KIET
GROUP OF INSTITUTIONS
Connecting Life with Learning

www.kiet.edu
Delhi-NCR, Ghaziabad

Customer Behavior Prediction

A Project Report

Submitted by

Rachit Garg (Roll No -202401100300188)

in partial fulfilment for the award of the degree of

Bachelor of Computer Science and Engineering (Artificial Intelligence),

KIET Group of Institutions, Ghaziabad

Affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow

1. Introduction:

Understanding customer behavior is crucial for businesses aiming to refine their strategies and improve customer satisfaction. This report explores predictive analytics to classify customers into two categories: 'bargain hunters' and 'premium buyers,' based on their purchasing history. Leveraging machine learning techniques such as logistic regression for classification and KMeans clustering for segmentation, the analysis delves into critical metrics like total_spent, avg_purchase_value, and visits_per_month.

To ensure robust results, evaluation metrics like accuracy, precision, and recall are employed, alongside visualization techniques like confusion matrix heatmaps for classification performance and scatterplots for segmentation insights. This report serves as a comprehensive study of applying data-driven methodologies to decode customer preferences, enhance business decision-making, and pave the way for personalized marketing approaches.

2. Methodology:

The approach followed to solve the problem involves:

1. Prepare the Dataset

Ensure the dataset (customer_behavior.csv) is clean and preprocessed. Here, you already have features like total_spent, avg_purchase_value, visits_per_month, and the target label buyer_type. Split the data into training and testing sets for evaluation.

2. Build the Classification Model

You can use a machine learning algorithm like Logistic Regression, Decision Trees, Random Forests, or Support Vector Machines (SVM) to classify customers as 'bargain hunters' or 'premium buyers'. Here's a Python snippet for using Logistic Regression.

3. Segmentation and Clustering

For unsupervised clustering (if you're analyzing customer segments), KMeans clustering can be helpful

3. Code:

The code for this project was written on Google Colab. Here's how you can approach the classification task for predicting customer behavior.

```
# Step 1: Ask the user to upload a CSV file
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
data = pd.read_csv('customer_behavior.csv')

# Encode target variable
data['buyer_type'] = data['buyer_type'].apply(lambda x: 1 if x == 'premium_buyer' else 0)
```

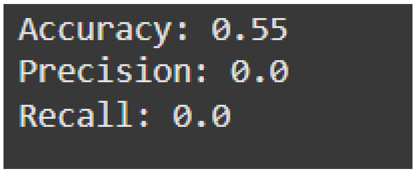
```
# Features and target variable
X = data[['total_spent', 'avg_purchase_value',
'visits_per_month']]
y = data['buyer_type']

# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Train the model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Evaluation Metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
```

```
recall = recall_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)


print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")


# Confusion Matrix Heatmap
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',
xticklabels=['Bargain Hunter', 'Premium Buyer'],
yticklabels=['Bargain Hunter', 'Premium Buyer'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix Heatmap')
plt.show()


from sklearn.cluster import KMeans
# Applying KMeans
kmeans = KMeans(n_clusters=2, random_state=42)
clusters = kmeans.fit_predict(X)
```

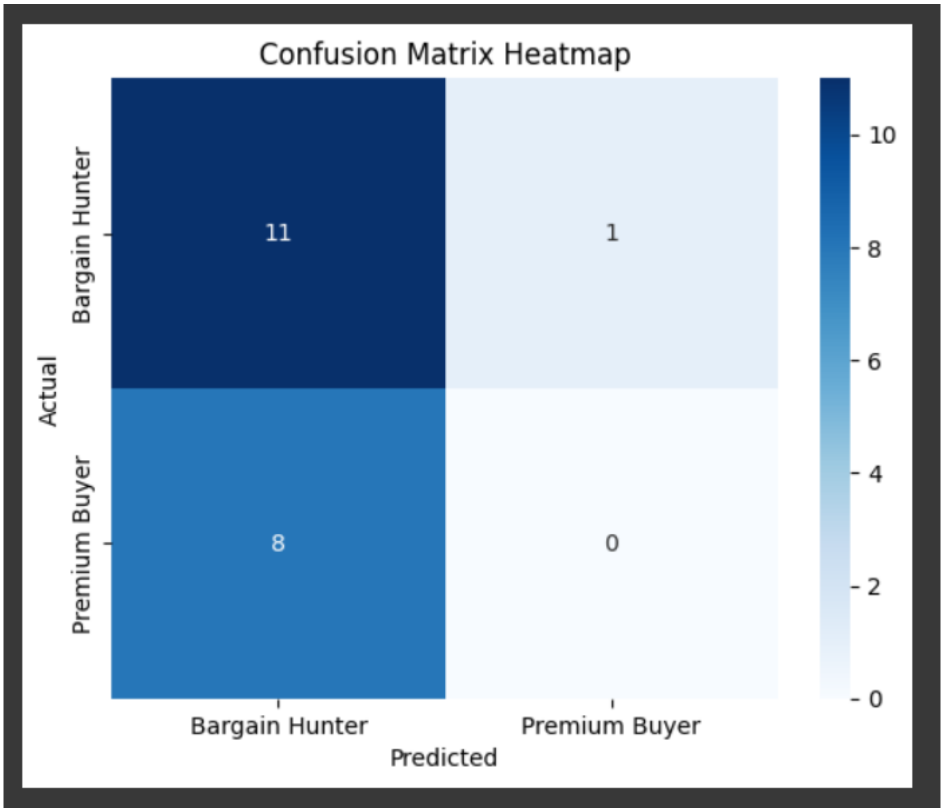
```
# Adding cluster labels to the data
data['Cluster'] = clusters

# Visualizing clusters
sns.scatterplot(data=data, x='total_spent',
y='avg_purchase_value', hue='Cluster', palette='viridis')
plt.title('Customer Segmentation')
plt.show()
```

