CUSTOMER SEGMENTATION USING MACHINE LEARNING

Report By Sania Shaikh for Feynn Labs

ABSTRACT

The emergence of newer competitors has caused a lot of tension among competing businesses to find new buyers and ensure customer retention. Exceptional customer services catering to personalized needs of each and every customer is of utmost importance. Furthermore, the ability of any business to understand the needs of each of its customers will provide greater customer support in providing targeted customer services and developing customized customer service plans. This understanding is possible through structured customer service. Each segment has customers who share the same market features. Big data ideas and machine learning have promoted greater acceptance of automated customer segmentation approaches in favour of traditional market analytics that often do not work when the customer base is very large. Through this paper, it is attempted to tackle this issue using the k-means clustering algorithm is used for this purpose.

KEYWORDS

data mining; Machine learning; Big data; Customer segment; K-Mean algorithm; Sklearn; Extrapolation

PROBLEM STATEMENT

In the modern day and age, increasing competition between similar companies has led to the increase in importance of customer retention and acquisition. In order to successfully do this, businesses need a clear idea of the demographic of the customers. Different demographics have different success ratio for various marketing methods.

However manually collecting and analysing such large amounts of data requires a lot of time and can sometimes be inaccurate due to human error.

Therefore, we require a faster, more efficient and accurate method to carry out customer segmentation.

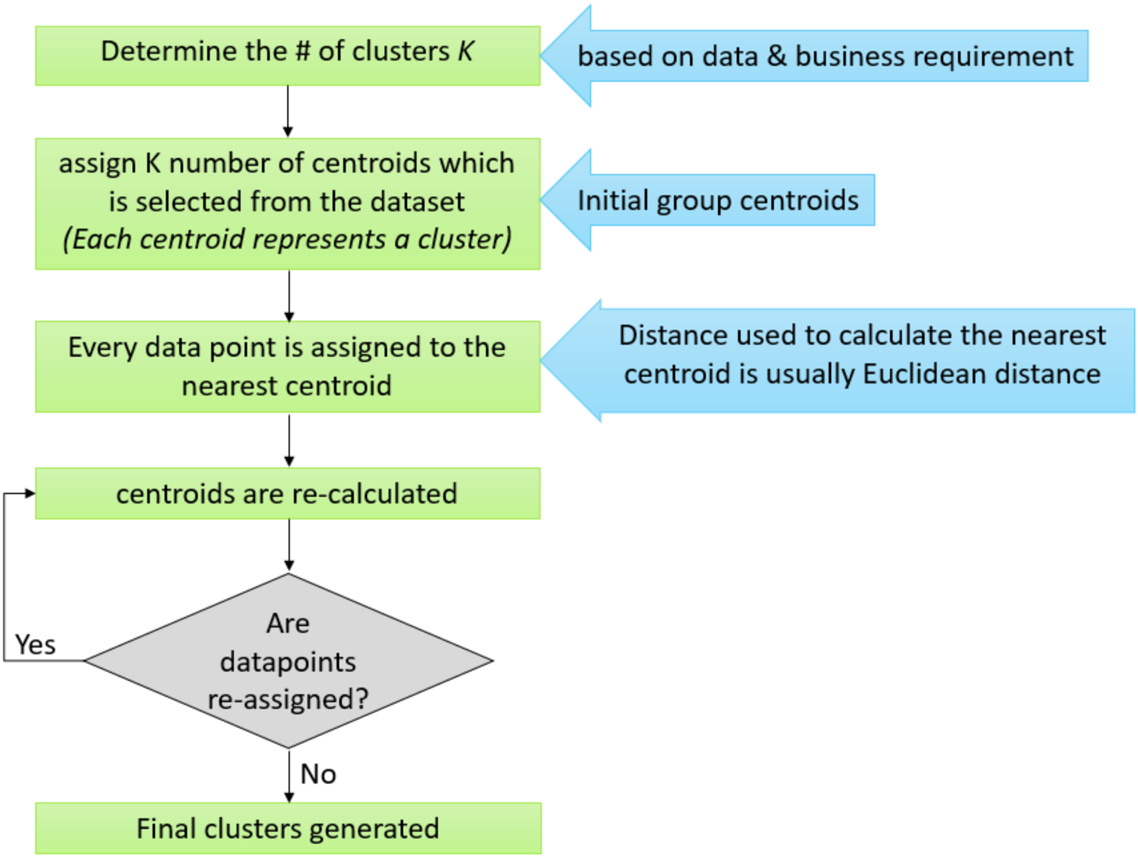
MARKET/ NEED ANALYSIS

Customer segmentation allows you to learn about your customers on a deeper level. With this information, you can tailor your content to each group's unique needs and challenges. You'll also be able to create targeted campaigns and ads that resonate with and convert certain segments of customers.

