Customer Segmentation using k-means Clustering in SAS

A marketing campaign in a company for a product or service is never easy. It involves a lot of research by the team, one of the important ones is the customer background study. Knowing your audience is very crucial and analytics is helping businesses a lot to come up with better techniques of customer segmentation. In this article, we are sharing our experience with such project we did at University of Illinois at Chicago as a part of a course for Masters in Management Information Systems.

One of the way to group your customers is by the unsupervised data mining concept of Clustering. The case that we worked on involved demographics data for one of the states in US and the problem statement given to us in the project was to formulate, cluster and analyze clusters of people. The software we used was SAS 9.4. The data provided to us had around 36 variables with all numeric values (expect state code) and over 34300 records.

To start with the analysis, we cleaned the data, as any analysis’ first step. Missing values were identified and replace with mean of the respective zip-code of the record. You can also do an outlier analysis, which we skipped as it was not necessary for our problem statement.

Moving forward…  
Before going on to the clustering analysis, since the number of columns available were too high, we had to cut down our variables in the dataset to a usable number of at most 6 to 7, using Factor Analysis. This is always helpful, or else the closely related variables present in the dataset will lead to multicollinearity results. We also standardized the numeric variables to their normal forms using mean = 0 and standard deviation =1 before running the factor analysis. This is a good practice with numeric variables to balance out the difference in the ranges of variables (like age and salary) and so that the variable with larger range do not dominate the results.

Once the factor analysis was completed, SAS gave us the correlated grouped columns as total of 6 factors like the Factor 1 explaining Education, Factor 2 explaining Affluence, Group 3 as Age etc. SAS also allows you to select the number of factors you want at the end. So, you can keep iterating the number to finally arrive at a fine set of groups that makes sense.

Last Step…  
After all the data preparation, we moved forward with the Cluster Analysis. Since we decided to use k-means clustering, there was a question of predefined number of clusters. To decide upon that, we build a macro in SAS, testing the optimal number of end clusters (between 4 to 9). Further, we ran another macro that would iterate the records and choose the centroids as the starting points for k-means, as we should not select that randomly ourselves. Once we finished this, we plotted a graph between pseudoF and distance between the clusters (distance between the centroids). PseudoF is described as

(between-cluster-sum-of-squares) / (within-cluster-sum-of-squares)

The best selection is when the pseudoF and distance between clusters have high values. We could find an ideal case from the graph and finalize the k value for our k-means clustering.

As per our data, we obtained a good set of 6 well distinct clusters. But the job didn’t end here. Once we segmented the records into clusters, we had to classify them with some characteristics. Hence, we analyzed the factors and tried to give properties to each cluster. Looking at the average values of factors for each cluster separately, we could identify their characteristics easily. For example, Cluster 1 had a higher average age than other clusters, with lesser income index and lesser education and high rural household index which meant that they belonged to a low income retired people living in suburbs. Similarly, for other clusters as well.

The corresponding results were handed over to the Marketing team to form the campaigning strategies as per the demographics and target their customers accordingly. As a course project, we used the general demographics of people in one of the states in US, but the techniques and procedure could be used for a specific set of customers for a company for segmenting them, analyzing their behavior based on the variables available and form decisions not only in marketing but several other business strategies like personal selling, employee management, feedback analysis etc.

We are glad to share our SAS code with the readers, which can be found on the github link below. Please let us know in case of any questions regarding the content or code, in the comments. Thank you for reading.

GitHub Code:

https://github.com/datatalesblog/Customer-Segmentation---SAS.git