EBA5003 Practice Module in Customer Analytics

Proposal for Discount Marketing Strategy

Project Report

22 November 2020

Group 4

Disha Jaya Shetty (A0215469X)

Nitin Lal Das (A0215434M)

Sayandip Ghosh (A0215460N)

Shivangi Verma (A0215504R)

INDEX

1. Business Understanding and Objectives

"Mom&Pop's Shop" is a local retail shop selling apparel, jewelry, house wares, small appliances, electronics, groceries, pharmaceutical products – with basic motto being – find everything you need under one roof – at discounted prices. It is a family business being run through generations - a local heritage and cultural insignia. Currently, business is facing fierce pricing competition from large conglomerates and e-commerce. Mom&Pop's Shop wants to leverage analytics to stay ahead in the competition.

SetShop Inc. is a data analytics startup that focuses on helping these brick & mortar shops to increase business by getting loyal customers with high CLTV, through targeted marketing strategies. It is helping a local retail store "Mom&Pop's Shop" to identify strategy for a discount marketing process using the concepts of customer analytics. The goal of this project is to deepen Mom&Pop's Shop's relationship with its customers and expand the business organically.

Mom&Pop's Shop has run several discount campaigns/coupon redemption campaigns in the past and now wishes to reassess its strategy, to see if they have been successful in their past endeavors and to optimize their future plans. SetShop.Inc Analytics team was tasked to four major purposes: business requirement understanding, analysis of previous coupon discount campaigns, to model the behavior and to provide suggestions based on these three steps.

The above business goals can be obtained through the following technical goals:

Customer Segmentation

a) RFM analysis to identify profitable customers
b) Clustering to identify customer groups based on demographics, who redeem the coupons
c) CLTV to overlay the identified customer segments with a currency value
d) Propensity modelling to identify customers who are more likely to respond to discount marketing campaigns

Campaign Analytics

a) Market Mix Model analysis to find out the profitable marketing channels
b) Customer Journey Mapping to understand the different phases a customer goes through and the various touchpoints involved

2. Data Processing & Feature Engineering

Below is the data schema

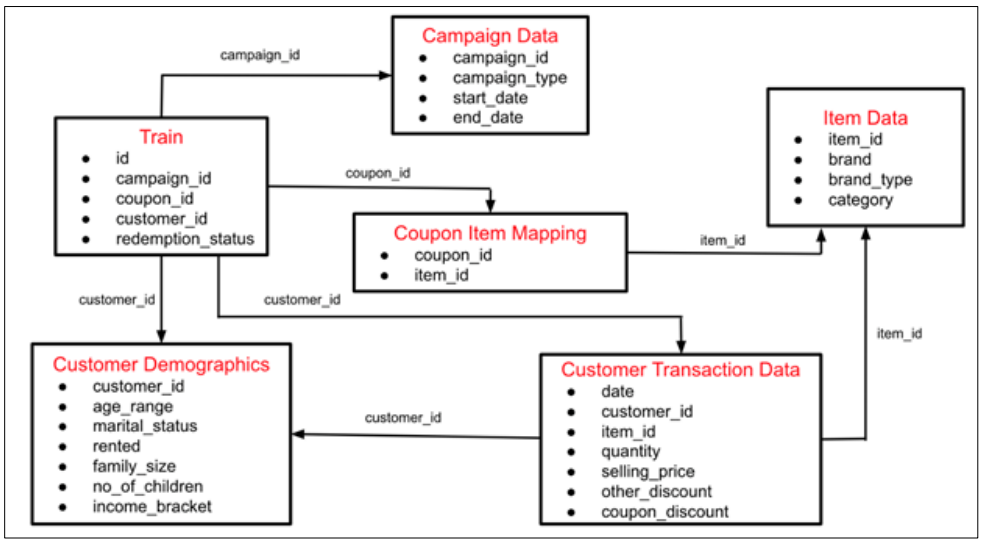


Figure 2.1

Overview:

a) Data was imbalanced with most values having redemption status value=1
b) For RFM analysis, we used Customer Transaction Data column without merging
c) For further analysis, we created a master file by using primary key as mentioned in Figure 2.1

Treating NAs and blanks in the dataset:

d) Under Customer Demographics table, marital_status column had large number on NaN values,so we imputed the NaN by using the following logic:
    i) customers with family size =1 will be single
    ii) customers whose family size - no of children == 1, will also be single
    iii) Wherever the difference of customers whose family size - no of children is 2 and marital status is NaN and No of Children is NaN we impute the Marital Status with Married
e) Under Customer Demographics table, no_of_children column had missing values, so we imputed them by using the following logic:
    i) Married people with family_size ==2 will have 0 children
    ii) Customers with family size 1 will have zero children
    iii) Singles with family size == 2, will probably have 1 child

Data type changes:

f) Under Customer Demographics table, the family_size and no_of_children columns had values like '5+' and data type was object, so we removed '+' and converted it to int data type
g) We used label encoder to convert age_range column in Customer_Demographics as 18-25 is 0, 26-35 is 1, 36-45 is 2, 46-55 is 3, 56-70 is 4 and 70+ is 5
h) Under Campaign_Data we converted the start_date and end_date to datetime type

3. Customer Segmentation

3.A. RFM Analysis

RFM (Recency, Frequency and Monetary) analysis was carried out to better understand the spending habits and nature of customers of Mom&Pop's Shop. This will help in identifying the segments of customers who are more valuable to the business and which segment of customers need focus.
The entire RFM analysis was conducted in Python. To conduct RFM analysis a baseline date was set which is one day after the campaign ended. This was used to calculate Recency and Frequency. Monetary is the summation of the total amount spend by the customer till baseline date.

