BAN620 Data Mining

Project Report

TOPIC : SUPERSTORE MARKETING CAMPAIGN

As the **year-end sale** approaches, the management of a superstore is strategizing ways to make the most of their campaign efforts. They have decided to launch a new offer called the "**gold membership**" exclusively for their **existing customers**. This membership comes with an enticing **20% discount** on all purchases and is available at a significantly reduced price of $499, compared to its regular price of $999 on other days.The management feels that the best way to **reduce the cost** of the campaign is to make a predictive model which will classify customers who might purchase the offer.

Data Source

The data source for our project is the “Superstore Marketing Campaign Dataset” available on Kaggle at https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset.

Problem Statement

To identify customers with a high likelihood of a positive response, factors such as demographic characteristics, purchase history, customer engagement, personalization, psychographic traits,

and feedback sentiment can be analyzed. By considering these factors, customer segments can be identified based on purchasing history and demographic characteristics, enabling targeted marketing approaches for improved campaign effectiveness.

Data Field Description

The data set includes 2240 records which includes data on the customer's demographics, spending, and method of purchases.The 22 attributes in the dataset are as below:

- Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise
- ID - Unique ID of each customer
- Year_Birth - Age of the customer
- Complain - 1 if the customer complained in the last 2 years
- Dt_Customer - date of customer's enrollment with the company
- Education - customer's level of education
- Marital - customer's marital status
- Kidhome - number of small children in customer's household
- Teenhome - number of teenagers in customer's household
- Income - customer's yearly household income
- MntFishProducts - the amount spent on fish products in the last 2 years
- MntMeatProducts - the amount spent on meat products in the last 2 years
- MntFruits - the amount spent on fruits products in the last 2 years
- MntSweetProducts - amount spent on sweet products in the last 2 years
- MntWines - the amount spent on wine products in the last 2 years

- MntGoldProds - the amount spent on gold products in the last 2 years
- NumDealsPurchases - number of purchases made with discount
- NumCatalogPurchases - number of purchases made using catalog (buying goods to be shipped through the mail)
- NumStorePurchases - number of purchases made directly in stores
- NumWebPurchases - number of purchases made through the company's website
- NumWebVisitsMonth - number of visits to company's website in the last month
- Recency - number of days since the last purchase

Data Pre-Processing

We did several steps to pre-process data.

Missing Value

Firstly, we identified and dropped missing values in the income column. This allowed us to work with clean data, which is essential for accurate analysis.

Missing Values in Each Variables:

```
Id                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                24
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
Response               0
Complain               0
dtype: int64
```

Transform Columns

Next, we transformed the Birth Year column into Age, which made it easier to analyze and understand our customer demographics. We also converted the Date of Customer Enrolled column into Enrolled Days, which helped us understand the length of time each customer has been with the company.

```
Index(['Id', 'Education', 'Marital_Status', 'Income', 'Kidhome', 'Teenhome',
       'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts',
       'MntFishProducts', 'MntSweetProducts', 'MntGoldProds',
       'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases',
       'NumStorePurchases', 'NumWebVisitsMonth', 'Response', 'Complain', 'Age',
       'Enrolled_Days'],
      dtype='object')
```

