Automated cluster generation and labelling of peer groups for marketing reporting

Received (in revised form): 13th June, 2023

Dakota Crisp

Analytics Manager, OneMagnify, USA

Dakota Crisp is an analytics manager at OneMagnify. With a PhD from the University of Michigan, his hypothesis-driven approach to integrating concepts from neural engineering into the data science space provides a distinctive take on consumer behaviours.

OneMagnify, 912 N Main Street #100, Ann Arbor, MI 48104, USA
Tel: +888 294 1512; E-mail: dcrisp@onemagnify.com

Jonathan Prantner

Chief Analytics Officer, OneMagnify, USA

Jonathan Prantner is the Chief Analytics Officer at OneMagnify. His approach to applied mathematics has pushed analytics to the limits for over two decades. Jonathan's career has spanned educational research, automotive, consumer packaged goods, travel and healthcare. At OneMagnify, he leads efforts surrounding applied artificial intelligence and machine learning as well as integrating advanced analytics with data visualisation platforms. Jonathan is a celebrated thought-leader and recipient of multiple data science patents.

OneMagnify, 912 N Main Street #100, Ann Arbor, MI 48104, USA
Tel: +888 294 1512; E-mail: jprantner@onemagnify.com

Grant Miller

RXA, USA

Grant Miller uses his programming and statistical approaches to try and tackle standardisation, automation and deep insights.

RXA, 912 N Main Street #100, Ann Arbor, MI 48104, USA
Tel: +888 294 1512; E-mail: grant.miller@rxa.io

Jack Claucherty

Analytics Manager, OneMagnify, USA

Jack Claucherty is an analytics manager at OneMagnify with a BS and MSE in industrial engineering from the University of Michigan. His background as an industrial engineer helps him understand complex systems and he loves helping his clients find value in their data and the models created.

OneMagnify, 912 N Main Street #100, Ann Arbor, MI 48104, USA
Tel: +888 294 1512; E-mail: jclaucherty@onemagnify.com

Tom Thomas

Vice President of Data Strategy, Analytics & Business Intelligence, FordDirect, USA

Tom Thomas is Vice President of Data Strategy, Analytics & Business Intelligence at FordDirect. Tom has 30 years of experience as an executive, consultant and entrepreneur delivering digital advertising, mobile application and enterprise resource planning solutions to the automotive, consumer packaged goods, hospitality and public utilities industries.

FordDirect, 4 Parklane Boulevard, Dearborn, MI 48126, USA
Tel: +844 889 7367; E-mail: tthom213@forddirect.com

Danielle Barnes
Analytics Director, OneMagnify, USA

Danielle Barnes is an analytics director at OneMagnify. She is an accomplished analytics leader with extensive experience across the entire data lifecycle. Her work directing enterprise analytics initiatives for companies across various industries has made her a powerhouse for realising visions in complex environments. She is a Spartan superfan with a BA in mathematics, MS in statistics and currently pursuing a PhD in data science, all from Michigan State University.

OneMagnify, 912 N Main Street #100, Ann Arbor, MI 48104, USA
Tel: +888 294 1512; E-mail: dbarnes@onemagnify.com

Abstract In today's data-driven marketing landscape, clustering data helps businesses better understand themselves and their customers. However, clusters derived from machine learning can be difficult to interpret and obtain buy-in from stakeholders. This paper details a method for automated cluster generation and labelling using machine learning. Two automotive case studies are provided where clustering enhanced business value and gained stakeholder buy-in. The first details segmenting dealerships based on their media environment to produce higher quality media models for lead generation. The second entails the creation of peer groups to enhance performance reporting across a diverse set of dealerships.

KEYWORDS: clustering, automotive, labelling, peer groups, segmentation

INTRODUCTION

In today's data-driven marketing landscape, grouping customers, stores or markets into clusters allows for better reporting and customised communications. Marketers want to understand which customers to target, how to best communicate with them and how much each segment is worth. Historically this process has often been focused on the creation of personas. Marketers have traditionally created these customer personas based on intuition, personal experience and marketing strategies rather than data-driven insights. This approach may lead to personas that are based on stereotypes or assumptions about customers, rather than on actual customer behaviour and preferences.[1]

Likewise, most traditional store-based comparisons used in performance reporting are based off comparisons to other stores in the same region. This approach ignores the fact that geographically diverse stores may have more in common than stores within a closer proximity. Additionally, the smallest store within a region may consistently find itself at the lower end of performance metrics, which can effectively limit the business value of performance reporting.

