Documentation of

‘Online Retail Segmentation & Analysis’ Project

Supervised by:

Eng. Mohamed Hanafy

Team Members:

Omar Elsayed Darwesh

Abdelrahman Saad

Hajar Khaled

Table of contents:

* Overview
* Business Understanding
* Data Understanding
* Data Preparation
* Data Segmentation & Analysis

**Overview**

Many small online retailers and new entrants to the online retail sector are keen to practice data mining and consumer-centric marketing in their businesses yet technically lack the necessary knowledge and expertise to do so.

In this project, a case study of using data mining techniques in customer-centric business intelligence for an online retailer is presented. The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively.

On the basis of the Recency, Frequency, and Monetary model, customers of the business have been segmented into various meaningful groups using customized segmentation based on special conditions as business needs and also using the k-means clustering algorithm, and the main characteristics of the consumers in each segment have been clearly identified.

Accordingly, a set of recommendations is further provided to the business on consumer-centric marketing. Python with data analysis packages is used in the project.

1. **Business Understanding**

The online retailer considered here is a typical one: a small business and a relatively new entrant to the online retail sector, knowing the growing importance of being analytical in today's online businesses and data mining techniques, however, lacking technical awareness and recourses.

The online retailer under consideration in this project is a UK-based and registered non-store business with some 80 members of staff. The company was established in 1981 mainly selling unique all-occasion gifts. For years in the past, the merchant relied heavily on direct mailing catalogues, and orders were taken over phone calls.

It was only in late of 2009 that the company launched its own web site and shifted completely to the Web. Since then, the company has maintained a steady and healthy number of customers from all parts of the United Kingdom and Europe, and has accumulated a huge amount of data about many customers. The company also uses Amazon.co.uk to market and sell its products.

So, the business problem is:

An e-commerce company wants to segment its customers and determine marketing strategies towards these segments.

To achieve this objective:

we will define the purchase behavior of customers,

create different segments based on these behaviors, and

suggest some valuable marketing strategies for each segment.

1. **Data Understanding**

The dataset contains all the transactions occurring in December of 2009 and also the whole years of 2010 and 2011.

The date range of orders (the first and last ever purchase):

* The first order: 2009-12-01
* The last order: 2011-12-09

The size of dataset (Records& Columns):

* Number of Records: 1067371
* Number of Columns: 8

There is a problem in the Quantity and the price columns:

* some records of quantity and price are negatives and zeros.

Number of Missing Values:

* In the Description column, there is 4382 records with null values.
* In the Customer ID column, there is 243007 records with null values.

Some invoices are cancelled and some are adjusted:

* it can be recognized by the first letter in the Invoice column. If it starts with ‘A’, then it’s adjusted. If it starts with ‘C’, then it’s cancelled.

Data Set Information:

Attribute Information:

InvoiceNo:

Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the 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 a transaction was generated.

UnitPrice:

Unit price. Numeric. Product price per unit in sterling (Â£).

CustomerID:

Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.

Country:

Country name. Nominal. The name of the country where a customer resides.

1. **Data Preprocessing**

In order to conduct the required RFM model-based clustering analysis, the dataset needs to be pre-processed. The main steps and relevant tasks involved in the data preparation are as follows:

1. Removing records with cancelled & Adjusted orders
2. Slicing the dataframe to contain only the rows with data in the customer id column. That’s, removing records with null values.
3. Removing records with price and quantity that are zeros or negatives.
4. Create an aggregated variable named a total price to use it in developing metrics, by multiplying Quantity with Price, which gives the total amount of money spent per product/item in each transaction.
5. Extracting a separate date column from the InvoiceDate column to make it easier to handle date analysis.
6. Creating three essential aggregated variables Recency, Frequency and Monetary:

* Creating a new dataframe only for the RFM analysis
* First, grouping the dataframe based on Customer IDs, then Creating lastpurchase column, then calculate the first factor of RFM (Recency)
* Dropping duplicates based on the invoice number and the customer id columns for the purpose of Frequency column, then creating Frequency column.
* Creating Monetary column - Summing the amount of money.

1. Splitting RFM elements into 4 parts and setting a score for each part depending on their weight/value.
2. Converting the three columns of RFM RecencyScore, FrequencyScore and MonetaryScore to use them in summing later.
3. Adding another two new columns: RFMScore, RFMScoreSum

* Concatenating all RFM score columns in one column per customer as a string to use it in checking the sizes of each RFM combination.
* Summing all RFM scores per customer to use in checking the sizes of each RFM total based on the sum of scores and other metrics.

1. **Data Segmentation & Analysis**
2. Checking the biggest segments in size based on each RFM combination of the newly created column RFMScore.
3. Checking the number of customers of each total unique sum of scores based on the newly created column RFMScoreSum.
4. Checking the size and the average of each group based on each total unique sum of scores based on the newly created column RFMScoreSum.
5. Segmentation with conditions According to business needs:

Ranked Segments:

* Setting general four segments of all customers based on total sum of scores of the newly created column RFMScore.
* Checking the average value for the actual RFM scores per each segment, then check the size of each segment.

Extreme Segments:

* Adding Extreme\_segm column to segment the customers in a way to highlight only the most recent, frequent or spending customers.
* Suggested segmentation of customers based on higest extreme score of each RFM factor per customer like 444.

Customer Level Segments:

* Adding another column CustValue\_Level to segment the customers in a way to show their value levels based on the RFM scores.
* Suggested segmentation of customers based on the same score of each RFM factor like 444, 333, 222 or 111.

Customized Segments:

* Adding another column Customized\_seg to segment the customers in a very deep and customized way with different conditions and different RFM scores for each.
* Suggested segmentation based on customized needs of business

1. Suggested Marketing Strategies for eight different segments to direct marketing efforts efficiently:

* Within each segment, identified what is RFM Score of this very segment, who are these customers, suggested marketing strategies, and an illustrating example.

1. Segmentation of customers using K-means Clustering way:

Checking for each one of the k-means assumptions:

* Checking if the recency column is following normal distribution or is skewed.
* Checking if the Frequency column is following normal distribution or is skewed.
* Checking if the Monetary column is following normal distribution or is skewed.
* Checking the average and the standard deviation of each RFM factor.

Slicing only the RFM values to apply log transformation and scaling on them.

* Applying logarithmatic transformation on each RFM factor
* Centering and scaling the average and the variance/standard deviation respectively at the same time using StandardScaler from sklearn library

Checking the optimum number of clusters using The Elbow Method.

* Dividing the customers based on RFM values three times differently:

Two Segments, then Three Segments, and then Four Segments.

After each segmenting, we are going to see:

* Summary Statistics for each cluster,
* Relative importance of cluster attributes compared to population,
* and a snake plot to notice the differences of RFM values between segments with different number of clusters.