Check Values in Categorical Variables

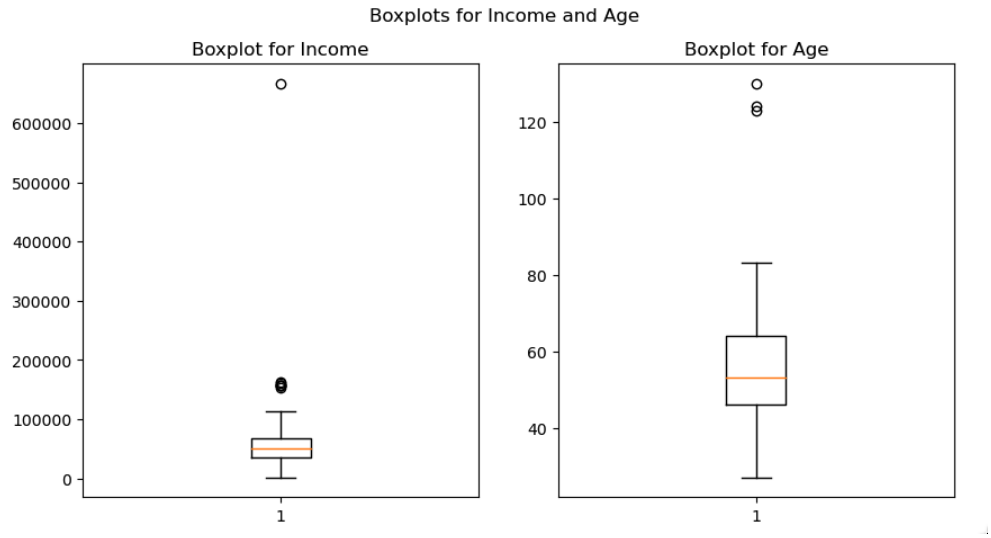
We then checked the values in our categorical variables, which allowed us to better understand the distribution of customers across different categories such as education level, and marital status.

```
Graduation    1116
PhD            481
Master         365
2n Cycle       200
Basic           54
Name: Education, dtype: int64
```

```
Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
YOLO          2
Absurd        2
Name: Marital_Status, dtype: int64
```

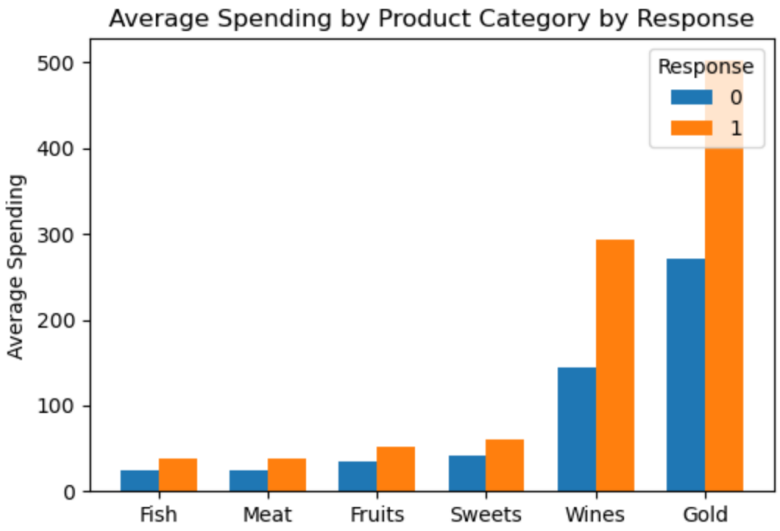
Outlier

One of the most important steps we took was identifying and dropping outliers in the Income and Age columns, using a boxplot. Outliers can skew our analysis and lead to inaccurate conclusions, so it was important to remove them before proceeding with our analysis.



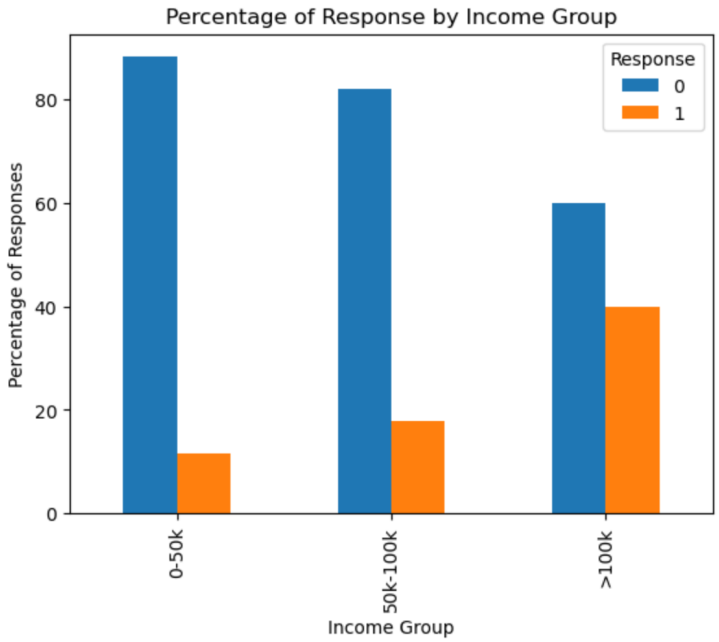
Dataset Exploration

We conducted several explorations of our dataset and discovered some interesting findings.