With the increasing availability of customer data and advanced analytical tools, marketers can now create customer personas that are based on data-driven insights and are more accurate representations of their target audience. From the store reporting side, the creation of peer groups is a mechanism for creating data-driven comparisons that allow for more actionable reporting. Identifying and understanding appropriate peer groups can provide valuable insights into store specific consumer behaviour and preferences. However, the process of manually creating and labelling these groups can be time-consuming and error-prone, particularly for large datasets.

Automated cluster generation and labelling can provide a solution to this challenge. By leveraging machine learning algorithms and statistical models, marketers can quickly and accurately group either consumers or stores based on shared attributes and behaviours, and then assign meaningful labels to these groups for reporting and analysis.

This paper explores the benefits and challenges of automated cluster generation and labelling, as well as examining different approaches and tools available for marketers to use. Two real-world, client case studies are provided to show how automated cluster generation and labelling can be used to improve marketing strategies and drive business results. These methods helped our clients gain insights into their target audience and make informed marketing decisions.

BACKGROUND

While customer segmentation takes place at a low level of analysis, either at the customer or purchase level, store segmentation takes place at a more aggregated level. There are three main strategies for segmenting stores for performance reporting: geographic, store format and purchase behaviour.

Geographic segmentation separates stores based on their location and the demographic and economic characteristics of the surrounding area. For example, stores in urban areas may cater to a younger, more diverse demographic than stores in rural areas. By understanding the local market, retailers can tailor their product offerings and marketing strategies to better meet the needs and preferences of customers in each location.

Store format segmentation groups stores based on their format, such as club stores, big box retailers, supermarkets, convenience stores and discount stores. In the automotive space this segmentation would consider the brands sold, whether they focus on fleet or retail, luxury or general market, or whether they sell heavy trucks. This type of segmentation is used in the retail industry, where different store formats cater to different customer needs and shopping behaviours. By understanding the preferences of customers who shop at each store format, retailers can optimise their product offerings and store layouts to better serve each segment.

Purchase behaviour segmentation separates stores based on the purchase behaviour of their customers. For example, stores may be segmented based on the frequency of customer visits (ie a high-volume store), the types of products purchased (ie discount grocery) or the average transaction value (ie high-end retail). By analysing customer purchase data, retailers can identify patterns and trends in customer behaviour, which can be used to develop targeted marketing campaigns and promotions.

METHODOLOGY

The methodology applied is comprised of four high-level steps: preprocessing, determining the feature set, clustering and labelling. These steps are typical of many data science workflows and allow for multiple paths depending on data types and goals. Therefore, this approach can be customised and adapted to related problems. See Figure 1 for an overview of the workflow.

Preprocessing

Effective cluster generation and labelling requires high-quality, clean data. Preprocessing involves cleaning, transforming and scaling data, and is critical to ensuring accurate and meaningful results. Properly preprocessing data reduces the risk of bias and improves the accuracy of clustering, which can help identify meaningful patterns and insights that might have been obscured by noise or irrelevant information.

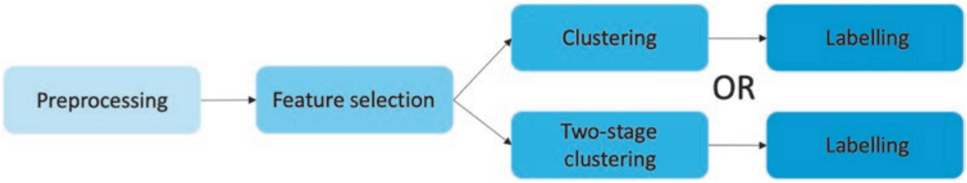

Figure 1: Algorithm selection options for clustering

Feature selection

In cases where there is a specific outcome or target variable (ie sales), it is important to identify features that are relevant to that variable. These features should capture information about customer behaviour or characteristics that are likely to influence the outcome. For example, if clustering customers based on their likelihood to make a purchase, consider features like past purchase history, browsing behaviour, demographics or geographic location. To determine which features are most important, statistical techniques like correlation analysis or feature importance ranking are helpful. Including the most important features in a clustering model can increase the accuracy and relevance of the resulting clusters.

In cases without a defined target variable, determining the appropriate features to use for clustering can be more challenging. In this case, it is important to identify features that are relevant to the business problem that is being solved. For example, clustering customers to identify segments for targeted marketing campaigns should consider features like product preferences, channel usage or customer lifetime value. To identify relevant features, utilise domain expertise or perform exploratory data analysis. For example, principal component analysis can be used to identify features that maximise variance, which may be indicative of natural segmentation. Additionally, plotting the per cent contribution of categorical variables over time may identify trends that signal changes in customer opinion. Finally, Voice of Customer analysis can also be helpful to identify topics that the customer is interested in, how they view a business's products and how they engage with the brand. Selecting features that are relevant to the business problem can ensure the clusters are actionable and meaningful to the end user.

Clustering

Determining the optimal number of clusters is a crucial step in automated cluster generation for marketing reporting. Too few clusters overgeneralise the data and lose the ability to target specific groups. Too many clusters can lead to a loss of distinction between the clusters. The overall goal is to create clusters that are internally homogenous and externally heterogeneous. In other words, the individuals in the clusters share similar traits and the clusters appear distinct from each other. The more variables included in clustering the more difficult it becomes to balance these two aspects.

There are several methods available to identify the mathematically appropriate number of clusters for a given dataset. The Elbow method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters. The optimal number of clusters is the point where the rate of decrease in WCSS slows down and bends.[2] The Silhouette method measures the similarity of each data point to its assigned cluster versus its similarity to the other clusters. The optimal number of clusters is the one that maximises the average Silhouette width.[3] The Gap statistic method compares the within-cluster dispersion with its expected value under a null reference distribution. The optimal number of clusters is the one that maximises the Gap statistic value.[4] These methods provide statistical justification for determining the appropriate number of clusters, but do not represent the end of the analysis. A significant part of finding the optimal cluster number is getting buy-in from the client. By making sure that the cluster number makes sense within their industry and problem statement, you build trust and increase the likelihood of your models being used and having an impact on the organisation.

In general, clustering is a technique used in data mining and machine learning to identify groups of similar data points within a dataset. The goal is to identify patterns in the data that may not be immediately apparent, and to group similar data points together so that they can be analysed and compared more easily. Once the data is prepared, the feature set is determined and the optimal number of clusters is known, attention can be turned to the development of the clusters. There are several predominant clustering techniques that can be used, each with its own strengths and weaknesses. Two frequently used methods are outlined: K-means and hierarchical clustering.

K-means is a widely used algorithm for clustering. It partitions a dataset into K number of clusters. The algorithm begins by randomly selecting K initial centroids, and then iteratively assigns each data point to the nearest centroid. It then updates the centroid based on the mean of the data points in the cluster. This process continues until the centroids no longer move, or a maximum number of iterations is reached. K-means clustering is fast and efficient but is sensitive to the initial placement of the centroids and can be biased towards spherical clusters.[5]

Hierarchical clustering approaches segmentation from either a top-down or a bottom-up approach There are two main types of hierarchical clustering: agglomerative and divisive. Agglomerative clustering starts with each data point as its own cluster, and iteratively merges the closest clusters until a single cluster is left. Divisive clustering starts with all the data points in one cluster, and recursively splits the clusters until each data point is in its own cluster. Hierarchical clustering can be more interpretable than K-means, but it can be computationally expensive, especially for large datasets.[6]

Once a dataset is clustered, cluster distribution checks should be deployed to ensure the number of samples in each cluster are not skewed. For example, if there are K clusters and one cluster had less than 5 per cent of the samples, it may be prudent to test K – 1 clusters to determine if there is a more equal balance. When using automated algorithms, these simple checks can help ensure the accuracy and quality of the clustering process.

Lastly, it is important to check the accuracy of the model in classifying the data points into the correct clusters. This can be done by using a holdout set of data points that were not used in the training of the model. By measuring the classification accuracy of the model on the holdout set, analysts can assess the reliability of the model's predictions and make any necessary adjustments to improve accuracy. This will help to ensure that the clusters are internally homogenous and externally heterogeneous, not only across all metrics used for clustering, but also for those specific metrics used in the labelling process detailed below.

Two-stage clustering

In the second case study below, the clustering was done in two stages. A two-level top-down hierarchical approach entails first dividing the dataset into larger, more general groups, and then further dividing those groups into smaller, more specific subgroups. Here are three advantages of this approach:

- Better representation of the dataset: By first dividing the dataset into larger, more general groups, the clustering algorithm can capture the broad characteristics of the dataset before moving on to the more specific subgroups. This can result in more accurate and representative peer groups, as the algorithm takes into account the larger context of the data.
- More efficient and manageable: Dividing the dataset into larger groups initially can make the clustering process more efficient and manageable, especially when dealing with large datasets. It can also make it easier to interpret and explain the resulting clusters, as the overall structure of the peer groups is more straightforward.
- Flexibility in creating peer groups: By dividing the dataset into larger groups first, analysts have the flexibility to create different levels of granularity in the peer groups as needed. This approach allows for a more customisable and adaptive clustering process that can be tailored to the specific needs and goals of the business.

Labelling

After clusters have been identified, the next step is to determine which features or metrics contributed most to the formation of each cluster. SHAP (SHapley Additive exPLanations) values can be used to identify which features have the most significant impact on the cluster formation. By examining the feature importance of each cluster, analysts can identify the most relevant attributes and use them to label the clusters. One approach is to rank (high, medium, low) each cluster by each significant attribute, creating meaningful labels.

It is also important to note that labelling clusters is an iterative process that requires ongoing evaluation and refinement. As new data becomes available or business goals change, the clustering analysis may need to be updated, and labels may need to be revised accordingly. By continuously monitoring the accuracy of the model and evaluating the effectiveness of the cluster labels, businesses can ensure that they are getting the most value from their clustering analysis.

CASE STUDIES

Below are two case studies in the tier three automotive marketing space that demonstrate how clustering and segmentation can be applied in practice. The first case study focuses on how to better model dealership advertising efforts based on their unique market conditions. The second details clustering dealers for performance reporting based on aspects of how their dealership operations compare to others. Both examples showcase how clustering and segmentation can be used to improve marketing strategies and operational performance in the automotive industry.

For both case studies, data was sourced across several internal and external databases used by our client. Vehicles sales, service volumes, brand makeup and other store configurations were internal, while market area demographics, population sizes and competitive levels were external. External demographic data like mortgage costs within a zip (post) code was used to infer clustering features like vehicle price sensitivity. This could lead to accidental bias by indirectly and inadvertently discriminating against certain groups within the market area with advertising content or discount (offer) ranges that could be considered unfair. It may also cause the dealer to spend less for display placements and/or under-bid for search terms that could constrain market awareness and competitiveness. Some strategies to avoid these issues include transparency with data and methods, review and discuss results and seeking guidance from external non-profit organisations that promote best practices for avoiding segmentation bias like the Algorithmic Justice League or the AI Now Institute.

MEDIA PERFORMANCE SEGMENTATION

Background

This case study examines automotive dealership execution of programmatic tier three media. Tier three media is focused on the dealership itself and in this study is driving consumers to a dealer's website through search, display, pre-roll video and social media. The client wanted a model that would estimate the number of sales leads driven by different levels of media spend across 1,000+ individual dealerships.

One model would be too generic to effectively capture the differences in media effect between the numerous dealerships. Conversely, creating a model for every dealer (1,000+) would be cumbersome and not provide reliable results for lower spending dealers due to lack of data. The ideal solution would be to create several models that focused on similar types of dealerships.

Solution

Dealerships were clustered based on their media environments and a separate media model was created for each cluster. This minimised the number of models used, making it more manageable while also effectively capturing the differences between different types of dealerships.

Determining the feature set

To segment dealerships based on their media perspective, characteristics were identified that would have the most impact on sales leads from dealership to dealership. These broad characteristics are listed below.

