PROJECT:CUSTOMER SEGMENTATION USING DATA SCIENCE

PHASE-5

PROBLEM STATEMENT

In an increasingly data-driven business landscape, companies are looking to leverage data science to gain a competitive edge and enhance customer relationships. The problem at hand is to develop an effective customer segmentation strategy that allows a company to better understand its diverse customer base and tailor its marketing and operational approaches to improve customer satisfaction and business performance.

PROBLEM DEFINITION

The problem is to implement data science techniques to segment customers based on their behaviour, preferences, and demographic attributes.

Customer segmentation is the practice of dividing a customer base into group of individuals that have similar characteristics relevant to marketing, such as age, gender, interests and spending habits.

The goal of customer segmentation is to reach out to customers more effectively, thereby leading to more sales or customer conversions. Companies also hope to gain a deeper understanding of their customers' preferences and needs by discovering what each segment finds most valuable and more accurately tailoring marketing materials toward that segment.

The goal is to enable businesses to personalize marketing strategies and enhance customer satisfaction. This project involves data collection, data pre processing, feature engineering, clustering algorithms, visualization, and interpretation of results.

DESIGN THINKING

1. **Data Collection:** Collect customer data, including attributes like purchase history, demographic information, and interaction behaviour. Gather from various sources of data to do the customer segments such as surveys, interviews, feedback forms, web analytics, CRM reports, social media, and email campaigns ,kaggle.
2. **Data Preprocessing:** Clean and preprocess the data, handle missing values, and convert categorical features into numerical representations.

   The raw data we downloaded is complex and in a format that cannot be easily ingested by customer segmentation models. We need to do some preliminary data preparation to make this data interpretable.

   The informative features in this dataset that tell us about customer buying behavior include “Quantity”, “InvoiceDate” and “UnitPrice.” Using these variables, we are going to derive a customer’s RFM profile - Recency, Frequency, Monetary Value.

   RFM is commonly used in marketing to evaluate a client’s value based on their:

   Recency: How recently have they made a purchase?

   Frequency: How often have they bought something?

   Monetary Value: How much money do they spend on average when making purchases?

3. **Feature Engineering:** Create additional features that capture customer behavior Preferences, such as total spending, frequency of purchases.

   Selecting the most important existing features.

Creating new features from existing features.

Aggregating features by customer.

4. **Clustering Algorithms:** Apply clustering algorithms like K-Means,DBSCAN,GMM algorithm ,Mean shift or hierarchical clustering to segment customers.

5. **Visualization:** Visualize the customer segments using techniques like scatter plots, bar charts, graphs ,histograms, pie charts and heatmaps.

6. **Interpretation:** Analyze and interpret the characteristics of each customer segment to derive actionable insights for marketing strategies.Customer segmentation is the process of dividing customers into groups based on common characteristics, such as demographics or behaviours. The goal is to help marketing and sales teams reach out to customers more effectively.

OBJECTIVES OF THE PROJECT

The objectives of customer segmentation using data science include:

1. Reaching the right people.
2. Targeting different segments with different products.
3. Introducing new products that meet the needs of customers.
4. Identifying new segments to target.
5. Making better marketing strategies.
6. Offering a specific group with more customized products or services.
7. Understanding customers' preferences.
8. Presenting customers with better-targeted advertisements.
9. Targeting customers with the highest potential value first.
10. Identifying the most active users/customers.
11. Optimizing your application/offer towards their needs.
12. Maximizing the value of each customer to the business.

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. It involves implementing data science methods to divide the customer base into smaller groups based on certain characteristics.

INNOVATION OF THE PROJECT

Innovation

Consider incorporating dimensionality reduction techniques like PCA or t-SNE to visualize high-dimensional customer data and discover underlying patterns.

Short explanation for the project

Customer segmentation using data science is a process of dividing a company's customer base into distinct groups or segments based on common characteristics and behaviours. This approach leverages data analysis, machine learning, and statistical techniques to identify meaningful patterns and insights within customer data. The goal is to better understand customers, tailor marketing efforts, and optimize business strategies. By segmenting customers, businesses can personalize their

interactions, products, and services, ultimately improving customer satisfaction and driving growth. Data science helps identify hidden trends and correlations in large datasets, making it a valuable tool for businesses seeking to enhance their understanding of their diverse customer base.

