Machine Learning Engineer Nanodegree

Capstone Project

Esteban Dib Puelma
September 22nd, 2018

I. Definition

Project Overview

Day to day, retailers make decisions about which products they need to sell, at what price, where to offer them, in what colors, and many more. These decisions are sometimes blindly made, obeying just rules of thumbs or the managers' guts. Luckily for them, we now have tools to help in this decision making process, and one of them is to segment customers and know them much better than we could without these tools.

I live in Chile, where retail competition has being local for many years, but it is turning more and more global, so it is imperative to apply the best available tools. My dad's company is struggling with some of the judgments mentioned above, because they lack detailed and informed knowledge about the customers that are buying on the stores and I want to make this project the starting point of a Data Strategy that would help compete through the provision of the best possible products and experience to the customers.

In terms of academic research and other studies considering customer segmentation. The closest example is the unsupervised learning project of this Nanodegree, which is also available at Kaggle[1]. Another research that aims to do something similar is a paper called "A Comparative Quantitative Analysis of Contemporary Big Data Clustering Algorithms for Market Segmentation in Hospitality Industry"[2]. They used a variety of density-based algorithms, including DBSCAN, to segment customers from the hospitality industry. They found that the best choice for their problem was EnDBSCAN to identify nested and embedded clusters, but OPTICS was better to identify adjacent nested clusters.

Problem Statement

The problem that wants to be solved in this project, is that the company knows nothing about their offline customers. They have some data: identification number, sex, transactions and others, but this data has never been transformed into knowledge. Currently, a customer goes to a store, buys something and that transaction goes to a database and is forgotten. This company has stores all around Chile, and the managers attempt to put the best suitable

[1] https://www.kaggle.com/samratp/creating-customer-segments-unsupervised-learning
[2] https://arxiv.org/pdf/1709.06202.pdf

products available in each location, but they not always succeed in this mission. The aim is to make a segmentation of the customers, to give managers tools to take more informed decisions, and in the end, increase sales.

To solve the problem a Gaussian Mixture Model and DBSCAN will be used to define clusters based on the age of the customers, and the amount of money and total items they bought in a period of two years, from January 2016 to August 2018. My hope is to find some clearly defined segments to make some marketing and operational decisions within the company. For example, is it could be found that people in their 20s tend to buy cheap stuff and people in their 50s go for more expensive shopping.

Metrics

To evaluate both models, 2 classical metrics will be used and one more qualitative approach. The classical metrics are the silhouette score, which is calculated with the mean of the distance between each point in a cluster and the mean distance to the nearest cluster, and will allow to understand if each cluster is very different from the others or not so much; and the Calinski-Harabasz index of the clusters, which evaluates the cluster validity based on the average between-cluster sum of squares and the average within-cluster sum of squares, and will allow to understand how related clusters are among themselves combined with their relationship between each other[3]. The "qualitative" approach will be to look at the graphs that both algorithms create when clustering the data, and understand visually the flaws and virtues each one has.

The metrics chosen are important because the first one will give information about how different are customer segments from each other, the second one will give information about the integrity of each customer segment and how related each customer is with the cluster and the third one, the visualization, will give an easy way to understand the distribution of the clusters and how the dimensions analyzed impact the segmentation.

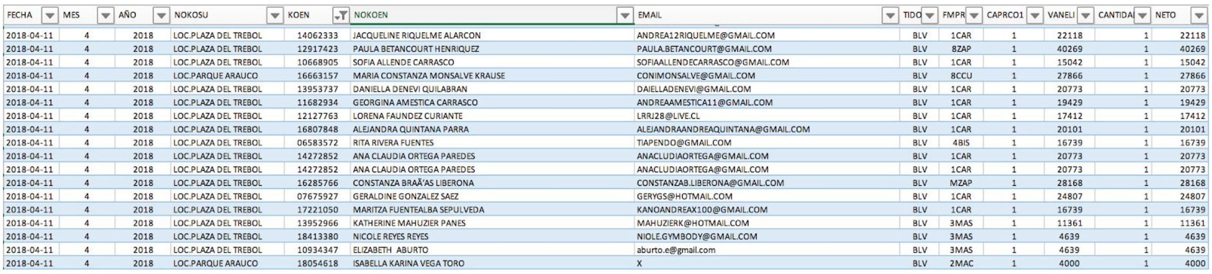
II. Analysis

Data Exploration

The data used to solve the problem comes from two main sources:

The first one will be given by the company and is the history of purchase from 4 of the company's stores. This dataset includes the customer's email and RUT (Unique Tax Number in Chile) which gives a estimation of the age of the person and is a unique identifier of who is making the purchase, the date of the transaction, the amount of items, the price of each item and in which store the purchase was made. It has 9528 rows of data in total. This dataset is available in the .zip file of this project and a sample is shown below:

[3] http://datamining.rutgers.edu/publication/internalmeasures.pdf



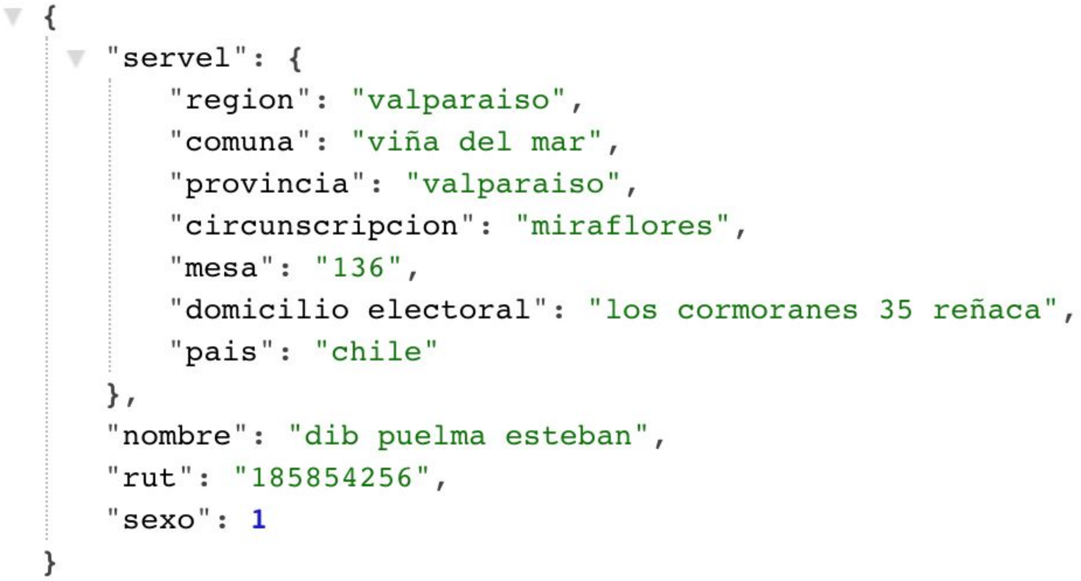
FECHA	MES	AÑO	NOKOSU	KOEN	NOKOEN	EMAIL	TIDO	FMPR	CAPRCO1	VANELI	CANTIDA	NETO
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	14062333	JACQUELINE RIQUELME ALARCON	ANDREA12RIQUELME@GMAIL.COM	BLV	1CAR	1	22118	1	22118
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	12917423	PAULA BETANCOURT HENRIQUEZ	PAULA.BETANCOURT@GMAIL.COM	BLV	8ZAP	1	40269	1	40269
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	10668905	SOFIA ALLENDE CARRASCO	SOFIAALLENDECARRASCO@GMAIL.COM	BLV	1CAR	1	15042	1	15042
2018-04-11	4	2018	LOC.PARQUE ARAUCO	16663157	MARIA CONSTANZA MONSALVE KRAUSE	CONIMONSALVE@GMAIL.COM	BLV	8CCU	1	27866	1	27866
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	13953737	DANIELLA DENEVI QUILABRAN	DAIELLADENEVI@GMAIL.COM	BLV	1CAR	1	20773	1	20773
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	11682934	GEORGINA AMESTICA CARRASCO	ANDREAAMESTICA11@GMAIL.COM	BLV	1CAR	1	19429	1	19429
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	12127763	LORENA FAUNDEZ CURIANTE	LRRJ28@LIVE.CL	BLV	1CAR	1	17412	1	17412
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	16807848	ALEJANDRA QUINTANA PARRA	ALEJANDRAANDREAQUINTANA@GMAIL.COM	BLV	1CAR	1	20101	1	20101
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	06583572	RITA RIVERA FUENTES	TIAPENDO@GMAIL.COM	BLV	4BIS	1	16739	1	16739
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	14272852	ANA CLAUDIA ORTEGA PAREDES	ANACLUDIAORTEGA@GMAIL.COM	BLV	1CAR	1	20773	1	20773
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	14272852	ANA CLAUDIA ORTEGA PAREDES	ANACLUDIAORTEGA@GMAIL.COM	BLV	1CAR	1	20773	1	20773
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	16285766	CONSTANZA BRAÃ'AS LIBERONA	CONSTANZAB.LIBERONA@GMAIL.COM	BLV	MZAP	1	28168	1	28168
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	07675927	GERALDINE GONZALEZ SAEZ	GERYGS@HOTMAIL.COM	BLV	1CAR	1	24807	1	24807
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	17221050	MARITZA FUENTEALBA SEPULVEDA	KANOANDREAX100@GMAIL.COM	BLV	1CAR	1	16739	1	16739
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	13952966	KATHERINE MAHUZIER PANES	MAHUZIERK@HOTMAIL.COM	BLV	3MAS	1	11361	1	11361
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	18413380	NICOLE REYES REYES	NIOLE.GYMBODY@GMAIL.COM	BLV	3MAS	1	4639	1	4639
2018-04-11	4	2018	LOC.PLAZA DEL TREBOL	10934347	ELIZABETH ABURTO	aburto.e@gmail.com	BLV	3MAS	1	4639	1	4639
2018-04-11	4	2018	LOC.PARQUE ARAUCO	18054618	ISABELLA KARINA VEGA TORO	X	BLV	2MAC	1	4000	1	4000

Figure 1: data sample

The fields mean the following:

- Fecha: date of the purchase
- Mes: month of the purchase
- Año: year of the purchase
- NOKOSU: store location
- KOEN: RUT of the customer
- NOKOEN: full name of the customer
- EMAIL: email of the customer
- TIDO: type of sale (ticket, bill, purchase order)
- FMPR: type of the item (wallet, purse, etc)
- CAPRCO1 and CANTIDAD: amount bought
- VANELI and NETO: total money spent

The second one is a public API hosted in https://api.rutify.cl/, which gives you the address and sex of a person based on it's RUT. An example of the API response is shown below:



```
{
  "servel": {
    "region": "valparaiso",
    "comuna": "viña del mar",
    "provincia": "valparaiso",
    "circunscripcion": "miraflores",
    "mesa": "136",
    "domicilio electoral": "los cormoranes 35 reñaca",
    "pais": "chile"
  },
  "nombre": "dib puelma esteban",
  "rut": "185854256",
  "sexo": 1
}
```

Figure 2: API example

The sex of the person is in the field "sexo" and the address is in the field "domicilio electoral". These fields were going to be used in the clustering algorithm, but it was not possible for to get the family income based on only the address and the sex did not give any interesting information to the clusters.

To clean the data and make a better and focused segmentation, some data manipulation was made.

1.- Group the transactions of each person to get the amount of items that person bought and the amount of money he or she spent in the whole period of analysis.

2.- Drop all the features that did not contribute to the segmentation, such as date, year, month, rut, purchase_type among others.

3.- Remove all the ids generated when the data was consolidated. This was done in the data_cleaner.py file and consisted on getting the age of each customer and leaving the wholesalers out of the customer base.

4.- Remove all rows with NA values.

This data processing left 3 important features for each client: age, number of items bought and amount of money spent. More details about this can be found in the jupyter notebook capstone.ipynb.

The outlier removal is explained in the Data Preprocessing section.

