INTEGRATED APPROACH ON CUSTOMER SEGMENTATION USING RFM AND MACHINE LEARNING ALGORITHMS

1. ABSTRACT:

In any Retail and E-Commerce industry one of the key elements in shaping the business strategy of a firm would be the understanding of customer behaviour. More specifically, understanding their customers based on different business metrics: how much they spend (revenue), how often they spend (frequency), are they new or existing customer, what are their favourite products, etc. Such understanding would in turn help direct marketing, sales, account management and product teams to support better the customer and improve the product offering in turn improving the business.

RFM model proves to be one such way of understanding the customers. RFM analysis is a data-driven customer behaviour segmentation technique where RFM stands for Recency, Frequency, and Monetary value. The idea is to segment customers based on their last purchase (Recency), how often they've purchased in the past (Frequency), and how much they have spent (Monetary). All three of these measures have proven to be effective predictors of a customer's which is also been used to perform better campaign's which in turn helps to increase the business ROI.

2. INDUSTRY REVIEW:

RFM is one of the commonly practiced procedure in the process of segmenting the customers based on the calculated recency, frequency, monetary value and with those calculated scores customer segmentation can be performed for future sales. This procedure is carried out in e-commerce, social media, lead management, apps, etc. The stake holder group within the firm would get benefited from such analysis and the insights would be:

1. **Product/Services**: Products selling more than others, would be an opportunity to evaluate the product offering or improve specific product features.

2. **Operations/Logistics:** From stock management perspective, understanding which products are in demand would reduce storage costs and improve delivery/logistics operations.
3. **Marketing:** Understanding of the customer segments, would allow for more effective and targeted marketing to specific customer groups, by creating a base campaign with core content for the broader client base but specific variations depending on the segment.
4. **Sales/Account Management:** Identifying which customers are the most valuable and understanding their trends would go a long way towards building a genuine relationship, thus retaining existing customers and attracting new with the ideal customer profile.

2.1 Literature Survey

1. **Title:** RFM customer analysis for product-oriented services and service business development: an interventionist case study of two machinery manufacturers
   Publication: Springer - 25 Jan 2019
2. **Title:** Customer segmentation by using RFM model and clustering methods: a case study in retail industry
   Publication: IJCEAS - Volume :8, Issue: 1, Year:2018
3. **Title:** RFM analysis: an effective customer segmentation technique using python
   Publication: Medium- Article
4. **Title:** Introduction to customer segmentation in python
   Publication: Datacamp- Article
5. **Title:** Integrated approach of RFM, Clustering, CLTV, and Machine Learning Algorithms for forecasting
   Publication: Analytics Vidya - Article

3. DATASET AND DOMAIN:

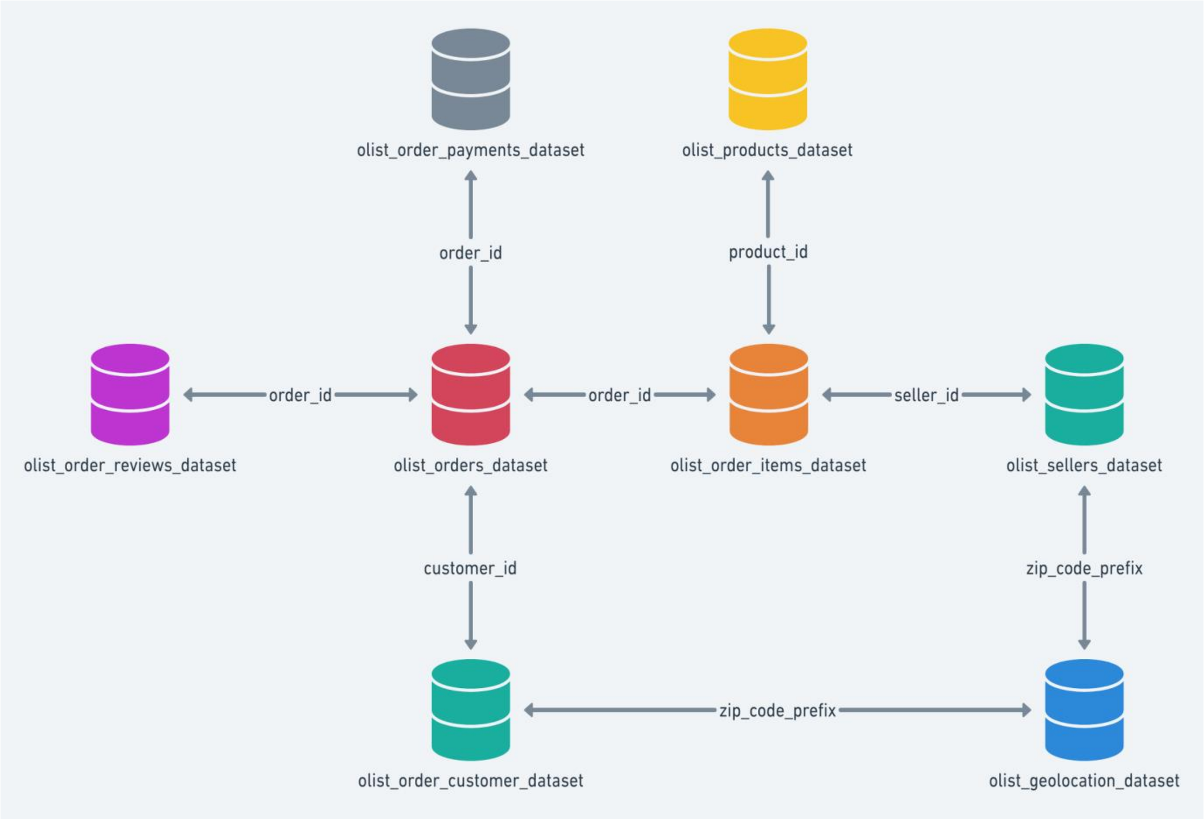
The Dataset has been taken from Kaggle; it is provided by the largest Brazilian online department store called Olist. The Dataset contains 8 separate tables which stored multi-dimensional data about over 100k order's information about the customers from Brazil between the year from 2016 to 2018.

