Paper 151-2022

Data mining for the online retail industry: Customer segmentation and assessment of customers using RFM and k-means

Gowtham Varma Bhupathiraju, Veeram Reddygari Tharun Sai Raghavendra,

Oklahoma State University

ABSTRACT:

This study applies to identifying potential wholesalers, providing relevant and timely data to the company. To enable the company to understand its customers and to conduct efficient customer-centric promotion. Based on the Monetary, Frequency, and Recency of customers. Further customers segmented using the k means clustering algorithm into different significant groups, and the primary attributes of customers have been determined in each segment. Accordingly, the company is even more supplied with a set of suggestions on consumer-centric marketing and advertising.

INTRODUCTION:

Most businesses used a product-centric strategy as a marketing strategy, which mainly focused on the manufacturing to create a better product and reduce the manufacturing costs rather than paying attention to the customers who used them, since they could gain a lot of profits from the market share, according to the principle that economies of scale and scope apply. In the latter half of the 20th century, the third industrial revolution, also called the digital revolution or information technology (IT) revolution, introduced a new way to collect, store, process, and transmit digital information (Forester, 1986). This was the era of customer-centric marketing, which shifted from focusing on product design and delivery to one focused on individuals as customers.

For this reason, the customer relationship management (CRM) was therefore developed in order to handle the needs of clients and also to reinforce the income as well as marketing capabilities of the company. The CRM has 4 dimensions: Customer identification, Customer development, customer attraction and customer retention. Buyer segmentation was set up in the very first dimension of CRM, client identification, in order to recognize groups of people that share common qualities & behaviors. Market variables & Recency, Frequency along with Monetary (RFM) have been employed for client retention as well as client advancement dimension.

To understand the customer segmentation better nothing is better than the online retail business data. For the past years, we've witnessed a strong and steady expansion of internet retail product sales. Based on the Interactive Media online buyers in the United Kingdom invested an estimated £ fifty billion in the year 2011, an over 5000 percent increase in contrast to the year 2000.This remarkable increase of online sales suggests that the way consumers look for and make use of financial solutions has fundamentally changed. In contrast to regular shopping in retail shops, online shopping has quite a few exceptional characteristics: every customer's shopping activities and process may be tracked accurately and instantaneously, every customer's order is generally linked to a delivery address along with a billing address, along with every buyer has an online shop account with crucial payment and contact details. These attractive, particularly online shopping attributes have enabled online retailers for treating each client as a person with a personalized understanding of every consumer and also to build upon customer-centric occupation intelligence.

RESEARCH OBJECTIVES:

1) Which customers are the most and least valuable to the business? What are their characteristics?
2) Who are the most / least loyal customers, and how are they characterized?
3) What are the sales patterns in terms of various perspectives such as products / items, regions and time (weekly, monthly, quarterly, yearly and seasonally), and so on?

In order to handle these business issues, data mining methods happen to be commonly used in the online retail industry, combined with a pair of popular business metrics regarding buyers profitability as well as values, like the Recency, Frequency, as well as Monetary (RFM) model. Data mining is now an essential component of many business processes within the United Kingdom & worldwide, especially for big retailers like Amazon, Walmart, Sainsbury's, Argos, marks and Spencer in creating customer-centric business intelligence and supporting customer-centric marketing.

We present a case study on using data mining techniques in customer-centric business intelligence for an online retailer. Suppose we consider a typical online retailer: a small business and a relatively new player in the sector, knowing the growing importance of analytics in today's online businesses and data mining techniques, but lacking technical know-how and resources. Through this analysis, the business is better able to understand its customers and therefore conduct customer-centric marketing more effectively. The K-means algorithm has been used to segment the customers of the business according to the RFM model, and the main characteristics of customers in each segment have been clearly identified. As a result, the business is provided with recommendations regarding customer-centric marketing and further data analysis.

The remainder of this article will likely be arranged as follows. The following portion provides background info about the online retailer studied in the article, together with the associated dataset. And then, the section discusses in detail about the key steps and tasks for information pre processing to produce a suitable objective dataset just for the essential further analyses. The k means clustering examination is done in the following section, and a pair of substantial clusters and sections of the target dataset were identified. A comprehensive conversation is provided on every one of the clusters and the segmentation. The penultimate portion summarizes the important consumer centric business intelligence according to the results and also provides several concrete recommendations to the online retailer aiming at maximizing earnings for the company. In the last portion, the concluding remarks are at last offered.

