**Predictive Customer Behaviour Modelling using AI**

**MCA - IV SEMESTER, FINAL PROJECT REPORT**

*Submitted by*

**SOUMEN CHATTERJEE**

**[EC2432251010407]**

**Under the Guidance of**

**Dr. G. Babu**

(Assistant Professor, Directorate of Online Education)

*in partial fulfilment for the award of the degree of*

**MASTER OF COMPUTER APPLICATIONS**



DIRECTORATE OF ONLINE EDUCATION

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR- 603 203

**JAN 2024 MCA Batch**

**NOV-DEC 2025 Exam (4th SEM)**

DIRECTORATE OF ONLINE EDUCATION

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY** KATTANKULATHUR – 603 203

**BONAFIDE CERTIFICATE**

This Final Project (SEM IV) Report titled **“Predictive Customer Behaviour Modelling using AI”** of **“Soumen Chatterjee [EC2432251010407]”**, who carried out the Semester IV Final Project Work under the supervision of Program Coordinator of Online Education along with the company mentor.

Certified further that to the best of my knowledge, the work reported herein does not form any other project report or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

A close-up of a signature

AI-generated content may be incorrect.

Signature of the Student

(Soumen Chatterjee)

**ACKNOWLEDGEMENTS**

I would like to express my sincere gratitude to **Dr. C. Muthamizhchelvan**, Vice Chancellor, SRM Institute of Science and Technology, for providing the necessary facilities and continued support that enabled the successful completion of this project.

I extend my heartfelt thanks to **Prof. Dr. Manoranjan Pon Ram**, Director, Directorate of Online Education, SRM Institute of Science and Technology, for his valuable guidance and encouragement throughout the course of this Final Project (SEM IV).

I am deeply thankful to **Dr. G. Babu**, Program Coordinator, Directorate of Online Education, for his insightful feedback during project reviews and his unwavering support. I am especially grateful to him for serving as my project guide. His mentorship, encouragement, and the freedom he offered to explore research topics aligned with my interests were instrumental in shaping this project. His dedication and passion for solving real-world problems have been truly inspiring.

I also extend my sincere appreciation to the faculty, staff, and fellow students of the **Directorate of Online Education**, SRM Institute of Science and Technology, for their cooperation and assistance during various stages of the project.

Lastly, I would like to acknowledge the unconditional love, encouragement, and support of my **parents, family members, and friends**, without whom this journey would not have been possible.

**SOUMEN CHATTERJEE**

Table

**Table of Contents**

[1. Abstract 5](#_Toc213630325)

[2. Introduction 6](#_Toc213630326)

[3. System Analysis 7](#_Toc213630327)

[4. Analysis and Requirements 10](#_Toc213630328)

[5. Problem Description / Module Description 11](#_Toc213630329)

[6. Design 13](#_Toc213630330)

[6. a) System Design 13](#_Toc213630331)

[6. b) UML Diagrams 14](#_Toc213630332)

[6. c) Database Design 17](#_Toc213630333)

[7. Implementation 19](#_Toc213630334)

[7.1. Implementation Of Additional Module(s) 34](#_Toc213630335)

[8. Testing 36](#_Toc213630336)

[9. Output Screens 41](#_Toc213630337)

[10. Tools and Technologies 47](#_Toc213630338)

[11. Conclusion 49](#_Toc213630339)

[12. Appendices 51](#_Toc213630340)

[13. References / Bibliography 55](#_Toc213630341)

1. Abstract

Predictive Customer Behaviour Modelling using AI is a data-driven approach designed to understand consumer behaviour, preferences, and purchasing trends. The project focuses on analysing customer data to segment consumers into distinct groups, enabling businesses to optimize targeted strategies. By applying machine learning techniques such as clustering (unsupervised learning) and classification (supervised learning), the study extracts valuable insights from customer demographics and purchasing behavior.

The major accomplishments of this project are as follows:

* **Data preprocessing**, including data cleaning, handling missing values, and feature engineering.
* **Unsupervised learning** implementation using K-Means clustering to group customers into meaningful segments.
* **Supervised learning techniques’** application such as Decision Tree, K-Nearest Neighbours (KNN), and Random Forest to predict customer responses.
* **Visualization and analysis** of critical customer attributes, including income, spending behaviour, and family structure.

My key contributions involved data preparation, model selection, and interpretation of results. The insights generated support targeted marketing efforts and enhance customer relationship management strategies.

2. Introduction

* **Background**

Businesses increasingly rely on data-driven decision-making to understand their customers better. Predictive Customer Behaviour Modelling using AIplays a crucial role in personalizing marketing strategies, optimizing customer engagement, and improving product recommendations.

* **Problem Statement**

Companies often struggle with segmenting their customers effectively, leading to inefficient marketing campaigns. Instead of marketing a product to the entire customer base, a company can analyse which customer segment is most likely to buy the product and target them specifically.

* **Development Process**

The project follows a structured development approach:

1. **Data Collection & Preprocessing**: Cleaning data, handling missing values, and engineering useful features.
2. **Exploratory Data Analysis (EDA)**: Understanding data distribution and identifying patterns.
3. **Unsupervised Learning (Clustering)**: Using K-Means clustering to segment customers.
4. **Supervised Learning (Classification)**: Using decision trees, KNN, and random forests to predict customer responses.
5. **Model Evaluation & Insights**: Assessing model performance and deriving business insights.

3. System Analysis

System analysis represents a fundamental stage in the development of any software or data-driven application. This phase focuses on evaluating the current operational environment, identifying the shortcomings of existing methodologies, and formulating an enhanced solution through the integration of advanced technologies and innovative strategies. The analysis encompasses a review of the existing system (where applicable), the design of the proposed framework for customer personality analysis, and a feasibility assessment to ensure the practicality and sustainability of the project.

