Customer Segmentation and Marketing Strategy Advice from a Statistical Perspective

Siying Yang, Wenyu Ji and Huakang Yang

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 are able to 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.

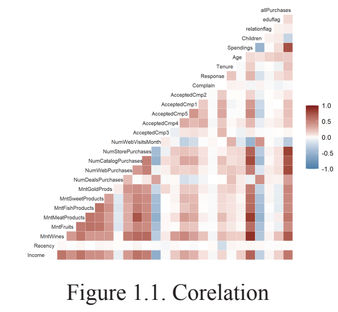

Figure 1.1. Corelation

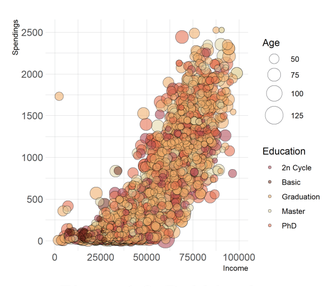

Figure 1.2. Bubble plot

2. Hierarchical Clustering

We use hierarchical clustering to segment consumers into distinct groups and analyze the characteristics of each group to better understand our customer base and tailor our strategies accordingly.

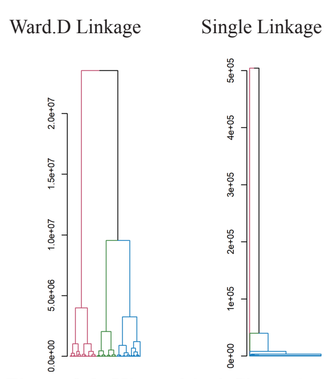

Figure 2.2. Hierarchical Clustering

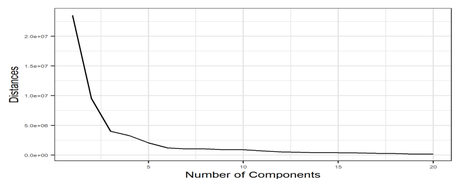

Figure 2.1. Elbow finding in HCLustering

Figure 2.3. Income and spending with clusters

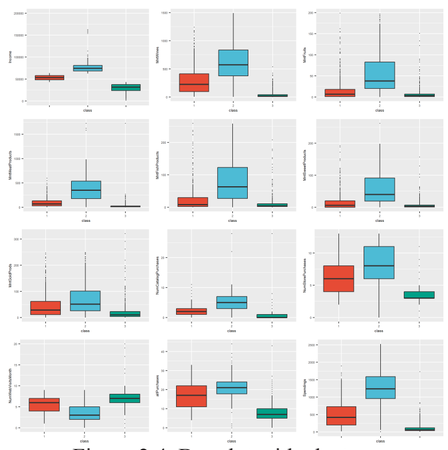

Figure 2.4. Boxplot with clusters

As an example, we will discuss the characteristics of Group 2 – The Generous People:

- Highest Income(70,000 to 80,000) with several outliers;
- Highest Spendings;
- Generous buyer with significant variance within group;
- Most purchases are made through catalogs and in-store, with minimal use of the web for shopping;
- High likelihood of accepting promotional campaigns;
- Most of them have no children.

3. Multiple Linear Regression

3.1 Relationship between total consumption and each variable

In the regression analysis section, we explore the relationship between total consumption and each individual variable. We conduct separate regression analyses to produce nine scatter plots. To obtain the final regression results, we follow a three-step process:

1. Removing outliers
2. Transforming the independent and dependent variables
3. Using weighted least squares estimation

By applying the final regression results, we generate a new scatter plot. Fig.3.1 shows the scatter plots before and after the process.

