USING DATA MODELING TO ENHANCE CUSTOMER SEGMENTATION

A CASE STUDY ON LOYALTY PROGRAM DATA ANALYSIS FOR JCD FASHION RETAIL COMPANY

APRIL 18TH, 2022

Submitted to Kate Ashley
MISM 6206 – Modeling for Business

Submitted by:
Team America's Next Top Modelers
Moeez Abbas
Laura Seeger
Dhanush Srinath
Neeti Oberoi

TABLE OF CONTENTS

Contents

Abstract: 3

Introduction 3

Data and Pre-processing 3

Methodology 4

Customer Segmentation through Clustering 4

Predictive Analysis through Logistic Regression: 4

Correlation Matrix 5

Conclusion and Business Insights 5

Appendix 6

Plot 1 6

Plot 2 6

Plot 3 7

Plot 4 7

Plot 5 8

Plot 6 8

Plot 7 9

Plot 8 9

Plot 9 10

Plot 10 10

Table 1 – Demographics of Clusters 11

Table 2 – Shopping Behavior of Clusters 11

Figure 1 – Summary Table for Base Model 12

Figure 2 – Confusion Matrix for Base Model 12

Figure 3 – Summary Table for Model 1 13

Figure 4 – Confusion Matrix for Model 1 13

Figure 5 – Summary Table for Model 2 14

Figure 6 – Confusion Matrix for Model 2 14

Figure 7 – Correlation Plot 15

Abstract:

This report focuses on the application of customer segmentation through data modeling techniques for JCD, a fashion retail company. The objective of the analysis was to gain insights into consumer behavior and factors driving loyalty to identify customer segments and build a predictive model that could be used for targeting. The report describes the steps taken to explore the loyalty program data, including data cleaning, exploratory data analysis, clustering analysis, logistic regression, and linear regression. The report presents the findings of the analysis, including the characteristics of the customer segments and the factors that drive customer loyalty. The report concludes with recommendations for JCD to improve its performance and competitiveness by better understanding its customers and targeting its marketing efforts more effectively.

Introduction

Customer loyalty is crucial for businesses to stay competitive and thrive. Customer segmentation helps businesses understand their customers and market their products effectively by grouping customers based on their characteristics, behavior, and other factors. However, the process can be challenging due to the vast number of customers and their different demographics. Companies are turning to data modeling to overcome this challenge, using historical data to build predictive models for customer segmentation through supervised and unsupervised methods.

In this project, we explore how customer segmentation and data modeling can help JC Dollars (JCD), a fashion retail company, address a significant business challenge. JCD has experienced a sharp decline in earnings and growth over the past few quarters, making it increasingly challenging for the company to compete based on price, product, and location. To address this, JCD's executives have invested more than $10 million in a loyalty program to increase customer loyalty, believing that it will lead to more purchases and increased revenue. By analyzing data generated through the loyalty program, JCD hopes to understand customer purchase behavior, determine key factors that drive loyalty, identify customer segments, and build a predictive model for targeting.

The report aimed to assist JCD in solving a business challenge by using customer segmentation. The team used modeling techniques to analyze loyalty program data, identify customer segments, and build a predictive model for targeting. The study's significance lies in its ability to help JCD improve customer loyalty and business results through data-driven decision making. The analysis of loyalty program data provided valuable insights into customer behavior, and the predictive model helps JCD develop targeted marketing strategies. The study highlights the importance of leveraging analytics to improve business performance and remain competitive, and the team recommended creating a broader analytics strategy to effectively compete in the future.

Data and Pre-processing

To tackle the business challenge faced by JC Dollars (JCD), our team focused on analyzing the customer and transaction data provided by the company. The dataset encompassed information on over 102,000 customers, featuring numerous characteristics such as age, profession, income, and other pertinent data points. To effectively analyze and derive insights from this data, it was crucial to pre-process and clean it for use in models.

The first step in the analysis was to explore the given data, with the team generating descriptive statistics and visualizations to begin understanding the differences between customers who did and did not respond to the promotions. The visualizations could not produce meaningful results hence they were omitted from the final report (see the details on visualizations in the EDA code attached).

During the exploratory data analysis (EDA) phase, the team identified that the customer ID was the unique identifier for each customer, and over 5% of the customer income values were missing. It was evident that the customer ID would not be included in our analysis. However, we faced a decision regarding the treatment of customer income. Ultimately, the team decided to exclude customer income from the models since imputing the missing values would lead to unreliable results during the clustering process. Nevertheless, we still analyzed the mean income of both the clusters, after the clusters were identified, as it could still offer valuable insights on the purchasing power of each customer segment. Moreover, to carry out the clustering analysis, we dropped all categorical variables and proceeded with the analysis using the numerical variables.

Methodology

Customer Segmentation through Clustering

Our first objective was to perform a customer segmentation analysis using clustering methodology. We employed the k-means method for clustering. Prior to implementing the clustering model, we had to determine the appropriate number of clusters. Consequently, we generated an elbow plot, but the outcome was not entirely conclusive. It indicated a possibility of four clusters, which did not, however, fit the data optimally (see plot 1). We then used to analyze internal validation metrics like Silhouette Scores (see plot 2) and Calinski Harabasz Scores (see plot 3) for a range of number of clusters. Both the scores peaked at 2 clusters. After conducting several iterations, we decided to select two clusters, as they provided the clearest differentiation and segmentation between the customers. Subsequently, we removed all categorical variables and ran the clustering analysis.

After creating the cluster segmentation, we integrated it into the original data frame to calculate descriptive statistics for each cluster. Specifically, we computed the mean values for all numerical variables and percentages for all categorical variables.

Although the differences between the two clusters were minor, it was found that cluster 2 tended to be more recent visitors, newer customers, and spent more money, but had a slightly lower basket margin than cluster 1. They also had a slightly higher response rate to the loyalty program.

Based on the evaluation of all the characteristics of both clusters, it appears that Cluster 2 customers are more lucrative to the business. (See Tables 1 & 2 and Plots 4, 5, 6, 7, 8, 9 and 10)

Predictive Analysis through Logistic Regression:

We performed predictive analysis for JC Dollars using logistic regression models to predict the 'response flag,' indicating a customer's response to promotion in the past 12 months. We created three models: Model 1 (Base Model) with the entire dataset, Model 2 using data from Cluster 0, and Model 3 using data from Cluster 1.

