MKT 4220 FINAL PRESENTATION

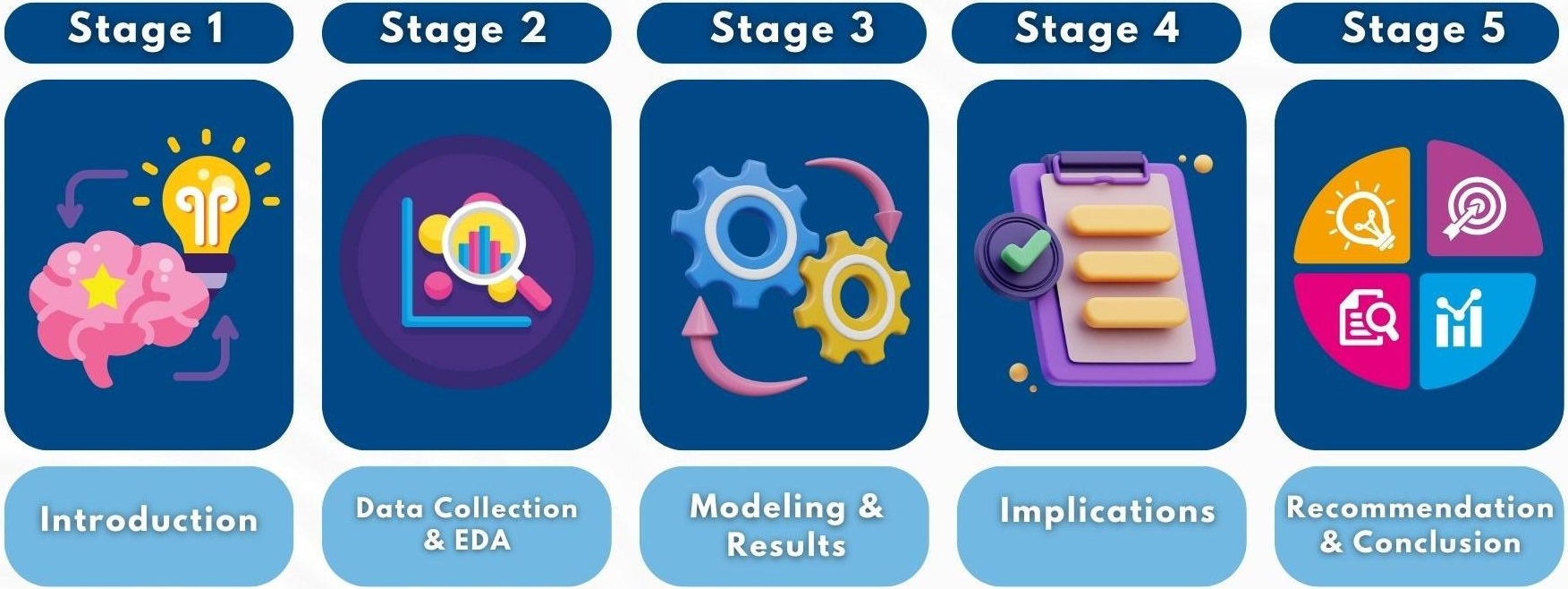
Unlocking Insight : A Comprehensive Customer Segmentation Study

An Empirical Study of Enhancing Decision Making through Segmentation Findings

Dharma Setiawan 伍铠佑
Josep Ariel Christopher Teja 黄熙
Darren Boesono 吴凡

PRESENTED BY GROUP 9

WORKFLOW ML

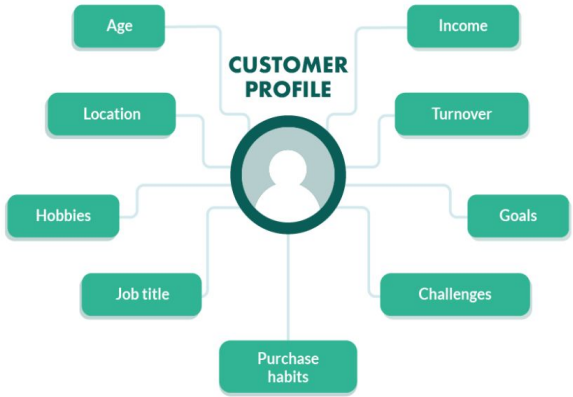

Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Introduction
Data Collection & EDA
Modeling & Results
Implications
Recommendation & Conclusion

Background

Change or Die? → Customer Segmentation

What is customer segmentation?

- Practice of dividing customers into different groups based on different marketing methods and specific perspectives.
- Used to predict customer behavior patterns. Once identified, it is possible to make predictions about the groups' responses to different situations, to align marketing strategies and types of policy, and to allow more creative and better targeted policies to emerge.
- Most commonly used attributes: location, age, gender, income, lifestyle, and previous purchasing behavior.



Background

Why customer segmentation is important?

Pareto principle: 20% of customers contribute more to the company's revenue than the rest. (Srivastava, 2016)

Delivering more targeted information

Marketing personnel can directly target different customer groups, delivering products and marketing information that better suit their interests and needs to resonate more.

Reasonably allocate marketing resources

After segmenting customers, brands can effectively allocate marketing resources to specific customer groups and maximize cross selling and upward sales opportunities.

Maintain better customer relationships

Help marketers better understand their customers, providing personalized customer service experiences to improve customer loyalty and retention.

Adjusting marketing budget

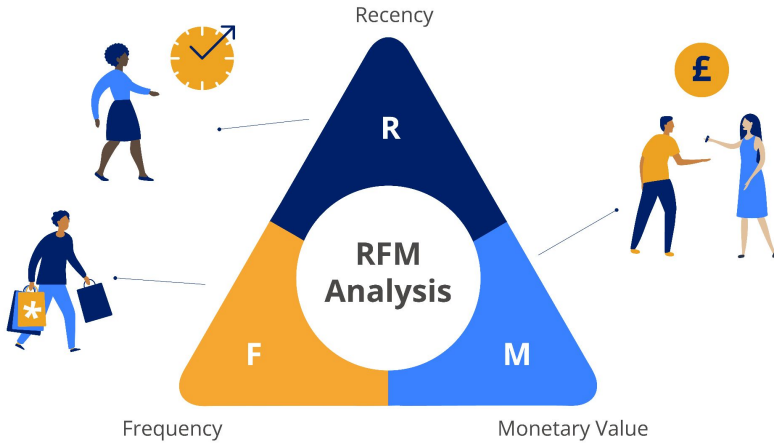
By analyzing its revenue and costs, the profit potential of each segmented market can be determined, helping the company determine the profit level of the segmented market, so that they can adjust their marketing budget accordingly.

RFM Analysis

RFM is a strategic approach employed in the analysis and estimation of a customer's worth, predicated upon the evaluation of three crucial data points, namely: **Recency, Frequency, and Monetary Value.**

Limitations of RFM

- It does not take into account key factors such as customer demographics or purchased items.
- More importantly, it depends entirely on the customer's history data, which means it may not be able to accurately predict the future activities of customers.
- Therefore, it is necessary to combine more advanced prediction methods.



Research Questions

Questions	Description
Question #1	Which is the best approach between RFM analysis, Clustering combined with classification, and a standalone classification model based on customer response to previous campaigns, in terms of their abilities to segment customers and provide actionable business insights for optimizing marketing strategies?
Question #2	Given our clusters, what segments do we get? How much the potential Gross Merchandise Value (GMV) and optimization cost? What is the personalized approach marketing strategy for each segment?
Question #3	How might a company implement a customer segmentation service in their internal stack?

Literature Review

Jiang and Tuzhilin (2009) proposed the K-Classifiers Segmentation algorithm, and found that this approach yielded far superior results compared to the traditional statistical approach.Shah and Singh (2012) proposed a new clustering algorithm which executes similar to the K-means algorithm and K-medoids algorithms. But the proposed algorithm does not provide an optimal solution in all cases. Zahrotun (2017) used the customer data from online to identify the finest customer using Customer Relationship Management (CRM).

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 3, MARCH 2009 305

Improving Personalization Solutions through Optimal Segmentation of Customer Bases

Tianyi Jiang and Alexander Tuzhilin

Abstract—On the Web, where the search costs are low and the competition is just a mouse click away, it is crucial to segment the customers intelligently in order to offer more targeted and personalized products and services to them. Traditionally, customer segmentation is achieved using statistics-based methods that compute a set of statistics from the customer data and group customers into segments by applying distance-based clustering algorithms in the space of these statistics. In this paper, we present a direct grouping-based approach to computing customer segments that groups customers not based on computed statistics, but in terms of optimally combining transactional data of several customers to build a data mining model of customer behavior for each group. Then, building customer segments becomes a combinatorial optimization problem of finding the best partitioning of the customer base into disjoint groups. This paper shows that finding an optimal customer partition is NP-hard, proposes several suboptimal direct grouping segmentation methods, and empirically compares them among themselves, traditional statistics-based hierarchical and affinity propagation-based segmentation, and one-to-one methods across multiple experimental conditions. It is shown that the best direct grouping method significantly dominates the statistics-based and one-to-one approaches across most of the experimental conditions, while still being computationally tractable. It is also shown that the distribution of the sizes of customer segments generated by the best direct grouping method follows a power law distribution and that microsegmentation provides the best approach to personalization.

Index Terms—Customer segmentation, marketing application, personalization, one-to-one marketing, customer profiles.

1 Introduction

Customer segmentation, such as customer grouping by the level of family income, education, or any other demographic variable, is considered as one of the standard techniques used by marketers for a long time [40]. Its popularity comes from the fact that segmented models usually outperform aggregated models of customer behavior [45]. More recently, there has been much interest in the marketing and data mining communities in learning *individual* models of customer behavior within the context of *one-to-one* marketing [37] and personalization [8], when models of customer behavior are learned from the data pertaining only to a particular customer. These learned individualized models of customer behavior are used as parts of customer profiles and are subsequently used for recommending and delivering personalized products and services to the customers [2].

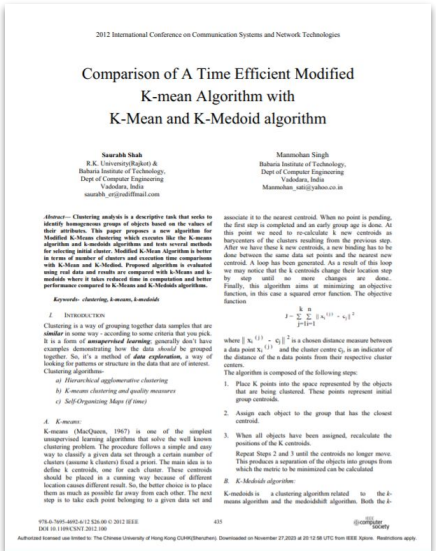
As was shown in [23], it is a nontrivial problem to compare segmented and individual customer models because of the tradeoff between the sparsity of data for individual customer models and customer heterogeneity in aggregate models: individual models may suffer from sparse data, while aggregate models suffer from high levels of customer heterogeneity. Depending on which effect dominates the other, it is possible that models of individual customers dominate the segmented or aggregated models, and vice versa.

A typical approach to customer segmentation is a distance-measure-based approach where a certain proximity or similarity measure is used to cluster customers. In this paper, we consider the *statistics-based* approach, a subclass of the distance-measure-based approaches to segmentation that computes the set of summary statistics from customer's demographic and transactional data [6], [23], [47], such as the average time it takes the customer to browse the Web page describing a product, maximal and minimal times taken to buy an online product, recency, frequency, and monetary (RFM values) statistics [34], and so forth. These customer summary statistics constitute statistical reductions of the customer transactional data across the transactional variables. They are typically used for clustering in the statistics-based approach instead of the raw transactional data since the unit of analysis in forming customer segments is the customer, not his or her individual transactions. After such statistics are computed for each customer, the customer base is then partitioned into customer segments by using various clustering methods on the space of the computed statistics [21]. It was shown in [21] that the best statistics-based approaches can be effective in some situations and can even outperform the *one-to-one* case under certain conditions. However, it was also shown in [21] that this approach can also be highly ineffective in other cases. This is primarily because computing different customer statistics would result in different h-dimensional spaces, and various distance metrics or clustering algorithms would yield different clusters. Depending on particular customer statistics, distance functions, and clustering algorithms, significantly different customer segments can be generated.

• *T. Jiang is with AvePoint Inc., 3 2nd Street, Suite 803, Jersey City, NJ 07302. E-mail: tjiang@stern.nyu.edu.*
• *A. Tuzhilin is with the Department of Information, Operations, and Management Sciences, Stern School of Business, New York University, 44 West 4th Street, Room 8-92, New York, NY 10012. E-mail: atuzhilin@stern.nyu.edu.*

Manuscript received 23 May 2007; revised 14 Feb. 2008; accepted 15 July 2008; published online 30 July 2008.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2007-05-0230.
Digital Object Identifier no. 10.1109/TKDE.2008.163.

1041-4347/09/$25.00 © 2009 IEEE Published by the IEEE Computer Society



2012 International Conference on Communication Systems and Network Technologies

Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm

Saurabh Shah
R.K. University(Rajkot) & Baharia Institute of Technology, Dept of Computer Engineering
Vadodara, India
saurabh_er@rediffmail.com

Manmohan Singh
Baharia Institute of Technology, Dept of Computer Engineering
Vadodara, India
Manmohan_sati@yahoo.co.in

***Abstract*— Clustering analysis is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. This paper proposes a new algorithm for Modified K-Means clustering which executes like the K-means algorithm and k-medoids algorithms and tests several methods for selecting initial cluster. Modified K-Mean Algorithm is better in terms of number of clusters and execution time comparisons with K-Mean and K-Medoid. Proposed algorithm is evaluated using real data and results are compared with k-Means and k-medoids where it takes reduced time in computation and better performance compared to K-Means and K-Medoids algorithms.**

Keywords- clustering, k-means, k-medoids

I. Introduction

Clustering is a way of grouping together data samples that are **similar** in some way - according to some criteria that you pick. It is a form of ***unsupervised learning***; generally don't have examples demonstrating how the data *should* be grouped together. So, it's a method of ***data exploration***, a way of looking for patterns or structure in the data that are of interest. Clustering algorithms-

a) Hierarchical agglomerative clustering
b) K-means clustering and quality measures
c) Self-Organizing Maps (if time)

A. K-means:

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2$$

where $\| x_i^{(j)} - c_j \|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centers.

