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.

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.

- **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.

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.

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.