In the **Base Model**, we found significant negative relationships between response flag and both days since last visit and months since first purchase. Customer's profession (employed/salaried) showed a significant positive relationship with response flag. The number of houseware purchases in the last 6 months had a significant positive relationship, and residence status as rental had a significant negative relationship with response flag. Customer's gender as male had a significant positive relationship with response flag, but with only 90% confidence. The same can be interpreted from the summary table of our model in R illustrated in Figure 1. The model's accuracy was 0.9878, but sensitivity and precision were not meaningful due to the skewed dataset. This can be seen from the confusion matrix for the model Figure 2.

Model 2, built with Cluster 0 data, had similar significant relationships to the Base Model, with the addition of margin on health & beauty purchases showing a significant negative relationship with response flag. The same can be interpreted from the summary table of our model in R illustrated in Figure 3. The model had an accuracy of 0.988, but other evaluation metrics were not insightful due to data skewness. The same has been illustrated in Figure 4.

Model 3, built with Cluster 1 data, showed similar significant relationships to the previous models, but customer's gender as male was not significant, while marital status as divorced emerged as having a significant positive relationship with response flag. The same can be interpreted from the summary table of our model in R illustrated in Figure 5. The model's accuracy was 0.9882, but other evaluation metrics were not insightful due to data skewness. The same has been illustrated confusion matrix in Figure 6.

Since the dataset was skewed, with only 1.2% of customers responding to promotions, we performed a linear regression to understand total dollars spent in the past 6 months. The regression identified significant relationships with total dollars spent, such as days since last visit, number of each type of good, months since first purchase, gender, residence status, marital status, and profession. The model had an r-squared value of 0.5 and an RMSE over 2,000, indicating room for improvement. Still, the linear regression supplements other models' findings and helps understand loyal customer characteristics.

Correlation Matrix

We have opted to perform a correlation matrix analysis to better comprehend the relationships between various factors and to derive significant insights. To achieve this, we selected a subset from the original dataframe, specifically focusing on customers who responded positively to the promotion (response flag = 1). This approach was chosen due to the highly skewed nature of the data, with the intention of identifying strong correlations among customers who responded affirmatively.

While acknowledging that correlation does not necessarily indicate causation, it is important to consider that variables exhibiting a strong correlation with the total number of purchases and total monetary expenditure in the past six months could potentially guide JC Dollars in determining which departments to prioritize. Furthermore, understanding why spending may be higher in these departments could be valuable.

Although a threshold value of 0.5 is typically employed to signify strong correlations, we opted to utilize a more stringent threshold of 0.7. Upon conducting the matrix analysis, we observed that houseware, health and beauty, apparel, and electronics (hw, haba, apparel, elec) demonstrated strong correlations with the total number of purchases made within the last six months. Consequently, it is advisable for JC Dollars to investigate the reasons behind these strong correlations and concentrate on promoting these specific departments.

Reference to the correlation plot in Figure 7.

Conclusion and Business Insights

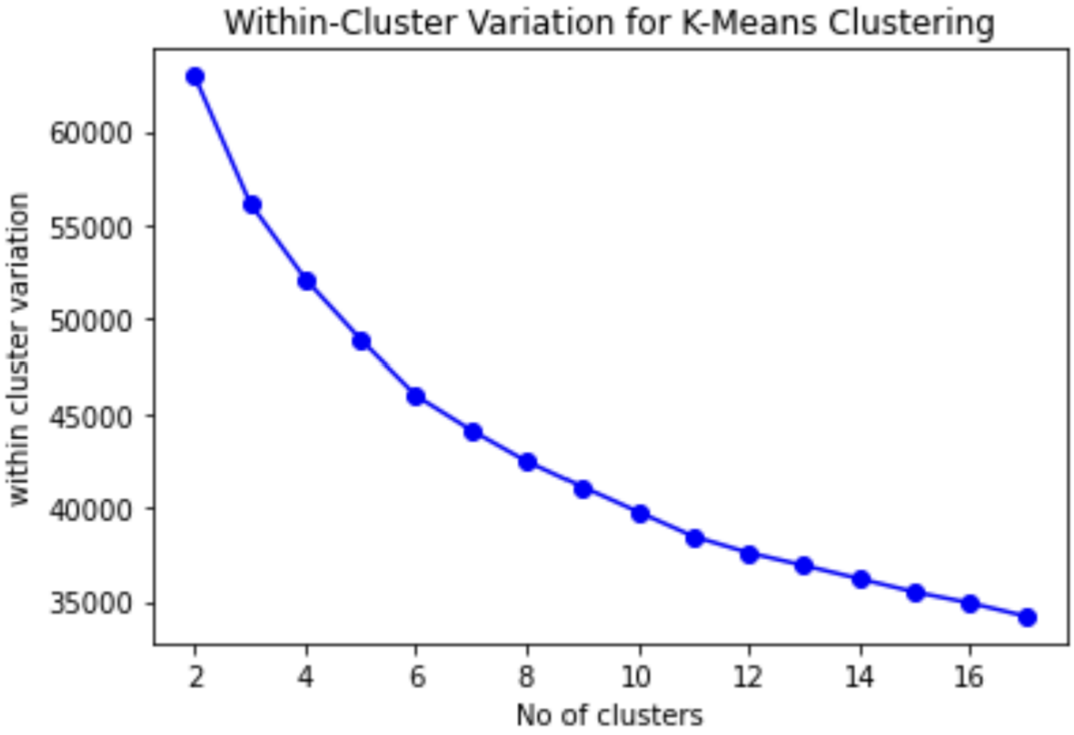
Based on the findings and recommendations provided in the report, some business insights and recommendations for JCD are as follows. For Cluster 0, it is important to focus on reducing the time since the last visit, as the analysis indicates that the number of days since the last visit has a significant negative impact on response_flag. Offering personalized promotions and incentives can encourage customers to visit the store or website more frequently. Targeting self-employed and salaried professionals, who show a higher likelihood of responding to the marketing campaign, can be effective by tailoring marketing messages and promotions to their specific needs and preferences. Emphasizing the importance of months since the first purchase can help engage newer customers and build loyalty early in their customer journey. By implementing welcome offers and special discounts, these customers can be kept engaged and encouraged to return. Additionally, offering incentives for purchasing higher-margin products can address the significant negative impact that margin on health and beauty products has on response_flag. Similarly, for Cluster 1, reducing the time since the last visit remains crucial. Offering personalized promotions and incentives will motivate customers to visit the store or website more frequently. As with Cluster 0, targeting self-employed and salaried professionals, as well as emphasizing the importance of months since the first purchase, can be beneficial. Divorced customers should also be targeted, as they show a higher likelihood of responding to the marketing campaign. Customizing marketing messages and offers can cater to their specific preferences and requirements.