- Dealers experience different cost per impressions in their area.
- Stores of different sizes have different name recognition.
- The competitors in the dealer's area influence the cost of certain search keyword terms.
- The population size and characteristics of a dealer trade area is related to the total lead volume.

A pool of potential dependent variables that related to the above characteristics was gathered. More than 20 features were taken from client data detailing business/media operations (internal features) and more than 50 control features were taken from public databases with various socioeconomic data (external features). With sales leads defined as the target variable, this pool of potential dependent variables was reduced to the most salient features that would be foundational for clustering.

From the pool, three internal features were selected. The first was regional average cost per lead (CPL) linked to media executions. This was calculated by taking the average cost per lead for all dealers that fell within a metropolitan area. The second feature was relative sales rank compared to other dealers of the same brand. The final internal feature was total competitors within the dealer's primary market area.

In addition to the internal metrics, two external features were used to help understand dealer characteristics. The first was median mortgage cost for homes that resided in the dealership's zip (post) code (see Figure 2). This served as a proxy for price sensitivity for a vehicle purchase. The second metric was total population in that zip (post) code, which served as a proxy for the size of the prospective audience.

As previously mentioned, it was important to balance the number of potential features considered with the ease of understanding the makeup of the clusters. The team worked to ensure that not only were the features mathematically related to the potential target variable but that they also represented meaningful information. The client's external users of the model were media strategists who were well versed in

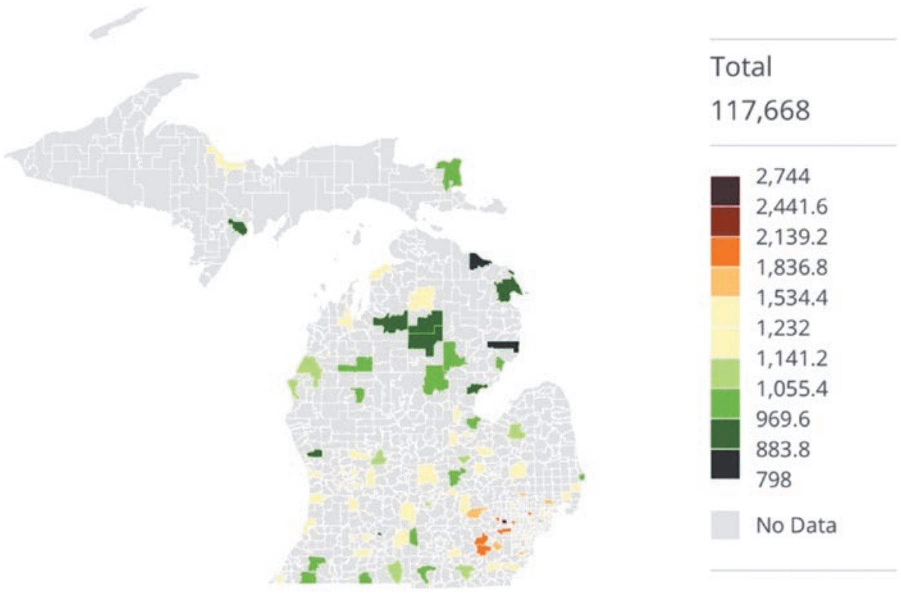

Figure 2: Median mortgage cost by zip (post) code for Michigan

marketing segmentation and customer personas. The application of data-driven cluster labelling via machine learning was both impressive and built further confidence in the broader media mix models.

Results

Using these five features across 1,000+ dealerships, clustering was performed, and eight dealership clusters were identified (see Figure 3). Automatic labels were generated for each cluster based on the five features. Ten dealerships from each cluster were withheld and the label model accuracy was determined to be 92 per cent. These clusters received client approval and were then used to generate eight separate models to more effectively predict sales leads.

SEGMENTATION FOR STORE PERFORMANCE REPORTING

Background

The client supported 2,500 dealers nationwide and needed to enhance their operations reporting process with regard to digital engagement, communication strategies and dealer enrollment across the products and services the dealers were enrolled in. Operations reporting in their industry can be difficult as there is often not a single metric that describes success. Their standard reporting was based on geographic and firmographic segmentation, where dealers are broken out by the brands sold, the type of service setup and their geographic location. While helpful, the segments generated had intra-segment variability that was too high for comparisons. The client wanted to focus on creating dealer peer groups, which would enable comparisons that are understandable to the field team and would stand up to the review of dealers themselves.