The plot above illustrates a comparison of the average spending on different product categories between customers who accepted and rejected the offer. It can be observed that customers who spent more on products had a higher response rate to the offer. Specifically, the likelihood of a positive response was nearly 5 times higher for customers who spent more on wines compared to those who spent on fish, meat, fruits, sweets, and gold products. Similarly, customers who spent more on meat had a positive response rate that was nearly 3 times higher than those who spent on the other product categories.

The above visualization displays the customer purchasing behavior by response to the offer. The results show that customers who make purchases directly in-store or from the company's website are more likely to have a positive response compared to those who purchase from catalog or discount channels.



The above plot illustrates the percentage of response by different income groups. It can be observed that customers with higher incomes are more likely to respond positively to the offer compared to those with lower incomes.

Analysis and Results

In order to predict customers' response and segment customers, we decide to use Classification Trees, Logit Regression and Clustering Models.

Classification Tree

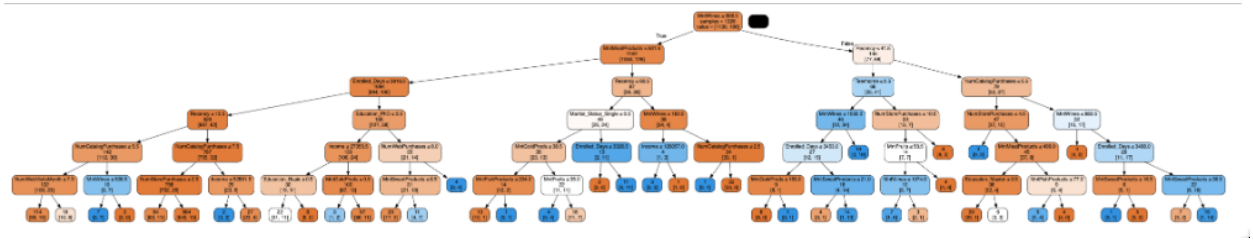
We choose to use the Classification Tree model to predict the customer's response. The first try we do a single tree and set a maximum depth as 6 to control the complexity of the tree. After

running the model, we calculate the confusion matrix to compare the performance for different models.

Confusion Matrix for Max_Depth=6

```
Confusion Matrix (Accuracy 0.8394)

         Prediction
Actual    0    1
     0 710   37
     1 105   32
```

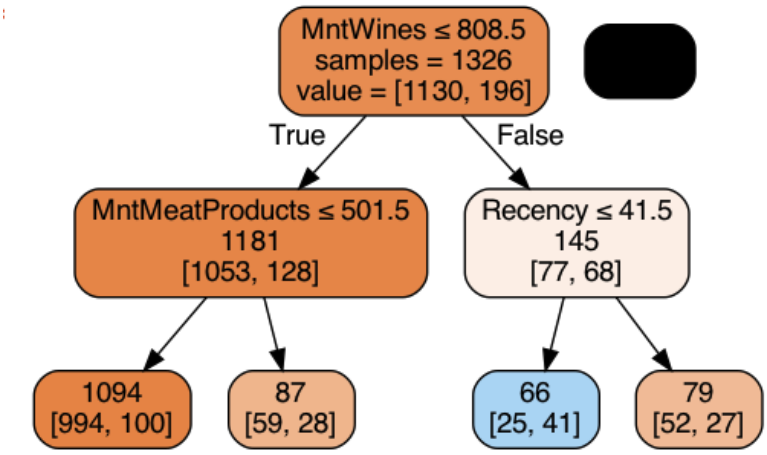


In order to optimize the performance of the classification tree, we also use a grid search method to find the best combination of factors.

Confusion Matrix after Grid Search

```
Confusion Matrix (Accuracy 0.8281)

         Prediction
Actual    0    1
     0 724   23
     1 129    8
```



Furthermore, other classification approaches we use are random forest, which uses multiple decision trees to improve accuracy, and boosted tree, which is a type of ensemble model. We calculate features importance for these two models since these models can no longer explain or implement due to loss of rules.