Exploratory Visualization

The visualization of the data was used to see the distribution of each feature and how each one related to one another. For this, two graphs were used, a scatter matrix and a heat map. Both images are shown below.

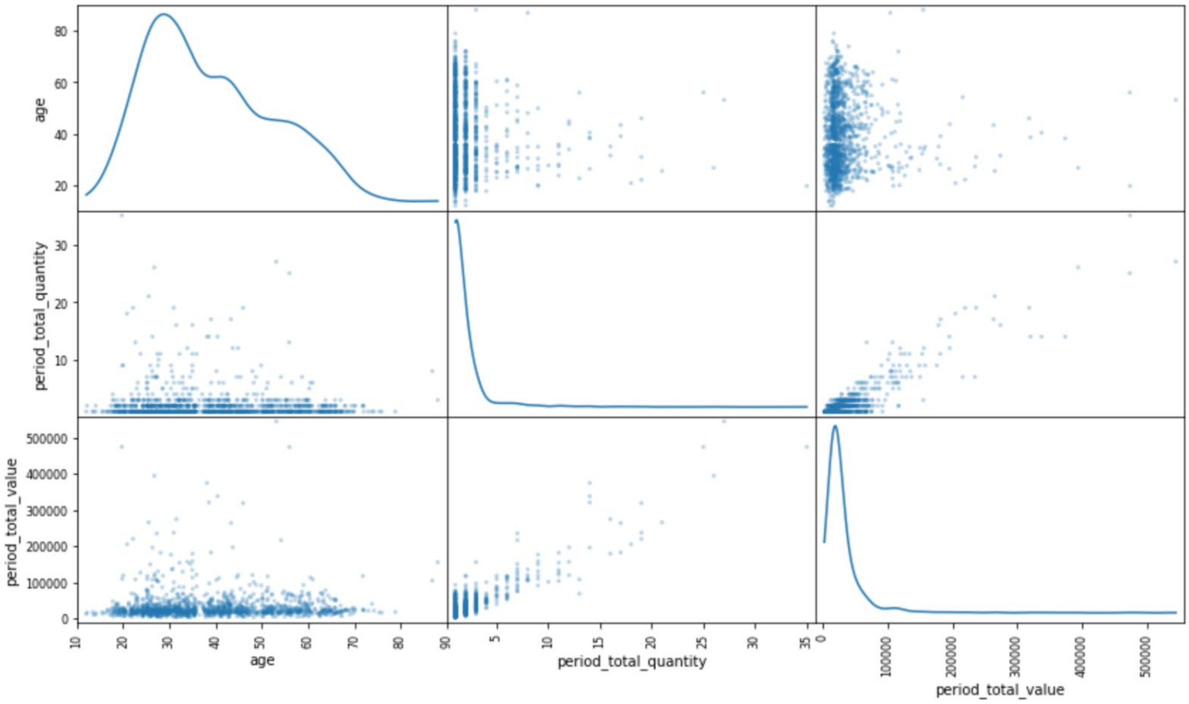
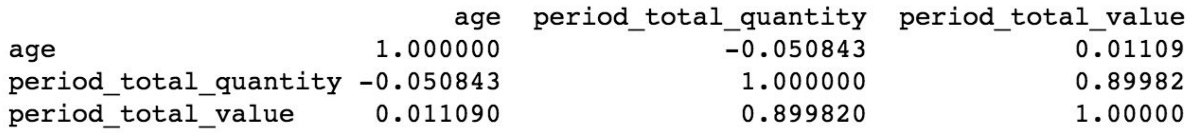


Figure 3: Scatter matrix



```
                            age  period_total_quantity  period_total_value
age                    1.000000              -0.050843             0.01109
period_total_quantity -0.050843               1.000000             0.89982
period_total_value     0.011090               0.899820             1.00000
```

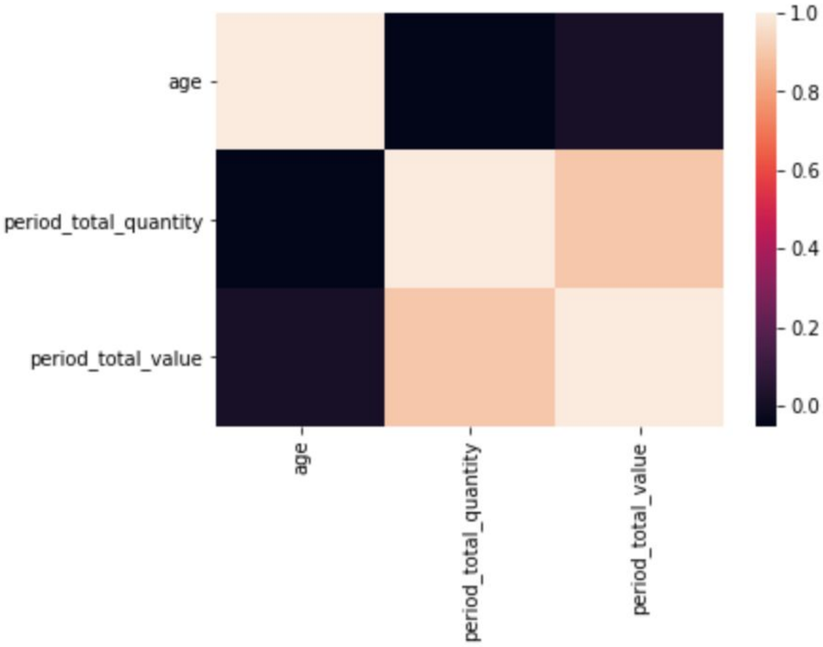


Figure 4: Heat map

As we can see, there is a huge correlation between the amount of money spent and the amount of items bought. This is pretty normal, as more items mean more money, but is important to have both features so we can differentiate customers that buy expensive things from the ones that buy cheap things in a bulkier way.

It is interesting that in the scatter matrix is it possible to see some outliers values in the amount of money spent (period_total_value) and in the amount of items bought (period_total_quantity). This outliers were deleted from the dataset.

