.00126373 | 0.00031344 |
| 1.08E-07 | -1.16E-07 | -0.0001579 | 0.00015815 | -0.0001473 | 0.00014709 |
| 0.00090004 | 0.00080539 | 0.00075629 | 0.00104378 | 0.00067153 | 0.0009393 |
| 2.49E-05 | 3.24E-05 | 9.49E-05 | -4.50E-05 | 9.76E-05 | -3.27E-05 |
| 1.67E-05 | 1.74E-05 | 0.00012321 | -8.98E-05 | 0.00011664 | -8.17E-05 |
| -8.94E-05 | -9.18E-05 | -0.0001349 | -4.39E-05 | -0.0001342 | -4.94E-05 |
| -0.0002364 | -0.000209 | -0.0001388 | -0.000334 | -0.0001181 | -0.0002999 |
| -5.25E-05 | -4.31E-05 | -0.0006606 | 0.0005556 | -0.0006095 | 0.00052323 |
| -4.03E-05 | -3.61E-05 | -9.11E-05 | 1.04E-05 | -8.34E-05 | 1.12E-05 |
| -0.0043182 | -0.0042611 | -0.0046653 | -0.0039711 | -0.0045844 | -0.0039378 |
| -0.0022796 | -0.0021015 | -0.0027178 | -0.0018413 | -0.0025097 | -0.0016934 |
| -9.58E-06 | -3.22E-06 | -0.0006321 | 0.0006129 | -0.000583 | 0.00057656 |
| -0.0002082 | -2.44E-04 | -0.0002998 | -0.0001165 | -0.0003291 | -0.0001584 |
| 0.00092168 | 0.00101103 | 0.00112591 | 0.00071745 | 0.00120125 | 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 | 6.03E-05 | -5.62E-05 | 5.62E-05 |
| 2.09E-06 | 1.99E-06 | -8.34E-05 | 8.76E-05 | -7.77E-05 | 8.16E-05 |
| 0.00330809 | 0.00330248 | 0.00363881 | 0.00297738 | 0.00361051 | 0.00299445 |
| 0.0009639 | 0.00103239 | 0.00039379 | 0.00153402 | 0.00050138 | 0.00156339 |
| -0.0026353 | -0.0025437 | -0.003126 | -0.0021446 | -0.0030008 | -0.0020867 |
| 0.00160456 | 0.0016475 | 0.00117913 | 0.00202999 | 0.00125125 | 0.00204375 |

* 1. **Bayes Distinction**

In the distance discrimination 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 discrimination. The Bayes distinction method modified on the basis of distance distinction, and the formula is as follows 35:



Among which represents a posteriori probability, represents a priori probability, represents the frequency at which the sample appears, and represents the total covariance matrix. 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 36:



Then we categorize into , among which is the ensemble, is the probability density function of , is priori probability of , which is the probability that it belongs a certain category when sample occurs, and is the number of . The solution formula for distinction analysis is as the following formulas 37-38:

and



In this case, represents the condition probability of wrongly distinction the sample of to the ensemble . is the loss caused by this distinction. is a division of a set of distinction samples. is the 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 samples as a learning sample, 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 and graph. For instance, the number “91” shows that there are 91 samples with 2G RAM are judged as click rate category 1.

Figure: 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 |

From the result 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 displays weaker sales performance tend to possess lower counterparts of the features listed above.

* 1. **BP Neural Network Fitting**

We utilizes 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.

We utilizes Tangent Sigmoid function as the transfer function; we use levenberg marquardt algorithm (trainlm) as the training althorithm; 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 is shown as follows.





We utilizes the properties after the entropy of information analysis to conduct the process. Using the MATLAB program, we still randomly samples as a learning sample, 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 samples 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 follows.



It can be seen that some of the predicted data runs an accuracy that is higher than 99%, which shows that this model can successfully reflect the trend.

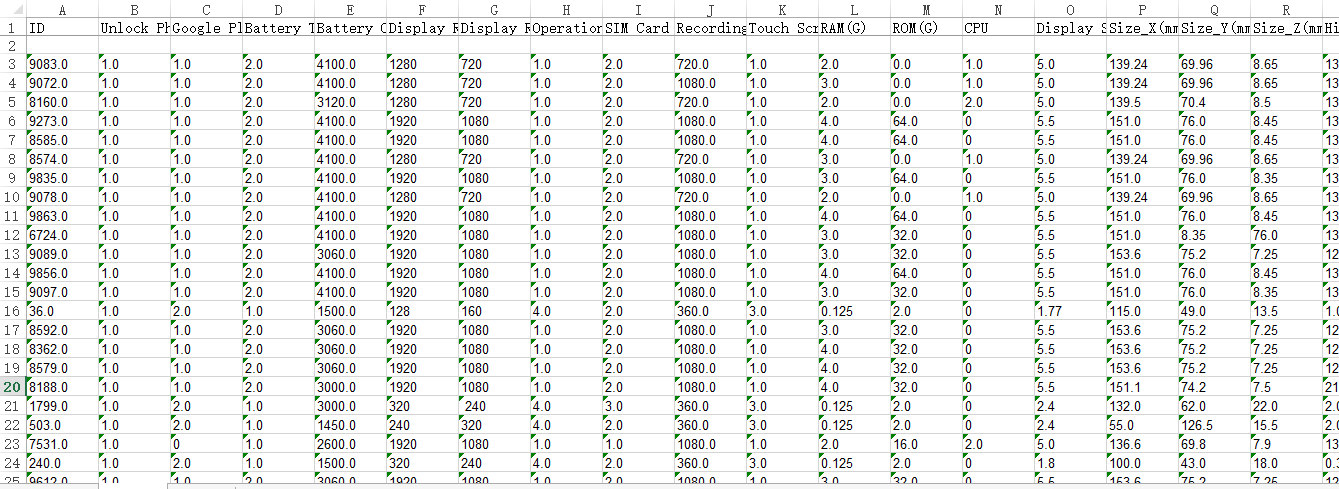
* 1. **BOOST Algorithm**

We utilizes BOOST Algorithm to obtain the average value of each methods of the samples. The formula is as follows

In the formula, denotes the original result of principle component analysis. denotes the result of Bayes distinction. 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.

1. **Application**

We use the data which has exactly one zero as the test data sets and conduct the process in part 5 to obtain the final result. Particularly, we use the data after Principal Component Analysis for Principal Component Regression and the data after Entropy of Information for Bayes distinction and BP neural network Fitting. The following figure is a part of the final result.



1. **Conclusion**
   1. **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.

* 1. **Strength and Weakness**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

* 1. **Conclusion**

blablablablablablablablablablablablablablablablablablablablablablablablablablabla

1. **Reference**

Blablablablablablablablablablablablablablablablablablablablablablablablablablabla

1. **Acknowledgement**

Blablablablablablablablablablablablablablablablablablablablablablablablablablabla

1. **Appendix**

Blablablablablablablablablablablablablablablablablablablablablablablablablablabla

**此页开始为致谢页**

**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。**

**此页为学术诚信声明**

**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**