Machine Learning Engineer Nanodegree

Capstone Proposal

Esteban Dib Puelma
August 28th, 2018

Proposal

Domain Background

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 in https://www.kaggle.com/samratp/creating-customer-segments-unsupervised-learning. 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", available here: https://arxiv.org/pdf/1709.06202.pdf. 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 products available in each location, but they not always succeed in this mission. By segmenting the customers, they will be able to take more informed decisions, and in the end, increase sales.

Datasets and Inputs

To solve the problem I will use data from three main sources. The first one will be given by the company and is the history of purchase from 4 of the companies 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.

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. The third one is the relationship between the address and the money earned by that household which, for this project will be a gross approximation that will correspond to the average salary of the commune that household belongs to.

By using the transaction history I will be able to classify the customers by purchase habits (quantity, amount, recurrence and favorite locations). By adding the API it is possible to segment customers by demographic characteristics such as sex and age. Last, but not least, by adding the commune-socioeconomic table it is possible to determine the socioeconomic group of the customers. An interesting thing to do is to cross all the data and understand the habits of the different socioeconomic and demographic groups.

Solution Statement

To solve the problem I will use unsupervised learning algorithms that allow to create segments of customers with the data described above. The idea is to try more than one to take profit from the advantages of each of them and make a comparison. For now, I am planning to use Gaussian Mixture and DBSCAN, but that can change when the implementation time comes. It is important to make clear that this project will be the starting point of the solution, and will provide the information necessary to take commercial actions toward the customers. With the clusters that will be found, the idea is to create marketing campaigns and personalized in-store and online experiences.

Benchmark Model

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. 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 than the human created model. Thus, better than the benchmark model.

In quantitative terms, I will compare my algorithms with a k-means algorithms with the best number of clusters, but with no other parameters optimized. This will help me get some insights on wether what I am doing is just a minor improvement or a much more fascinating approach.

Evaluation Metrics

To evaluate both models I will use 2 classical metrics 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 me to understand if each cluster is very different from the others or not so much; and the inertia of the clusters, which is calculated using the sum of the squares distances of the points within the cluster, and will allow me to understand how related clusters are among themselves. I have read that inertia is better when the clusters are convex, and DBSCAN does not necessarily form convex clusters, but I want to give this metric a try (https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818). 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.

Project Design

First of all I will consolidate all the purchases of a person by year and by location. This means adding all the items he or she bought and all the money he or she spent in one full year in one location and putting that in one row of the dataset. After that, I will get the address and sex for each person in the database using the API mentioned above and for each address. To do this, I will run a python script using the requests libraries and the csv with the transactional data. Last, I will gather the data for the communes that are near the 4 stores that will be studied in this project and add the socioeconomic data to each customer.

Another thing to do is to make some data exploration to understand deeply what does the data mean. In this exploration it is important to establish the relevance of each feature, so I would train a supervised learning algorithm to predict one feature based on the others. If that is possible, that means that probably the segmentation could be made without the predicted feature. This will allow to use less data, thus less processing power and time to segment the customers and it also will drive decision making on which data to ask the customers when they buy something and which data to save in the database. I will also visualize the distribution of one feature vs the other in terms of heatmap and scatter plots. This will also give insights of correlations, but in a more visual way.

After wholly knowing my data, I will analyze if it is worth to scale some features such as the money spent. Apart from that I will detect the outliers and leave them out of the analysis. The outliers will be analyzed in terms of the amount of items they bought and the amount of money spent.

One really important thing I want to do is principal component analysis. I want to know which features are the most relevant to segment clients. This has similar effects to the feature correlation, but in a much more potent way. It will really help me understand what are the most important things you need to know about your customers to understand what they do and what they need. Obviously this will help me experiment with some dimensionality reduction and try different combination of features, allowing me to visualize in 2d or 3d a multidimensional problem.

After all the data manipulation, I will cluster the data. My idea is to try two clustering algorithms one that needs to be specified with the number of clusters and one that does not. I am thinking about Gaussian Mixture and DBSCAN, although this could change when I get to the implementation part. I will use the silhouette score to compare different number of components in the Gaussian Mixture model and for the DBSCAN I will add the inertia of the different clusters. I think Gaussian Mixture should work well, because all these customers are real people, so their purchase habits, demographics and socioeconomics should have a normal distribution, we will see if this hypothesis turns out to be true.

Finally, I will visualize the clusters using 2d or 3d graphs if the PCA is successful or by using multidimensional graphical technics. If necessary, I will also pick the cluster centers and transform them back to the initial dimensions, so that I have the data of the average consumer of that cluster.