TECHNICAL AND BUSINESS WRITING

How Data Science in Customer Segmentation facilitates Targeting the Potential User Base: Exploration

MEMBERS:

K17-3711 Aymen Irum
K17-3730 Eisha Tir Raazia
K17-3639 Fatima Ibrahim
K17-3658 Maham Tariq
K17-3766 Samana Moosa
K17-2375 Syed Muhammad Ammar

1. ABSTRACT

Lately, customer segmentation has been receiving much attention due to its effectiveness in the marketing paradigm. Perceiving purchasing habits by client type assists with advertising suitably is a challenge these days. Marketers are constantly looking for ways to improve the effectiveness of their campaigns by targeting potential users/audiences e.g: they target customers with particular offers most likely to attract them back to the store and to spend more time and money on their next visit. Data science is always a good fit for customer segmentation where clustering and different types of data modeling help us identify the groups having some common traits. Therefore, the problem is that with the increase in user/customer records, the segmentation through the big data becomes way harder. Although, a number of published papers have addressed the technicalities of targeting potential customers to increase profitability. This paper adds to the literature an overview of how segmentation helps you target the right audience while applying different techniques and calculations on big user records in order to make better marketing/advertising decisions. It also discusses how customer segmentation may have different possible uses in the future.

2. INTRODUCTION

Market segmentation is the action of partitioning a broad consumer or business market, typically comprising existing and expected clients, into sub-gatherings of customers (known as segments) in light of some kind of shared qualities. **Data science** plays a very important role in achieving the goal of targeting potential users through insights gained by large user data.

Customer segmentation is done on different bases depending on the potential market, for example:

Segmentation base	A brief explanation of base (and example)
Demographic	Quantifiable population characteristics. (e.g. age, gender, income, education, socio-economic status, family size or situation).

Geographic	Physical location or region (e.g. country, state, region, city, suburb, postcode).
Geo-demographic or geoclusters	Combination of geographic & demographic variables.
Psychographics	Lifestyle, social, or personality characteristics. (typically includes basic demographic descriptors)
Behavioral	Purchasing, consumption, or user behavior. (e.g. Needs-based, benefit-sought, usage occasion, purchase frequency, customer loyalty, buyer readiness).
Contextual and situational	The same consumer changes in their attractiveness to marketers based on context and situation. This is particularly used in digital targeting via programmatic bidding approaches

3. PROBLEM STATEMENT

These days more consideration is paid on building concrete web-based advertising procedures, due to the gigantic development of web advancements. Technicalities of targeting potential customers, through data science, to increase profitability are mentioned at several places but this task becomes very crucial with the increase in user data, the segmentation through the big data becomes way harder and makes it difficult to target the right market. Therefore, creative methods to transfer data in awkward forms to a form you want it to be is most important before applying any algorithm to gain best insights of the potential user base

4. BACKGROUND

Before the 1950s it was believed that direct marketing, which used to be called “mail order” was a mass marketing strategy, in which a company or an organization decides to completely neglect the market segmentation and appeal the whole market with a single offer.[11] In fact, the development of direct marketing can be directly linked to technical advancement in those early years, which enabled the consumers to be approached as a group rather than as an individual by the tradesmen and artisans.[13]

By the late 1940s, R. L. Polk and Reuben H. Donnelley initiated to compile the data lists acquired by the external sources, consisting of contact numbers, auto registration, and driving license. Moreover, U.S. postal service allowed Donnelly to create the first list of “Occupants” which enables mail-order based companies and manufacturers to send their sample products to those all addresses which they compiled before or within the specific region regardless of knowing who would be the receiver of the parcel.[12], [13]

Compiled information helped the mail industry to become more sophisticated and it also enabled the marketers to make the procedure of sending a large number of mails in a more systematic way. Moreover, the compiled data really helped some companies to carry out the incredibly targeted mailings. For example, automobile registration data also helped to estimate the income proxy of families, as less wealthy

households were believed to own no car or drive a lower-end model. In the late 1940s Driving license records were used to target older persons with direct mail requests for life insurance. Since age is listed in these records and drivers were supposed to be wealthy enough and relatively good in health, old American insurance has been a success. [13]

With businesses that have a very large database, it is very challenging to perform segmentation.[1]. Therefore, data science is playing a great role in customer segmentation by giving insights into data in large databases and by making data as fine as possible through different Exploratory Data Analysis(EDA) techniques on it. In this era data science is the most

5. METHODOLOGY

EDA and data cleaning are the most important data science techniques which are applied in parallel with machine learning algorithms to get real insight into the data and helps in performing customer segmentation, as accurately as possible.

After conducting the data collection, the most important and critical part is 'Noise Removal' from the data. Data doesn't always come in ready-to-analyze format. The application of ML algorithms like clustering, K-Nearest Neighbours (KNN), DBSCAN, etc give you the results but they aren't accurate if provided data is not meaningful and provides the right learning attributes/features. Therefore the right approach for the customer segmentation is making the data set as clean and meaningful as possible, by following all the good practices of data preprocessing and choosing the right evaluation metrics to gain insights.

