Customer Segmentation and Marketing Strategy Advice from a Statistical Perspective

Group9: Yang Huakang, Yang Siying, Ji Wenyu

1 Introduction

Customer analysis is a vital component of any successful business strategy. By leveraging advanced data analysis techniques, we can gain a deeper understanding of our customers and tailor our products and operations to better meet their needs and preferences. In our project, we employ a range of analytical methods, including EDA, clustering, multiple linear regression, and machine learning for regression and classification. Through this analysis, we can gain valuable insights into the characteristics and spending behaviors of our customers. This information enables us to make informed decisions about marketing strategies and improve the overall success of the business. Our FlowChart in Appendix 1 presents our logic in a clear way.

2 Explorative Data Analysis

After cleaning the data, we conducted Exploratory Data Analysis (EDA) to gain a better understanding of it. The bubble plot below shows the relationship between Income, Spendings, Age, and Education. We can see that as Income increases, Spendings also increase rapidly. This suggests that Income may be an important variable when analyzing customers' spending behaviors. We could also consider transforming the data so that Spending and Income have a linear relationship for use in a multiple linear regression model. Additionally, the plot shows that people with lower levels of education generally have lower income and spendings.

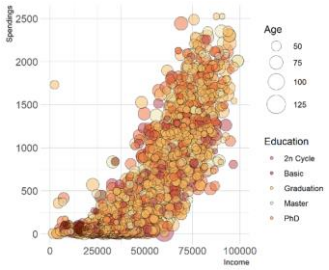


Figure1. Bubble Plot for EDA

From the correlation plot below, we can see that the variable Spendings has a strong relationship with the number of purchases made online, in-store, and through catalogs. Additionally, we can see the correlation between other variables, such as Income and Purchases. This correlation matrix provides insight into which variables we should consider including in our linear regression model.

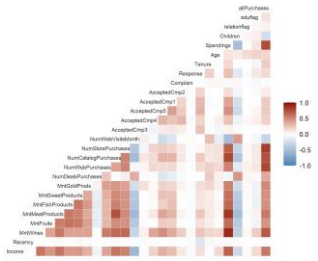


Figure2. Correlation

3 Hierarchical Clustering

In this section, we aim to use hierarchical clustering to gain a general understanding of the customer persona and develop a basic strategy for identifying our target customer group. After experimenting with different linkage methods, including Ward.D, Complete, Average, and Single linkage, we found that the Ward.D method was the most suitable. Additionally, by identifying the elbow point, we determined that it was appropriate to choose three as the number of clusters.

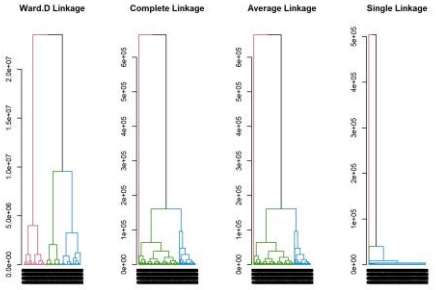


Figure3. Different Linkage Methods

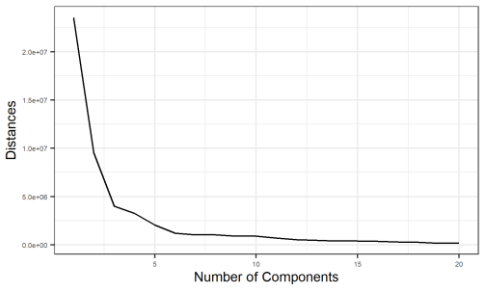


Figure4. Elbow finding

After performing the clustering, we assigned class labels to each data point and analyzed the differences and characteristics of each consumer group. The plot below shows the relationship between Income and Spending, with the clusters indicated on the graph. We can see that Class Two has the highest income and spending levels.

