DM Report

Engagement Clustering

Mostly customer segmentation is made based on business and common knowledge. It seems logically to assume that for example younger and older people have different pattern of behaviour, they have different priority in life. So in that case to approach those two group of customers there have to be developed difference marketing strategy. Although it seems logically correct this approach has two major disadvantage. First of all, this split of customer are not always the best one, in many cases age difference play not really significant role where there are other features, which splits customers in more efficient way. Additionally cost of wrong assumption in business can lead to loss in market share. Moreover assuming, that our a priori knowledge is correct it is possible that it won’t be enough. The additional knowledge which can give business advantage over competition can be obtain by performing unsupervised learning, from which you can have new, very valuable insides.

This being said we decided to try different clustering algorithm to find separate groups of customers, yet reasonable one. Clustering the customers based on their engagement we used features from the data set: Year of 1st bought policy (1stPol), Birth year of customer(BirthYear), Gross salary(Salary), Information whether they have children or no (HasChild), highest level of education (EduDegree), information where they live (GeoLivArea) and information about their claim rate (ClaimRate).

In the process of customer clustering, we observed that some variables are not important and even blur the division between groups of customers. That is why we decided to perform some feature engineering.

In EduDegree each customer had one of four possible labels: *“Basic”, HighSchool”, “Bs/Msc”, “PhD”.* Only 97 customers had PhD which was ~1% of our customers, so under a priori assumption that people with higher education and doctors should have very similar opinion about their expenditures on insurance, we decided to group those two labels together. A posteriori approach lead us to another conclusion that people with only basic and with high school education also have very similar customer behavior so we also grouped them together to simplify future analysis. We ended up with only two groups of customer which are evenly distributed.

Another a posteriori conclusion we came up with that living area of customer have no idea in any approach we made to clustering. That could be justify be reasoning that all of the customers are living in Portugal so different living area can be information about different city of neighbourhood but in all of them customers have the same attitude to insurance. That being said we decided to not include that variable in future analysis.

Another conclusion we came up with that the year of first year policy doesn’t distinguish customers significantly enough to include this information in final model. That could be justify by reasoning that customer profile does not change over the years of existence of our insurance company. That being said we decided to not include that variable in future analysis.

Another idea we pursue was to calculate on what age people starts being customers of insurance company, because the reasoning standing behind expenditure on insurance if someone is very young, in middle age or elderly person could be translated into different marketing strategy for each group. This information can be calculated as a difference between customers age and year of their first policy. We went one step further ( because of the level of incorrect data in Birth Year) we calculated predicted birth year using simple linear regression on gross salary ( level of correlation was 0.92) when we calculated age of each customer than age of their first policy using formula:

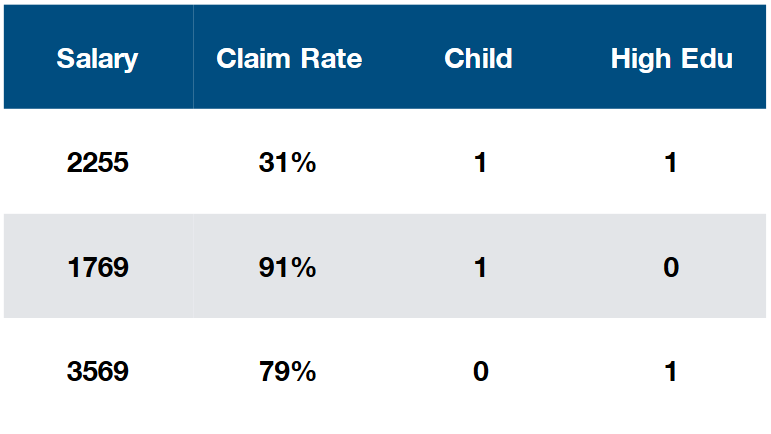
*( 2016 - predicted birth year) - year of first policy = age of their first policy*

Unfortunately we didn’t observe any significant improvement in our clustering outcome so we drop this idea.

Our final set of features was: Salary, HasChild, Grouped EduDegree, Claim Rate. To cluster the data we used a variety of algorithms. First approach was to use kMode method. The biggest advantage of this algorithm is ease of implementation and interpretation, but on the other hand we only can use categorical data. We discretized continues variable (claim rate and salary) into 3 category and used them as categorical information in kMode. This approach yield unsatisfying results, because our a priori assumption of what salary is considered to be low or high was incorrect and we lost useful information which continues variable contain.

Another approach we tried was to split data into categorical and continues variables and use different algorithm on each group. We used kModes on categorical data and kMeans on continues one. Unfortunately when we merge results from two models we received very unevenly distributed clusters, some of them were were big and other contain only small amount of customers. Instead of thinking how to merge clusters, which lead to significant loss in the amount of information about customers, we decided to use another clustering algorithm - k Prototype.

The biggest advantage about this algorithm is that it has ability to use categorical and clustering data in the same time. After trying with different number of clusters we observed that the best split can be obtain having three cluster. In the table below are presented centroid of each clusters.



*Figure X: Centroids of clusters obtain in kPrototype algorithm*

References:

<https://medium.com/@Chamanijks/k-prototype-in-clustering-mixed-attributes-e6907db91914>

<https://searchsalesforce.techtarget.com/definition/customer-segmentation>

<https://shapeofdata.wordpress.com/2014/03/04/k-modes/>

“clustering categorical data with k-Modes” Josuha Zhexue Huang, The university of Hong Kong