A PROJECT REPORT
on

Customer Personality Analysis System

(Detailed analysis of a company's ideal customers)

Submitted to Prof.

By

Ananya Thakur	**21052225**
Nihar Ranjan Sahoo	**21052165**

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

Acknowledgement

We would like to express my sincere gratitude to our guide and professor, Dr. SARITA TRIPATHY, who gave us this opportunity to implement this project and also guided us in the process in understanding the basic underlying principle of this project which helped us to complete the same. We used this medium to acknowledge and appreciate her as she contributed immensely and graciously for the achievement of this project work.

Contents

Abstract

Customer Personality Analysis is a thorough examination of a business's ideal clients. It allows a company to better understand its consumers and makes it simpler to alter goods according to the individual wants, habits, and concerns of various sorts of customers.

Customer personality analysis enables a company to adjust its product in response to the needs of its target customers from various categories. For example, instead of paying money to promote a new product to every client in the firm's database, a corporation may determine which customer group is most likely to purchase the product and then market it just to that segment.

Leveraging a comprehensive dataset consisting of 29 features originally , the system employs unsupervised machine learning techniques for efficient analysis. Through meticulous data preprocessing, including feature extraction and standardization, the project generates different clusters representing different types of customers.

The model uses PCA and k-means clustering to segment the data into different groups of customers for personality analysis. Elbow method was used to determine the three clusters, allowing more effective segmentation of customers.

Keywords: Unsupervised Machine learning, PCA, k-means clustering, Elbow Method.

Introduction

Understanding clients in today's commercial environment has evolved beyond demographic data. It is increasingly critical for businesses to dive deeper into the complexities of consumer behavior, interests, and expectations. Customer Personality Analysis appears as a critical method in this respect, providing a detailed insight of the different clients that firms serve.

This analytical approach not only helps businesses understand their consumers' diverse wants and behaviors, but it also allows them to precisely adjust their services. Companies may divide their consumer base into various personas by deconstructing a variety of criteria and applying modern data analytics tools. Each persona represents a unique set of features and interests.

This project makes use of techniques like PCA, k-means clustering and other data pre-processing techniques to segment customers into three clusters using the elbow method. At the end we were able to segment the customers into three clusters based on their total spending, determining if they are frequent customers or not and what is the range of customers spending more money.

A model like this helps the companies to segment their customers profitably on various conditions , thus, making it easier for them to design products according to the target customers.

Objective and Project Flow

1. **Project Scope and Objectives:**
   - Objective: Develop a customer personality analysis system using ML.
   - Scope: Extract, preprocess, and analyze customer data to help companies segment customers for more profit.

2. **Key Tasks and Milestones:**
   - **Data Preparation:**
     - *Datasets:* Used marketing_campaign.csv
     - *Data Visualization:* Plot histograms and Bar plots to check distribution of various features
     - *Handle missing values:* Ensure data consistency and address missing values.
     - *Handling outliers:* Remove outliers to ensure consistency and avoid irregularities in the dataset.
     - *Feature Extraction:* Merge features together and convert them into broader features to keep relevant data and drop the rest of the unnecessary columns.
   - **Algorithm Implementation:**
     - *Implement elbow method:* To find optimal numbers of clusters to be made.
     - *Apply PCA:* Transform original data into principal components for analysis.
     - *Apply K-means clustering:* Visualize 3 main clusters identified before.
3. **Resource Requirements:**
   - Tools: Python, Pandas, NumPy, Scikit-learn, seaborn, matplotlib.
   - Datasets:marketing_campaign.csv.

Implementation

This study aims to develop a Customer Personality System by integrating unsupervised machine learning techniques using open-source Python modules like Scikit-Learn. Efficient data preprocessing with Pandas resulting in 13 final columns. This System makes use of PCA and k-means clustering for customer segmentation. The detailed stepwise implementation is:

Data Collection: Gather a comprehensive dataset (marketing_campaign.csv) consisting of 29 features related to customer behavior, demographics, and preferences. This dataset may include information such as age, gender, purchase history, spending etc.

```python
df = pd.read_csv("marketing_campaign.csv",sep = "\t")

df.head()
```

[2] Python

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMont
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	

5 rows × 29 columns

Data Preprocessing: Conduct thorough data preprocessing tasks, including cleaning, feature extraction, missing-value handling, null-value removal and standardization. This ensures that the dataset is suitable for analysis and eliminates any inconsistencies or missing values.

