Final Project Report

Customer Segmentation Using Machine Learning

1.Introduction

1.1 Project overviews

In today's digital age, e-commerce has become a cornerstone of retail, with global online sales projected to surpass $6 trillion by 2024, emphasizing the need for businesses to effectively understand their customer base. With millions of users engaging with online platforms, the complexity of consumer behaviors necessitates advanced analytical techniques to segment customers meaningfully. Traditional methods of customer analysis often fall short, leading to generalized marketing strategies that fail to resonate with diverse consumer groups. This project aims to address these challenges by utilizing machine learning for customer segmentation, focusing on data derived from online e-commerce sites. Through comprehensive exploratory data analysis (EDA), we will uncover key insights and trends within the dataset, which will guide our feature engineering efforts.

To manage the high dimensionality of the data, Principal Component Analysis (PCA) will be employed, enabling us to distill the most informative features for clustering. We will then implement K-means clustering to categorize customers into four distinct segments, leveraging a similarity score of 83 to ensure meaningful differentiation among groups. By analyzing customer characteristics in depth, this project seeks to provide actionable insights that empower e-commerce businesses to tailor their marketing efforts, optimize product offerings, and enhance customer engagement, ultimately contributing to increased sales and customer loyalty in a competitive market landscape.

SMARTBRIDGE
Let's Bridge the Gap
a Veranda Enterprise

1.2 Objectives

The objective of this project is to leverage machine learning techniques for effective customer segmentation in e-commerce, enabling businesses to identify and understand distinct customer groups based on their behaviors and preferences. By employing exploratory data analysis, feature engineering, PCA for dimensionality reduction, and K-means clustering, the project aims to categorize customers into meaningful segments. This segmentation will assist e-commerce platforms in tailoring marketing strategies, optimizing product offerings, and enhancing customer experiences, ultimately leading to increased sales and improved customer retention.

Specific Objectives:

1. **Identifying Customer:**
    - **Objective:** Identify distinct customer groups within an e-commerce platform based on their behaviors and preferences.
    - **Specifics:** Use machine learning algorithms such as K-means clustering to group customers with similar characteristics. This enables the e-commerce platform to better understand different customer types, such as high spenders, deal seekers, or infrequent buyers.
2. **Optimizing Marketing Strategies:**
    - **Objective:** Enable the e-commerce business to develop personalized marketing strategies for each segment.
    - **Specifics:** Tailor promotions, offers, and messaging to each segment to enhance engagement and increase conversion rates. For example,

sending high-spending customers exclusive deals on premium products while offering frequent buyers discounts on their next purchase.

3. **Enhancing Customer Experience:**
   - Assist authorities in **synchronizing traffic signals**, planning road usage, and managing public transportation. ○ Enable **proactive responses** to potential congestion and disruptions.
4. **Reduced Environmental Impact:**
   - Promote **efficient routing** and minimize vehicle idle times to lower emissions and fuel consumption. ○ Support sustainable urban mobility through optimized traffic flows.
5. **Scalability and Adaptability:**
   - Design a flexible solution that can **scale across multiple cities** and adapt to various infrastructure and traffic patterns.
   - Ensure compatibility with new data sources and technologies such as **IoT devices**.
6. **Enhanced Commuter Experience:**
   - Provide actionable predictions through **web and mobile platforms**, allowing commuters to plan better routes. ○ Reduce travel time, frustration, and uncertainty for users by offering **real-time alerts**.
7. **Data-Driven Decision Support:**
   - Empower city planners and traffic managers with predictive insights to **improve long-term urban planning**.
   - Prepare for **unexpected events** such as accidents or weather disruptions, ensuring smooth traffic operations.

2. Project Initialization and Planning Phase

2.1 Define Problem Statements (Customer Problem Statement Template):

By implementing a robust machine learning approach to segment customers into distinct groups, this project aims to provide actionable insights that can drive targeted marketing efforts and enhance overall customer satisfaction.

Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	A marketing manager at a retail or service-based company	Segment my customer base effectively to tailor personalized marketing startegies and improve customer retention.	Traditional segmentation methods rely on simple demographics data and don't capture the complex behavior of customers, leading to generic marketing campaingns.	Customers have diverse needs, perferences, and purchasing behavior that change over time, making it hard to identify relevant customer groups using manual methods	Frustrated because of missed opportunities to engage customers effectively, leading to lower conversion rates and reduced customer loyalty.

SMARTBRIDGE
Let's Bridge the Gap
a Veranda Enterprise

I am	I'm trying to	But	Because	Which makes me feel
A marketing manager ata retail or service- based company	improve customer retention	The complex behavior of customers, leading to generic marketing campaingns.	Identify relevant customer groups using manual methods	leading to lower conversion ratesand reduced customer loyalty.

2.2 Project Proposal (Proposed Solution) template

This project proposal outlines a solution to address a specific problem. With a clear objective, defined scope, and a concise problem statement, the proposed solution details the approach, key features, and resource requirements, including hardware, software, and personnel.

Project Overview	
Objective	To leverage Machine Learning algorithms to segment a business's customer base into distinct groups based on behaviors, preferences, & demographics. It aims to automate customer segmentation, making it scalable, accurate, & actionable for businesses.
Scope	The boundaries define the limits of the project regarding time, data sources, technology, and stakeholder involvement. The extent outlines the key activities and deliverables like data collection & preparation, model development & evaluation.

Problem Statement	
Description	Businesses struggle to effectively understand and engage their diverse customer base due to outdated segmentation methods that rely on basic demographics. This results in generic marketing strategies that fail to resonate with customers, leading to decreased satisfaction and missed opportunities.
Impact	The impact of this project extends beyond improved marketing strategies; it will transform how businesses understand and engage with their customers, leading to increased satisfaction, loyalty, and ultimately, profitability.
Proposed Solution	
Approach	Data collection & preprocessing (cleaning & transformation), EDA, feature engineering, model development & evaluation, implementing & reporting.

Key Features	Data - Driven insights, dynamic segmentation, personalized customer profiles, integration of multiple data sources, scalability.

Resource Requirements

Resource Type	**Description**	**Specification/Allocation**
Hardware		
Computing Resources	GPUs for model training	2 x NVIDIA V100 GPUs
Memory	RAM for processing large datasets	8 GB RAM
Storage	Disk space for models and logs	512 GB SSD
Software		
Frameworks	Python frameworks	Flask
Libraries	Additional machine learning tools	Scikit-learn, pandas, numpy, matplotlib

Development Environment	IDE and version control tools	Jupyter Notebook, Git, Vscode.
Data		
Data	Data source, size, and format	Kaggle dataset, 48 KB Size, CSV format.

2.3 Initial Project Planning Template

	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members	Sprint Start Date	Sprint End Date (Planned)
Sprint-1	Data Collection & Preparation	USN-1	As a user, I want to collect & clean customer data from various sources & want to perform preprocessing like encoding, handling missing values.	3	High	Thota Sivalingeswari	15-08-2024	22-08-2024
Sprint-2	Exploratory Data Analysis(EDA)	USN-2	As a user, I want to identify important features & to visualize data distribution using histograms, pair plots for better decision making.	2	Medium	Shaik Mahira	23-08-2024	30-08-2024
Sprint-3	Model Building & Evaluation	USN-3	As a user, I want to build & train unsupervised ML models e.g., K-means, hierarchical clustering & evaluate them to find optimal number of segments.	3	High	Mettu Hari Chandana	31-08-2024	08-09-2024
Sprint-4	Model Deployment & Reporting	USN-4	As a user, I want to deploy models to a cloud platform e.g., AWS, Azure used in production & to generate reports that stakeholders can easily understand.	2	Medium	Kunchala Suresh	09-09-2024	16-09-2024

4. Data Collection and Preprocessing Phase

4.1 Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	This project aims to Segmentation will Assist E-Commerce Platforms in Tailoring Marketing Strategies, Optimizing Product Offerings, and Enhancing Customer Experiences, Ultimately leading to increased sales and improved Customer retention.
Data Collection Plan	The Data Collection Plan is to Surveys and Questionnaires to Gather Direct Feedback, Interviews or Focus Groups for Qualitative Insights, Web Analytics and tracking tools to monitor online Behavior.
Raw Data Sources Identified	Transactional Data (Purchase History), Customer Demographics Customer Profiles from CRM System), Web and Analytics (User behavior data from website tracking).

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Smart Internz Platform	The data consists ofmeta data	https://docs.google.com/spreadsheets/d/1NnUMX3sjJgRRerkJTAXemlfdyo2GiUhgE_m4w-fAhvs/edit?gid=1219451115#gid = 1219451115	CSV	48 MB	Public

Data Collection and Preprocessing Phase

4.2 Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Smart Internz Dataset	Categorical data in the dataset	Moderate	Encoding has to be done in the data