Other common benefits of customer segmentation include:

1. Improving your customer service and customer support efforts.
2. Helping internal teams prepare for challenges different groups are likely to experience.
3. Communicating with segments of customers through preferred channels or platforms.
4. Finding new opportunities for products, support, and service efficiently.

METHODOLOGY


Determine the # of clusters *K*
based on data & business requirement
assign K number of centroids which is selected from the dataset
(Each centroid represents a cluster)
Initial group centroids
Every data point is assigned to the nearest centroid
Distance used to calculate the nearest centroid is usually Euclidean distance
centroids are re-calculated
Are datapoints re-assigned?
Yes
No
Final clusters generated

DATASET

The data set used to implement clustering and k-means algorithm was collected from UCI Machine Learning Repository. The data set contains 541909 instances and 8 attributes. The attributes of data set consist of Invoice no, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country. The data is in raw format.

VISUALISING DATASET

Dataset contains data in raw format, hence it can contain anomalies like negative values, missing values. Visualising dataset help in gaining insight of data in it. Using information of data,we can pre-process data according to our needs. From Table I. we can see that there are certain null values in two columns:

1]Description (0.27%)

2] CustomerID (24.93 %)

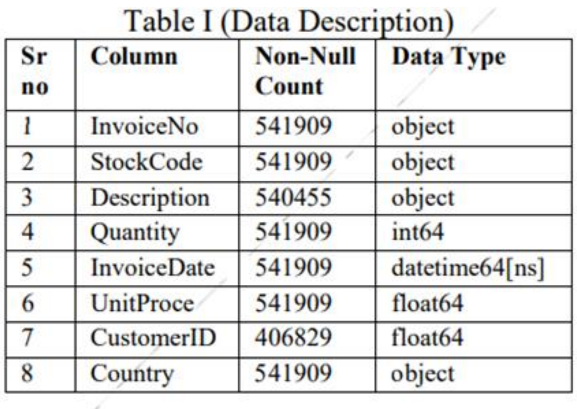


Table I (Data Description)

Sr no	Column	Non-Null Count	Data Type
1	InvoiceNo	541909	object
2	StockCode	541909	object
3	Description	540455	object
4	Quantity	541909	int64
5	InvoiceDate	541909	datetime64[ns]
6	UnitProce	541909	float64
7	CustomerID	406829	float64
8	Country	541909	object

From Table II. We can see that there are negative values in two columns;

1] Quantity

2] UnitPrice

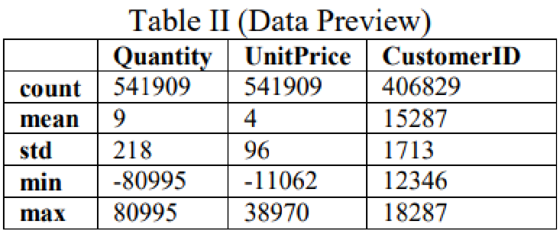


Table II (Data Preview)

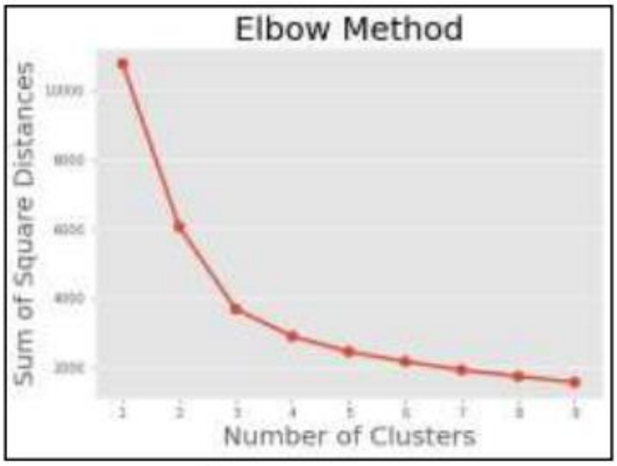
	Quantity	UnitPrice	CustomerID
count	541909	541909	406829
mean	9	4	15287
std	218	96	1713
min	-80995	-11062	12346
max	80995	38970	18287

K- MEANS CLUSTERING

k-means clustering model is one the vastly used model for clustering. Being unsupervised learning algorithm, it has many applications. It requires the number of cluster's to be formed. The optimal number of clusters can be found by different methods. One of them is Elbow method.

ELBOW METHOD

To determine the optimal number of cluster's required for k-means algorithm elbow method can be used. Elbow method generates plot which has an elbow like curve. The point where elbow is formed is taken as the optimal value of k. The elbow is formed at cluster number three. Hence, the number of clusters to use in k-means clustering are k = 3.



REFERENCES

[1] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics.

[2] R.C. de Amorim, C. Hennig (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors". Information Sciences.

[3] Leonard Kaufman; Peter J. Rousseeuw (1990). Finding groups in data : An introduction to cluster Aanalysis. Hoboken, NJ: Wiley-Interscience

[4] Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems.