[Team Members] Saketh Bandaru

Priyanka Emani

Roopam Verma

Prasanna Sundararajan Muthukumaran

[Instructor Name] Ramakrishna Koganti

[Course Number] CSE 5301-001 Data Analysis & Modeling Techniques

[Date] 04/30/2022

Auto-Insurance Customer Analytics

Abstract

Vehicle insurance companies must employ client segmentation and analysis to mine important customer-related data to get a competitive advantage. While a few attempts have been made to use client division to enhance vehicle insurance decision-making, their client division findings are influenced by the algorithm's characteristics and require multiple validations from different algorithms. This project aims to create a model that analyzes customer data linked with their vehicle insurance policies to segment clients of auto insurance companies using various clustering algorithms for targeted marketing. The customers are sorted into groups using K-Means Clustering and Hierarchical Clustering, and the outcomes are reviewed.

Introduction

Auto-Insurance companies have become an indispensable and integral part of everyone's life. Insurance is an intangible service to consumers, where they essentially are buying protection and peace of mind. It protects the consumers against unexpected damage to their motor vehicle and financial loss. However, It is not a product that consumers actively seek out - they know that they should have it, as it is mandatory to have one. The majority of the consumers are less involved in the purchase decision, primarily due to a lack of interest. As a result, the consumer typically sees very little differentiation between the various competitive offers. This consumer behavior shows that the price, brand, and distribution channel play a more significant role in the successful sale of insurance policies, as opposed to the details and specifications of the policy.

Since the customer is considered a critical factor for insurance companies in producing revenue and improving profitability, keeping customers satisfied is the primary concern of any Insurance company. The profitability of auto-insurance companies mainly depends on the services they offer and on meeting customer demand regularly, so they must find an excellent customer-related strategy to analyze the customer's features and preferences. Customer segmentation is a powerful means of dividing customers into different sets and analyzing their attributes. Auto insurance is one of the essential components of the insurance industry, which is profitable and lucrative. In this paper, we focus on customer segmentation research in auto insurance companies.

Methodology

The steps involved in coming up with a model for analyzing customer's data with their vehicle insurance policies are:

* Dataset Collection
* Data Pre-Processing & Exploratory Data Analysis
* Unsupervised Learning Techniques
  + Hierarchical Clustering
  + K-Means Clustering

Dataset Collection

The dataset comprises the socio-economic details of the customers and the vehicle insurance policies taken for their vehicle from an auto insurance company. This data is obtained from an open-source historical dataset from the Kaggle repository. The dataset consists of 9134 customer records, each described by 24 attributes. The customer lifetime value provided in the dataset, based on historical data, helps understand the customer purchase behavior.

Data Pre-Processing & Exploratory Data Analysis

First, the columns and their corresponding data types are displayed to check whether they contain NULL values or not before applying any data cleansing technique. Since there are no NULL values, no imputation technique is applied to the dataset.

| # | Column | Non-Null Count | Dtype |
| --- | --- | --- | --- |
| 0 | Customer | 9134 non-null | object |
| 1 | State | 9134 non-null | object |
| 2 | Customer Lifetime Value | 9134 non-null | float64 |
| 3 | Response | 9134 non-null | object |
| 4 | Coverage | 9134 non-null | object |
| 5 | Education | 9134 non-null | object |
| 6 | Effective To Date | 9134 non-null | object |
| 7 | EmploymentStatus | 9134 non-null | object |
| 8 | Gender | 9134 non-null | object |
| 9 | Income | 9134 non-null | int64 |
| 10 | Location Code | 9134 non-null | object |
| 11 | Marital Status | 9134 non-null | object |
| 12 | Monthly Premium Auto | 9134 non-null | int64 |
| 13 | Months Since Last Claim | 9134 non-null | int64 |
| 14 | Months Since Policy Inception | 9134 non-null | int64 |
| 15 | Number of Open Complaints | 9134 non-null | int64 |
| 16 | Number of Policies | 9134 non-null | int64 |
| 17 | Policy Type | 9134 non-null | object |
| 18 | Policy | 9134 non-null | object |
| 19 | Renew Offer Type | 9134 non-null | object |
| 20 | Sales Channel | 9134 non-null | object |
| 21 | Total Claim Amount | 9134 non-null | float64 |
| 22 | Vehicle Class | 9134 non-null | object |
| 23 | Vehicle Size | 9134 non-null | object |