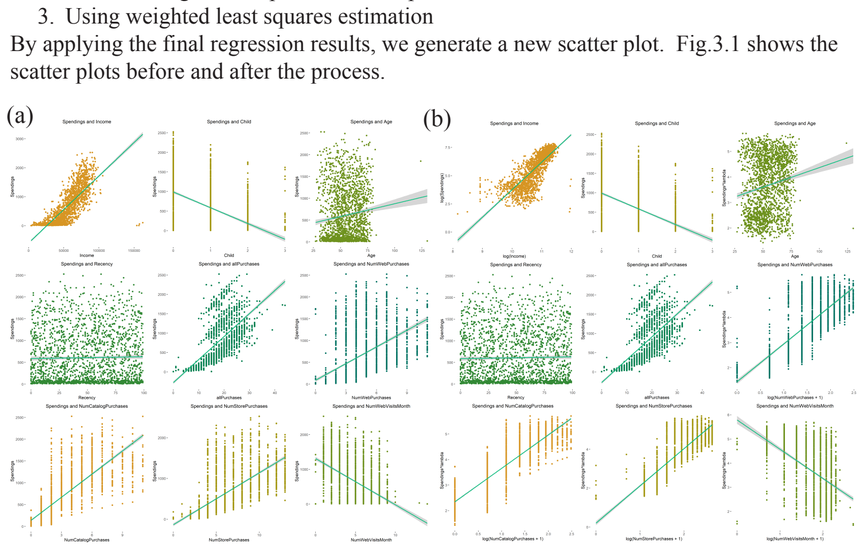

Figure 3.1. Regression (a) before transformation (b) after transformation

It can be observed that total expenditure has a strong relationship with seven variables: log(Income), Child, Age, log(NumWebPurchases+1), log(NumCatalogPurchases+1), log(NumStorePurchases+1), and log(NumWebVisitsMonth+1). However, the relationship with Recency is weak, so this variable was discarded.

By combining the regression equations into a single multiple linear regression equation and analyzing the diagnostic plots, it is revealed that the regression performed better for total expenditure values above 500.

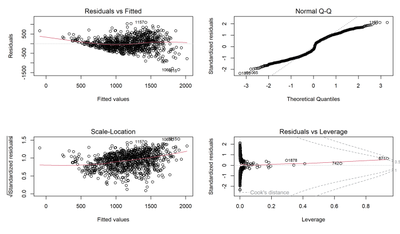

Figure 3.2. Diagnostic plots

The regression results have three characteristics:

1. The logarithm of income has a significant impact on consumption expenditure.
2. The more children a consumer has, the lower their consumption expenditure.
3. The effect on consumer spending varies across different channels.

Table 3.1. Regression results

num	term	estimate	std.error	statistic	p-value
1	(Intercept)	-1.444008e+04	225.0652784	-64.1595344	0.000000e+00
2	Age	-1.136748e-01	0.2180082	-0.5214245	6.021903e-01
3	Child	-2.327788e+02	7.0949730	-32.8089750	2.220772e-159
4	log(Income)	1.346189e+03	20.8678385	64.5102473	0.000000e+00
5	log(NumCatalogPurchases + 1)	2.430744e+02	8.0103886	30.3448923	1.031154e-142
6	log(NumStorePurchases + 1)	-6.708839e+00	8.3055620	-0.8077526	4.194312e-01
7	log(NumWebPurchases + 1)	-7.946649e+01	11.4440853	-6.9438920	6.983111e-12
8	log(NumWebVisitsMonth + 1)	3.468809e+02	10.9429728	31.6989642	7.108005e-152

3.2 Relationship between consumption and distribution channels for each category of goods

The regression results offer valuable insights for merchants regarding their sales channels:

1. For both enjoy and necessary goods, merchants in these categories should focus on strengthening their offline operations and promotions to boost sales.
2. For luxury goods, merchants in this category should focus on enhancing their online operations.
3. Across all three categories, merchants may need to improve their web design and promotion efforts and invest more in this area of marketing.

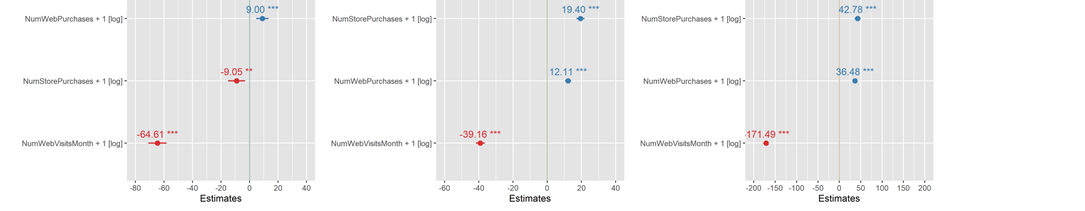

Figure 3.3. Relationship between consumption and sales channels

4. Machine learning for regression

Our objective is to utilize a machine learning model for consumption prediction and identify the most influential variables for regression prediction, in order to guide the company's marketing strategy. Consequently, we opt for the XGBoost algorithm for modeling. During the model construction process, we conduct feature selection, transform categorical variables into numerical ones, train the model using the input data, and tune the parameters to obtain the final model.

