**Customer Segmentation with**

**RFM Analysis and Kmeans**

1. **Main Objective**

Attemp to find reasonable customer cluster into three segments. In this report will focus on finding “best” **cluster** that match into business understanding using unsupervised learning technique named Kmeans.

1. **Data Definition and Description**

This dataset was provided open-public in Kaggle by Vijaykumar, containing customer transaction data that might contain “cluster” behavior. The customer id’s might not actual customer id and the dataset information is intended for educational purposes only.

The dataset contain 541909 observations and 8 attributes (variables).

**Data description**

InvoiceNo : unique value of invoice receipt

StockCode : unique value of product

Description : product description

Quantity : amount of items purchased in the transaction

InvoiceDate : date of transaction

UnitPrice : product price per unit

CustomerID : customer unique id

Country : customer’s country

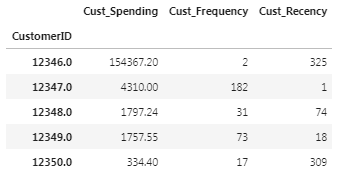
**Research Objective**

**Segment the customer into 3 segment** that can label the customer as such; “High Value Customer, Medium Value Customer, Low Value Customer”. Hence the **marketing can perform customized/ personalized ‘offer’** to the specific value type.

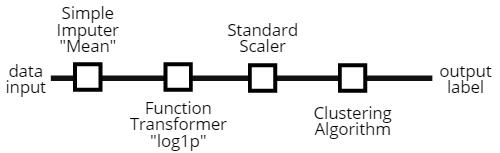
1. **Pre-processing and Feature Engineering Plan**

* Inspect the columns and its data type
* Inspect null value and drop/fix with reasonable value
* Fix the column data type and names
* Due to “best” clustering main objective, variable transformation will be performed
* Extract total value from “quantity” and “unit price”
* Extract customer recency value
* Perform RFM Analysis Transform by aggregating [total spending, recency value,invoiceid ] into CustomerID
* Extract recency score,frequency score, and monetary score from the data
* Extract overal rfm score
* Create machine learning pipeline for further transformation;

**Ready Dataset might look like:**

****

**Pipeline Diagram**

****

1. **Model Selection and Training Results**

To achieve best clustering that match the business understanding, we will conduct two different Models which are Kmeans and Agglomerative Clustering. The model training conducted with ;

* Pipeline technique
* Inertia metric

**Modeling**

1. **Agglomerative Clustering**

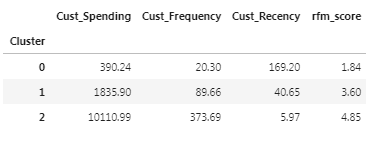
**Parameter used:**

N\_cluster = 3 as the business objective

Linkage = ward

Affinity/distance metric = euclidean

**Resulting Cluster:**

****

**Note ; 2 =** high value cust, **1 =** medium value cust, **0 =** low value cust

1. **Kmeans**

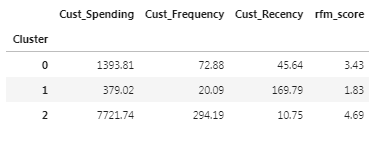
**Parameter used:**

N\_cluster = 3 as the business objective

Init = Kmeans ++

Algorithm = elkan

N\_init = 10



**Note; 2 =** high value cust, **1 =** low value cust, **0 =** medium value cust

Standard deviation comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| variables | Agglomerative | | | KMeans | | |
| High | Med | Low | High | Med | low |
| Cust\_Spending | 24131 | 8636 | 359 | 22238 | 4584 | 335 |
| Cust\_Frequency | 590 | 72 | 17 | 472 | 53 | 17 |
| Cust\_Recency | 7 | 44 | 105 | 12 | 46 | 105 |
| rfm\_score | 0.2 | 0.6 | 0.5 | 0.3 | 0.5 | 0.3 |

1. **Model Selection, Conclusion, and Summary**

Based on the distribution of cluster variables that measured with standard deviation value it is highly **recommended to choose Kmeans** as the clustering algorithm because it has “more” concistency on the value distributions. The model generate three segment label which are ;

* 1834 low value customers,
* 1702 medium value customers,
* 836 high value customers

From now on, this model will be mentioned as choosen model.

Due to clustering main objective, the model training will focus on how to generate a label with the most resemble with business sense. The unsupervised model can be evaluated by domain knowledge and business understanding. Hence **we can conclude below insights**:

1. We have 836 high value customers that on average will make a purchase on every eleven days
2. With the resulting label, perhaps the marketing can personalized the advertisements or any promotional offers
3. Low value customers dominated our customers, we might reconsider what is the causation that lead to low RFM score.

**Further Development**

This **dataset has a flaw** that heavily impacting research performace for RFM analysis. More than 100.000 missing values on customer id, hence any impute technique might end up violating research and analysis results hence the best option is delete them.

With dataset open-sourcity and broader techniques, any researcher can expand this report or conduct different research purposes, exploratory data analysis, and many more. Highly encouraged to develop further clustering technique on this dataset to generate more accurate to business sense.

**References**

Author’s notebook and source code

<https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/17c04566-afb1-4ecf-8d5e-431fb61e56ce/view?access_token=9cc4ad78aed1fcceadb27a8ce1996c86d36bca635e73b7a2974a634438eb370a>

datasets

<https://www.kaggle.com/vijayuv/onlineretail>

johar ashfaque’s similiar research

https://www.kaggle.com/ukveteran/clustering-here-we-go-again-jma