Code snippets for data pre-processing::

```python
# Replace 'Married' and 'Together' with 'Relationship'
df.loc[df['Marital_Status'].isin(['Married', 'Together']), 'Marital_Status'] = 'Relationship'

# Replace 'Single', 'Divorced', 'Widow', 'Alone', 'Absurd', and 'YOLO' with 'Single'
df.loc[df['Marital_Status'].isin(['Single', 'Divorced', 'Widow', 'Alone', 'Absurd', 'YOLO']), 'Marital_Status'] = 'Single'
```

```python
df.loc[:, 'Kids'] = df['Kidhome'] + df['Teenhome']
df.loc[:, 'Expenses'] = df['MntWines'] + df['MntFruits'] + df['MntMeatProducts'] + df['MntFishProducts'] + df['MntSweetProduc
df.loc[:, 'TotalAcceptedCmp'] = df['AcceptedCmp1'] + df['AcceptedCmp2'] + df['AcceptedCmp3'] + df['AcceptedCmp4'] + df['Accep
df.loc[:, 'NumTotalPurchases'] = df['NumWebPurchases'] + df['NumCatalogPurchases'] + df['NumStorePurchases'] + df['NumDealsPu
```

```python
columns_to_drop = ['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']
df = df.drop(columns=columns_to_drop)
```

```python
# Drop 'Kidhome' and 'Teenhome' columns
df.drop(columns=['Kidhome', 'Teenhome'], inplace=True)
```

```python
df['NumTotalPurchases'] = df['NumWebPurchases'] + df['NumCatalogPurchases'] + df['NumStorePurchases'] + df['NumDealsPurchases
```

Python

```python
# Drop the columns used to calculate 'NumTotalPurchases'
columns_to_drop = ['NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumDealsPurchases']
df.drop(columns=columns_to_drop, inplace=True)
```

```python
# Convert Marital_Status to binary (0, 1)
df['Marital_Status'] = df['Marital_Status'].map({'Single': 0, 'Relationship': 1})
```

Py

```python
# Convert Education to binary (0, 1)
df['Education'] = df['Education'].map({'Graduation': 0, 'PhD': 1, 'Master': 2, '2n Cycle': 3, 'Basic': 4})
```

Dataset before Pre-processing:

	ID	Education	Marital_Status	Income	Dt_Customer	Recency	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	Num
0	5524	Graduation	Single	58138.0	04-09-2012	58	3	8	10	
1	2174	Graduation	Single	46344.0	08-03-2014	38	2	1	1	
2	4141	Graduation	Relationship	71613.0	21-08-2013	26	1	8	2	
3	6182	Graduation	Relationship	26646.0	10-02-2014	26	2	2	0	
4	5324	PhD	Relationship	58293.0	19-01-2014	94	5	5	3	

5 rows × 25 columns

```python
df.columns
```

Python

```
Index(['ID', 'Education', 'Marital_Status', 'Income', 'Dt_Customer', 'Recency',
       'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases',
       'NumStorePurchases', 'NumWebVisitsMonth', 'AcceptedCmp3',
       'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2',
       'Complain', 'Z_CostContact', 'Z_Revenue', 'Response', 'Kids',
       'Expenses', 'TotalAcceptedCmp', 'NumTotalPurchases', 'Age'],
      dtype='object')
```

Dataset after Pre-processing:

```
df.describe()
```

✓ 0.0s Python

	Education	Marital_Status	Income	Recency	NumWebVisitsMonth	Complain	Kids	Expenses	TotalAcceptedCmp	NumTotalPurchases	
count	2236.000000	2236.000000	2236.000000	2236.000000	2236.000000	2236.000000	2236.000000	2236.000000	2236.000000	2236.000000	2236.000
mean	0.914132	0.644902	51961.906544	49.116279	5.318873	0.008945	0.950805	605.986583	0.447227	14.872540	55.101
std	1.113174	0.478650	21411.404811	28.957284	2.426886	0.094173	0.752204	601.865156	0.891113	7.677874	11.703
min	0.000000	0.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	5.000000	0.000000	0.000000	28.000
25%	0.000000	0.000000	35502.500000	24.000000	3.000000	0.000000	0.000000	69.000000	0.000000	8.000000	47.000
50%	0.000000	1.000000	51684.000000	49.000000	6.000000	0.000000	1.000000	396.500000	0.000000	15.000000	54.000
75%	2.000000	1.000000	68275.750000	74.000000	7.000000	0.000000	1.000000	1045.500000	1.000000	21.000000	65.000
max	4.000000	1.000000	162397.000000	99.000000	20.000000	1.000000	3.000000	2525.000000	5.000000	44.000000	84.000

```
df.columns
```

✓ 0.0s Python

```
Index(['Education', 'Marital_Status', 'Income', 'Recency', 'NumWebVisitsMonth',
       'Complain', 'Kids', 'Expenses', 'TotalAcceptedCmp', 'NumTotalPurchases',
       'Age', 'Time_Customer'],
      dtype='object')
```

Unsupervised Learning: Applied unsupervised machine learning techniques such as PCA and K-means clustering algorithm to the preprocessed dataset. These algorithms will group similar customers together based on their attributes, forming distinct clusters. Applied the elbow method to determine the number of clusters.

