ML – Unsupervised

ONLINE RETAIL CUSTOMER SEGMENTATION

Almabetter, Bengaluru

Team Members	
Varsha Rani	**Vivek Chandrakant Pawar**
Rabista Parween	**Tushar Gaikwad**

1. Problem Description:

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

We are given the following dataset: **Online Retail.xlsx**

2. Introduction:

Attribute Information:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

3. Objective:

To identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

4. Steps Involved:

Importing Libraries and Data Inspection

We have used the following Libraries:

- Pandas – Manipulation of tabular data in Dataframes
- Numpy – Mathematical operations on arrays
- Matplotlib – Visualization
- Seaborn – Visualization
- Sklearn – Data Modeling

We have done data inspection. The original Dataset contains 8 columns and 541909 rows.

Data Cleaning

After dropping the duplicate values, now we have 8 columns and 536641 rows.
After dropping the null values, now we have 8 columns and 401604 rows.
We added 7 more columns to our dataset:
Amount_spent = df['Quantity'] * df['UnitPrice’]
Year = df['InvoiceDate'].dt.year
Month = df['InvoiceDate'].dt.month
Day = df['InvoiceDate'].dt.day
Hour = df['InvoiceDate'].dt.hour
Minutes = df['InvoiceDate'].dt.minute
Day_of_week = df['InvoiceDate'].dt.dayofweek

Exploratory Data Analysis

- Countries which made the most transactions
- Yearly transactions count
- Monthly transactions count
- Weekly transactions count
- Hourly transactions count
- Log distribution of quantity
- Distribution of UnitPrice

RFM(Recency Frequency Monetary) Analysis

- Recency Evaluation
- Frequency Evaluation
- Monetary Evaluation
- Recency with Frequency
- RFM Quantiles
- RFM score
- Number of Different Types of customers

Model Preparation

- Outliers variable Distribution
- Outliers variable distribution (After removing outliers for Amount)
- Data distribution after data normalization for Recency
- Data distribution after data normalization for Frequency
- Data distribution after data normalization for Monetary

Data Modeling

- Apply Silhouette score method on Recency and Monetary
- Elbow method on Recency and Monetary
- Customer segmentation based on Recency and Monetary
- Apply Silhouette score method on Frequency and Monetary
- Elbow method on Frequency and Monetary
- Customer segmentation based on Frequency and Monetary
- Apply Silhouette score method on Recency, Frequency and Monetary
- Silhouette analysis for KMeans clustering on sample data with n_clusters = 2 to n_clusters = 10
- Elbow method on Recency, Frequency and Monetary
- Customer segmentation based on Recency, Frequency and Monetar

Count of number of customers in each cluster

Cluster 0 = 1735 customers

Cluster 1 = 2055 customers

Cluster 0 Analysis

- Number of invoices per month
- Number of invoices per day of the week

Cluster 1 Analysis

- Number of invoices per month
- Number of invoices per day of the week

Conclusion

We have got 2 clusters by applying k means algorithm. So the customers got segmented into 2 clusters. Online Retail Customer marketing team can now use different approaches to acquire the customers

Cluster 0

Key Figures

- Frequency : 28.68
- Recency : 230
- Monetary : 3070
- RFM Score : 10.71

Top 5 Products

- WHITE HANGING HEART T-LIGHT HOLDER : 339
- REGENCY CAKESTAND 3 TIER : 268
- ASSORTED COLOUR BIRD ORNAMENT : 235
- PARTY BUNTING : 229
- REX CASH+CARRY JUMBO SHOPPER : 202

Cluster 1

Key Figures

- Frequency : 37.67
- Recency : 134.64
- Monetary : 447.40
- RFM Score : 5.90

Top 5 Products

- WHITE HANGING HEART T-LIGHT HOLDER : 344
- REGENCY CAKESTAND 3 TIER : 271
- ASSORTED COLOUR BIRD ORNAMENT : 239
- PARTY BUNTING : 232
- REX CASH+CARRY JUMBO SHOPPER : 204