ML1000 FINANCE GROUP

Assignment 2

Customer Segmentation using
Unsupervised Learning

10th March 2019

ML1000 Finance Group

Daniyal Shamim

Juan Calvillo

Mark Lewis

Salman Amin

Table of Contents

1.0 Introduction

E-Commerce is a tough, competitive, and constantly evolving industry. The barriers to entry are low and larger firms have an inherent advantage as they can afford to lower their prices and reach more customers without a significant impact to their profit margins. Larger firms, particularly Amazon and Walmart, can also consistently expand their product offerings as they have achieved economies of scale with their massive warehouses and distribution networks.

There is still room for niche players to survive if they can consistently add new customers and retain existing ones by using creative tactics that can differentiate them from mass retailers. One of these tactics is to offer targeted incentives to certain segments of customers to increase revenue and retain customers for the longer term.

In this report, we demonstrate how to apply this tactic for one of our clients, a niche retailer of gift items. We are going to employ the CRISP-DM framework along and Machine Learning algorithms on Point of Sale (POS) data to uncover trends using unsupervised learning and segment customers to implement a discount program.

2.0 Business Background & Context

2.1 Background

Our client is a U.K. based e-commerce firm specializing in the distribution of unique, all-occasion gifts. Its customers are spread throughout the globe and sales are made through its own website as well as other online marketplaces such as Amazon and Ebay. Facing stiff competition from other online retailers that are selling their own branded products, our client is looking for creative solutions to increase its stagnating revenue.

One of these creative solutions is the development of a discount program that applies a specific amount of discount targeting specific segment of customers, increasing the likelihood of them returning to buy more to take advantage of the discount. In order to do so, we need a deep understanding of the customers and their buying patterns.

2.2 Objective

The objective of this research is to identify distinct segments of customers to determine how to create a discount program for enticing more sales from these customers. We will analyze everything contained in the transactions data and apply appropriate unsupervised learning models to identify these customer segments.

3.0 Data Analysis

We are going to use the Online Retail Data Set, sourced from the UCI Machine Learning Repository[1]. This dataset contains 541,909 records of transactions occurring between 01/12/2010 and 09/12/2010.

3.1 Data Dictionary

There are 8 variables in the dataset as shown in Table 3.1 below:

Table 3.1

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

4.0 Data Exploration

Now we will delve into the details of the dataset and explore its contents. First, we load the raw dataset into RStudio as a dataframe.

```
data <- read.csv("Online Retail.csv", header = TRUE, na.strings =
c("NA","","#NA"),sep=",")
summary(data)
```

We notice that even though we have the Quantity and Price, there is no column with the total amount spent. We add that into our dataset as a new column:

```
data$totalSpent <- data$UnitPrice * data$Quantity
```

The dates of all the transactions appear to be as strings instead of the correct Date and Time format, so we reformat these values to be used in further analysis. We also create separate columns for Year, Month, Day, Hour and Minute:

[1] https://archive.ics.uci.edu/ml/datasets/Online%20Retail

```
dates <- as.character(data$InvoiceDate)
datesX <- strsplit(dates, " ")
datesX <- matrix(unlist(datesX), ncol=2, byrow=TRUE)
datesY <- strsplit(datesX[,1], "/")
datesY <- matrix(unlist(datesY), ncol=3, byrow=TRUE)
datesZ <- strsplit(datesX[,2], ":")
datesZ <- matrix(unlist(datesZ), ncol=2, byrow=TRUE)

data$month <- datesY[,1]
data$day <- datesY[,2]
data$year <- datesY[,3]
data$hour <- datesZ[,1]
data$minute <- datesZ[,2]
```

Now that we have all the values correctly formatted and included as separate columns in the dataset, we can explore the data more to make sense of it.

