October 5th, 2020

Dunnhumby: Let's Get Sort-of-Real

Aditya Kamboj

Agenda

1. Objective
2. Data Models
   - Consistency vs Monetary Value
   - K-means Clustering
3. Model Results
4. ROI Analysis
5. Recommendations
6. Appendix
   - RFM(C) Model

Objective

Analyzing retail data to identify the customer segments who would be the most beneficial recipients for the store.

DATA MODELS OVERVIEW

Popular segmentation models / methodologies used in the industry

Demographics

Identifying key demographics, and delivering content based on that segment. It can be as simple as gender, or a complex model leveraging several demographic features like age, race, ethnicity, income.

Geographical

Customers can be targeted based on their location (postal code or FSA) as similar shopping patterns can be deduced from shoppers in the same geographical region.

Behavioral

Leveraging past customer behaviour to predict future actions. E.g. Purchasing for certain occasions, buying certain brands, or significant life events like moving, getting married, or having a baby.

Psychological

Psychological customer segmentation tends to involve softer measures such as attitudes, beliefs, or even personality traits. Like Last minutes shopper, weekly planning, buying only certain brands

Customer Status

This methodology is used to bucket customers in two key segments i.e. active and lapsed customers.
Every vertical defines these segments differently.

RFM

This model is often used in the direct marketing or database marketing campaigns.
Customers are segmented on the basis of Recency (last visit) , Frequency and Monetary value.

Three models were evaluated for the analysis; the M-C model was selected for the final output

Consistency - Monetary Value

A simple two factor model that provides a quick and easy perspective. However, it misses the granularity aspect compared to other options.

This model is used for the analysis

- Strengths:
  - Easy to construct
  - Less effort
  - Provides strong high-level recommendations
- Weakness:
  - Lacks granularity

RFM-C

RFM is a popular customer centric model that utilizes three factors important in targeting customers. It can be improved upon by adding the consistency Metric.

- Strengths:
  - Provides more granular insights
  - It comparatively considers more parameters
- Weakness:
  - Output may produce more than required number of segments

K-means Clustering

Simple yet popular, unsupervised algorithm. Can be powerful and is easy to implement. Several iterations are needed to get robust results

- Strengths:
  - Simple Implementation
  - Easy to scale to larger datasets
- Weakness:
  - Choosing K
  - Different partitions can result in different outputs
  - Sensitive to outliers

The analysis included five distinct steps

Data Cleaning → Exploratory Analysis → Computing Consistency → M+C | RFMC Analysis → Segment aggregation

Input

- ❑ Fixed the data types, checked for missing values
- ❑ Eliminated missing values where necessary

- ❑ Checked basic statistics and performed EDA to understand the data

- ❑ Calculated lag between each visit for all customers
- ❑ Used standard deviation to establish measure of consistency
- ❑ Segmented customers in quantiles.

- ❑ Calculated reach, frequency and monetary value of customers and segmented all the customers in 4 Quantiles.

Output

The Consistency and Value model provided some interesting initial insights

CONSISTENCY

- Measure of spread in purchase frequency
- Calculated standard deviation of lag between each visit for every customer

	Cust Code	Shop Date	Diff
Row 1	CUST000013	2007-04-23	Nat
Row 2	CUST000013	2007-05-22	29 days
Row 3	CUST000013	2007-06-01	10 days
Row 4	CUST000013	2007-07-19	9 days

Cust Code	Std
CUST000013	27.631010
CUST000055	57.188189
CUST0000679	59.435054
CUST001058	20.222441

695 customers who visited the store only once were eliminated. Customers who visited twice, their first visit difference value was imputed as 0 for the convenience of calculations

Description

Data Head

Notes

VALUE

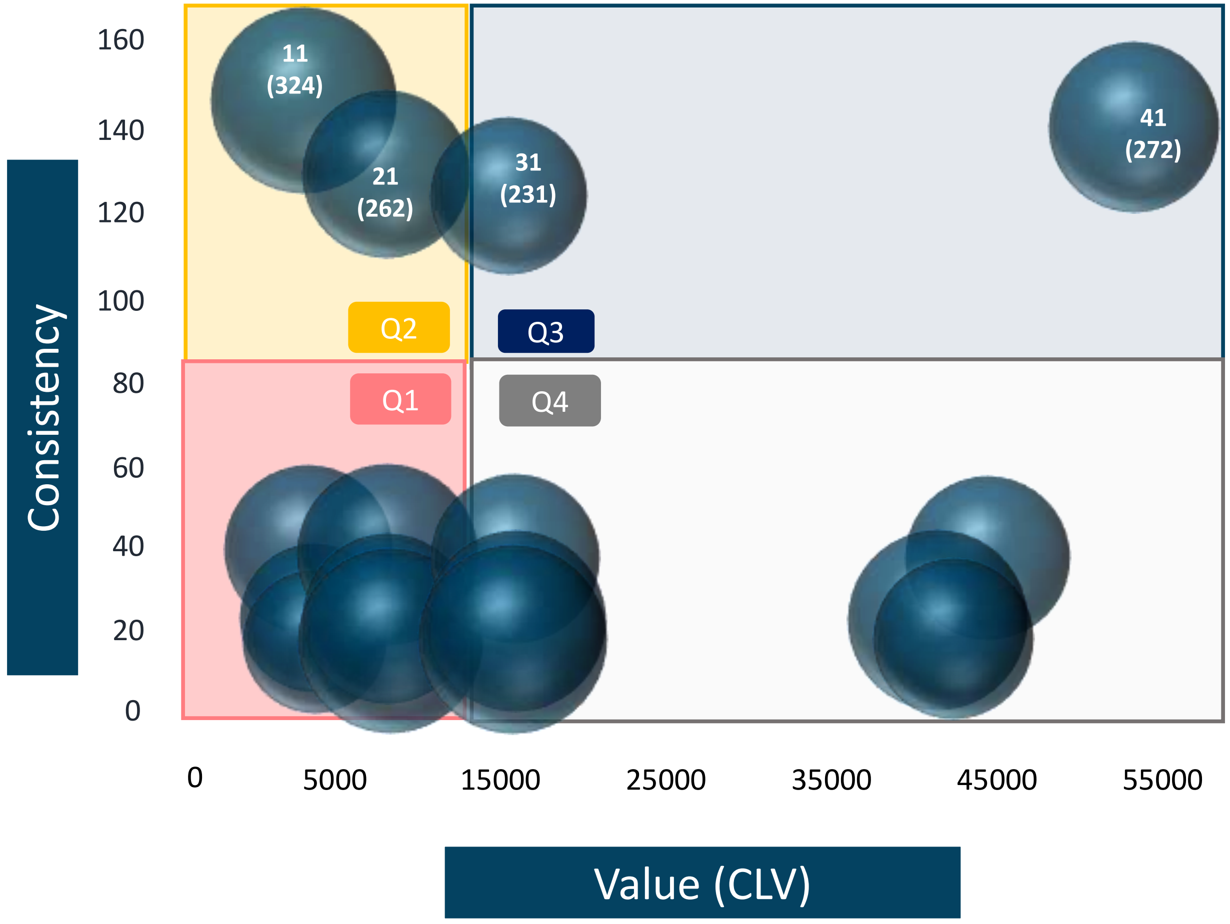
- Measure of value of each customer
- Calculated customer life value (CLV)

Cust Code	Days	Units	Spends	Avg order Value	CLV
CUST000013	139	122	$261.12	$21.76	$13270.88
CUST000055	19	320	$2671.55	$178.10333	$108620.77
CUST0000679	73	78	$141.73	10.123571	$6174.11
CUST001058	45	98	$316.90	10.222521	$6234.49

Customer value = Average order value * Purchase Frequency

The model provides some interesting results; a majority of customers are consistent but not high value (Q1)

CONSISTENCY VS VALUE OUTPUT



Quadrant 1 Q1 : Low value but Consistent shoppers

Quadrant 2 Q2 : Low value and inconsistent shoppers

Quadrant 3 Q3 : High value but inconsistent shoppers

Quadrant 4 Q4 : High value and consistent shoppers

The target segments on the graph would be Q2 and Q3 as these quadrants comprise of inconsistent shoppers. The goal of the campaign is to convert inconsistent shoppers to consistent shoppers. However, maximum value is achieved when the inconsistent shoppers are also high value customers.
Bubbles to target have been identified by their segment number and should be prioritized as: 21,31,11 and 41.

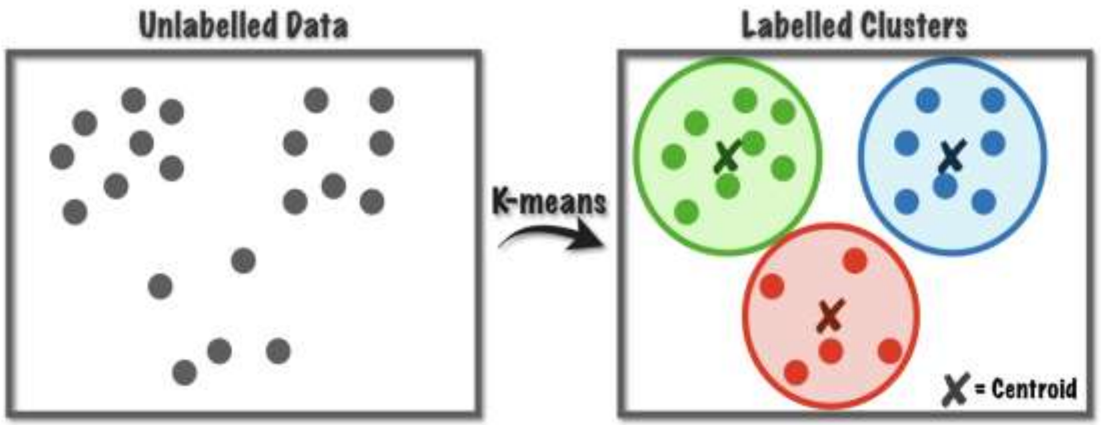
Test, Test, Test....