**A. Existing System**

Traditionally, organizations have depended on basic customer profiling techniques that utilize a limited set of demographic variables such as age, gender, and geographic location. These conventional methods are often manual or rule-based, offering insufficient depth and precision to capture the complexity of customer behaviour. The primary limitations of such systems include:

* **No Personalization**: Marketing initiatives tend to be generic rather than customized for distinct customer segments.
* **Inefficient Targeting**: In the absence of advanced insights, resources are frequently allocated to customers with low engagement potential.
* **Data Underutilization**: Although businesses gather substantial customer data, they often fail to leverage advanced analytical models for informed decision-making.
* **No Predictive Capability**: Traditional frameworks lack the ability to forecast customer responses to campaigns or predict purchase likelihood.

**B. Proposed System**

The proposed system, titled *Predictive Customer Behavior Modeling using Artificial Intelligence*, is developed to address the shortcomings of traditional approaches by employing advanced data science methodologies. This framework adopts a data-driven strategy for customer segmentation and behavioural forecasting, enabling more accurate and actionable insights.

**Key Features of the Proposed System:**

* **Data-Driven Insights**: Utilizes historical customer information, including expenditure patterns, demographic attributes, and digital interaction data, to generate actionable insights.
* **Customer Segmentation**: Employs clustering algorithms such as *K-Means* to categorize customers into distinct groups based on behavioural and preference similarities.
* **Marketing Response Prediction**: Implements supervised machine learning models, including *Random Forest*, to estimate customer responsiveness to promotional campaigns.
* **Feature Engineering**: Constructs derived variables such as *Age*, *Total\_Spent*, and *Family\_Size* to enhance the analytical depth and predictive accuracy of the model.
* **Visualization**: Delivers graphical representations to facilitate intuitive interpretation of complex patterns and interrelationships within the data.
* **Model Evaluation**: Assesses predictive performance using metrics such as accuracy scores to ensure reliability and robustness of the results.

**Benefits:**

* **Facilitates Personalized Marketing:** Enables the design of marketing strategies tailored to individual customer profiles.
* **Enhances Customer Engagement:** Strengthens interaction and long-term relationships with customers through targeted initiatives.
* **Optimizes Marketing Expenditure:** Reduces unnecessary costs by focusing resources on high-potential customer segments.
* **Improves Conversion Rates:** Increases the likelihood of transforming prospects into active customers through data-driven targeting.
* **Supports Strategic Decision-Making:** Provides actionable insights that inform and refine organizational marketing strategies.

**C. Feasibility Study**

To ensure the practicality and success of the project, a feasibility study was conducted in the following areas:

| **Feasibility Type** | **Assessment** |
| --- | --- |
| **Technical Feasibility** | ✅ The tools and technologies used (Python, Jupyter Notebook, scikit-learn, etc.) are well-established and supported. The project was successfully implemented on standard hardware without requiring high-end computational resources. |
| **Operational Feasibility** | ✅ The solution is user-friendly, modular, and adaptable. Business users and analysts can adopt this model for campaign optimization with minimal training. The outputs are easy to interpret via visualizations and performance metrics. |
| **Economic Feasibility** | ✅ As the project uses open-source tools, the cost of development is minimal. Implementation can lead to higher ROI by optimizing marketing strategies and improving customer targeting. |
| **Schedule Feasibility** | ✅ The project was completed within the academic timeline. A well-structured plan with clear milestones ensured timely completion of each phase—data cleaning, modeling, and evaluation. |

**Conclusion of System Analysis**

The analysis clearly shows that the proposed machine learning-based system significantly improves upon the traditional approach to customer profiling. With minimal investment and strong analytical capabilities, this system is feasible, scalable, and aligns well with modern business intelligence needs.

4. Analysis and Requirements

**Problem Analysis**

The primary challenge in this project is to segment customers based on their demographic and purchasing data. The dataset consists of 29 features, including customer age, income, education, marital status, and spending behaviour across various product categories.

**UML Analysis Model**

The analysis can be represented using the following UML models:

* **Use Case Diagram**: Represents interactions between the system and different user roles (e.g., Data Analyst, Business Manager).
* **Activity Diagram**: Shows the step-by-step process of data preprocessing, clustering, and classification.
* **Class Diagram**: Defines key entities such as Customer, Purchase History, and Segmentation Model.

**System-Level and Software-Level Requirements**

* **System Requirements**
  + Python environment (Jupyter Notebook, Anaconda)
  + Libraries: pandas, NumPy, seaborn, matplotlib, scikit-learn
  + Computational resources for machine learning processing
* **Software Requirements**
  + Data preprocessing module
  + Clustering module (K-Means)
  + Classification module (Decision Tree, KNN, Random Forest)
  + Visualization module for insights

5. Problem Description / Module Description

The project consists of the following key modules:

* 1. **Data Preprocessing**
* Importing libraries and dataset
* Handling missing values (e.g., replacing missing income values with mean)
* Feature engineering (creating new features like Age, Total\_Spent, Family\_Size)
* Encoding categorical variables
  1. **Exploratory Data Analysis (EDA)**
* Understanding data distribution using histograms and scatter plots
* Visualizing customer spending behaviour
* Analysing correlations using heatmaps
  1. **Unsupervised Learning: Clustering**
* Applying K-Means clustering to segment customers
* Finding the optimal number of clusters using the Elbow Method
* Visualizing clusters based on income and spending behaviour
  1. **Supervised Learning: Classification**
* Training machine learning models to predict customer response
* Implementing Decision Tree, KNN, and Random Forest classifiers
* Comparing model performance and accuracy
  1. **Model Evaluation and Business Insights**
* Evaluating classification models using accuracy scores
* Understanding customer segments for targeted marketing strategies