General recommendations for both clusters include enhancing the loyalty program by investing more in personalized rewards and benefits, which will encourage repeat purchases and help retain loyal customers. Improving customer experience is also essential, particularly by addressing issues that led to the decline in customer satisfaction after changes in merchandise mix, store layouts, and pricing strategy. Ensuring a seamless in-store and online experience is crucial. Leveraging data analytics to better understand customer preferences and behavior can help personalize marketing campaigns, product recommendations, and customer communications. Finally, enhancing the digital presence by strengthening online sales channels and investing in digital marketing will enable JC Dollars to compete with e-commerce giants like Amazon and eBay. By implementing these targeted marketing strategies for each cluster, JC Dollars can better address customer preferences, improve loyalty, and ultimately drive growth in the face of increasing competition.

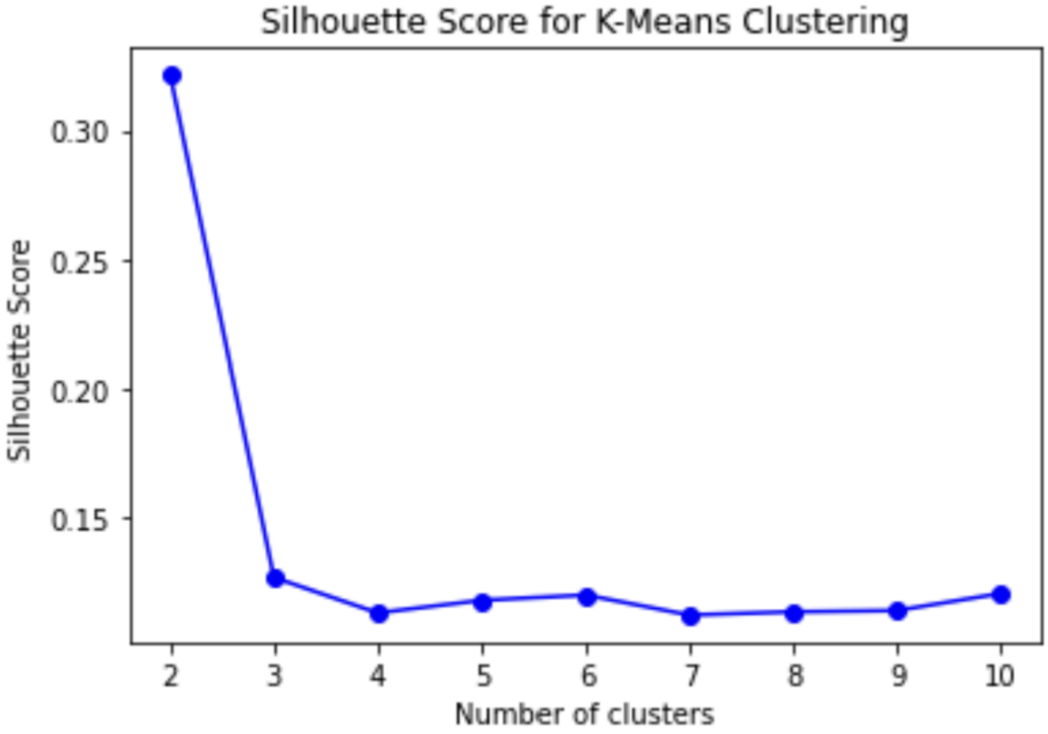
Moreover, the company should use both supervised and unsupervised machine learning methods to derive better results while gathering additional data from customer to strengthen the analysis.