Algorithms and Techniques

As this is a segmentation problem, clustering algorithms will be used. Specifically two of them; a Gaussian Mixture Model, because, as the data comes from human population and human interaction, it should be composed of a finite number of Gaussian distributions; and a DBSCAN to see if a density based clustering technique is adequate for this type of problems. Both algorithms will take the data points and assign them to clusters.

The first one will use k-means or random positions to establish the first centroids and then iterate to move those centroids in order to try to find a better clustering. The second one will start in a random location and use the epsilon and the minimum number of points to decide whether a group of points belongs to a cluster or not[4].

[4] https://www.naftaliharris.com/blog/visualizing-dbscan-clustering

In terms of hyper-parameters, for the GMM, the metrics will be used to find the best number of components and then use a grid search algorithm to find the best type of covariance between full, tied, diagonal or spherical, the best type of parameter initialization between k-means or random, a good value for the convergence threshold and a good value for the maximum amount of iterations. For the DBSCAN a grid search algorithm will be used to find a good value for the maximum distance between to points to be considered of the same cluster (epsilon) and a good value for the minimum amount of samples in the neighborhood of a data point for it to be considered a core point (minimum samples).

Benchmark

In qualitative terms, to consider the analysis successful, the clusters created by the model should help take commercial decisions and should represent types of customers. Today, the company is guessing that they have 4 types of clients based on an psychographic description and this could be reflected in their expenditure habits. These types are: sport casual, basic classical, sophisticated classical, and ethnic. If the algorithm is capable of accurately describe these four types or make a better segmentation of the clients, it will beat the human created model. Thus, better than the benchmark model.

In quantitative terms, the algorithms will be compared with a k-means algorithms with the best number of clusters based on silhouette score and the Calinski-Harabasz index, but with no other parameters optimized. This will help get some insights on whether the analysis made in this project is just a minor improvement or a much more fascinating approach. If the k-means algorithm achieves a better value in the metrics, then the use of the proposed algorithms is not valuable.

III. Methodology

Data Preprocessing

The first preprocessing of the data was done to complete the dataset. The initial idea was to incorporate the sex and socioeconomic status of the customer to the segmentation, and although it was not possible to include these characteristics, it was possible to obtain the data for at least the sex and the address of the person, which was going to be used to calculate an estimate of the socioeconomic status. A script written in python was used to achieve this, found in the data_cleaner.py file. The objective of each function is the following:

- clean_majorists: clean the wholesale customers from the database, these are considered to buy 3 or more items in only one purchase.
- get_age: get the age of the customer from the RUT using a regression that says that $year\ of\ birth\ =\ 1930.3\ +\ 3.46\ *\ \frac{Rut}{10\hat{}6}$ [5]. With this, $age\ = current\, year - year\ of\ birth$
- get_verify_number: add a verification number to each RUT

[5] https://twitter.com/rodrigowagnerb/status/410907897019629568?lang=es

- get_data: use the verification number and the RUT to get the address and the age of the customer
- clean_columns: rename the columns on the dataset for a better readability

After having the complete dataset and readable column names, the cleanup mentioned in the Data Exploration section was made.

In terms of preprocessing made for the algorithm to work better, feature transformation was used, applying the natural logarithm function for every feature in every datapoint. After that an outlier detection was made using the interquartile range, there were many points outside the range, but only seven points where sufficiently away from the range to be removed from the dataset. Last, but not least, a feature transformation was made using PCA. The three features were turned into three dimensions at first to understand how much of the variance each component represented and after that the decision was made to use two dimensions that meant a total explained variance of 88.79%. From the two chosen dimensions, one of them explained the variance from the total quantity and the total expenditure, and the other one explained mainly the variance that came from the age. This can be seen in the following figure

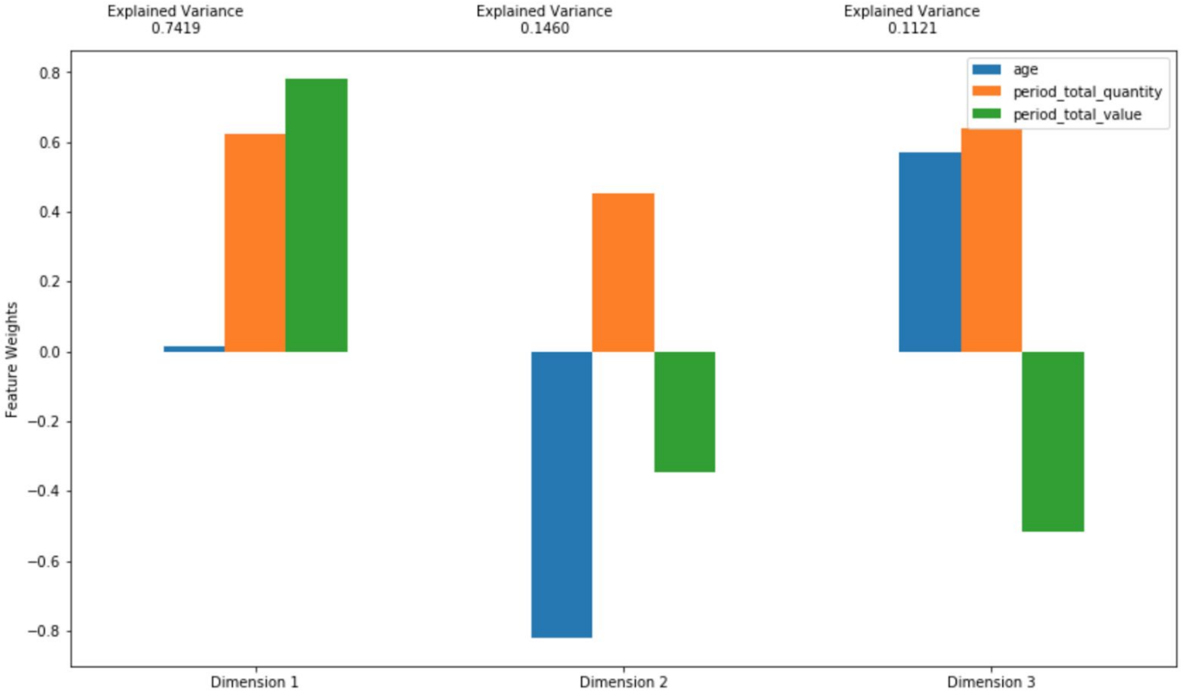


Figure 5: explained variance for each dimension

Implementation

As said before, the analysis made use of two algorithms to solve the problem, a Gaussian Mixture Model and a DBSCAN Model. Both of them were compared to a K-means model to understand if an improvement was being made.

