CUSTOMER SEGMENTATION BASED ON REVIEWS ON DIFFERENT ELECTRIC VEHICLES

(REPORT BY - ABHISHEK KUMAR – ML INTERN AT FYENN LABS)

Objective: To do the customer segmentation based on their reviews about the different electric vehicles.

Dataset Source: The dataset consists of reviews that about the different electric cars available in the Indian market from the car review websites like Carwale.com. The dataset was uploaded on Kaggle and publicly available and the same is uploaded on my GitHub profile.

1. DATA PREPROCESSING (STEPS AND LIBRARIES USED)

In this section, we outline the data pre-processing steps carried out on the datasets using various Python libraries. The pre-processing steps include text data transformation, categorical data encoding, and feature extraction.

Libraries Used:

- `numpy`
- `pandas`
- `matplotlib.pyplot`
- `seaborn`
- `sklearn.feature_extraction.text.TfidfVectorizer`
- `sklearn.preprocessing.MultiLabelBinarizer`
- `collections.Counter`
- `spacy`
- `scipy`
- `re`
- `sklearn.cluster.KMeans`
- `string`
- `gensim.parsing.preprocessing.preprocess_string`

Pre-processing Steps:

1. Text Data Transformation:

Text data from the 'Review' column of the dataset was transformed using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer. This step involves converting text data into numerical vectors, allowing for analysis and clustering based on the content of reviews.

2. Categorical Data Encoding:

Categorical data in the 'Model' column was encoded using one-hot encoding. This was done to convert categorical data into numerical format suitable for analysis. The resulting encoded features represent the different car models present in the dataset.

3. Multi-label Binarization:

For the 'Attributes Mentioned' column, multi-label binarization was performed using the `MultiLabelBinarizer` from `sklearn.preprocessing`. This step converted the attributes mentioned in reviews into binary features, indicating the presence or absence of specific attributes.

4. Feature Matrix Creation:

The pre-processed text data, encoded categorical values, and attributes matrix were combined into a feature matrix. This matrix serves as the input for further analysis and clustering. The 'Rating' column, representing the rating given to each review, was also included in the feature matrix.

5. Data Transformation for Four Wheeler Data from Carwale:

Similar pre-processing steps were applied to the 'four_wheeler_carwale' dataset. Text data from the 'review' column was transformed using TF-IDF vectorization. Categorical values in the 'model_name' column were encoded, and the 'driven' and 'condition' columns were binarized. Numerical attributes, such as 'Exterior', 'Comfort', 'Performance', 'Fuel Economy', and 'Value for Money', were included in the feature matrix.

2. Segment Extraction (ML techniques used)

In this section, we describe the process of segment extraction using the K-means clustering algorithm. K-means clustering is a popular unsupervised machine learning technique that groups similar data points into clusters based on their feature similarity.

K-means Clustering and Determining Optimal K:

The process of segment extraction begins with determining the optimal number of clusters (K). The elbow method is employed to identify the optimal K value. The elbow method involves calculating the sum of squared distances (inertia) for a range of K values and observing the point at which the inertia starts to level off, resembling an "elbow."

We performed the following steps for K-means clustering:

1. **Elbow Method for Optimal K:**

- The column names of the feature matrix were converted to string data types.
- A range of K values (from 1 to 10) was selected for consideration.
- Inertia values were calculated for each K value using the KMeans algorithm.
- An elbow curve was plotted to visualize the inertia values against the number of clusters.

2. **Applying K-means Clustering:**

- Based on the elbow curve, we selected an optimal K value (3 clusters).
- The KMeans algorithm was applied to the feature matrix using the chosen K value.
- Cluster labels were assigned to each data point using the fitted model.

Cluster Assignment and Visualization:

The cluster labels obtained from the K-means clustering were added to the respective datasets. A count plot was created to visualize the distribution of reviews among the identified clusters. This visualization provides insights into the number of reviews in each cluster, helping us understand the segmentation of reviews based on similar attributes.

4. Profiling and Describing Potential Segments

After segment extraction, we proceed to profile and describe the potential market segments. This step involves a comprehensive analysis of each cluster's characteristics and preferences, enabling us to derive actionable insights.

4.1 Distribution of Reviews Among Clusters

To visualize the distribution of reviews among the clusters, a countplot is created. This plot provides an overview of the number of customer reviews within each cluster, offering an initial understanding of segment sizes.

4.2 Average Ratings by Cluster

Analyzing average ratings for each cluster offers insights into overall customer satisfaction. A barplot showcasing average ratings per cluster provides a clear comparison of satisfaction levels.

4.3 Model Preferences Among Clusters

Understanding the most preferred car models within each cluster is essential. By visualizing model preferences, businesses can tailor marketing efforts to align with customer choices.

5. Selection of Target Segment

From the insights gained in the previous steps, we can identify the segment(s) that present the greatest potential for growth or improvement. This selection process guides marketing efforts, ensuring that resources are directed towards segments with the highest impact.

6. Customizing the Marketing Mix

With target segments identified, the next step involves customizing the marketing mix for each cluster. This entails tailoring product positioning, pricing strategies, promotional activities, and distribution channels to cater to the specific needs and preferences of each segment.

7. Potential Customer Base for Business Markets

Analyzing the clusters within the business market provides insights into segments with potential for significant business growth. By understanding these segments' characteristics, businesses can formulate strategies to attract and retain customers within these markets.

8. The Most Optimal Market Segments

Based on various factors such as profitability, growth potential, and alignment with business objectives, the most optimal market segments are determined. These segments are poised to offer the best return on investment and strategic focus.

9. Analyzing Each Cluster

Further analyzing each individual cluster helps extract specific insights:

9.1 Cluster 0

- Characteristics: Lower average ratings, frequent mentions of attributes like "price," "mileage," and "service."
- Recommendation: Focus on improving perceived value and addressing concerns related to pricing and service quality.

9.2 Cluster 1

- Characteristics: Higher average ratings, emphasis on attributes such as "comfort," "performance," and "looks."
- Recommendation: Leverage strengths in comfort and performance, highlighting these attributes in marketing campaigns.

9.3 Cluster 2

- Characteristics: Balanced average ratings, attention to attributes including "price" and "looks."
- Recommendation: Capitalize on the balanced satisfaction levels by reinforcing pricing advantages and aesthetics.

10. Conclusion

Through segment extraction, profiling, and analysis, we have gained valuable insights into the electric vehicle market. By understanding customer preferences and behaviors within different clusters, businesses can tailor strategies, enhance product offerings, and create impactful marketing campaigns. This data-driven approach positions companies for success in an increasingly competitive EV market.