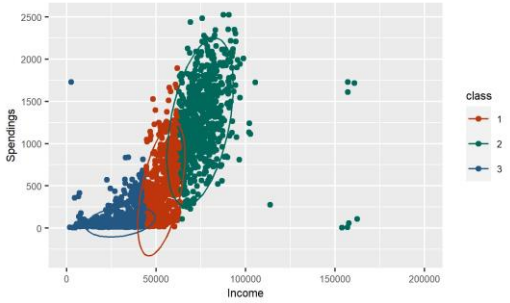


Figure5. Distribution of Clusters on Spendings and Income

In addition to Income and Spending, we also created boxplots for the other non-factor variables to visualize the differences between clusters. Below are the plots for the variables that show significant differences between clusters.

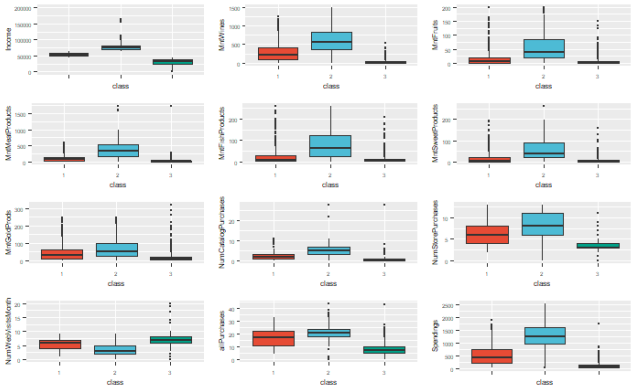


Figure6. Boxplot of variables with clusters

We can also see differences in the number of times a customer accepted promotions and the number of children they have.

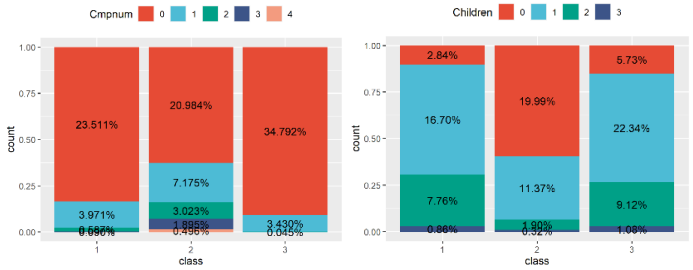


Figure7. Number of accepted promotions and children

Based on the plots above, we can summarize the characteristics of the three customer groups. The groups can be described as follows: Group 2 - The Generous People, Group 3 - The Low Income People, and Group 1 - The Middle Group.

As an example, we will discuss the characteristics of Group2 & Group3.

Here are the characteristics of Group 2 – The Generous People:

(1) Highest Income(70,000 to 80,000) with several outliers of very high income;(2) Highest Spendings, far surpassing that of the other two groups;(3) Generous buyer of all types of products, but with significant variance within group. This means that individuals in Group 2 have the financial freedom to choose whether to spend a lot on products. As a result, the actual number of purchases can vary significantly from person to person within this group;(4) Most purchases are made through catalogs and in-store, with minimal use of the web for shopping;(5) High likelihood of accepting promotional campaigns;(6) Most of them have no children.

Here are the characteristics of Group 3 – The Low Income People

(1) Lowest Income, mostly around 25,000 to 37,500, also has outliers with very low income;(2) Lowest Spendings, very low, the average of the Generous Group is about 12 times of this group, but some outliers can reach the income level of the Middle Group;(3) Smallest group of buyer of every type of products, i.e. they have the lowest num of purchases of each product. However, their purchase behavior also varies a lot from person to person, which is especially obvious when purchasing gold, the data of which presents many outliers showing that some of this group spend a lot in gold;(4) They have lowest Catalog and Store Purchases, while they visit the web a lot, and is the most frequent visitor among the three groups;(5) They are very unlikely to accept promotion campaigns;(6) Most of them have one or two children. The characteristics of the Middle Group lie between those of the other two groups, so further discussion of Middle Group is not necessary here.

In summary, we can conclude that consumers with high incomes, fewer children, and a preference for catalog and in-store purchasing channels have greater spending power and fall into the Generous Group. They also tend to readily accept promotional campaigns. Companies can target individuals with these characteristics when advertising and promoting their products.