DOMAIN: Marketing Analytics

SOURCE: www.kaggle.com/olistbr/brazilian-ecommerce

3.1 Olist:

Olist operates an online e-commerce site for sellers, that connects merchants and their products to the main marketplaces of Brazil. It has developed a platform for shopkeepers of all sizes and segments to register their products to be sold at the Olist store within Brazil's top retailers. It contains 9 different tables with different variables that contains the customer, seller, product details.



Only 4 datasets out of 9 is required to carry out the entire process. Those datasets are

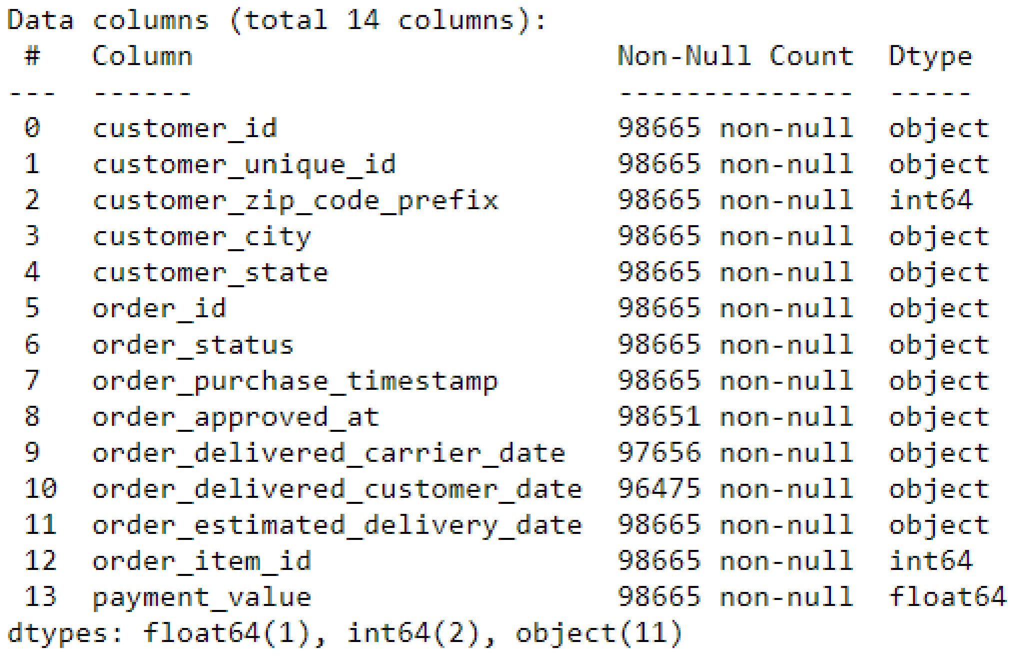
- olist_order_payments_dataset
- olist_orders_dataset
- olist_customers_dataset
- olist_order_items_dataset