6. Design

6. a) System Design

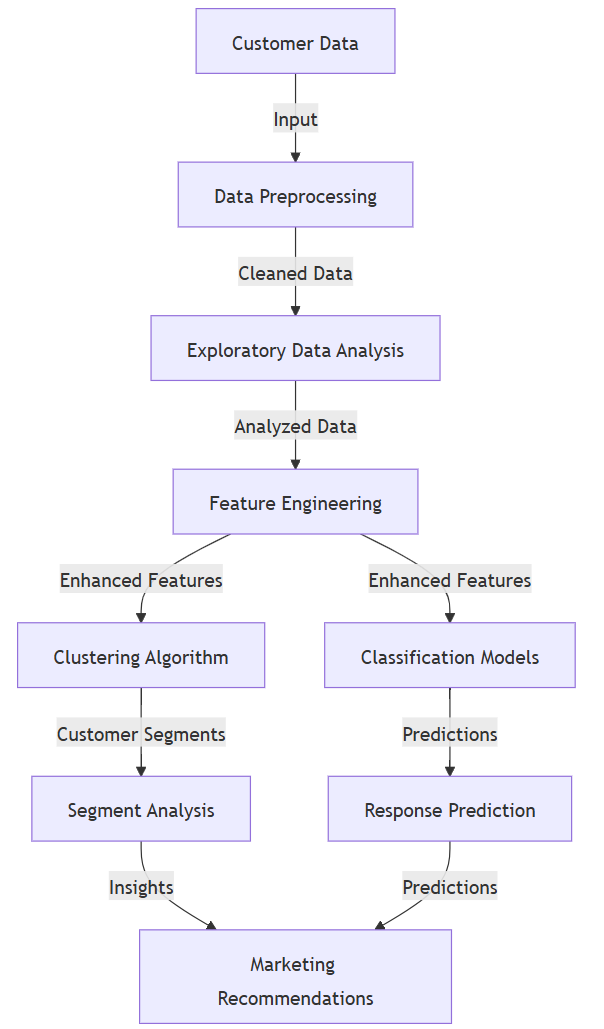
**Data Flow Diagram (DFD):**

Level 0 DFD - Predictive Customer Behavior Modeling using AI

A diagram of customer data

AI-generated content may be incorrect.

Level 1 DFD - Predictive Customer Behaviour Modelling using AISystem



6. b) UML Diagrams

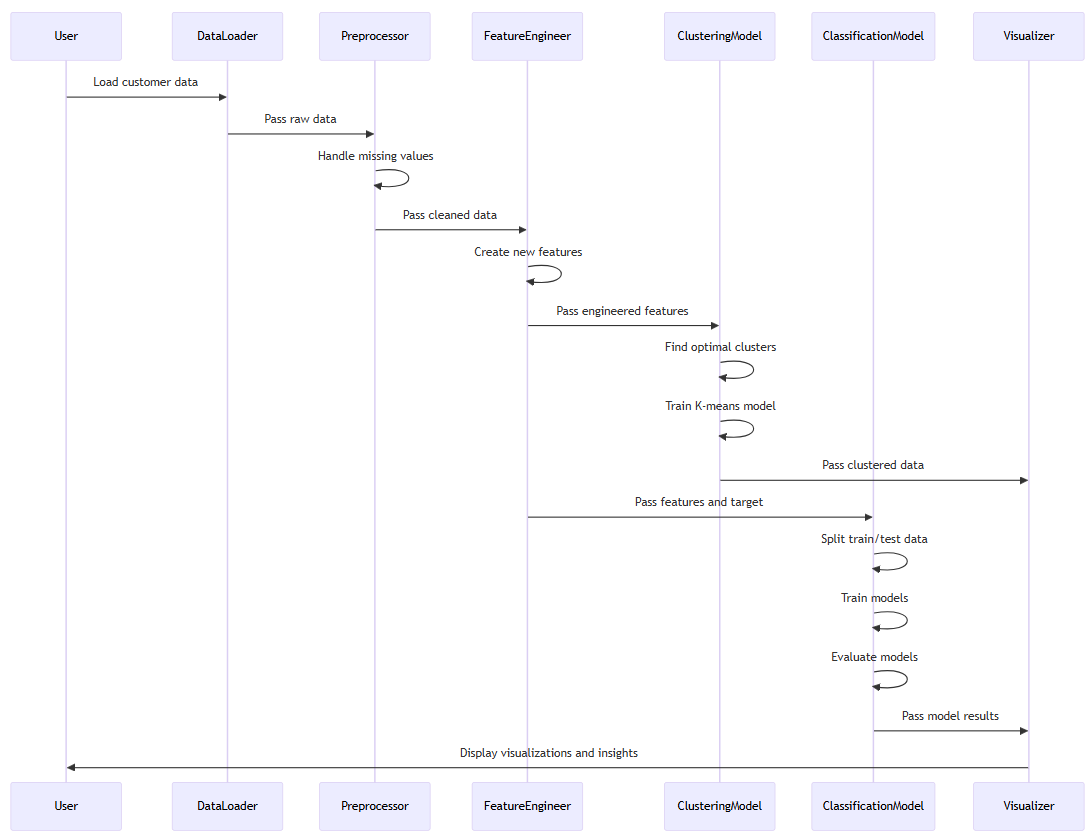
**Class Diagram** - Predictive Customer Behavior Modeling using AI

