Contents

Objective

The main objective of this project is to apply k-means clustering to the scaled dataset (X_train) in order to determine the optimal number of clusters using the Elbow Method, train K-Means clustering Models and Visualising and labelling the clusters to facilitate interpretation and actionable insights. Furthermore, this report discusses how a dimensionality reduction technique like Principal Component Analysis (PCA) helps us resolve visualization problems in high-dimensional data.

Clustering Analysis

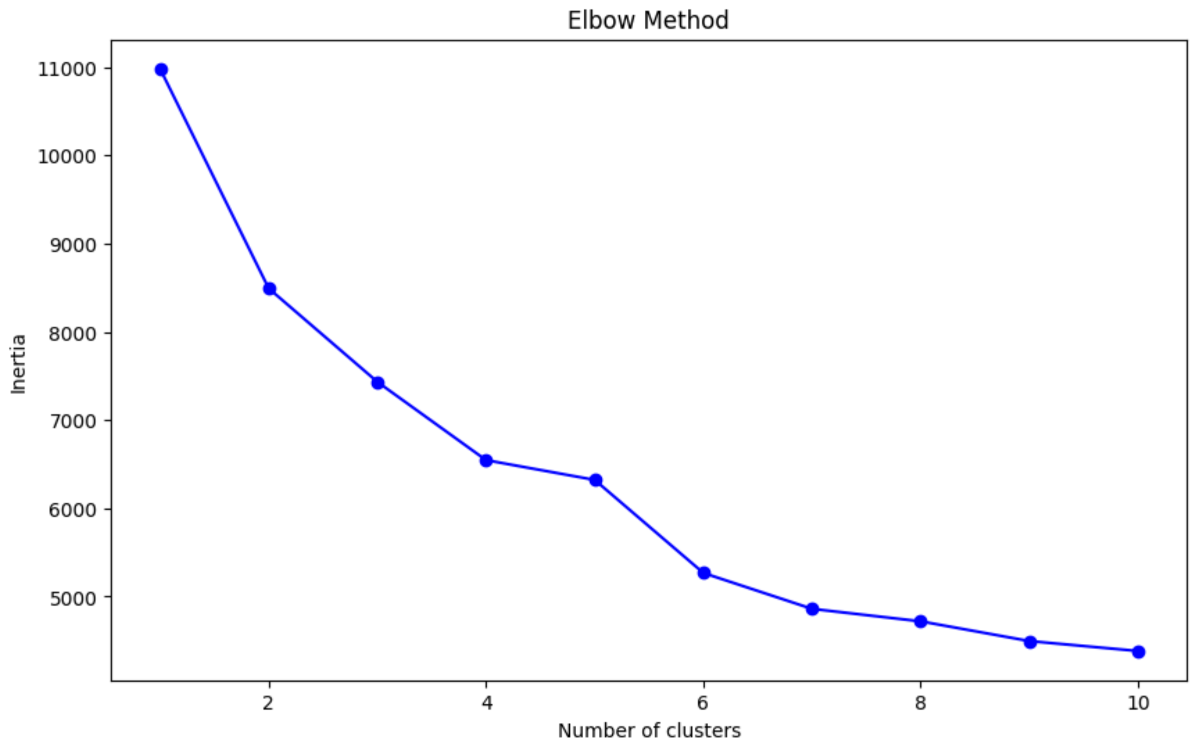
Clustering analysis is a valuable technique of processing data in such a way that organizes data into groups also called clusters based on how closely they are associated. The main goal of this analysis is to find distinct customer segments by grouping customers with similar characteristics,

Elbow Method:

We applied The Elbow Method in our dataset which is a clustering validation technique used to identify the optimal number of clusters for customer segmentation. Here we calculate the inertia which is the sum of squared distances between data point and its corresponding cluster centres, for a range of cluster numbers from 1 to 10 clusters.

We then use the elbow method in which the elbow in the inertia plot represents the point where the rate of decrease in inertia sharply changes forming an elbow-like point. Through this method we find the elbow appears at 4 clusters, suggesting the optimal number of clusters for the data set.

