DM Report

Consumption Clustering

Customer segmentation relies on identifying key differentiators that divide customers into groups that can be targeted[[1]](#footnote-1). With the given data we decided to cluster the customers based on demographical and on behavioral attributes.

Clustering the customers based on their consumption behavior we used only the continues product[[2]](#footnote-2) specific features of the dataset:

﻿PremLOBMotor, PremLOBHousehold, PremLOBHealth, PremLOBLife and PremLOBWorkCompensation representing the money spent by on that product.

Given these 5 features and their specific meaning to the company we decided to not apply PCA as a priori we want clear cluster information for each product offered and assume there is valuable information allowing a good split for each product. Further we regard 5 dimensions as low enough to avoid the course of dimensionality.

In regard to clustering, we assumed a priori that clusters do not necessarily have to be convex as well as not equally in terms of size. This and in particular the correlation of the different features is reason enough for us to regard the *K-Means* algorithm and alterations as not suitable for us.

In a trial and error process we tried clustering using the *DBSCAN* algorithm and *Mean shift* algorithm with various parameters, but they did not allow for a good split. Both of them yielded results either identifying one major cluster only or many undefined data points one major cluster and many small clusters. All these clusters are not useful for our case of customer segmentation for marketing which is why we discarded them for our final clustering.

Using the Expectation Maximization clustering method with 3 mixture models allowed a first good split among the data points. As this algorithm clusters data points in terms of likelihood of belonging to one cluster we introduced an 80% threshold for hard clustering following the Pareto principle[[3]](#footnote-3).

Distribution of LOB Household for

3 Clusters 4 Clusters 2 Clusters

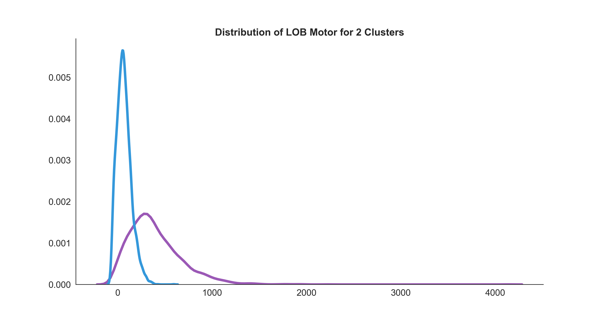
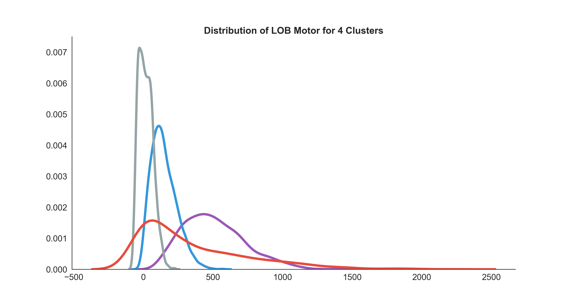
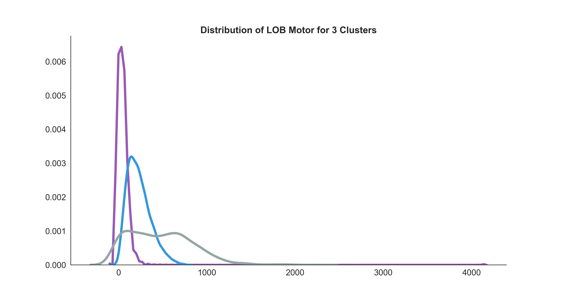


Figure 1: Distribution of LOB for n Clusters

However, the split was not perfect for all variables cf. Figure 1 In order to give more flexibility, we decided to introduce 4 mixture models with the intention to discard one with less meaning or/and combine clusters manually. Clearly, adding an additional cluster did not allow good manual work, hence we reduced to 2 mixture models. As a result, the clusters split the consumption of all products except for Health Insurance (cf. Figure 2).