A diagram of a data processing process

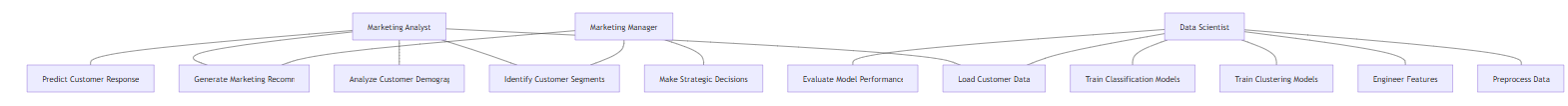
AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

**Sequence Diagram** - Predictive Customer Behaviour Modelling using AI



**Use Case Diagram** - Predictive Customer Behavior Modeling using AI



Object Diagram

Object Diagram - Customer Segments

A screenshot of a computer screen

AI-generated content may be incorrect.

**Control Flow Diagram** - Predictive Customer Behavior Modeling using AI

A diagram of a company

AI-generated content may be incorrect.

6. c) Database Design

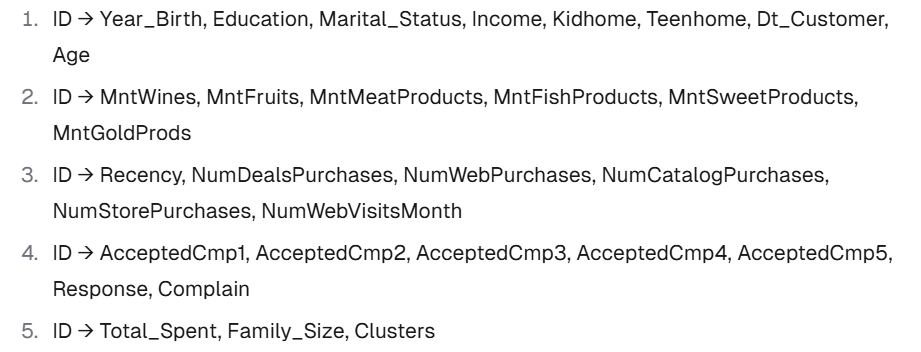
**E-R Diagram** - Predictive Customer Behavior Modeling using AI

A screenshot of a computer

AI-generated content may be incorrect.

**Functional Dependencies and Normalization**

**Functional Dependencies:**



**Normalization Process:**

**A. First Normal Form (1NF):**

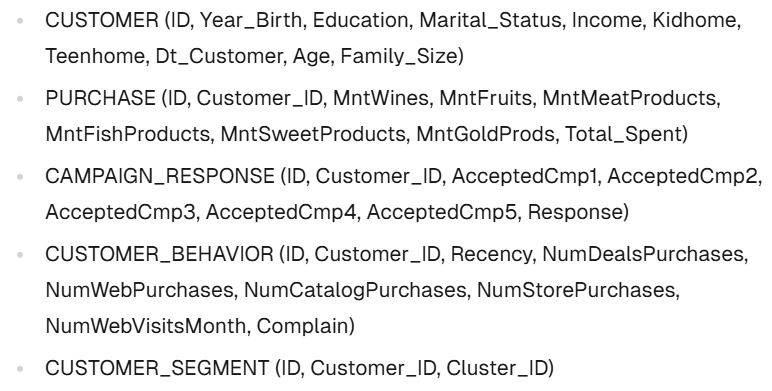
1. All attributes contain atomic values
2. No repeating groups
3. Primary key identified (ID)

**B. Second Normal Form (2NF):**

1. 1. Already in 1NF
2. 2. No partial dependencies (all attributes depend on the entire primary key)

**C. Third Normal Form (3NF):**

1. Already in 2NF
2. No transitive dependencies
3. Decomposed into:



7. Implementation

**Implementation Approach**

The project was implemented in a structured, modular fashion using Python within the Jupyter Notebook environment. Its primary objective was to perform customer data analysis through both supervised and unsupervised machine learning approaches, ensuring a clear distinction between the stages of data preprocessing, model development, and performance evaluation.

**The stages of implementation included**:

* Data Ingestion
* Data Cleaning and Preprocessing
* Feature Engineering
* Data Visualization
* Modeling (Clustering and Classification)
* Evaluation and Interpretation of Results

Each phase utilizes well-structured, reusable code segments designed for clarity and adaptability, enabling easy modifications and experimentation throughout the process.

**Software Reuse and Libraries Used**

The project extensively leveraged software reuse by incorporating widely adopted open-source libraries, which helped minimize development time while enhancing overall reliability. The primary Python libraries utilized included:

| **Library** | **Purpose** |
| --- | --- |
| pandas | Data manipulation and tabular data processing |
| NumPy | Numerical operations and statistical functions |
| matplotlib | Data visualization through plots and charts |
| seaborn | Enhanced statistical data visualization |
| scikit-learn | Machine Learning algorithms for clustering, classification, and preprocessing |
| yellowbrick | Visual analysis of ML model performance (e.g., Elbow method for clustering) |

These libraries adhere to established industry standards and enjoy broad adoption, ensuring that the codebase remains scalable and easy to maintain.

**Special Tools Used**

| **Tool** | **Usage** |
| --- | --- |
| Jupyter Notebook | Main IDE used for developing, testing, visualizing, and documenting the project |
| Yellowbrick | Used specifically for KElbowVisualizer, which helps to find the optimal number of clusters (k) |
| scikit-learn Models | Used for KMeans Clustering, Decision Tree, KNN, and Random Forest classifiers |

Leveraging Jupyter Notebook offered an interactive environment that facilitated iterative code development and debugging, enabled visualization of intermediate outputs, and allowed detailed step-by-step documentation using Markdown.

**Design Patterns and Coding Techniques**

Although the project does not follow an object-oriented paradigm, several fundamental design principles and patterns were incorporated:

* **Modularity:** Each stage of the analysis—such as data loading, cleaning, visualization, and modeling—was implemented in distinct code blocks, adhering to the Separation of Concerns principle.
* **DRY (Don’t Repeat Yourself):** Common operations like aggregations and visualizations were structured for reuse, minimizing redundancy.
* **Reusable Functions (Future Enhancement):** The project can be further improved by converting repetitive tasks (e.g., plotting or model evaluation) into callable functions or class methods.
* **Encapsulation of Data Transformations:** Preprocessing activities, including label encoding, handling missing values, and feature engineering, were encapsulated before passing data to machine learning models.

**Data Transformation and Preprocessing Techniques**

Specialized coding and preprocessing techniques applied in the project included:

* **Handling Missing Data:** Missing income values were imputed using the mean to retain records without discarding rows.
* **Feature Engineering:**
  + Age was computed from the *Year\_Birth* attribute.
  + *Total\_Spent* was derived by summing expenditures across all product categories.
  + *Family\_Size* was calculated by combining *Kidhome*, *Teenhome*, and marital status information.
* **Categorical Encoding:** Education levels were transformed into numeric form using LabelEncoder from sklearn.preprocessing.
* **Feature Scaling (Advanced Option):** StandardScaler was employed to normalize feature ranges, which is advantageous for algorithms sensitive to data magnitude.

**Model Implementation Summary**

 **Unsupervised Learning (K-Means Clustering):**

* The optimal number of clusters (k = 4) was identified using the Elbow Method, implemented via Yellowbrick’s KElbowVisualizer.
* K-Means was then applied to segment customers based on demographic and behavioral attributes.

 **Supervised Learning (Classification Models):**

* Three classifiers were employed: Decision Tree, K-Nearest Neighbors (KNN), and Random Forest.
* Among these, the Random Forest model achieved the highest accuracy of approximately **90.4%**, demonstrating strong predictive capability for customer response.

**Summary**

The implementation adopted a structured, modular, and adaptable approach that delivered:

* **Efficient development** through Python and widely used industry-standard libraries
* **Reliable outcomes** using well-tested machine learning algorithms
* **Clear, reproducible, and extensible code** organized within a Jupyter Notebook environment

This robust foundation provides scope for future enhancements, such as integrating additional models, incorporating real-time data streams, or deploying the solution as an API or interactive dashboard.

**Code Modules and Functionality**

**Module 1: Data Loading and Exploration**

A screenshot of a computer code

AI-generated content may be incorrect.

**Functionality**: This module begins by importing the necessary libraries and loading the dataset. It then examines the fundamental structure of the data, including its dimensions, column details, and associated data types.

**Input**: Marketing campaign CSV file

**Output**: DataFrame object with loaded data and basic information about the dataset

**Module 2: Data Analysis**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Functionality**: This module focuses on exploratory data analysis (EDA) of the dataset. It identifies and addresses missing values, engineers new features, and visualizes key aspects such as age distribution, education levels, marital status, income ranges, and family composition. Additionally, it examines variable correlations and investigates the relationship between income and spending patterns.

**Input**: DataFrame with customer data

**Output**: Visualizations and insights about customer demographics and behaviour

**Module 3: Data Cleaning and Feature Engineering**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**Functionality**: This module handles data cleaning by addressing missing values and performs feature engineering to create new attributes such as *Total\_Spent* and *Family\_Size*. It also encodes categorical variables and prepares the dataset for scaling to ensure compatibility with machine learning algorithms.

**Input**: DataFrame with raw customer data

**Output**: Cleaned DataFrame with engineered features

**Module 4: Clustering (Unsupervised Learning)**

A screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

**Functionality**: This module applies unsupervised learning through K-Means clustering. It determines the optimal number of clusters using the Elbow Method, executes K-Means with the selected cluster count, and visualizes the resulting customer segments.

**Input**: Cleaned DataFrame with engineered features

**Output**: DataFrame with cluster assignments and visualizations of the clusters

**Module 5: Classification (Supervised Learning)**

**A screenshot of a computer code

AI-generated content may be incorrect.** **A screenshot of a computer code

AI-generated content may be incorrect.**

**Functionality**: This module focuses on supervised learning using multiple classification algorithms. It begins by defining the dependent and independent variables, splits the dataset into training and testing subsets, and then trains and evaluates three classifiers: Decision Tree, K-Nearest Neighbors (KNN), and Random Forest.

**Input**: DataFrame with features and target variable

**Output**: Trained classification models and their performance metrics

**Database Tables**

**Database Table Explanation -** CustomerProfile

The dataset comprises detailed records of customer demographics, lifestyle attributes, purchasing patterns, and responses to marketing campaigns. Each entry corresponds to an individual customer.