Each of the model received the data already processed with feature scaling and feature transformation, so there were only two dimensions and all of the values were passed to the natural logarithm function before entering the clustering process.

At first, the idea was to use silhouette score and inertia as metrics of how good the algorithms were working on the clustering problem, but sklearn does not provide an implementation of the inertia score, so a decision was made to change this to the Calinski-Harabasz score, which is not the same, but can give us similar information.

One of the biggest complications was to understand how to use the data for the clustering. At first the intention was to use the sex, quantity, purchase amount, commune, age, category of the item and in which store was the purchase being made. But soon it was realized that such a broad clustering method was not going to work and that there were too much discrete variables confusing the algorithms in their clustering processes. This is why, after some trial and error, the approach was to only use the age, total quantity of items and total amount of purchase to segment the customers expenditure habits and relate them to their ages.

Refinement

For the Gaussian Mixture Model and to create the benchmark k-means model, a measurement of the silhouette score and the Calinski-Harabasz index was made for each number of clusters, from 2 to 50. The graphs are presented below

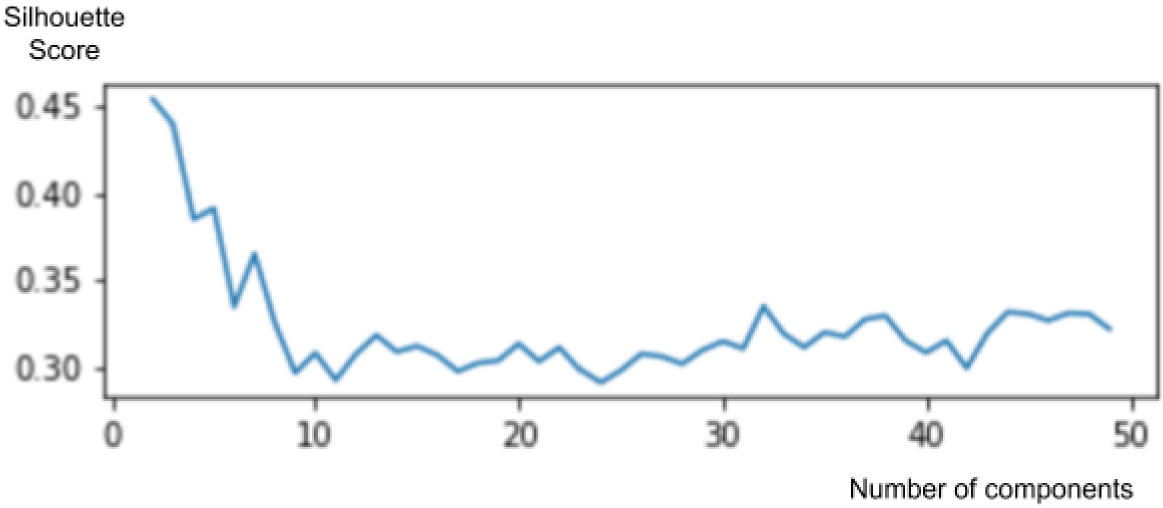


Figure 6: Gaussian Mixture Model silhouette score v/s number of components

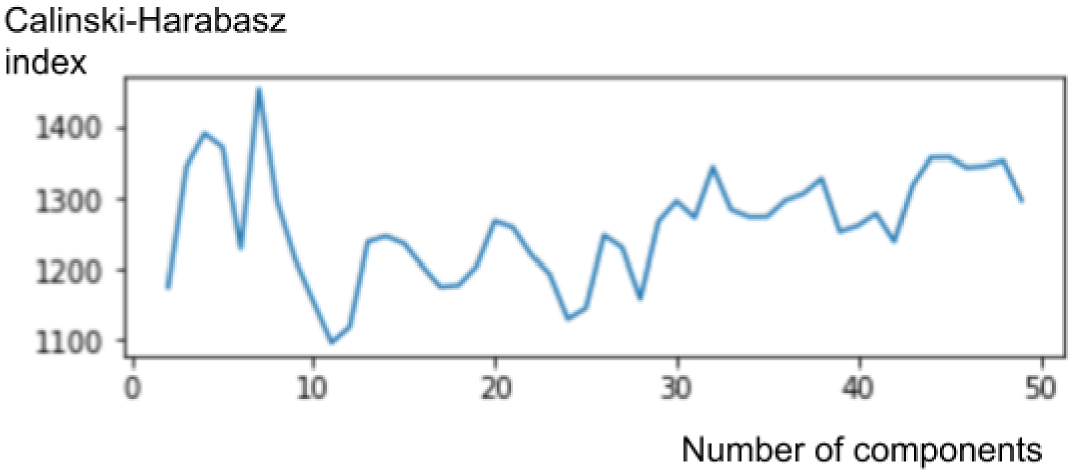


Figure 7: Gaussian Mixture Model Calinski-Harabasz index v/s number of components

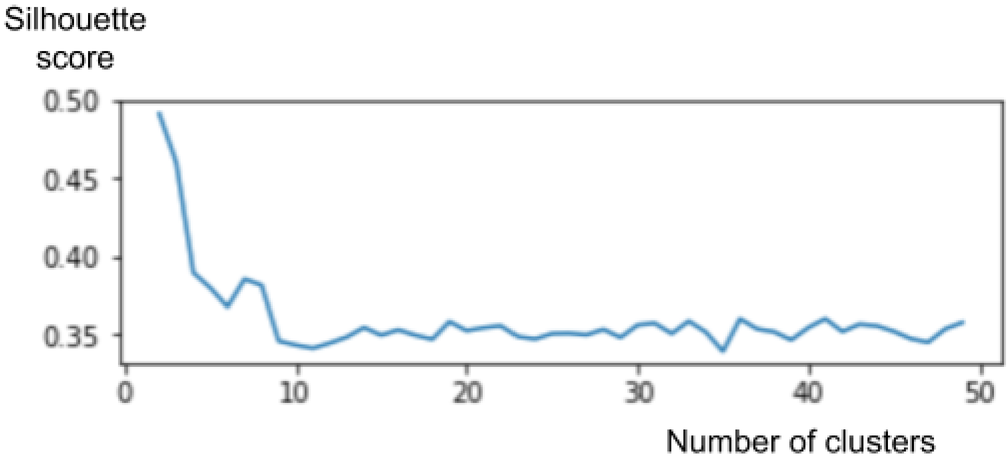


Figure 8: K-means silhouette score v/s number of clusters

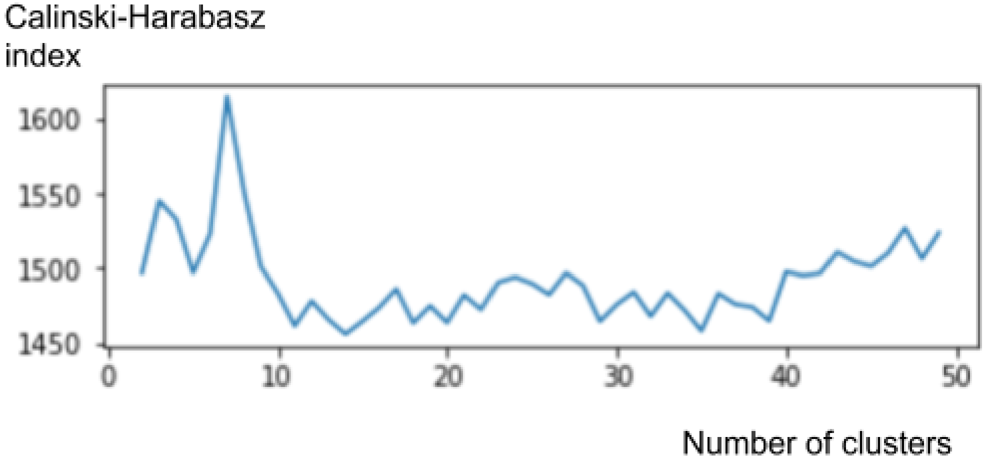


Figure 9: K-means Calinski-Harabasz index v/s number of clusters

After defining the best number of components, the hyper-parameters of the GMM were refined with a grid search, varying the 4 types of covariance, the 2 types of parameters initializations, 4 different values for the convergence threshold and 3 different values for the maximum amount of iterations. Based on the analysis, the best silhouette score is presented when the covariance is full, the convergence threshold is 0.1, the max iterations made by the algorithm are 50 and the initial distribution is made by k-means. The best Calinski-Harabasz score is presented in the same circumstances, but with the covariance set to spherical. It is important to notice that if we continue decreasing the value of the max iteration up until 1, we achieve the best scores. This means that k-means, the initialization of the clusters, is a better model.

For the DBSCAN, a manual grid search was used, considering 8 different values for the eps and seven different values for the minimum samples. A measurement of the silhouette score and the Calinski-Harabasz index for each combination helped chose the ones with the best values. Based on the analysis, the best silhouette score is presented with two clusters, when eps is 0.9 and the minimum number of samples is 60. The best Calinski-Harabasz score is presented with two clusters, when eps is 0.5 and the minimum number of samples is 60.

IV. Results

Model Evaluation and Validation

The following values were achieved by each proposed model

	Gaussian Mixture Model	DBSCAN (eps=0.9, min_samples=60)	DBSCAN (eps=0.5, min_samples=60)
Silhouette Score	0.3835	0.6758	0.6337
Calinski-Harabasz Index	1590.6632	64.3602	741.0070
Nº clusters/components	7	2	2

Table 1: GMM v/s DBSCAN

As seen in the table, the Gaussian Mixture Model has a much better Calinski-Harabasz index, which means that the points of each cluster relate a lot to each other and do not relate much with the points in other clusters. The DBSCAN has a much better silhouette score, which means that the clusters created by DBSCAN are more different between each other than the ones created by the GMM. When we look at the graphs of each clustering, shown below, we find that the DBSCAN algorithm did not actually cluster the data.

