**Capstone Project**

Data Science Nanodegree

**5/26/2024**

**Definition**

**Project Overview:**

The Problem: Define the problem as a clustering task using K-means. The objective is to identify distinct user segments, such as discount enthusiasts, BOGO enthusiasts, and disinterested users, to tailor marketing messages accordingly.

The ultimate goal is to clearly define which is to maximize engagement and conversion rates by delivering personalized marketing messages to each segment.

Understand the business objective thoroughly. In this case, the goal is to optimize direct marketing campaigns by segmenting app users based on their responsiveness to promotional offers.

Libraries used:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from datetime import datetime

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

**Metrics:**

In this project, we primarily used descriptive statistics to assess the clustering result. These statistics include mean and standard deviation for various attributes across different clusters. Additionally, we examined the distribution of categorical variables such as gender.

For model evaluation, we did not explicitly use any quantitative metrics such as silhouette score, inertia, or Davies–Bouldin index. Instead, we relied on visual inspection of cluster centroids and data points, along with some knowledge, to interpret the clustering result and assess its practical utility for optimizing direct marketing campaigns.

**Methodology:**

In this project, our primary goal was to segment app users into distinct groups based on their responsiveness to various promotional offers using K-means clustering.

Here's a detailed description of the overall approach and methodology we employed to solve the problem:

**Overall Approach**

1. **Data Preparation and Preprocessing**:
   * We started by exploring and cleaning the dataset, ensuring that all relevant features were included and any missing or irrelevant data was handled appropriately.
   * We encoded categorical variables, such as gender, to numerical values suitable for K-means clustering.
   * Features related to user behavior and demographics were scaled to ensure they were on the same scale, which is crucial for distance-based algorithms like K-means.
2. **Feature Selection**:
   * We selected features that were relevant to understanding user behavior in response to promotional offers. This included features like offer interaction events, transaction amounts, and user demographics such as income and age.
3. **Clustering with K-means**:
   * **K-means Clustering**: We chose K-means as our clustering algorithm due to its simplicity and effectiveness in partitioning data into distinct groups. K-means is suitable for this problem because it aims to minimize the within-cluster variance, thereby grouping similar users together based on their feature values.
   * **Elbow Method for Optimal K**: To determine the optimal number of clusters (K), we used the elbow method. This involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the point where the rate of decrease slows down, indicating the optimal K.
4. **Model Tuning**:
   * We experimented with different numbers of clusters to find the optimal segmentation of users.
   * After selecting the optimal number of clusters, we examined the centroids of each cluster to understand the characteristics of each segment.
   * We visualized the clusters using scatter plots to inspect the separation between clusters and the distribution of data points within each cluster.
5. **Model Evaluation**:
   * **Visual Inspection**: We plotted the data points with their cluster labels and centroids to visually inspect the clustering results.
   * **Cluster Characteristics Analysis**: We analyzed the mean and standard deviation of features within each cluster to understand the distinct patterns and behaviors of users in each segment.
   * **Interpretation of Clusters**: We interpreted the clusters in the context of our business objective—optimizing direct marketing campaigns by tailoring promotional messages to different user segments.

**Techniques and Algorithms:**

* **K-means Clustering**: Suitable for partitioning data into distinct, non-overlapping clusters based on feature similarity. Its simplicity and interpretability make it a good choice for this segmentation task.
* **Elbow Method**: Used to determine the optimal number of clusters by analyzing the within-cluster sum of squares (WCSS) and identifying the point where adding more clusters yields diminishing returns.
* **Standard Scaling**: Ensured that all features were on the same scale, which is crucial for distance-based clustering algorithms.

**Justification:**

* **K-means Clustering**: Chosen for its efficiency and effectiveness in creating distinct clusters. It works well with the type of data we have and helps in understanding the natural grouping of users based on their behavior and demographics.
* **Elbow Method**: Provided a systematic approach to selecting the optimal number of clusters, ensuring that we balance between underfitting and overfitting the data.
* **Feature Scaling**: Necessary for ensuring that the distance calculations in K-means are not dominated by features with larger scales.