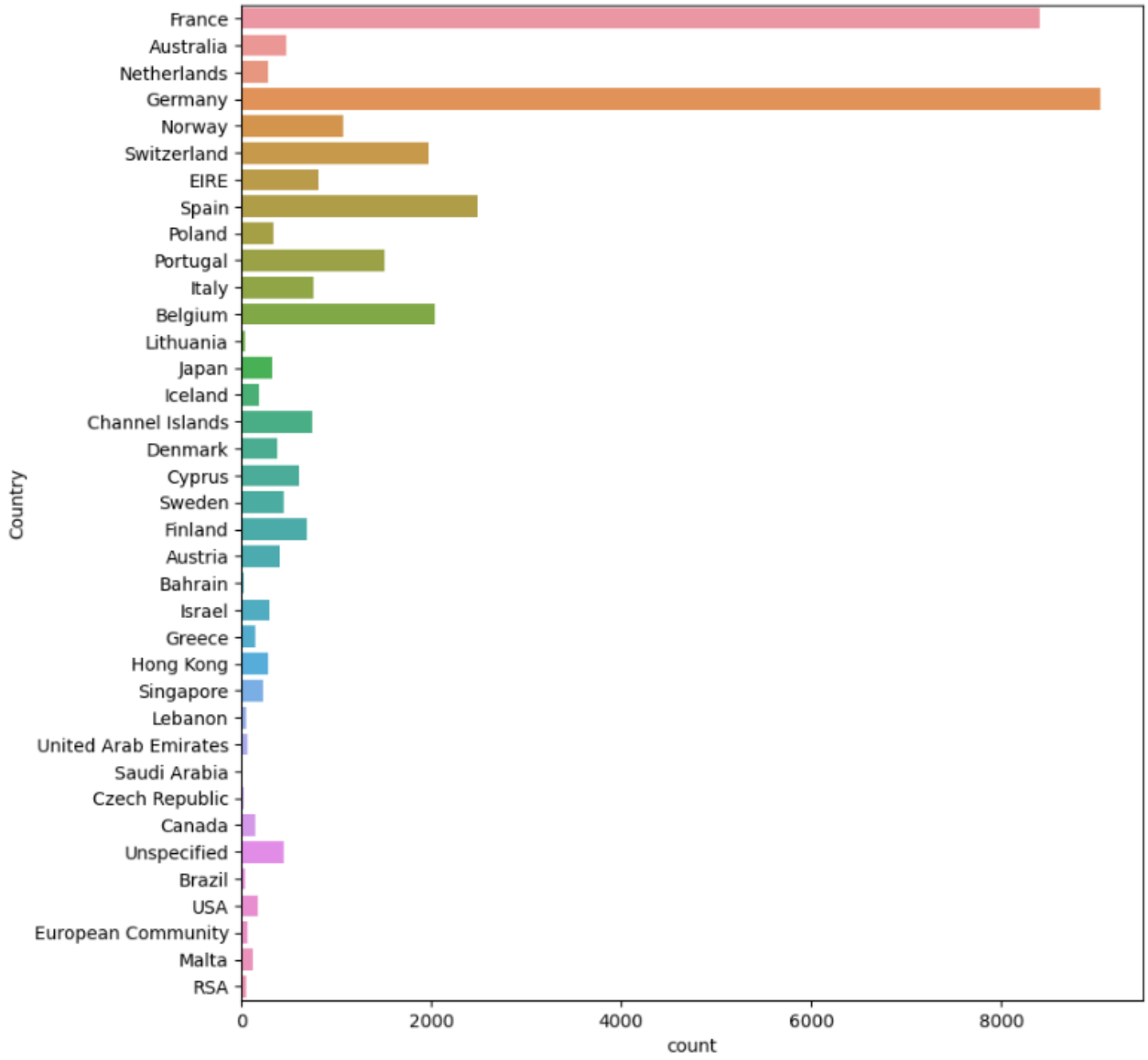
Data Collection and Preprocessing Phase

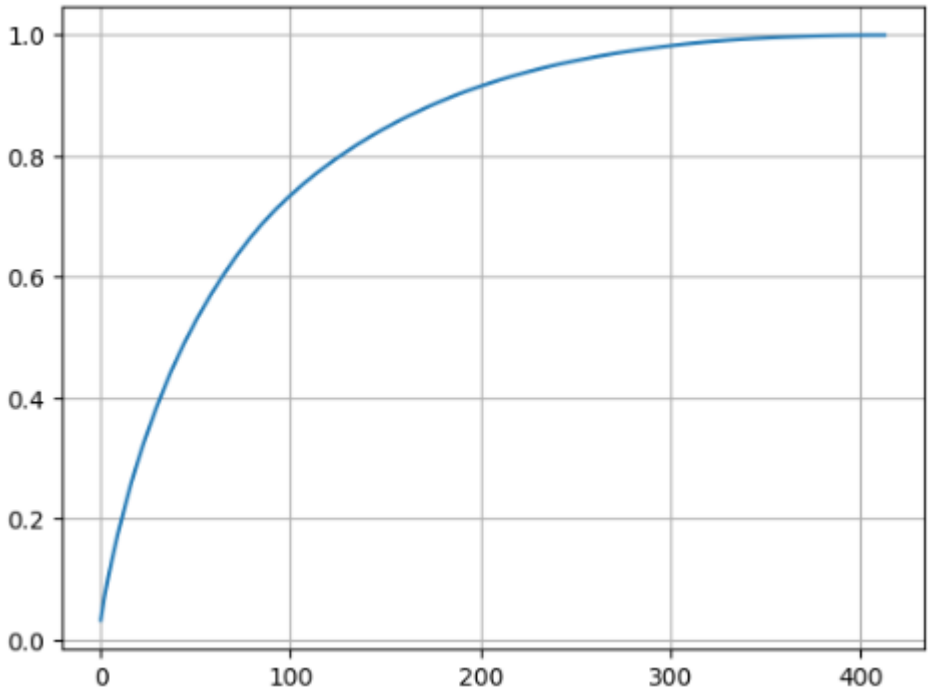
4.3 Data Exploration and Preprocessing Template