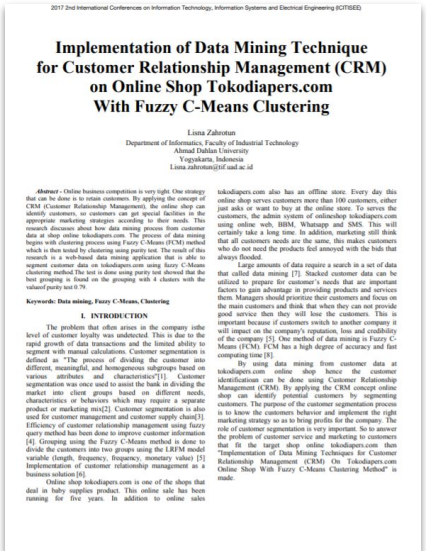
The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
   Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated

B. K-Medoids algorithm:

K-medoids is a clustering algorithm related to the k-means algorithm and the medoidshift algorithm. Both the k-

978-0-7695-4692-612 $26.00 © 2012 IEEE
DOI 10.1109/CSNT.2012.100



2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)

Implementation of Data Mining Technique for Customer Relationship Management (CRM) on Online Shop Tokodiapers.com With Fuzzy C-Means Clustering

Lisna Zahrotun
Department of Informatics, Faculty of Industrial Technology
Ahmad Dahlan University
Yogyakarta, Indonesia
Lisna.zahrotun@if.uad.ac.id

Abstract **- Online business competition is very tight. One strategy that can be done is to retain customers. By applying the concept of CRM (Customer Relationship Management), the online shop can identify customers, so customers can get special facilities in the appropriate marketing strategies according to their needs. This research discusses about how data mining process from customer data at shop online tokodiapers.com. The process of data mining begins with clustering process using Fuzzy C-Means (FCM) method which is then tested by clustering using purity test. The result of this research is a web-based data mining application that is able to segment customer data on tokodiapers.com using fuzzy C-Means clustering method.The test is done using purity test showed that the best grouping is found on the grouping with 4 clusters with the valued purity test 0.79.**

Keywords: Data mining, Fuzzy C-Means, Clustering

I. INTRODUCTION

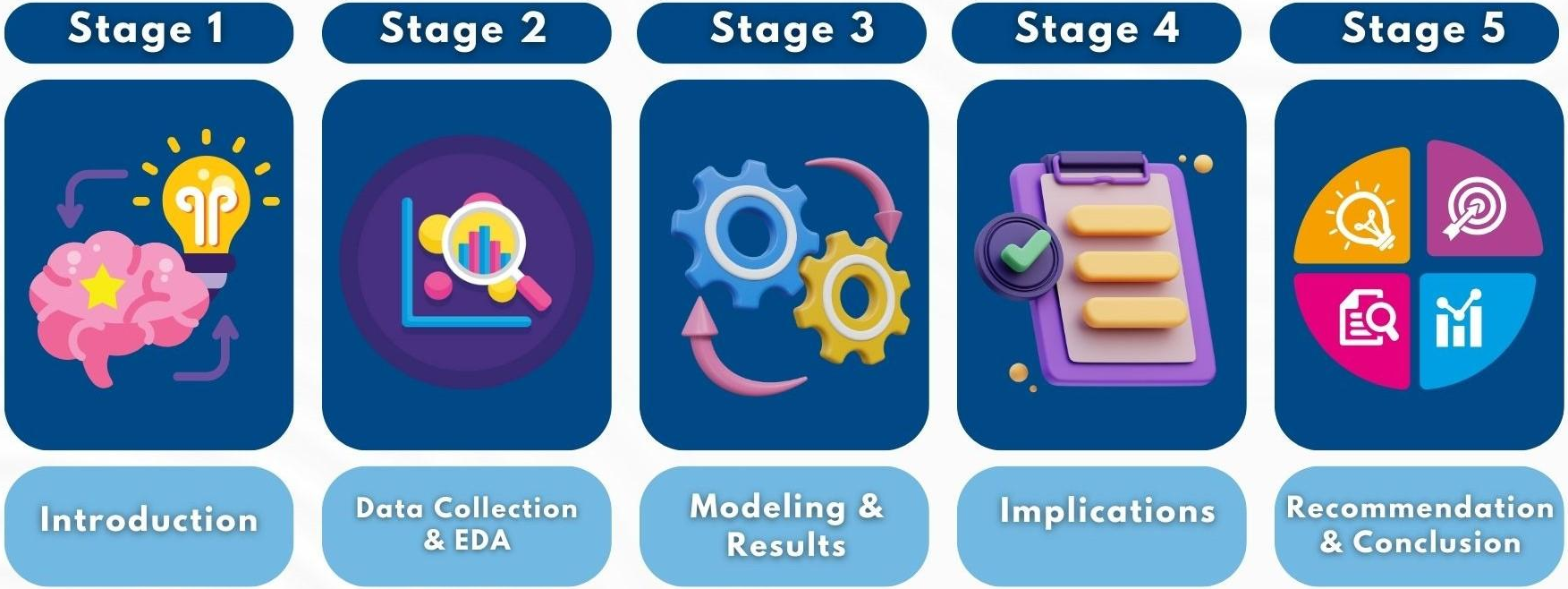
The problem that often arises in the company isthe level of customer loyalty was undetected. This is due to the rapid growth of data transactions and the limited ability to segment with manual calculations. Customer segmentation is defined as "The process of dividing the customer into different, meaningful, and homogeneous subgroups based on various attributes and characteristics"[1]. Customer segmentation was once used to assist the bank in dividing the market into client groups based on different needs, characteristics or behaviors which may require a separate product or marketing mix[2]. Customer segmentation is also used for customer management and customer supply chain[3]. Efficiency of customer relationship management using fuzzy query method has been done to improve customer information [4]. Grouping using the Fuzzy C-Means method is done to divide the customers into two groups using the LRFM model variable (length, frequency, frequency, monetary value) [5] Implementation of customer relationship management as a business solution [6].

Online shop tokodiapers.com is one of the shops that deal in baby supplies product. This online sale has been running for five years. In addition to online sales tokodiapers.com also has an offline store. Every day this online shop serves customers more than 100 customers, either just asks or want to buy at the online store. To serves the customers, the admin system of onlineshop tokodiapers.com using online web, BBM, Whatsapp and SMS. This will certainly take a long time. In addition, marketing still think that all customers needs are the same, this makes customers who do not need the products feel annoyed with the bids that always flooded.

Large amounts of data require a search in a set of data that called data mining [7]. Stacked out customer data can be utilized to prepare for customer's needs that are important factors to gain advantage in providing products and services them. Managers should prioritize their customers and focus on the main customers and think that when they can not provide good service then they will lose the customers. This is important because if customers switch to another company it will impact on the company's reputation, loss and credibility of the company [5]. One method of data mining is Fuzzy C-Means (FCM). FCM has a high degree of accuracy and fast computing time [8].

By using data mining from customer data at tokodiapers.com online shop hence the customer identifications can be done using Customer Relationship Management (CRM). By applying the CRM concept online shop can identify potential customers by segmenting customers. The purpose of the customer segmentation process is to know the customers behavior and implement the right marketing strategy so as to bring profits for the company. The role of customer segmentation is very important. So to answer the problem of customer service and marketing to customers that fit the target shop online tokodiapers.com then "Implementation of Data Mining Techniques for Customer Relationship Management (CRM) On Tokodiapers.com Online Shop With Fuzzy C-Means Clustering Method" is made.

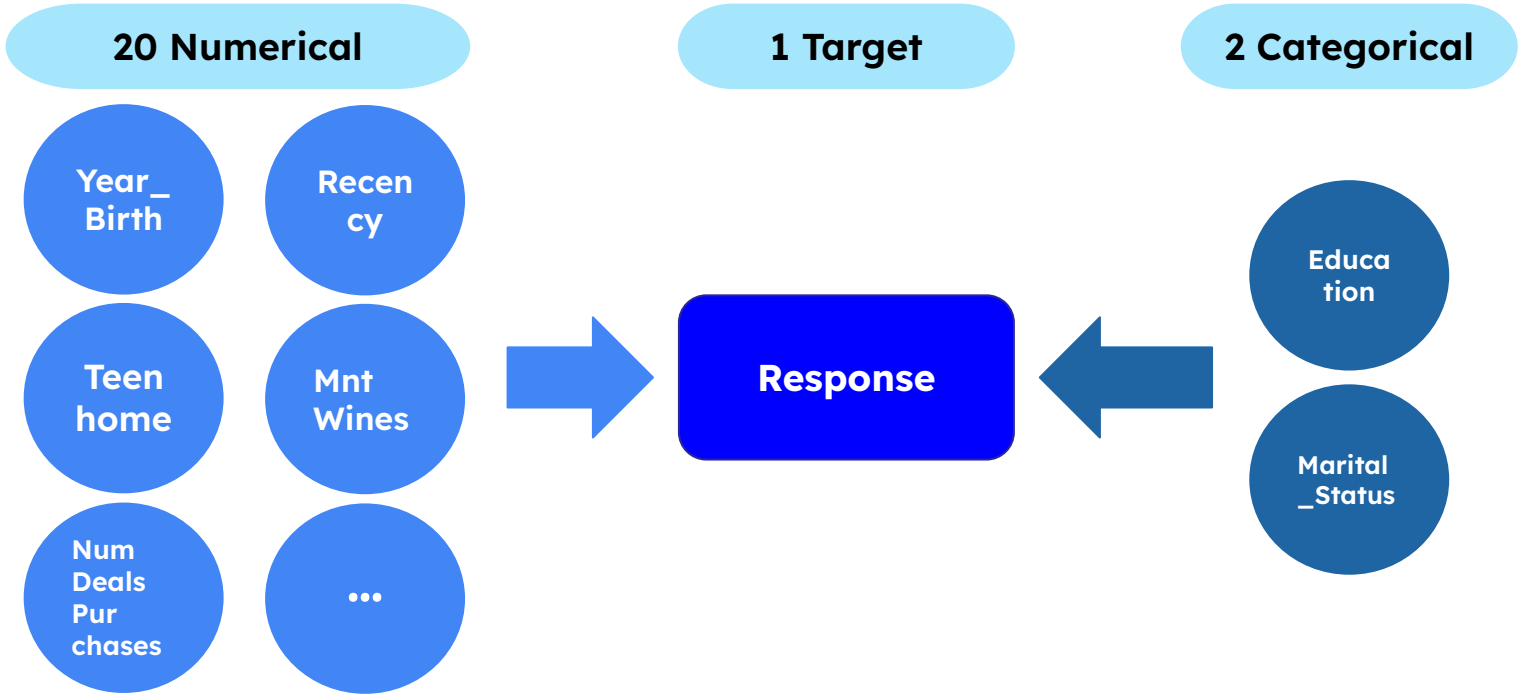
WORKFLOW ML


Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Introduction
Data Collection & EDA
Modeling & Results
Implications
Recommendation & Conclusion

Data Collection

https://www.kaggle.com/datasets/rodsaldanha/marketing-campaign/data

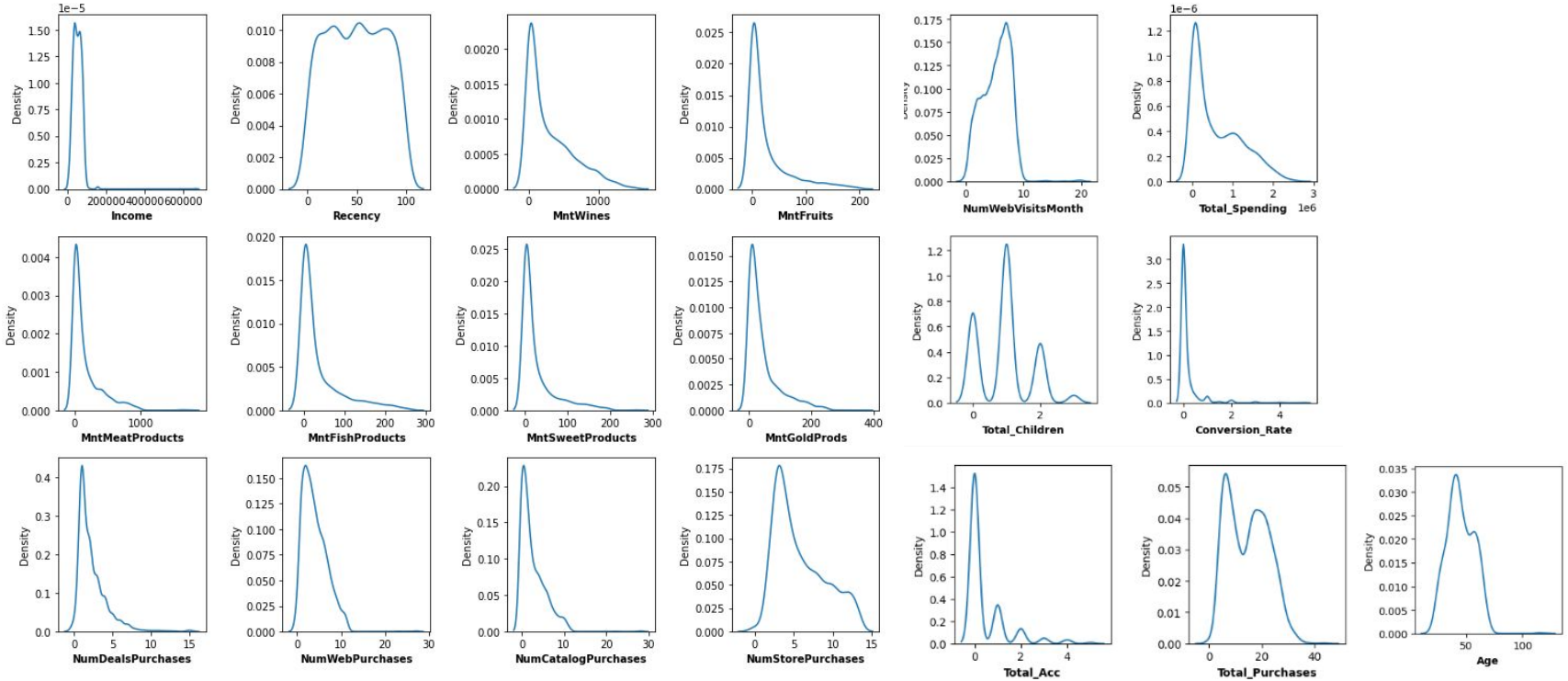
2240 Rows | 29 Features | 0 Duplicate | 24 Missing Values