4 Multiple Linear Regression

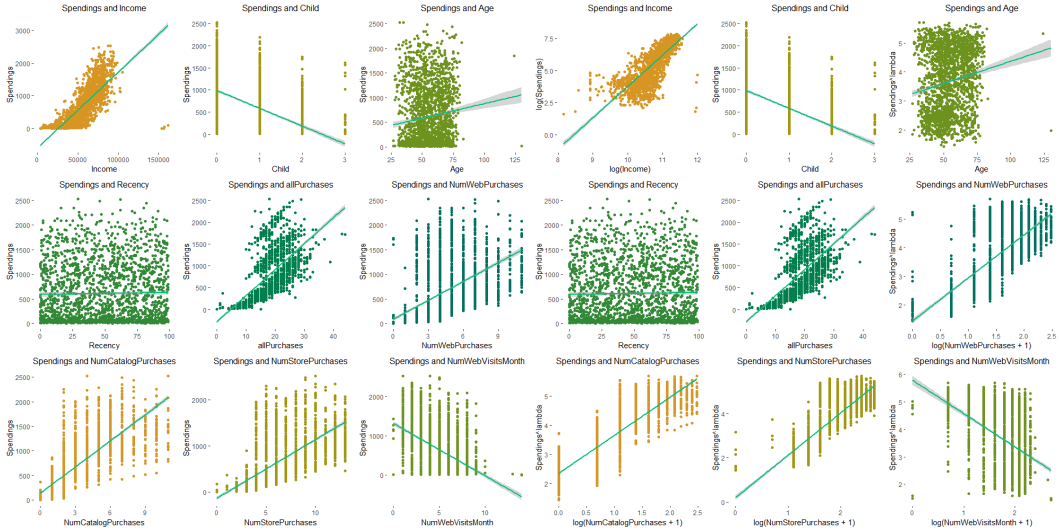
After a basic understanding of customer characteristics is obtained from the sections above, this section focuses on Multiple Linear Regression.

4.1 Relationship between total consumption and the variables

Firstly scatter plots were done between total consumption and each variable separately to observe the basic relationship between the variables. Then regression analysis was done between total consumption and each variable separately to obtain the final regression results through the following three steps:

1. remove the outliers according to the diagnostic diagram
2. transform the independent and dependent variables according to the diagnostic and scatter plots
3. determine whether to use weighted least squares estimation based on the diagnostic plots

A new scatterplot is made based on the transformations in step 2. The two scatterplots are shown below:



(a) (b)

Figure8. Scatterplot before(a) and after(b) Transformation

The regression results show that total expenditure has a strong relationship with log(Income), Child, Age, log(NumWebPurchases+1), log(NumCatalogPurchases+1), log(NumStore-Purchases+1), log(NumWebVisitsMonth+ 1) These seven variables have a strong relationship with Recency and a weak relationship with Recency, so the Recency variable is discarded.

The logarithm number is taken to obtain a more linear relationship, and the variable +1 is taken to prevent a base 0 situation.

The variables chosen and the way each variable is transformed will be obtained from each regression equation, and then a multiple linear regression will be done between total consumption and the transformed variables. The analysis of the diagnostic plots showed that the regression was better for the portion of total expenditure above 500, and the regression was only done for the portion of total expenditure above 500, yielding the following diagnostic plots:

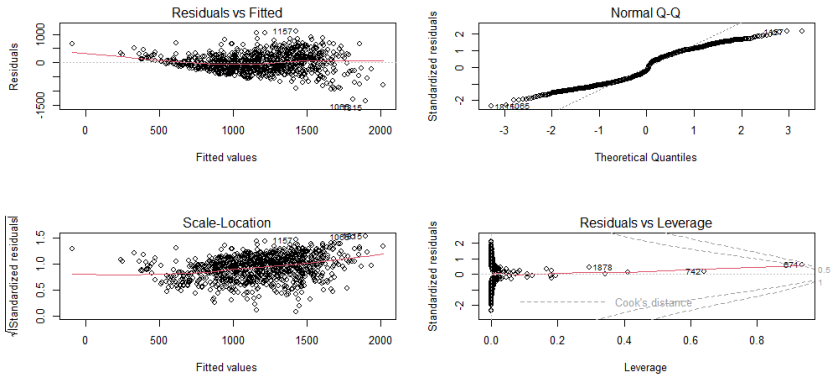


Figure9. Diagnostic Plots

The regression results are shown in the table below:

	Dependent variable:
	Spendings
log(Income)	1,352.904*** (23.439)
Child	-230.957*** (7.443)
Age	-0.042 (0.221)
log(NumWebPurchases + 1)	-80.695*** (11.616)
log(NumCatalogPurchases + 1)	245.172*** (8.173)
log(NumStorePurchases + 1)	-0.759 (8.504)
log(NumWebVisitsMonth + 1)	347.195*** (10.837)
Constant	-14,534.930*** (266.931)
Observations	980
R^2	0.911
Adjusted R^2	0.911
Residual Std. Error	15.641 (df = 972)
F Statistic	1,424.350*** (df = 7; 972)
Note:	*p**p***p<0.01

Table1. Regression Results

The regression results have three characteristics:

1. The logarithm of income has a greater impact on consumption expenditure. This may be because consumers' marginal propensity to consume tends to diminish, i.e. as consumers' income increases, consumption expenditure on consumption decreases, one of the three fundamental psychological laws proposed by Keynes, which we have successfully verified here.
2. The greater the number of children, the lower the consumption expenditure. This may be due to the higher share of other costs of raising children, which has a crowding out effect on the type of consumption expenditure within this dataset.
3. The impact on consumer spending varies across channels and this point is important for businesses to attract consumers.

4.2 The relationship between consumption of various types of goods and distribution channels

The next step is to analyze the impact of each channel on the consumption of each type of goods. Using the four variables NumWebPurchases, NumCatalogPurchases, NumStorePurchases, and NumWebVisitsMonth as independent variables, these four independent variables are transformed according to the previous transformations into log(NumWebPurchases+1),log(NumCatalogPurchases+1),log(NumStorePurchases+1), log(NumWebVisitsMonth+1). Multiple regressions were done on these transformed four variables for each category of goods consumption and the regression results were as follows:

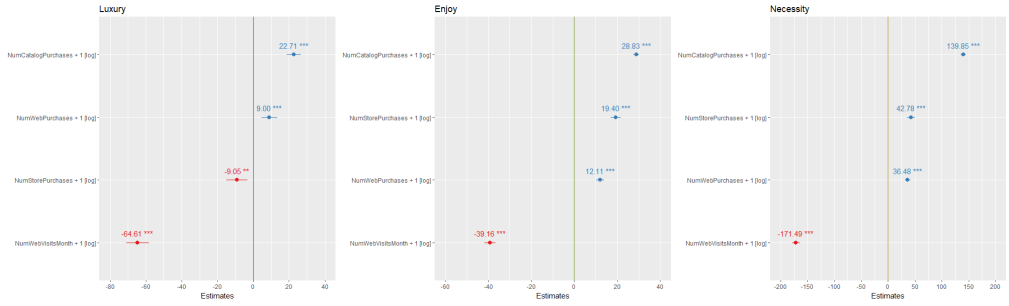


Figure10. Relationship between consumption and sales channels

The regression results provide some suggestions for merchants in terms of sales channels:

1. The two dominant channels for both recreational and essential goods are promotional and offline channels. Merchants in these two categories need to strengthen their offline channel operations and promotions, which have a positive effect on boosting sales.
2. For luxury goods, the online channel has a positive effect on sales, while the offline channel has a negative effect on sales, so merchants in these categories need to strengthen their online channel operations.
3. For all three categories, there is no positive relationship between web hits and sales, which means that merchants are still lacking in web design and promotion, and it is recommended that they invest more in this area of marketing.

