GRT INSTITUTE OF ENGINEERING AND TECHNOLOGY, TIRUTTANI - 631209

Approved by AICTE, New Delhi Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

PHASE 2

PROJECT TITLE

Customer Segmentation using Data Science

COLLEGE CODE : 1103

DIWAKAR M G

3rd yr, 5th sem

Reg no. : 110321104007

diwakarmg03@gmail.com

2.1 SHORT EXPLAINATION ABOUT CUSTOMER SEGMENTATION USING DATA SCIENCE

Customer segmentation using data science is the process of dividing a company's customer base into distinct groups or segments based on common characteristics or behaviors, with the goal of tailoring marketing, product development, and customer service strategies to better meet the needs of each segment. Here's a brief explanation of how it works:

1. Data Collection: Gather relevant data about your customers, which can include demographic information (age, gender, location), psychographic data (lifestyle, interests), transaction history, website behavior, and more.

2. Data Preprocessing: Clean and prepare the data by handling missing values, outliers, and ensuring data consistency. This step is essential for accurate analysis.

3. Segmentation Techniques: Apply data science techniques such as clustering, classification, or dimensionality reduction to group customers with similar characteristics or behaviors together. Common methods include k-means clustering, hierarchical clustering, and decision tree algorithms.

4. Feature Selection: Identify the most important features (variables) that contribute to the segmentation. This helps in focusing on relevant aspects of customer behavior.

5. Model Validation: Evaluate the quality of your segmentation model using metrics like silhouette score, Davies-Bouldin index, or domain-specific criteria to ensure that the segments are meaningful and distinct.

6. Segment Profiling: Once segments are defined, create detailed profiles for each group. Understand their needs, preferences, and pain points to develop tailored marketing strategies and product offerings.

7. Targeted Marketing: Design and implement personalized marketing campaigns and strategies for each segment. This can involve customizing product recommendations, messaging, and advertising channels.

8. Monitoring and Iteration: Continuously monitor customer behavior and segment performance. Adjust your strategies as needed based on changing trends and feedback to ensure the segments remain relevant.

Customer segmentation using data science can lead to more effective marketing, increased customer satisfaction, higher retention rates, and improved overall business performance. It allows companies to treat different customer groups in a way that resonates with their unique preferences and characteristics, ultimately driving better results and stronger customer relationships.

2.2 WHERE I GOT THE DATASETS AND ITS DETAILS

You can find datasets for customer segmentation and various other data science projects from several reputable sources.

KAGGLE : Kaggle is a popular platform for data science competitions and dataset sharing. It hosts a wide range of datasets on various topics, including customer data. You can browse datasets, read their descriptions, and download them for free. Kaggle also provides a community where you can discuss and collaborate on data science projects.

Website : *https://www.kaggle.com/datasets/akram24/mall-customers*

NAME OF THE DATASET : Mall Customers

DATA DESCRIPTION :

Customer segmentation is a common application of data science in the retail industry, including malls. To perform customer segmentation effectively, you need relevant data about mall customers. Once you have collected and cleaned the relevant data, you can apply various data science techniques such as clustering, classification, and regression to segment mall customers effectively. The goal is to identify groups of customers with similar characteristics and preferences to tailor marketing strategies, promotions, and store layouts to meet their needs and maximize the mall's revenue.

2.3 DETAILS ABOUT COLUMNS

CustomerID – in this column fill the ID details belongs to the customer that was given

Gender – Mention the gender of the customer

Age – Mention the age of the customer

Annual Income (k$) – Mentioning the customer annual income for a count purpose and calculating the spending score.

Spending Score (1-100) – In this by using the annual income column the spending score is calculated for the customer in that specified mall

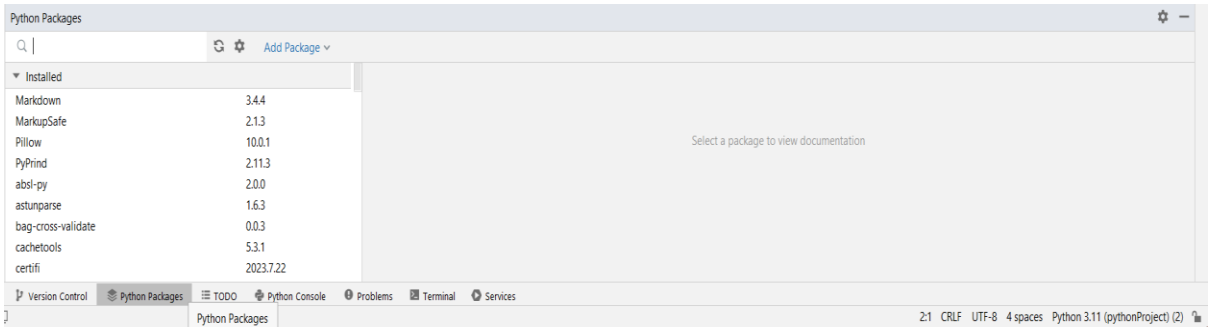
2.4 DETAILS OF LIBRARIES TO BE USED AND WAY TO DOWNLOAD

LIBRARIES TO BE USED

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- import seaborn as sns
- from sklearn.cluster import KMeans

WAY TO DOWNLOAD THE LIBRARIES

1. Click the python packages in the bottom of your project in pycharm



2. Type the required library in the search box and click install package in the right end top of the python packages.

3. After installation process finished it shows the package was installed in the python packages.

2.5 HOW TO TRAIN AND TEST THE DATASET

To train and test a machine learning model using a dataset of mall customers with the given column names (CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100)), you can follow these steps:

Data Preprocessing :

Load your dataset into a data analysis or machine learning environment (e.g., Python with libraries like pandas and scikit-learn).