EDA in General

Univariate Analysis: Distplot Analysis

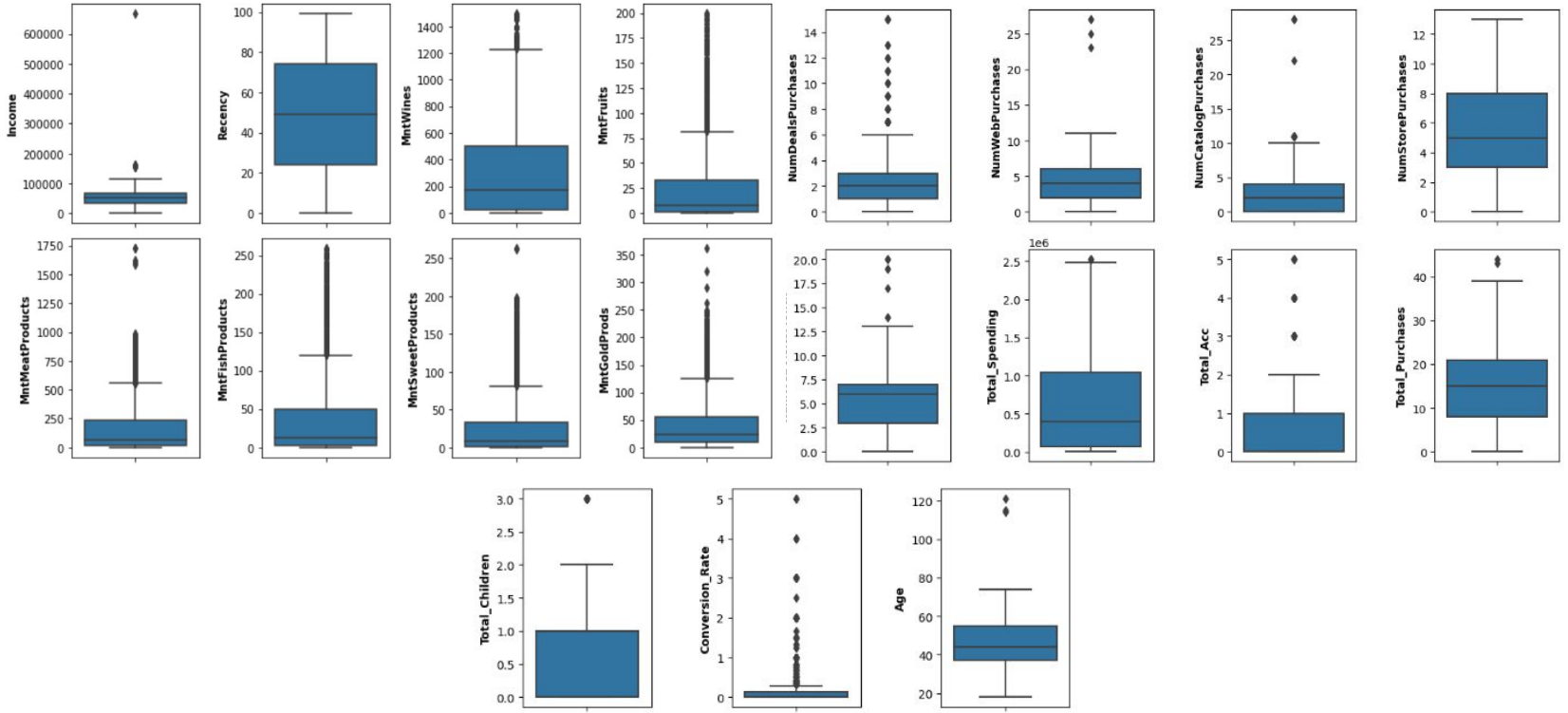
Most of the dataset has highly positively skewed distribution



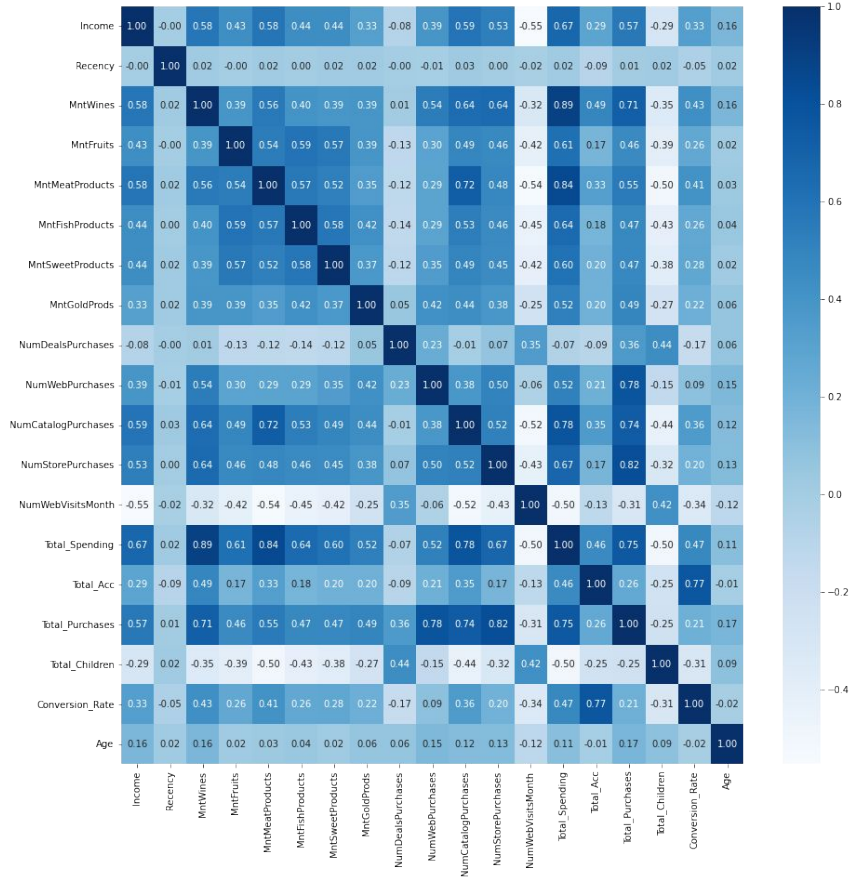
EDA in General

Univariate Analysis: Box Plot Analysis

There are some of outliers in each features and most features are not symmetric



EDA in General



Multivariate Analysis: Correlation Heatmap

- Customers are more likely to buy wine and meat product
- Customers tend to buy products directly in stores
- Number of visits to company's web site and number of children in customers' household contribute little to the response

Feature Engineering

Method	Description
customer_year & customer_month	
age & join_at_age	
total_children & is_parents	

```
#Convert date to date object
df["Dt_Customer"]= pd.to_datetime(df["Dt_Customer"])

#Create Month and Year variables from Dt_Customer
df["Customer_year"]= df["Dt_Customer"].dt.year
df["Customer_month"]= df["Dt_Customer"].dt.month

#Removing ID, Date from further analysis
df.drop(["ID"], axis= 1, inplace=True)
df.head()
```

```
# age_range
df['age'] = 2023 - df['Year_Birth']
df['join_at_age'] = df['Dt_Customer'].dt.year - df['Year_Birth']
df.loc[(df['age'] >= 0) & (df['age'] < 12), 'age_range'] = "child"
df.loc[(df['age'] >= 12) & (df['age'] < 18), 'age_range'] = "teens"
df.loc[(df['age'] >= 18) & (df['age'] < 36), 'age_range'] = "young_adults"
df.loc[(df['age'] >= 36) & (df['age'] < 55), 'age_range'] = "middle_aged_adults"
df.loc[(df['age'] >= 55), 'age_range'] = "older_adults"
```

```
df['total_children'] = df['Kidhome'] + df['Teenhome']
df['is_parents'] = np.where(df['total_children'] > 0, 1, 0)
```

Data Preprocessing Feature Transformation

Method	Description	Example

Normalization	• Hard feature scaling: change the range of the feature with the valid or exact range
	• Doesn't change the distribution of the data
		`scaler= MinMaxScaler()`
		`scaled = scaler.fit_transform(df_enc)`
		`df_enc_scaled= pd.DataFrame(scaled, columns= df_enc.columns)`
		`df_enc_scaled.head()`

Standardization	• Soft feature scaling: change the range of the features with not exact range
	• Change the distribution of the data to the normal distribution
	• Transformation will result in mean = 0 with sd = 1
		`from sklearn.preprocessing import StandardScaler`
		`df_standard = df.copy()`
		
		`for i in numerical_cols:`
		`  df_standard[i] = StandardScaler().fit_transform(df_standard[i].values.reshape(len(df_standard), 1))`
		
		`df_standard`

Data Preprocessing Feature Encoding

Method	Description	Example

Label Encoding	Label Encoding is a categorical feature change to numeric by giving a different number for each value is unique	`#label encoder`
		`mapping_education = {`
		`'Basic' : 0,`
		`'Graduation' : 1,`
		`'2n Cycle': 2,`
		`"Master":2,`
		`"PhD":3`
		`}`
		
		`df['Education_mapped'] = df['Education'].map(mapping_education)`
		
		`le= LabelEncoder()`
		`enc_cols= ["Marital_Status", 'age_range', 'is_parents']`
		`for c in enc_cols:`
		`df_enc[c]= le.fit_transform(df_enc[c])`
		
		`df_enc.drop(["Education"], axis= 1, inplace=True)`

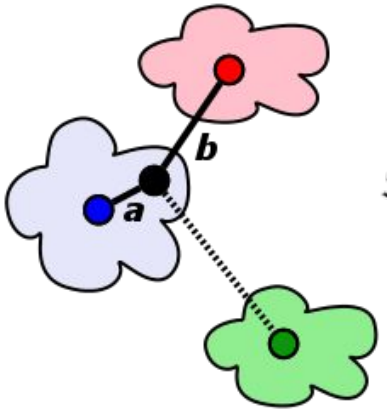
Model Evaluation (Silhouette Score)

Silhouette coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

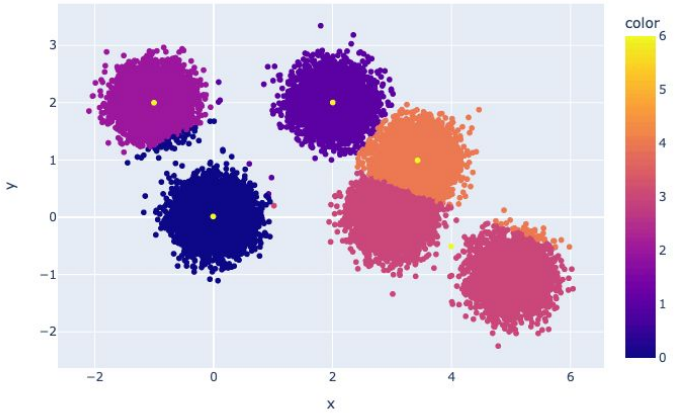
1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

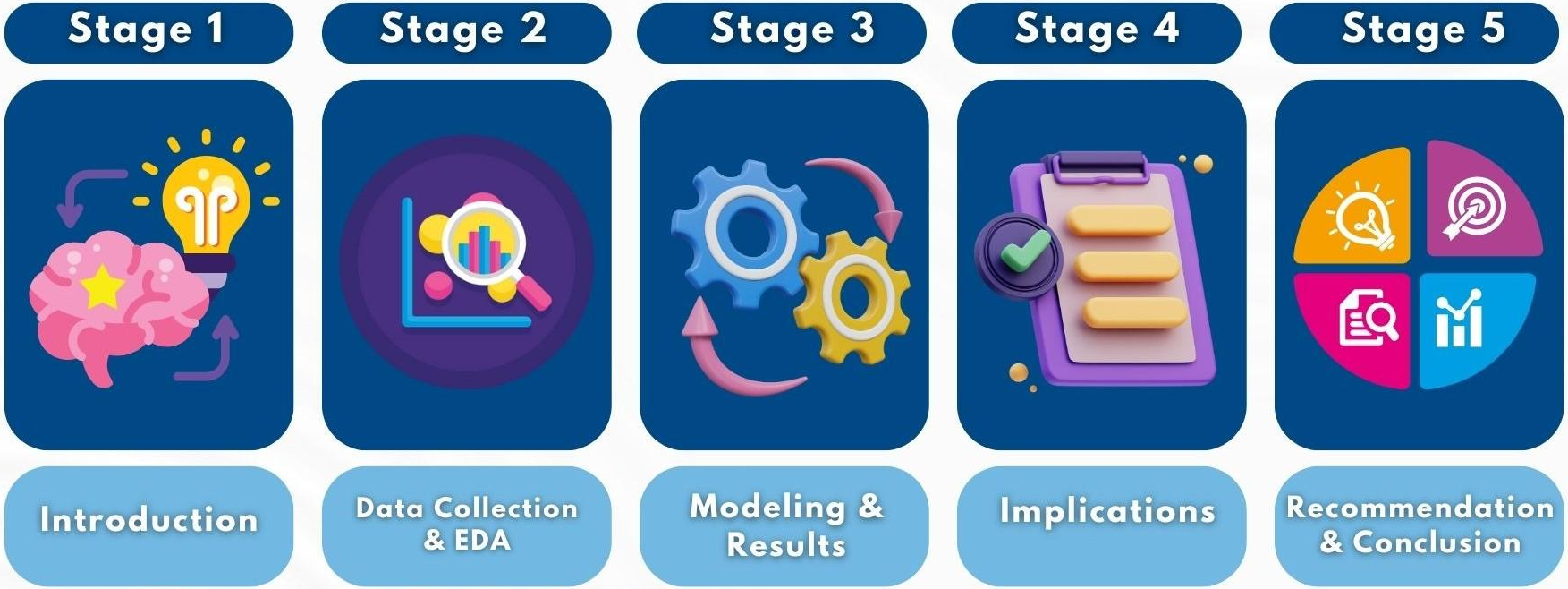
-1: Means clusters are assigned in the wrong way.