By employing this methodology, we aimed to uncover meaningful segments within the user base, which would allow for more targeted and effective marketing strategies, ultimately enhancing user engagement and conversion rates.

**Data Understanding:**

* **Data Collection:** Data was provided from Udacity – Starbucks.

The data is contained in three files:

1. portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
2. profile.json - demographic data for each customer
3. transcript.json - records for transactions, offers received, offers viewed, and offers completed

* **Data Exploration:** Explore the datasets to understand its structure, size, and variables. Identify potential challenges such as missing values, outliers, or skewed distributions.

**What is the features of the data we have and how they are significance to address our objective?**

**Features in the Dataset**

1. **person:** Unique identifier for each user.

**Significance:** Helps track individual user behavior across different events and offers.

1. **time:** Time at which the event occurred.

**Significance:** Useful for understanding the temporal patterns of user interactions with offers.

1. **value:** Type of event (e.g., offer received, offer viewed).

**Significance:** Helps in categorizing different types of user interactions.

1. **amount:** Amount spent in a transaction.

**Significance:** Directly related to the user’s spending behavior, critical for assessing offer effectiveness.

1. **age: Age of the user.**

**Significance:** Important for segmenting users based on age demographics, which can influence their responsiveness to different offers.

1. **became\_member\_on:** Date when the user became a member.

**Significance**: Helps calculate the membership duration, indicating user loyalty.

1. **income:** Annual income of the user.

**Significance:** Critical for understanding the purchasing power and spending behavior of users.

1. **difficulty:** Difficulty level of redeeming an offer.

**Significance:** Can influence user engagement with different types of offers.

1. **duration:** Duration for which the offer is valid.

**Significance**: Important for analyzing how the length of offer availability affects user interaction.

1. **reward:** Reward received from the offer.

**Significance:** Incentive for the user to engage with the offer, crucial for understanding offer attractiveness.

1. **event\_offer completed, event\_offer received, event\_offer viewed, event\_transaction:** Binary indicators of different user events.

**Significance:** These features track specific user actions, providing insight into user engagement and behavior patterns.

1. **offer\_type\_bogo, offer\_type\_discount, offer\_type\_informational:** Types of offers.

**Significance:** Helps differentiate user preferences for various types of promotional offers.

1. **web, email, mobile, social:** Channels through which offers are communicated.

**Significance:** Important for understanding which communication channels are most effective.

1. **gender\_encoded, gender\_M, gender\_O:** Encoded gender information.

**Significance:** Helps in segmenting users based on gender, which can influence offer preferences.

1. **membership\_duration:** Calculated duration of user membership.

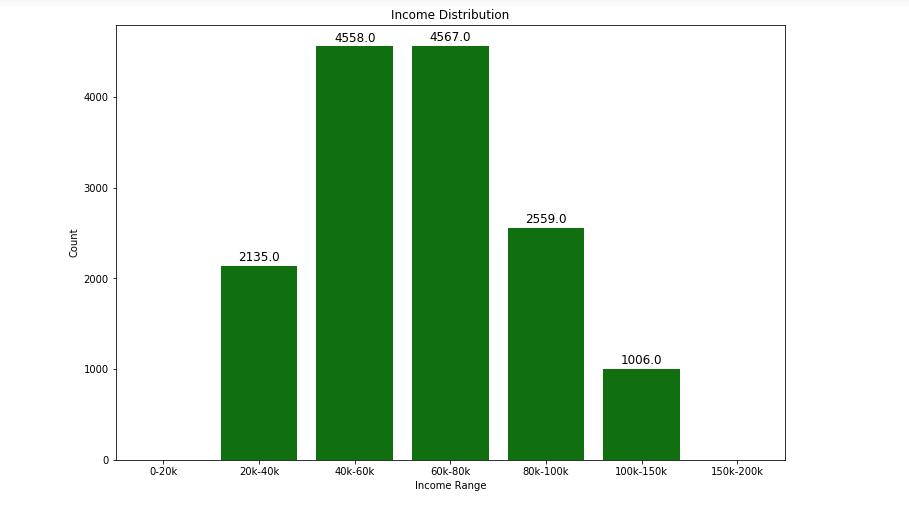
**Significance:** Indicates user loyalty and engagement over time.