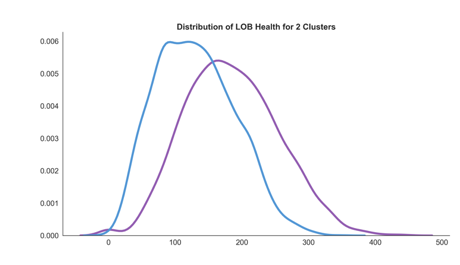


Figure 2: Distribution of LOB Health for 2 Clusters

To deal with this problem we first tried clustering neglecting the consumption of Health Insurance discarding our hypothesis that all consumptions on all products can be split. However, we found much more clusters, which lead us to the conclusion that the consumption of Health Insurance is an important factor for clustering but is not a product where consumption behavior is ambiguous. This is where speculation began:

*How is the Portuguese health system organized/financed?*

*Does Health Insurance fill a different place in Maslow's hierarchy of needs than the other products?*

We considered further investigation in this out of scope and accepted the consumption of Health Insurance being non-split able but important.

For our Final clustering in consumption behavior we used 2 mixture models with an 80% threshold for hard clustering. As a result, we found 3896 data points being regarded as cluster 1 and 4443 labeled as cluster 2 and 1667 not clustered. Looking at the distribution of consumption values, exemplified in Figure 3 we conclude that customers in cluster 1 spend less on Motor Insurance than customers in cluster 2. Contrary, customers in cluster 1 spend more on all other products than customers in cluster 2, but the consumption in Health Insurance is not fundamentally different. Hierarchical clustering was not discovered further, as the results using expectation maximization already topped our expectations.

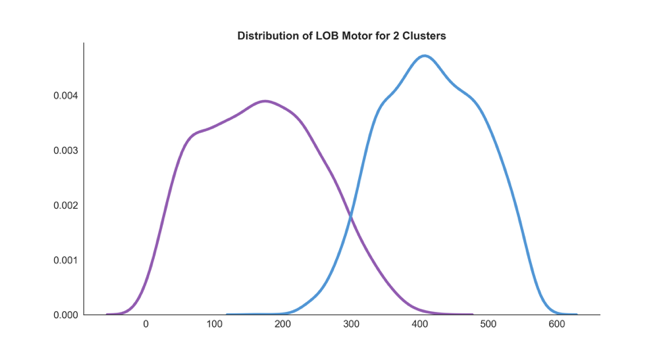


Figure 3: Distribution of LOB Motor for 2 Clusters

[1],[2],[3]

1. *Bunkley, Nick (March 3, 2008).*[*"Joseph Juran, 103, Pioneer in Quality Control, Dies"*](https://web.archive.org/web/20170906182706/http:/www.nytimes.com/2008/03/03/business/03juran.html)*. New York Times. Archived from*[*the original*](https://www.nytimes.com/2008/03/03/business/03juran.html)*on September 6, 2017. Retrieved 25 January 2018.*
2. [**^**](https://en.wikipedia.org/wiki/Pareto_principle#cite_ref-2) *Box, George E.P.; Meyer, R. Daniel (1986). "An Analysis for Unreplicated Fractional Factorials". Technometrics.****28****(1).*[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier)*:*[*10.1080/00401706.1986.10488093*](https://doi.org/10.1080%2F00401706.1986.10488093)*.*
3. [**^**](https://en.wikipedia.org/wiki/Pareto_principle#cite_ref-NYT_3-0) *Bunkley, Nick (March 3, 2008).*[*"Joseph Juran, 103, Pioneer in Quality Control, Dies"*](https://www.nytimes.com/2008/03/03/business/03juran.html)*.*[*The New York Times*](https://en.wikipedia.org/wiki/The_New_York_Times)*.*

1. https://searchsalesforce.techtarget.com/definition/customer-segmentation [↑](#footnote-ref-1)
2. Insurance is regarded as an intangible product in business [↑](#footnote-ref-2)
3. The Pareto principle (also known as the 80/20 rule, the law of the vital few, or the principle of factor sparsity)[1][2] states that, for many events, roughly 80% of the effects come from 20% of the causes.[3]

   The Marketing department might find 80% not high enough or too strict, therefore in reality setting the threshold will heavily rely on the given problem. [↑](#footnote-ref-3)