$$SSI_i = \frac{b_i - a_i}{max(a_i, b_i)}$$



WORKFLOW ML


Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Introduction
Data Collection & EDA
Modeling & Results
Implications
Recommendation & Conclusion

RFM Architecture

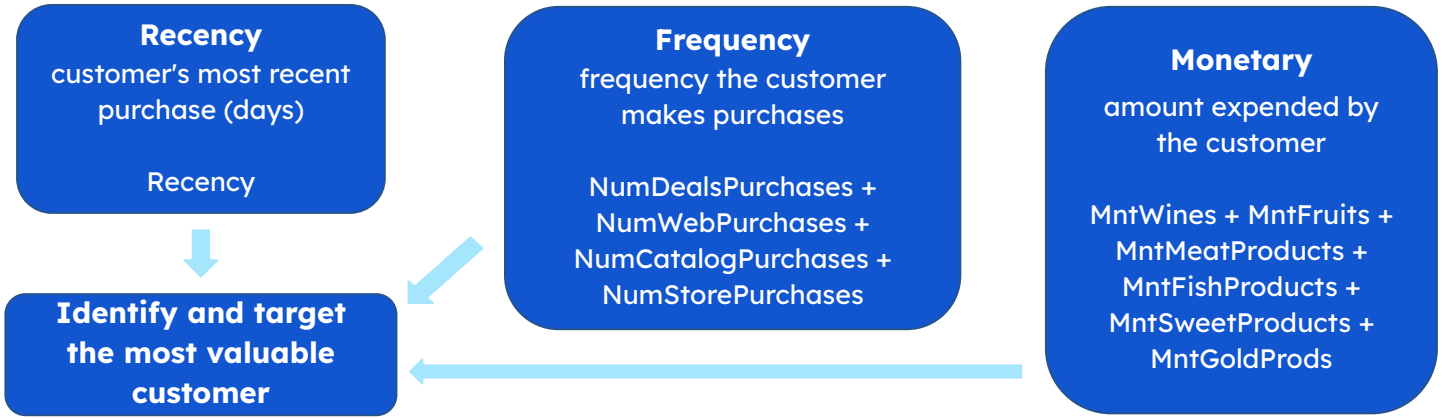
Motivation

RFM is a straightforward and highly effective way to comprehend and assess consumer behavior predicated on purchase.

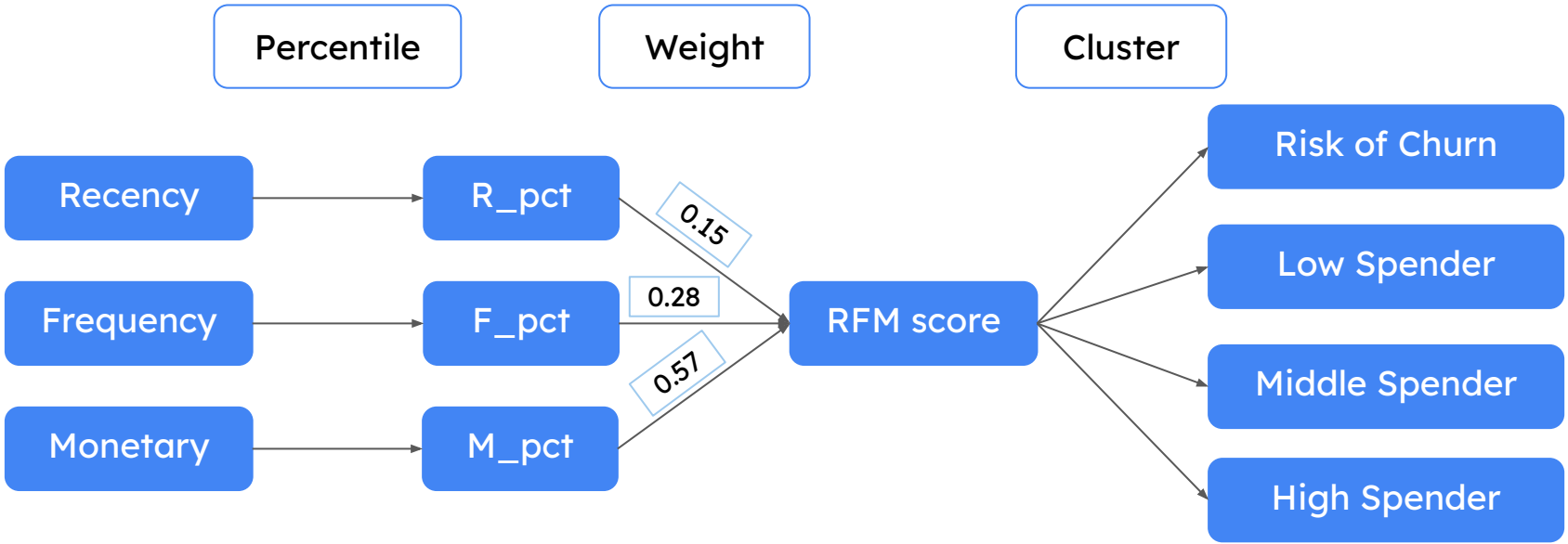
Idea

It functions by quantitatively categorizing and classifying patrons based on the recency, frequency, and monetary total of their most recent transactions, with the end goal of identifying and targeting the most valuable customers for the purposes of performing focused and precision-targeted marketing campaigns.

Implementation



RFM Scores

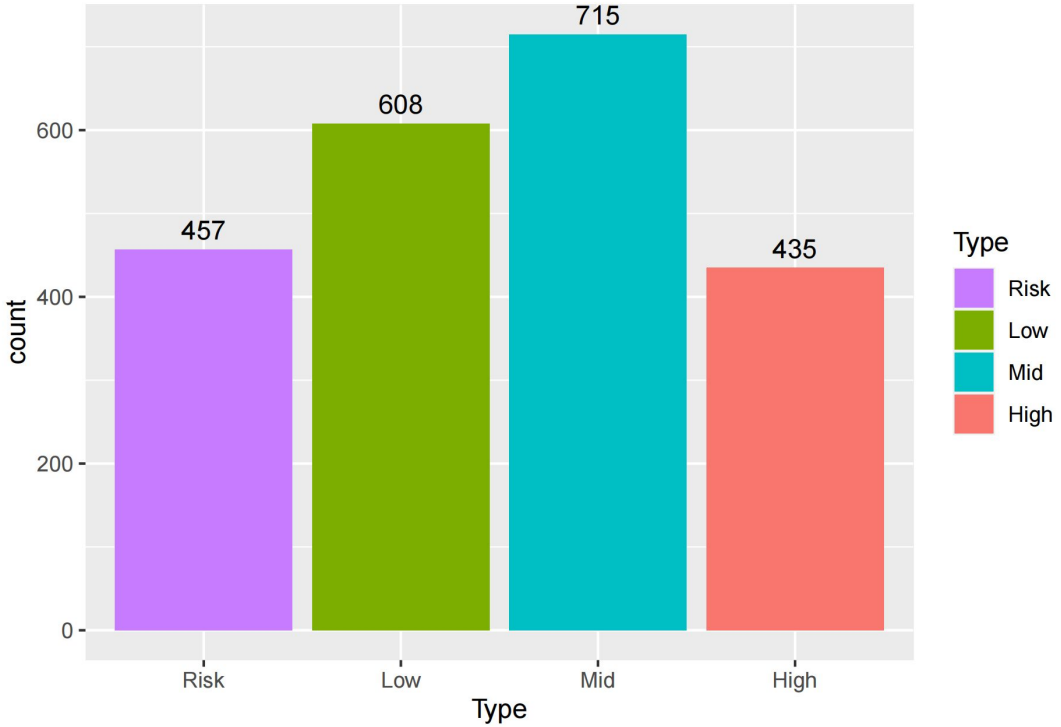


For clustering convenience, we assume RFM score = RFM score*4

Cluster criteria:

RFM score <= 1, Risk of Churn
1 < RFM score <=2, Low Spender
2 < RFM score <= 3, Middle Spender
3 < RFM score, High Spender

RFM Results



After calculating the RFM score and clustering, we have

457 risk of churn

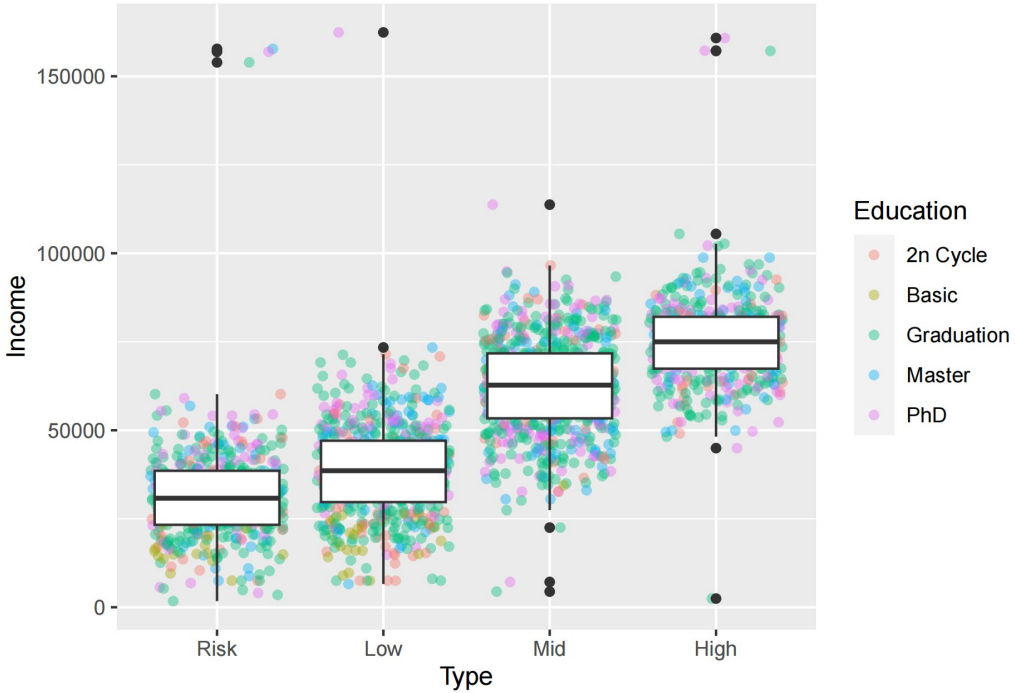
608 low spenders

715 middle spenders

435 high spenders

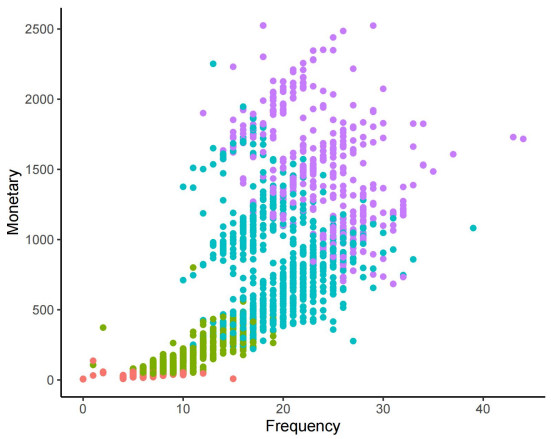
Silhouette score: 0.528

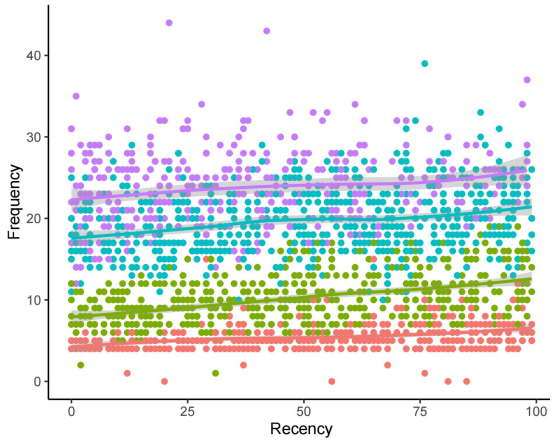
RFM Cluster Interpretation

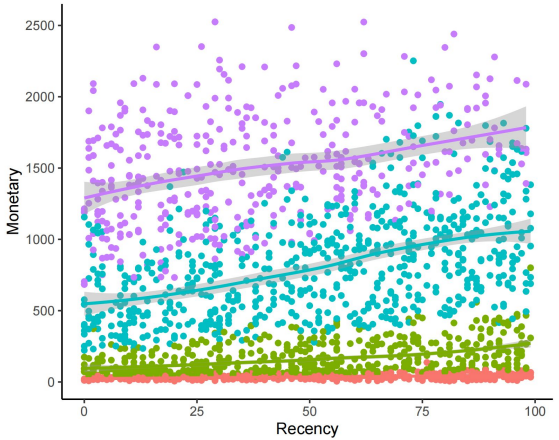


- High spenders typically have higher income, while low spenders usually have lower income.
- Type of customer is education independent.