Principle Component Analysis-

```
pca = PCA(n_components=2, whiten=True)
pca.fit(df)
data_pca = pca.transform(df)
```

✓ 0.0s

Selecting Number of clusters to make-

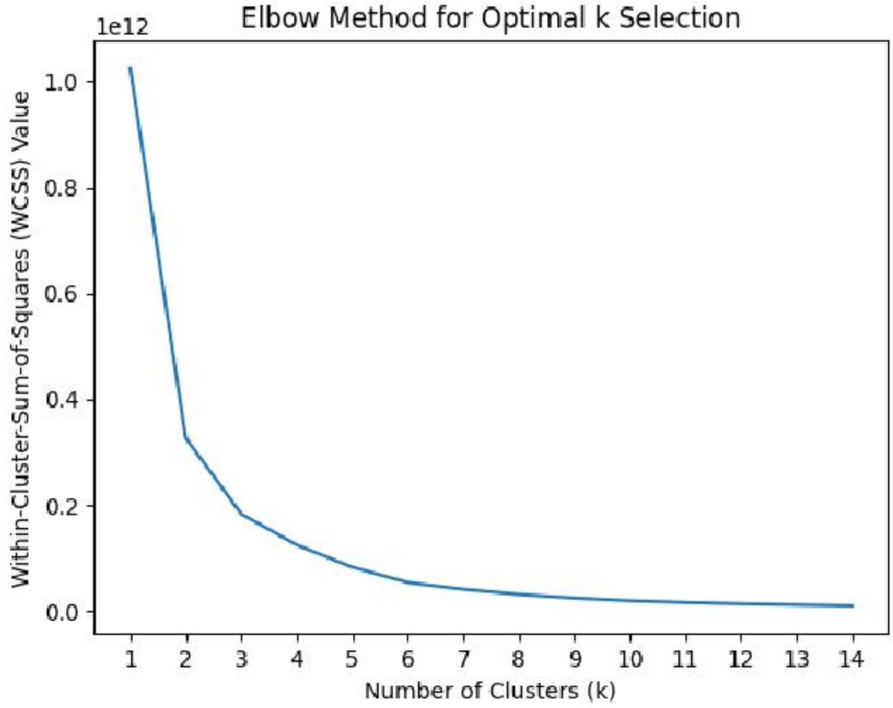
```
from sklearn.cluster import KMeans

# List to store the Within-Cluster-Sum-of-Squares (WCSS) values for different values of k
wcss = []

# Iterate through different values of k (number of clusters)
for k in range(1, 15):
    # Create a KMeans clustering model with the current value of k
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(df)
    wcss.append(kmeans.inertia_)

# Plot the WCSS values against the number of clusters (k)
plt.plot(range(1, 15), wcss)
plt.xlabel("Number of Clusters (k)")
plt.xticks(range(1, 15, 1))
plt.ylabel("Within-Cluster-Sum-of-Squares (WCSS) Value")
plt.title("Elbow Method for Optimal k Selection")
plt.show()
```

✓ 2.4s



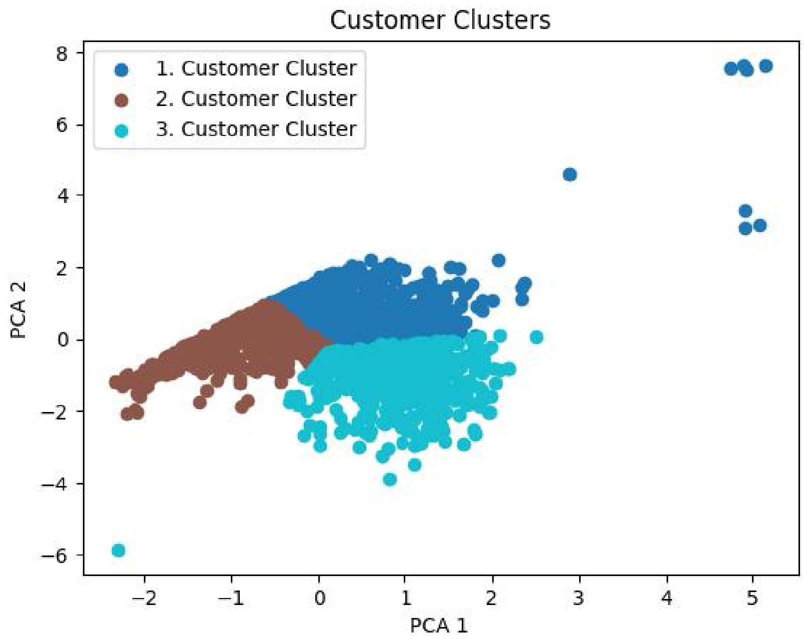
Cluster Analysis: Analyze the resulting clusters to understand the characteristics and preferences of each segment. This involves visualizing those clusters by plotting scatter plots.