Out[15]:

customer_id	Recency	Frequency	MonetaryValue	R	F	M	RFM_Segment_Concat	RFM_Score	RFM_Level
1	18	849	102776.01	1	3	2	132	6.0	Potential
2	22	292	39149.47	1	1	1	111	3.0	Require Activation
3	17	624	145323.86	1	3	2	132	6.0	Potential
4	33	210	48797.59	1	1	1	111	3.0	Require Activation
5	10	650	110403.01	3	3	2	332	8.0	Champions

Figure 3.A.1: RFM Segments with score and values for Recency, Frequency and Monetary Values

In Figure 3.A.1, we can see the Recency, Frequency and Monetary values; the RFM scores and the customer segments for the first five customers. The entire calculations were carried out at the customer ID Level. R, F and M scores were assigned by grouping the Recency, Frequency and Monetary values in quantiles. The scores range from 1-4 for four quantiles: 4 denoting the highest score and 1 is the lowest score. Net RFM Score of each customer was calculated by aggregating the individual R, F and M score of a customer. For segmentation of customers; six different score ranges were selected and were named uniquely; these score brackets represent the customer segments.

Segment No (Based on Priority)	Segment Name	RFM Score	Description	Count
1	**Can't Lose**	9 or above	Very high: Extremely important customer. Loyal and high spending	601
2	**Champion**	8	High: Very important and with proper attention can be extended further	212

3	**Needs Attention**	6,7	Medium: Needs further investigation to identify which area is lagging and accordingly targeted	108
4	**Potential**	5	Lower: Needs to be reached out to learn more about their grievances and desires - feedback needs to be taken	157
5	**Promising**	4	Low: Needs attention and evaluation at a case by case level to develop into a Long-Term Customer	400
6	**Require Activation**	3	Very Low: The long-term goals are centered towards converting these leads	104

Table 3.A.1: RFM Segments based on RFM Score Ranges (Bracket)

In Table 3.A.1, the RFM Segments based on RFM score ranges are listed. The segments are listed in decreasing order of priority; that is Segment 1 – Can't Lose is of the Highest Priority consisting customers with very high RFM Scores whereas the Segment 6 named 'Require Activation' has customers with the least customer scores. The count column shows the number of customers in each segment.

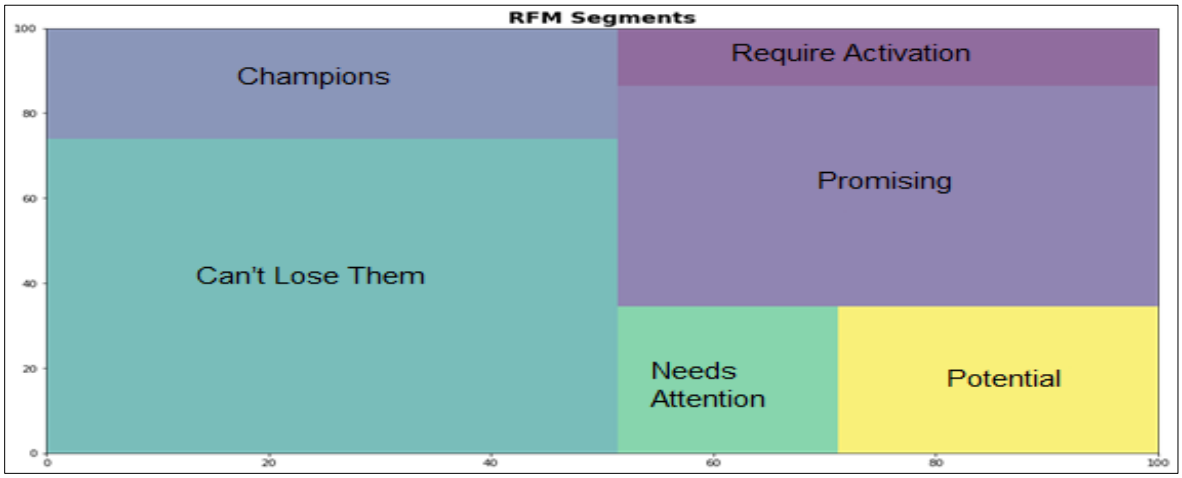


Figure 3.A.2; The RFM Segments represented in square plots

In Figure 3.A.2, we can observe the segments represented in the square plots such that the area of the rectangle denoting a segment is proportional to the count of customers in the segment. In the figure, it is clearly evident that about 50% of the customer belong to the segments 'Can't Lose them' and 'Champions', which are high score segments. From this we can infer that Mom&Pop's Shop has been doing well in retaining valuable customers; these segments can further be targeted for discount coupon campaigning and Cross-Sell or Up-Sell tactics can be used to strengthen the relationship further. A major proportion of the customer falls in the segment 'Promising' which has low RFM Score but can be targeted to engage more customers with the business. A Three-Dimensional Plot of the RFM Score has been included in the Appendix Section- Figure 3.A.3: to visualize the concept applied.