Confusion Matrix for Random Forest

Confusion Matrix (Accuracy 0.8609)

Actual	Prediction 0	1
0	733	14
1	109	28

Confusion Matrix for Boosted Tree

Confusion Matrix (Accuracy 0.8620)

Actual	Prediction 0	1
0	714	33
1	89	48

We calculate sensitivity, specificity and overall accuracy for these Classification Tree models. The sensitivity and specificity represent the true positive rate and true negative rate, respectively, while the overall accuracy indicates the percentage of correctly classified instances. We have compared the confusion matrix results from four different machine learning models, the Boosted Tree has the highest overall accuracy and sensitivity. The Random Forest has the highest specificity and a higher overall accuracy.

Performance Matrix Table for Four Classification Tree

Confusion Matrix	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
CM_6Branch	23.36%	95.05%	83.94%
CM_Grid	5.84%	96.92%	82.81%
CM_RF	20.44%	98.13%	86.09%
CM_Boost	35.04%	95.58%	86.20%

Since overall accuracy is close for Random Forest and Boosted Tree. We compared the feature importance of different variables in the dataset, based on Random Forest and Boosted Tree models. The left side is for Random Forest and the right side is for Boosted Tree.

Random Forest Feature Importance

Boosted Tree Feature Importance

	feature	importance
4	MntWines	0.110992
3	Recency	0.087615
0	Income	0.084070
6	MntMeatProducts	0.079061
17	Enrolled_Days	0.075154
9	MntGoldProds	0.061528
16	Age	0.056988
12	NumCatalogPurchases	0.053814
13	NumStorePurchases	0.047873
7	MntFishProducts	0.047210
5	MntFruits	0.045357
8	MntSweetProducts	0.044947
14	NumWebVisitsMonth	0.036695
11	NumWebPurchases	0.036503
10	NumDealsPurchases	0.029804

	feature	importance
4	MntWines	0.195336
3	Recency	0.131468
6	MntMeatProducts	0.094687
17	Enrolled_Days	0.077608
13	NumStorePurchases	0.073232
12	NumCatalogPurchases	0.071289
0	Income	0.065115
9	MntGoldProds	0.045609
7	MntFishProducts	0.041196
10	NumDealsPurchases	0.030414
14	NumWebVisitsMonth	0.029878
16	Age	0.024967
21	Education_PhD	0.023042

From both lists, we can see that the top features in terms of importance are related to the customer's spending behavior, such as amount spent on wines, meat, fish and sweet products, and number of purchases made from catalog, store, web or with discount, as well as their recency of purchase and income level. This suggests that these factors play a significant role in determining the customer's response.

Logit Regression

In addition to classification models, we also use logit regression, which is a method for analyzing a dataset in which explains what and how much magnitude of a factor's impact. We also calculate the Confusion Matrix for this model.

Confusion Matrix for Logit Model

Confusion Matrix (Accuracy 0.8394)

Actual	Prediction 0	Prediction 1
0	719	28
1	114	23

Performance Matrix Table for Logit Regression

Confusion Matrix	Sensitivity (%)	Specificity (%)	Total Accuracy (%)
CM_Logit	16.79%	96.25%	83.94%

Logit model is not very good at identifying the positive cases, but it performs well in identifying the negative cases. However, the overall accuracy of the model is not very high. Therefore, it may not be the best model for this particular task, and we will consider use Boosted Tree. However, the magnitude of the coefficients can give an indication of the strength of the impact of each factor on the response. The left table is variables that have the positive impact and the table on the right shows the negative impact.

	Coefficients
NumCatalogPurchases	0.075743
NumWebVisitsMonth	0.067851
NumWebPurchases	0.035844
Education_PhD	0.034540
Marital_Status_Single	0.028159
Marital_Status_Widow	0.009098
Marital_Status_Divorced	0.005821
MntSweetProducts	0.004395
MntGoldProds	0.002659
MntWines	0.002449
MntMeatProducts	0.001789
Kidhome	0.001287
Marital_Status_Alone	0.001020
MntFishProducts	0.000800
Complain	0.000188
Marital_Status_YOLO	0.000000

Income	-0.000015
Enrolled_Days	-0.000027
MntFruits	-0.001797
Education_Basic	-0.004955
Education_Master	-0.006612
NumDealsPurchases	-0.013264
Education_Graduation	-0.018598
Marital_Status_Together	-0.018810
Age	-0.019797
Recency	-0.023490
Teenhome	-0.026259
Marital_Status_Married	-0.030747
NumStorePurchases	-0.162190

The variables with the highest positive coefficients are the ones that have the strongest positive correlation with the likelihood of a customer giving a positive response. From the table, we can see that the variables with the highest three positive coefficients are: number of purchases made using category, number of visits to company's website in the last month and number of purchases made through the company's website.