Appendix

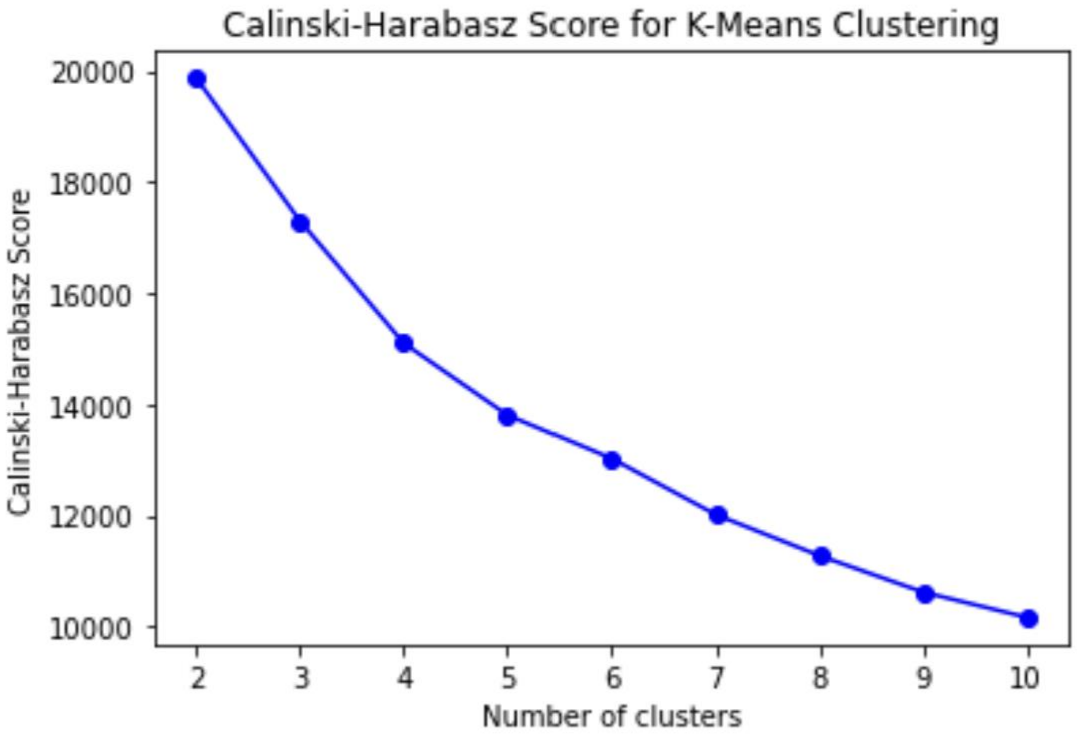
Plot 1



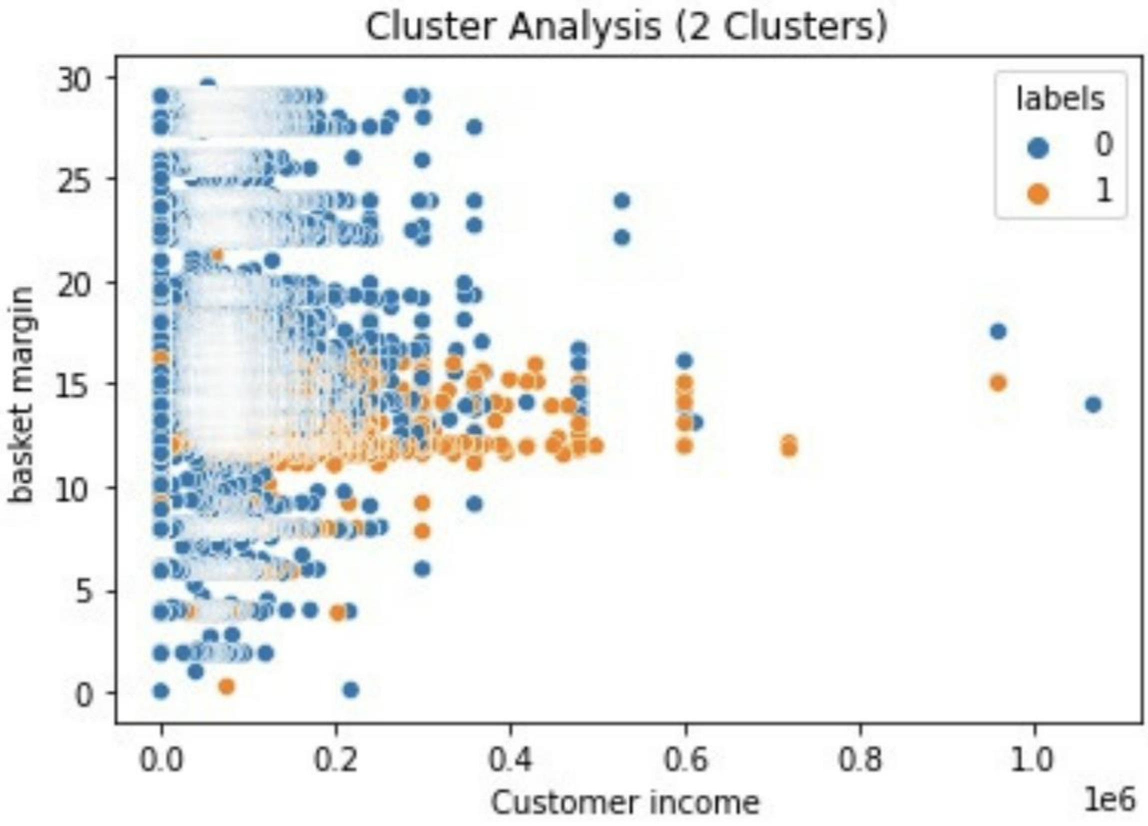
Plot 2



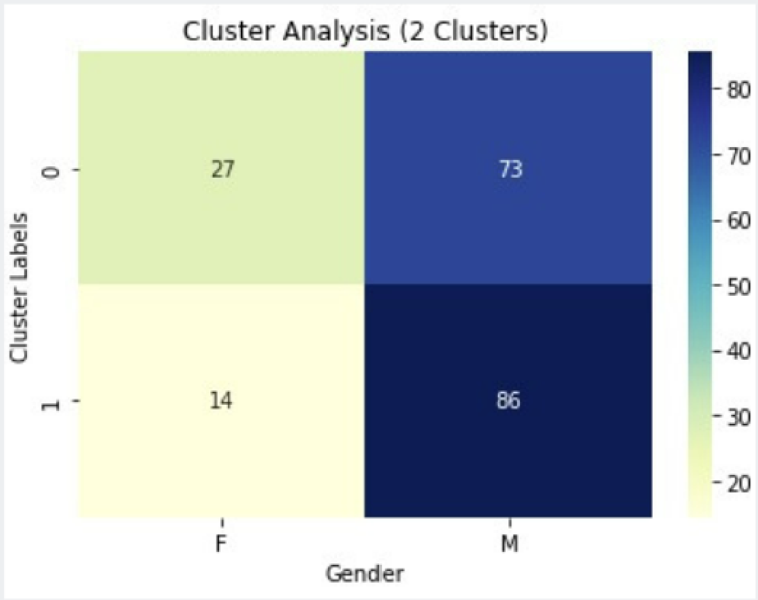
Plot 3


Calinski-Harabasz Score for K-Means Clustering
Calinski-Harabasz Score
20000
18000
16000
14000
12000
10000
2
3
4
5
6
7
8
9
10
Number of clusters

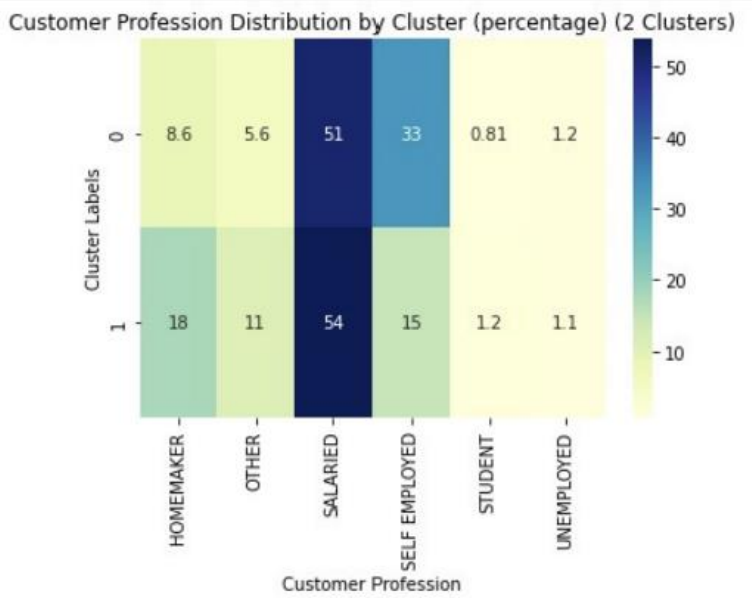
Plot 4


Cluster Analysis (2 Clusters)
basket margin
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Customer income
labels
● 0
● 1