Segment Name	Recommended Action
Can't Lose	Send Vouchers and Coupons via Direct Mail and personalized emails on Birthday/Anniversary/ Special Occasion- Offer discounts, Cross Sell, Up-Sell- Extend
Champion	Send Discount offers via Direct Mail, Personalized Emails, SMS - Focus more on Cross Sell
Needs Attention	Send Coupon Code via SMS, Email- More focus on discount on items purchased previously
Potential	Send offers via Email- Focus more limited time discount offers discounts
Promising	Attempt to build brand awareness with them, try for short feedbacks by mail
Require Activation	Most probably has fallen off- needs further investigation and different marketing strategy

Table 3.A.2, The recommended actions for each segment

Table 3.A.2, shows the Recommended Action for the different Customer Segments. The higher segments vis a vis 'Can't Lose' and 'Champion' consists of customers with high RFM Scores and the business needs to keep them happy; so, a more personalized approach for them is required whereas the segments 'Needs Attention' and 'Potential' requires some push to increase their engagements with the business. This can be achieved by availing more discount on the products they bought recently. For the Lower Segments of 'Promising' and 'Require Activation'; promoting brand awareness is paramount and further investigation or surveys are required to identify the appropriate marketing strategy.

3.B. Clustering

Clustering involves grouping of data points. The data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering was used to segment the customers. This is important as it helps to develop more focused marketing strategies.

Customers were segmented on their basis of demographics: age range, marital status, type of accommodation, family size, no of children and income bracket. The table below (Table 3.B.1) gives a brief description on the variables of demographics.

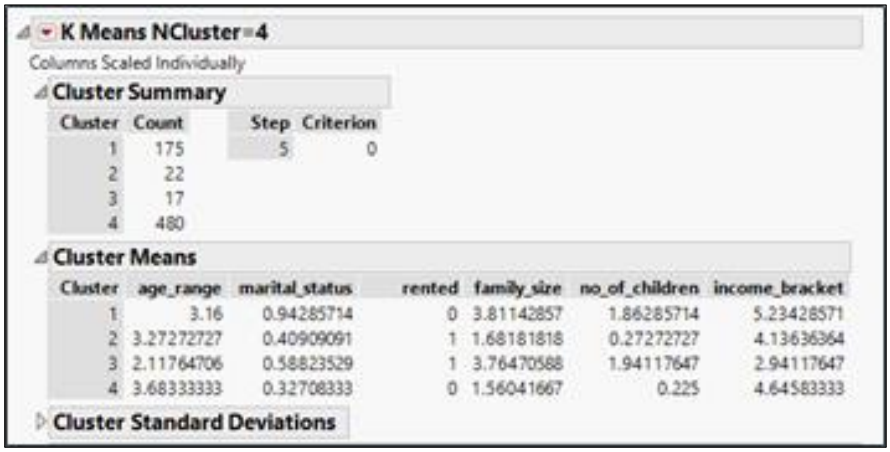
Age range	18-25: 1, 26-35: 2, 36-45: 3, 46-55: 4, 56-70: 5, 70+ : 6
Marital Status	Single :0, Married: 1

Type of accommodation	Rented: 0, Non-rented: 1
Family size	1,2,3,4,5+
No of children	1,2,3+
Income bracket	1-12

Table 3.B.1. Customer demographics

In the given dataset, it was observed unmarried middle-aged people living in non-rented accommodations were most likely to redeem the coupons.

K-means clustering was used to group the customers who have not redeemed the data. By careful analysis, clusters with k =4 was chosen as they had the least overlap. Figures 3.B.1 shows the count of each cluster and the average mean. Figure 3.B.2 shows the Biplot of the four clusters.



K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	175	5	0
2	22		
3	17		
4	480		

Cluster Means

Cluster	age_range	marital_status	rented	family_size	no_of_children	income_bracket
1	3.16	0.94285714	0	3.81142857	1.86285714	5.23428571
2	3.27272727	0.40909091	1	1.68181818	0.27272727	4.13636364
3	2.11764706	0.58823529	1	3.76470588	1.94117647	2.94117647
4	3.68333333	0.32708333	0	1.56041667	0.225	4.64583333

Cluster Standard Deviations

Figure 3.B.1: Count of Clusters

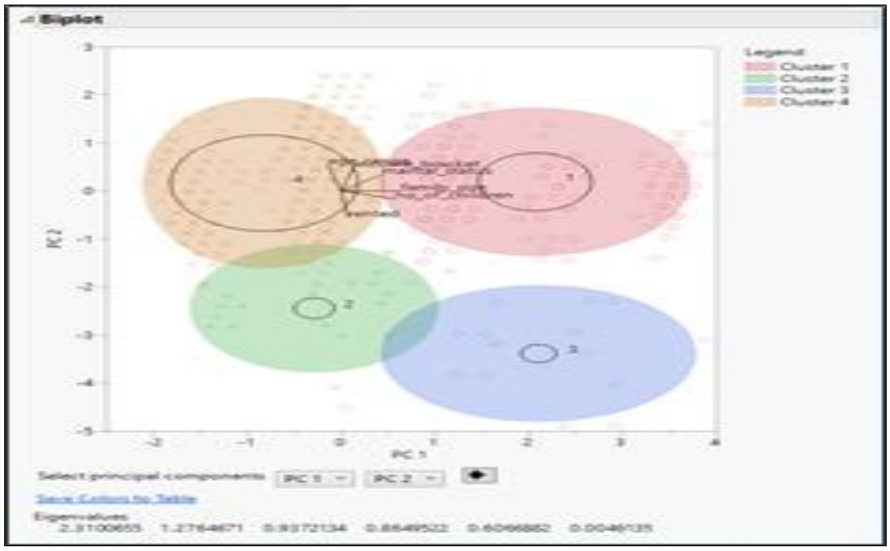


Figure 3.B.2: Biplot of the four clusters

From Figure 3.B.1, the following descriptions of the clusters can be deduced:

Cluster 4: Singles with no child in the age range of 46-55, living in non-rented accommodation and having average income bracket – Wealthy Singles

Cluster 1: Married with two children in the age range of 36-45, living in non-rented accommodation and having average income bracket – Loyal customers

Cluster 2: Mix of both married and singles with no child in the age range of 36-45, living in rented accommodation and having average income bracket – Carefree customers

Cluster 3: Mix of both married and singles with two children in the age range of 26-35, living in rented accommodation and having a low-income bracket – Young emerging customers

After careful analysis, it can be inferred that it would be best to target the customers in Cluster 4 as the characteristics of their demographics align with the people who have redeemed the coupons.

3.C. Customer Lifetime Value (CLTV)

The concept of Customer Lifetime Value or CLTV has been used to attach a currency value to the Customer Segments and get a notion about how much revenues can be expected from the customers.

For calculating Customer Lifetime value, the following Formulae were used:

a) CLTV = ((Average Order Value x Purchase Frequency)/Churn Rate) x Profit margin
b) Average Order Value= Total Revenue/Total Number of Orders
c) Purchase Frequency=Total number of orders/Total number of customers
d) Churn Rate: Churn Rate is the percentage of customers who have not ordered again.

The CLTV were calculated at Customer Id Level, and following assumptions were made:

Profit Margin is 10%; Churn rate is 1,5 and 10% for First, Second and Third trials respectively.

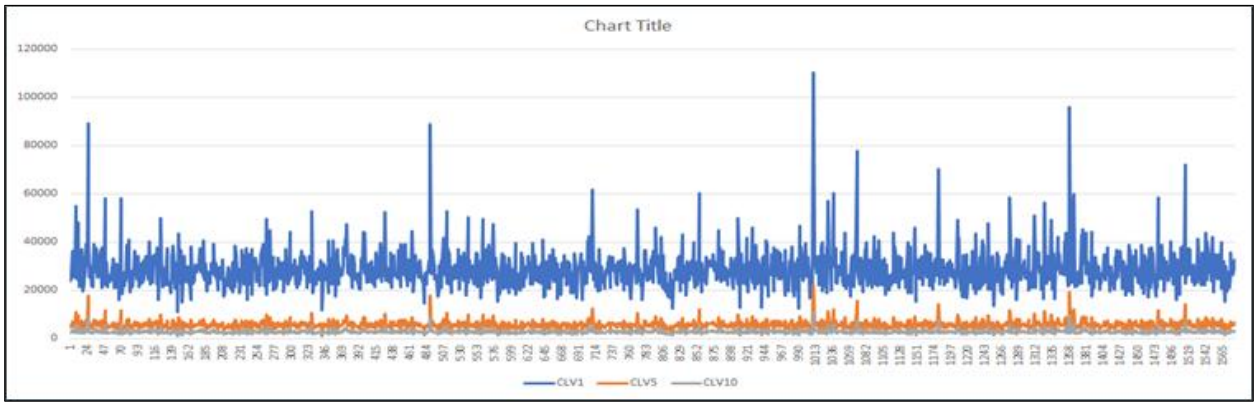


Figure 3.C.1: Customer ID VS CLTV

From Figure 3.C.1, we can infer that at churn Rate 5, the maximum CLV value observed was 22026.05 SGD for Mom&Pop's shop.

3.D. Propensity Model

Propensity model was applied to identify customers who are most likely redeem discount coupons.

To calculate scores probabilities derived from Machine learning models were used.

Three different models used were:

a) Naive Bayes
b) SVM
c) Logistic Regression

From, Figure 3.D.1, based on Evaluation metrics, Logistic Regression is observed to be the best performing model and was selected for the project.

