D212: Data Mining II Task 1

Justin Huynh

Student ID: 012229514

M.S. Data Analytics

A1. Proposal of Question

My research question for this project is "Can we segment customers into different groups based on their usage patterns (monthly charges, bandwidth usage) to better understand who is at risk of churn?" We will use the k-means clustering technique to answer our research question using only the continuous variables MonthlyCharge and Bandwidth_GB_Year.

A2. Defined Goals

The goal of this analysis is to identify customer segments based on usage patterns. By understanding these segments, an organization can tailor their retention strategies more effectively by targeting specific groups of customers who might have a higher risk of churn. This segmentation will help an organization take action by creating personalized offers and enhance customer satisfaction to reduce churn rates.

B1. Explanation of the Clustering Technique

K-means clustering is a partitioning method that divides a dataset into distinct non-overlapping clusters based on similarity. It works by iteratively assigning data points (customers) to clusters such that the sum of the squared distances between data points and the cluster centroids is minimized. For this analysis, k-means will be applied to the continuous variables MonthlyCharge and Bandwidth_GB_Year. The algorithm will do the following:

- Initialize by randomly selecting k centroids (where k is the number of clusters).
- Assign each customer to the nearest centroid based on their usage patterns.
- Recalculate the centroids of the newly formed clusters.

- Repeat the process until the centroids no longer change significantly, indicating that the clusters are stable.

The expected outcome will have the distinct customer segments identified based on their monthly charges and bandwidth usage. Each segment will represent a group of customers with similar usage patterns. By examining the distribution of churn across these segments, an organization can identify which segments are at higher risk of churn.

B2: Summary of the Technique Assumption

K-means clustering assumes that the clusters are spherical in shape and that each cluster has roughly the same size or equal variance. This assumption implies that the data within each cluster is uniformly distributed around the centroid, and the clustering boundaries are equally distant from the centroids. If the actual clusters in the data are not spherical or have varying sizes, k-means may not perform optimally, leading to suboptimal cluster assignments.

B3: Packages or Libraries List

The benefits of using Python, which I will be using for this project's analysis, are the comprehensive libraries and the visualization capabilities. The libraries such as Pandas and NumPy are great at facilitating statistical analyses and making data manipulation more efficient. The scikit-learn library is great for robust implementations of the k-means algorithm, and it's the primary library for performing clustering analyses. The visualizations that we can create from libraries such as Matplotlib and Seaborn are great at creating graphs and plots that we need to not only visualize our data but to also

identify trends and understand data distributions. This makes communicating results to stakeholders a lot simpler and effective from a technical perspective.

C1: Data Preprocessing

The primary preprocessing goal for the k-means clustering is to ensure that the continuous variables used in the analysis are standardized. K-means clustering is sensitive to the scale of the input features because it relies on Euclidean distance to assign data points to clusters. So it's crucial to standardize the continuous variables, such as MonthlyCharge and Bandwidth_GB_Year, so that they have a mean of 0 and a standard deviation of 1. This will help prevent variables with bigger ranges from influencing the clustering results disproportionately.

C2: Data Set Variables

For the k-means clustering analysis, the following variables will be used:

- MonthlyCharge (continuous): represents the average monthly charge for the customer.
- Bandwidth_GB_Year (continuous): represents the average amount of data used by the customer in a year

These continuous variables are directly related to the customers' usage patterns and are key to segmenting the customers based on their behavior.

C3: Steps for Analysis

Here are the steps to prepare the data for analysis:

- We first select the relevant continuous variables, MonthlyCharge and Bandwidth_GB_Year, for the clustering analysis.

```
# select the relevant continuous variables for clustering
df_cluster = df[['MonthlyCharge', 'Bandwidth_GB_Year']]
```

- We check for any missing values in the selected columns and handle them accordingly. If any missing values are found, we might drop the rows or impute the missing values.

```
# check for missing values in the selected columns
missing_values = df_cluster.isnull().sum()
```

```
# drop rows with missing values (if any)
df_cluster_clean = df_cluster.dropna()
```

- Lastly, to ensure that the variables are on the same scale, we standardize the MonthlyCharge and Bandwidth_GB_Year columns. This step is crucial for k-means clustering because it ensures that each variable contributes equally to the distance calculation.

```
# standardize the data
scaler = StandardScaler()
df_cluster_scaled = scaler.fit_transform(df_cluster_clean)
```

```
# convert back to dataframe for ease of use
df_cluster_scaled = pd.DataFrame(df_cluster_scaled, columns=['MonthlyCharge', 'Bandwidth_GB_Year'])
```

See code attached in WGU_D212_Task_1.ipynb.

C4: Cleaned Data Set

A copy of the fully prepared data set will be submitted as ‘prepared_data_d212_task1.csv’.

