Information entropy:

The information gain algorithm based on the information entropy theory can effectively help us judge the relationship between dependent and independent variables and sift out the most important ones. Information entropy is used to reflect the complexity of the information being processed, and the meaning of information entropy is first defined in Shannon's Principal of General Mathematics, in which he implied that it couldbe used to quantify the specific value of data. The theory of information entropy nowadays is applied in risk management, optimization of model and information safety, possessing high social value. Information gain, on the other hand, reflects how much more information we can get with the appearance of an additional independent variable. For instance, in order to calculate the information gain of ROM, we should first gain the information of sales volume itself, which is represented by the global entropy of Category Click and Conversion Rate. Next, we divide ROM into 5 groups and calculate the information entropy of each group, then multiply them respectively with the possibility in order to get the total information entropy of ROM relates to Category Click Rate. The difference of two values above defines ROM's information gain. Apply the algorithm to each independent variables and thus we can rank the most important ones. In the modeling process later, top-ranking variables in the chart are taken into consideration.

The key issues concerning reducing the independent variables are how to determine the relationship between the new variables and original ones, and how to define the importance of each new variables. To solve these two problems, we apply the Principal Component Analysis. The essence of PCA is singular value decomposition, and the core technology of it is reducing the dimension of data and find the weight of original variables to each new variables. We standardize the original variables, set up the correlation coefficient matrix and solve the eigenvalues and eigenvectors. For the first problem, the eigenvectors are extracted as the coefficients of the new expression which concerns the new PCs in terms of original variables. For the second problem, the importance is defined by the contribution rates of eigenvalues. We extract first several large eigenvalues until the sum of them reaches 80%. In this way, the new variables are formed and the important independent variables can be extracted.

We first use Weight Determination Technique, imitating the idea of AHP, but with only two layers, to determine the weight between each two groups. We also apply Logistic Regression and KNN Algorithm. We add a transfer function on Linear Regression to determine the output result. But we discover that the correlation coefficients and accuracy are relatively lower, which means we need to use other algorithms to optimize the result.

As a regression based on the PCA, Principal Component Regression aims to reduce noisy data and highlight the relationship between important variables. The model we apply is logistic model. It shows that the dependent variables lie in the range [0,1), which means the model is defined and valid. Then, we calculate the correlation coefficients, and the result we get from PCR is higher than those from pure logistic regression, indicating the model we construct is more effective. However, with both discrete and continuous variables presented, the margin of error will inevitably increase when applying this model. Therefore, other optimized models will be taken into consideration.

conclusion:

Through the data extraction, modeling and optimization process, we have gained different conclusions. First, as for the data extraction, we synthesize the results mainly from information entropy algorithm and PCA, and the ranking is shown above. We can see that Comment Count, Good Comment, and Search Count rank respectively the top three, which indicates that they are more important to the sales volume. Then we apply the variables we extracted to the modeling and optimization and yield qualitative results based on the weight determining technique. They are shown directly with the help of straightforward diagrams. Take Display Resolution for instance; a higher group number indicates higher sales volume, so we can see that phones in Display Resolution Group 3, which represent middle display resolution level, sell the best. As for the quantitative conclusion, which concerns phone's specific traits, we discover that phones with a middle level of battery capacity, gold or white color and high camera resolution is the best model predicted. During the compilation of data we also discover that RAM, ROM and CPU should be the top choices when improving the phone's overall quality. Besides, the public isnot sensitive to the change in Display and Camera Resolution, so manufacturers should give more thought to lowering the price when these two factors are taken into consideration. Furthermore, the whole modeling process and methods we apply possess several advantages as well as pragmatic value. Besides helping the manufacturers understand the demands from the public more precisely and giving valuable guidance on the improvement of specific cell phone traits by both qualitative and quantitative methods, our model can also ground-breakingly predict the cell phone sales with given traits, aiming to assistant the manufacturers with insightful analysis and help them gain maximum benefits. Therefore, our model possesses high application value in the field of economy.