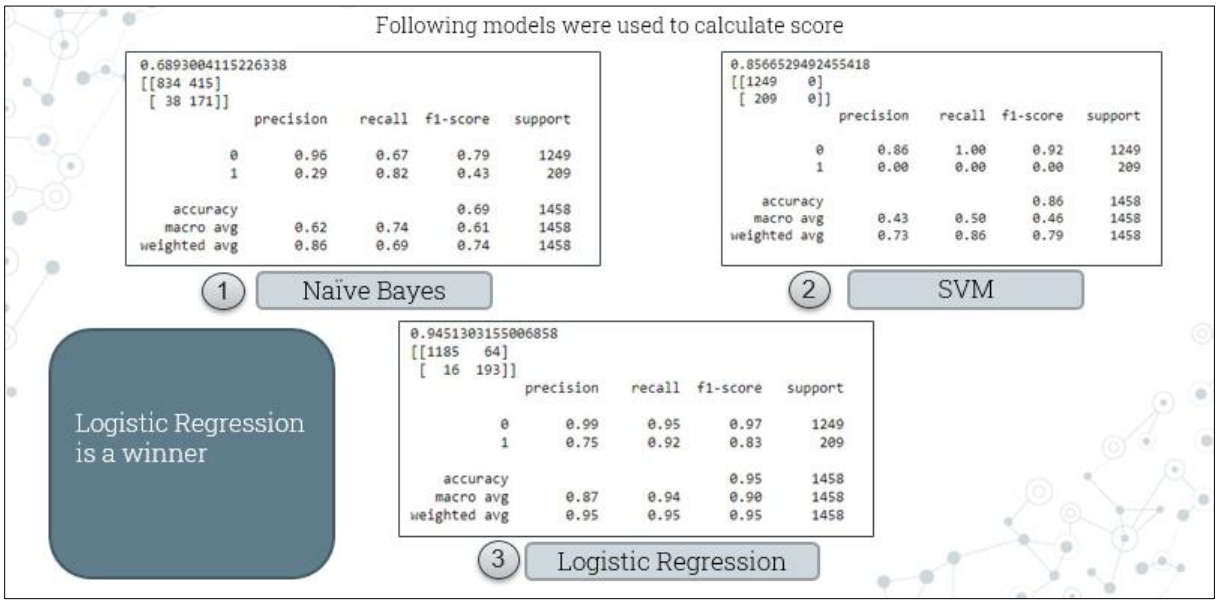


Figure 3.D.1: Evaluation Metrics of Naïve Bayes, SVM, Logistic Regression Models

The model displayed best performance in both train and test sets as seen in Figure 3.D.2.

```
print("Accuracy on training set: {:.2f}".format(model.score(X_train, y_train)))
print("Accuracy on test set: {:.2f}".format(model.score(X_test, y_test)))

Accuracy on training set: 0.96
Accuracy on test set: 0.95
```

Figure 3.D.2: Accuracy on Training and Test Set

The scores generated from the Logistic Regression were binned into several segments and Figure 3.D.3, depicts the criterion for the binning.

```
In [10]: pd.cut(df.Totalscore,bins=5,right=True).head()

Out[10]: 0    (5.029, 5.765]
         1    (5.029, 5.765]
         2    (5.029, 5.765]
         3    (5.765, 6.502]
         4    (5.029, 5.765]
         Name: Totalscore, dtype: category
         Categories (5, interval[float64]): [(3.552, 4.292] < (4.292, 5.029] < (5.029, 5.765] < (5.765, 6.502] < (6.502, 7.238]]

In [11]: df['cut']=pd.cut(df.Totalscore,bins=5, labels=['1', '2', '3', '4', '5'],right=True)
```

Figure 3.D.3: Binning Criterion

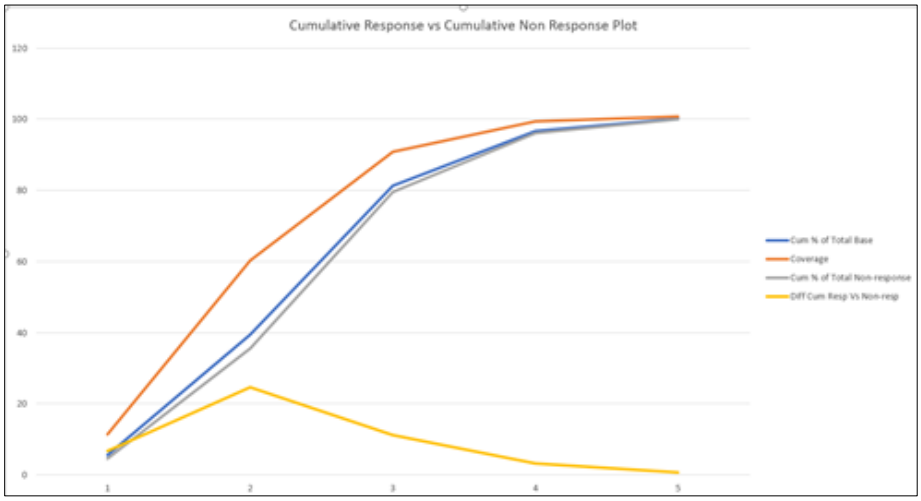
Based on the score bins MIS Report was generated and is shown in Table 3.D.1

Decile	# Obs	Min Score	Max Score	Average Score	Actual # Response	Actual Response Rate	Lift	% of Total Base	Cum % of Total Base	marginal Coverage	Coverage	# Non-Response	% of Total Non-response	Cum % of Total Non-response	Diff Cum Resp Vs Non-resp
1	273	6.644	7.46	7.052	83	30.4029304	2.026862027	5.617283951	5.617283951	11.38545955	11.38545955	190	4.599370612	4.599370612	6.786088921
2	1639	5.828	6.644	6.236	357	21.78157413	1.452104942	33.72427984	39.34156379	48.97119342	60.35665295	1282	31.03364803	35.63301864	24.72363431
3	2039	5.012	5.828	5.42	222	10.88769004	0.725846003	41.95473251	81.2962963	30.4526749	90.80932785	1817	43.98450738	79.61752602	11.19180182
4	745	4.196	5.012	4.604	62	8.322147651	0.554809843	15.32921811	96.6255144	8.504801097	99.31412894	683	16.53352699	96.15105301	3.16307593
5	169	3.376	4.196	3.786	10	5.917159763	0.394477318	3.477366255	100.1028807	1.371742112	100.6858711	159	3.848946986	100	0.685871056
	4860				729	15						4131			

Table 3.D.1: MIS Report

From the Table 3.D.1, a maximum lift of 2 was observed. Response rate and Lift was found to be decreasing monotonically however the model is still weak as by this model on targeting 5% of population we will be targeting only 11% of actual responders and overall, we need to target almost 40% of the customers to get a response rate of 60%. More over the maximum lift value of 2 is not a sign of a good model. Also, decile three was not considered here because according to it after targeting 81% of total base we get 90% response which does not make any business sense.

