BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Customer Segmentation

– MIRPR report –

Team members
Ciprian Cuibus - Back End - Report - ciprian.cuibus1@gmail.com
Aurel Nicolescu - Back End -Report - nicolescu.aurel.7@gmail.com
Titus Trif - Front End - ttitus_97@yahoo.com
Alexandru Pintican - Front End -andreipintican@gmail.com

2020-2021

Abstract

Text of abstract. Short info about:

- project relevance/importance,
- inteligent methods used for solving,
- data involved in the numerical experiments;
- conclude by the the results obtained.

Please add a graphical abstract of your work.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 What? Why? How?

- The world of e-commerce is a tough one, as there are hundreds of thousands of online shops, hundreds targeting the same clients and trying to out compete their rivals. For doing this, one of the best ways is having loyal customers, and what is a better way to keep a customer buying products from you than having targeted discounts that are exactly what the person wants?
- But how do businesses know what ads/recommendations to send their clients? This process of âgetting to know the customerâ is named customer segmentation
- The project is used to segment customers into similar groups, for a better recommendation of different products.
- There is actually no real need to use an intelligent algorithm for this kind of task, someone can just spend days and weeks finding good recommendations for each customer, but since this is no real solution that plagues all sale companies/ markets, the competition dictates a smart solution for this problem.
- The scientific problem was finding a good algorithm for dividing the users into groups. This can be done through a number of algorithms, but after researching a bit on qualified articles we have found that K Means Clustering (more on that later in the report) is at the moment the best fit for this.

1.2 Paper structure and original contribution(s)

The research presented in this paper advances the use of K Means Clustering into the Customer Segmentation problem. As this is a real world problem we have used a real world dataset to help us test our results using different parameters. And for this specific dataset we came up with a specific number of clusters that we obtained through a specific clearance of data, adding new derived features and normalization

We have implemented into a small application the algorithm to help us suggest to different users the right kind of items

The present work contains *xyz* bibliographical references and is structured in five chapters as follows. //TODO

- The first chapter is a short introduction in the problem at hand and a small idea of how to work around it.
- The second chapter describes the problem in more details and a scientific solution that is possible at this time and we would like to test it out.
- The third chapter details the current state of the art in the world of e-commerce businesses.
- The fourth chapter describes our algorithm and how we came up with it and with different parameters that we have used in implementing it.
- The fifth chapter is a discussion on our application, methodology and dataset used, and the results of it being used on the aforementioned UCI dataset.
- The sixth and final chapter illustrates the conclusion of our experiment and possible the future work, with which we can expand our application.

Chapter 2

Scientific Problem

2.1 Problem definition

As stated before. When a retail shop wants to keep, or expand their clients base, the easiest way to do so is by "buying" their loyalty. And how do you get loyal customers? By giving them assurance that you have the products they need at a cost they can afford, or they are really happy with. So how do you know what each client wants? Basically you don't. Or you can know, but this comes with a tremendous effort attached to it. You better observe their purchasing behaviour, and compare it to other clients so you are able to group them into similar categories, and then make appropriate suggestions for each category, not each client.

Why should this problem be solved by an intelligent algorithm? Because we, as humans don't have enough time and energy at our disposal to do this task. And since we have a great computing machine at our disposal instead, we may as well use it to do this task. The traditional (ancient) approach for this was measuring the RFM coefficient for the users and ranking them by this number.

The advantage of using K Means is that it is a fairly simple algorithm to implement, having a set of meaningful data. It requires minimum assistance from the programmer side, as he/she can easily analyze the correctness of the results and adjust the parameters accordingly. The disadvantage is that if we don't have a good enough dataset our results may be not good enough or outright wrong. But in the case of e-commerce, a small dataset means that the business is fairly small and may not even require such analysis and/or effort for using Artificial Intelligence solutions

For our solution we are using K Means Clustering, which is an unsupervised learning algorithm. This means that we do not have to train our data with correct or incorrect labels, it just finds groups of

similar entities. Of course, we do need to tend to it a bit, because the proper number of clusters must be found out by ourselves, through thorough testing and measuring different coefficients, but that is mostly all that we need to do, as there are already dedicated tools that implement most of the hard calculations that must be made.