The goal of the peer grouping was to create groups of dealers which are similar in some key measures of dealership operations to be used for comparison of engagement at the customer level. The client requested that these groups include 20–50 dealers each. With over 2,500 dealers across the country, this means there would be between 50 and 125 separate peer groups. Not only were the peer groups to be internally homogeneous and externally heterogeneous, but each comparison set within each peer group

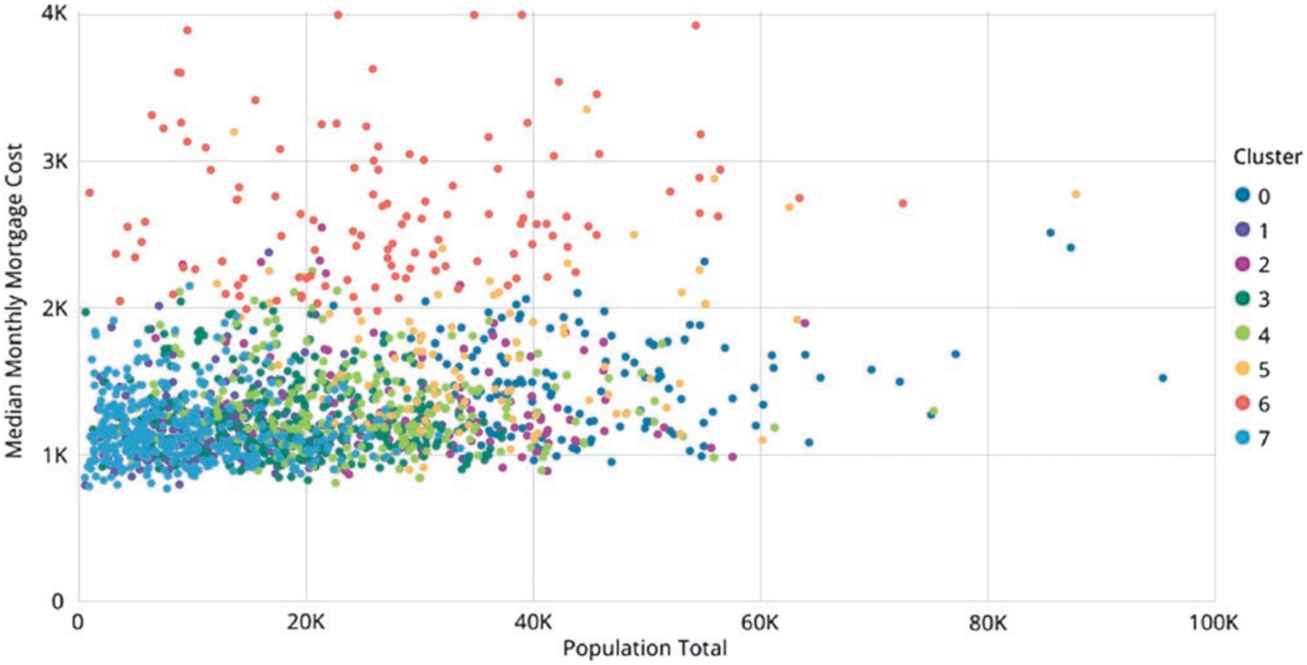

Figure 3: Clustering plotted against two of the variables used, median monthly mortgage cost and population total

needed to be appropriate and understandable to the end user.

Finally, preliminary analysis identified the optimal number of traditional clusters to be between 6 and 10, which did not align with the problem statement (50–150). The client needed a new solution that would create peer groups that satisfied their business conditions.

Solution

To create the desired number of peer groups, a two-layer approach to clustering was used. The process deployed was to first make a 'super group' of hundreds of dealers and then make smaller peer groups inside of each super group.

Determining the feature set

The following datapoints were available for each dealer:

- Vehicle sales (retail and fleet sales)
- Repair orders
- Population in area
- Average income in area
- Competitors in area.