OVERVIEW OF K-MEANS CLUSTERING

- K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data
- It aims to partition *n* observations into k clusters in which each observation belongs to the cluster with nearest mean
- Used already computed R,F,M,C metrics to create and analyze the clusters

Why K-means?

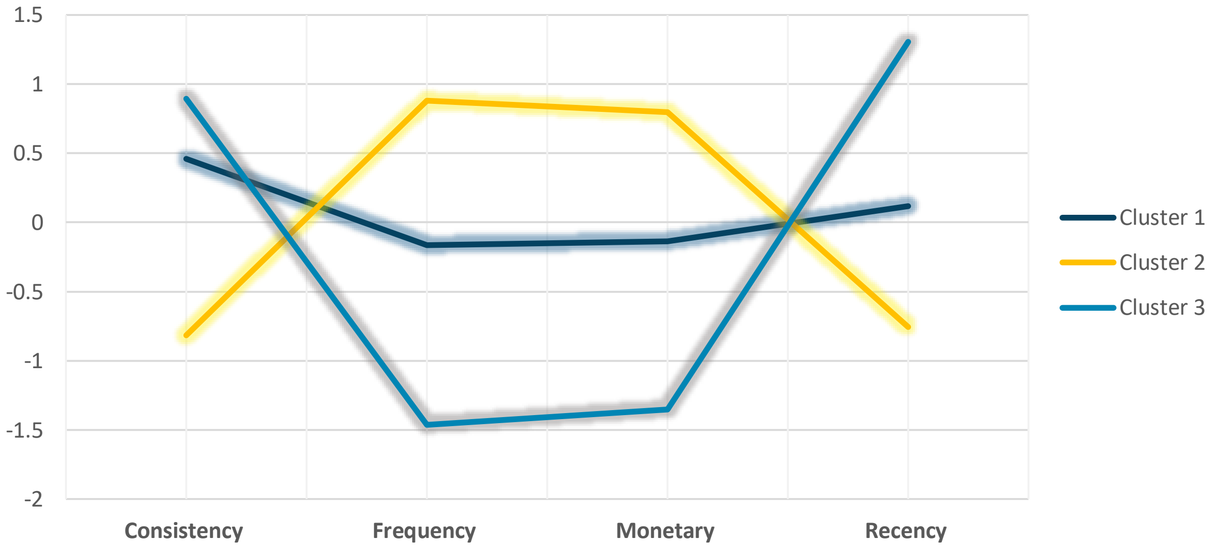
- It's fast and simple to implement
- It can be scaled to large data sets
- Easy to interpret and explain



K-means clusters

Result from first K-Means attempt

Snake plot of standardized variables



Summary table

Cluster	Recency (Mean)	Frequency (Mean)	Consistency (Mean)	Monetary (Mean)	Monetary (Count)
1	23..960376	38.60371	38.915357	899.723338	1489
2	5.486031	171.934634	6.284950	4584.283637	1897
3	213.664948	6.217526	105.884265	140.183247	970

- It is evident from the plot and the summary table that the cluster 2 is composed of our loyal customers, cluster 3 are the customers that are least frequent and inconsistent customers i.e. cluster that is showcasing churn behaviour and cluster 1 is the ideal cluster that we should be targeting in this campaign.

ROI ANALYSIS

ROI Calculation

Consistency + Monetary value

1

Segment	Resp	Cost	Spend
11	32	$486	$4
21	26	$393	$12
31	23	$347	$21
41	27	$408	$49
Grand Total	**109**	**$1634**	**$21**

Total Revenue	ROI
$128	-74%
$301	-23%
$480	38%
$1328	226%
$2238	**37%**

K-means Clustering

2

Segment	Resp	Cost	Spend
Cluster 1	148.9	$2233.5	$44.98

Total Revenue	ROI
$6698	200%

Target customers i.e. customers who visit inconsistently to the stores were selected from both consistency + monetary segmentation analysis and k-means clustering exercise as both showed promising ROI.

As expected, segment with inconsistent customers with strong monetary value generated higher ROI.

ROI breakdown for RFMC model isn't featured as the selected segment did not produce a viable return.

RECOMMENDATIONS FOR FURTHER ANALYSIS

Recommendation for future optimization

- ❑ Inconsistency is the key parameter to be considered while targeting customers for this marketing campaign, and our models validates that.
- ❑Further investigation required for the RFM model - as it’s parameters (reach, frequency) coupled with consistency can offer a robust solution.
- ❑Test and iterate K-mean again to find optimal clusters