|  |  |
| --- | --- |
| **DATA 430 Technical Report Assignment 4: Clustering** | **Matt Schnorr** |
| **K Means Clustering for Telecommunications** | |
| **URL to dataset: https://www.kaggle.com/datasets/prathamtripathi/customersegmentation** | |

This template should be used in conjunction with the assignment instructions. The size of the text area below will expand to the length of your response; the area should not be interpreted as a required or suggested length of response. Responses within the text area should be single spaced with Times New Roman 12pt font. The body of the document will likely be 6-9 pages, not including the Appendix; length may vary depending on specifics of the analysis and the dataset. As needed, APA format in-text citations should be included, along with a full references list at the end of the document.

|  |
| --- |
| **Overview** |
| **Problem Domain**: give some background and context about the problem domain (application area). For instance, if you are doing the analysis for predicting heart disease, provide some context about the disease and include some interesting statistics about it. Also, discuss how the method is relevant for the chosen problem. |
| Customer segmentation is a crucial aspect for businesses to understand and cater to the diverse needs of their clientele. In the telecommunications industry, customer segmentation can help identify groups of customers with similar characteristics, enabling targeted marketing strategies and personalized service offerings. By analyzing demographic and usage attributes of customers, companies can tailor their products and services to meet specific needs, improve customer satisfaction, and enhance retention rates. For instance, identifying high-value customers or those likely to churn can inform proactive measures to maintain customer loyalty. K-means clustering is particularly relevant here as it helps partition customers into homogeneous groups based on their attributes, facilitating a deeper understanding of customer segments. |
| **Objective**: clearly state the objective of the analysis in relation to the kind of algorithm you are employing. Use specific language as to what question(s) you are trying to answer using the specific analysis/modeling type. |
| The objective of this analysis is to employ K-means clustering to segment the customers in the Telecust1 dataset into distinct groups based on their demographic and usage attributes. The specific question I aim to answer is: "How can we group customers into meaningful clusters that reveal patterns in their demographics and service usage, to better target marketing strategies and improve customer satisfaction?" |
| **Analysis** |
| **Exploratory Analysis**: describe the data including the source, the collection method, and variables. Perform exploratory analysis. Also, select few key variables (including the target variable for supervised learning) and study their distributions using plots such as histograms, box plot, bar chart, etc. |
| The Telecust1 dataset comprises 1000 entries with 12 columns, including region, tenure, age, income, marital, address, ed, employ, retire, gender, reside, and custcat (customer category). The data was likely collected through customer surveys and service records.   * **Data Source**: The dataset originates from a telecommunications company. * **Collection Method**: Likely through customer interaction records and surveys. * **Variables**: The dataset includes both demographic variables (e.g., age, income, marital status) and service-related variables (e.g., tenure, number of services used).   Visualizations (See Figure 1-3):   * **Age Distribution**: Histogram showing the spread of ages among customers. It appears to be a right skewed distribution with a heavy concentration of the values among the lower portion of the age range. * **Income Distribution**: Histogram displaying the income levels of customers. Another heavily skewed view (right skewed) where a vast majority of incomes are at the lower end of the range. This matches with the common sense perspective of wealth distribution in the overall population. * **Tenure Distribution**: Histogram illustrating how long customers have been with the company. This appears flat as there is a near equal distribution across all tenure levels. |
| **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed. |
| * **Handling Categorical Data:** The dataset contains categorical variables (e.g., region, marital, retire, gender, custcat). I dropped these values before generating the model. * **Standardization**: K-means clustering is sensitive to the scale of data. Features with larger scales can dominate the distance calculations, leading to biased clusters. To avoid this, we standardize the data using StandardScaler, which scales the data to have a mean of 0 and a standard deviation of 1. This ensures that each feature contributes equally to the clustering process. * **Justification**: Converting categorical data to numeric codes allows the K-means algorithm to compute distances properly. Standardizing the features ensures that no single feature disproportionately influences the clustering results due to differences in scale. |
| **Cluster Development**: explain the key steps and activities you perform to develop the clusters. Experiment (as appropriate) with parameters tuning. This is key, what separates highly accurate model from a less accurate ones is the amount of performance tuning performed. |