3.2 DATA DICTIONARY:

COLUMNS	**DESCRIPTION**
customer_id	Key to the orders dataset. Each order has a unique customer_id
customer_unique_id	Unique identifier of a customer
customer_zip_code_prefix	First five digits of customer zip code
customer_city	Customer city name
customer_state	Customer state name
order_id	Order unique identifier
order_status	Reference to the order status (delivered, shipped, etc)
order_purchase_timestamp	Shows the purchase timestamp
order_approved_at	Shows the payment approval timestamp
order_delivered_carrier_date	Shows the order posting timestamp. When it was handled to the logistic partner
order_delivered_customer_date	Shows the actual order delivery date to the customer
order_estimated_delivery_date	Shows the estimated delivery date that was informed to customer at the purchase moment
order_item_id	Sequential number identifying number of items included in the same order
payment_value	Transaction value

3.3 VARIABLE CATEGORIZATION:

Initially the dataset gets merged together and had 14 different columns with different datatypes and data counts.



```
Data columns (total 14 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   customer_id                    98665 non-null  object 
 1   customer_unique_id             98665 non-null  object 
 2   customer_zip_code_prefix       98665 non-null  int64  
 3   customer_city                  98665 non-null  object 
 4   customer_state                 98665 non-null  object 
 5   order_id                       98665 non-null  object 
 6   order_status                   98665 non-null  object 
 7   order_purchase_timestamp       98665 non-null  object 
 8   order_approved_at              98651 non-null  object 
 9   order_delivered_carrier_date   97656 non-null  object 
 10  order_delivered_customer_date  96475 non-null  object 
 11  order_estimated_delivery_date  98665 non-null  object 
 12  order_item_id                  98665 non-null  int64  
 13  payment_value                  98665 non-null  float64
dtypes: float64(1), int64(2), object(11)
```

3.4 DATA PRE-PROCESSING:

The data pre-processing involves in modification of data types and dropping columns which is not required for the analysis and identifying the null values. In the data set date columns has been converted into timestamp values and the customer_id has been dropped since customer_unique_id is the primary key to the dataset. The presence of null values in the entire dataset is in negligible amount and can be ignored.

3.5 PROJECT JUSTIFICATION:

3.5.1 Problem Statement

- Identifying the customer segments based on the overall buying behavior of the client.
- To create an unsupervised model that generates the optimum number of segments for the customer base.
- Reducing the cost of acquiring customers
- Improving the ROI of marketing efforts

3.5.2 Complexity Involved

- The process of merging the datasets needed complete understanding of entire dataset and the key that connects the different dataset should be identified.
- Understanding of customer_id and customer_unique_id.
- Understanding of order_item_id and corresponding payment_value.
- Complexity in merging dataset because of sequential order_id.

3.5.3 Project Outcome

- Effective customer segmentation.
- Targeting the right customers.
- Improved marketing strategies.
- Improved campaigning strategies.
- Reducing customer churn.
- Obtaining highest ROI for the business.

4. DATA EXPLORATION (EDA):

In order to perform RFM analysis we need the customer_unique_id as the primary key. Based on that RFM values will be calculated. In this case we are considering only the customer whose products got delivered.

Recency:

The recency is calculated with the process of grouping the customer and the purchase date with the last date available in the dataset. This process will group the recent date of the customer and the last date in the dataset and provides the count of the days the customer recently purchased.

Frequency:

The frequency is calculated by grouping the customer_unique_id with the count of the order_id which gives the frequent number of purchases by the customer.

Monetary:

The monetary is calculated by grouping the customer_unique_id with the sum of amount that has been paid for the purchase. This gives the sum of payment_value by individual customers.

5. FEATURE ENGINEERING:

To perform RFM analysis it requires only minimal features to obtain the required scores. In this project the features namely customer_unique_id, order_id grouped with order_item_id and the payment_value based on the order_id from the entire dataset. These features will help in calculating the scores for the analysis. The remaining feature can be utilized once the customer segmentation is done. Based on the segmented customer each type of customers can be analyzed with the help of remaining features from the dataset.

6. ASSUMPTIONS:

In RFM analysis the RFM scores plays a major role. Once the score has been calculated it can be verified for the correlations and for the outliers' present. In this dataset there is no high correlation present between the scores and there exists few outliers which can be treated using IQR and other methods. Once these outliers have been treated k-means will be applied to attain the clusters for segmenting the customers and the segmented customers can be further analyzed based on the business requirements.