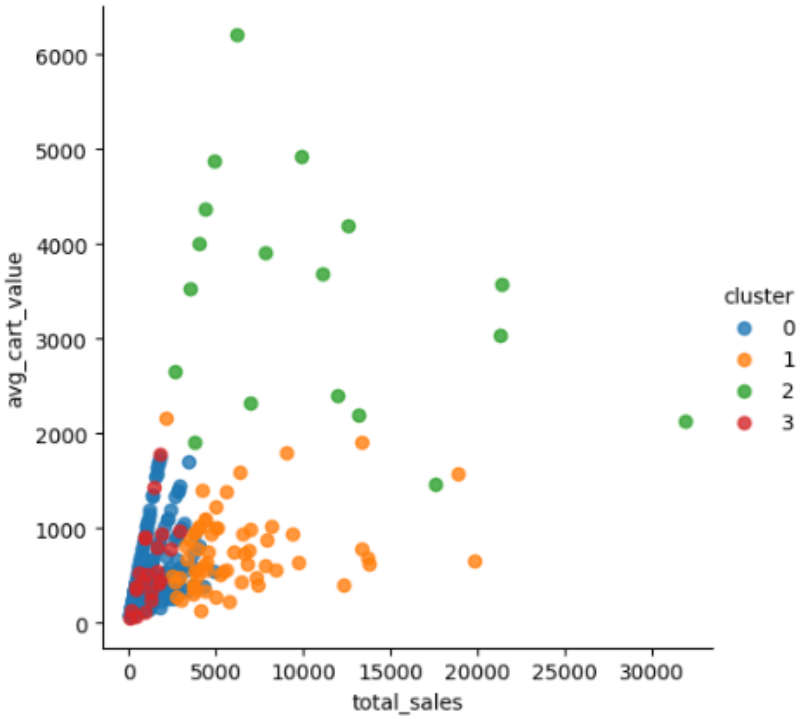
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
1. Data Collection	The process of gathering data from various sources, such as surveys, databases, APIs, or web scraping. This step ensures that relevant and sufficient data is obtained for analysis.
2. Data Inspection	Analyzing the collected data to understand its structure, characteristics, and quality. This includes checking for missing values, data types, and general statistics to identify initial issues.
3.Exploratory Data Analysis (EDA)	A visual and quantitative analysis of the dataset to uncover patterns, trends, and insights. EDA techniques include statistical summaries, data visualization, and correlation analysis, helping to inform further analysis and model selection.
4. Data Cleaning	The process of identifying and correcting errors or inconsistencies in the data. This includes handling missing values, correcting inaccuracies, removing duplicates, and ensuring data integrity.
5. Data Balancing	Addressing class imbalance in the dataset to ensure that models are trained effectively. Techniques may include oversampling minority classes, under sampling majority classes, or using

	synthetic data generation methods like SMOTE.
Data Preprocessing Code Screenshots	
6. Text Preprocessing	Preparing text data for analysis, which involves tasks such as tokenization, removing stop words, stemming or lemmatization, and converting text to lower case. This step enhances the quality of text inputs for further processing.
7. Label Encoding	Transforming categorical labels into numerical values, allowing models to process these categories effectively. This step is essential for machine learning algorithms that require numerical input.
8. Data Splitting	Dividing the dataset into training and testing subsets. The training set is used to build the model, while the testing set evaluates its performance.
9. Model Building	The phase where machine learning or statistical models are constructed using the training data. Various algorithms can be applied based on the problem type (e.g., classification, regression).
10.Model Evaluation	Assessing the performance of the built model using the testing dataset. Evaluation metrics may include accuracy, precision, recall, F1 score, and AUC-ROC for classification problems.