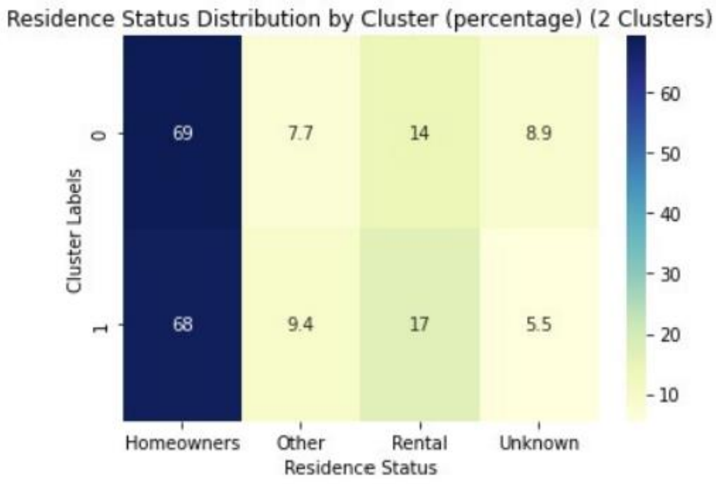
Plot 5


Cluster Analysis (2 Clusters)
Cluster Labels
0
1
27
73
14
86
F
M
Gender
80
70
60
50
40
30
20

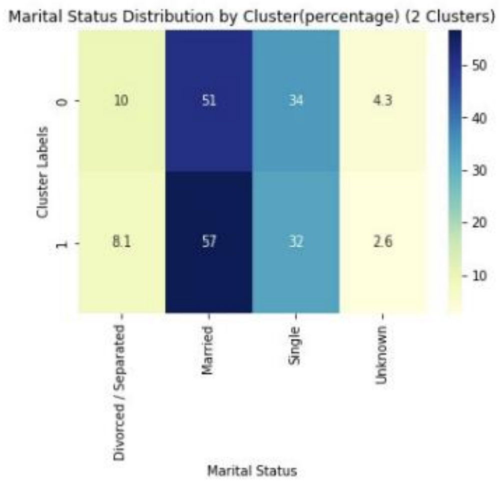
Plot 6


Customer Profession Distribution by Cluster (percentage) (2 Clusters)
Cluster Labels
0
1
8.6
5.6
51
33
0.81
1.2
18
11
54
15
1.2
1.1
HOMEMAKER
OTHER
SALARIED
SELF EMPLOYED
STUDENT
UNEMPLOYED
Customer Profession
50
40
30
20
10

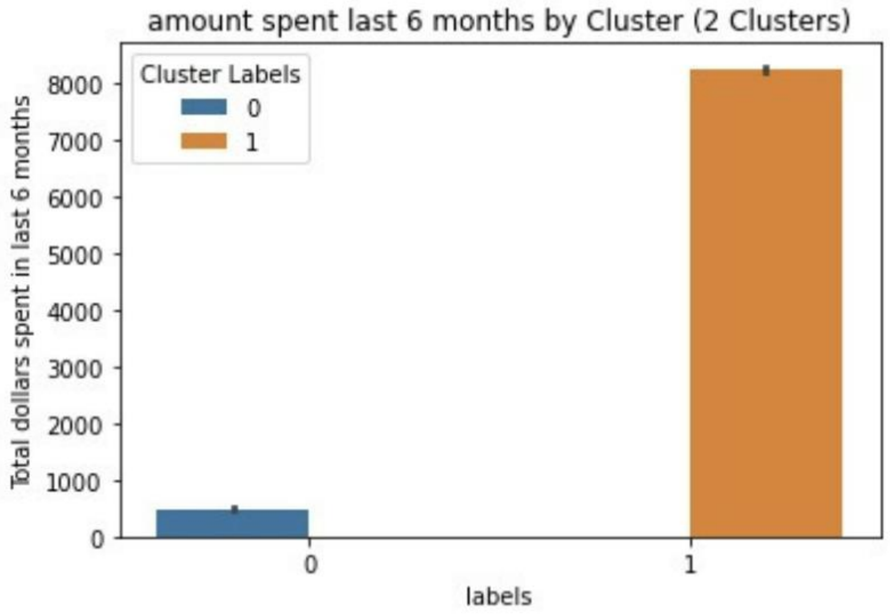
Plot 7


Residence Status Distribution by Cluster (percentage) (2 Clusters)
Cluster Labels
0
1
69
7.7
14
8.9
68
9.4
17
5.5
Homeowners
Other
Rental
Unknown
Residence Status
60
50
40
30
20
10

Plot 8


Marital Status Distribution by Cluster(percentage) (2 Clusters)
Cluster Labels
0
1
10
51
34
4.3
8.1
57
32
2.6
Divorced / Separated
Married
Single
Unknown
Marital Status
50
40
30
20
10