Dataset and its detail

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

A dataset for customer segmentation is a collection of data about customers that can be used to divide them into smaller groups based on their shared characteristics or behaviours. This data can include demographic information (e.g., age, gender, location, income), purchase history and other factors.

Here we used this dataset

- Mall Customer Segmentation Data (Kaggle): This dataset contains information about 20,000 customers of a shopping mall, including their age, gender, annual income, spending score, and purchase history.

Demographic variables:
Age: Segmenting customers by age can help target products and marketing messages to specificage groups.
For example, products for children, teenagers, adults, or seniors may differ significantly.
Gender: Gender-based segmentation can be useful for businesses offering gender-specific products.
Income: Income-based segmentation can help target products and services to customers with different spending power.
Customer id: In customer segmentation, it's common to include a unique customer identifier, often referred to as "Customer ID" or "Customer Number," in the dataset.
Spending score: In customer segmentation, the "spending score" is often an essential feature used for clustering customers based on their spending behaviour

Details of libraries for innovation

1)Numpy-NumPy is a Python library for numerical and mathematical operations.Go to command prompt and give “pip install numpy”

2)Pandas-Pandas is a Python library used for working with data sets.Go to command prompt and give “pip install pandas”

3)Matplotlib-Matplotlib is a Python library for creating static, animated, and interactive visualizations in various formats, such as line charts, scatter plots, bar charts, histograms, and more.Go to command prompt and give “pip install matplotlib”

4)Sklearn- It provides a wide range of tools and algorithms for tasks such as classification, regression, clustering, dimensionality reduction, and more.Go to command prompt and give “pip install sklearn”

Algorithm explanation which we used in innovation

1.PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in data analysis and machine learning. It helps simplify the complexity in high-dimensional data while retaining trends and patterns. PCA transforms the original variables into a new set of orthogonal variables called principal components, which are linear combinations of the original variables.

2.t-SNE

T-distributed Stochastic Neighbor Embedding (t-SNE) is a dimensionality reduction technique used in machine learning and data visualization. It is particularly useful for visualizing high-dimensional data in a lower-dimensional space while preserving the pairwise similarities between data points. t-SNE is commonly employed for exploratory data analysis, clustering, and visualization tasks.

Code:

```
 import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler, OneHotEncoder

from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline

from sklearn.decomposition import PCA

from sklearn.manifold import TSNE

data = pd.read_csv('D:\project\Mall_Customers.csv')

numerical_features = [ 'Age', 'Annual Income (k$)','Spending Score (1-100)']

categorical_features = ['Genre']

numerical_transformer = Pipeline(steps=[('scaler', StandardScaler())])

categorical_transformer = Pipeline(steps=[('onehot', OneHotEncoder(drop='first')) ])

preprocessor = ColumnTransformer(transformers=[('num', numerical_transformer, numerical_features),('cat',
categorical_transformer, categorical_features)   ])

X = data[numerical_features + categorical_features]

X_processed = preprocessor.fit_transform(X)

pca = PCA(n_components=2, random_state=42)

X_pca = pca.fit_transform(X_processed)

tsne = TSNE(n_components=2, perplexity=30, random_state=42)

X_tsne = tsne.fit_transform(X_processed)

plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)

plt.scatter(X_pca[:, 0], X_pca[:, 1], alpha=0.5)

plt.title('PCA Visualization')

plt.xlabel('PCA Dimension 1')

plt.ylabel('PCA Dimension 2')

plt.subplot(1, 2, 2)

plt.scatter(X_tsne[:, 0], X_tsne[:, 1], alpha=0.5)

plt.title('t-SNE Visualization')
```