The Cumulative response against Cumulative Non response plot is shown in Plot 3.D.1



Plot 3.D.1: Cumulative response vs Cumulative Non response plot

The Observation period, Scoring Month, Performance period is as follows:

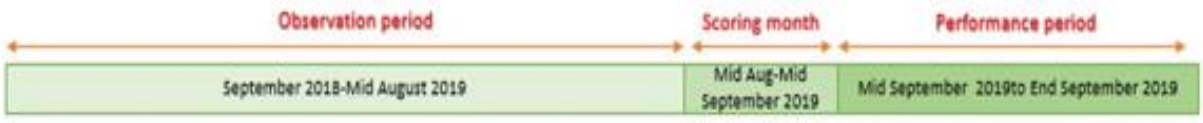


Figure 3.D.5: Observation period, Scoring Month, Performance period

4. Campaign Analytics

Transaction data of 1,048,576 single purchases and a turnover of SGD 56,827,606 were available from past campaigns, a Media Mix analysis was carried out to analyze the spending efficiency of campaigns over different modes of advertisement, to assess the performance of individual media channels. Mom&Pop's Shop being a local shop, have by far tried the cheapest mode of advertising, Digital Marketing- Facebook, Email, Instagram and YouTube -to advertise themselves.

A simple linear regression model was developed based on the data at hand to see the contributions and importance of each of the channels.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.932283
R Square	0.869152
Adjusted R	0.846396
Standard E	293913.2
Observatio	28

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	4	1.32E+13	3.3E+12	38.19424	7.66E-10
Residual	23	1.99E+12	8.64E+10		
Total	27	1.52E+13			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	7317.048	239460.5	0.030556	0.975887	-488045	502678.8	-488045	502678.8
GRP(Email)	32349.57	12915.33	2.504742	0.019786	5632.177	59066.96	5632.177	59066.96
GRP(Faceb	44290.98	19197.05	2.307176	0.030391	4578.848	84003.11	4578.848	84003.11
GRP(Youtu	2594.886	19412.98	0.133668	0.894828	-37563.9	42753.7	-37563.9	42753.7
GRP(Instag	10487.13	18490.73	0.567156	0.576102	-27763.9	48738.12	-27763.9	48738.12

Figure 4.1: Linear Regression results

In Figure 4.1, it can be observed that p-Values of E-Mail & Facebook are very significant, but those of YouTube and Instagram are not very significant. These may be dropped for the purpose of regression but considering the long-term goal of media visibility for brand imagery aspect, all the online channel was included in further analysis. Adjusted R-Sq is good. Figure 4.2, shows the plot of Fitted Sale against the Actual sale. Fitted sale is quite close to actual sale value.

Figure 4.2: Plot of Fitted Sale Vs Actual Sale

	ROI	Spend Efficiency
GRP(Email)	7.24969636	45.83974076
GRP(Facebook)	12.55395536	87.57127835
GRP(Youtube)	-0.470607413	-7.372057074
GRP(Instagram)	2.209278605	20.00524119

Figure 4.3: ROI & Spend Efficiency of the past campaigns via 4 channels.

From Figure 4.3, it can be observed that ROI for YouTube is negative but all existing online channel for marketing were included to enhance brand awareness through increased visibility for building brand equity in long term.

	Contribution	
	Vol	%
Intercept	7317.048	7.54%
GRP(Email)	32349.57	33.34%
GRP(Facebook)	44290.98	45.64%
GRP(Youtube)	2594.886	2.67%
GRP(Instagram)	10487.13	10.81%
Sum	97039.61	

Volume	
Base	7.54%
Incrementa	92.46%

Figure 4.4: Volume contributions of all Channels

From Figure 4.4, it is observed that Base Volume Contribution is low, this can be attributed to the factor that the store run campaigns very frequently and relies on Discount Marketing heavily as it is attempting to establish its brand image as a 'One Stop, Budget Retail Store'.

5. Customer Journey Map

Customer Journey Map will help to represent the different phases of customer experience visually. We will be able to understand the path and channels our customers take to get to our product which can be used to forecast the path of future customer. The various touchpoints mentioned in the journey map can be used to further improve customer experience.

In Figure 5.1, we can see various stages of the customer journey and the corresponding touchpoints. To improve customer experience business can conduct surveys to find out drawbacks in specific touchpoints.