BUSINESS BACKGROUND AND THE ASSOCIATED DATA

The online retailer under consideration in this particular article is an UK based and registered non store company employing roughly eighty individuals. Founded in 1981, the business sells primarily special all occasion gifts. The merchant utilized to depend on direct mail catalogues and orders have been taken over by calls for a long time. The company began using a site a bit over two years back, when it'd its first Internet presence. Since that time, the company has developed a solid and healthy client base from all of aspects of Europe and also the United Kingdom and possesses amassed a huge amount of information on almost all of its buyers. The organization likewise marketplaces and also offers its products and services through Amazon.co.uk.

The merchant's buyer transaction dataset has eleven variables, as found in Table one, and possesses most transactions for the many years 2010 as well as 2011 as shown below.

Variable	Data-Type	Description
Invoice No	Nominal	Invoice number; a 6-digit integral number uniquely assigned to each transaction
StockCode	Nominal	Product (item) code; 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	Numeric	The day and time when each transaction was generated;
UnitPrice	Numeric	Product price per unit in sterling;
CustomerId	Nominal	Customer Unique Id for any given user
Country	Nominal	Delivery address country

Table 1. Variables Details

DATA PRE-PROCESSING:

The original dataset must be pre-processed in order to perform the required RFM model-based clustering analysis. There are several steps and relevant tasks involved in data preparation:

1. Select appropriate variables of interest from the given dataset. In our case the following six variables have been chosen: Invoice, Stock Code, Quantity, Description, Unit Price , Invoice Date and Stock Code.
2. Create an aggregated variable named Total Price, by multiplying Quantity with Price, which gives the total amount of money spent per product / item in each transaction.
3. Remove the null values in the given dataset.
4. Remove the inappropriate data such as the data with given quantity less than 0.
5. Change the data types appropriately for the given variables for example as such Date time variable from text to datetime data type.

RFM MODEL-BASED CLUSTERING ANALYSIS:

The RFM (Recency, Frequency along with Monetary) analytic design is among the most crucial models utilized by businesses to produce marketing and advertising methods. The RFM design symbolizes consumption behaviors of customers depending on the transaction database, that is made simple as follows into 3 variables (attributes):

1) Recency (R): R means recency, that describes time period beginning from the latest purchasing behavior (last purchase) along with existing. The closer the date is, the much more likely it's the consumer is going to make a purchase once again.

It'll therefore have a better value in the variable recency.

2) Frequency (F): F could be the frequency which describes the number of transactions in a specific time. It's anticipated that the taller the loyalty of people, the taller the buy frequency of customers and the taller the consumer value will be for the business. The bigger the frequency is, the higher the value on the frequency variable.

3) Monetary (M): M represents financial that describes the quantity of total usage cash in a specific time. The bigger the monetary value, the taller the profit contributions of the buyer to the business as well as the higher the consumer value.

With all the prepared goal dataset, we plan to identify if customers in the view of monetary values, frequency, and recency might be segmented meaningfully. For this job the k means clustering algorithm was utilized and could be quickly applied in Python. For the RFM table, the raw transaction should be converted based on the RFM analysis model. Three important variables, monetary, Recency and Frequency. Frequency variables are determined by evaluating the very last month where each buyer makes a transaction coming from the Transaction Date variable. Thus, the Recency adjustable must be changed into a numerical information sort by counting the quantity of transactions produced by a specific buyer. The monetary variable can be estimated by including all the information values coming from the cost variable related to a specific customer.

	Recency	Frequency	MonetaryValue
CustomerID			
12346	326	1	77183.60
12747	2	103	4196.01
12748	1	4592	33707.73
12749	4	199	4090.88
12820	3	59	942.34

Output 1. Output for RFM values generated for each Customer.

With above given values, we were able to generate the RFM score for the each of the customer id and according with those values we have generated the RFM quartile values using these values we were able to generate the RFM Score which is addition of the all three RFM values for any given individual customer.