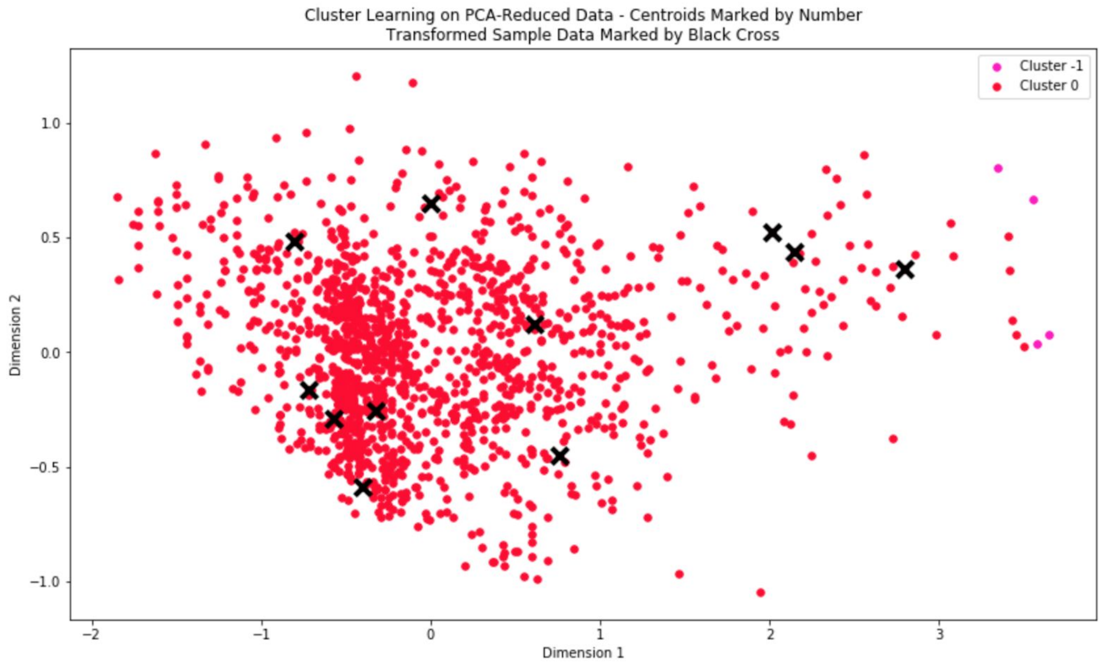


Figure 10: DBSCAN (eps=0.9, min_samples=60) clustering

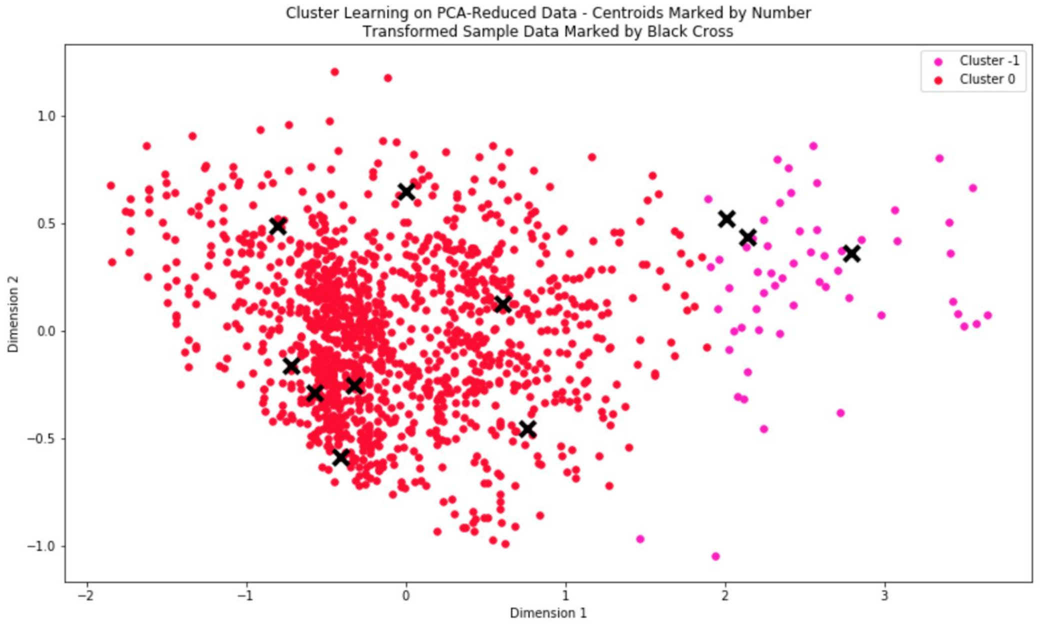


Figure 11: DBSCAN (eps=0.5, min_samples=60) clustering

It is true that the scores for the Gaussian Mixture Model are not incredibly good, but they seem reasonable to say that we can trust on the customer segmentation that was made, and that those segments are probably a good start to achieve a better understanding and this is reflected

Justification

In the following table, we can see that GMM did not have better results than the proposed k-means benchmark model.

	Gaussian Mixture Model	K-Means
Silhouette Score	0.3835	0.3855
Calinski-Harabasz Index	1590.6632	1615.5515
Nº clusters/components	7	7

Table 2: GMM v/s K-Means

Although grid search was used to look for the best hyper-parameters for both the Gaussian Mixture Model and the DBSCAN, neither of them could get a better score than the k-means model. This is disappointing, but encouraging at the same time. It means that with a simple and well-known clustering model it is possible to get interesting information from the company's customers.

In terms of the qualitative benchmark, about the 4 segments the company's managers think they have, this is analysis is much better. It tells that there are more than 4 groups, and that is only by analyzing the expenditure habits. For a forthcoming project, it would be really interesting to classify each product and see how is the relationship between the customers and the types of products. This will definitely improve much more the conceptual understanding of customers.

As it was mentioned before, this is a start point for a better customer understanding and despite this project did not give all the answers, it illuminated the path to follow. To further validate the analysis and solve the problem, the data from each store should be disaggregated. They are all located in different environments, with different people that can have different habits. So to make better decisions and solve the presented problem, it is necessary to understand the segments for each of the stores.

V. Conclusion

Free-Form Visualization

The first important quality to discuss, has to do with the problem, is that segmenting customers is not an easy task and that it can be made in an infinite number of dimensions. In this project, the first approach was to use all the data possible, feed it to the algorithm and hope for something magic to happen. This was incredibly wrong and is not at all the way to approach these types of problem. It is necessary to have a more narrowed down objective, such as, for example, get the relationship between the age of the customers and their expenditure habits. In the end, when segmenting customers, it is important to make more than one clustering, using different combinations of features and that is a paramount discovery made in this project.

The second important quality has to do with the solution, and are the segments found after the clustering. In figure 12 we can see in colors the different clusters, with their respective centroids surrounded by a circle. To see this is kind of magical, each dimension has its meaning and this graph tells the relation between each cluster, the relation between each point within a cluster and how each cluster looks.