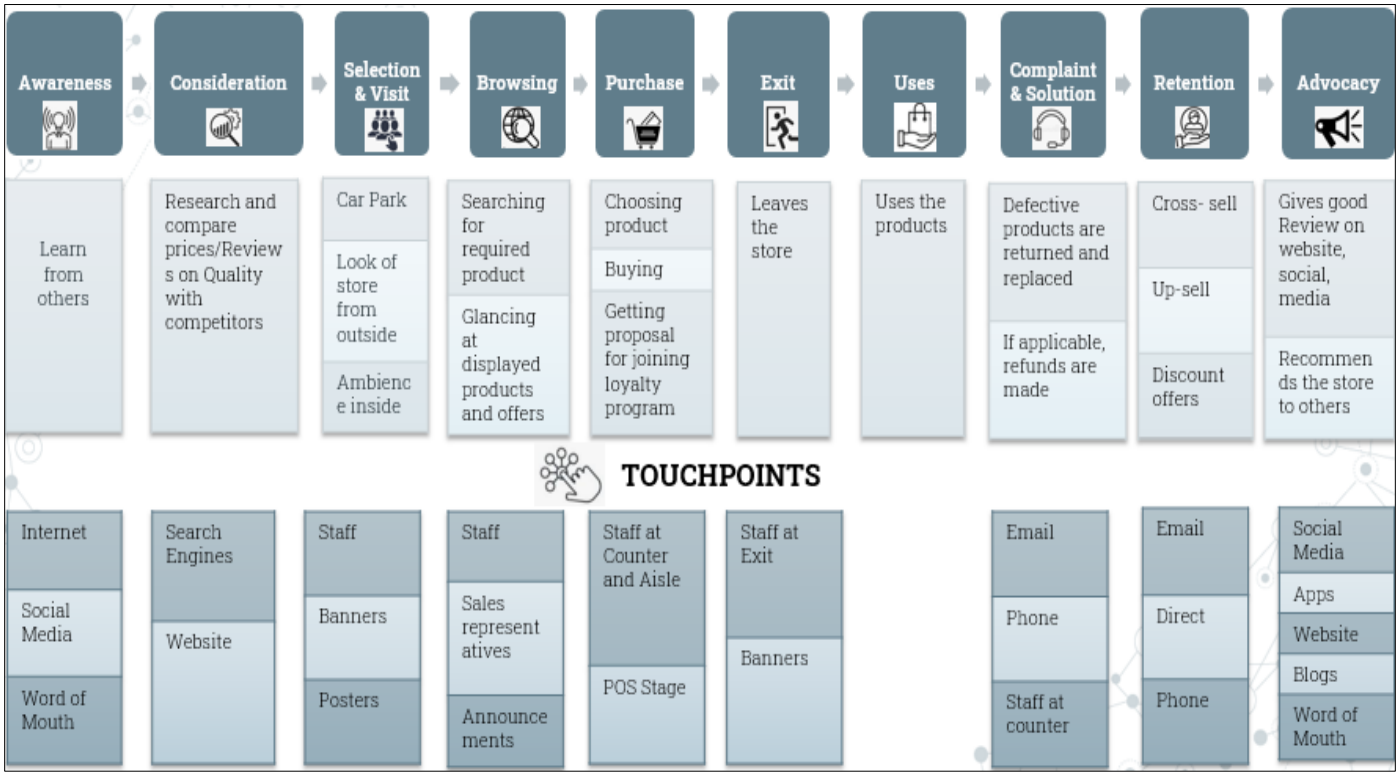


Figure 5.1: Customer Journey Map

6. Conclusions

Based on the observations made from various analytics techniques used in the project it can be concluded that - Mom&Pop's shop have done well in the past to build a profitable customer base but there are certain flaws in their approach which when resolved will bring in new customers and will also boost loyalty of existing customers and increase the revenue. The conclusions can be summarized as:

a) Customer Segment Target (RFM)- Based on the RFM segmentation, specific segments can be targeted using the different actions recommended in Table 3.A.2. Targeting the right segment, in right time using the right channel and offering the right incentive to buy will definitely help the business to expand the customer base and turn more customers profitable.
b) Customer Demographics Target (Coupon Redemption)- Cluster Wealthy Singles should be targeted as the demographics of this cluster are in line with customers who have redeemed the coupons
c) Spending needs be increased in Facebook & E-Mail channels, whereas for YouTube and Instagram, spending should be reduced but these channels are not to be dropped completely as they can help in our long-term goal of media visibility for brand imagery considerations.
d) Short surveys and Feedbacks on services, cleanliness and ambience at the store should be collected to improve customer experience.
e) Techniques like playing soothing music in the stores can be used to enhance customer mood which may result in better customer experience and thereby increase the purchase.

f) Using the results from the Propensity Model, we can conclude that customers in decile 1 and 2 should be targeted as they cover 60% of the responders, this will lead to better response for discount coupon redemption.

7. Gap Analysis

a) Clustering

For Clustering, as the original data set provided has mostly categorical values, like that of salary and age ranges instead of continuous values, and considering the fact that: Mom&Pop Shop has a large customer base of all types of customers, who come under one roof for all ranges of products, when the Clustering was done using demographics data and RFM scores, the resulting clusters were highly overlapping (as seen in figure 7.1 of appendix) and no clear distinction of pattern could be made out from it. There is scope in RFM Segmentation for Clustering which needs to be explored further.

As there are similar types of customers (in terms of demographics) who have and have not redeemed coupons, we intend to device clustering of people who have not redeemed coupons, compare them (in terms of demographics) with people who have redeemed coupons and try to find out meaningful groups of customers – to target – for coupon redemption.

In propensity model, we saw that the model was still weak even after monotonic decrease in lift values and response rate as uneven distribution in values under each bin was there. A maximum lift of only 2 could be achieved which might not be taken into consideration according to business standards.