The advantages are clear, we don't have to train our data, and given a big enough dataset almost assures us of decent results. As stated before, a well organised e-commerce business will mostly have this kind of dataset at it's disposal.

The disadvantage is that the possibility of a corrupt, too small or completely disorganised data is still there, but in this case we don't really have an actual solution. Only time and some human input from the company will make the use of an intelligent solution possible

Chapter 3

State of the art/Related work

The theory of the methods utilised until now in order to solve the given problem.

Answer the following questions for each piece of related work that addresses the same or a similar problem.

- What is their problem and method?
- How is your problem and method different?
- Why is your problem and method better?

In order to cite a given work you can use a bib file (see the example) and the *cite* command: [?], [?], [?], [?], [?].

Customer segmentation on a bank database

For our case study we have read some articles regarding customer segmentation, and one of those was [1]. In it is discussed that market segmentation is one of the most important area of knowledge-based marketing. Regarding banks there was once a huge barrier regarding the study of collected data, as data bases are rather large and multidimensional. With the emergence of data mining methods, this process was fastened, giving us access to more options regarding the work with these databases. In the paper it is discussed the use of 3 algorithms, DBSCAN (Density Based Spatial Clustering of Applications with Noise), K Means and a 2 step clustering process, then comparing their effectiveness and scalability. They even took into account noise possibilities as well as high dimensionality of data. In another cited paper, it was discussed the use of K Means Compared with a Neural Network. In it, neural networks gave poorer results, but a combination of these two was actually the better solution.

The main advantage of DBSCAN consists on minimal number of parameters, and the ability of outlier detection. However even though it has good efficacy for large data bases, this property is not fulfilled in high dimensional spaces. On the other hand, the main advantage og K means is it;s simplicity and effectiveness on large data sets. The disadvantage is the reliance of the results on the initial assignments, that may lead to not actually finding the optimal cluster allocation at the end of the process. The 2 step algorithm consists of a modified k means algorithm in the first stage, then, for the second one, the authors have used an agglomerative hierarchical clustering technique

During tests, algorithms were examined depending on number of dimensions (attributes), efficacy in outlier detection, scalability and behavior in case of standardized and non-standardized data. The considered features of the customers were age, income, deposit, credit, profit/loss As shown in the tables, the K Means performs better than the DBSCAN, but the combination of the algorithms produces more accurate results. Even though K Means is efficient and fast, it is noise sensitive, and it does not produce outliers, this even a small number of unusual objects in the data set may drastically alter the results, compared to the other two.

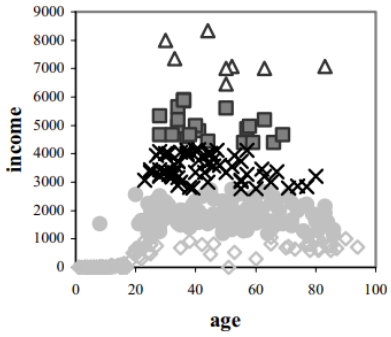


Figure 1. K-means algorithm (k=5).

Figure 3.1: K Means Clustering

Regarding scalability, tests were done on each algorithm, modifying the number of clients that were taken into consideration. From these results we observe that K Means is still the most efficient algorithm. DBSCAN is a lot slower, and by combining them, the results is more than the sum of each method taken separately

After some standardisation of data, the authors got even more accurate clusters, given that by standardisation we are smoothing the differences of data attributes values