Some data science techniques that help us in getting the best out of the data are:

5.1 Data preprocessing and data cleaning

Data preprocessing is the most critical step in the data science process. This step aims at noise removal by using statistical tests and data transformation methods to get data into better shape for modeling. Data preprocessing involves several techniques to handle the data such as:

- Identical rows (IDs) or rows with missing IDs. It is advised to assign the IDs through a method like Decision Tree, a classifier, which recognizes a customer's ID based on its purchases.
- Punctuation is removed and capital letters are lowered.
- Data imported is explored for outliers, NaN values, and domain ranges.
- Different features have different domain ranges, it can make the model biased. To avoid this, feature scaling is used to limit the domain of every feature from 0 to 1.
- NaN values are generally treated in two ways. One, observations with at least one NaN value are dropped. However, this might pose a problem if the number of observations available is small. In such a scenario, the second approach to treating NaN values comes in handy. NaN values are replaced with the average of some past and future values (of the same feature).
- Majority machine learning models work only with numeric data type which means, categorical variables with string data type are to be encoded to a numeric datatype. Label encoder is used to transform variables into numeric data types by assigning a number to every category in a feature (variable).

- Unbalanced data also called skewed data is a kind of data in which observations with a certain label are more than all other observations with other labels collectively. If a model is trained on skewed data, the model is most likely to predict new observations with the most abundant label. To avoid this, data resampling and data mining techniques are used to get data into better shape.
- Outliers are detected using a z-test. Z-scores are assigned to each observation of every feature in the data frame. Observations (rows) with z-score greater/less than +3/-3 are labeled as outliers and thus removed.

5.2 Exploratory Data Analysis (EDA)

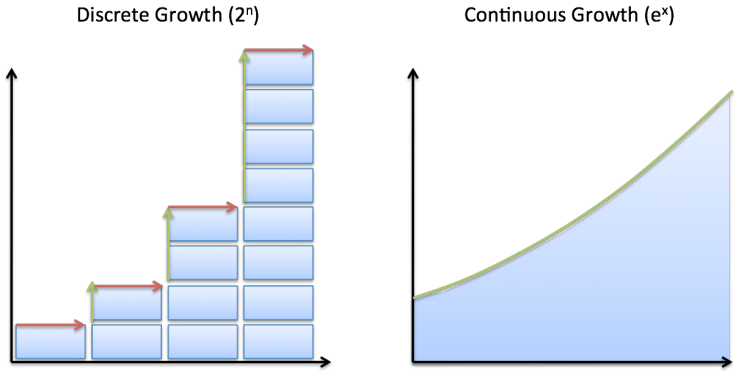
EDA is another critical step in the data science process. It involves:

- Univariate, Bivariate, and Multivariate analysis is done to find the correlation among features in the dataset.
- EDA provides insight on how one feature changes with another by observing correlation values and trends on plotted graphs.

5.3 Data visualization

Data visualization is performed using different libraries to see the trend and behavior of data. Two types of data are visualized:

- Continuous
- discrete.



After performing all the above tasks and making sure that data has no unmeaningful fields or features left, that adds to the noise, different machine learning algorithms are applied on it, but in case of Market segmentation clustering and data mining techniques are always a good idea.

5.4 Clustering And Data Mining Techniques:

The rapid development of data mining methods enables using large databases of customer data to extract knowledge, supporting the marketing decision process. Now clustering is a critically important data mining technique and can be used for the segmentation applications in the market

forecasting and planning research. Clustering is a statistical technique, which is very similar to classification. It organizes raw data into relevant clusters and homogeneous observation groups. [2]

Since we are looking into how data mining techniques like clustering can facilitate targeting the potential user base, we must know that there are many approaches to cluster analysis.

The most popular is the k-means algorithm, which together with its modifications was broadly investigated by different authors. [5], [6], [10]

Algorithms:

In our report, we will focus on three algorithms that have been previously discussed in different papers to obtain the goal of market segmentation.

- K-means
- Two-phase clustering algorithm
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

1. **K-means:**

   Out of all the algorithms mentioned above, K-means is the one used most often. When you have a large dataset, this would be the best choice for segmentation. It partitions the data into a given number of clusters. In the start, K-means move the observations into random clusters but with each new iteration, the algorithm keeps transferring the data to the nearest cluster. This process stops only after all the observations are located in the closest cluster. [1]

2. **Two-phase clustering algorithm:**

   S.S. Tseng, C.M. Su, and M.F. Jiang introduced a Two-phase clustering algorithm. It was originally developed to detect outliers. [9]

   The algorithm consists of two phases.

   Phase 1: In this phase, the K-means algorithm is modified. The original one calculates cluster centers after allocating all the objects but in modified one, these centers are calculated after every object's allocation.