	R_val	F_val	M_val	R_quartile	F_quartile	M_quartile	RFM_Seg	RFM_Score
CustomerID								
12346.0	326	2	0.00	1	1	1	111	3
12347.0	2	182	4310.00	4	4	4	444	12
12348.0	75	31	1797.24	2	2	4	224	8
12349.0	19	73	1757.55	3	3	4	334	10
12350.0	310	17	334.40	1	1	2	112	4

Output 2. Output for RFM quartile for the RFM scores

With those created RFM score we are Visualizing the total count of each RFM Segment:

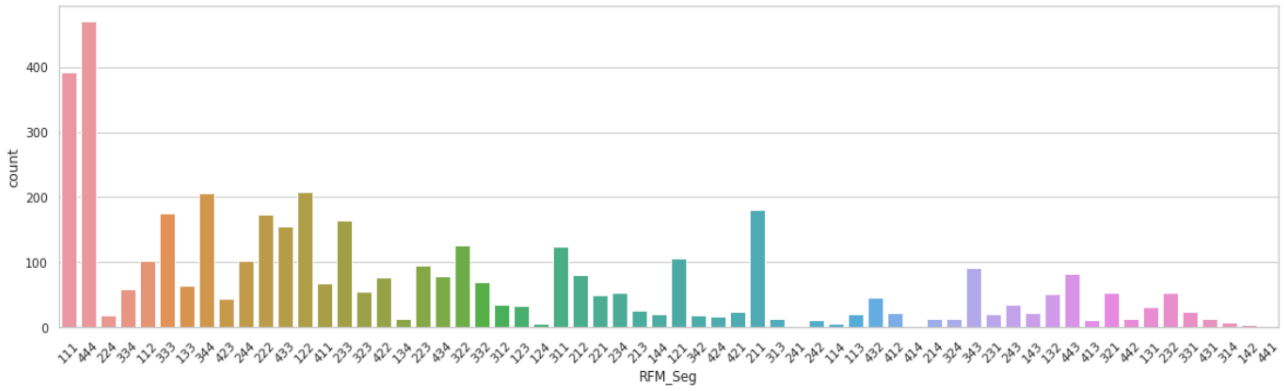

Figure 1. Histograms for Segmentation for RFM Groups

Now analyze the data better, we group the customers into different tiers based on their total RFM Score. In decreasing order of their RFM scores, they are as follows:

After grouping the customers with their RFM segment score, Now, we will be approaching the segmentation using K-Means Clustering, which is unsupervised learning. Grouping similar objects into clusters is the process of clustering. To begin, we should process the data in accordance with the following assumptions for K-Means Clustering: K-Means assumes that the variables are not skewed. R, F, M values will be tested. If they are skewed, we will use logarithmic transformation to remove the skewness. Below are the graphs for the RFM values.

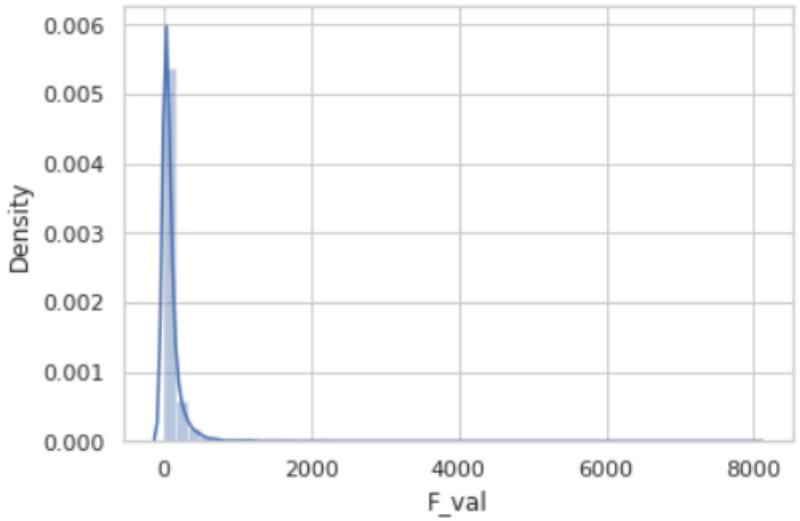

Figure 2. Positive (Right) Skewness graph of the Frequency value.