Plot 9



Plot 10

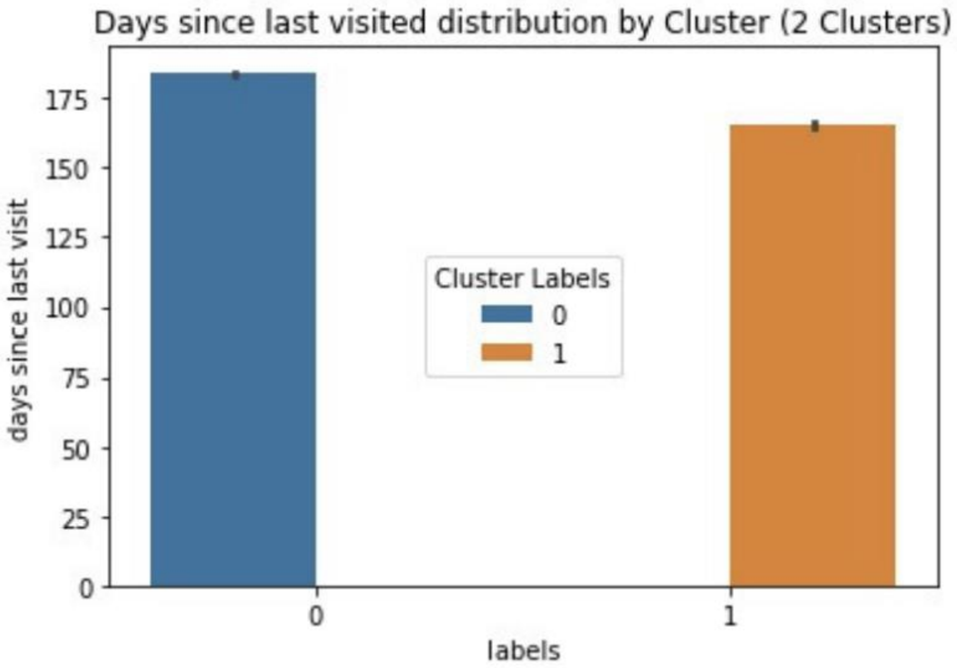
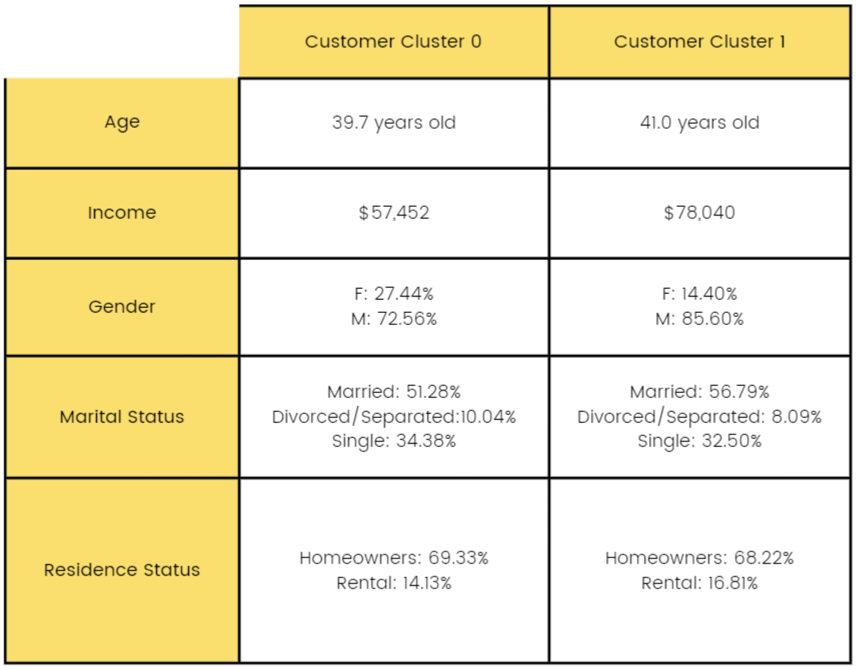
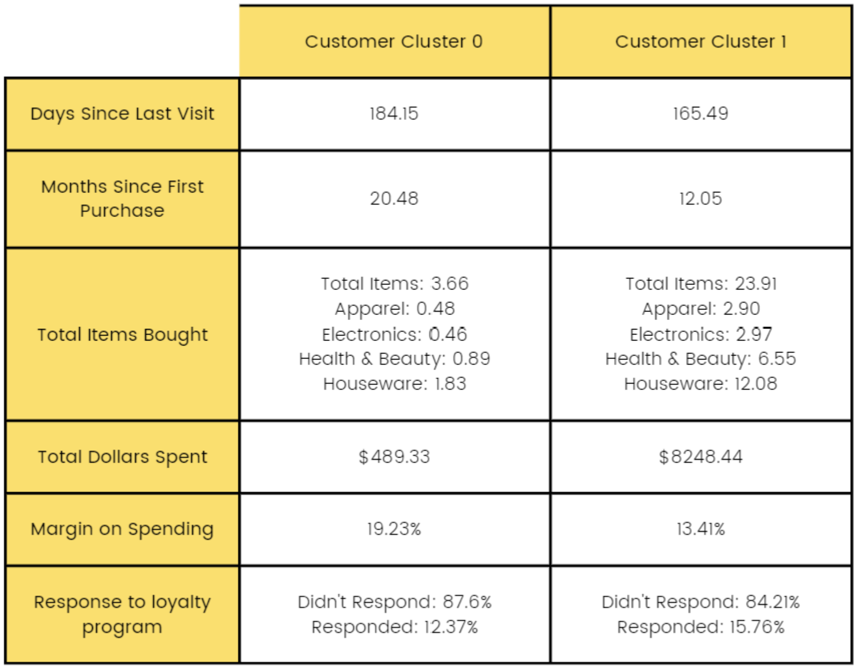


Table 1 – Demographics of Clusters



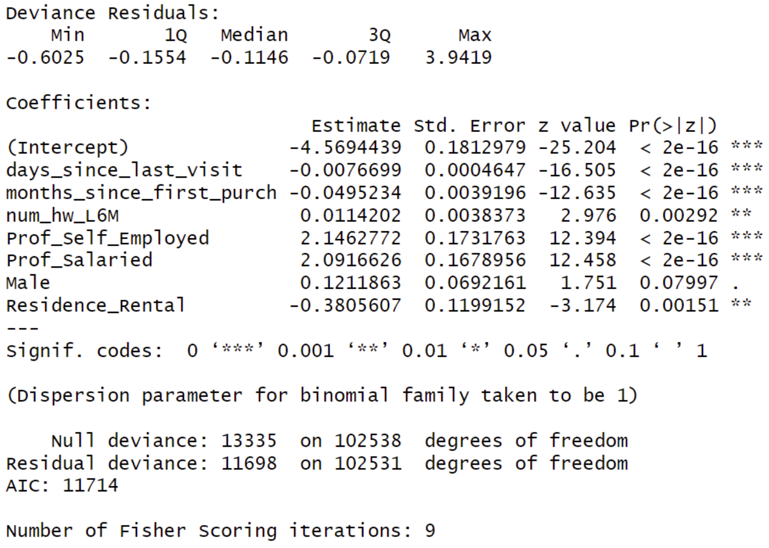
	Customer Cluster 0	Customer Cluster 1
Age	39.7 years old	41.0 years old
Income	$57,452	$78,040
Gender	F: 27.44% M: 72.56%	F: 14.40% M: 85.60%
Marital Status	Married: 51.28% Divorced/Separated:10.04% Single: 34.38%	Married: 56.79% Divorced/Separated: 8.09% Single: 32.50%
Residence Status	Homeowners: 69.33% Rental: 14.13%	Homeowners: 68.22% Rental: 16.81%

Table 2 – Shopping Behavior of Clusters



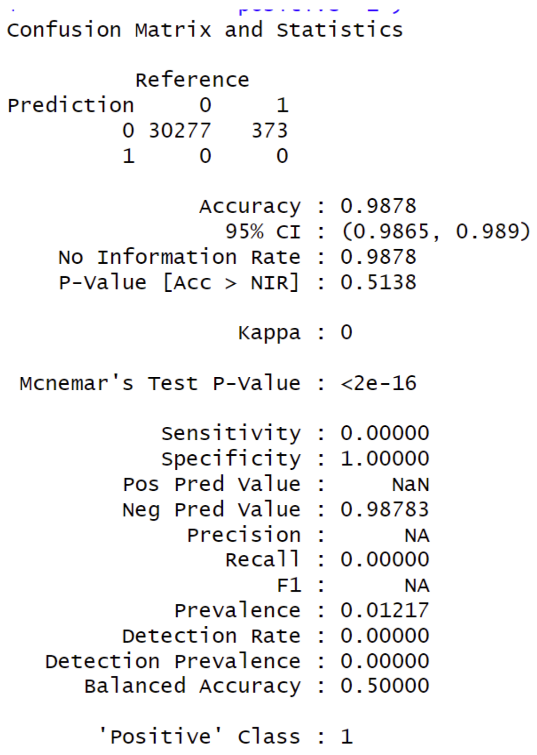
	Customer Cluster 0	Customer Cluster 1
Days Since Last Visit	184.15	165.49
Months Since First Purchase	20.48	12.05
Total Items Bought	Total Items: 3.66 Apparel: 0.48 Electronics: 0.46 Health & Beauty: 0.89 Houseware: 1.83	Total Items: 23.91 Apparel: 2.90 Electronics: 2.97 Health & Beauty: 6.55 Houseware: 12.08
Total Dollars Spent	$489.33	$8248.44
Margin on Spending	19.23%	13.41%
Response to loyalty program	Didn't Respond: 87.6% Responded: 12.37%	Didn't Respond: 84.21% Responded: 15.76%