France
Australia
Netherlands
Germany
Norway
Switzerland
EIRE
Spain
Poland
Portugal
Italy
Belgium
Lithuania
Japan
Iceland
Channel Islands
Denmark
Cyprus
Sweden
Finland
Austria
Bahrain
Israel
Greece
Hong Kong
Singapore
Lebanon
United Arab Emirates
Saudi Arabia
Czech Republic
Canada
Unspecified
Brazil
USA
European Community
Malta
RSA
Country
0
2000
4000
6000
8000
count


1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400


6000
5000
4000
3000
2000
1000
0
avg_cart_value
0
5000
10000
15000
20000
25000
30000
total_sales
cluster
0
1
2
3

Smart Internz

5.Model Development Phase Template

5.1 Feature Selection Report Template

In the forthcoming update, each feature will be accompanied by a brief description. Users will indicate whether it's selected or not, providing reasoning for their decision. This process will streamline decision-making and enhance transparency in feature selection.

Feature	Description	Selected (Yes/No)	Reasoning
Age	The age the individual in year	Yes	Age Influences Purchasing Behavior, Preferences, and Brand Loyalty. Younger Customers may Prioritize Trends, While older Customers might Value Quality and Service.
Family Size	It includes the number of people Living together	Yes	Family Size influences buying habits. Larger Families may Prioritize Bulk purchases and values, While Smaller Families might Focus on covenience and Premium Products.
Work Experience	It provides valuable Insights of Different Customer Groups	Yes	Work Experience often correlates with income, influencing purchasing power. This helps Businesses Tailor Products and Pricing Strategies to different segments.
Profession	It is essential for understanding consumer Behavior	**Yes**	Different Professions have Unique needs and Preferences. Tailoring marketing messages to specific Industries (e.g., healthcare, tech, education) can increase relevance and engagement.

Gender	It enhances marketing strategies	**Yes**	Gender insights can inform product design and features, ensuring offerings meet the preferences of different gender (e.g., cosmetics, clothing, personal care products).
Married	It provides valuable insights for tailoring marketing strategies	**Yes**	Married individuals may have different needs compared to singles, such as home goods, family planning products, or joint Financial Services.
Graduated	It has two Category graduated and not graduated	**Yes**	Analyzing customer interactions, purchase history, and Engagement levels to identify patterns and preferences.
Spending score	It quantifies customers spend and their purchasing behavior	**Yes**	The spending sore typically ranges from 0 to 100 and reflects a customer's spending habits over a specific period. Higher scores indicate higher spending levels.

Model Development Phase Template

5.2 Model Selection Report

In the forthcoming Model Selection Report, various models will be outlined, detailing their descriptions, hyperparameters, and performance metrics, including Accuracy or F1 Score. This comprehensive report will provide insights into the chosen models and their effectiveness.

Model Selection Report:

Model	Description	Hyperparameters	Performance Metric (e.g., Accuracy, F1 Score)

K-Mean	Select K initial centroids randomly from the data points. Assign each data points assigned to each cluster. Recalculate the centroids as the mean of all points assigned to each cluster.	-	Accuracy score=80%
Hierarchical Clustering	It is an effective method for customer segmentation, providing insights into the structure of customer data	-	Accuracy score=80%
DBSCAN	It is the framework for segmentation that focuses on the key dimensions to better understanding and categorizes	-	Accuracy score=80%

Model Development Phase Template

5.3 Initial Model Training Code, Model Validation and Evaluation Report

The initial model training code will be showcased in the future through a screenshot. The model validation and evaluation report will include classification reports, accuracy, and confusion matrices for multiple models, presented through respective screenshots.

Initial Model Training Code:

Paste the screenshot of the model training code

```
[97] from sklearn import linear_model
     from sklearn import tree
     from sklearn import ensemble
     from sklearn import svm
     import xgboost
```

```
[98] lin_reg=linear_model.LinearRegression()
     Dtree=tree.DecisionTreeRegressor()
     Rand=ensemble.RandomForestRegressor()
     svr=svm.SVR()
     XGB=xgboost.XGBRegressor()
```

```
     lin_reg.fit(x_train,y_train)
     Dtree.fit(x_train,y_train)
     Rand.fit(x_train,y_train)
     svr.fit(x_train,y_train)
     XGB.fit(x_train,y_train)
```

```
[100] p1=lin_reg.predict(x_train)
      p2=Dtree.predict(x_train)
      p3=Rand.predict(x_train)
      p4=svr.predict(x_train)
      p5=XGB.predict(x_train)
```

```
[101] from sklearn import metrics
```

```
[102] print(metrics.r2_score(p1,y_train))
      print(metrics.r2_score(p2,y_train))
      print(metrics.r2_score(p3,y_train))
      print(metrics.r2_score(p4,y_train))
      print(metrics.r2_score(p5,y_train))
```

```
[103] p1=lin_reg.predict(x_test)
      p2=Dtree.predict(x_test)
      p3=Rand.predict(x_test)
      p4=svr.predict(x_test)
      p5=XGB.predict(x_test)
```

```
print(metrics.r2_score(p1,y_test))
print(metrics.r2_score(p2,y_test))
print(metrics.r2_score(p3,y_test))
print(metrics.r2_score(p4,y_test))
print(metrics.r2_score(p5,y_test))
```

```
[105] MSE=metrics.mean_squared_error(p3,y_test)
```