1. **income\_bin\_20k-40k, income\_bin\_40k-60k, income\_bin\_60k-80k, income\_bin\_80k-100k, income\_bin\_100k-150k, income\_bin\_150k-200k:** Binned income ranges.

**Significance**: Useful for segmenting users into income brackets, aiding in targeted marketing.

Interesting insights were discovered but what really worth mentioning here is income ranges for Starbucks users *(based on the provided dataset)*.

Down below is a chart shows the income distribution for Starbucks Customers:



From first look, we can see that most customers’ income range falls between

(40K – 80k); this information would be very helpful to tailor marketing campaigns.

**Statistical Analysis of Features:**

To better understand the significance and distribution of these features, we conducted statistical analyses and visualized key attributes:

1. **Income Distribution:**
   * **Objective**: To identify potential income class imbalances.
   * **Visualization**: Bar plot showing the count of users in different income bins.
2. **Age Distribution:**

* **Objective:** To analyze the age demographics of users.
* **Visualization:** Histogram of user ages

1. **Gender Distribution:**
   * **Objective**: To identify gender representation in the dataset.
   * **Visualization**: Bar plot of gender counts.

**4. Event Type Distribution:**

* **Objective:** Visualize the distribution of different event types in the dataset.
* **Visualization:** Bar plot of event types.

**Data Examination for Abnormalities:**

**1. Missing Values:** To check if there are any missing values in the dataset, as they can significantly impact the analysis.

**2. Distribution of Categorical Variables:** Check for any imbalances or unusual distributions in categorical variables like gender and offer types.

* **Data Description**: Document the features in the dataset, including their data types, descriptions, main characteristics, and potential relationships.
* **Data Quality Assessment:** Assess the quality of the data by checking for inconsistencies, errors, or anomalies. Ensure that the data is clean, reliable, and suitable for analysis.

**Data Pre-processing:**

Pre-process the data by handling missing values, encoding categorical variables, and scaling numerical features. Prepare the dataset for modeling by transforming it into a format suitable for the used machine learning algorithms.

To summarize the steps you've taken:

1. Filled the missing values in the 'channels', 'duration', 'amount', and 'reward' columns using the previously defined strategies.
2. Handled the remaining missing values in the 'age', 'became\_member\_on', 'gender', and 'income\_bin' columns by filling them with the mean (for 'age'), mode (for 'became\_member\_on' and 'gender'), and median (for 'income').
3. Dropped the 'income' column since you already have the 'income\_bin' column.

**Analysis:**

* **Feature Engineering**: Create new features or transform existing ones to extract meaningful information from the data. For example, binning income into categories or calculating membership duration.
* **Data Visualization:** Explore the distribution of variables, detect patterns, and identify relationships between variables using visualizations such as bar plots, and correlation matrices.
* **Statistical Analysis:** Conduct statistical tests to understand the significance of relationships and identify potential variables for modeling.

**Modeling:**

* **Model Selection**: Choose appropriate machine learning models based on the problem type, data characteristics, and business objectives. Consider models like logistic regression, random forests, and clustering algorithms like K-means.