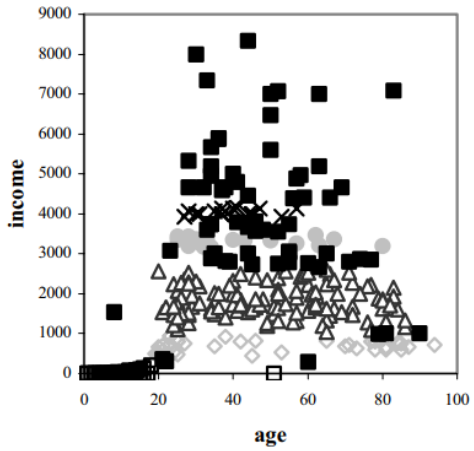


Figure 3. DBSCAN algorithm

Figure 3.2: DBSCAN Clustering

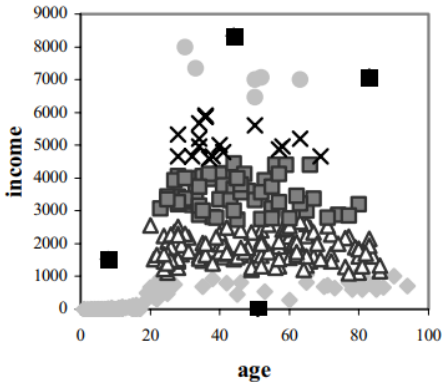


Figure 2. Two-phase clustering algorithm(k'=5)

Figure 3.3: 2 Step Clustering

Table 7. Run times in seconds

Number of objects	K-means	Two-phase	DBSCAN
100	0,1563	0,7813	0,1563
500	0,625	3,4688	2,3438
1000	1,25	15,375	9,0625

Figure 3.4: Scalability

In the end, the authors of the paper concluded that there may not be only one algorithm that is the proper way of grouping customers, but more methods combined may result in better results, with the expense of scalability

Customer segmentation in e-commerce and the importance of it.

The other article brought into discussion is regarding a review on customer segmentation techniques in

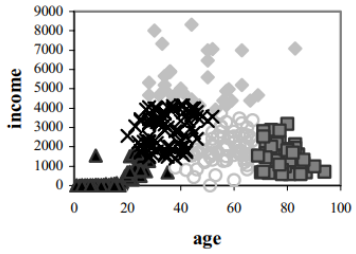


Figure 4. K-means with standardization

Figure 3.5: Standardisation

e-commerce [2]. This paper discusses about the importance of the customer segmentation in maintaining, gaining and satisfying the customers, thus obtaining more revenue for the e-commerce businesses. Moreover, is discusses several variables on which to perform different customer segmentation methods such as: product, transaction, geographic, hobbies and page viewed variables. Furthermore, the methods on which these variables are used are pointed out as follows: Business Rule, Quantile membership, Supervised Clustering, Unsupervised Clustering, Customer Profiling, RFM Cell Classification Grouping, Customer Likeness Clustering and Purchase Affinity Clustering [2].

Customer Segmentation Intelligence is brought into light by pointing out the improvements added to the market in offering products or services that meet the need of each group of customers. Moreover, the process of categorize or classify an item into a group with similar characteristics is used to determine the similarities between the customers by segmenting the records from the customer database [2]. In the process of segmenting customers there were categorized 2 type of data, internal, which are customer registration, profile, purchase history and external are media browsing, surveys and market search, cookies, web and social media analysis [2].

There are some methods stated for dealing with customer segmentation as follows:

1. *Business Rule:* describes the arrangement of customers into specific groups based on a predefined class such as:

    - Grouping based on demographic data, mainly age, gender, income etc [2].
    - Grouping based on customer interaction with the company based on purchase pattern such as the type of product or service provided or RFM (R is Recency (when customer last shopped), F is Frequency (how often the customer shops) and M is Monetary (how much the customer spends)) [2].

2. *Quantile Membership:* it uses the RFM methodology where recency will be divided into groups

of time interval which are classified with labels to determine the very valuable and low-value customer, than maps 2 components of RFM to a table. Finally, the result can be interpreted to check whether a discount or promotion should be offered to an old customer to come back [2].

3. *Supervised Clustering with decision tree:* uses a specific target to predict differences in independent variables. Data utilized in this method is previous purchase pattern and customer demographic. As stated in the paper this method is only able to show one aspect of the customer behaviour [2].

4. *Unsupervised Clustering:* it uses any number of customer attributes then it measures the similarities between the customers. Attributes of the customer uses the Euclidean distance then cluster the customer using k-means algorithm [2].

Based on the above researches, customer segmentation methods can be classified into: simple techniques, statistical data, target techniques and Unsupervised techniques. A shown below:

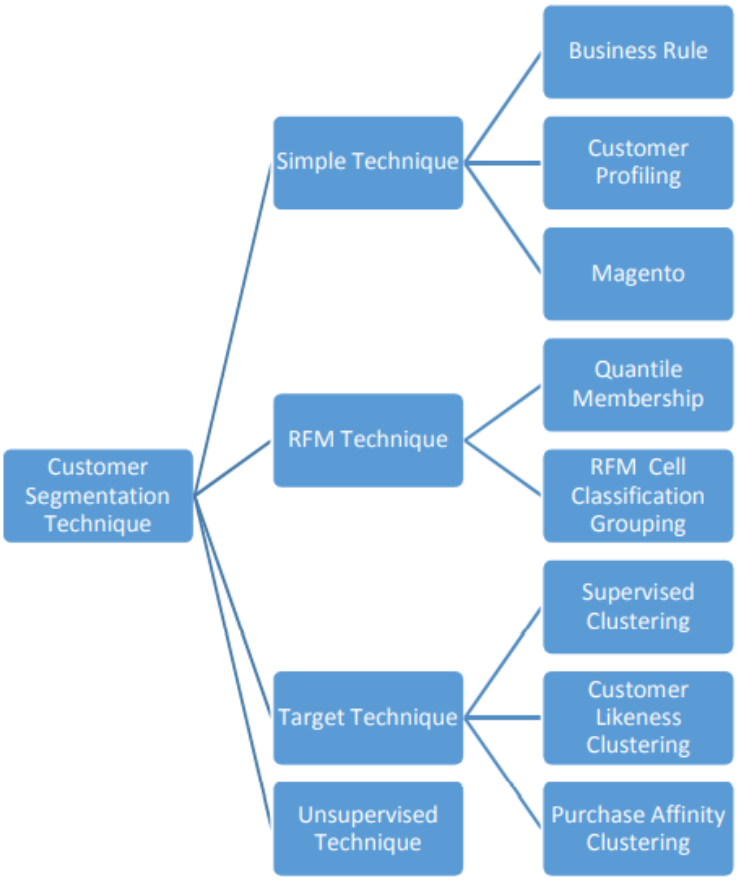


Figure 3.6: Customer Segmentation Classification

One last aspect discussed in the paper is the process of customer segmentation. They discuss about identifying the high profitable customer groups and gathering data from data-mining. After that collecting more 'sensitive' data such as demographic data, transaction data, and promotional data. Lastly choosing the right algorithm and based on that compute other metrics [2]. As shown below:

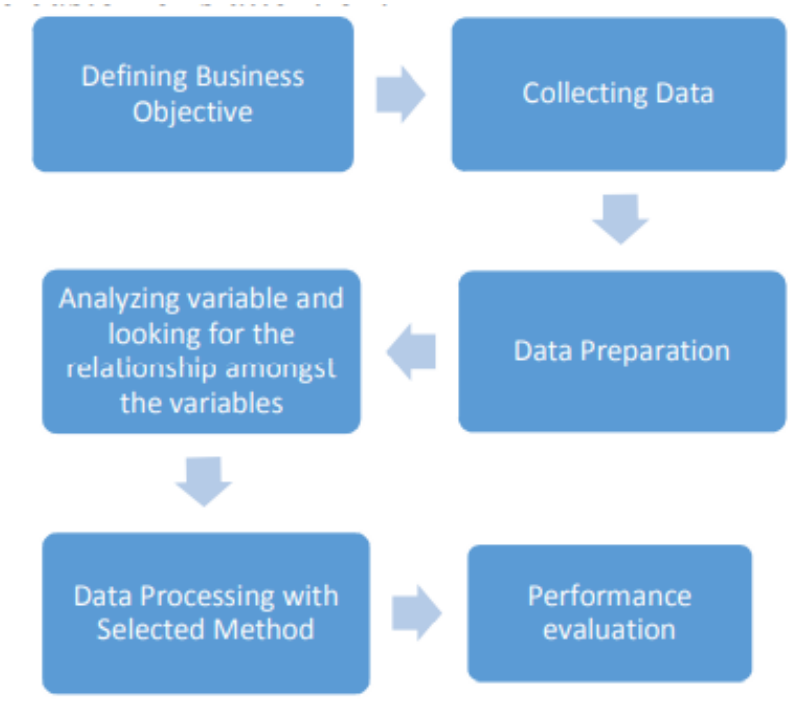


Figure 3.7: Customer Segmentation process

In the end, the authors of the paper concluded that customer segmentation is a good way to improve the customer satisfaction, communication and it can bring new customers to the business. Moreover, business which used customer segmentation got an increase in profits.

Our proposed solution

Our problem is regarding the grouping of clients of a e-commerce site, in order to offer them the best deals that we can find for each particular group/individual

As we do not posses that much knowledge regarding the subject of artificial intelligence algorithms, we have settled on only trying the K-means Clustering algorithm on our particular set of data, and trying to find out the best parameters to use and feature engineer our data for the best performance. As a future improvement we might try out combining 2 different algorithms in the prospect of actually improving upon our first results

Chapter 4

Investigated approach

Describe your approach!

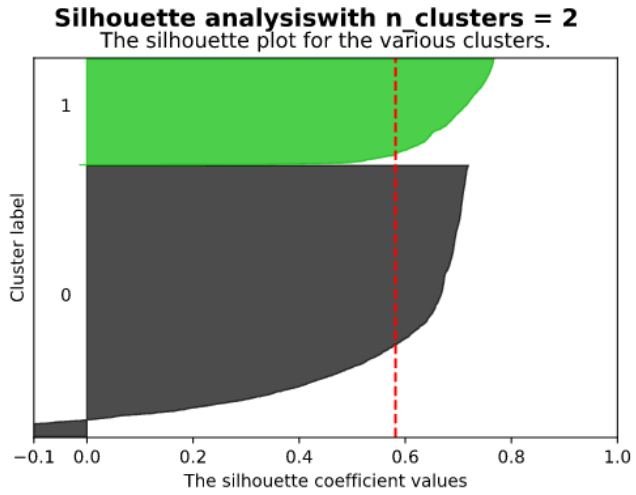
Describe in reasonable detail the algorithm you are using to address this problem. A psuedocode description of the algorithm you are using is frequently useful. Trace through a concrete example, showing how your algorithm processes this example. The example should be complex enough to illustrate all of the important aspects of the problem but simple enough to be easily understood. If possible, an intuitively meaningful example is better than one with meaningless symbols.

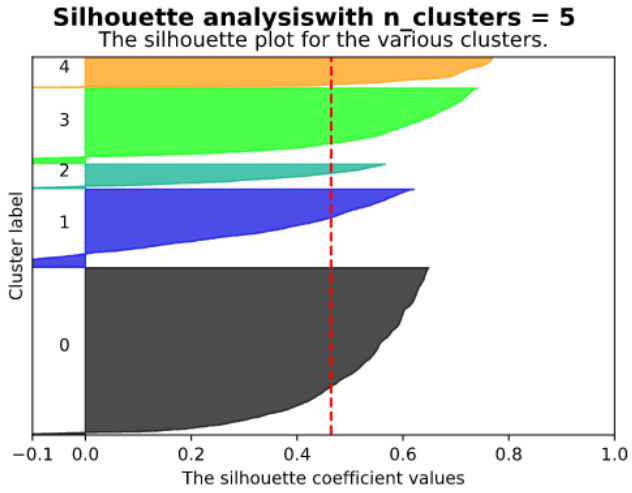
Our approach revolves around using K Means Clustering algorithm. After some research we concluded that it may be the best solution for our current problem.

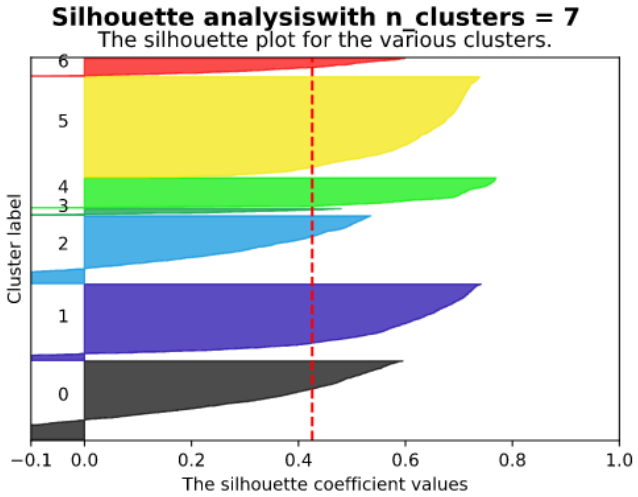
Firstly, we have cleaned up the data that may have been just invalid. We removed the rows that had missing customer IDs and product descriptions, immediately follow by the removal of negative value invoices, as they are most probably returned invoices, because we are interested in what the user ordered and not what he/she returned. Furthermore, in the cleanup process we removed maximum values for quantity and price, as they may be related to wholesalers, again, that type of customer not being of importance to us. As the products that have stock code that starts with a letter, and they are not fitting the rule of 5 digits may be variations of the same item, we have also dropped them, as they are unusual items that may not really be purchased at a given time from the site.

After we have cleaned up the data, we have started to "feature engineer" our data, as for our recommendations we needed additional attributes. We have obtained a set of intermediate attributes, like Total Price per Line item, Total Price per Invoice, Time between Orders, Total Quantity per Order, all of these so we can eventually get to the RFM(Recency, Frequency, Monetary value) segmentation.

Using these new features, we have employed the K Means Clustering algorithm, using 5 clusters, and have successfully managed to obtain some modest results, with well defined centroids. The evaluation was done using the Silhouette analysis applied to a number of clusters between 2 to 7. Some of our testing with clusters, as given by the following graphs:







As a final method of obtaining some recommendations, and this is by no real means final, as it is quite primitive. We have gotten 10 users that are a part of the same cluster with any other given user, and we have recommended the first item that was part of the most recent invoice of the selected users. Because the users are similar by their type of purchases, we have concluded that users of similar clusters may want the same type of products, henceforth this logic was the one applied by us

	CustomerID	Frequency	Monetary Value	Recency	Clusters	total_orders	avg_spend
0	12347	7	4310.00	1	2	7	615.714286
1	12348	4	1437.24	74	0	4	359.310000
2	12349	1	1457.55	18	0	1	1457.550000
3	12350	1	294.40	309	1	1	294.400000
4	12352	7	1385.74	35	2	7	197.962857
...	...	...	...	...	...	...	...
4289	18280	1	180.60	277	1	1	180.600000
4290	18281	1	80.82	180	4	1	80.820000
4291	18282	2	178.05	7	0	2	89.025000
4292	18283	16	2088.93	3	2	16	130.558125
4293	18287	3	1837.28	42	0	3	612.426667

4294 rows × 7 columns

Chapter 5

Application (numerical validation)

Explain the experimental methodology and the numerical results obtained with your approach and the state of art approache(s).

Try to perform a comparison of several approaches.

Statistical validation of the results.

For this application we are using in the backscenes a typical K-Means algorithm for segmenting our users into different Clusters. These Clusters are made using our pre-given data that we got from the UCI website, along with some custom made features that we manually created. These features have in mind the RFM (Recency Frequency Monetary value) model that is mostly used. We have tried using different numbers of clusters, so we could test out the Silhouette and the ARI (adjusted rand index) values, and came up with the proper number of 4 clusters.

5.1 Methodology

- Well there is no surefire way of testing our hypothesis, as we are not actually recommending items to a real client, but we are confident in our algorithm that it found real types of customers, and in a real world situation these recommendations will hold up
- As for the scientific part of the question of evaluation, we tried measuring the Silhouette coefficient, the elbow method and the value of the ARI . As the silhouette coefficient was decreasing as the ARI was rising, we had to balance them out and came up with middle ground, or in this case a middle number of 4 clusters
- What are the dependent and independent variables? //TODO

- For our test data we have used the data set offered by UCI (the Machine Learning Repository), that tracked the invoices of a large number of customers of a retail shop on a period of almost one year. The dataset is realistic, as each invoice holds all the proper information of an actual purchase of a customer. We have analyzed it and came up with the conclusion that some adjustments could be made, for example clearing out noise data, not-proper invoices, duplicates, etc. Competing methods, for example is manually going through each and every invoice, collecting it's data, finding out each customer's purchases and calculating the RFM coefficient. We are sure that using even an unsupervised learning algorithm is more efficient that the classical way of doing things

5.2 Data

The UCI OnlineRetail data is a straightforward collection of invoices that are given to us in an easy to use, organised spreadhsheet. Contents

- Invoice Number
- Stock Code - These were tricky, because they were not in the same format, some of them even had letters at the end of the code. We supposed that they actually were variations of the same product and treated them accordingly
- Item Description
- Quantity - We had to check if there were negative quantities, as they were probably returns of the clients that may have not enjoyed the product
- Invoice Date - very useful for Frequency feature
- Unit Price - Again, the same issue with the Quantity, treated it accordingly
- Customer ID - Some of the rows were having missing customer ID, maybe they were products bought from the actual, physical retail shop, but as we could not map them to a client, we tried to ignore them
- Country - This was one of the less useful features that the dataset offered

5.3 Results

As stated in the algorithm presentation, we have tried with a different number of clusters and checked for their Silhouette coefficient, resulting with our decision to stick to only just 5 clusters.

Of course there may be better methods of obtaining our results, but for this project/ experiment we were contempt with our findings. As a future work we can try out different ways, but that is a mission for another day

5.4 Discussion

As stated before this was a small toy project that gave us an insight in the world of recommender systems. Our method may be primitive, but we have found decent enough results for us to propose this idea.

On the dataset that we have used, using the number of 5 clusters gave us what we thought to be the most balanced and real customer segmentation. Based on that we made up our idea of how to suggest item. The dataset not being one of the most complete we needed to clean it and add some features that we have engineered by ourselves

Chapter 6

Conclusion and future work

The main strength of the approach is that it is a fairy simple one, and with a bit of parameter "play" and analysis using some measurements of the previously mentioned coefficients (Silhouette and ARI) we can find a decent segmentation method.

The main drawback of our method is that it's (like most unsupervised algorithms) having issues when dealing with the cold -start problem. This problem is an infamous one, that relates to too little data being given to the algorithm and it not being able to correctly classify the elements. This can not really be overcame, we just need to patiently wait for the dataset to be large enough for us to use it. Only then we can hope for correct/ appropriate results

Nonetheless we have proved that the K Means algorithm Customer Segmentation is an efficient way of grouping similar customers into classes, and retrieve product suggestions as close as a 1 to 1 relationship between a shop and its client

These results are not groundbreaking, as they are already used by most e-commerce companies, even though their approaches are fairly more complex or better suited for their business model. This approach, being simple enough can be actually implemented by new or small retails for a better relationship with their clients.

Bibliography

[1] Zakrzewska, D., Murlewski, J. (2005). Clustering algorithms for bank customer segmentation. 5th International Conference on Intelligent Systems Design and Applications (ISDAâ05). doi:10.1109/isda.2005.33

[2] Juni Nurma Sari, Lukito Edi Nugroho, Ridi Ferdiana, P. Insap Santosa. *Review on Customer Segmentation Technique on Ecommerce*. American Scientific Publishers, Advanced Science Letters Vol.4, 400-407, 2011.