RFM Cluster Interpretation







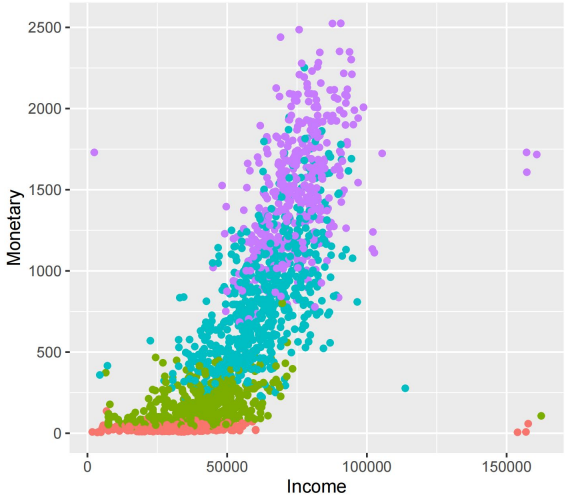
	Frequency	Monetary
Recency	Corr: 0.007 High: 0.167*** Low: 0.437*** Mid: 0.250*** Risk: 0.298***	Corr: 0.020 High: 0.334*** Low: 0.435*** Mid: 0.472*** Risk: 0.404***
Frequency	Corr: 0.756*** High: −0.359*** Low: 0.782*** Mid: 0.044 Risk: 0.452***	

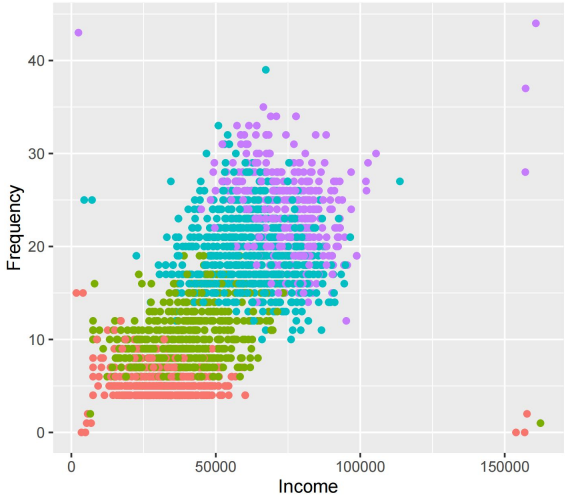
" *** " $p < 0.001$ " ** " $p < 0.01$ " * " $p < 0.05$ " . " $p < 0.10$ " " otherwise

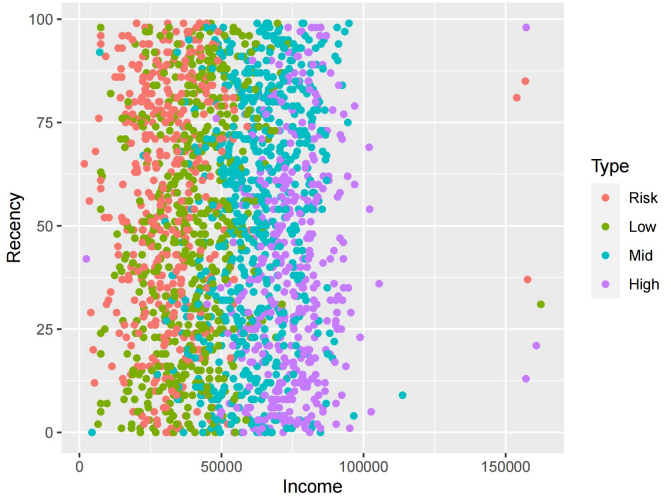
To avoid Simpson's Paradox, we stratify the data with customer segments

- High spenders typically have higher frequency and monetary, but recency can be varied.
- Frequency and recency; monetary and recency are significantly positively correlated within customer segments.
- Although monetary and frequency seem to be positively correlated, they are significantly negatively correlated within high spender segment.

RFM Customer Interpretation







- Customers with higher income may not necessarily spend a lot on purchases
- But low monetary customers generally have lower incomes

- Generally, customers with higher income purchase products more frequently

- The correlation between recency and income is not significant
- High spenders own higher income

Clustering + Classification Architecture

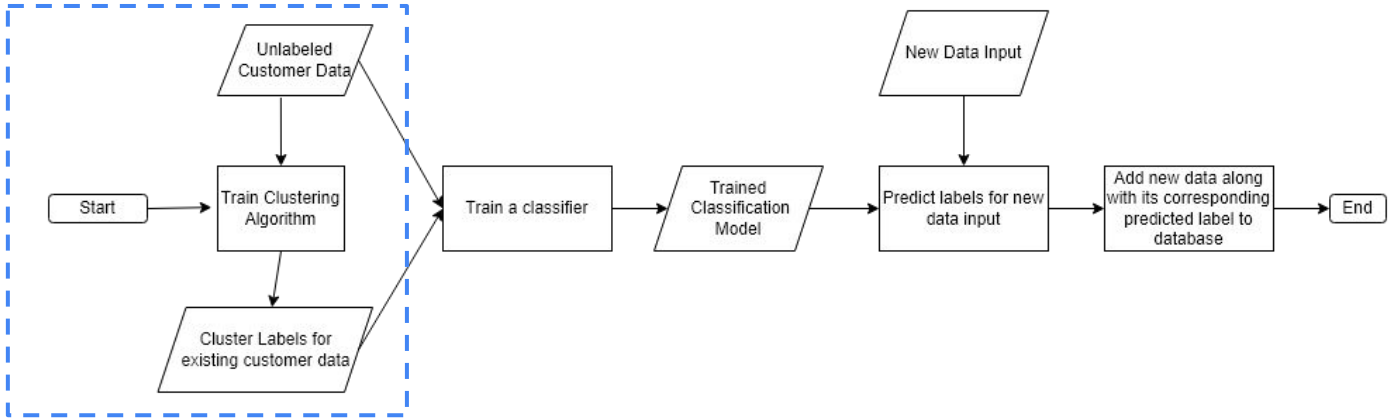
Motivation

Although clustering can be done to segment existing customers, not all clustering algorithms are able to efficiently cluster new data points (future customer data).

Idea

Train a robust classifier using the clustered customer data to efficiently segment new customers.

Implementation



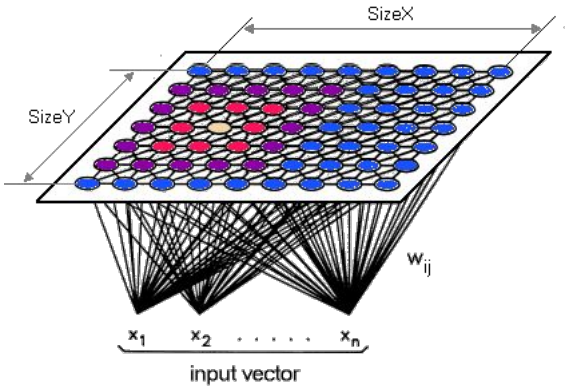
Self-Organizing Map (SOM), (Kohonen)

- It is an **unsupervised neural network** that is trained using unsupervised learning techniques to produce a low dimensional, discretized representation from the input space of the training samples, known as a map and is a method to reduce data dimensions.

- Unlike other neural networks, SOMs apply **competitive learning technique** instead of error-correction learning methods like backpropagation with gradient descent.

- Unlike other clustering techniques, SOM uses a neighbourhood function to **preserve the topological properties** within the input space

- Intuitively, you can think of a SOM as something like k-means where the k-centroids are constrained in a 2D manifold.



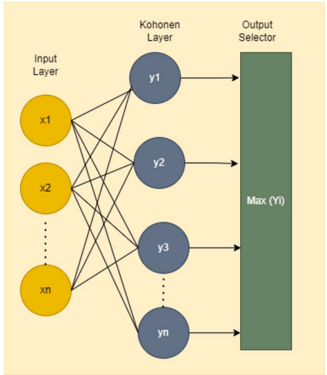
SOM as a Neural Network

- There are two layers: 1. **Input Layer** and 2. **Kohonen (Competitive) Layer** which acts as an output node

- The **Input Layer** is responsible for receiving input data and transmitting it to the second layer.

- The **Kohonen Layer** acts as the output layer where each neuron competes to become the winner.
  - Neurons in this layer are interconnected, forming a competitive network by restricting feedback.
  - Neurons have **excitatory links to immediate neighbors**, fostering local competition, and **inhibitory connections to distant neurons**, preventing excessive influence.
  - The layer uses a "**winner takes all**" approach, where the most responsive neuron to a given input set becomes the sole winner, generating a binary output (1 for the winner, 0 for others).

- Unlike feed forward and back propagation neural networks, SOM does not use threshold or activation function, nor does it have any hidden layers.



SOM: Training Process

- Parameters: Map size, Neighbourhood size, learning rate, iterations
- SOMs are trained with neighbourhood size, update parameters and weight.

Training Steps:

1. Network Initialization: Initialize $w_{ij}(t)$ $(0 \le i \le n-1)$ is the weight from i^{th} input node to j^{th} node at iteration t. the n input nodes are assigned with the weights randomly with smaller value. The neighborhood radius around the node j is $N_j(0)$.

2. New Input Initialization: input $X_0(t)$, $X_1(t)$, $X_2(t)$, ...$X_{n-1}(t)$ where $X_i(t)$ is the input from node to node i at iteration t.

3. Distance calculation: the distance d_j between input and j^{th} output node is computed using the equation

$$d_j = \sum_{i=1}^{n-1} (X_i(t) - w_{ij}(t))^2$$

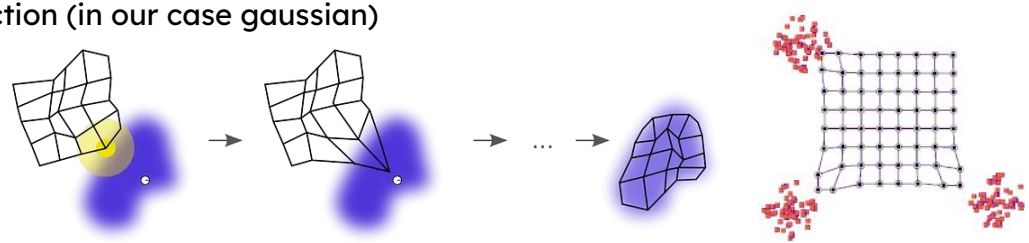
Minimum distance is selected as the winning node with the comparison and it is selected as the output node j'.

4. Weight update: the output node j' weight and its neighborhood node weights are updated using the following equation

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t).\, h_{ij}(t)[X_i(t) - w_{ij}(t)]$$

This process is repeated for j in $N_{j'}(t)$ $(0 \le i \le n-1)$, where $N_{j'}(t)$ is neighborhood size of output neuron j', η is learning rate in the range [0,1] which decreases over the time and the N value also decreases over time with the localization of maximum activity. h_{ij} is the neighbourhood function (in our case gaussian)

5. Until maximum iteration repeat step 2-4



SOM: Methods & Evaluation Metrics

Treatments

- There are two ways in which we used SOM for our case
  1. Directly using it as a **clustering algorithm**.
  2. Uses it as a **dimensionality reduction technique** followed by the hierarchical agglomerative clustering algorithm for clustering into 4 clusters.

- Due to unsatisfactory initial results, we explored the impact of reducing feature dimensionality on algorithm performance. We repeated both methods, limiting features to 4 which are:
  - Income, Recency, Total transaction, Total amount spent.

Parameters

- Neighbourhood Size = 0.5 ; Learning Rate = 0.5 ; Neighbourhood Function = Gaussian ; Maximum number of iterations = 1000
- Map Size: Treatment 1 = (4,1) ; Treatment 2 = (20,20)

Evaluation Metrics

- We used three evaluation metrics to measure model performance

  1. **Quantization error:** Measures the distance from each data vector to its best matching unit.
  2. **Topographic error:** Measures how well the topographic structure of the data is preserved on the map.
  3. **Silhouette score:** Used to get a reference for comparison with the other algorithms

SOM: Results

Treatment	Quantization Error	Topographic Error	Silhouette Score
SOM + No Feature Trimming	1.084	0.69	0.16
SOM + Agglomerative + No Feature Trimming	0.597	0.97	0.143
SOM + Feature Trimming	0.229	0.84	0.387
SOM + Agglomerative + Feature Trimming	0.038	0.98	0.372

SOM: Remarks

1. Contrary to literature, SOM seems to perform quite badly on this particular dataset.

2. The poor performance of SOM in our case is likely due to the large number of categorical variables and outliers in our dataset.

3. Feature reduction increases silhouette score and reduces quantization error.

4. Having a larger map size reduces quantization error but increases topographic error.

Modeling

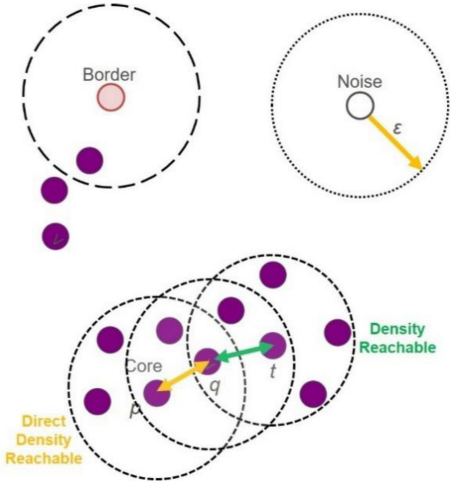
Model	Description	Silhouette Score
K-Means	K-Means is a centroid-based clustering algorithm that partitions data into K clusters, assigning each data point to the cluster with the nearest centroid.	0.1666
K-Means PCA	K-Means with PCA helps to handle high-dimensional data and improve clustering performance.	0.5786
K-Means t-SNE	K-Means with t-SNE incorporates t-SNE dimensionality reduction to visualize and cluster data points in a lower-dimensional space, preserving local structures.	0.6723
Mini Batch K-Means	Mini Batch K-Means is a variant of K-Means that processes random subsets (mini-batches) of the data at each iteration, making it computationally more efficient for large datasets.	0.6702
Agglomerative Clustering	Agglomerative Clustering is a hierarchical clustering algorithm that starts with individual data points and progressively merges them into clusters based on similarity, forming a tree-like structure.	0.6681
DBSCAN	DBSCAN is a density-based clustering algorithm that identifies clusters based on the density of data points, discovering clusters of varying shapes and handling noise effectively.	0.6651
OPTICS	OPTICS (Ordering Points To Identify the Clustering Structure) is a density-based clustering algorithm that extends DBSCAN by providing a hierarchical view of clusters and identifying varying densities within the data.	0.7199
BIRCH	BIRCH is a hierarchical clustering algorithm designed for large datasets, using a tree structure to efficiently group data points and reduce memory requirements.	0.6714

DBSCAN

The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

Algorithm:

1. Find all the neighbor points within eps (defines the neighborhood around a data point) and identify the core points or visited with more than MinPts (minimum number of neighbors within eps radius) neighbors.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density-connected points and assign them to the same cluster as the core point.
4. A point a and b are said to be density connected if there exists a point c which has a sufficient number of points in its neighbors and both points a and b are within the eps distance. This is a chaining process. So, if b is a neighbor of c, c is a neighbor of d, and d is a neighbor of e, which in turn is neighbor of a implying that b is a neighbor of a.
5. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.