Table 1. Data types of the columns in the dataset

Then, the Exploratory Data Analysis (EDA) is performed using the describe() function. This function displays the count, mean, standard deviation, quantiles, and the minimum and maximum values of the data.

|  | Customer Lifetime Value | Income | Monthly Premium Auto | Months Since Last Claim | Months Since Policy Inception | Number of Open Complaints | Number of Policies | Total Claim Amount |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 9134.00 | 9134.00 | 9134.00 | 9134.00 | 9134.00 | 9134.00 | 9134.00 | 9134.00 |
| mean | 8004.00 | 37657.38 | 93.219 | 15.09 | 48.06 | 0.38 | 2.96 | 434.08 |
| std | 6870.96 | 30379.90 | 34.40 | 10.07 | 27.91 | 0.91 | 2.39 | 290.50 |
| min | 1898.00 | 0.00 | 61.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.09 |
| 25% | 3994.25 | 0.00 | 68.00 | 6.00 | 24.00 | 0.00 | 1.00 | 272.26 |
| 50% | 5780.18 | 33889.50 | 83.00 | 14.00 | 48.00 | 0.00 | 2.00 | 383.94 |
| 75% | 8962.17 | 62320.00 | 109.00 | 23.00 | 71.00 | 0.00 | 4.00 | 547.51 |
| max | 83325.38 | 99981.00 | 298.00 | 35.00 | 99.00 | 5.00 | 9.00 | 2893.24 |

Table 2. Exploratory Data Analysis of the auto-insurance customer dataset

A few key insights from EDA are:

* The Mean value of every column is higher than its corresponding Median values (represented as the 50th percentile), showing that the population graph is slightly right-skewed.
* Since there is a significant difference between the maximum value and the 75th percentile value of some columns like 'Monthly Premium Auto'; 'Total Claim Amount' etc., it suggests that the dataset consists of outliers.

Unsupervised Learning Techniques

Unsupervised learning is an algorithm using which machines learn patterns from unlabelled data. One such technique used here is the Clustering technique. Cluster Analysis, also called Data Segmentation, is an exploratory method that identifies the homogeneous records group. Similar items are grouped into homogeneous groups, whereas dissimilar things are separated and placed into heterogeneous groups. In the case of homogeneous groups, the within-group similarity should be higher, i.e., the distance among the records/items should be less (High intra-class similarity). Heterogeneous groups have high inter-class similarity (Distinctive between clusters).

Hierarchical Clustering (Agglomerative or Divisive Clustering)

Algorithm:

* Start with 'n' records and find all the distances between them using Euclidean or Manhattan distance.
* Then, merge similar records or groups of records until all form a larger group.

The hierarchical cluster model is visualized using a dendrogram below.