4.1 Orders

Here's a summary view of all the Orders in our dataset:

Table 4.1

Total Number of Orders	25,900
Average number of products	
Minimum Order Amount	-168,469.6
Maximum Order Amount	168,469.6
Average Invoice Amount	17.99
Median Invoice Amount	9.75

We notice the large negative value, and a corresponding positive value for the minimum and maximum, respectively. Further inspection reveals that there are some cancelled orders in the dataset. These are indicated by a prefix of "C" before the InvoiceNo. We will address these cancelled orders as part of our Data Preparation.

4.2 Products

The following table displays the summary about all the Products contained in our dataset.

Table 4.2

Total Number of Products	4, 223
Minimum Product Price	-11,062.06
Maximum Product Price	38,970.00
Average Product Price	4.61
Median Product Price	2.08

Again, we notice the extreme values in our dataset and further investigation leads reveals that the large negative amount is due to the Cancelled orders.

Among the 4,223 products sold, the top 10 selling products sellers are summarized in Table 4.3 below:

Table 4.3

Top 10 Selling Products	Total Quantity
WORLD WAR 2 GLIDERS ASSTD DESIGNS	53847
JUMBO BAG RED RETROSPOT	47363
ASSORTED COLOUR BIRD ORNAMENT	36381
POPCORN HOLDER	36334
PACK OF 72 RETROSPOT CAKE CASES	36039
WHITE HANGING HEART T-LIGHT HOLDER	35317
RABBIT NIGHT LIGHT	30680
MINI PAINT SET VINTAGE	26437
PACK OF 12 LONDON TISSUES	26315

When analyzing the Stock Codes, we notice that there are some unusual records. Table 4.3 below provides more details on these stock codes, we will address these as well in the Data Preparation section:

Table 4.4

Stock Code	Description	Details
AMAZONFEE	AMAZON FEE	Fees charged by Amazon Marketplace
D	Discount	Discount applied on certain transactions
CRUK	CRUK Commission	Commissions for products sold by CRUK
C2	Carriage	Shipment charges to certain jurisdictions
S	Samples	Free Samples provided to customers
POST	Postage	Postage paid on delivering orders
M	Manual	Unknown. We are unable to infer what these transactions refer to.
DOT	Dotcom Postage	Postage applied to the customers of an online marketplace
Gift_0001_##	Dotcomgiftshop Gift Voucher ##	Various discount vouchers that reduce the total amount of an Invoice by the amount indicated by ##

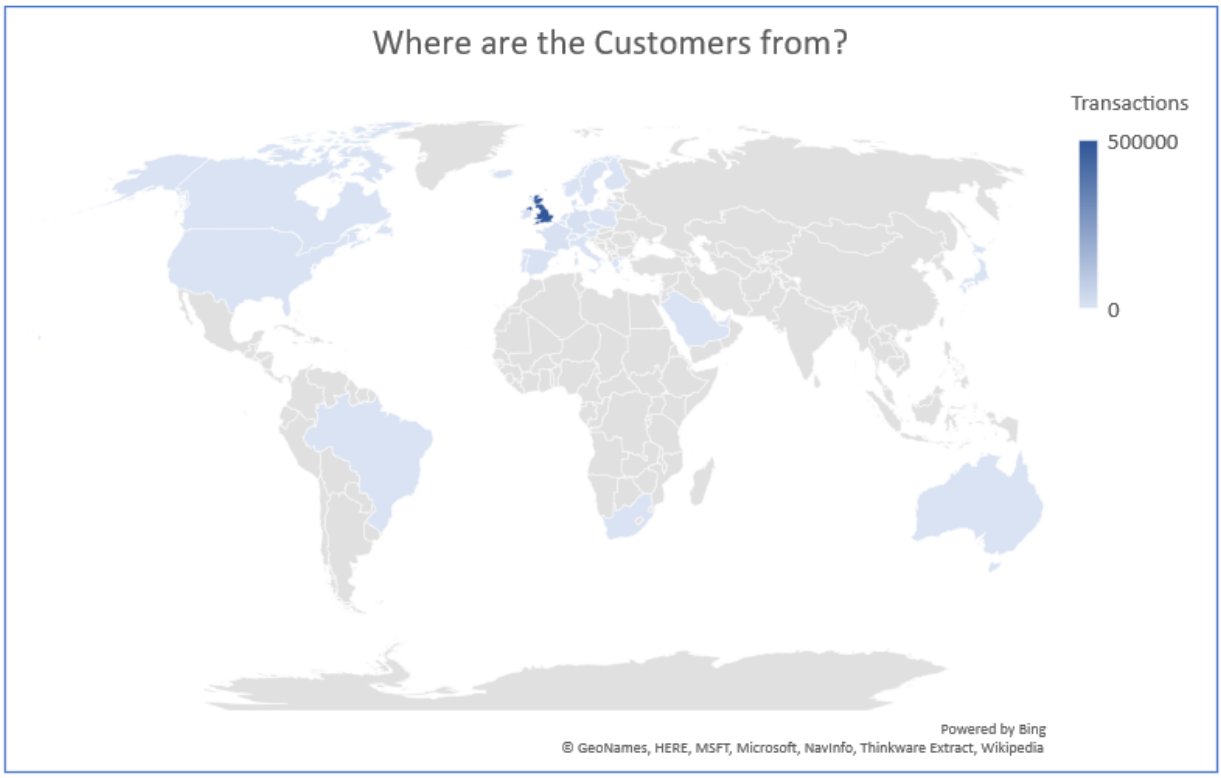
We also notice that there is a large amount of unusual Product Descriptions. These include blank values, "?", and other random values which indicate inconsistencies in data input. We will remove all of these during Data Preparation as well.