DBSCAN vs K-Means

DBSCAN

- In DBSCAN we do not need to specify the number of clusters
- Clusters formed in DBSCAN can be of any arbitrary shape
- DBSCAN can work well with datasets having noise and outliers
- In DBSCAN two parameters are required for training the model

K-Means

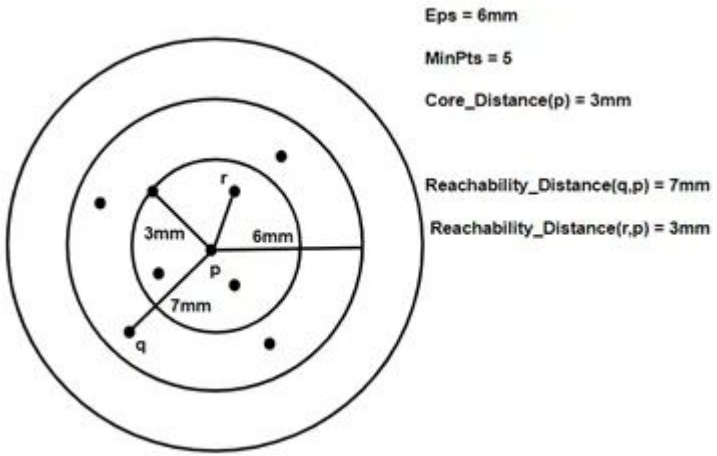
- K-Means is very sensitive to the number of clusters so it needs to be specified
- Clusters formed in K-Means are spherical or convex in shape
- K-Means does not work well with outliers data. Outliers can skew the clusters in K-Means to a very large extent
- In K-Means only one parameter is required for training the model

OPTICS

Ordering Points To Identify Cluster Structure (OPTICS) is a density-based clustering technique that allows partitioning data into groups with similar characteristics (clusters). It addresses one of the DBSCAN's major weaknesses. The problem of detecting meaningful clusters in data of varying density. In a density based clustering, clusters are defined as dense regions of data points separated by low-density regions.

OPTICS adds two more terms to the concepts of DBSCAN clustering:

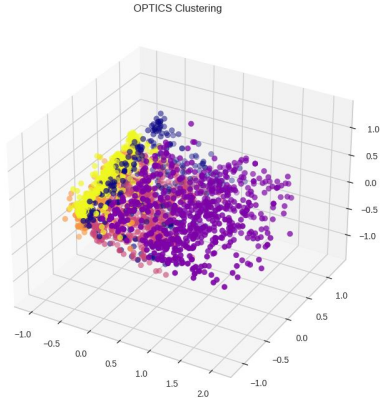
1. **Core Distance:** it is the minimum value of radius required to classify a given point as a core point. if the given point is not a Core point, then its Core Distance is undefined.
2. **Reachability Distance**: it is defined with respect to another data point 'q'. The Reachability distance between a point p and q



Best Model

```
OP = OPTICS(eps=0.4,min_samples = 250)
```

- OPTICS produces a hierarchical clustering result, providing a more **comprehensive view** of the data structure. This hierarchical information can be valuable for understanding varying densities and shapes of clusters within the dataset.
- **Robust to noise and outliers** in the data. It can identify clusters of varying shapes and sizes while gracefully handling data points that do not belong to any cluster.
- Well-suited for datasets with **clusters of varying densities**. It can adapt to clusters that have different levels of density, making it effective in scenarios where traditional methods like K-Means may struggle

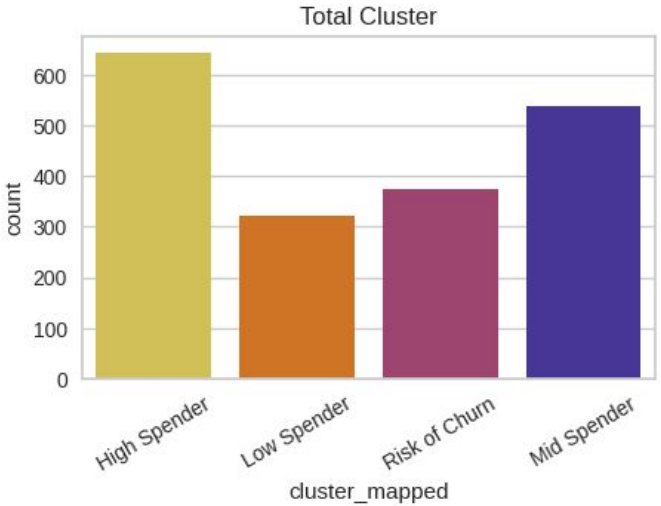


Best Model

```
OP = OPTICS(eps=0.4,min_samples = 250)
```

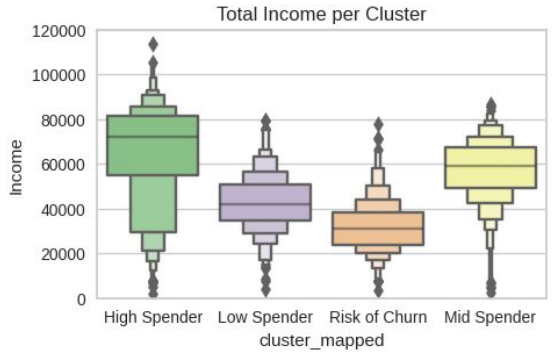
- eps: We set the number of estimators to 0.4. This parameter controls the maximum distance between two samples for one to be considered as in the neighborhood of the other. In OPTICS, it is often referred to as the "reachability distance." The eps parameter influences the sensitivity of the algorithm to the density of clusters. Smaller values of eps result in more fine-grained clusters. **We chose a relatively smaller number of estimators to ensure that the model has better separation in cluster whilst still ensuring that we are not making it too small which will cause most of the data points to be treated as noise. Setting eps as 0.4 results in only about 10% of data points to be treated as noise.**
- min_samples: We set the min_samples to 250. This parameter specifies the minimum number of samples in a neighborhood for a data point to be considered a core point. A core point is a point that has at least min_samples within its epsilon neighborhood. Increasing min_samples can lead to a more robust identification of dense regions, filtering out smaller clusters and noise. **We chose a moderate value of 250 to balance the model's capacity to capture complex patterns while avoiding overfitting.**

OPTICS Cluster Interpretation

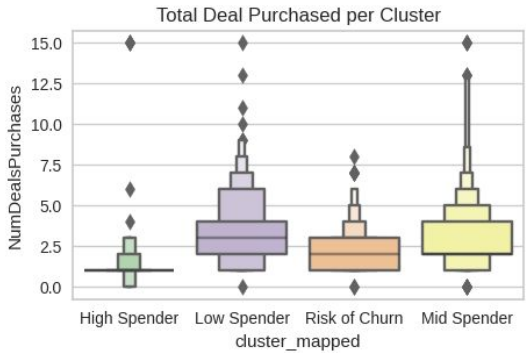


	total_amount_spent							
	count	mean	std	min	25%	50%	75%	max
cluster_optics								
0	643.0	1113.525661	645.533234	6.0	647.00	1193.0	1599.00	2525.0
1	539.0	679.024119	453.989709	15.0	305.00	653.0	1001.00	2092.0
2	322.0	141.897516	170.814355	8.0	38.25	68.0	185.75	1019.0
3	374.0	118.919786	158.494218	5.0	37.00	57.0	119.75	960.0

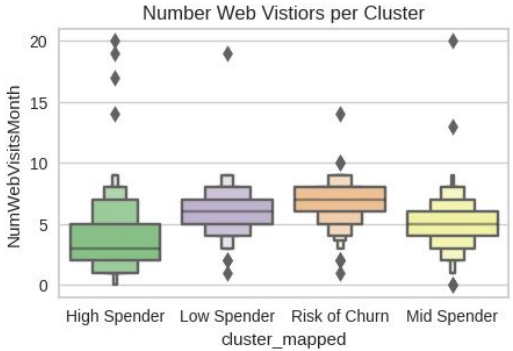
OPTICS Cluster Interpretation



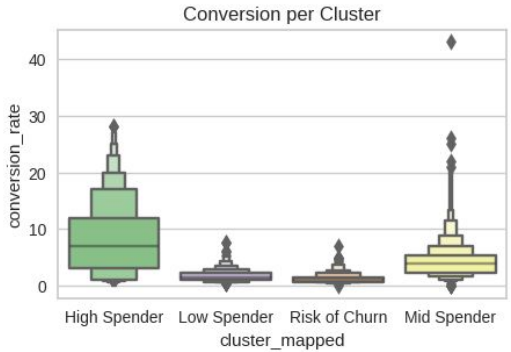
High spender customers have the highest income as expected



High spender customers rarely make purchases with discounts.

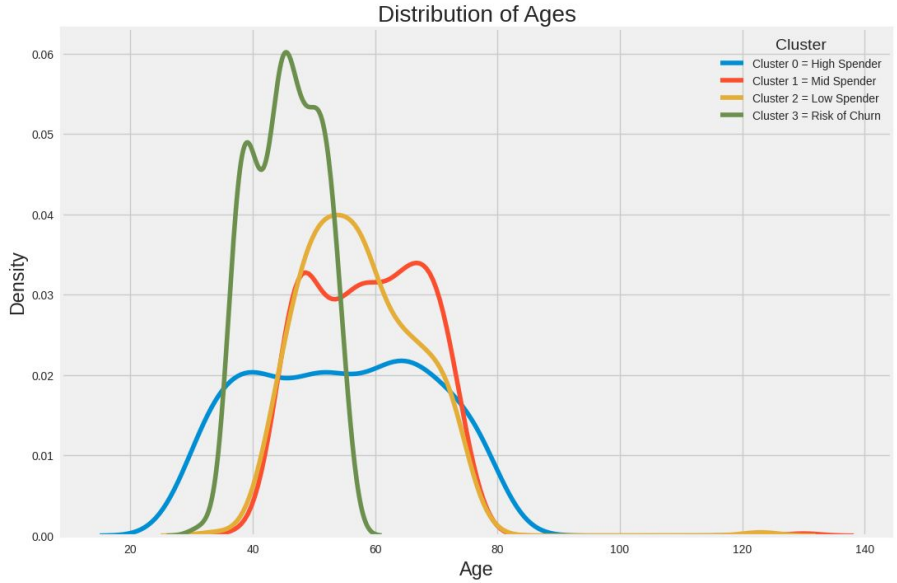


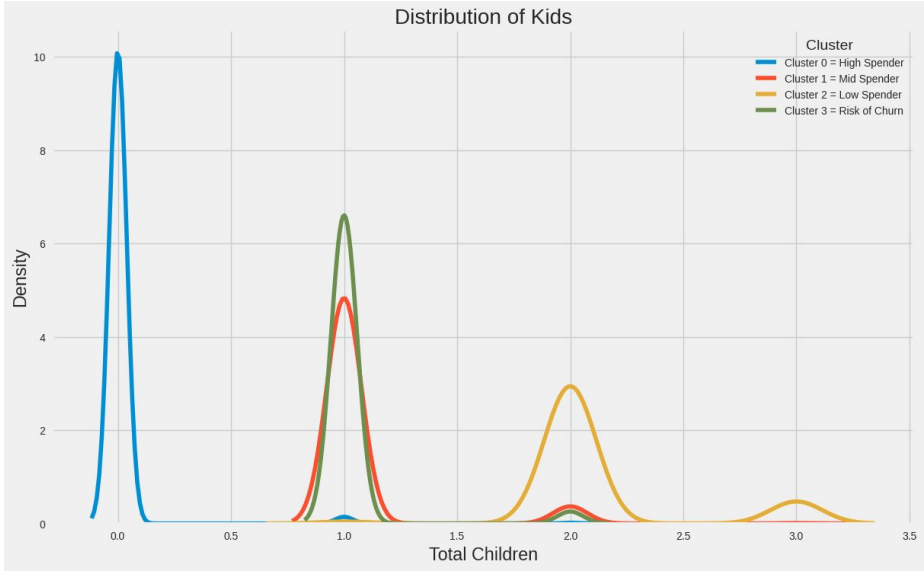
However, High spender customers rarely visit the web.



High spender customers also have high conversion rate.

OPTICS Cluster Interpretation





High spender customers are highly distributed within their age range. The most important pattern is that young generations (Gen Y) are most likely to churn. When they don't like the product anymore, they will go.

Almost all the high spender customers do not have kids. On the other side, low spender customers have from 2-3 kids.