Structure of the CustomerProfile Table

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| ID | Integer | Unique identifier for each customer |
| Year\_Birth | Integer | Year the customer was born |
| Education | Categorical | Education level (e.g., Graduation, PhD, Master) |
| Marital\_Status | Categorical | Marital status (e.g., Married, Single, Divorced) |
| Income | Float | Annual income of the customer |
| Kidhome | Integer | Number of children living at home |
| Teenhome | Integer | Number of teenagers living at home |
| Dt\_Customer | Date | Date the customer enrolled with the company |
| Recency | Integer | Number of days since last purchase |
| MntWines | Integer | Amount spent on wine products |
| MntFruits | Integer | Amount spent on fruit products |
| MntMeatProducts | Integer | Amount spent on meat products |
| MntFishProducts | Integer | Amount spent on fish products |
| MntSweetProducts | Integer | Amount spent on sweet products |
| MntGoldProds | Integer | Amount spent on gold products |
| NumDealsPurchases | Integer | Number of purchases made using a discount deal |
| NumWebPurchases | Integer | Number of purchases made via the company website |
| NumCatalogPurchases | Integer | Number of purchases made using a catalog |
| NumStorePurchases | Integer | Number of purchases made in a physical store |
| NumWebVisitsMonth | Integer | Number of visits to the website in the last month |
| AcceptedCmp1 to AcceptedCmp5 | Binary | Indicates if the customer accepted each of 5 previous marketing campaigns |
| Response | Binary | Indicates if the customer accepted the last campaign |
| Complain | Binary | Indicates if the customer complained in the last 2 years |
| Z\_CostContact | Constant | Cost of customer contact (constant for all entries) |
| Z\_Revenue | Constant | Revenue from customer contact (constant for all entries) |
| Age | Integer | Derived field: Customer's age |
| Total\_Spent | Integer | Derived field: Total amount spent across product categories |
| Family\_Size | Integer | Derived field: Total number of family members (self + kids/teens + partner) |
| Education (encoded) | Integer | Label-encoded version of Education |
| Clusters | Integer | Cluster ID assigned after KMeans clustering |

**Notes on Derived Fields**

* **Age:** Calculated as 2022 - Year\_Birth.
* **Total\_Spent:** Derived by summing all product spending columns.
* **Family\_Size:** Computed by adding *Kidhome*, *Teenhome*, and the inferred relationship count.
* **Clusters:** Represent customer segments obtained through K-Means clustering.
* **Response:** Serves as the target variable for classification in supervised learning models.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a white box

AI-generated content may be incorrect.

A screenshot of a data type

AI-generated content may be incorrect. A screenshot of a phone

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a phone

AI-generated content may be incorrect.

A screenshot of a data type

AI-generated content may be incorrect. A screenshot of a phone

AI-generated content may be incorrect.

A white screen shot of a document

AI-generated content may be incorrect.

7.1. Implementation Of Additional Module(s)

Other technologies/analytics can be used to make this project more useful for marketing campaigns, segmenting customers for better targeting –

**A. Market Basket Analysis (Association Rule Mining)**

* **What:** Discover which products are frequently bought together by the customer segments.
* **How:** Usage of the Apriori algorithm or FP-Growth (available in mlxtend or apyori Python packages).
* **Value:** Helps marketing teams design cross-selling strategies and personalized offers.

**B. Customer Lifetime Value (CLV) Prediction**

* **What:** Estimate the future value each customer brings to the company.
* **How:** Usage of regression models or probabilistic models (e.g., BG/NBD, Gamma-Gamma).
* **Value:** Enables prioritization of high-value customers for retention and upselling.

**C. Market Segmentation with External Data**

* **What:** Enrich the customer clusters by integrating external market research data (e.g., industry trends, competitor pricing, regional demographics).
* **How:** Use public datasets (from Kaggle, government sources, or Statista) and merge with the clusters for deeper insights.
* **Value:** Shows how your company’s customer base compares to the broader market.

**D. Churn Prediction Module**

* **What:** Predict which customers are likely to stop buying or engaging.
* **How:** Usage of classification models (logistic regression, Random Forest, XGBoost) with features like Recency, Frequency, Monetary value (RFM).
* **Value:** Supports proactive retention campaigns.

**E. Sentiment Analysis on Customer Feedback (if available)**

* **What:** Analyze customer reviews or feedback for sentiment trends.
* **How:** Use NLP techniques (VADER, TextBlob, or transformer models).
* **Value:** Adds qualitative market research to existing quantitative analysis.

Based on the current data what we have –

 Demographics (age, education, marital status, income, etc.)

 Household info (kids, teens at home)

 Detailed product spending (wines, fruits, meat, fish, sweets, gold)

 Purchase channels (web, catalog, store)

 Campaign responses (accepted campaigns, response to last campaign)

 Recency, frequency, and engagement metrics

**Customer Lifetime Value (CLV)** Prediction will be the best additional analytical module we can have in this project.

**Why CLV?**

* **Directly uses the available data:** Necessary all the features needed (spending, frequency, recency, demographics) are available.
* **No need for external data:** We can calculate and model CLV with just the current dataset.
* **Business impact:** CLV is a key metric for marketing and customer management, showing the ability to connect analytics to real business value.

**What Would This Module Include?**

1. **Feature Engineering:** Calculate total spend, average spend per purchase, frequency, recency, etc.
2. **CLV Calculation:** Use formulas such as:



(We can estimate "Customer Lifespan" as the time between first and last purchase, or use Recency as a proxy.)

1. **Segmentation:** Group customers by predicted CLV (e.g., high, medium, low value).
2. **Visualization:** Show CLV distribution, segment profiles, and actionable insights.
3. **(Optional) Predictive Modeling:** Usage of regression or classification to predict which features drive high CLV.

**Why Not Other Modules?**

* **Market Basket Analysis:** The data is not transactional (no product-level basket per order), so this is not feasible.
* **Churn Prediction:** The dataset doesn’t have explicit churn labels or time-series engagement data.
* **Sentiment Analysis:** No text data or customer feedback in the dataset.
* **External Market Segmentation:** No availability of external market data.

**How to Add the CLV Module**

**Section Title:** Customer Lifetime Value (CLV) Prediction and Segmentation

**Steps:**

1. Calculate total spend and frequency for each customer.
2. Estimate customer lifespan (if possible, using Dt\_Customer and Recency).
3. Compute CLV for each customer.
4. Segment customers by CLV (e.g., quartiles or k-means).
5. Visualize and interpret results.
6. (Optional) Build a regression model to predict CLV from demographics and behavior.