Figure 1 – Summary Table for Base Model



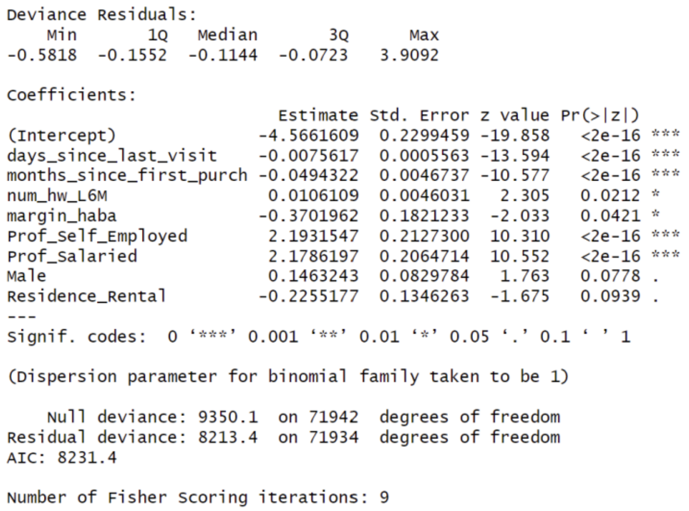
```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.6025  -0.1554  -0.1146  -0.0719   3.9419  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -4.5694439  0.1812979 -25.204  < 2e-16 ***
days_since_last_visit    -0.0076699  0.0004647 -16.505  < 2e-16 ***
months_since_first_purch -0.0495234  0.0039196 -12.635  < 2e-16 ***
num_hw_L6M                0.0114202  0.0038373   2.976  0.00292 ** 
Prof_Self_Employed        2.1462772  0.1731763  12.394  < 2e-16 ***
Prof_Salaried             2.0916626  0.1678956  12.458  < 2e-16 ***
Male                      0.1211863  0.0692161   1.751  0.07997 .  
Residence_Rental         -0.3805607  0.1199152  -3.174  0.00151 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13335  on 102538  degrees of freedom
Residual deviance: 11698  on 102531  degrees of freedom
AIC: 11714

Number of Fisher Scoring iterations: 9
```

Figure 2 – Confusion Matrix for Base Model



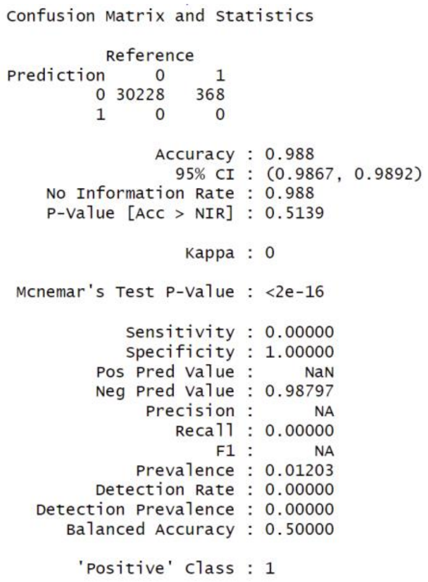
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 30277   373
         1     0     0
                                          
               Accuracy : 0.9878          
                 95% CI : (0.9865, 0.989) 
    No Information Rate : 0.9878          
    P-Value [Acc > NIR] : 0.5138          
                                          
                  Kappa : 0               
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 0.00000         
            Specificity : 1.00000         
         Pos Pred Value :     NaN         
         Neg Pred Value : 0.98783         
              Precision :      NA         
                 Recall : 0.00000         
                     F1 :      NA         
             Prevalence : 0.01217         
         Detection Rate : 0.00000         
   Detection Prevalence : 0.00000         
      Balanced Accuracy : 0.50000         
                                          
       'Positive' Class : 1               
```

Figure 3 – Summary Table for Model 1



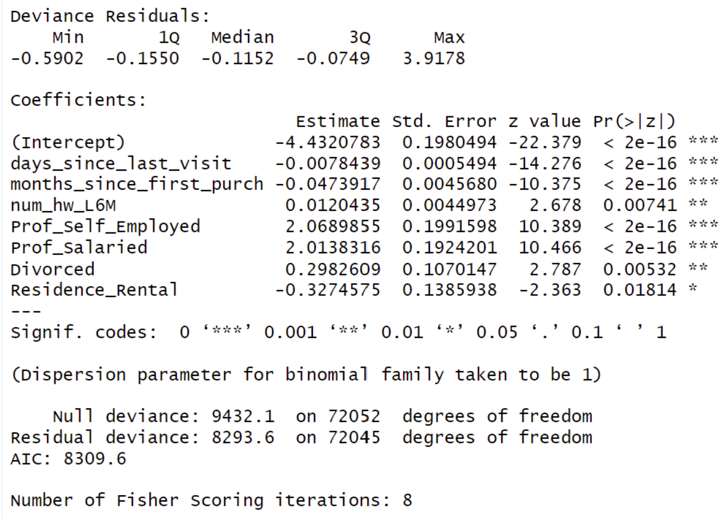
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5818  -0.1552  -0.1144  -0.0723   3.9092

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.5661609  0.2299459 -19.858   <2e-16 ***
days_since_last_visit    -0.0075617  0.0005563 -13.594   <2e-16 ***
months_since_first_purch -0.0494322  0.0046737 -10.577   <2e-16 ***
num_hw_L6M                0.0106109  0.0046031   2.305   0.0212 *
margin_haba              -0.3701962  0.1821233  -2.033   0.0421 *
Prof_Self_Employed        2.1931547  0.2127300  10.310   <2e-16 ***
Prof_Salaried             2.1786197  0.2064714  10.552   <2e-16 ***
Male                      0.1463243  0.0829784   1.763   0.0778 .
Residence_Rental         -0.2255177  0.1346263  -1.675   0.0939 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9350.1  on 71942  degrees of freedom
Residual deviance: 8213.4  on 71934  degrees of freedom
AIC: 8231.4

Number of Fisher Scoring iterations: 9
```