Clustering + Classification Architecture

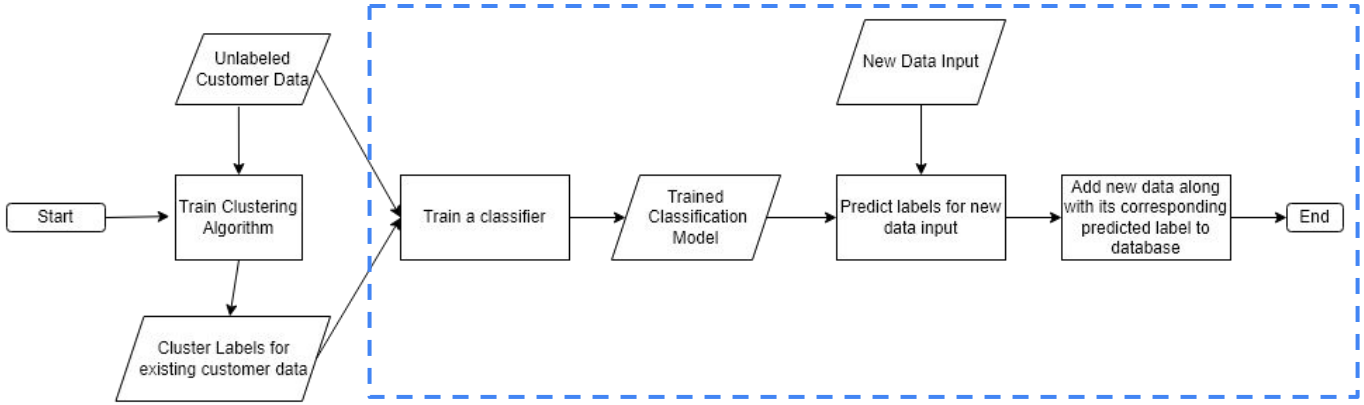
Motivation

Although clustering can be done to segment existing customers, not all clustering algorithms are able to efficiently cluster new data points (future customer data).

Idea

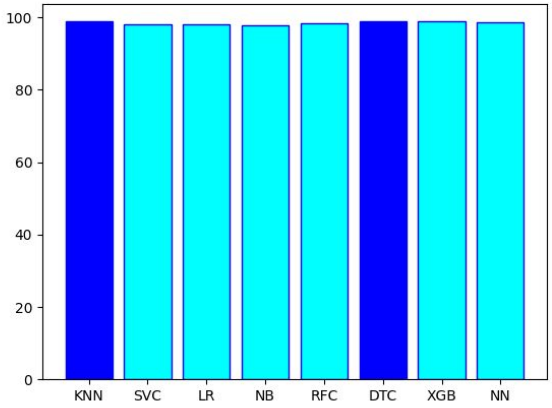
Train a robust classifier using the clustered customer data to efficiently segment new customers