8. Testing

Testing is essential for validating the correctness, accuracy, and performance of the implemented system. In this project, which emphasizes data analysis and machine learning, the testing strategy ensures that:

* Data preprocessing and transformations are performed accurately.
* Machine learning models function as intended.
* Predictions are reliable and aligned with project objectives.
* Code components generate valid and interpretable outputs.

**Testing Approach**

Given the scope of the project, the following testing methods were employed:

* **Unit Testing:** Validated individual functions such as data cleaning, feature engineering, encoding, and model training to ensure they operate correctly.
* **Data Validation Testing:** Confirmed that data loading, handling of missing values, and transformations maintain data integrity.
* **Functional Testing:** Verified that the entire pipeline—from raw data through visualization, modeling, and prediction—executes as intended.
* **Model Evaluation Testing:** Assessed classification model accuracy by comparing predicted results with actual values using metrics such as accuracy score.

**Lessons Learnt from Testing**

* Conducting early tests on preprocessing steps helps prevent downstream issues during model training.
* Data imbalance and feature skew can impact model performance, making it crucial to validate assumptions through visualizations.
* Regular checks for data types and missing values are essential components of a robust ML pipeline.
* Evaluating multiple algorithms revealed Random Forest as the most reliable and high-performing model.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a test

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

9. Output Screens

Data Exploration and Analysis

Figure 1: Age distribution of customers

A graph of a red line

AI-generated content may be incorrect.

Figure 2: Pie chart showing education distribution

A pie chart with different colored circles

AI-generated content may be incorrect.

Figure 3: Count plot of marital status

A graph of different colored bars

AI-generated content may be incorrect.

Figure 4: Distribution of customer income

A graph of income

AI-generated content may be incorrect.

Figure 5: Correlation heatmap of numerical variables

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 6: Scatter plots showing relationship between income and spending categories

A group of blue dots

AI-generated content may be incorrect.

Figure 7: Elbow method for determining optimal number of clusters

A graph with a line and a line

AI-generated content may be incorrect.

Figure 8: Distribution of customers across clusters

A graph with different colored bars

AI-generated content may be incorrect.

Figure 9: Scatter plot showing cluster profiles based on income and total spending

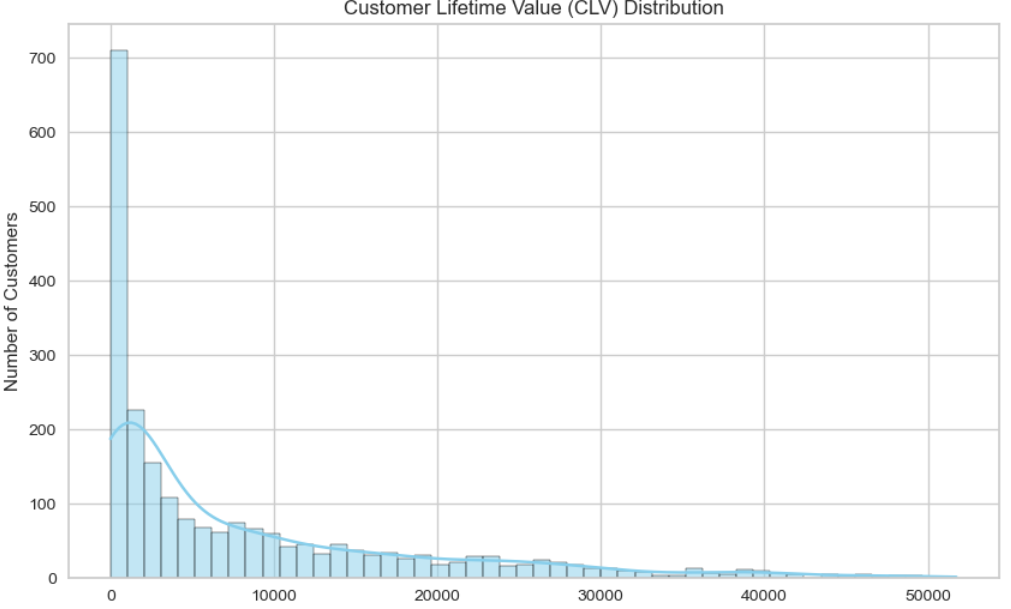
A graph showing a number of colored dots

AI-generated content may be incorrect.

**Classification Results**

A screenshot of a computer

AI-generated content may be incorrect.



A chart of different colored bars

AI-generated content may be incorrect.

10. Tools and Technologies

**Programming Language:**

- Python: Served as the core programming language for tasks such as data analysis, preprocessing, visualization, and machine learning modeling.

**Data Analysis and Manipulation:**

- Pandas: Utilized for efficient data manipulation and analysis, offering powerful data structures and operations to handle numerical tables and time-series data.

- NumPy: Utilized for high-performance numerical computations, providing efficient handling of large, multi-dimensional arrays and matrices.

**Data Visualization:**

- Matplotlib: Used for generating static, animated, and interactive visualizations within Python.

- Seaborn: Built on top of Matplotlib, it is used to create visually appealing and informative statistical graphics.

- Yellowbrick: Designed for machine learning visualizations, particularly useful for applying the Elbow Method in clustering analysis.