D1: Output and Intermediate Calculations

The Elbow Method is a technique used to determine the optimal number of clusters in a dataset. The idea is to run the k-means clustering algorithm for a range of values of k (number of clusters), and for each value of k, calculate the sum of squared distances (also known as inertia) between data points and their assigned cluster centroids.

The optimal number of clusters is determined by identifying the "elbow point" in the plot of the sum of squared distances against the number of clusters. The elbow point is where the inertia begins to diminish at a slower rate, indicating that adding more clusters beyond this point does not significantly improve the model.

```
# determine the optimal number of clusters using the elbow method
inertia = []
k_range = range(1, 11)

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    kmeans.fit(df_cluster_scaled)
    inertia.append(kmeans.inertia_)

# plotting the elbow method results
plt.figure(figsize=(8, 5))
plt.plot(k_range, inertia, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('Elbow Method for Determining Optimal k')
plt.show()

# perform K-means clustering with the optimal number of clusters
optimal_k = 3
# replace with the optimal k determined from the plot
kmeans = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
df_cluster_scaled['Cluster'] = kmeans.fit_predict(df_cluster_scaled)
```

See code attached in WGU_D212_Task_1.ipynb.

D2: Code Execution

See code attached in WGU_D212_Task_1.ipynb.

E1: Quality of the Clustering Technique

The quality of the clusters created using the k-means clustering technique can be assessed by examining the degree of separation between clusters and the consistency within each cluster. In this analysis, the elbow method was used to determine that three clusters provided an optimal balance between the number of clusters and the inertia (sum of squared distances from each point to its assigned cluster center). The drop in inertia at three clusters, followed by a slower rate of decrease, suggests that these clusters are well-defined and appropriately capture the underlying patterns in the data. The k-means algorithm works effectively in this context because the selected variables, MonthlyCharge and Bandwidth_GB_Year, are continuous and well-suited to the euclidean distance metric that k-means relies on.

However, even though the clusters seem to be well-separated, the inherent assumption of k-means that clusters are spherical and of similar size could be a limitation if the true shape of the clusters in the data is not spherical. Despite the potential limitation, the clustering technique used in this analysis provides a meaningful segmentation of customers that can be leveraged for decision-making.

E2: Results and Implications

The results show the k-means clustering identified 3 distinct customer segments based on their monthly charges and annual bandwidth usage. These segments could be characterized as:

- Cluster 1: High spenders with high bandwidth usage. These customers are probably heavy internet users and may require higher-end services and faster internet plans.
- Cluster 2: Moderate spenders with moderate bandwidth usage. These customers are likely typical users who may benefit from standard plans.
- Cluster 3: Low spenders with low bandwidth usage.

These customers might be light internet users and probably looking for budget-friendly options. The implications suggest that an organization could use targeted marketing to tailor its strategies to each cluster. For example, high spenders might be offered premium packages or additional services, while low spenders could be targeted with budget-friendly promotions. Another implication is understanding these segments can help identify which clusters are more likely to churn. For instance, low spenders might be more sensitive to price changes and could benefit from retention efforts focused on cost-effective solutions. Lastly, an organization can optimize its service offerings based on the needs of each cluster to make sure that the customers receive the most relevant services according to their usage patterns.

E3: Limitations

One limitation is the analysis was conducted using only two continuous variables: MonthlyCharge and Bandwidth_GB_Year. While this provides a useful segmentation based on usage patterns, it does not consider other important factors such as customer satisfaction, technical support interactions, etc. Including these additional variables could lead to a more nuanced and accurate clustering that reflects the complexity of customer behavior more in-depth.

E4: Course of Action

There are several courses of action an organization can take based on the results and implications. One being understanding the lower spenders' reason for low usage and tailoring retention strategies to prevent churn. Since they're price-sensitive, small incentives could help improve churn rate significantly. Another action could be upsell opportunities for higher spenders such as higher bandwidth, exclusive promos, etc. since they're already less price-sensitive and most likely focused on value-added services. Lastly, an organization can customize different services per different clusters to maintain quality and consistent services that'll help retain or improve churn rate. By implementing these strategies, an organization can enhance customer satisfaction, reduce churn, and increase overall revenue by catering more precisely to the needs of each customer segment.

F: Panopto Video of Code

The URL link will be submitted in the PA task submission.

F1: Panopto Video of Programs

The URL link will be submitted in the PA task submission.

G: Sources for Third-Party Code

1. "K-means Clustering - Introduction" Retrieved from https://www.geeksforgeeks.org/k-means-clustering-introduction/

H: Sources

1. "Understanding K-means Clustering in Machine Learning" Retrieved from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

2. “KMeans” Retrieved from

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html