Implementation



Classification for New Data Point Results

Model Names	Accuracy
K-Nearest Neighbours	98.91%
Support Vector Machine (Linear Kernel)	98.23%
Logistic Regression	98.07%
Naive Bayes	97.92%
Random Forest Classifier	98.43%
Decision Tree Classifier	98.91%
XGBoost Classifier	98.89%
Sequential Neural Network (1 Hidden Layer)	98.77%




```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 15)                570

 dense_1 (Dense)             (None, 15)                240

 dense_2 (Dense)             (None, 4)                 64

=================================================================
Total params: 874 (3.41 KB)
Trainable params: 874 (3.41 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Neural Network Structure

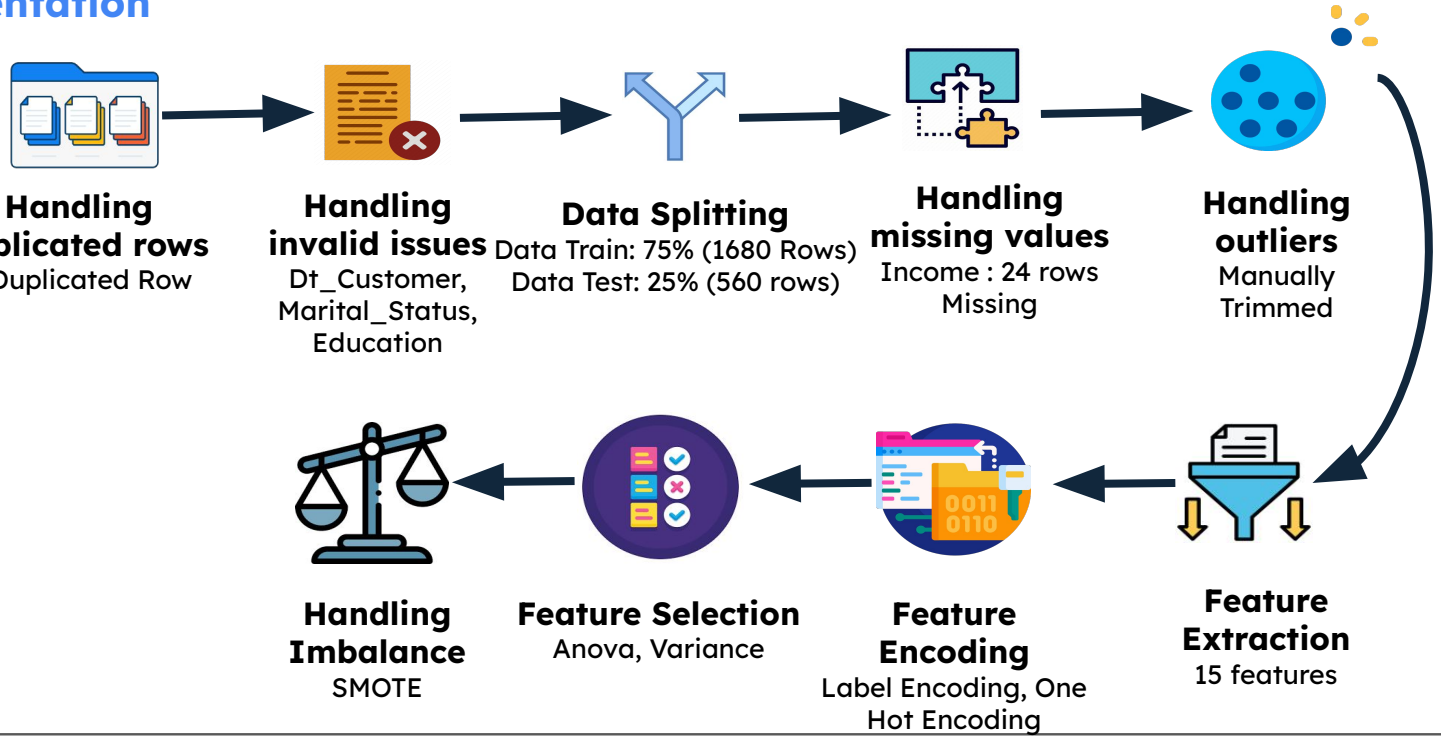
- **Input Layer**: 15 neurons; Rectified Linear Unit (ReLU) activation function; 37 input features.
- **Hidden Layer**: 15 neurons; ReLU activation function.
- **Output Layer**: 4 neurons;Softmax activation function.
- **Compilation**:
  - Loss function: categorical cross entropy
  - Optimizer: Adaptive Moment Estimation (Adam)
  - Evaluation Metric: Accuracy

Classification based on Customer Response to Past Campaigns Architecture

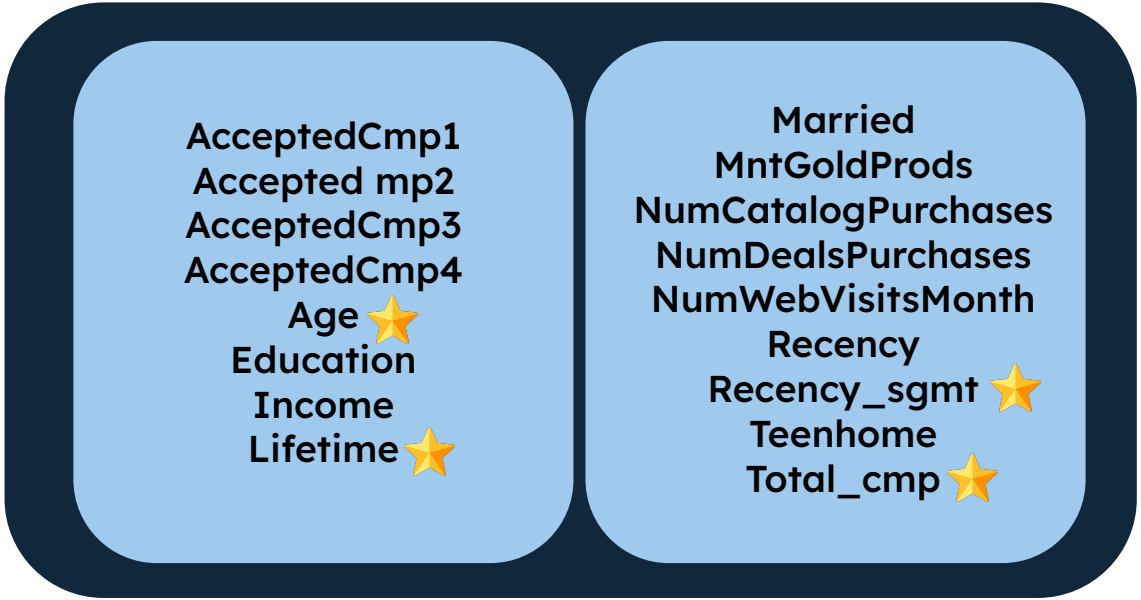
Motivation

After getting the cluster, we still need to predict whether the customers will join in the next campaign or not. We will use supervised learning binary classification methods to segment the customer.

Implementation



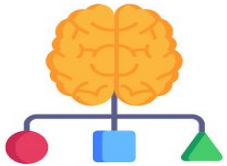
Classification Attributes



There are 17 features that will be used for modeling.
The stars is the result of Feature Engineering

Classification Modeling & Evaluation Metrics

Positive = Response Campaign
Negative = No Response Campaign

	 Model Prediction	 Actual / Reality
False Positive Main target to be reduced	Response	No Response
False Negative Second target to be reduced	No Response	Response
	FALSE	**TRUE**

Classification Modeling & Evaluation Metrics

1.

Precision

Primary Metrics Evaluation

Reducing False Positive
Customers are predicted to respond ✓
but in reality they don't ✗

↑ **Increase Response Rate**
↓ **Minimize Cost Marketing**

2.

Recall

Secondary Metrics Evaluation

Reducing False Positive
Customers are predicted not to respond ✗
but are interested in responding campaign ✓

Optimizing Revenue Rates

3.

F1 Score

Combined Metrics Evaluation

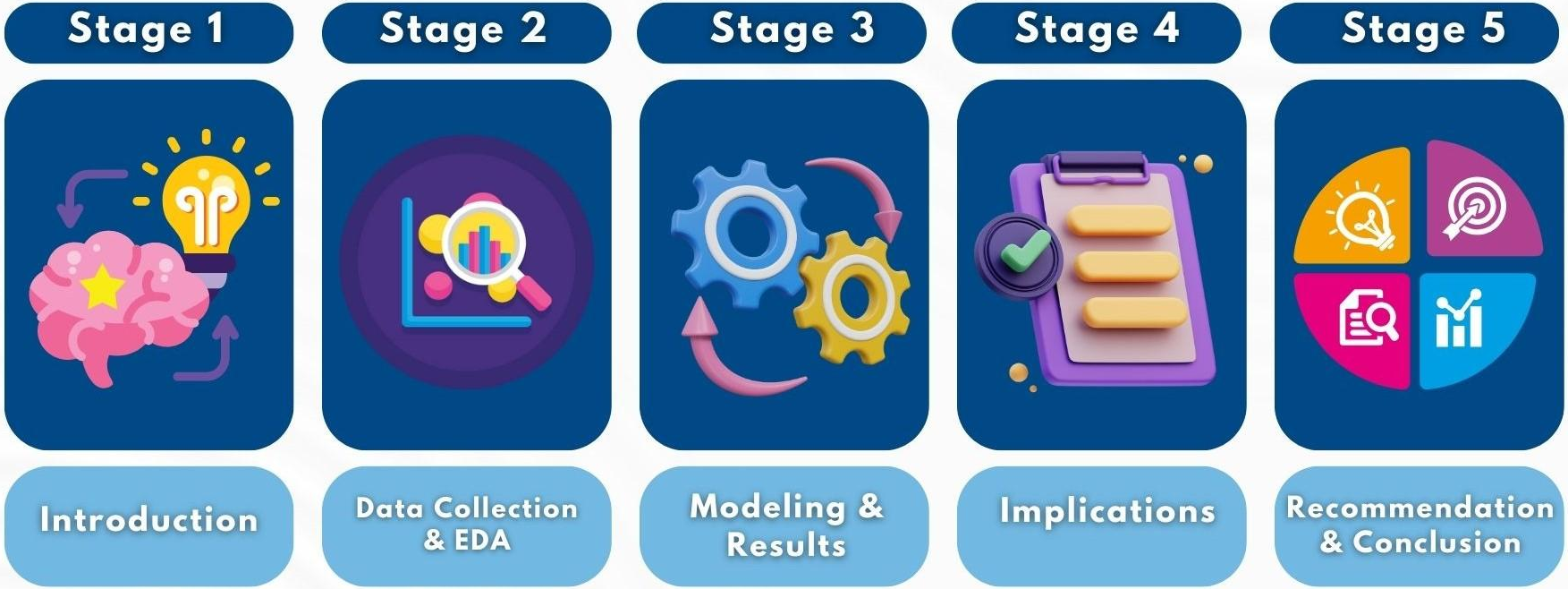
Necessary for our problem, we need a
Trade-off between Precision and Recall
F1 score is defined as the harmonic average of
Precision and Recall

Classification Results

Model	Accuracy	Precision	Recall	F1 Score	Cross Val F1 (k=5)	ROC AUC	Cross Val ROC AUC (k=5)
LogisticRegression	0.861000	0.523000	0.675000	0.589000	0.510000	0.900000	0.885000
LinearSVC	0.859000	0.520000	0.614000	0.564000	0.448000	0.895000	0.884000
MLPClassifier	0.846000	0.483000	0.518000	0.500000	0.558000	0.830000	0.887000
GradientBoostingClassifier	0.846000	0.482000	0.494000	0.488000	0.487000	0.725000	0.752000
KNeighborsClassifier	0.845000	0.482000	0.639000	0.549000	0.400000	0.842000	0.810000
DecisionTreeClassifier	0.838000	0.456000	0.494000	0.474000	0.488000	0.699000	0.718000

The Precision, Recall, and F1 Score values on the best evaluation are produced by **Logistic Regression** and **Support Vector Machine**

WORKFLOW ML


Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Introduction
Data Collection & EDA
Modeling & Results
Implications
Recommendation & Conclusion

Implications

Theoretical Implications

Which is the best approach between RFM analysis, clustering combined with classification, and a standalone classification model based on customer response to previous campaigns, in terms of their abilities to segment customers and provide actionable business insights for optimizing marketing strategies?

Marketing Implications

Given our clusters, what segments do we get? How much the potential GMV and optimization cost? What is the personalized approach marketing strategy for each segment?

Managerial Implications

How might a company impact a customer segmentation microservices in their internal stack?

Theoretical Implications

Approach	Performance (silhouette score - accuracy)	Results Interpretability	Method	Computational Cost
RFM Analysis	0.528 - N/A	Resulting clusters can be interpreted and turned into actionable insights	Data Analysis (No Machine Learning)	Low
Clustering + Classification (OPTICS + KNN/DTC)	0.72 - 98.91%	Resulting clusters can be interpreted and turned into actionable insights. Hard to find feature importance.	Unsupervised + Supervised Machine Learning	High
Classification based on customer response to past campaigns (Logistic Regression)	N/A - 86.1%	Low model interpretability, challenging to derive actionable insights.	Supervised Machine Learning	Moderate

Marketing Implications

AVERAGE TOTAL SPEND MID &HIGH CLUSTER

HIGH SPENDER (N=643)

1113,5

MID SPENDER (N=539)

679

POTENTIAL REVENUE INCREASE PER PERSON MID-> HIGH

FORMULA

MEAN TOTAL AMOUNT HIGH-
MEAN TOTAL AMOUNT MID

434

ASSUMING 15% CONVERSION RATE FROM MID-> HIGH, IT WILL GET US

POTENTIAL ADDITIONAL REVENUE

FORMULA

NUMBER OF CONVERSION X
POTENTIAL REVENUE INCREASE PER PERSON

35,154

Assuming $40 budget per person, we will get

PROFIT

FORMULA

POTENTIAL ADDITIONAL REVENUE
-TOTAL PROMOTIONAL COST

13,594

Marketing Campaign for Each Clusters

High Spender
"VIP Shoppers"

- Personalized Concierge Service
- Exclusive Preview Sales
- Elite Loyalty Program
- Premium Product Bundles

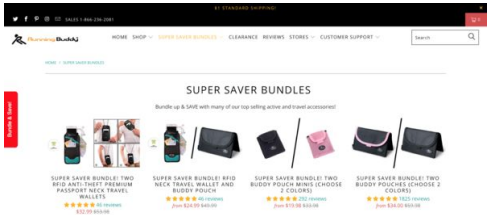
Middle Spender
"Smart Shoppers"

- Upgrade & Save Campaign
- Flash Sales & Limited Offers
- Referral Rewards Program
- Personalized Product Recommendation

Marketing Campaign for Each Clusters

Low Spender
"Value Seekers"

- First Purchase Discounts
- Frequent Shopper Discounts
- Bundle and Save
- Social Media Contests



Risk of Churn
"Win-Back Warriors"

- Reactivation Email Campaign
- Survey and Feedback Collection
- Exclusive Offers to Win-Back Customers

Managerial Implications

Goal: Building an automated, scalable, efficient, and accurate customer segmentation service.

- Data Processing Considerations: Batch Processing vs **Real-Time Stream Processing**
  - Batch Processing: Processing data in batches in regularly scheduled intervals
  - Stream Processing: Real-time processing of data streams as it is continuously generated or received
- If we use batch processing at long intervals, reclustering the entire customer database might be feasible, albeit inefficient
- If we use stream processing, reclustering the entire customer database is infeasible

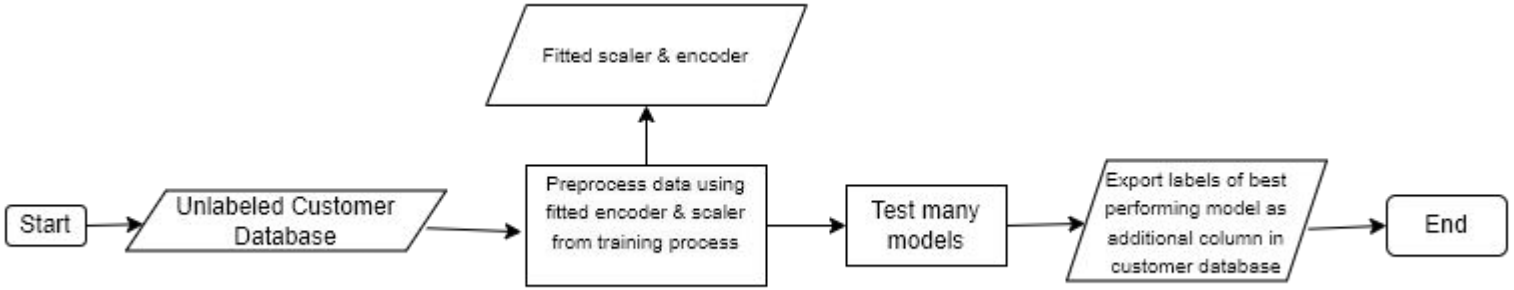
Solution

4 Step Process:

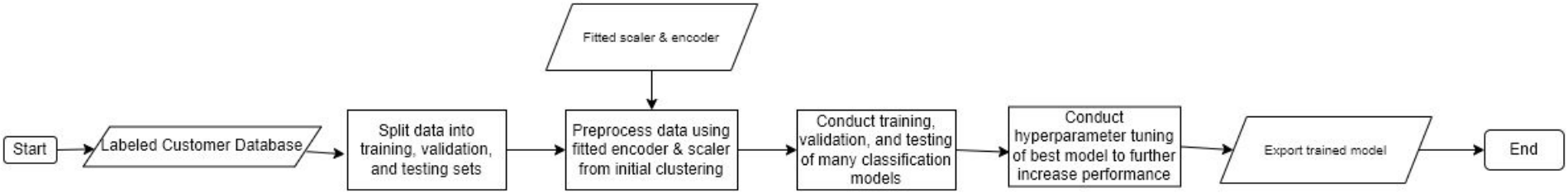
1. Segmentation of existing customers using clustering methods
2. Train a classification model using the cluster labels from the previous step as ground truth label for efficiency and scalability of future inference
3. Segment new customers using the trained classification model
4. Do model monitoring & maintenance to ensure reliability of results

Implementation of a Clustering + Classification Solution For Customer Segmentation

1. Segmentation of existing customers using clustering methods

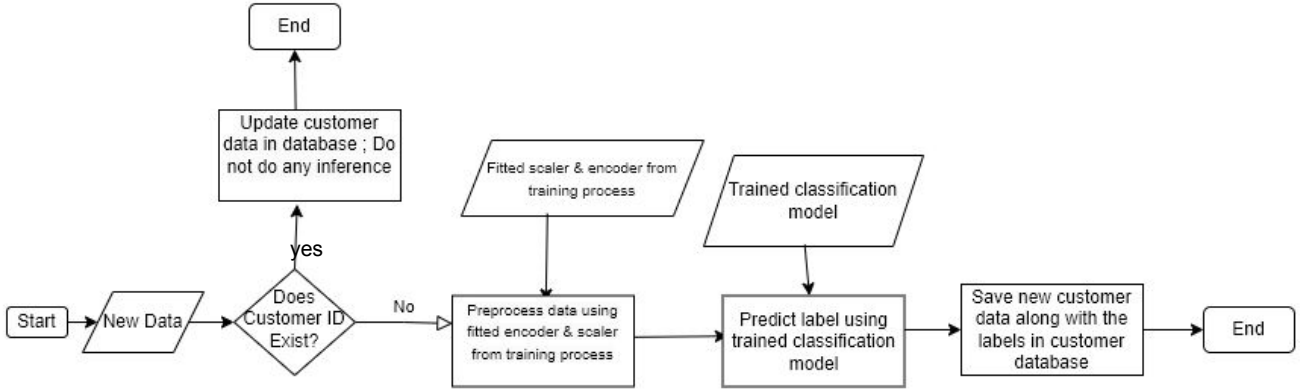


2. Train a classification model using the cluster labels from the previous step as ground truth label for efficiency and scalability of future inference



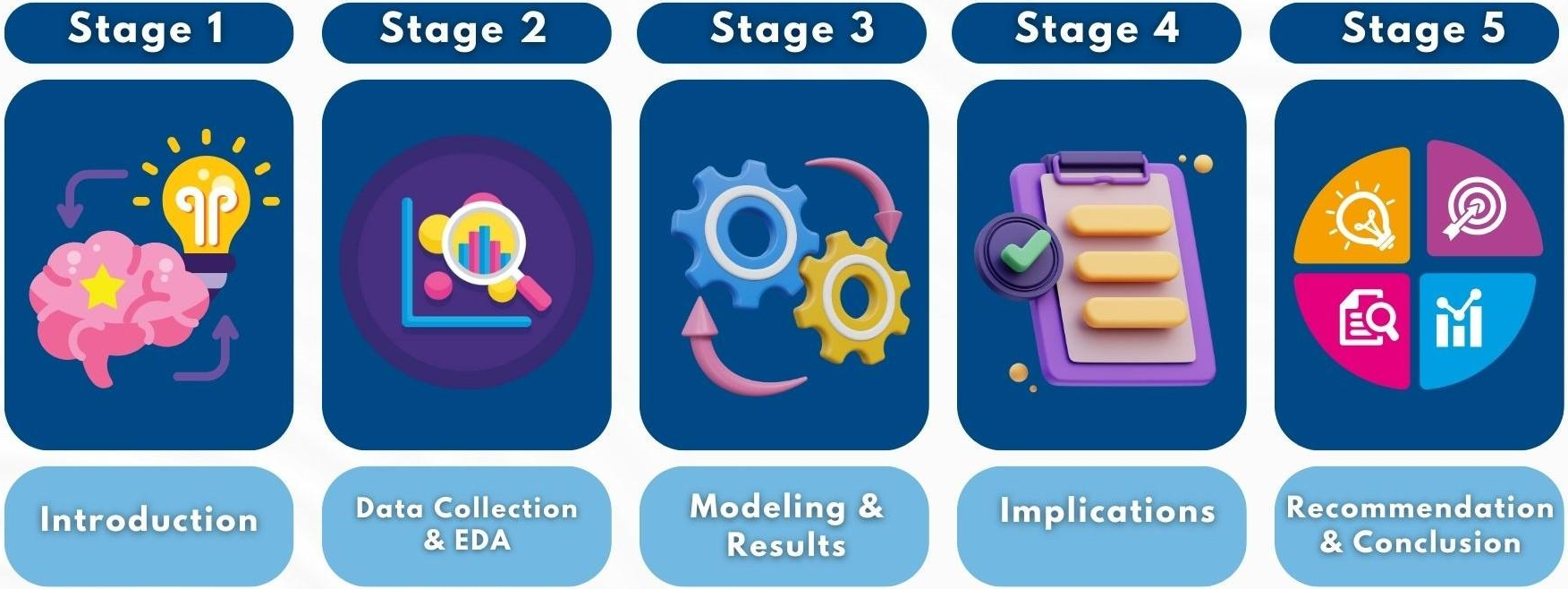
Implementation of a Clustering + Classification Solution For Customer Segmentation

3. Segment new customers using the trained classification model



4. Do model monitoring & maintenance to ensure reliability of results. Examples:
   - **Data Drift Monitoring:** Monitor Changes in statistical properties of input features. If changes are too drastic, reclustering the entire database might be needed to ensure accuracy of customer segmentation.
     - Example Tools: Tensorflow Data Validation, Apache Drift
   - **Cluster Stability Monitoring:** Monitor stability metrics of our clusters over time. In this case, an example is silhouette score. If cluster stability degrades significantly, reclustering is necessary.
   - **Classification Model Performance Monitoring:** Monitor the performance of the classification model over time. Using performance metrics such as accuracy, Gini and KS -Statistics, to ensure model reliability. If metrics degrades, then retraining our classification model is necessary.

WORKFLOW ML


Stage 1
Stage 2
Stage 3
Stage 4
Stage 5
Introduction
Data Collection & EDA
Modeling & Results
Implications
Recommendation & Conclusion

Novelty & Significance

Novelty

1

Employ different architecture

In this project we employed many different architectures. We used 10+ different models and tune several models to have better silhouette score result

2

Pipeline for Customer Segmentation problem

We also created a pipeline of customer segmentation that can be used to deploy the dashboard in the next stage.

Significance

1

Low Computational Cost

Our model can be converged very fast within just several iterations

2

Offering Advanced Method to solve bigger problem

OPTICS can be used to solve bigger problem of customer segmentation with high silhouette score.

Limitation & Future Work

Limitation

1. **Limited Time**
   We were only given a month to develop a comprehensive end to end project.

2. **Model is still under development**
   As can be seen in the performance report, our models' capabilities are still towards better development. We will use explore more hyperparameter tuning

3. **Limited Data**
   As can be seen in the data collection, the data that we have is relatively small.

Future of Work

1. **Enhancing model**
   Create much more varieties of models and maybe some combinations between models

2. **Trying on different dataset**
   The results from different dataset will make the results more generalized and applicable

3. **Making a better pipeline**
   A user friendly dashboard must be made so that the users can use the dashboard easily.

Q&A

ASK AWAY!

THANK YOU!

For Being part of our Capstone Journey!