The following is the elbow plot visualizing the inertia for different cluster numbers:



Training K-Means Clustering Model:

K-Means Clustering is an unsupervised learning algorithm meaning that data points do not have a defined classification structure, and is used to partition the data into k distinct clusters which is 4 that we find using the elbow method and these clusters are based on their similarity feature.

Here, we assign each customer to one of the four clusters based on the similarities in their characterises. Then, the model is trained on the prepared dataset (X_train), by assigning a cluster label to each customer.

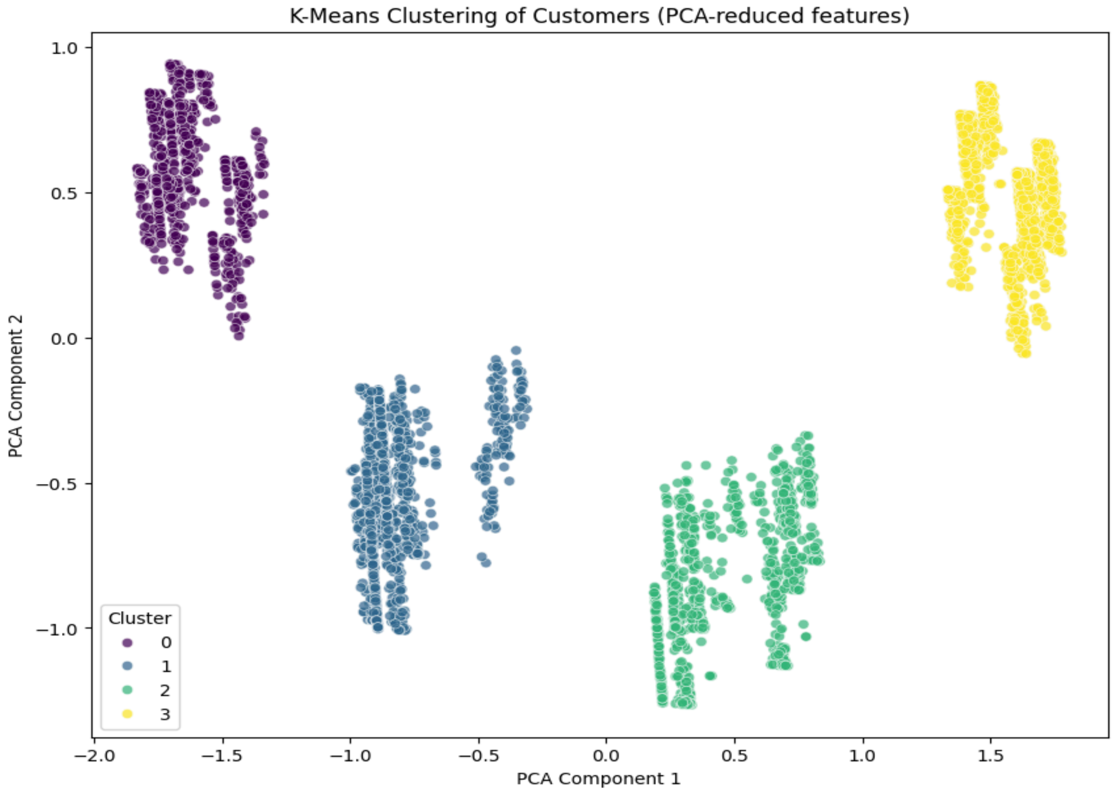
Issue: Overlapping Clusters in High-Dimensional Data

In the previous attempts to visualize the clusters in the scatter plot, we found that the points in the scatter plot overlapped significantly, making it difficult to distinguish between the clusters

Principal Component Analysis (PCA) for Dimensionality Reduction

The overlapping of clusters in the scatter plot was arisen due to the high dimensionality of our dataset. To solve the problem, we applied Principal Component Analysis (PCA) which reduced the dataset into 2 dimensions, retaining most of the variance in the data. Although, it transforms a large set of variables into a smaller one it still contains most of the information in the large set. Then the PCA-reduces data was plotted using Seaborn, and we achieved the distinct cluster separation.