   Phase 2: MST (minimum spanning tree) is created, nodes in the spanning tree are the clusters that are found in phase 1. Outliers are detected by removing the longest edge. Agglomerative hierarchical clustering technique was then introduced in phase-2 because the goal was not only to find outliers but also to obtain high-quality clusters. [1]

3. **DBSCAN:**

   Ester at al. introduced DBSCAN [4], which uses a density-based approach. It finds arbitrary shape clusters and outliers in data. Two concepts followed by DBSCAN are

density-reachability and density-connectivity of points. It performs very well when dealing with large spatial databases with noise. [1]

Experiments:

Many experiments have been performed to test the efficiency of data mining's clustering algorithms. One such experiment had been conducted and written by the computer science department of Technical University of Lodz, Poland.

Total objects: 300

Objects features (characteristics) considered for the test: age, income, deposit, credit, and profit/loss.

The following tables show no. of clusters found by all algorithms and no. of objects in each cluster.

Table 1: K-means

Cluster	Number of objects
1	126
2	11
3	26
4	6
5	70
6	19
7	42

Table 2: Two-phase clustering

Cluster	Number of objects
1	236
2	25
3	6

4	9
5	4
6	11
7	4
Outliers	5

Table 3: DBSCAN

Cluster	Number of objects
1	166
2	18
3	8
4	7
5	6
Outliers	95

Results of the experiment:

- **K-means:**
  K means is highly dependent on its input parameter ‘k’. Although it is very efficient for large datasets, it is not recommended in the case of datasets with noise.

- **Two-phase clustering:**
  Two-phase clustering performs really well in case of noise but it is more suitable for a small amount of dimensions because in multidimensional cases, it may provide low-quality clusters.

- **DBSCAN:**
  Incorrect choice of input parameters for this algorithm may result in a very bad quality of obtained clusters the detection of too many outliers but with correct input parameters, this algorithm works really well.[1]

6. CONCLUSION

Data Analysis and its insights are most important in customer segmentation that helps to evaluate and prioritize your best customer segments. Once, the desired insights are gained and we start to get maximum wins on different tests then it's the right time to translate all the gained information into action.

In light of the previously researched papers and results of the experiments conducted, it is recommended to increase the usage of data mining techniques for market segmentation, and using them for online systems with a large user base would seem to be a very good option.

7. RECOMMENDATIONS

Before COVID-19, the world did not require focusing much on learning online but now we have entered an era where we must prepare ourselves for distant learning. A lot of educational material is available online for kids and adults but most of the parents and students themselves always preferred face-to-face interaction by going for traditional campus-based educational systems. Now when things are unclear on when we will be able to bring 1000s of students together, we must prepare for the worst and start developing better online educational systems for students all over the world. Keeping this in mind, we propose an online educational system where students will have access to multiple courses. They may be offered a survey when they register to keep a track of their interests and then they will be offered related courses. The system may offer the functionality to add other people as friends and suggest adding students showing similar interests. This way, students will be able to communicate with others and study in groups as well.

A newly updated analysis projects growth of nearly 200% in global higher education enrollments through 2040 [3] which means we will have more students than ever in the Post-COVID-19 era. For a large user base like this, we need techniques that can handle such load and are able to provide us with fast results. Therefore, it is recommended to use data mining techniques for such an online educational system.

8. REFERENCES

[1] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," Proc. - 5th Int. Conf. Intell. Syst. Des. Appl. 2005, ISDA '05, vol. 2005, pp. 197–202, 2005, doi: 10.1109/ISDA.2005.33.

[2] K. R. Kashwan and C. M. Velu, "Customer Segmentation Using Clustering and Data Mining Techniques," Int. J. Comput. Theory Eng., no. January 2013, pp. 856–861, 2013, doi: 10.7763/ijcte.2013.v5.811.

[3] Angel J Calderon, “Massification of higher education revisited,” 2018.

[4] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, 1996, pp. 226-231.

[5] P. V. Balakrishnan, M. C. Cooper, V. S. Jacob, P. A. Lewis, “Comparative Performance of the FSCL Neural Net and K-means Algorithm for Market Segmentation”, European Journal of Operational Research, Elsevier Ltd., 93,1996, pp. 346-357.

[6] H. Hruschka, M. Natter, “Comparing Performance of Feedforward Neural Nets and K-means for Cluster-Based Market Segmentation”, European Journal of Operational Research, Elsevier Ltd., 114,1999, pp. 346-353.

[9] M.F. Jiang, S.S. Tseng, C.M. Su, “Two-Phase Clustering Process for Outliers Detection”, Pattern Recognition Letters, Elsevier Ltd., 22,2001, pp. 691-700.

[10] R. J. Kuo, L. M. Ho, C. M. Hu, “Cluster Analysis in Industrial Market Segmentation through Artificial Neural Network”, Computers & Industrial Engineering,, Elsevier Ltd., 42, 2002, pp. 391-399.

[11] Ross, Nat (1992), “A History of Direct Marketing,” Unpub- lished Paper, NY: Direct Marketing Association.

[12] Roel, Raymond (1988), “Direct Marketing’s 50 Big Ideas,” Direct Marketing,(May).

[13] LISA A. PETRISON ROBERT C. BLATTBERG PAUL WANG