Milestone3: Data Modeling

CONTENTS

INTRODUCTION

This milestone aims to implement and analyze a K-means clustering model to group customers based on their financial behavior. Clustering is used to identify patterns or segments within customer data that can be used to better understand customer behavior and inform business decisions. This model will provide insights into customer segments that exhibit similar financial profiles.

MODEL SELECTION

We chose K-means clustering as the model to group customers based on similar financial patterns because it is a simple, efficient, and effective algorithm for segmentation tasks. K-means partitions the data into a predefined number of clusters where each group shares comparable properties. Its ability to identify patterns based on similarity makes it well-suited for customer segmentation, allowing us to group customers based on financial attributes such as annual salary, years of residence, and other key variables. This approach is particularly useful for identifying trends and patterns within customer data, which can then be used for targeted marketing, financial planning, or service offerings.

WHY K-MEANS?

- Simplicity and Efficiency: K-means clustering is computationally efficient for large datasets and provides clear, interpretable results.
- Handling Numeric Data: Since our dataset contains continuous numeric attributes like annual salary, gross pay, and years of residence, K-means clustering was appropriate.
- Centroid-Based Interpretation: K-means allow for easy interpretation of clusters through their centroids, which represent the average characteristics of each cluster. This is particularly useful for understanding customer segments based on financial stability and affordability.
- Unsupervised Learning: Since we do not have a predefined target variable, clustering is an appropriate technique. K-means helps us discover hidden patterns in the data by grouping customers into clusters based on similar financial attributes.
- Efficient for Large Datasets: K-means is computationally efficient, especially for large datasets like ours, which contain over 190,000 rows. This makes it suitable for initial exploration and segmentation tasks.
- Interpretability: K-means is intuitive as it partitions the data into a predefined number of clusters where each customer is assigned to the nearest centroid. The centroids represent the average values of the financial attributes for that cluster, making it easy to interpret.
- Business Relevance: Clustering customers based on financial attributes like Annual Salary, Gross Pay Last Paycheck, and Gross Year-To-Date provides valuable insights for business strategies such as identifying high-income customers or customers with irregular pay patterns.

DATA PREPARATION

The dataset was cleaned and processed in Milestone 2. The following columns were retained for analysis:

- Year of Birth: Removed due to irrelevance for clustering.
- Age Group: Removed as it was derived from "Year of Birth."

- Gross Pay Last Paycheck and Gross FRS Contribution: Removed to focus on variables directly tied to yearly financial stability.

FINAL DATASET COLUMNS USED:

- Marital Status
- Education
- Occupation
- Annual Salary
- Gross Year To Date
- Household Size
- Years of Residence

DATA NORMALIZATION

To ensure that all features are on a similar scale, we normalize the data using the scale () function. Normalization is crucial for clustering because it prevents features with larger ranges from dominating the results.

which plots the total within-cluster sum of squares against the number of clusters. The elbow was identified at k, indicating that clusters were optimal for our data.

MODEL BUILDING

Principal Component Analysis (PCA)

- Purpose: PCA was performed to reduce the complexity of the dataset and improve clustering by focusing on the most significant patterns. Dimensionality reduction allowed us to focus on the two principal components that capture the most variance.
- Results: The first two principal components were extracted and used for visualizing and clustering the data. These components reflect the major financial patterns of customers, such as the relationship between their annual salary and gross pay.

K-MEANS CLUSTERING

Number of Clusters: The Elbow Method was used to determine the optimal number of clusters. This method plots the total within-cluster sum of squares (WSS) against the number of clusters (k). The "elbow" of the curve, where the decrease in WSS begins to level off, indicates the ideal number of clusters. In this case, the elbow was found at k = 6, suggesting that 6 clusters are optimal.

Cluster Sizes:

- Cluster 1: 40,227 data points
- Cluster 2: 25,757 data points
- Cluster 3: 31,537 data points

- Cluster 4: 26,292 data points
- Cluster 5: 31,158 data points
- Cluster 6: 30,085 data points

Cluster Centers: Each cluster has a centroid, which is the average position of all data points in that cluster. These centroids summarize the characteristics of the cluster in terms of financial attributes such as salary and gross pay. For instance, clusters with higher centroid values for salary and gross pay represent wealthier customer segments.

TEST DESIGN AND MODEL EVALUATION

MODEL EVALUATION:

Within-Cluster Sum of Squares (WSS): The performance of K-means clustering is evaluated based on the WSS for each cluster. The goal is to minimize WSS, which indicates that data points within a cluster are close to the centroid and therefore like each other.

Between-Cluster Sum of Squares (BSS): The ratio of between-cluster sum of squares to total sum of squares (BSS/TSS) was 76.8%, meaning that the clusters explain 76.8% of the variability in the data.

Affordability-Based Clustering: We introduced an additional feature, affordability, to assess financial stability. Customers were categorized as either "Affordable" or "Not Affordable" based on their annual salary. A threshold of $50,000 was set to distinguish between the two groups.

Visualization: Using scatter plots, we visualized the relationship between Gross Year-To-Date and Annual Salary, colored by affordability. This revealed clear patterns in financial behavior, with "Affordable" customers clustered together and "Not Affordable" customers forming distinct groups.

RESULTS INTERPRETATION

ELBOW METHOD FOR OPTIMAL CLUSTERS

- The **Elbow Method** helped identify 6 clusters as the optimal number. The WSS decreased significantly when increasing from 2 to 6 clusters, but further increases in cluster count did not reduce WSS, indicating that 6 clusters effectively capture the data's structure.

CLUSTER CENTERS

- The **centroids** of each cluster represent the average values for the financial variables. For instance, customers in Cluster 1 may have lower salaries and gross pay compared to those in Cluster 6, where the average salary and gross pay are significantly higher.
- By examining the centroids, we can infer which clusters contain wealthier or more financially stable customers.

AFFORDABILITY ANALYSIS

- Customers classified as **Affordable** (Annual Salary >= $50,000) were found to cluster together, while those classified as **Not Affordable** tended to form separate groups.
- **Gross Year-To-Date vs. Annual Salary**: The scatter plot showed that affordable customers have higher gross earnings to date, indicating more stable financial behavior.

ADDITIONAL VISUALIZATIONS

- **Affordability vs. Years of Residence**: This visualization provided insights into how long customers have lived at their current address, revealing that customers with higher gross pay and salaries also tended to have longer tenures at their residence, indicating stability.
- **Affordability vs. Household Size**: By examining household size, we found that larger households (3+ members) were more likely to be classified as "Affordable," due to higher combined incomes.

CONCLUSION

In conclusion, K-means clustering proved to be an effective method for segmenting customers based on financial attributes. The clusters revealed distinct customer groups with varying financial behaviors, helping to identify wealthier, more stable customers versus those with more unpredictable earnings patterns. The affordability classification added an extra layer of insight, enabling us to pinpoint which customers were likely to have stronger financial health.

Future work will involve refining these clusters by experimenting with additional variables and testing the model's applicability to real-world business decisions, such as targeted marketing or customer support strategies.