Jonathan Gordon April 26, 2024
Customer Segmentation Article Review

Introduction

In the article," Improving customer segmentation via classification of key accounts as outliers" (Spoor, 2022), the researcher presented a methodology to improve customer segmentation using a two staged approach. The first stage removes key accounts that businesses typically handle with separate sales, marketing and support teams. The second stage uses cluster analysis to segment the remaining customers. The results of this two-stage approach are compared to an earlier study that used the same data but with only a single stage clustering methodology. The clusters derived from the two-stage approach show better differentiation than the one stage approach on the same data set with the added benefit of identify key customers that should be treated differently than the rest.

Article Summary

Customer segmentation exercises typically use clustering algorithms on the entire data set to group customers with similar attributes. This methodology does not take advantage of the fact that many businesses manage their top customers with separate sales, marketing and support teams. Including these customers in the cluster analysis is detrimental to the exercise as the key customers will continue to get special treatment no matter what cluster they end up in, yet they have an oversized influence on the remaining cluster assignments. This study proposed to include domain-specific knowledge in the data analysis and then use outlier detection to identify special accounts. Although a business may have already identified top accounts, this outlier detection may propose a different set of accounts that the business should consider as special.

This study used a two-step approach where the first step identified the key accounts by looking at low-density areas of the data using cluster algorithms such as DBSCAN. Once the key accounts were removed, they perform cluster analysis using Gaussian Mixture Model (GMM). GMM is a soft clustering algorithm that assigns data points based on probabilities they belong to a cluster. Some refer to GMM as a fuzzy assignment. This differs from algorithms such as K-Means which use hard clustering. In K-Means each data point belongs to one and only one cluster. The article points out that DBSCAN is not typically used in customer segmentation as it tends to create one large cluster of customers. This may not be good for segmentation, but it is very good for identifying outliers – which in this application are the key accounts. The article refers to other segmentation articles, which use K-Means, Ward, Euclidean Hierarchical, Recency, Frequency, and Monetary (RFM), and Principal Component Analysis (PCA) to cluster customer data. Euclidean Hierarchy measures the Euclidean distance between clusters. Ward, unlike the other agglomerative methods, analyzes the variance of clusters rather than measuring the distance directly. RFM is useful for segmenting sales data. K-Means clusters based on data similarities – a friend versus enemies' approach. PCA reduces data dimensionality while attempting to maintain variance between variables.

There is no universal approach to identify key customers. The ABC approach groups customers into three buckets, based on their revenue generation. A more nuanced approach, proposed in this study, is to consider not only high revenue producers but also customers that are highly important to the business regarding one or more product area (often called strategic accounts), and customers with unique purchasing behaviors. The number of such accounts must be reasonably limited. The article mentions that PCA or logarithmic scaling could be used to reduce the impact of the outliers, rather than removing them, but then you would lose the benefit of identifying customers that are key to the business and should receive special attention. The key accounts should stand out as less dense areas when using a density-based detection algorithm such as DBSCAN. DBSCAN uses core data points and all reachable data points to form clusters. Data points not reachable are identified as outliers in this study and removed from the second stage of clustering. The key parameters in DBSCAN are MinPts and Eps. A k-dist-graph is created, and the elbow of the curve used to select the optimal value for Eps. After the outliers are removed a GMM is used to cluster the remain accounts. The estimation of parameters for GMM is performed using EM (expectation and maximization). The final parameters describe the clusters' centers and covariance. The study uses a Bayesian approach to select the number of clusters. The silhouette coefficient is used to measure the quality of the clusters.

The proposed model was applied to a data set of a Portuguese food wholesaler. This data was used in a previous clustering paper, analyzing the performance of GMM, but without removing key accounts. The quality of the clusters without removing key accounts is later compared to this studies approach of first removing key customers. In addition to revenue, other variables used in the clustering include product segment, region, and industry sector. The k-dist-graph identifies the elbow of the curve at 12000. DBSCAN is run using Eps set to 12000 and MinPts set to 4. Sixteen of the four hundred and forty customers are identified as outliers. All 16 of them are either high overall revenue generators or high within an industry segment. Revenue was not the only variable used to identify outliers. High revenue customers with behaviors that matched smaller revenue accounts were left in the cluster. The outliers are identified as key customers and removed from the next stage of clustering.

The next step in the methodology was to assign the number of clusters for GMM. BIC and AIC are plotted for differing number of clusters. In the study that did not remove key accounts it is very hard to identify an optimal number of clusters. The BIC and AIC curves do not have an obvious elbow. Three clusters were chosen, but they could have chosen anywhere between three and ten. In the study with key accounts removed, the elbow of the curve was clearly defined at four clusters.

Analyzing the results of the three-cluster solution, only one of the clusters clearly differentiates between the retail and hotel industry segments. The clusters also do not differentiate well between purchasing behavior. The clusters appear to mainly differentiate by overall revenue. The two-stage, four-cluster solution shows clusters that are more clearly delineated by industry mix as well as their purchasing behavior. Using ABC analysis, group one are C customers with an even product revenue mix which they identify as the small café customers. Group two are hotel industry B customers, group three are hotel industry A customers, and group four are A and B retail customers with an even mix of product purchasing behaviors. The silhouette coefficients were also larger (better) for clustering without key accounts. The advantage disappears when a larger number of clusters is used. The result points to using smaller number of clusters in customer segmentation.

Critique

The methodology proposed in this study makes logical sense in the context of how businesses treat their key customers. It is common practice for medium to large sized businesses to assign separate sales, marketing and support teams to key accounts. One advantage for businesses using the methodology described in this article is that customers are not identified as key solely by revenue. Other factors are considered, such as industry segment, importance to individual product lines and geographic location. Companies can use this methodology to test if they are identifying the optimal key accounts.

The article does not go into detail as to why GMM was selected for stage two versus K-Means, PAM-TSNE, or any of the agglomerative methods for clustering the remaining accounts. If they had applied multiple methods, they could have then used Rand index to measure the similarity between the resulting clusters. PAM-TSNE would have allowed us to see the medoids and confirm if these accounts looked like they should be in different clusters.

Using GMM after removing outliers removed did show improved results compared to earlier work done using GMM on the same data set but without the removal. Improvement was measured quantitatively using silhouette scores and visually looking at the industry and product purchase distribution in the clusters. It is my belief that a two-stage methodology is a valid choice for doing customer segmentation when customers confirm that they wish to treat key customers differently than the mainstream.

Conclusion

Using a two-stage approach to customer segmentation showed impressive results. The customers identified as outliers in the first stage had the attributes one would expect for key accounts. They were either high overall revenue producing or high within their industry. This mirrors typical business behavior of identifying key customers based on multiple factors. The second stage of the methodology created clusters that more clearly delineated the types of customers. This differentiation provided more insight into the customer segmentation as they grouped customers by multiple important factors. Removing the key customers made it easier to select the right number of customers using BIC and AIC analysis. The resulting clusters are more useful for marketing campaigns as they more clearly delineate the type of customer attributes and behaviors. The end results is a statistical model that both identifies key accounts and segments the remaining accounts based on complex attributes as well as revenue. This approach needs to be tested on other industries, but the results are encouraging.

References

Improving customer segmentation via classification of key accounts as outliers

Jan Michael Spoor, June 2022

Journal of Marketing Analytics