Figure 4 – Confusion Matrix for Model 1



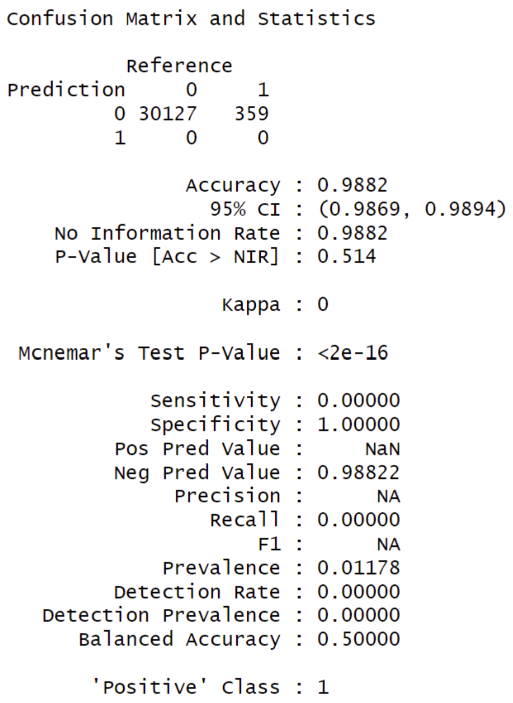
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 30228   368
         1     0     0

               Accuracy : 0.988
                 95% CI : (0.9867, 0.9892)
    No Information Rate : 0.988
    P-Value [Acc > NIR] : 0.5139

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.98797
              Precision :      NA
                 Recall : 0.00000
                     F1 :      NA
             Prevalence : 0.01203
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 1
```

Figure 5 – Summary Table for Model 2



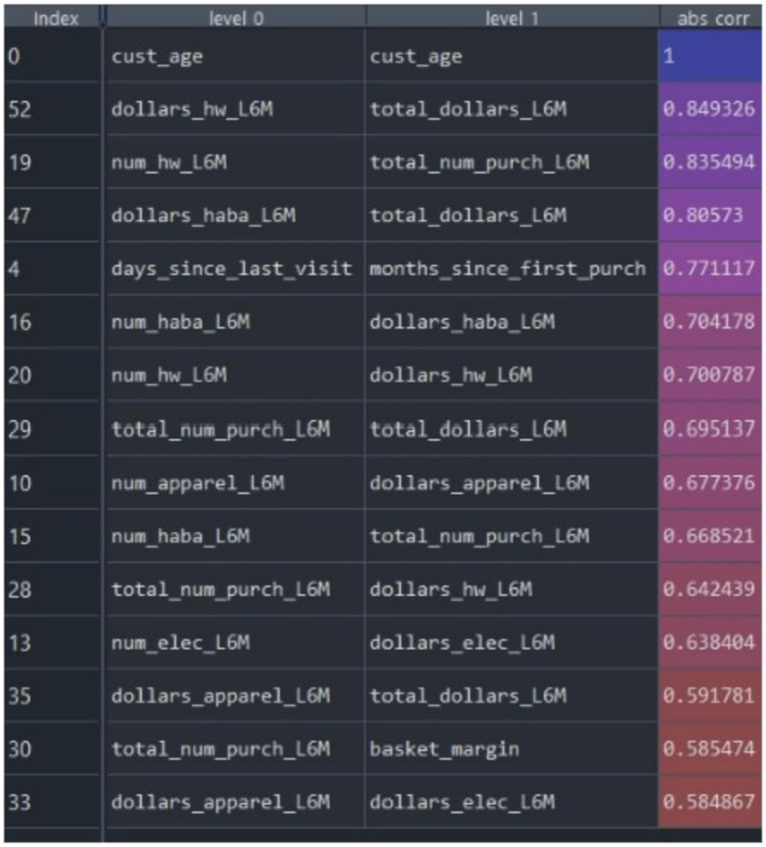
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5902  -0.1550  -0.1152  -0.0749   3.9178

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.4320783  0.1980494 -22.379  < 2e-16 ***
days_since_last_visit    -0.0078439  0.0005494 -14.276  < 2e-16 ***
months_since_first_purch -0.0473917  0.0045680 -10.375  < 2e-16 ***
num_hw_L6M                0.0120435  0.0044973   2.678  0.00741 **
Prof_Self_Employed        2.0689855  0.1991598  10.389  < 2e-16 ***
Prof_Salaried             2.0138316  0.1924201  10.466  < 2e-16 ***
Divorced                  0.2982609  0.1070147   2.787  0.00532 **
Residence_Rental         -0.3274575  0.1385938  -2.363  0.01814 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9432.1  on 72052  degrees of freedom
Residual deviance: 8293.6  on 72045  degrees of freedom
AIC: 8309.6

Number of Fisher Scoring iterations: 8
```

Figure 6 – Confusion Matrix for Model 2



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 30127   359
         1     0     0

               Accuracy : 0.9882
                 95% CI : (0.9869, 0.9894)
    No Information Rate : 0.9882
    P-Value [Acc > NIR] : 0.514

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.98822
              Precision :      NA
                 Recall : 0.00000
                     F1 :      NA
             Prevalence : 0.01178
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 1
```

Figure 7 – Correlation Plot



Index	level 0	level 1	abs corr
0	cust_age	cust_age	1
52	dollars_hw_L6M	total_dollars_L6M	0.849326
19	num_hw_L6M	total_num_purch_L6M	0.835494
47	dollars_haba_L6M	total_dollars_L6M	0.80573
4	days_since_last_visit	months_since_first_purch	0.771117
16	num_haba_L6M	dollars_haba_L6M	0.704178
20	num_hw_L6M	dollars_hw_L6M	0.700787
29	total_num_purch_L6M	total_dollars_L6M	0.695137
10	num_apparel_L6M	dollars_apparel_L6M	0.677376
15	num_haba_L6M	total_num_purch_L6M	0.668521
28	total_num_purch_L6M	dollars_hw_L6M	0.642439
13	num_elec_L6M	dollars_elec_L6M	0.638404
35	dollars_apparel_L6M	total_dollars_L6M	0.591781
30	total_num_purch_L6M	basket_margin	0.585474
33	dollars_apparel_L6M	dollars_elec_L6M	0.584867