E-COMMERCE CUSTOMER SEGMENTATION AND PURCHASE PREDICTION

Machine Learning Laboratory Project Report

Submitted by

Kayam Mahasvin Reddy(AP23110010075)

Department of Computer Science and Engineering
SRM University - AP
Amaravati, Andhra Pradesh

Academic Year 2024–2025

SRM University - AP

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project report titled

"E-COMMERCE COSTUMER SEGMENTATION AND PURCHASE PREDICTION**"**

has been completed as part of the

Machine Learning Laboratory

by

Kayam Mahasvin Reddy(AP23110010075)

during Academic Year 2024–2025.

Faculty Signature: _____

Lab Incharge: _____

Place: Amaravati **Date: _____**

Abstract

Customer segmentation and behaviour prediction play a crucial role in designing effective marketing strategies and improving business profitability. This project focuses on analysing customer behaviour using the *Customer Personality Analysis* dataset and developing machine learning models for segmentation, spending prediction, and campaign response prediction. The data undergoes preprocessing that includes handling missing values, feature scaling, one-hot encoding of categorical variables, and feature engineering such as Total Spending and Family Size.

Unsupervised learning is applied using **K-Means clustering**, supported by **PCA** visualization, to segment customers into meaningful groups based on demographic attributes and purchasing behaviour. For supervised learning, **Linear Regression** is used to predict total customer spending, while **Logistic Regression** and **Gradient Boosting** are implemented to predict campaign response. Ensemble models demonstrate superior performance by capturing non-linear customer patterns and delivering higher accuracy.

The project provides actionable insights such as identifying high-value customers, understanding key drivers of campaign acceptance, and optimizing targeted marketing strategies. The results highlight the effectiveness of combining clustering, regression, and classification to support data-driven decision-making in modern customer relationship management.

Contents

CHAPTER 1

Introduction

In today's competitive business environment, understanding customer behaviour has become essential for organizations aiming to improve marketing effectiveness and enhance customer satisfaction. With the rapid growth of digital platforms and e-commerce, companies now collect vast amounts of customer-related data, including demographics, transaction history, purchase frequency, and responses to marketing campaigns. Leveraging this data through machine learning enables businesses to segment customers accurately, predict purchase behaviour, and tailor marketing strategies to maximize return on investment (ROI).

Customer Personality Analysis is a powerful approach that helps identify unique customer groups based on shared characteristics and behavioural patterns. These insights allow companies to differentiate between high-value customers, low-engagement users, and potential churners, enabling personalized communication and targeted marketing campaigns. Machine learning techniques such as clustering, regression, and classification enhance the decision-making process by uncovering hidden patterns, predicting future spending, and estimating the probability of campaign acceptance.

This project integrates multiple machine learning approaches to develop a comprehensive customer analytics system. It performs customer segmentation using K-Means clustering, spending prediction using regression models, and campaign response prediction using classification and ensemble techniques. The goal is to provide actionable insights and demonstrate the effectiveness of machine learning in modern customer relationship management.

This project serves as a practical application of machine learning concepts in the field of environmental science and public health, providing students with hands-on experience in real-world data analysis and predictive modeling.

CHAPTER 2

Objectives

Businesses often struggle to understand diverse customer behaviour and predict how customers will respond to marketing efforts. Without proper segmentation and behaviour prediction, companies may waste resources on ineffective campaigns and fail to identify high-value customers who contribute significantly to revenue. Traditional marketing strategies rely heavily on intuition and generic segmentation methods, which are insufficient in handling large-scale and complex customer datasets.

Objectives of the Project:

(1) **To segment customers into meaningful groups** using K-Means clustering based on their demographics and purchasing behaviour.

(2) **To predict customer spending** using regression models such as Linear Regression and Random Forest Regressor.

(3) **To predict customer response to marketing campaigns** using classification models like Logistic Regression, Random Forest, and Gradient Boosting.

(4) **To evaluate and compare model performance** using metrics such as R^2, MSE, Accuracy, Precision, Recall, and ROC-AUC.