Explore and clean the data to handle any missing values, duplicates, or outliers.

Encode categorical variables like "Gender" into numerical values (e.g., 0 for Male, 1 for Female) if needed.

Splitting the Data :

Divide your dataset into two parts: a training set and a testing set. A common split is 80% for training and 20% for testing, but you can adjust this ratio as needed.

Ensure that the split maintains a representative distribution of data, especially if you have imbalanced classes or segments.

Selecting a Machine Learning Model :

Choose an appropriate machine learning model for your task. Since you want to segment customers, unsupervised learning techniques like clustering (e.g., K-Means, hierarchical clustering) are often used.

Feature Selection :

In this case, you can use features like "Age," "Annual Income," and "Spending Score" for customer segmentation. Exclude "CustomerID" and "Gender" if they do not contribute to the segmentation.

Training the Model :

Fit your chosen machine learning model to the training data using the selected features.

For example, if you're using K-Means clustering in Python with scikit-learn, you can use the following code:

```
kmeansmodel = KMeans(n_clusters = 5 , init = "k-means++",random_state=0)
y_kmeans = kmeansmodel.fit_predict(x)
```

Model Evaluation :

Since clustering is unsupervised, traditional evaluation metrics like accuracy do not apply. Instead, you can use within-cluster sum of squares (WCSS) or silhouette score to assess the quality of the clusters.

Testing the Model :

Use the trained model to predict the clusters for the samples in your testing dataset.

```
plt.scatter(x[y_kmeans==0,0],x[y_kmeans==0,1],s=100,c="red",label = "cluster 1")
plt.scatter(x[y_kmeans==1,0],x[y_kmeans==1,1],s=100,c="blue",label = "cluster 2")
plt.scatter(x[y_kmeans==2,0],x[y_kmeans==2,1],s=100,c="green",label = "cluster 3")
plt.scatter(x[y_kmeans==3,0],x[y_kmeans==3,1],s=100,c="cyan",label = "cluster 4")
plt.scatter(x[y_kmeans==4,0],x[y_kmeans==4,1],s=100,c="magenta",label = "cluster 5")
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s=300,c="yellow",label="centroids")
plt.title("clusters of customers")
plt.xlabel("Yıllık gelir")
plt.ylabel=("harcama skoru")
plt.legend()
plt.show()
```

Analysing the Results :

Examine the clustering results to understand how customers are grouped based on their features.

Visualize the clusters, if possible, to gain insights.

Utilizing Segmentation :

Once you have clustered customers, you can tailor marketing strategies, promotions, and store layouts for each segment based on their characteristics.

2.6 REST OF EXPLAINATION

Certainly! Continuing from where we left off, here is the rest of the explanation on how to work with a mall customer dataset for customer segmentation using data science:

Validation and Fine-Tuning : you may want to fine-tune your model by adjusting hyperparameters or exploring different clustering algorithms. Cross-validation can help you assess the robustness of your model and choose the best configuration.

Interpreting the Segmentation : Interpret the meaning of each customer segment. What are the distinguishing characteristics of each group? For example, do you have a high-income, high-spending segment and a low-income, low-spending segment? Use visualizations like scatter plots or bar charts to illustrate the differences between segments.

Targeted Marketing and Strategy : Develop tailored marketing strategies for each customer segment. For example, create promotions or advertisements that resonate with the unique

preferences and behaviours of each group. Optimize store layouts, product placements, and inventory based on the identified segments.

Monitoring and Feedback : Continuously monitor the effectiveness of your strategies and promotions for each segment. Collect feedback from customers in each segment and use it to make data-driven improvements.

Retraining the Model : Over time, as new data becomes available, consider retraining your customer segmentation model. Customer preferences and behaviours can change, and your model should adapt accordingly.

Integration with Customer Relationship Management (CRM) : Integrate the segmentation results with your CRM system to ensure that customer interactions and communications are personalized and consistent with the identified segments.

Privacy and Compliance : Ensure that you handle customer data with care and in compliance with relevant privacy regulations (e.g., GDPR, CCPA). Anonymize or pseudonymize customer data as needed to protect privacy.

A/B Testing : Implement A/B testing for marketing campaigns to measure the impact of changes on different customer segments accurately.

Documentation and Reporting : Document your data pre-processing steps, model selection, and results thoroughly. This documentation is essential for future reference and model maintenance.

Scaling and Scalability : Consider how your customer segmentation process can scale as the dataset and business grow. Ensure that your infrastructure and tools can handle larger volumes of data.

2.7 WHAT METRICS USED FOR THE ACCURACY CHECK

When performing customer segmentation using data science, traditional accuracy metrics like classification accuracy are not applicable because customer segmentation is an unsupervised learning task. In unsupervised learning, there are no ground truth labels to compare predictions against. Instead, you use different metrics to evaluate the quality of the segmentation. Here are some commonly used metrics for assessing the accuracy of customer segmentation:

Silhouette Score : The silhouette score measures how similar each data point in one cluster is to the data points in the same cluster compared to the nearest neighboring cluster. A higher silhouette score indicates better-defined clusters.

Davies-Bouldin Index : This index measures the average similarity between each cluster and its most similar cluster. Lower values indicate better clustering, with a lower Davies-Bouldin Index representing more distinct clusters.