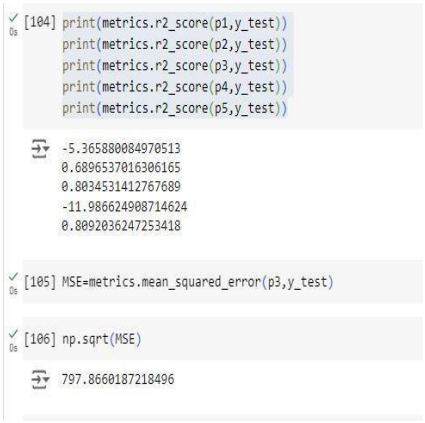
```
np.sqrt(MSE)
```

```
797.8660187218496
```

```
[109] import pickle
```

```
pickle.dump(Rand,open("model.pk1",'wb'))
pickle.dump(le,open("encoder.pk1",'wb'))
```

Model Validation and Evaluation Report:

Model	Classification Report	Accuracy	Confusion Matrix
Random Forest		80%	---

6. Model Optimization and Tuning Phase Template

Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

Hyperparameter Tuning Documentation (6 Marks):

Model	Tuned Hyperparameters	Optimal Values
----	-----	

Performance Metrics Comparison Report (2 Marks):

Model	Baseline Metric	Optimized Metric
----	-----	

Final Model Selection Justification:

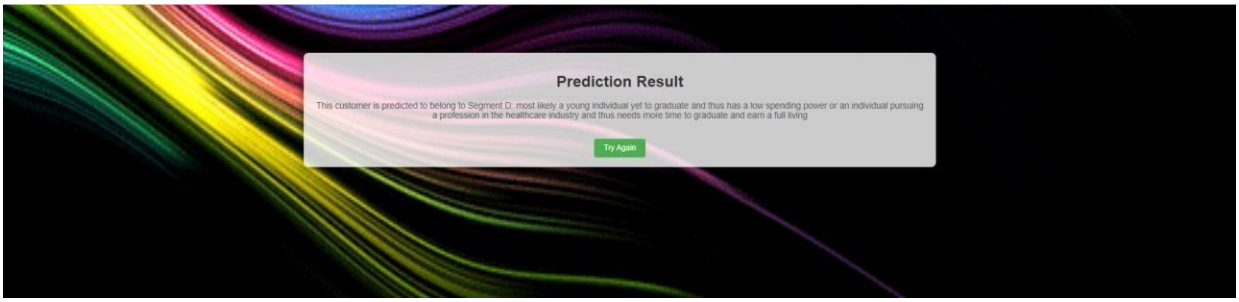
Final Model	Reasoning

Gradient Boosting	XG Boost was choosen as the final optimized model due to its high predictive accuracy and effieciency.Its built-in regularization helped prevent overfitting,while its ability to handle missing values simplified preprocessing.Additionally,XGBoost provides valuable insights into feature importance,enhancing interpretability and model refinement to align with project objectives justifying it as a final model.

Smart Internz

7. Results

7.1 Outputs screenshots



8. Advantages & Disadvantages

ADVANTAGES:

- **Enhanced Accuracy:** By integrating exploratory data analysis, feature engineering, and dimensionality reduction, the proposed method improves the accuracy of customer segmentation, enabling more precise identification of distinct customer groups and their specific needs.
- **Dynamic Adaptability:** The use of advanced machine learning techniques allows for continuous updates to customer segments, ensuring that the model adapts to changing consumer behaviors and market trends, thus maintaining relevance over time.
- **Actionable Insights:** The comprehensive approach provides deeper insights into customer preferences and behaviors, empowering businesses to implement targeted marketing strategies, optimize product offerings, and ultimately enhance customer satisfaction and loyalty.

DISADVANTAGES:

- **Limited Insight Depth:** Traditional methods, such as RFM analysis, often provide superficial insights, failing to capture the complexity of customer behaviors and preferences, which can lead to ineffective targeting.

- **Assumption of Spherical Clusters:** Algorithms like K-means assume that clusters are spherical and evenly sized, which may not accurately reflect the true distribution of customer data, resulting in misclassification of customer segments.
- **High Dimensionality Challenges:** Existing methods may struggle with high-dimensional data, leading to the "curse of dimensionality," where performance deteriorates and meaningful patterns become harder to identify.
- **Static Segmentation:** Many traditional segmentation techniques do not adapt over time, failing to account for changing customer behaviors or market trends, which can render segments outdated and ineffective.