Based on the cluster results, here's a summary of each cluster:

* **Cluster 0:**
  + This cluster has below-average values for 'time', 'amount', 'age', and 'income'.
  + The 'difficulty' and 'reward' for offers are slightly above average.
  + Most offer events (completed, received, viewed) are well above average.
  + This cluster has a high proportion of mobile and social offers.
  + Gender distribution and membership duration are relatively balanced.
  + Income distribution is mostly concentrated in the lower income bins.
* **Cluster 1:**
  + This cluster has above-average values for 'time', 'amount', and 'income', but slightly below-average for 'age'.
  + The 'difficulty' and 'reward' for offers are close to the average.
  + Offer events (completed, received, viewed) are slightly below average.
  + There is a lower proportion of mobile and social offers compared to other clusters.
  + Gender distribution and membership duration vary.
  + Income distribution is spread across different income bins.
* **Cluster 2:**
  + This cluster has above-average values for 'time', 'age', and 'income', while 'amount' is slightly below average.
  + The 'difficulty' and 'reward' for offers are close to zero.
  + Offer events (completed, received, viewed) are very high compared to other clusters.
  + There is a relatively low proportion of mobile and social offers.
  + Gender distribution is slightly skewed towards males.
  + Membership duration varies, and income distribution is spread across different income bins.
* **Cluster 3:**
  + This cluster has below-average values for 'time', 'amount', 'age', and 'income'.
  + The 'difficulty' for offers is slightly above average, while 'reward' is below average.
  + Offer events (completed, received, viewed) are below average.
  + There is a high proportion of mobile and social offers compared to other clusters.
  + Gender distribution is relatively balanced.
  + Membership duration varies, and income distribution is spread across different income bins.

This summary provides a high-level overview of the characteristics of each cluster based on their respective cluster centers.

**Model Tuning:** In the model tuning phase, we focused on optimizing the parameters of the K-means clustering algorithm to improve the quality of the clustering solution. Here's a summary of what we did:

Determined the Optimal Number of Clusters: We used the elbow method to identify the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters. The "elbow" point on the plot, where the rate of decrease in WCSS slows down, indicated the optimal number of clusters.

Fit the K-means Model: After determining the optimal number of clusters, we instantiated the K-means algorithm with the chosen number of clusters and fit it to the scaled feature data.

Analyzed Clustering Results: We examined the clustering results by inspecting the distribution of data points among clusters and the characteristics of each cluster, such as centroid values and cluster sizes.

Visualized Clusters: To gain further insights into the clustering solution, we visualized the clusters by plotting the data points with cluster labels and centroids in a scatter plot.

Evaluated Clustering Performance: Instead of traditional evaluation metrics like silhouette score or inertia, which are not always suitable for K-means clustering, we relied on visual inspection and domain knowledge to assess the quality and interpretability of the clustering solution.

Overall, the model tuning phase aimed to refine the clustering solution by optimizing the number of clusters and interpreting the results in the context of the business problem.

**Model Evaluation:** In the model evaluation phase, our primary objective was to assess the quality and effectiveness of the clustering solution obtained from the K-means algorithm. Here's a summary of what we did:

* + Visual Inspection: We visually inspected the clustering results by plotting the data points with cluster labels and centroids in scatter plots. This allowed us to observe the distribution of data points among clusters and understand the characteristics of each cluster.
  + Interpretation of Clusters: Based on the visual inspection, we interpreted the clusters in the context of the business problem and domain knowledge. We analyzed the centroid values and the distribution of features within each cluster to understand the distinct patterns and behaviors of users.
  + Comparison with Business Objectives: We compared the clustering solution with the objectives of the direct marketing campaign optimization project. We assessed whether the clusters aligned with the desired segmentation goals and whether they provided actionable insights for tailoring marketing messages.
  + Assessment of Cluster Characteristics: We evaluated the cluster characteristics, such as cluster sizes, centroid values, and feature distributions, to determine if they were meaningful and interpretable in the context of the business problem.
  + Feedback and Iteration: Based on the evaluation results, we provided feedback on the clustering solution and identified areas for potential improvement. If necessary, we iterated on the modeling process by adjusting parameters or features to enhance the clustering solution.

Overall, the model evaluation phase aimed to validate the clustering solution and its relevance to the business problem, ensuring that it provided actionable insights for optimizing direct marketing campaigns.