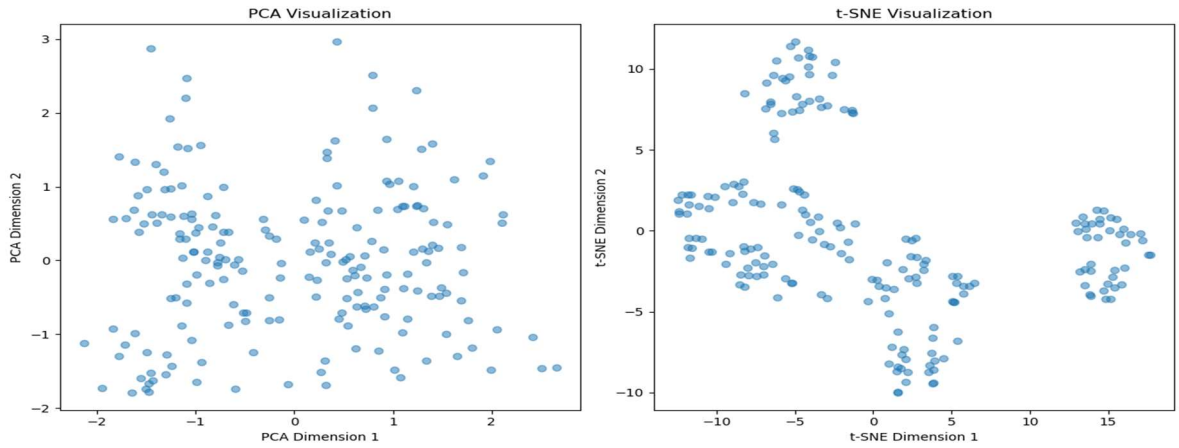
plt.xlabel('t-SNE Dimension 1')

plt.ylabel('t-SNE Dimension 2')

plt.tight_layout()

plt.show()

Output:



DEVELOPMENT FOR THE PROJECT

1)Load the dataset

Loading a dataset for customer segmentation involves reading the data from an external source, such as a CSV file, a database, or an API, into your program or environment for further analysis and segmentation.

CSV File (Pandas in Python): If your dataset is in a CSV file, you can use the pd.read_csv() function from the Pandas library in Python to load the data into a DataFrame.

import pandas as pd data = pd.read_csv('your_dataset.csv')

2)Preprocess the dataset

Here are some common preprocessing steps:

Data Exploration:Examine the dataset by checking its structure, the first few rows, data types, and summary statistics.

Handling missing values: You can use methods like data.fillna() to fill in missing data or data.dropna() to remove rows with missing values.

Encoding categorical features: If your dataset contains categorical variables, you can use one-hot encoding or label encoding to convert them into numeric values.

Standardizing numerical features: Standardization ensures that all numerical features have a mean of 0 and a standard deviation of 1.

Outlier Detection:Identify and handle outliers in your data.

Code for loading and pre processing the dataset:

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
data = pd.read_csv("D:\IBM project\Mall_Customers.csv")
print(data.head())
print(data.info())
data.fillna(data.mean(), inplace=True)
data = pd.get_dummies(data, columns=['Gender'], drop_first=True)
numerical_features = ['Age', 'Annual_Income_(k$)',’Spending_Score’]
scaler = StandardScaler()
data[numerical_features] = scaler.fit_transform(data[numerical_features])
X = data[['Age', 'Annual_Income_(k$)',’Spending_Score’]]
from sklearn.ensemble import IsolationForest
outlier_detector = IsolationForest(contamination=0.05)
data['IsOutlier'] = outlier_detector.fit_predict(X)
data = data[data['IsOutlier'] != -1]
```

Output:

Head data

	CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Info data

None

3)Feature engineering

This process involves carefully selecting and crafting features that provide valuable insights into customer behaviour and characteristics. It will Create additional features that capture customer behaviour Preferences, such as total spending, frequency of purchases.

Code for feature engineering

```
import pandas as pd
import numpy as np
```

```
data = pd.read_csv("D:\IBM project\Mall_Customers.csv")
data['Gender'] = data['Genre'].map({'Male': 0, 'Female': 1})
print('Gender\n',data['Genre'])
data['Income_Spending'] = data['Annual Income (k$)'] * data['Spending Score (1-100)']
print('Income spending\n',data['Income_Spending'])
age_bins = [0, 25, 35, 45, 55, 100]
age_labels = ['18-25', '26-35', '36-45', '46-55', '56+']
data['Age_Group'] = pd.cut(data['Age'], bins=age_bins, labels=age_labels)
print('Age group\n',data['Age_Group'])
```

Output