On the other hand, the variables with the highest negative coefficients have the strongest negative correlation with the target variable. From the table, we can see that the variable with the highest negative coefficient is: number of purchases made directly in stores.

Clustering

Clustering is another model we use that helps us group similar data points together. We tested different method and found that ward, which is a method for grouping similar objects into clusters based on a measure of similarity, and K-means which is a type of unsupervised learning that partitions a set of data points into K clusters based on their distance to the mean of each cluster us better clusters. The results from these two methods are similar.

Profile plots of normalized means of each input variable for each cluster – K-means

	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	\
0	-0.166	0.780	0.680	-0.020	-0.069	-0.353	-0.271	
1	0.485	-0.698	0.656	-0.015	0.764	0.028	-0.019	
2	1.160	-0.754	-0.677	0.023	0.843	1.152	1.386	
3	-1.070	0.790	-0.891	-0.028	-0.821	-0.513	-0.643	
4	-0.475	0.318	0.797	0.029	-0.730	-0.556	-0.649	

	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	\
0	-0.348	-0.316	0.165	2.184	
1	-0.016	0.023	0.393	0.211	
2	1.219	1.135	0.666	-0.570	
3	-0.522	-0.510	-0.527	-0.295	
4	-0.577	-0.550	-0.589	-0.145	

	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	\
0	0.605	-0.190	-0.036	0.858	
1	0.920	0.378	0.794	-0.044	
2	0.338	1.158	0.809	-1.097	
3	-0.715	-0.760	-0.852	0.678	
4	-0.732	-0.702	-0.702	0.124	

	Age	Enrolled_Days	Response	Cluster
0	0.034	0.618	0.223	Cluster 0
1	0.412	0.160	0.129	Cluster 1
2	-0.016	-0.014	0.276	Cluster 2
3	-0.823	-0.040	0.121	Cluster 3
4	0.510	-0.348	0.042	Cluster 4

```
Cluster 0 (193 members) with total responses 43
Cluster 1 (482 members) with total responses 62
Cluster 2 (518 members) with total responses 143
Cluster 3 (537 members) with total responses 65
Cluster 4 (480 members) with total responses 20
```

Profile plots of normalized means of each input variable for each cluster – Ward Linkage

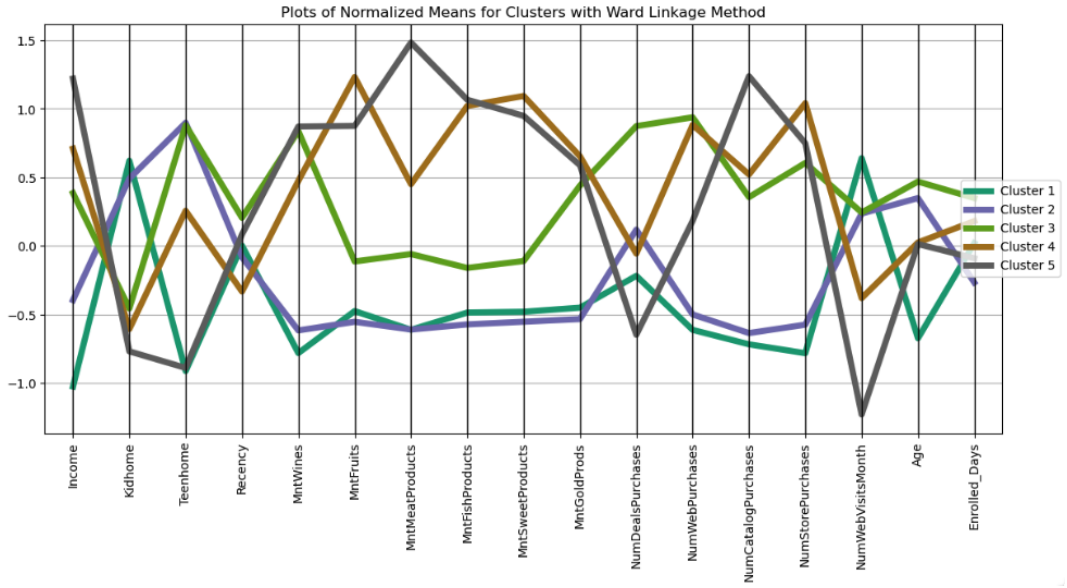
	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	\
1	-1.027	0.621	-0.913	0.002	-0.782	-0.476	-0.612	
2	-0.396	0.491	0.898	-0.083	-0.617	-0.554	-0.611	
3	0.385	-0.457	0.881	0.204	0.838	-0.114	-0.060	
4	0.710	-0.607	0.256	-0.333	0.474	1.232	0.450	
5	1.220	-0.770	-0.890	0.084	0.871	0.875	1.483	