5 Machine Learning Models for Regression

In this section, we use two machine learning models to make predictions about consumption and identify the variables that are most important to the predictions. The goal is to make recommendations on the company's market strategy. We split the data into training and testing sets. We choose XGBoosting algorithm and Randomforest algorithm, and obtain the results through steps such as data type conversion, training the model on the training set, and testing the model on the test set.

After running multiple models for prediction, we choose the Randomforest model with the smallest mean square error (MSE) to analyze the importance of each variable to the prediction. The MSE of each model is shown in the table.

Model	MSE
Randomforest	280.78
XGBoosting	299.96

Table2. MSE of two models

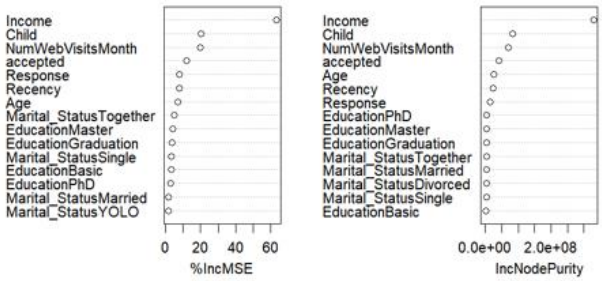


Figure11. Variable Importance

As the figure shows, Income is the most important variable for the forecast, which also agrees with our previous analysis. In addition, Child, NumWebVisitsMonth and accpted is also important factor. And educational background and marital status have basically no effect on the regression prediction.

This means: when a company reaches out to new consumers and wants to predict their spending, it should focus on their income, number of children, and website views, regardless of their education and marital status. Combined with the previous analysis, we can know that the higher the income, the smaller the number of child, the smaller the number of views, and the higher the consumption may be.

6 Classification of Promotion Response

We are interested in whether one person will eventually accept the promotion campaign or not. So we use classification models to predict one's behavior. In this case, we perform both Logistic Regression and Decision Tree Classification.

First of all, let's determine the dependent variable, which represents whether someone will eventually accept the promotion campaign or not. If the consumer accepts at least one promotion campaign, then we denote the dependent variable as "yes", while if he/she does not accept any campaign, we denote the variable as "no".

6.1 Logistic Regression

We transform the factor variables into dummy variables and use the train dataset to build the Logistic Regression model. The results are listed in Appendix2. Many variables are not significant, so we use stepwise search with the AIC criterion to perform feature selection. The results are shown below, 9 variables and no interactions is chosen by AIC.

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.4500104  0.3187013  -4.550 5.37e-06 ***
Recency          -0.0110439  0.0023053  -4.791 1.66e-06 ***
MntWines          0.0035272  0.0003124  11.289  < 2e-16 ***
MntFruits        -0.0030782  0.0021458  -1.435 0.151423
MntMeatProducts   0.0010245  0.0004405   2.326 0.020025 *
MntSweetProducts  0.0037100  0.0019891   1.865 0.062164 .
NumCatalogPurchases  0.0582661  0.0341820   1.705 0.088272 .
NumStorePurchases -0.1830207  0.0296081  -6.181 6.35e-10 ***
NumWebVisitsMonth  0.1579509  0.0344331   4.587 4.49e-06 ***
Child            -0.3719780  0.1127194  -3.300 0.000967 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure11. Logistic Regression Result after Stepwise Search

The variables Child, NumStorePurchases, and NumWebVisitsMonth have a significant impact on whether an individual will accept a promotional campaign. Based on this, we can target our promotions towards individuals who frequently visit the web, make more catalog purchases, make fewer in-store purchases, and have fewer children.

An AUC of 0.78 is obtained, the confusion matrix and ROC Curve can be seen in the figure.