```
Gender
 0 Male
1 Male
2 Female
3 Female
4 Female
 ...
195 Female
196 Female
197 Male
198 Male
199 Male
Name: Genre, Length: 200, dtype: object
Income spending
 0 585
1 1215
2 96
3 1232
4 680
 ...
195 9480
196 3528
197 9324
198 2466
199 11371
Name: Income_Spending, Length: 200, dtype: int64
```

Age group

0 18-25

1 18-25

2 18-25

3 18-25

4 26-35

...

195 26-35

196 36-45

197 26-35

198 26-35

199 26-35

Name: Age_Group, Length: 200, dtype: category

Categories (5, object): ['18-25' < '26-35' < '36-45' < '46-55' < '56+']

4) Applying clustering algorithms

The next phase involves the application of clustering algorithms.

These algorithms group customers into distinct segments based on the features we've engineered.

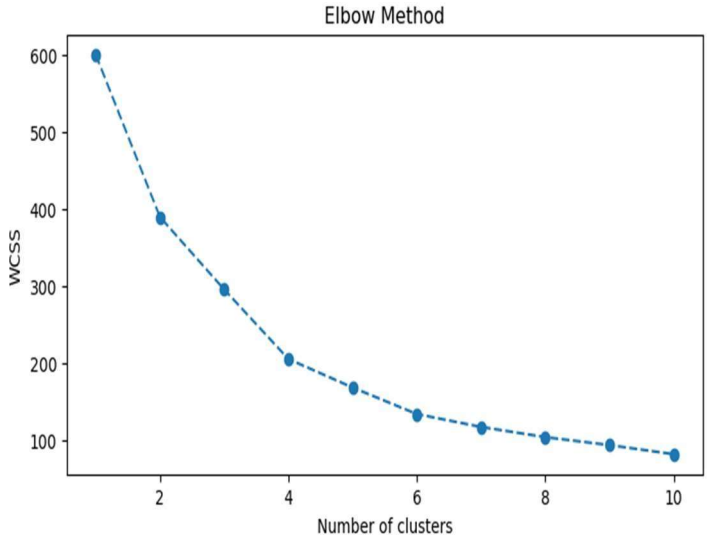
Popular clustering algorithms for customer segmentation include K-Means, hierarchical clustering, and DBSCAN. Each cluster generated represents a segment of customers who share common characteristics or behaviours.

Code:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

customer_data = pd.read_csv("D:\IBM project\Mall_Customers.csv")

features = customer_data[['Age', 'Annual Income (k$)','Spending Score (1-100)']]

scaler = StandardScaler()

scaled_features = scaler.fit_transform(features)

wcss = []

for i in range(1, 11):

 kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)

 kmeans.fit(scaled_features)

 wcss.append(kmeans.inertia_)
```

```
plt.figure(figsize=(8, 4))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
optimal_clusters = 3 # Adjust as needed
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300, n_init=10, random_state=0)
customer_data['Cluster'] = kmeans.fit_predict(scaled_features)
plt.figure(figsize=(8, 6))
for cluster in range(optimal_clusters):
 plt.scatter(reduced_features[customer_data['Cluster'] == cluster][:, 0],
 reduced_features[customer_data['Cluster'] == cluster][:, 1],
 label=f'Cluster {cluster}')
```

Output



5) Visualization

To gain a deeper understanding of the customer segments, visualization is indispensable.

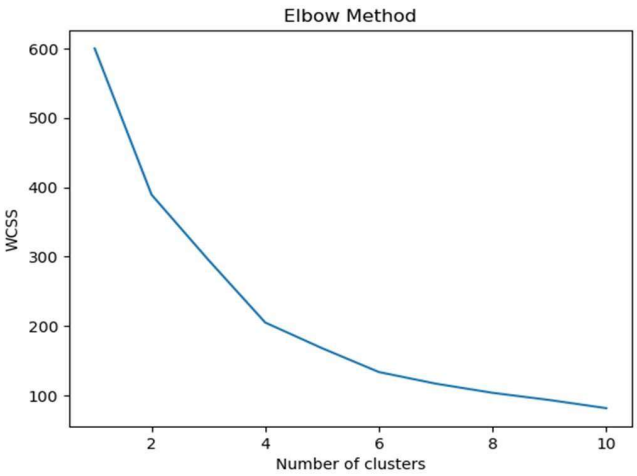
Visual representations help us comprehend and convey the results effectively.