(5) **To generate actionable business insights** that help companies design targeted marketing strategies and improve customer engagement.

CHAPTER 3

Dataset Description

This project uses a real-world **E-commerce Customer Behavior dataset** that contains detailed information about customer demographics, purchasing patterns, spending behavior, and satisfaction levels. The dataset is well-structured and supports both **exploratory data analysis (EDA)** and **machine learning model development** for tasks such as customer segmentation, satisfaction prediction, and purchase behavior analysis.

Dataset Source and Size

- **Filename:** Ecommerce Customer Behavior -Sheet1.csv

- **Total Records:** 12000
- **Number of Features:** 9

The dataset is clean and well-balanced, with no major missing values. It is suitable for direct use in analytics and machine learning experiments without extensive preprocessing.

Key Features in the Dataset

The dataset includes attributes related to **customer demographics, purchasing behavior, discounts, recency of purchase, and satisfaction level**.

Customer Demographics

- **Customer ID:** Unique identifier for each customer.

- **Gender:** Gender of the customer (Male/Female).

- **Age:** Age of the customer.

- **City:** City where the customer resides.

Engagement Information

- **Days Since Last Purchase:** Number of days since the customer last made a purchase.

Purchase and Spending Behavior

- **Total Spend:** Total amount of money spent by the customer.
- **Items Purchased:** Total number of items purchased.
- **Average Rating:** Average rating given by the customer to products.
- **Discount Applied:** Whether a discount was applied (True/False).

Customer Satisfaction

- **Satisfaction Level:** Overall satisfaction status of the customer (Satisfied, Neutral, Unsatisfied).

Target Variable Used

Customer Satisfaction (Multi-Class Classification)

The main target variable used for model training is:

- **Satisfaction Level**, categorized into:
    - High
    - Medium

    - Low

This variable is suitable for **multi-class classification models** such as:

- Logistic Regression
- K-Nearest Neighbors

Features Used for Machine Learning

Selected predictors for training the model include:

- Gender
- Age
- City
- Total Spend
- Items Purchased
- Average Rating
- Discount Applied
- Days Since Last Purchase

Target Variable:

- Satisfaction Level

Importance of the Dataset

This dataset provides a strong foundation for understanding **customer purchasing behavior and satisfaction trends in e-commerce platforms**. The combination of demographic, behavioral, and engagement features helps in:

- Predicting customer satisfaction
- Identifying high-value customers
- Improving marketing strategies
- Optimizing discount policies
- Enhancing customer retention

The dataset supports accurate and meaningful **machine learning models for business decision support systems**.

CHAPTER 4

Methodology

The proposed solution for Customer Personality Analysis focuses on applying machine learning techniques to understand customer behaviour, segment customers, predict their spending, and estimate the likelihood of responding to marketing campaigns. The project integrates unsupervised and supervised learning to generate actionable business insights.

Machine Learning Models Used

1. Logistic Regression

A supervised classification algorithm used for predicting binary outcomes.
It models the probability of customer response using a sigmoid function.
This model is used as a **baseline classifier** to understand how different customer features influence the final prediction outcome.
It is simple, fast, and highly interpretable.

2. Support Vector Machine (SVM)

A powerful supervised learning algorithm used for classification.
It works by finding an optimal hyperplane that best separates different classes with the maximum margin.
SVM performs well in **high-dimensional feature spaces** and is effective when the number of features is large.

3. K-Nearest Neighbors (KNN)

A simple, instance-based supervised learning algorithm.
It classifies a data point based on the majority class among its K nearest neighbors.
This model works well for datasets with **clear cluster boundaries** and helps understand neighborhood-based customer behavior.

4. Naïve Bayes Classifier

A probabilistic classification algorithm based on Bayes' Theorem with an assumption of feature independence.

It is highly **efficient and fast**, even with large datasets.
Naïve Bayes performs well in real-world classification problems and is especially effective when features contribute independently to the target prediction.