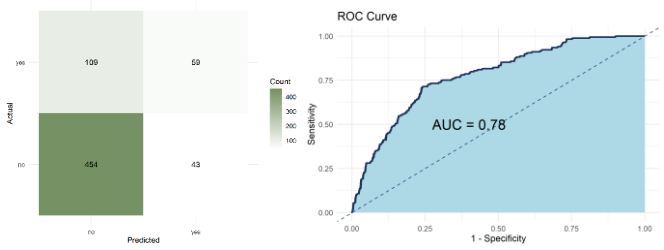


Figure12. Model Performance

6.2 Decision Tree

A decision tree model is also established to analyze the problem. The growth of the tree and the resulting leaf nodes show the most useful variables and how they were split. The model achieved an accuracy of 78.6%.

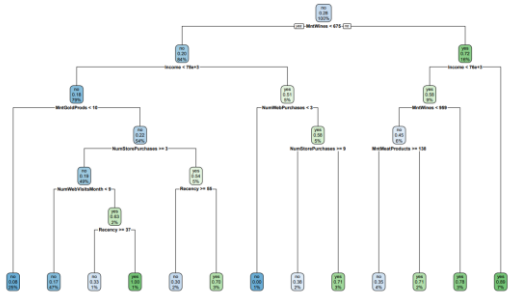


Figure13. Decision Tree

In summary, based on the analysis in this section, the following promotional recommendations can be made for companies: Promotional activities should target consumers with fewer children who prefer web and catalog channels for shopping.

7 Conclusion

First, we use hierarchical clustering to Identify the characteristics of different customer groups, divide customers into three groups - generous people, low-income people, Middle Group, and summarize the characteristics of the three groups of people, which will help the company improve marketing and promotional strategies.

Then we use the method of multiple linear regression to conduct multiple regression analysis on consumption to study the relationship between each variable and consumption. We found that the logarithm of income has a significant impact on consumption, which is consistent with the classic psychological law; the larger the number of children, the lower the consumption; different commodities depend on different sales channels, online and offline sales should be more targeted.

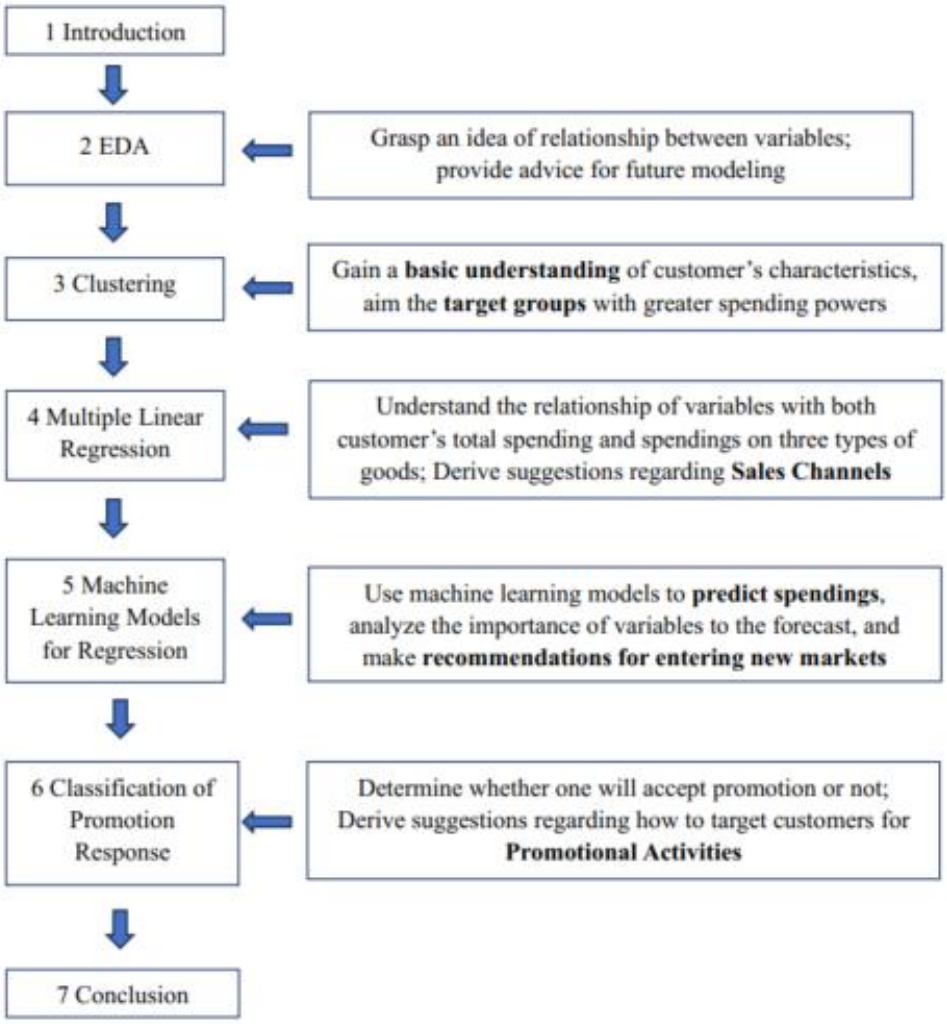
Next, we use the machine learning model to predict the consumption, get a random forest model with better effect, and discuss the importance of each variable to the regression.