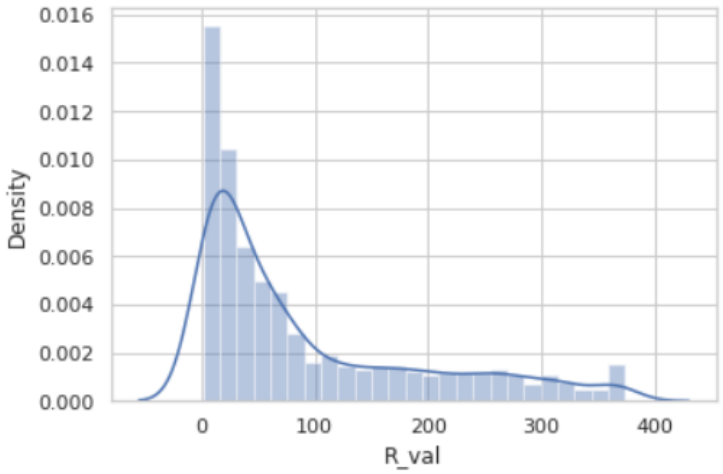

Figure 3. Positive (Right) Skewness graph of the Recency value.

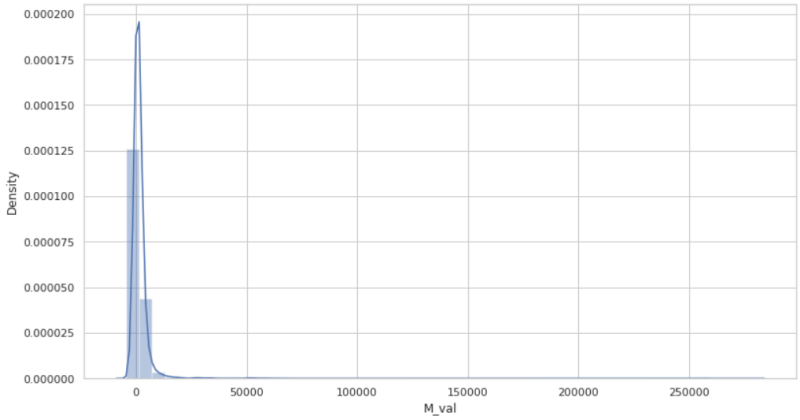

Figure 4. Positive (Right) Skewness graph of the Monetary value.

From the above given graphs we could see that Recency has the data right skewed so applying the log transformation is going to be fix that and here is the graph after applying the transformation.

The next step would be picking the right number of clusters as It is a major issue to determine the optimal number of clusters, k, in a dataset Partitioning clustering, especially in K-means clustering. There have been a variety of approaches proposed in the literatures for evaluating the number of clusters. Here we have used the elbow method to pick the idle number of clusters for the given data.

Elbow-Method- The elbow technique is probably the most typical method. In this case, the graph on the amount of squares is estimated at each selection of clusters. If the incline of the graph switches from high to shallow, the perfect amount of clusters, k, is going to be driven, at that “elbow" point. Here the graph describing the elbow- method for our data.

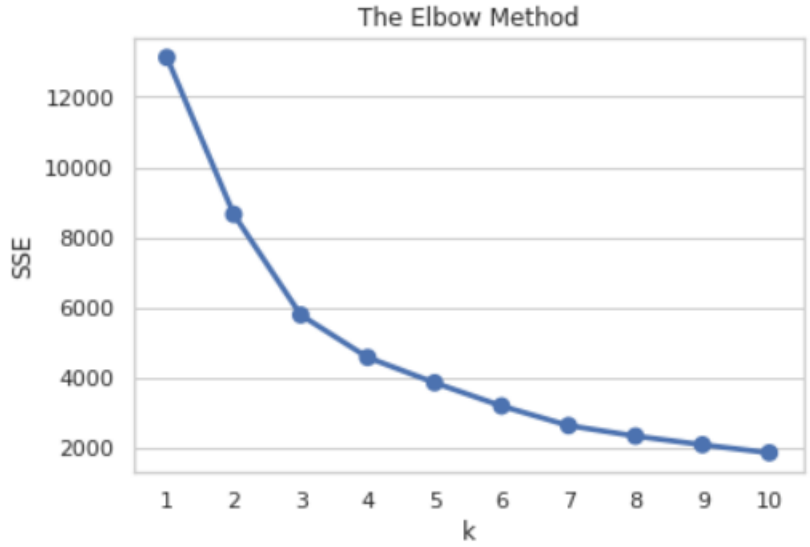

Figure 5. Elbow Method for finding no of cluster

We could see that the 3 is the idle number of clusters for our given data. Below are found the clustering and segment outcomes with 3 clusters and the distribution of situations within every cluster. This segmentation by 3 clusters appears to possess a better interpretation of the goal dataset than those by any other clusters.

	R_val	F_val	M_val
Cluster			
0	147.61	28.23	567.25
1	7.67	1828.33	182181.98
2	21.07	170.74	3039.99

Cluster -0 will be the number of people has a reduced frequency throughout the entire year along with a significantly smaller average financial worth, indicating a significantly lesser quantity of spending per customer. This particular cluster could be categorized as medium monetary, high frequency, along with minimal recency, with a moderate spending every consumer.

The consumers in this group have a reasonable frequency value. There is a lower but reasonable monetary value for this group. As this group seems to represent ordinary consumers, it has a certain level of uncertainty regarding profitability. Depending on how you view the long-term picture, some of the consumers may be highly profitable or may be completely unprofitable.

RECOMMENDATIONS

For each one of these consumer groups, it's crucial to further discover what items the consumers in each team have bought, which items are bought together most often and where sequence the items have been bought. The business can produce a clear understanding of its clients by checking out the associations between their consumer organizations and the items they purchase. The association could be investigated also at the number of items or products and at the degree of product groups.

Most of the customers of the company were organizational customers with a high volume of merchandise per transaction. The company is going to benefit significantly by examining what products they've purchased in the different seasons, what kinds of products they've bought and just how frequently they actually do this. It'll additionally be interesting to compare as well as contrast the shopping patterns of various groups of buyers, organizational and individual namely clients. In the long term, monitoring the variety of probably the most diverse client group and predicting which consumer will possibly get associated with the most or maybe the least profitable team will be extremely helpful for the company. Linking customer organizations to geographic locations is yet another issue which could be appealing to researchers. In case there's a correlation, this may assist the company think about how other things like tradition, culture, and economics might influence a consumer 's purchasing choice.

CONCLUDING REMARKS

This article uses a case study to demonstrate how customer-centric business intelligence can be created for online retailers using data mining techniques. Using the case study as a guide, the business can identify the different customer segments that contribute to its profitability and accordingly, develop appropriate marketing strategies to target them. It has been proven in this study that the most crucial and time-consuming steps in the entire data mining process are the preparation of the data and the interpretation and evaluation of the models. Further research for the business entails conducting association analysis to determine which products have been purchased together frequently by which customers and which customer groups; enhancing the merchant's website so that a consumer's shopping activities can be captured and tracked instantly and accurately; and predicting a customer's lifecycle value in order to quantify how diverse a customer is.

REFERENCE:

1. Interactive Media in Retail Group (IMRG) . (2012) Press archive , http://www.imrg.com , accessed January 2012.

2. Kumar, V .and Reinartz , W . J.(2006) Customer Relationship Management: A Databased Approach , Hoboken, NJ: John Wiley &Sons .

3. Hughes, A. M.(2012) Strategic Database Marketing 4e: The Masterplan for Starting and Managing a Profi table, Customer-based Marketing Program, McGraw-Hill Professional, USA

4. Davenport, T. H.(2009) Realizing the Potential of Retail Analytics: Plenty of Food for Those with the Appetite . Working Knowledge Report, Babson Executive Education.

ACKNOWLEDGMENTS

We would like to express gratitude to professor Chakraborty Goutam and McGaugh Miriam who guided us throughout this project.

CONTACT INFORMATION:

Your comments and questions are valued and encouraged. Contact the author at:

Gowtham Varma Bhupathiraju

gowtham.bhupathiraju@okstate.edu

Pursuing MS in Business Analytics and Data Science from Oklahoma State University. Prior to this he was working as a Software Engineer in IT People Cooperation.

Veeram Reddygari Tharun Sai Raghavendra

tveeram@okstate.edu

Pursuing MS in Business Analytics and Data Science from Oklahoma State University. Prior to this he was working as a Software Engineer in IT Cooperation.