The following is our scatter plot diagram of PCA-reduced data:



Visualizing Clusters:

After applying the PCA, the PCA-reduced dataset was plotted in 2D, where each point was colour-coded based on its cluster assignment, all the clusters were well separated which confirms that our dataset had successfully segmented into meaningful groups using the K-Means Clustering Model.

Labelling Clusters:

After plotting the dataset into a scatter plot, we use cluster labels to calculate the average values of each feature within the clusters. These values provided useful insights into the characteristics of each customer segment. These insights could be used for critical business decisions like targeted marketing and customer retention strategies. The following is the interpretation of each cluster:

Cluster 0:

- **Gender (0.503)**: it means there are roughly 50% male and female.
- **Senior Citizen (0.100)**: This means 10% of the customers are senior citizens.
- **Dependents (0.258)**: 25.8% of the customers have dependents.
- **Tenure (0.198)**: short tenure period (19.8% of the maximum tenure).
- **Phone Service (0.774)**: 77.4% of the customers have phone service.
- **Multiple Lines (0.151)**: 15.1% of the customers have multiple phone lines.
- **Internet Service (0.000)**: No internet service for this cluster.
- **Monthly Charges (0.236)**: Lower monthly charges (23.6% of the maximum charges).

- **Contract Type**: All customers are on a month-to-month contract.

Cluster 1:

- **Gender (0.513)**: Balanced gender distribution.
- **Senior Citizen (0.062)**: Only 6% are senior citizens.
- **Dependents (0.446)**: 44.6% of customers have dependents.
- **Tenure (0.583)**: Medium-length tenure.
- **Phone Service (0.821)**: 82.1% have phone service.
- **Multiple Lines (0.007)**: Almost none of the customers have multiple lines.
- **Internet Service (0.108)**: Low internet service usage just 10.8%.
- **Monthly Charges (0.272)**: Slightly higher monthly charges than Cluster 0 (27.2%).
- **Contract Type**: 54.5% of customers are on one-year contracts, while 45.5% of customers are on two-year contracts.

Cluster 2:

- **Gender (0.496)**: Again, balanced gender distribution.
- **Senior Citizen (0.154)**: 15.4% of the customers are senior citizens.
- **Dependents (0.379)**: 37.9% of the customers have dependents.
- **Tenure (0.804)**: Longer tenure compared to other clusters.
- **Phone Service (1.000)**: All customers have phone service.
- **Multiple Lines (0.950)**: Most customers have multiple lines.
- **Internet Service (0.522)**: 52.2% of the customers have internet service.
- **Monthly Charges (0.626)**: Higher monthly charges.
- **Contract Type**: 59.2% of customers are on two-year contracts, while 40.8% of customers are on one-year contracts.

Cluster 3:

- **Gender (0.504)**: Again, balanced gender distribution**.**
- **Senior Citizen (0.308)**: 31% senior citizens, highest among the clusters.
- **Dependents (0.167)**: Only 16.7% have dependents.
- **Tenure (0.305)**: Short tenure.
- **Phone Service (1.000)**: All customers have phone service.
- **Multiple Lines (0.581)**: 58.1% of the customers have multiple lines.
- **Internet Service (0.996)**: Almost all customers have internet service.
- **Monthly Charges (0.683)**: Highest monthly charges across the clusters.
- **Contract Type**: All on month-to-month contracts.

Observations:

- **Cluster 0**: it represents the cluster with no internet service, short tenure, lower monthly charges and all with month-to-month contracts.
- **Cluster 1:** It represents the cluster with moderate tenure and equally preferred both one-year and two-year contracts.

- **Cluster 2:** It represents the cluster with long tenure, multiple lines and high monthly charges. Most of the customers prefer two-year contracts (59.2%), with a significant proportion of one-year contracts (40.8%).
- **Cluster 3:** It represents the cluster with a high number of senior citizens, short tenure, almost all customers use phone and internet services, pay high monthly charges and all with month-to-month contracts.

Conclusion

In conclusion, we successfully applied the Elbow Method to find the optimal number of clusters for customer segmentation. We used K-Means Clustering to train the model and visualise the clusters using PCA and also labelled the clusters for easy interpretation. These insights are valuable for better business decision-making like; targeted marketing and customer retention strategies.