To determine which two metrics to use to create the super groups, every combination of metrics listed above was tried and the silhouette scores for the supergroups were calculated. From the results, retail sales and population were the best performing metrics. The project stakeholders liked this combination because it included one metric a dealer could control (sales), and one which was variable based on their location (population). Using these metrics, the dealers were ranked, segmented into thirds, and labelled (high, medium, low) for each metric. The combination of their labels in each metric created nine dealer super groups (see Figure 4).

Results

Once the super groups were created, a size constrained clustering algorithm was utilised

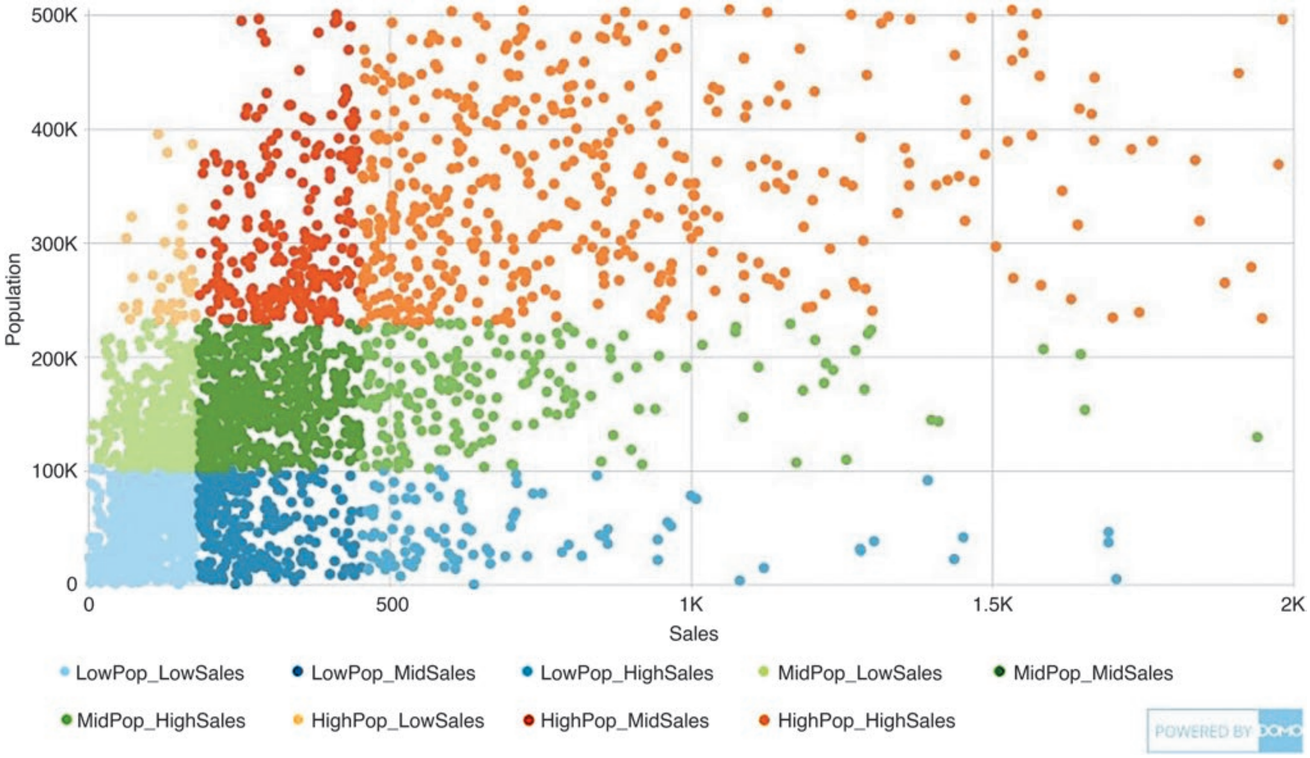

Figure 4: Super group segmentation. Dealerships were ordered and segmented in three groups (high, medium, low) for both population and sales, creating nine distinct segments across both metrics

separately inside each super group to form the peer groups. Size constrained clustering minimises pairwise distances within clusters while adhering to minimum cluster sizes. This ensured there would be enough dealers inside each group to provide meaningful comparisons. The creation of super groups was critical for field use adoption. Explaining to dealers how they were in one of nine peer groups was simple, in their terms and credible.