9. Conclusion

In conclusion, this project on customer segmentation using K-means clustering has demonstrated the significant potential of advanced analytical techniques to unlock valuable insights from e-commerce data. By integrating exploratory data analysis, feature engineering, and dimensionality reduction through PCA, we have established a robust methodology for effectively categorizing customers into meaningful segments. This structured approach not only enhances the accuracy of segmentation but also provides a deeper understanding of customer behaviors, preferences, and motivations, which are critical for businesses seeking to optimize their marketing strategies and enhance customer engagement. Furthermore, the findings from this project underscore the importance of adopting a data-driven mindset in decision-making processes.

As the e-commerce landscape continues to evolve, leveraging sophisticated analytical tools will be essential for businesses to remain competitive and responsive to customer needs. The segmentation model developed in this project serves as a foundation for future initiatives, allowing organizations to refine their strategies based on real-time insights and adapt to changing market dynamics. Ultimately, this work highlights the transformative power of machine learning and data analytics in shaping the future of customer relationship management and driving business success in the digital age.

10. Future Scope

For future enhancements of the customer segmentation model, several avenues can be explored to increase its effectiveness and applicability in the dynamic e-commerce landscape. One significant improvement could involve the integration of additional advanced clustering algorithms, such as DBSCAN or hierarchical clustering, which may offer better performance in identifying non-spherical clusters or accommodating varying densities among customer segments. Additionally, incorporating temporal data analysis could provide insights into how customer behavior evolves over time, allowing for the identification of trends and seasonal patterns that may influence purchasing decisions. Leveraging deep learning techniques, such as autoencoders, could also facilitate more nuanced feature extraction, enhancing the representation of complex customer behaviors and improving clustering outcomes. Furthermore, implementing real-time data processing capabilities would enable businesses to continuously update customer segments based on the latest interactions and transactions, ensuring that marketing strategies remain relevant and effective. Finally, expanding the scope of the analysis to include external factors, such as socio-economic trends and competitive market dynamics, could further enrich the understanding of customer segments and

drive more strategic decision-making. Together, these enhancements could significantly bolster the model's utility and adaptability, empowering businesses to meet the evolving needs of their customers in a rapidly changing digital environment.

11. Appendix

11.1 Source Code

Customer Segmentation.ipynb

```
from flask import Flask, render_template, request, jsonify

import numpy as np

import joblib

import json

import pandas as pd

from sklearn.preprocessing import MinMaxScaler



app = Flask(__name__)



# Load project descriptions

with open("../notebooks/descriptions.json") as f:

    segment_descriptions = json.load(f)
```

```
segment_descriptions = pd.DataFrame(segment_descriptions.values(),
index=segment_descriptions.keys(), columns=["description"])


# Load the model and scaler
model = joblib.load("../models/model.joblib")
scaler = MinMaxScaler()
scaler.min_, scaler.scale_ = np.load('minmax_scaler_params.npy')


# Encoding mappings
encodings = {
    'Married': {'Yes': 1, 'No': 0},
    'Graduated': {'Yes': 1, 'No': 0},
    'Gender': {"Female": 0, "Male": 1},
    'Profession': {'Artist': 0, 'Doctor': 1, 'Engineer': 2, 'Entertainment': 3, 'Executive':
4, 'Healthcare': 5, "Lawyer": 6, "Other": 7},
    'Spending Score': {'Low': 2, 'Average': 0, 'High': 1}
}


# List of features and column names
num_features = ['Family Size', 'Age', 'Work Experience']
cat_features = ['Spending Score', 'Profession', 'Gender', 'Graduated', 'Married']
columns = ['Family_Size', 'Age', 'Work_Experience', 'Spending_Score',
```

```
        'Profession_Artist', 'Profession_Doctor', 'Profession_Engineer',
        'Profession_Entertainment', 'Profession_Executive',
        'Profession_Healthcare', 'Profession_Lawyer', 'Profession_Other',
        'Gender', 'Graduated', 'Ever_Married']

@app.route('/')
def home():
    return render_template('index.html')

@app.route('/predict', methods=['POST'])
def predict():
    # Get input data from the form
    age = request.form['age']
    married = request.form['married']
    work_experience = request.form['work_experience']
    graduated = request.form['graduated']
    work_profession = request.form['profession']
    gender = request.form['gender']
    spending_score = request.form['spending_score']
    family_size = request.form['family_size']
```

```
# Prepare the input dictionary
inputs = {
    'Family Size': family_size,
    'Age': age,
    'Work Experience': work_experience,
    'Spending Score': spending_score,
    'Profession': work_profession,
    'Gender': gender,
    'Graduated': graduated,
    'Married': married
}

# Preprocess numerical inputs
num_inputs = {k: v for k, v in inputs.items() if k in num_features}
num_df = pd.DataFrame.from_dict(num_inputs, orient='index').T
scaled_inputs = scaler.transform(num_df)
num_df = pd.DataFrame(scaled_inputs)

# Process categorical inputs
num_professions = len(set(encodings['Profession'].values()))
num_onehot_encoded_features = 1
```

```
    cat_df = np.zeros((1, len(cat_features) - num_onehot_encoded_features +
num_professions))

    for i, feature in enumerate(cat_features):
        if feature == 'Spending Score':
            cat_df[0, i] = encodings[feature][inputs[feature]]
        elif feature == 'Profession':
            profession = np.zeros(num_professions)
            profession[encodings[feature][inputs[feature]]] = 1
            cat_df[:, i:i+num_professions] = profession.reshape(1, num_professions)
        elif feature in ['Married', 'Graduated', 'Gender']:
            cat_df[0, i+num_professions-num_onehot_encoded_features] =
encodings[feature][inputs[feature]]

    cat_df = pd.DataFrame(cat_df)
    predict_df = pd.concat([num_df, cat_df.add_suffix('_2')], axis=1)

    # Rename columns
    predict_df.columns = columns

    # Make the prediction
    prediction = model.predict(predict_df)
```

```
    # Get the description for the predicted segment
    predicted_segment = prediction[0]
    description = segment_descriptions.loc[predicted_segment]['description']

    # Return the result to the user
    return render_template('result.html', segment=predicted_segment, description=description)

if __name__ == '__main__':
    app.run(debug=True)
```