Overall Methodology

1. **Data Collection:** The Customer Personality Analysis dataset is collected from Kaggle, containing demographic, behavioural, and spending information.
2. **Feature Engineering:** New features such as Total_Spent, Age, and FamilySize are created, and categorical variables are encoded.
3. **Data Preprocessing:** Missing values are handled, numerical features are scaled, and categorical features are one-hot encoded.
4. **Customer Segmentation:** K-Means clustering is applied to identify customer groups, supported by PCA for visualization.
5. **Data Splitting:** The processed dataset is divided into training and testing sets for both regression and classification tasks.
6. **Model Training:** Regression models and classification models are trained to predict spending and campaign response, respectively.
7. **Model Testing:** The trained models are tested on unseen data to evaluate their predictive performance.
8. **Performance Evaluation:** All models are assessed using metrics such as R^2, MSE, Accuracy, Precision, Recall, and ROC-AUC.
9. **Model Comparison:** The performance of each model is compared to identify the best regression and classification models.
10. **Visualization:** PCA cluster plots, confusion matrices, ROC curves, and actual vs predicted graphs are generated for interpretation.

Pseudo-code

BEGIN

 # STEP 1: LOAD DATA

Load dataset from CSV file

Display first few rows and dataset summary

Remove 'CustomerID' column

 # STEP 2: HANDLE MISSING VALUES

Check for null values in all columns

If missing values exist:

Remove or fill missing entries

 # STEP 3: ENCODING CATEGORICAL VARIABLES

For each categorical column (Gender, City, MembershipType):

Apply LabelEncoder and store the encoder for future use

 # STEP 4: FEATURE SCALING

Select numerical columns: (Age, TotalSpend, ItemsPurchased, AverageRating, DaysSinceLastPurchase)

Apply MinMaxScaler to normalize values

Save the fitted scaler

 # STEP 5: PREPARE TARGET VARIABLE

Convert SatisfactionLevel into integer classes:

Medium → 0

High → 1

Low → 2

 # STEP 6: DATA SPLITTING

Split data into Training set (80%) and Testing set (20%)

 # STEP 7: MODEL TRAINING

Initialize models:

Logistic Regression

Decision Tree Classifier

Support Vector Machine

K-Nearest Neighbors (k = 3)

Naive Bayes

For each model in models:

Train model on training data

Predict on test data

Calculate accuracy

Store accuracy in list

 # STEP 8: SELECT BEST MODEL

Find model with highest accuracy

Set BestModel = model with max accuracy (e.g., KNN or Random Forest)

 # STEP 9: SAVE MODEL & ENCODERS

Save:

BestModel → "final_model.pkl"

Label Encoders → (gender_le, pay_le)

Scaler → "scaler.pkl"

 # STEP 10: PREDICTION FUNCTION FOR NEW CUSTOMER

Define PredictSatisfaction(customer_input):

Apply LabelEncoders to categorical fields

Apply MinMaxScaler to numerical fields

Reorder feature values to match training format

Predict satisfaction class using BestModel

Convert numeric class back to text label

Return predicted Satisfaction Level

END FUNCTION

END

CHAPTER 5

Implementation

The project was implemented using **Python in Google Colab**, leveraging widely used machine learning and data processing libraries. The dataset, containing 350 customer records and 10 features, was loaded from a CSV file. The implementation followed a structured pipeline involving preprocessing, encoding, scaling, model training, evaluation, and saving the final model.

Tools and Environment

- **Platform:** Google Colab
- **Language:** Python
- **Libraries Used:**
  - *Pandas, NumPy* — Data handling
  - *Matplotlib, Seaborn* — Visualizations
  - *Scikit-learn* — Preprocessing & ML models
  - *Joblib* — Saving trained models
  - *JSON* — Exporting categorical values (Cities list)

Preprocessing Steps

1. Removed **Customer ID**, as seen on page 1

   771286b1-ba89-47dd-90dd-1060892…

2. Checked missing values — only **2 missing entries** in *Satisfaction Level* were removed (page 1).
3. Categorical columns (**Gender, Payment Method**) encoded using **LabelEncoder** (page 5).
4. Numerical columns (**Age, Total Spend, Items Purchased, Average Rating, Days Since Last Purchase**) scaled using **MinMaxScaler** (page 4).
5. Encoded target variable *Satisfaction Level* into 3 classes:
   - 0 → Medium
   - 1 → High
   - 2 → Low
     (mapping shown on page 6)