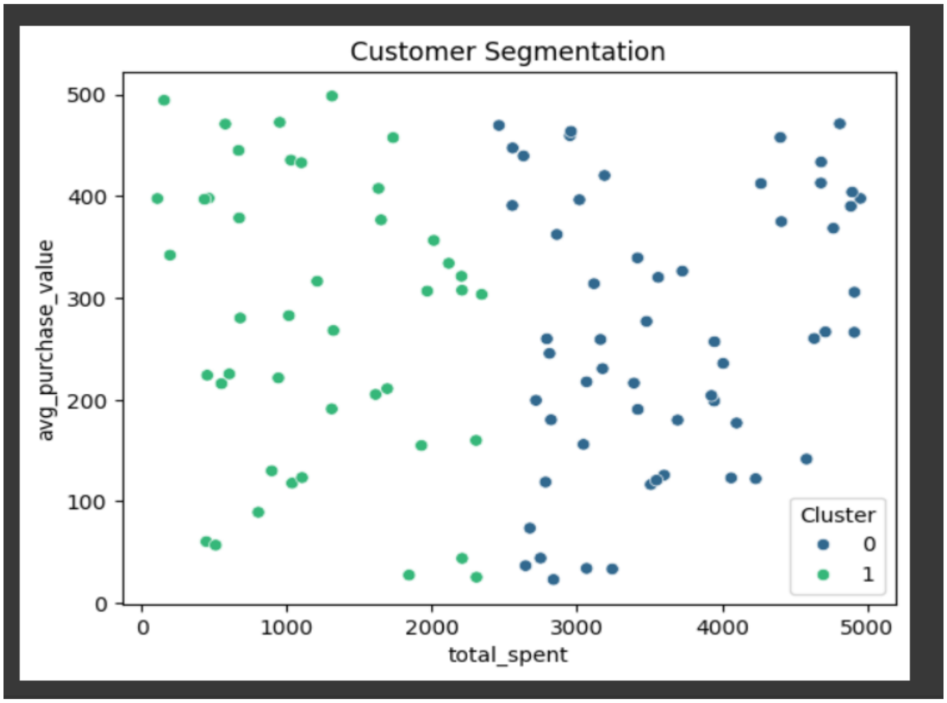
4. Output/Result:

The following results were obtained:



```
Accuracy: 0.55
Precision: 0.0
Recall: 0.0
```





5. References/Credits:

1. Dataset Source: Google Drive link

2. Libraries Used:

- **Pandas**
  Used for data manipulation and analysis, including reading the CSV file, preprocessing data, and exploring features

- **NumPy**
  Supports numerical operations, which are essential for certain calculations and transformations.

- **Scikit-learn**
  Includes tools for machine learning models like Logistic Regression, as well as clustering algorithms like KMeans. It also provides utilities for splitting data, evaluating models (accuracy, precision, recall), and generating confusion matrices.

- **Seaborn**
  Facilitates advanced data visualization, particularly for creating heatmaps of confusion matrices and scatterplots.

- **Matplotlib**
  Used for plotting graphs and visualizing clusters or metrics in combination with Seaborn.

6. Files Uploaded to GitHub:

- Jupyter Notebook (.ipynb)
- PDF Report
- README.md