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Automobile Customer Segmentation</title>
    <style>
        body {
            background-image: url('https://www.corporatevision-news.com/wp-content/uploads/2022/10/Customer-Segmentation.jpg');
```

```
    background-size: cover;
    background-position: center;
    font-family: Arial, sans-serif;
}
.container {
    background-color: rgba(255, 255, 255, 0.8);
    width: 50%;
    margin: 100px auto;
    padding: 20px;
    border-radius: 10px;
    box-shadow: 0px 0px 20px rgba(0, 0, 0, 0.5);
}
h1 {
    text-align: center;
    color: #333;
}
form {
    display: flex;
    flex-direction: column;
    gap: 15px;
}
```

```
label {
    font-size: 18px;
    color: #333;
}
input, select {
    padding: 10px;
    font-size: 16px;
    border: 1px solid #ccc;
    border-radius: 5px;
}
input[type="submit"] {
    background-color: #4CAF50;
    color: white;
    border: none;
    padding: 10px 20px;
    font-size: 18px;
    cursor: pointer;
}
input[type="submit"]:hover {
    background-color: #45a049;
}
```

```
    </style>
</head>
<body>
    <div class="container">
        <h1>Customer Segmentation Prediction</h1>
        <form action="/predict" method="POST">
            <label for="age">Age:</label>
            <input type="number" id="age" name="age"><br><br>

            <label for="married">Have you ever been married?</label>
            <select id="married" name="married">
                <option value="Yes">Yes</option>
                <option value="No">No</option>
            </select><br><br>

            <label for="work_experience">Work Experience:</label>
            <input type="number" id="work_experience"
name="work_experience"><br><br>

            <label for="graduated">Have you ever graduated?</label>
            <select id="graduated" name="graduated">
                <option value="Yes">Yes</option>
```

```
    <option value="No">No</option>
</select><br><br>

<label for="profession">What is your profession?</label>
<select id="profession" name="profession">
    <option value="Artist">Artist</option>
    <option value="Doctor">Doctor</option>
    <option value="Engineer">Engineer</option>
    <option value="Entertainment">Entertainment</option>
    <option value="Executive">Executive</option>
    <option value="Healthcare">Healthcare</option>
    <option value="Lawyer">Lawyer</option>
    <option value="Other">Other</option>
</select><br><br>

<label for="gender">What is your gender?</label>
```

```
        <select id="gender" name="gender">
            <option value="Female">Female</option>
            <option value="Male">Male</option>
        </select><br><br>

        <label for="spending_score">How big of a spender on automobiles are
you?</label>
        <select id="spending_score" name="spending_score">
            <option value="Low">Low</option>
            <option value="Average">Average</option>
            <option value="High">High</option>
        </select><br><br>

        <label for="family_size">Family Size:</label>
        <input type="number" id="family_size" name="family_size"><br><br>

        <input type="submit" value="Submit">
      </form>
   </div>
</body>
</html>
```

1.1 GitHub & Project Demo Link

https://github.com/siva494/Customer-Segmentation-Using-Machine-Learning