4.3 Customers

Equipped with a better understanding about the Orders and Products, we now take a closer look at our Customers. First, we notice that there are several blank Customer IDs. These will be removed during Data Preparation.

Looking at the distribution of customers, we notice that there are 4,372 number of unique customers in the dataset. The geographic distribution of the customers reveal that sales are concentrated across certain regions. The U.K. has the biggest share, followed by the rest of Europe. There are also some pockets in Asia, Australia, and the Americas where there is a moderate number of customers. Figure 4.1 below illustrates this geographic distribution.

Figure 4.1



5.0 Data Preparation

In order to prepare our data for unsupervised learning models, there is quite a bit of cleanup that we need to do. We are unable to impute data in all of these cases reliably so we remove them from our dataset to get a reliable, unbiased input for our models.

While the details of all the data preparation tasks are included in the R Markdown, here are the steps we take to to clean up the dataset:

1. Remove large negative amounts. These have a description of "Adjust bad debts"
2. Remove blank Product Descriptions
3. Remove other invalid Product Descriptions such as "?", "amazon fees", etc. See the R Markdown for details
4. Remove invalid Product Codes
5. Remove Cancelled transactions
6. Remove transactions with a blank Customer ID

6.0 Modeling and Evaluation

Given the structure of our data and the number of records, we were able to employ only the KMeans algorithm to create clusters and segment the data as illustrated in the R Markdown.

Based on our model, we were able to segment customers in three distinct clusters based on their Item and Quantity they purchase. These segments in turn determine the likelihood of the customer returning for another purchase.

Each one of the segments translate into the percentage of discount the customer would receive on their next purchase, as shown in Table 6.1 below:

Figure 6.1

Segment	Discount
1	0%
2	10%
3	20%

We attempted to run other unsupervised algorithms but the size of our dataset prevented us from doing so. We consistently got the error in RStudio with the message: "cannot allocate vector size…"

7.0 Model Deployment

We have developed a simple application for our client that determines the amount of discount a customer should be eligible for based on their purchase. Simply enter an eligible Stock Code and Quantity and click submit to determine how much discount the customer should receive on the next purchase.

The Shiny App for our model can be found at:
https://tenochinc.shinyapps.io/CustomerSegmentationAssignment2/

The GitHub Repository for this Assignment can be found at:
https://github.com/markglewis/MLCustomerSegmentation