| * **Cluster Elbow Method: T**o determine the optimal number of clusters, we use the elbow method. This involves plotting the sum of squared errors (SSE) for a range of cluster counts (e.g., from 1 to 10) and identifying the "elbow point" where the SSE starts to level off. This point indicates a balance between the number of clusters and the variance within each cluster. * **K-means Clustering:** Based on the elbow plot, we select the optimal number of clusters (e.g., 4 clusters). We then apply K-means clustering with this chosen number of clusters. The K-means algorithm partitions the data into k clusters by iteratively assigning each data point to the nearest cluster center and then recalculating the cluster centers. * **Parameter Tuning**: The primary parameter in K-means clustering is the number of clusters (k). The elbow method helps determine this optimal value. Other parameters such as the number of initializations (n\_init) and the maximum number of iterations (max\_iter) can also be tuned, but default values often suffice. |
| **Results** |
| **Cluster Properties:** explain the properties of the clusters by leveraging distance measures and discuss the clusters characteristics (differences and similarities). Produce appropriate cluster plots and discuss the output. |
| **Cluster Centers**: The cluster centers represent the mean values of all features for the data points in each cluster. These centers provide a summary of the typical characteristics of each cluster.  **Cluster Characteristics**: By examining the cluster centers, we can describe the common traits of customers in each cluster. For example, one cluster might have customers who are younger with higher incomes, while another might consist of older customers with longer tenure.  **Cluster Visualization**: Visualizing clusters helps understand the distribution and separation of clusters. We can use scatter plots to project the clusters in 2D space (e.g., using age and income as axes).  See figure 5 for the values of the cluster centers. |
| **Output Interpretation**: explain the result and interpret the overall clusters using terms that reflect the application area and in relation to the stated objective. This is where you check whether or not the stated objective is met. |
| The clustering results indicate distinct groups of customers with specific demographic and usage patterns. For instance, one cluster might represent young professionals with high income and short tenure, while another could represent retired individuals with long tenure and moderate income. These insights can help the telecommunications company tailor its marketing strategies and service offerings to better meet the needs of different customer segments. (See Figure 6) |
| **Evaluation**: employ appropriate metrics (measures) to quantitatively evaluate the performance of the clusters. For unsupervised classification, this primarily involves distance metrics. |
| The silhouette score, which ranges from -1 to 1, measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters.  In this use case, the silhouette score is .24 (Figure 6). Since the score is somewhat close to zero, it indicates overlapping clusters, which is clear in the visualization.  Therefore, the model either needs fine-tuned, or another machine learning model would be preferred. |
| **Conclusion** |
| **Summary**: highlight the main findings in relation to the stated objective. You don’t need to discuss the details of the analysis and the model such as accuracy here, just focus on the key findings. |
| The K-means clustering successfully segmented the customers into distinct clusters based on their demographic and usage attributes. Key findings include the identification of customer groups with similar characteristics, such as age, income, and tenure, which can inform targeted marketing strategies.  I experimented with going against the elbow recommended number of clusters (both three and five), but the elbow method valuation of four was clearly the best.  The near zero silhouette score tells me the model has overlap, and that is supported with the visualization. More data, more detailed data, or another model may be preferential in this case. |
| **Limitations & Improvement areas**: discuss the limitations of the analysis and identify potential improvement areas for future work. This could be related to the data, algorithm, or a combination of the two. |
| * **Data Limitations**: The dataset might not capture all relevant features influencing customer behavior. Including additional features like customer preferences or detailed service usage data could improve the clustering quality. * **Algorithm Limitations**: K-means assumes clusters are spherical and equally sized, which might not always be the case. Exploring other clustering algorithms like hierarchical clustering or DBSCAN could provide better results. * **Parameter Tuning**: Further tuning of K-means parameters, such as the number of clusters, could enhance the clustering performance. Additionally, using advanced methods like Gaussian Mixture Models (GMM) could improve the clustering accuracy. |

|  |
| --- |
| **Appendix** |
| Figure 1-3:        Figure 4:    Figure 5:    Figure 6: |