Customer Personality Analysis

Problem Statement

Customer Personality Analysis is a detailed analysis of a company’s ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customers.

Business Objective

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company’s database, a company can analyse which customer segment is most likely to buy the product and then market the product only on that particular segment.

Dataset Details

ATTRIBUTES

People

- ID: Customers unique identifier
- Year_Birth: Customers birth year
- Education: Customers education level
- Marital_Status: Customers marital status
- Income: Customers yearly household income
- Kidhome: Number of children in customers household
- Teenhome: Number of teenagers in customers household
- Dt_Customer: Date of customers enrollment with the company
- Recency: Number of days since customers last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Product

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Dataset Details

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company’s website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores

- □ NumWebVisitsMonth: Number of visits to company’s website in the last month

Data Preparation Process

Handling missing values

Missing values in the 'Income' column were handled by removing corresponding rows. This approach ensures data quality and prevents potential bias in subsequent analyses by using complete data only

Feature Engineering

Feature engineering involves creating new features from existing ones to enhance analysis. In the provided code, new features like 'TotalAmountSpent' and 'Year_Customer' were generated, providing additional insights for analysis and modeling tasks

Encoding categorical variables

Encoding categorical variables involves converting them into numerical format. In the provided code, LabelEncoder from sklearn.preprocessing was used to transform categorical variables like 'Education' and 'Marital_Status', making them suitable for analysis and modeling

Exploratory Data Analysis (EDA)

- Exploratory Data Analysis is a crucial phase where we gain insights into our dataset. In this section, we explored various aspects of our data.
- Summary Statistics: Calculated summary statistics for numerical attributes like 'Year_Birth', 'Income', and 'Recency’.
- Frequency Counts: Examined the frequency of categories in categorical attributes such as 'Education' and 'Marital_Status'

Exploratory Data Analysis (EDA)

EDA with Visualizations

Utilized visualizations to understand distributions, relationships, and correlations.

Boxplot to identify outliers in numerical attributes.

Pie chart to visualize spending on different product categories.

Bar plot to compare purchases through different channels.

Histogram to understand the distribution of customer responses.

Pair plot to explore relationships between numerical attributes.

Heatmaps to visualize correlations between accepting offers in different campaigns.

Insights: Derived insights from visualizations to understand customer behaviour, preferences, and potential patterns within the data.

Through EDA, we gained a comprehensive understanding of our dataset, laying the groundwork for further analysis and modelling.

Automated Machine Learning (Auto ML)

1. ## Pandas

   Used for data manipulation and analysis, including reading the dataset from an Excel file, handling missing values, and creating new features..

2. ## Matplotlib & Seaborn

   Utilized for data visualization, including creating box plots, pie charts, bar plots, histograms, scatter plots, pairplots, and heatmaps to explore various aspects of the dataset.

3. ## Sklearn

   Specifically, the LabelEncoder from the preprocessing module was used to encode categorical variables into numerical format.

Automated Machine Learning (Auto ML)

4 Standard Scaler

Used from the Sklearn preprocessing module for standardizing numerical features.

5 PCA

Employed from the Sklearn decomposition module for dimensionality reduction.

Manual Machine Learning Techniques

K-means Clustering

K-means clustering is a manual machine learning technique used for unsupervised clustering.

It involves partitioning the data into 'k' clusters based on the similarity of data points to cluster centroids.

The algorithm iteratively assigns data points to the nearest cluster centroid and updates centroids until convergence.
K-means clustering is used to identify natural groupings or clusters within the data.

Therefore, K-means clustering is the manual machine learning technique applied in the provided code snippet for clustering the data into distinct groups.

Model Evaluation & Performance

K-means Clustering

While K-means clustering is primarily a clustering algorithm, it can also be evaluated based on certain metrics such as within-cluster sum of squares (WCSS)

Elbow method

The elbow method is used to determine the optimal number of clusters for K-means by plotting the WCSS values against different numbers of clusters and identifying the point where the rate of decrease in WCSS slows down, indicating the optimal number of clusters.

Visual Inspection

The resulting clusters from K-means clustering are visually inspected using scatter plots to evaluate how well data points are grouped into clusters.

Deployment: User Interactive Web App

1

Pickle for Model Persistence

Utilized pickle for model serialization, ensuring the trained model's persistent storage and retrieval for seamless deployment.

2

Flask Web Application

Leveraged Flask to develop a user interactive web application, enabling users to predict the Yearly Amount Spent based on the deployed machine learning model.

Comprehensive Statistical Analysis Overview

Data Preparation & EDA

The initial phases of the project involved meticulous data preparation, validation, and detailed exploratory data analysis, laying the groundwork for the subsequent modeling and deployment stages.

Modeling & Evaluation

Implemented a wide array of machine learning techniques, both automated and manual, followed by rigorous evaluation to select the most optimal models for deployment.

Interactive Web App Deployment

The final phase included the deployment of an interactive web application using Flask, providing users with a seamless experience for predicting the target variable.

Final Insights and Conclusions

In summary, our analysis revealed crucial predictors affecting yearly customer spending, emphasizing the significance of session length, time on app, time on website, and membership duration. The developed multiple linear regression models demonstrated satisfactory accuracy, as measured by MSE and R-squared, offering valuable predictive insights for decision-making.

The findings underscore the importance of enhancing user engagement, both on the app and website, as a strategic approach to positively influence customer spending. Notably, a positive correlation was identified between membership duration and spending, suggesting the potential benefits of encouraging longer memberships to foster loyalty.

In conclusion, this project contributes to a deeper understanding of customer behavior for the Ecommerce company, providing actionable insights. Future recommendations include refining models, incorporating additional features, and exploring advanced techniques to further enhance predictive capabilities.

The practical application of deploying predictive models through Flask offers real-time insights, creating a seamless integration of data science findings into operational decision-making processes. This holistic approach emphasizes the bridge between data-driven insights and practical business optimization.