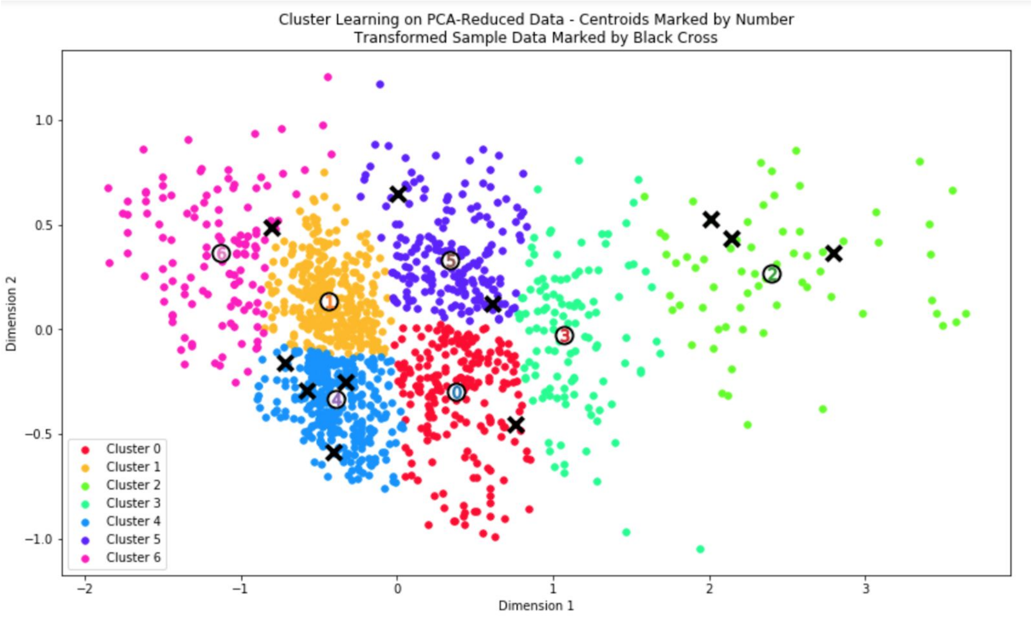
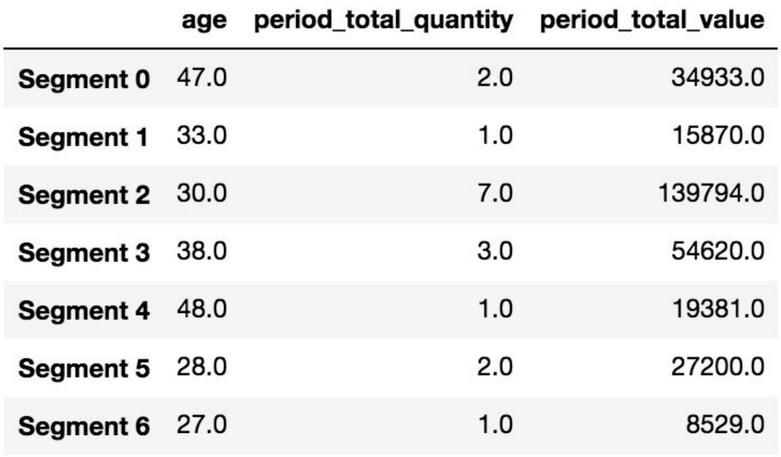


Figure 12: Gaussian Mixture Model clustering

Moreover, when each centroid is analysed, it is possible to get interesting insights. For example, table 2 shows the seven centroids and their represented values.



	age	period_total_quantity	period_total_value
Segment 0	47.0	2.0	34933.0
Segment 1	33.0	1.0	15870.0
Segment 2	30.0	7.0	139794.0
Segment 3	38.0	3.0	54620.0
Segment 4	48.0	1.0	19381.0
Segment 5	28.0	2.0	27200.0
Segment 6	27.0	1.0	8529.0

Table 3: discovered segments

From this is it possible to tell that:

- Segment number 0 represents older customers that have a moderate frequency as buyers.
- Segment number 4 represents older customers that do not buy much stuff in the store, but what they buy tend to be expensive.

- Segment number 5 represents younger customer that have a moderate frequency as buyers.
- Segment number 6 represents younger customers that not only buy a small amount of items, but they also buy cheap items.
- Segment number 1 represents customers in an average age that buy a small amount of items at an average price.
- Segment number 3 represents recurrent customers that make more than average purchases in terms of the expensiveness of the items.
- Segment number 2 represents the best customers, those who buy a lot of items and spend a lot of money in the stores.

Reflection

The end-to-end problem solution can be summarized as follows:

1. Complete the data
2. Process the data to only use the useful features
3. Understand the relevance of each feature and the correlation between them
4. Scale the features so that the distance based algorithms work as expected
5. Remove outliers to aim for more dense clusters
6. Perform a PCA to decrease dimensionality and at the same time understand the variance explained by each dimensions and the features it represents.
7. Optimize the hyper-parameters using grid search and the metrics
8. Cluster the data
9. Visualize the clustering made
10. Analyze the centroids that represent each cluster

The most interesting aspect of the project was that in the end, it was possible to take a real problem, from a real company and apply some machine learning algorithms to it. It is true that the solution is not definite and by its own, the segmentation does not allow to make business decisions, but with no doubt it is a good start and an important thing to know about the customers.

It was really difficult to know what data should be used to solve the problem and it is important to note that there is not a unique solution to it. As mentioned above, it is necessary to make more than one segmentation, using different data and different approaches to really understand a company's customer base and their habits. There is much more to it than just the transactional history. In this project the idea was to use some socioeconomic data, but it was not possible, and apart from that, other things, such as people likes, psychographics and habits, among others, are necessary to fully understand customers and their actions.

Improvement

In terms of algorithms, there is a lot to improve. Although k-means turned out to be the best of the three algorithms, there are many more clustering algorithms that could be use. For another project Affinity Propagation is a must try because the customers are more like a bunch of points together than different figures in space, and according to the classical analysis of clustering algorithms, AP is good on that[6].

There are no further improvements for the algorithms for this problem in particular. On one hand, when the grid search on the Gaussian Mixture Model stated that when the maximum numbers of iterations was 1 and the initial distribution was k-means, the best scores were presented, so obviously that is telling that when the GMM is actually a K-Means, the clustering is better. On the other hand, it does not matter how much tuning was made to the hyper-parameters on DBSCAN, it was not possible to get good results, and now that the knowledge about the data distribution is higher, and looking at the clustering comparison image, is it appropriate to tell that DBSCAN was a bad choice.

If this solution is used as the benchmark, a better solution would arise almost instantly: k-means. There is still an opportunity to use Affinity Propagation or maybe Ward and have better results than k-means and in the future it will be necessary to experiment with those algorithms in another project.

[6] http://scikit-learn.org/0.16/_images/plot_cluster_comparison_0011.png