K-means clustering-

```
kmeans2 = KMeans(n_clusters=3)
clusters = kmeans2.fit_predict(data_pca)
colors = plt.cm.get_cmap('tab10', 3)

for cluster_num in range(3):
    plt.scatter(data_pca[clusters == cluster_num, 0],
                data_pca[clusters == cluster_num, 1],
                label=f'{cluster_num + 1}. Customer Cluster',
                color=colors(cluster_num))

plt.title('Customer Clusters')
plt.xlabel('PCA 1')
plt.ylabel('PCA 2')
plt.legend()
plt.show()
```
✓ 0.2s

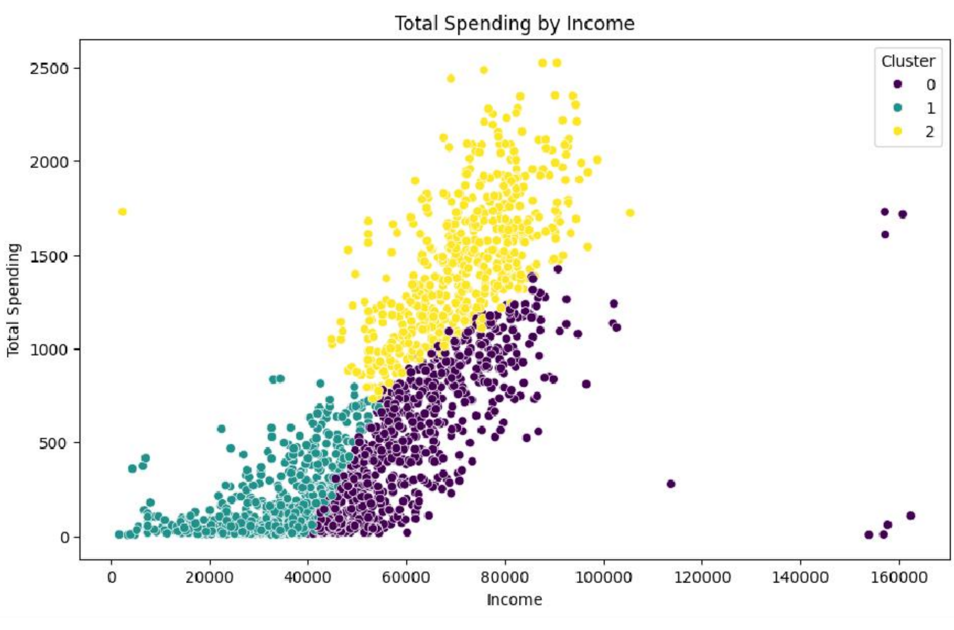


Results and Insights:

Using the mentioned dataset, the K-means algorithm effectively identified numerous different consumer groups. Each cluster has distinct features, such as income, shopping patterns, and preferences. Businesses may better target certain clusters of customers by knowing their unique identities.

The system was able to segment the customers into three clusters using K-means clustering and Principal Component Analysis .