Finally, we focus on how to improve the company's promotional activities. By training the classification model to find the characteristics of people who are easy to accept promotional activities, and then make targeted recommendations to the company.

In this way, we gain a full understanding of data, and also provide comprehensive suggestions for companies to understand their customers.

Appendix

1. Flow Chart of our Project



2. Results of Logistic Regression before AIC

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.747e+00	6.211e-01	-2.813	0.0049	**
EducationBasic	-1.811e-02	5.204e-01	-0.035	0.9722	
EducationGraduation	7.982e-03	2.517e-01	0.032	0.9747	
EducationMaster	-1.425e-01	2.901e-01	-0.491	0.6233	
EducationPhD	-1.646e-01	2.874e-01	-0.573	0.5667	
Income	8.871e-06	6.345e-06	1.398	0.1621	
Recency	-1.119e-02	2.326e-03	-4.809	1.52e-06	***
MntWines	3.453e-03	3.472e-04	9.946	< 2e-16	***
MntFruits	-3.475e-03	2.244e-03	-1.549	0.1214	
MntMeatProducts	9.700e-04	4.734e-04	2.049	0.0404	*
MntFishProducts	-1.725e-03	1.636e-03	-1.055	0.2917	
MntSweetProducts	3.536e-03	2.107e-03	1.679	0.0932	.
MntGoldProds	1.301e-03	1.460e-03	0.891	0.3728	
NumDealsPurchases	-5.391e-02	4.323e-02	-1.247	0.2124	
NumWebPurchases	-3.267e-03	3.292e-02	-0.099	0.9209	
NumCatalogPurchases	5.669e-02	3.675e-02	1.543	0.1229	
NumStorePurchases	-1.800e-01	3.114e-02	-5.780	7.49e-09	***
NumWebVisitsMonth	1.994e-01	4.634e-02	4.303	1.69e-05	***
MarriageMarried	-1.922e-01	2.305e-01	-0.834	0.4045	
MarriageSingle	1.644e-01	2.501e-01	0.657	0.5109	
MarriageTogether	-2.164e-01	2.428e-01	-0.891	0.3728	
MarriageWidow	2.191e-01	3.884e-01	0.564	0.5726	
Tenure	-1.690e-04	3.726e-04	-0.454	0.6502	
Age	-6.862e-04	5.755e-03	-0.119	0.9051	
Child	-3.346e-01	1.317e-01	-2.540	0.0111	*

Figure1. Results of Logistic Regression before AIC

2. Division of Labor

Yang Huakang: 5 Machine Learning Models for Regression; 7 Conclusion; report drafting
Ji Wenyu: 4 Multiple Linear Regression; report drafting
Yang Siying: 1 Introduction; 2 Explorative Data Analysis; 3 Hierarchical Clustering; 6 Classification of Promotion Response; report drafting and combination