**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 below:

where E(X) represents the set of incidents, X, taken into consideration (in the formula above the total number of incidents is n) and 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 are the top factors 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 contributing or crucial factors for the sale of the phone including Google play, battery type, brand, RAM, ROM, dual camera, front camera, display size, etc. Then, types of data representing the actual sale of cellphones are regarded as the bases for calculating the information gain. Instead of choosing the actual sales volume, we considered 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 the formula above 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 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 as following:

|  |  |  |
| --- | --- | --- |
|  | 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 consists 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 as following:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 represents the line in the table of the forty groups (as distinguished by double cross lines), represents the column in the table, and represents the number in the unit of the line and column. Thus, in the unit , the number 110 represents 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 equals 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 the formula presented at the beginning of this part. The information entropy of ROM in each line is as the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | Information entropy E(ROM i) |
| 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 A1,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 as the following:

|  |  |  |
| --- | --- | --- |
| Information entropy E(i) | Possibility P(i) | 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 the 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:

IGain(Category Click Rate, ROM) = E(global) -ΣE(i)\*P(i)

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 as the following:

|  |  |  |
| --- | --- | --- |
|  | 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 Cnt(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 extracted, it is obvious that the data in Search Cnt 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 in to 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 part are similar as that for the independent variables in group one. The following table presents the grouping of Search Cnt after the data division.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 Cnt related to the Category Click Rate is as the following:

|  |  |  |
| --- | --- | --- |
|  | Sum of the products | IGain |
| Search Cnt | 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 the two tables following:

table: information gain of each individual variables related to the Category Click Rate

|  |  |
| --- | --- |
| Comment Count | 0.732792417 |
| Good Comment Count | 0.680453664 |
| Search Cnt | 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: information gain of each individual variables related to the Category Convert Rate

|  |  |
| --- | --- |
| Comment Count | 0.950131659 |
| Good Comment Count | 0.910616696 |
| Search Cnt | 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 |

The higher the information gain of the individual variable indicates the higher the importance of that factor contributing to the sale of the product. 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.