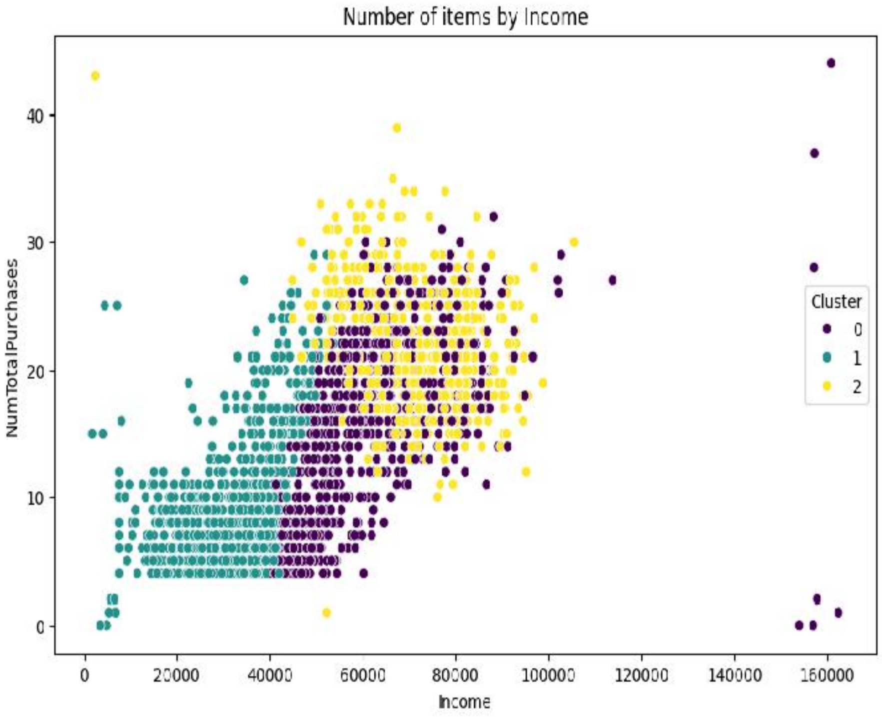
Comparison of Total spendings w.r.t income



From this graph we can observe that:-

- Cluster 0 most likely represents people with a high income and low total spending.
- Cluster 1 most likely represents people with a low income and low total spending.
- Cluster 2 most likely represents people with a high income and high total spending

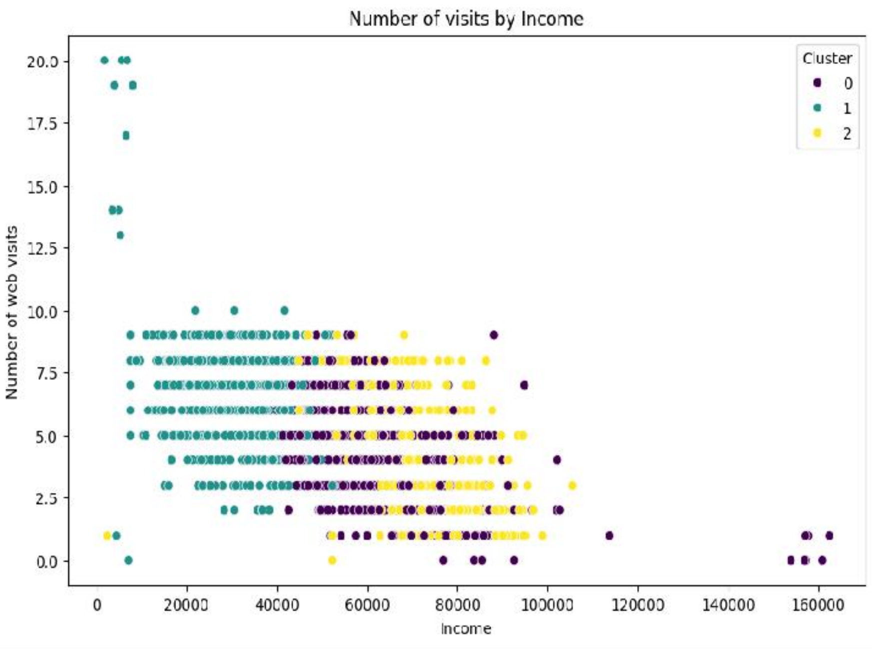
Comparison of the number of purchases w.r.t income :



From this graph, we can observe that:-

- Cluster 0 (High Income Low spending):- These customers have high income but most likely have the same spending habits as Cluster 0 as not only they tend to spend less but also buy nearly the same amount of items.
- Cluster 1 (Low income Low spending):-While this group has low income and spends less, they don't seem to be buying fewer items, indicating they might be buying cheaper items.
- Cluster 2 (High Income High spending):- These customers have high income, and spend a lot, and tend to buy more items.

Comparison of number of Web visits w.r.t income



While most of the groups have similar numbers of visits, there is a slight trend where we can observe that the higher income groups tend to visit the website less often.

This analysis concludes that the companies should try to focus on trying to increase the engagement of their higher income customers and push more products onto them.

Conclusion:

In conclusion, the customer personality analysis project effectively illustrates the efficiency of unsupervised machine learning algorithms for understanding and segmenting client personalities. By evaluating a large dataset comprising 29 attributes, three client clusters were found, each reflecting a different type of consumer with distinctive qualities and interests. The findings of this research give significant information for firms looking to improve their marketing tactics, optimize product development, and increase consumer engagement. Finally, by adapting solutions to the individual demands of different client segments, firms may increase customer happiness, loyalty, and market competitiveness. Moving forward, continual refining and application of these insights will be critical for maintaining customer-centric methods and generating long-term success in today's changing business environment.