Mehrdad Baradaran

99222020

Third Assignment

Machine Learning

Supermarket dataset for predictive marketing

Unsupervised problems

Outline

- Explore and preprocess the dataset
    - handling categorical variables
    - normalizing or scaling numerical features
    - feature engineering
- Use K-means clustering
    - Experiment with different values of K
    - use elbow method
    - use silhouette score
- Visualize the clusters and analyze their characteristics
    - plotting the clusters in 2D or 3D
        - using PCA or t-SNE
- Experiment other algorithms
    - DBSCAN
    - hierarchical clustering
- compare other algorithms performance with K-means
- reduce data dimensionality (PCA)
- Interpret the results and provide insights to the store owners

Abstract:

We followed a comprehensive workflow to understand and cluster customer purchasing data. We started by loading the data and performing exploratory data analysis (EDA) to gain insights into the dataset. Afterward, we preprocessed the data by encoding categorical variables and applying feature scaling to make it suitable for clustering algorithms.

Using the preprocessed data, we applied the K-means algorithm to cluster the customers. To determine the optimal number of clusters (k), we utilized the elbow method and silhouette score as evaluation metrics. Additionally, we experimented with other clustering algorithms such as DBSCAN and hierarchical clustering to compare their performance against K-means.

In an attempt to reduce data dimensionality, we employed PCA (Principal Component Analysis) and explored different numbers of components. This allowed us to visualize the data in lower-dimensional spaces using techniques like PCA or t-SNE.

Finally, we interpreted the clustering results and provided insights to the store owners. We identified distinct customer segments based on their purchasing behavior and characteristics. The store owners can utilize this information to enhance their marketing strategies, tailor their product offerings, or improve the overall customer experience based on the specific needs and preferences of each customer segment.

Introduction:

Our goal was to gain insights into customer purchasing behavior and identify distinct customer segments to help store owners optimize their business strategies. To achieve this, we followed a systematic approach that involved data preprocessing, clustering using various algorithms, dimensionality reduction using PCA, and visualization techniques.

The key question we aimed to address was how to effectively cluster customers based on their purchasing patterns and identify meaningful segments. By doing so, store owners can tailor their marketing strategies, product offerings, and overall customer experience to meet the specific needs and preferences of different customer segments.

To accomplish our goal, we started by loading the customer purchasing data and conducting exploratory data analysis (EDA). This allowed us to understand the structure of the dataset and uncover any patterns or trends. Next, we performed necessary preprocessing steps, including encoding categorical variables and applying feature scaling to ensure compatibility with clustering algorithms.

We then applied the K-means algorithm, which is a popular unsupervised clustering technique, to partition the customers into distinct clusters. To determine the optimal number of clusters, we employed evaluation metrics such as the elbow method and silhouette score. Additionally, we explored alternative clustering algorithms such as DBSCAN and hierarchical clustering to compare their performance with K-means.

Recognizing the potential benefits of reducing data dimensionality, we utilized Principal Component Analysis (PCA) to transform the data into lower-dimensional representations. This allowed us to visualize the clusters and explore the relationships between variables in a more interpretable manner.

Throughout the analysis, we aimed to provide actionable insights to store owners based on the identified customer segments. By understanding the distinct characteristics and purchasing behaviors of each segment, store owners can tailor their marketing strategies, optimize product offerings, and enhance the overall customer experience.

In summary, our objective was to analyze customer purchasing data, cluster customers into meaningful segments, and provide valuable insights to store owners. By employing a combination of preprocessing techniques, clustering algorithms, dimensionality reduction, and visualization, we aimed to empower store owners to make informed decisions and optimize their business strategies based on a deeper understanding of their customers.

Methods:

1. Data Loading and Exploratory Data Analysis (EDA):
   - Loaded the customer purchasing data into the analysis environment.
   - Conducted exploratory data analysis to understand the data's structure, distribution, and relationships between variables.
   - Explored statistical measures such as mean, median, standard deviation, and correlation coefficients to gain insights into the data.

2. Data Preprocessing:
   - Handled missing values by employing appropriate techniques such as mean imputation, median imputation, or forward/backward filling based on the nature of the missingness.
   - Performed feature engineering to create new features or transform existing features to extract more meaningful information. This could include deriving aggregate statistics, creating interaction terms, or binning continuous variables into categorical groups.
   - Identified and handled outliers in the data by applying techniques such as Winsorization or outlier detection algorithms.

3. Encoding Categorical Variables:
   - Identified categorical variables in the dataset, which are variables that represent discrete categories or labels.
   - Utilized one-hot encoding to transform categorical variables into binary indicators, creating separate binary columns for each category.
   - Alternatively, used label encoding to assign numerical labels to each category, enabling the algorithms to work with the encoded categorical data.

4. Feature Scaling:

   - Applied feature scaling to ensure that variables with different scales or units do not dominate the clustering process.
   - Employed StandardScaler from the scikit-learn library to standardize the numerical features. This technique rescales the data to have zero mean and unit variance, bringing all variables to a similar scale.
   - Standardization is particularly useful when variables have different measurement units or varying ranges, as it prevents certain variables from dominating the clustering based solely on their magnitude.

5. Clustering Algorithms:

   - Utilized the K-means algorithm as the primary clustering algorithm.
   - K-means is a centroid-based algorithm that aims to partition the data

6. Determining Optimal Number of Clusters:

   - Used the elbow method to determine the optimal number of clusters for the K-means algorithm.
   - The elbow method involves fitting the K-means model with different numbers of clusters and plotting the sum of squared distances (inertia) against the number of clusters.
   - Looked for a point in the plot where the decrease in inertia starts to level off, indicating diminishing returns of adding more clusters. This point is often referred to as the "elbow," and it represents a good balance between capturing meaningful clusters and avoiding overfitting.

7. Evaluation Metrics:
    - Calculated the Silhouette Score to evaluate the quality of the clustering results.
    - The Silhouette Score measures how similar each sample is to its own cluster compared to other clusters. Higher scores indicate better-defined clusters.
    - Interpreted the Silhouette Scores to assess the clustering performance, with scores closer to 1 indicating well-separated clusters, scores around 0 indicating overlapping clusters, and negative scores indicating samples assigned to the wrong clusters.

8. Dimensionality Reduction using PCA:
    - Performed dimensionality reduction using Principal Component Analysis (PCA) to reduce the number of features while preserving the most important information.
    - Explored different numbers of components in PCA and evaluated their effects on the clustering performance.
    - Visualized the data in the reduced-dimensional space to gain insights into the cluster structures and relationships between data points.

9. Comparison with Other Clustering Algorithms:
    - Experimented with alternative clustering algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and hierarchical clustering.
    - Evaluated the performance of these algorithms using appropriate metrics and compared them with the results obtained from the K-means algorithm.

10. Visualization and Interpretation:

- Visualized the clustering results using scatter plots, 2D or 3D visualizations, or other suitable techniques.
- Analyzed the characteristics and patterns of each cluster, such as the distribution of data points, centroids, or cluster boundaries.
- Interpreted the clusters to identify distinct customer segments or groups and derived insights from their purchasing behavior or preferences.
- Provided recommendations or insights to the store owners based on the identified customer segments, such as improving marketing strategies, tailoring product offerings, or enhancing the overall customer experience.

11. Feature Engineering:

- Performed feature engineering to derive new features or transform existing features in order to capture more relevant information for the clustering task.
- Specifically, dropped features that had "id" in their name, as these features may not contribute significantly to the clustering process or may not provide meaningful information about customer behavior.
- Conducted a careful analysis of the remaining features to ensure they capture relevant information and improve the clustering results.

Overall, these methods allowed for the exploration, preprocessing, scaling, clustering, and evaluation of the customer purchasing data. The goal was to identify meaningful customer segments and gain insights to support decision-making for marketing strategies, product offerings, and customer experience improvements

Conclusion:

we aimed to uncover distinct customer segments within a store's customer base using clustering techniques. We conducted an in-depth exploration of the dataset and applied various preprocessing steps to prepare the data for clustering. We dropped features that had "id" in their names, as they did not significantly contribute to the clustering process.

For scaling, we utilized the StandardScaler, which transformed the features to have zero mean and unit variance, ensuring that all features were on a similar scale. This preprocessing step helped to prevent features with larger values from dominating the clustering process.

We applied the K-means algorithm to cluster the customers based on their purchasing behavior. Using the elbow method, we determined the optimal number of clusters to be 5. The original K-means model achieved a Silhouette Score of 0.1021 and an inertia value of 16,815,753.34.

To further enhance the clustering results, we performed dimensionality reduction using Principal Component Analysis (PCA). By reducing the dimensionality of the data while preserving important information, we were able to visualize the clusters in lower-dimensional spaces. The filtered K-means model, after dropping irrelevant features, achieved a higher Silhouette Score of 0.3221 and a lower inertia value of 22,253,031.81.

The comparison between different clustering algorithms showed that K-means outperformed DBSCAN and hierarchical clustering in terms of Silhouette Score and inertia. This indicates that K-means was the most suitable algorithm for our dataset and objective.

The clustering analysis revealed distinct customer segments characterized by their purchasing behavior. These segments can be summarized as follows:

- Cluster 0: High-frequency and high-value customers.
- Cluster 1: Moderate-frequency and moderate-value customers.
- Cluster 2: Low-frequency and low-value customers.
- Cluster 3: Medium-frequency and high-value customers.

- Cluster 4: Low-frequency and high-value customers.

The store owners can leverage these insights to optimize their marketing strategy, personalize product offerings, and improve the overall customer experience. For example, they can tailor promotional campaigns to target high-value customers, implement loyalty programs to retain moderate-value customers, and develop strategies to engage and encourage low-value customers to increase their purchase frequency.

In conclusion, this analysis successfully identified distinct customer segments using K-means clustering and provided valuable insights for the store owners to make data-driven decisions. By understanding the characteristics and behaviors of different customer segments, the store owners can optimize their resources, improve customer satisfaction, and drive business growth.

References:

1. Scikit-learn: Machine Learning in Python. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Journal of Machine Learning Research, 12(Oct), 2825-2830.
2. Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.
3. Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.
4. Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027-1035.
5. Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference, 57-61.
6. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

"Note: For each stage of the analysis, including exploratory data analysis (EDA), changing models, and data processing, a more detailed analysis, along with comprehensive conclusions and results, has been documented in the accompanying Jupyter Notebook. Due to the limited space in this report, the focus has been on providing a concise summary. Readers are encouraged to refer to the Jupyter Notebook for a deeper understanding of the steps taken and the detailed findings obtained."