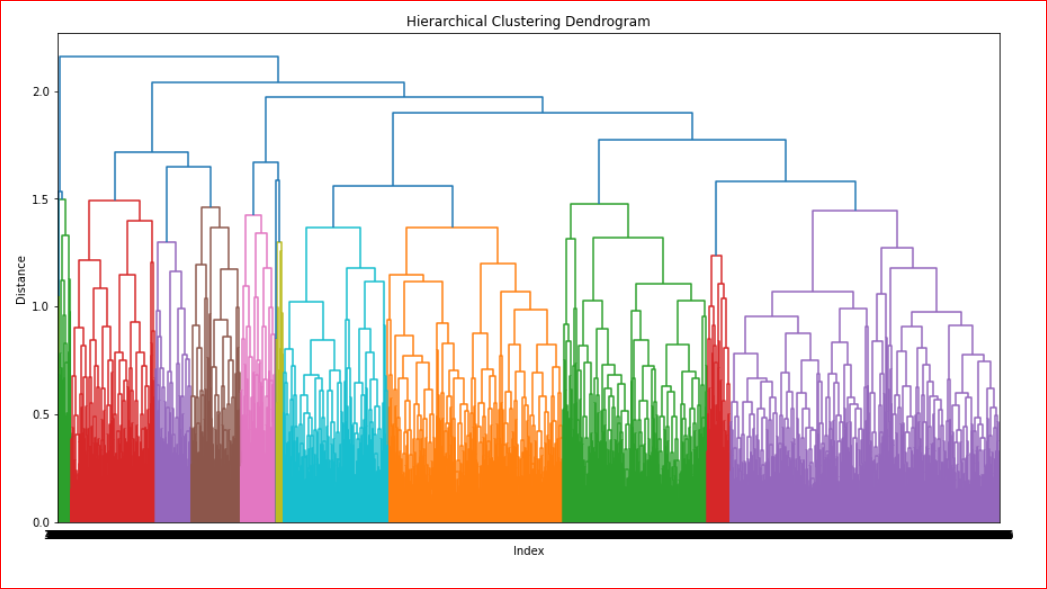


Fig. 1. Dendrogram representing the clusters formed using Hierarchical Clustering

Observation

It is found that the number of clusters created is 11. The various clusters are represented by different colors.

K-Means Clustering

Algorithm:

The number of clusters must be decided upfront, based on

* Square-root(n/2); Where 'n' is the number of data points (For a smaller dataset)
* Using domain knowledge & customer requirements (For a larger dataset)
* Select 'K' data points as 'K' cluster centroids
* Find the distance between all data points to the 'K' cluster centroids
* Assign each of the data points to one of the 'K' centroids based on the nearest distance
* Re-calculate the new centroids by taking the average of the points which form a cluster
* Repeat the above step until the algorithm stops converging

The K-Means cluster model is visualized using the Scree plot or Elbow Curve

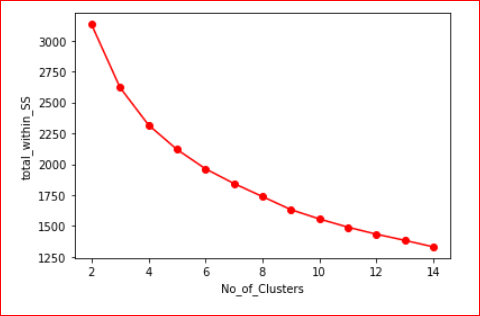


Fig. 2. Scree Plot representing the Elbow Point formed using K-Means Clustering

Observation

The ideal number of clusters to be formed is 3 (Elbow Point), as the maximum drop (slope) is between points 2 and 3.

Conclusion

In this project, the customer data along with their motor insurance policies were analyzed to form clusters using Hierarchical & K-Means Clustering techniques. From the results obtained, it is observed that Hierarchical Clustering works well with datasets containing a small number of records whereas the K-Mean model helped identify the ideal number of clusters to be formed even for very large datasets. For the dataset used above, the ideal number of groups that the customers can be put into is 3. Therefore, the auto insurance company can break down the larger market into 3 smaller segments to serve the customers with the most relevant and valuable experiences possible. Thus it helps in market segmentation and advanced targeting.

Works Cited

[1] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., ―An efficient enhanced k-means clustering algorithm,‖ Journal of Zhejiang University Science A., pp. 1626–1633, 2006

[2] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2001

[3] Hand, D. J., Mannila, H., and Smyth, P. Principles of Data Mining. MIT Press, 2001

[4] Dataset - <https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data>