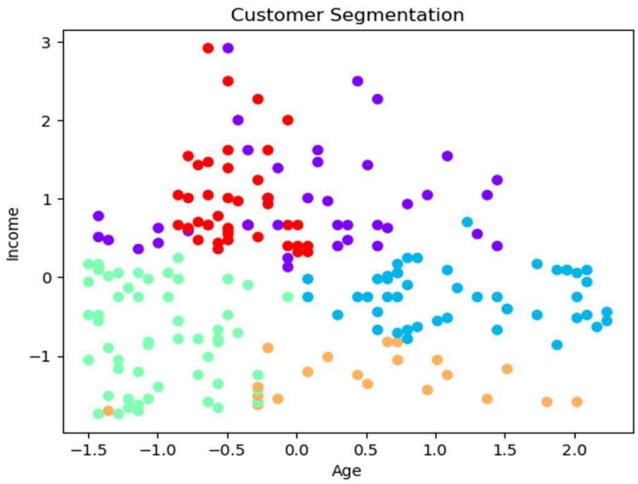
Visualization techniques may include scatter plots to visualize how customers cluster, heatmaps to show the similarity between customers in different clusters, bar charts to display the sizes of each segment, and dimensionality reduction plots like T-SNE or PCA, as well as distribution plots like box plots or histograms to understand feature distributions within clusters.

Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
data = pd.read_csv("D:\IBM project\Mall_Customers.csv")
X = data[['Age', 'Annual Income (k$)','Spending Score (1-100)']]
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
wcss = []
for i in range(1, 11):
 kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
 kmeans.fit(X_scaled)
 wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS') # Within-cluster sum of squares
plt.show()
k = 5
kmeans = KMeans(n_clusters=k, init='k-means++', max_iter=300, n_init=10, random_state=0)
kmeans.fit(X_scaled)
data['Cluster'] = kmeans.labels_
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=kmeans.labels_, cmap='rainbow')
plt.xlabel('Age')
plt.ylabel('Income')
plt.title('Customer Segmentation')
plt.show()
```

Output





6) Interpretation:

The final step is interpretation, where we derive meaningful insights from the clusters created. This

involves:

Assigning labels or names to each cluster based on common customer characteristics.

Distinguishing the unique traits of each segment, such as high spenders, infrequent shoppers, tech-savvy individuals, or price-sensitive customers.

Leveraging domain knowledge to explain the patterns uncovered.

Evaluating the quality of clustering results by using relevant metrics and aligning these insights with the business objectives.

Already interpretation is done in the previous code

KEY FINDINGS AND INSIGHTS

Some of the key findings and insights that can be gained from customer segmentation using data science include:

Customer needs and preferences: Customer segmentation can help businesses to identify the different needs and preferences of their customers. This information can then be used to develop more targeted products and services, and to create more relevant marketing campaigns.

Customer behaviour: Customer segmentation can also be used to understand customer behaviour, such as their purchase patterns, frequency of visits, and response to different marketing campaigns. This information can be used to improve the customer experience and to increase sales.

Customer profitability: Customer segmentation can also be used to identify the most profitable customers. This information can then be used to target these customers with special offers and promotions, and to develop strategies to retain them.

RECOMMENDATIONS

Once a business has segmented its customers, it can use this information to make a number of recommendations, such as:

Develop targeted products and services: Businesses can develop targeted products and services to meet the specific needs of each customer segment.

Create more relevant marketing campaigns: Businesses can create more relevant marketing campaigns by targeting each customer segment with messages that are tailored to their specific needs and interests.

Improve the customer experience: Businesses can improve the customer experience by understanding the different needs and preferences of each customer segment.

Increase sales: Businesses can increase sales by targeting the most profitable customers with special offers and promotions.

Retain customers: Businesses can develop strategies to retain the most profitable and valuable customers.

SOME REAL TIME EXAMPLES FOR CUSTOMER SEGMENTATION

Netflix: Netflix uses customer segmentation to recommend movies and TV shows to its users. By understanding the different genres and types of content that each customer segment watches, Netflix can provide more personalized recommendations.

Amazon: Amazon uses customer segmentation to recommend products to its customers. By understanding the different purchase patterns of each customer segment, Amazon can provide more relevant recommendations.

Disney: Disney uses customer segmentation to personalize the experience of its visitors. By understanding the different needs and preferences of each customer segment, Disney can provide more targeted attractions, dining options, and entertainment.

CONCLUSION FOR THE PROJECT

By grouping customers into segments based on shared characteristics, businesses can identify patterns and trends that would be difficult to see if all customers were treated as one group. By understanding the different needs, preferences, and behaviours of each customer segment, businesses can develop more targeted products and services, create more relevant marketing campaigns, improve the customer experience, and increase sales.