Using K-Means Clustering to Identify Customer Segments and Reduce Churn at Telelink

Marisela Monterroza

D212:Task 1

Western Governors University (WGU)

September 2024

Part 1: Research Question:

A.

1. How can we use K-Means clustering to classify Telelink customers based on tenure, income, and monthly charges? We want to find out what long-term customers have in common and come up with ways to lower the churn rate.
2. The goal of this data analysis is to use K-Means clustering to classify Telelink customers based on their time with the company, income, and monthly charges. By finding different groups of customers, we hope to see patterns that help us understand why some customers stay longer and why others leave. This will help stakeholders create strategies to keep more customers, like offering special deals or improving services.
    a. It is important to specify, when it comes to clusters the correlation coefficient is not the main target. I will be focusing on the elbow method instead to analyze how tightly points fit within a cluster.

Part II: Technique Justification

B.

1. **K-means clustering** works by dividing customers into groups based on how similar they are to one another, using the variables like tenure, income, and monthly charges.
    i. K-means selects a number of clusters and assigns each customer to the nearest cluster based on the distance between their data points and centroid.
    ii. Then it adjusts the cluster centers until the groupings stabilize.
    b. **Expected outcomes** include identifying distinct customer segments, such as:
        i. Long-term customers who pay higher monthly charges.
        ii. Customers that could benefit from targeted marketing strategies to reduce churn.

2. One assumption of K-means clustering is that the number of clusters (k) must be chosen ahead of time. This assumes that the data can be divided into a set number of distinct groups, and the data points within each cluster are more similar to each other than to points in other clusters.
3. List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

    NumPy: This library will help in performing numerical operations and manipulating arrays, which is crucial for preparing and analyzing the dataset.

    Pandas: Pandas will be used to handle data preprocessing and manipulation. It makes it easy to read data from files and perform operations like filtering, sorting, and cleaning the data.

Scikit-learn: This package includes the KMeans algorithm, which will be used to perform the clustering analysis. It also offers methods to calculate the optimal number of clusters, such as the "Elbow Method" or "Silhouette Score."
Matplotlib/Seaborn: These libraries will be used for data visualization. They can plot scatterplots of clusters and centroids to help visualize the customer segments.
numpy: Used for numerical operations (like arrays and matrices).
pandas: Essential for handling structured data (like CSV files or dataframes).
seaborn: High-level interface for visualizing data.
matplotlib.pyplot: The foundational library for creating plots.
KMeans from scikit-learn: For performing K-means clustering.

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:

1. Describe **one** data preprocessing goal relevant to the clustering technique from part A1.

The primary preprocessing goal is to ensure that all the variables used in the clustering analysis are standardized and clean. Since K-means clustering is sensitive to the scale of data, variables like tenure, income, and monthly charges should be normalized to ensure that no single feature dominates the clustering process.

Goal: Normalize/standardize the tenure, income, and monthly charges to have similar scales, and handle any missing data in these columns.

2. Identify the initial data set variables you will use to perform the analysis for the clustering question from part A1, and label *each* as continuous or categorical.

- **Tenure** (continuous): Number of months a customer has been with Telelink.
- **Monthly Charges** (continuous): The amount a customer pays each month for Telelink services.
- **Income** (continuous): The customer's annual income.

Tenure:

- Reasoning: Tenure directly relates to how long a customer has been with the company. This is essential for understanding customer loyalty and retention trends.

Monthly Charges:

- Reasoning: Monthly charges reflect customer spending habits with the company, which can help identify high-value customers or those more likely to switch to competitors.

Income:

- Reasoning: Income levels can help identify which customers are more likely to respond to premium offers or discounts. It can also help segment customers by their purchasing power.

Note: All selected variables are continuous, as K-means clustering works best with continuous data.

3. Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.
Step 1: Load the data
Step 2: Handle missing values
Step 3: Standardize the data
Step 4: Prepare the data for clustering

4. Provide a copy of the cleaned data set.

Part IV: Analysis

D. Perform the data analysis, and report on the results by doing the following:

1. Determine the optimal number of clusters in the data set, and describe the method used to determine this number.
The optimal number of clusters in the dataset was determined using the **Elbow Method**. This method involves plotting the Sum of Squared Errors (SSE) against different values of kkk (the number of clusters). The point at which the SSE begins to decrease at a slower rate (forming an "elbow" shape) indicates the optimal number of clusters. In this analysis, the optimal kkk was determined to be **3**, as this point balances complexity and cluster quality.
2. Provide the code used to perform the clustering analysis technique

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
```

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Assuming df is your cleaned and prepared DataFrame
df_scaled = StandardScaler().fit_transform(df[['Tenure', 'MonthlyCharge', 'Income']])

# Using Elbow Method to determine optimal k
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(df_scaled)
    sse.append(kmeans.inertia_)

# Plotting Elbow Method
plt.figure(figsize=(8, 5))
plt.plot(range(1, 11), sse, marker='o')
plt.title('Elbow Method for Optimal K')
plt.xlabel('Number of clusters')
plt.ylabel('SSE (Sum of squared distances)')
plt.show()

# Performing K-means clustering with optimal k
kmeans = KMeans(n_clusters=3, random_state=42)
df['Cluster'] = kmeans.fit_predict(df_scaled)

# Print cluster centers
print("Cluster centers:")
print(kmeans.cluster_centers_)
```

1. Explain the quality of the clusters created.
   The clusters created in this analysis were evaluated using the **Silhouette Score**, which provides insight into the quality of the clusters. A Silhouette Score of **0.35** indicates that the clusters are moderately well-defined, with some overlap. The analysis also included box plots to visualize differences in Monthly Charges across the clusters, revealing that Clusters 0 and 1 are similar, while Cluster 2 is distinctly different.
2. Discuss the results and implications of your clustering analysis.
   The clustering analysis identified three distinct segments among Telelink customers based on tenure, monthly charges, and income.

- **Cluster 0**: Represents long-term customers with lower charges and income.
- **Cluster 1**: Includes new customers with moderate charges.

- **Cluster 2**: Comprises high-income customers with high monthly charges.

These insights can inform targeted marketing strategies to reduce churn, such as offering loyalty discounts to long-term customers and tailored services to high-income segments to encourage retention.

3. Discuss **one** limitation of your data analysis.
One limitation of the data analysis is the reliance on the selected features (tenure, monthly charge, income) for clustering. Other factors, such as customer satisfaction or service usage patterns, were not considered and may influence churn rates. This may lead to incomplete customer profiles and less effective strategies.
4. Recommend a course of action for the real-world organizational situation from part A1 based on the results and implications discussed in part E2.
Based on the clustering analysis results, I recommend that Telelink implement targeted marketing strategies:

- **For Cluster 0**: Introduce loyalty programs to reward long-term customers and reduce churn.
- **For Cluster 1**: Develop introductory offers and services to increase satisfaction and retention among newer customers.
- **For Cluster 2**: Offer premium services or upgrades that align with their higher spending capacity to enhance customer loyalty.

https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
https://www.ibm.com/topics/k-means-clustering
https://trcmarketresearch.com/whitepaper/cluster-analysis-gets-complicated/
https://www.mbmlbook.com/ModelAnalysis_K-means_Clustering.html
https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/
https://www.investopedia.com/terms/c/correlationcoefficient.asp