Using the scclust package in R,[7] 93 unique peer groups were created and are now being used as a normalised way to compare dealers. Results were validated and Shapley values were used to rank feature importance. Labels for peer groups were automatically generated using the super group name and the top three features from the second phase of clustering. An example group name is Mid_Sales_Mid_Population_High_Competitors_Low_AvgIncome_High_RepairOrders.

DISCUSSION

Clustering falls into the category of unsupervised learning solutions because it involves grouping similar data points together without any predefined categories or labels. This approach allows marketing analysts to identify patterns and relationships in their data that may not be immediately apparent through other methods. However, it is crucial to choose a clustering methodology that promotes buy-in from key stakeholders. By involving stakeholders in the selection and implementation of the clustering methodology, marketers can ensure that the results are relevant, actionable and aligned with the overall goals of the organisation.

The first case study demonstrates how clustering and segmentation can be applied to improve marketing strategies and operational performance in the automotive industry. In this case there was a key target variable, sales from leads, which could be used as the justification for selecting certain metrics to cluster upon. By clustering dealerships based on their media environments, the team created separate media models for each cluster. This made it easier to estimate the media impact on driving sales leads. This process not only improved the accuracy of the models but also helped develop meaningful names for each cluster.

The second case study demonstrates how clustering can be used to create comparisons that are understandable to the field team and would stand up to the review of dealers themselves. Clustering was used to create dealership peer groups, which helped better identify relative performance and highlight strengths and weaknesses. This case study demonstrates that clustering can be used to create meaningful segments that provide useful information to stakeholders.

Overall, clustering and segmentation can be used to improve marketing strategies and operational performance in the automotive industry and beyond. These case studies demonstrate how clustering can be applied in practice and how it can help improve the accuracy of models and the understanding of dealership operations. The success of these efforts relies on selecting appropriate features, utilising the right clustering algorithms and ensuring the accuracy and reliability of the resulting models. Choosing the right metrics for evaluating the cluster assignments helps validate accuracy and alignment to business goals, while clarifying the rigor at which cluster accuracy was measured can provide confidence in use.

The names given to the clusters are often equally as important as the clusters themselves. Auto-labelling clusters based on attributes of the data allows businesses to gain a deeper understanding of why a particular consumer is in each cluster. This enables highly targeted marketing campaigns and personalised customer experiences, which can result in increased customer loyalty and brand advocacy. Auto-labelling of clusters also encourages the use of segmentation, as business analysts are better able to understand the clusters and the behaviours of consumers within them. By utilising automated methods of naming clusters based on the characteristics of their input data, and ensuring that the data is understood by stakeholders, clustering can move beyond the 'black box' modelling realm into something that is widely adopted and touted within organisations.

References

1. Mijač, T., Jadrić, M. and Ćukušić, M. (2018) 'The Potential and Issues in Data-driven Development of Web Personas', in '2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018', pp. 1237–42, available at https://doi.org/10.23919/MIPRO.2018.8400224
2. Ketchen, D. J., Jr. and Shook, C. L. (1996) 'The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique', *Strategic Management Journal*, Vol. 17, No. 6, pp. 441–58.
3. Rousseeuw, P. J. (1987) 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis', *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53–65.
4. Tibshirani, R., Walther, G. and Hastie, T. (2001) 'Estimating the Number of Clusters in a Data Set via the Gap Statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 2, pp. 411–23.
5. Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) 'Data Clustering: A Review', *ACM Computing Surveys (CSUR)*, Vol. 31, No. 3, pp. 264–323.
6. Bertrand, P. and Friggit, V. (2015) 'Cluster Analysis: A Method for Grouping Objects or Variables', *Revue Française de Gestion*, Vol. 41, No. 249, pp. 131–46.
7. Savje, F., Higgins, M. and Sekhon, J. (2018) 'Scclust: Size-constrained Clustering in R', available at https://github.com/fsavje/scclust-R (accessed 6th June, 2023).