8. Sources and References

a) Data source: https://www.kaggle.com/

b) References:

i) https://www.datacamp.com/community/tutorials/customer-life-time-value

ii) https://towardsdatascience.com/recency-frequency-monetary-model-with-python-and-how-sephora-uses-it-to-optimize-their-google-d6a0707c5f17

iii) https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-future-of-retail-how-to-make-your-bricks-click

iv) https://www.dnb.com/ca-en/perspectives/small-business/3-reasons-brick-mortar-businesses-need-analytics.html

APPENDIX

Section 3.A) RFM Analysis:

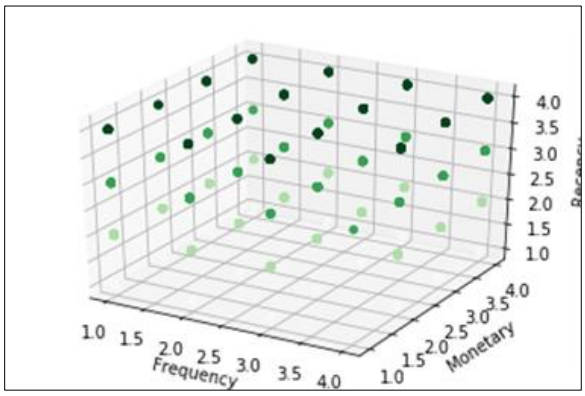
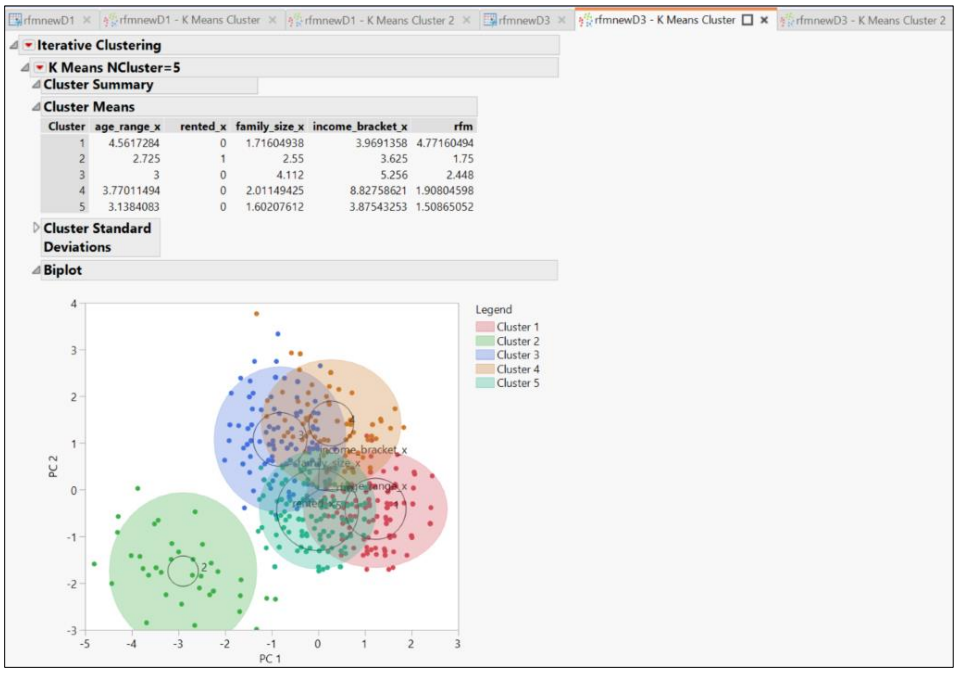


Figure 3.A.3: Three-Dimensional Plot of Recency, Frequency and Monetary (RFM) Scores

Section 7) Gap Analysis:



Cluster	age_range_x	rented_x	family_size_x	income_bracket_x	rfm
1	4.5617284	0	1.71604938	3.9691358	4.77160494
2	2.725	1	2.55	3.625	1.75
3	3	0	4.112	5.256	2.448
4	3.77011494	0	2.01149425	8.82758621	1.90804598
5	3.1384083	0	1.60207612	3.87543253	1.50865052

Figure 7.1: Overlap seen in customer segments when using RFM with Demographics.

Data dictionary:

campaign_data.csv: Campaign information for each of the 28 campaigns

Variable	Definition
campaign_id	Unique id for a discount campaign
campaign_type	Anonymised Campaign Type (X/Y)
start_date	Campaign Start Date

end_date	Campaign End Date

coupon*item*mapping.csv: Mapping of coupon and items valid for discount under that coupon

Variable	Definition
coupon_id	Unique id for a discount coupon (no order)
item_id	Unique id for items for which given coupon is valid (no order)

customer_demographics.csv: Customer demographic information for customers

Variable	Definition
customer_id	Unique id for a customer
age_range	Age range of customer family in years
marital_status	Married/Single
rented	0 - not rented accommodation, 1 - rented accommodation
family_size	Number of family members
no*of*children	Number of children in the family
income_bracket	Label Encoded Income Bracket (Higher income corresponds to higher number)

customer*transaction*data.csv: Transaction data for all customers for duration of campaigns in the train data

Variable	Definition
date	Date of Transaction
customer_id	Unique id for a customer
item_id	Unique id for item
quantity	quantity of item bought
selling_price	Sales value of the transaction
other_discount	Discount from other sources such as manufacturer coupon/loyalty card
coupon_discount	Discount availed from retailer coupon

item_data.csv: Item information for each item sold by the retailer

Variable	Definition
item_id	Unique id for itemv
brand	Unique id for item brand

brand_type	Brand Type (local/Established)
category	Item Category

train.csv: Train data containing the coupons offered to the given customers under the 28 campaigns

Variable	Definition
id	Unique id for coupon customer impression
campaign_id	Unique id for a discount campaign
coupon_id	Unique id for a discount coupon
customer_id	Unique id for a customer
redemption_status	(target) (0 - Coupon not redeemed, 1 - Coupon redeemed)

MarketMix.csv: Impressions and GRP Data of the 28 campaigns

campaign_id	Unique id for a discount campaign
GRP & Digital Impressions	GRP & Digital Impressions (Email, Facebook, YouTube & Instagram)