	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	\
1	-0.487	-0.481	-0.451	-0.219	
2	-0.573	-0.554	-0.534	0.118	
3	-0.161	-0.110	0.440	0.874	
4	1.021	1.094	0.653	-0.057	
5	1.066	0.948	0.588	-0.649	

	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	\
1	-0.613	-0.718	-0.783	0.640	
2	-0.501	-0.636	-0.575	0.236	
3	0.938	0.355	0.605	0.244	
4	0.883	0.521	1.041	-0.382	
5	0.182	1.238	0.748	-1.231	

	Age	Enrolled_Days	Cluster
1	-0.674	0.020	Cluster 1
2	0.349	-0.267	Cluster 2
3	0.469	0.349	Cluster 3
4	0.025	0.180	Cluster 4
5	0.013	-0.091	Cluster 5

```
Cluster 1 (591 members) with total responses 72
Cluster 2 (570 members) with total responses 38
Cluster 3 (403 members) with total responses 67
Cluster 4 (217 members) with total responses 31
Cluster 5 (429 members) with total responses 125
```



From the plot, we can see there are five clusters:

Cluster 1: This cluster has the lowest income and tends to spend less on all product categories.

They have a higher number of kids at home and a higher number of web visits per month.

Cluster 2: This cluster has lower than average income and tends to spend less on all product categories except for meat and fish products. They have a higher number of teenagers at home and a lower number of catalog and web purchases.

Cluster 3: This cluster has average income and tends to spend the most on wines and gold products. They have a lower number of kids and teenagers at home, and a higher number of catalog and web purchases.

Cluster 4: This cluster has above-average income and tends to spend the most on fruits, meat, fish, and sweet products. They have a lower number of kids at home, a higher number of store purchases, and a lower number of web visits per month.

Cluster 5: This cluster has the highest income and tends to spend the most on all product categories, except for deals purchases. They have a lower number of kids and teenagers at home, a lower number of web visits per month, and a higher number of catalog and store purchases.

In summary, the clusters can be characterized based on their income, family size, and purchasing behavior. This information can be used to tailor marketing strategies to each cluster to better meet their needs and preferences.

Conclusion and Recommendations

Through our analysis from various models we understand that Boosted Trees give us the maximum overall accuracy of 86.20% (with higher percentage of Sensitivity and Specificity) These factors might increase the odds of a positive response by their respective factors.

Based on our analysis we found that some of the most important factors affecting customer response are Number of catalog purchases made, number of web purchases made, amount spent on wines and amount spent on meat. The superstore could inform the customer that they have

increased the varieties of wine and meat collection added in the store or the updated website with all new products offering pre-orders placement.

We can observe from the clusters that Cluster 5 (who have high income, low kidhome, low teenhome, high web purchases made, high meat and high wines amount spent) and cluster 3 (who have decent income which is above 0.5, high recency, high number of web purchases, high teenhome, high amount spent on wine) can be a good customer segment to target for a future campaign. Targeting these specific audiences could help reduce cost for the overall campaign.

What we learnt from this project

- We gained immense knowledge about data mining.
- How to connect results with business outcomes.
- Work as a team and played on everyone's individual strengths.