Based on the evaluation executed, one of the main takeaway is when clustering features are changed to ('event\_offer viewed', 'amount'), we noted different behaviours of the 4 clusters.

The descriptive statistics for each cluster, here's an assessment of the clustering result:

* **Time, Amount, and Age**: The mean values for these attributes show some variation across clusters, indicating that the clusters may represent different groups of users in terms of their transaction behavior, age, and possibly time spent on the platform.
* **Income:** There's noticeable variation in income across clusters, suggesting that the clusters may capture different income groups among users.
* **Offer Difficulty and Duration:** It appears that clusters 2 and 3 have very low values for offer difficulty and duration, which could indicate a distinct group of users who tend to interact with offers that are easier to complete and have shorter durations.
* **Gender Distribution:** The gender distribution seems relatively consistent across clusters, with minor variations.
* **Offer Interaction:** Cluster 0 has the highest mean value for offer interaction, suggesting that users in this cluster are more engaged with offers compared to users in other clusters.

In summary, the clustering result provides valuable insights into user segmentation.

***What are the interesting Findings?***

1. **High Offer Views, No Spending**:
   * **Cluster Characteristics**: This cluster represents users who frequently view offers (high **event\_offer viewed** value around 2.5) but do not spend any money (**amount spent** value is 0).
   * **Centroid Position**: The centroid of this cluster is directly above these data points, indicating that the average characteristics of users in this cluster align with viewing offers frequently but not making any purchases.

**Insights and Actions**

1. **Engagement Without Conversion**:
   * These users are engaged with the promotional offers since they view them frequently. However, this engagement does not translate into actual spending.
   * Possible reasons could include:
     + Offers are not compelling enough to make a purchase.
     + Users might be looking for better deals.
     + Users could be browsing but not in a position to make a purchase (e.g., financial constraints).
2. **Tailoring Marketing Strategies**:
   * **Improve Offer Appeal**: Enhance the attractiveness of the offers. This could include increasing discounts, adding more value, or personalizing the offers based on user preferences.
   * **Follow-Up Communication**: Use follow-up communication strategies, such as reminder emails or notifications, to encourage these users to complete a purchase.
   * **Understanding Barriers**: Conduct surveys or analyze additional data to understand why these users are not converting. Are there common barriers to purchasing that can be addressed?
3. **Segment-Specific Offers**:
   * **Incentivize First Purchase**: Provide special incentives for first-time purchases. For example, offer a small reward or an additional discount for users making their first transaction.
   * **Targeted Campaigns**: Run targeted campaigns specifically designed for high-engagement, low-conversion users. This can include limited-time offers or personalized discounts.

***How can we improve the implementation?***

1. **Memory Management**: Monitor and manage memory usage throughout the implementation, especially when working with large datasets or computationally intensive tasks. This may involve periodically clearing variables, using memory-efficient data structures, or increasing the available memory if feasible.
2. **Code Profiling**: Use code profiling tools to identify bottlenecks and performance hotspots in the implementation. This can help pinpoint areas for optimization and prioritize improvements based on their impact on overall performance.
3. **Model Architecture**: Choose appropriate model architectures that are well-suited for the problem at hand. For example, in the context of clustering, consider different algorithms such as K-means, hierarchical clustering, or density-based clustering based on the characteristics of the data and the desired outcomes. Experimenting with different architectures can help identify the most effective approach.
4. Cross-Validation: Implement cross-validation techniques such as k-fold cross-validation or stratified cross-validation to assess the model's performance more reliably. Cross-validation helps estimate the model's generalization error and identify potential sources of bias or variance.

**Conclusion**

A cluster with high offer views and zero spending indicates a significant interest in offers without actual purchase activity. This segment presents an opportunity to convert interest into sales through strategic marketing efforts, improved offers, and targeted engagement tactics. By understanding and addressing the needs and barriers of this user segment, you can potentially increase conversion rates and overall campaign effectiveness.