Models Implemented

Your notebook trained **seven models** (page 6–7):

1. Logistic Regression
2. Support Vector Machine (SVM)
3. K-Nearest Neighbours (KNN)
4. Naïve Bayes
5. Linear Regression (not used for accuracy comparison; included for completeness)

Training & Testing

- Dataset split: **80% training (9600 samples)** and **20% testing (2400 samples)** as seen on page 6.
- Accuracy calculated for all models using accuracy_score().

Model Saving

The best-performing model (**Naive Bayes**) was saved using joblib

CHAPTER 6

Results and Discussion

Exploratory Data Analysis (EDA)

The notebook performed extensive EDA:

- **Pair plot visualizations** of numeric features .
- **Correlation heatmap** showing strong relationships such as:
  - Total Spend ↔ Items Purchased (0.97)
  - Total Spend ↔ Average Rating (0.92)
- **Distribution plots** for gender, membership type, and satisfaction levels (page 2).
- **City distribution**, histograms, scatterplots, and boxplots (page 3–4).

These visualizations show that purchasing behaviour is highly influenced by rating, spend, and membership type.

Model Performance Comparison

Model	Accuracy
Logistic Regression	0.960833
Naive Bayes	0.965000
KNN	0.887500
SVM	0.962083

Insights

- KNN works well because features were correctly **scaled using MinMaxScaler**, making distance-based methods effective.
- SVM and Logistic Regression also performed strongly

- Naive bayes worked also good as the Logistic and SVM

Final Model Selection

You selected **KNN (k=3)** as the deployment model, saved as final_model.pkl (page 7).
This aligns with the highest accuracy and simpler interpretability.

Prediction Demo

The notebook includes a prediction function (page 7–8) that:

- Scales numerical values
- Encodes categorical values
- Reorders columns
- Predicts satisfaction level for a new customer

Example inputs:

```
'Age': 32,

  'Gender': 'Male',

  'Annual Income (k$)': 75,

  'Spending Score (1-100)': 68,

  'Purchase Frequency': 12,

  'Payment Method': 'Credit Card',

  'Browsing Time (min)': 35,

  'Total Purchases': 18
```

CHAPTER 7

Conclusion and Future Work

This project successfully implemented a full machine learning workflow to analyse and predict customer satisfaction levels based on demographic and behavioural factors. After data cleaning, preprocessing, and feature transformation, multiple ML models were trained and evaluated. Among all models, some have achieved very good accuracy , indicating strong predictive capability for this dataset.

Extensive EDA revealed important behavioural patterns such as higher spending among gold members and correlations between ratings, spending, and items purchased. The final KNN model was saved along with encoders and scalers, enabling practical deployment for real-time customer satisfaction prediction.

Overall, this study demonstrates the efficiency of machine learning in understanding customer behaviour, supporting better marketing decisions, and improving customer relationship management. Future extensions may include hyperparameter tuning, cross-validation, larger datasets, or integrating clustering for deeper segmentation.

Future Work

- **Integration of Deep Learning Models:**
  Future work can explore neural networks or hybrid deep-learning architectures to improve predictive accuracy and handle larger, more complex customer datasets.
- **Deployment as a Real-Time Application:**
  The trained model can be integrated into a web or mobile dashboard using Streamlit or Flask to provide instant customer satisfaction predictions for business use.
- **Enhanced Customer Analytics with Clustering:**
  Additional unsupervised learning techniques such as DBSCAN or hierarchical clustering can be applied to discover deeper customer segments and improve targeted marketing strategies.

Bibliography

- Kaggle Dataset – *E-commerce Customer Behaviour Dataset* (CSV used in project)
- Scikit-learn Documentation – https://scikit-learn.org
- Pandas Documentation – https://pandas.pydata.org
- Seaborn Documentation – https://seaborn.pydata.org
- Google Colab – Cloud execution environment