Ultimately, we discover that Income, Child, Recency, and NumWebVisitsMonth are the most significant variables for regression prediction, with Income being particularly influential. On the other hand, educational background and marital status have minimal impact on the prediction. Hence, when approaching consumers in new markets, the company should prioritize the following aspects:

- Consumers' income
- The number of children in consumers' families

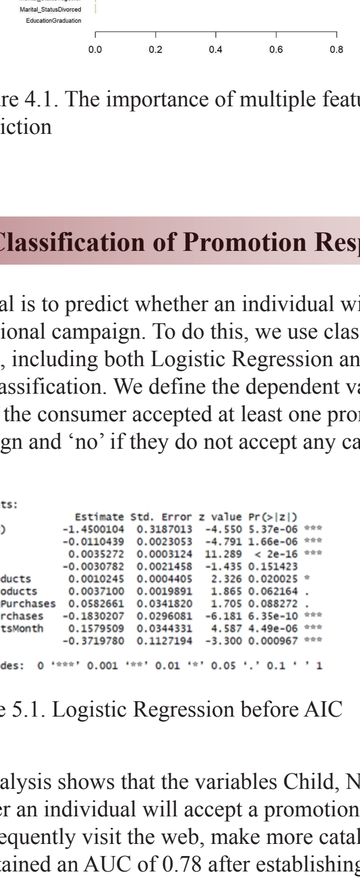

Figure 4.1. The importance of multiple features for prediction

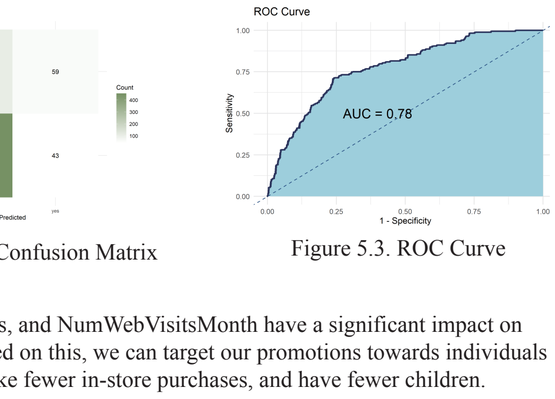

Figure 4.2. Visualization of XGBoost Decision Tree

5. Classification of Promotion Response

Our goal is to predict whether an individual will accept a promotional campaign. To do this, we use classification models, including both Logistic Regression and Decision Tree classification. We define the dependent variable as 'yes' if the consumer accepted at least one promotional campaign and 'no' if they do not accept any campaigns.

5.1 Logistic Regression

We send variables and interaction items into the model, and use AIC to do the feature selection. As is shown in the figure, 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
```

Figure 5.1. Logistic Regression before AIC

Figure 5.2. Confusion Matrix

Figure 5.3. ROC Curve

Our analysis shows that 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. We obtained an AUC of 0.78 after establishing this model.

5.2 Decision Tree

We also use a decision tree model 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%.

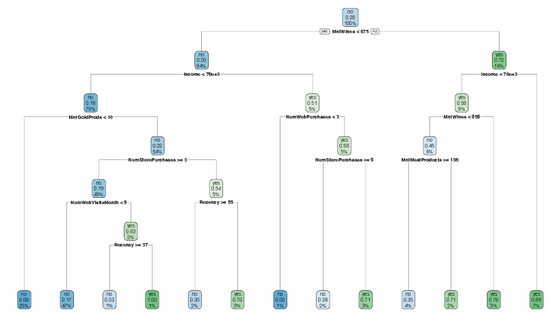

Figure 5.4. Decision Tree

6. Conclusion

Based on our analysis, we have drawn the following conclusions and recommendations:

1. Consumer characteristics: Consumers with high incomes, fewer children, and a preference for catalog and store purchasing channels have greater spending power.
2. Channel recommendations: All types of merchants should focus on enhancing their online marketing capabilities.
3. Promotional recommendations: Promotional activities should target consumers with fewer children who prefer web and catalog channels.