**Machine Learning:**

- Scikit-learn: Used for machine learning algorithms implementation, including:

- Preprocessing tools (LabelEncoder, StandardScaler)

- Clustering algorithms (KMeans, AgglomerativeClustering)

- Classification algorithms (DecisionTreeClassifier, KNeighborsClassifier, RandomForestClassifier)

- Model evaluation metrics

- Train-test splitting functionality

**Development Environment**

- Jupyter Notebook: Interactive computing environment used for developing and documenting the analysis.

- Anaconda: Distribution of Python used for scientific computing, which includes many of the packages used in this project.

**Version Control**

- Git: Used for version control and collaboration.

**Database**

- CSV: The data was stored in CSV format, which was processed using Pandas.

**Methodologies**

- Exploratory Data Analysis (EDA): Used to analyze and investigate data sets and summarize their main characteristics.

- Feature Engineering: Process of using domain knowledge to extract features from raw data.

- Unsupervised Learning: Used K-means clustering to segment customers without labelled data.

- Supervised Learning: Used classification algorithms to predict customer response to campaigns.

- Cross-validation: Used to evaluate model performance and prevent overfitting.

This comprehensive set of tools and technologies enabled efficient data processing, insightful analysis, and effective modelling for the Predictive Customer Behaviour Modelling using AI project.

11. Conclusion

**Summary and Key Achievements**

The central aim of this project, titled *Predictive Customer Behavior Modeling using AI*, is to derive actionable insights from customer datasets and uncover trends that support strategic segmentation and personalized targeting. By harnessing Machine Learning techniques, the study emphasizes the examination of customer profiles, behavioral patterns, and marketing interaction data to interpret personality characteristics and classify customers into meaningful groups.

The project was carried out in the following structured manner:

* **Data Preprocessing and Cleaning**: Addressed missing entries—particularly in the Income field—engineered new features such as Age, Total\_Spent, and Family\_Size, and applied label encoding to convert categorical variables into numerical format.
* **Exploratory Data Analysis (EDA)**: Explored and illustrated key data relationships, including age distribution, income disparities, household composition, and consumer spending patterns through visual analytics.
* **Unsupervised Learning – Clustering**: Utilized the K-Means Clustering algorithm to categorize customers into distinct segments based on variables such as income, expenditure habits, and family characteristics. The optimal cluster count (k = 4) was identified using the Elbow Method.
* **Supervised Learning – Classification Models**: Developed classification models using Decision Tree, K-Nearest Neighbors (KNN), and Random Forest algorithms to predict customer reactions to marketing initiatives. Among these, the Random Forest model demonstrated superior performance with an accuracy of approximately 90.4%, positioning it as the most effective option for forecasting marketing campaign outcomes.

**Limitations and Lessons Learnt**

**Limitations Encountered:**

* The dataset used was somewhat narrow in scope, lacking temporal and regional diversity—factors that could influence customer behavior but were not captured due to the static nature of the data.
* Certain categorical variables, like *Marital\_Status*, contained inconsistent entries that required manual standardization.
* Due to constraints in time and resources, the project did not incorporate deep learning techniques.

**Key Takeaways:**

* Data preprocessing and feature engineering play a crucial role in enhancing the accuracy and reliability of predictive models.
* Integrating exploratory data analysis (EDA) with domain expertise helps in crafting impactful features.
* Gained hands-on experience with how various machine learning algorithms perform when applied to real-world business datasets.

**Further Enhancements / Recommendations**

Future work in this area could consider the following improvements:

* Introduce **time-series analysis** to capture and evaluate how customer behavior changes over time.
* **Enhance the dataset** by integrating external sources such as transaction records, web engagement metrics, and location-based data for richer insights.
* Utilize **dimensionality reduction techniques** like PCA or t-SNE to improve both data visualization and model efficiency.
* Experiment with **advanced algorithms** including XGBoost, LightGBM, and deep neural networks to potentially boost predictive accuracy.
* Build an **interactive dashboard** using tools like Power BI or Tableau to make insights more accessible and actionable for marketing teams.

12. Appendices

This section presents additional resources that complement the core content of the report. These materials offer deeper insights into implementation specifics, usage instructions, and supporting elements that might disrupt the narrative flow if included in the main chapters.

**Appendix A: User Documentation**

**Project Title**: *Predictive Customer Behaviour Modelling using AI*   
**Objective**: To examine customer profiles and behavioral patterns using machine learning approaches, with the goal of helping marketing teams enhance their segmentation strategies and target audiences more effectively.

**Functionality Overview**:

* Load and clean the dataset
* Perform Exploratory Data Analysis (EDA)
* Engineer new features such as Age, Total\_Spent, and Family\_Size
* Cluster customers using K-Means
* Build classification models using Decision Tree, KNN, and Random Forest
* Visualize clusters and classifier results

**Tools Used**:

* Programming Language: Python 3.x
* Platform: Jupyter Notebook
* Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, yellowbrick

**Appendix B: Installation Instructions**

To run the project locally, follow these steps:

1. **Prerequisites**:
   * Python 3.x installed (preferably 3.8 or higher)
   * Jupyter Notebook (install via Anaconda or pip)
2. **Installation Steps**:

|  |
| --- |
| 1. pip install pandas 2. pip install numpy 3. pip install matplotlib 4. pip install seaborn 5. pip install scikit-learn 6. pip install yellowbrick |

**3. Launch the Jupyter Notebook**:

* Navigate to the project directory and open the notebook using:

|  |
| --- |
| jupyter notebook |

* Open the notebook file Customer Personality Analysis.ipynb and run all cells in order.

**Appendix C: README – How to Interact with the System**

**Step-by-step instructions to use the project:**

1. **Load the dataset**:  
   The dataset marketing\_campaign.csv is loaded using pandas.read\_csv(). Ensure the file is present in the same directory as the notebook.
2. **Execute Data Cleaning Cells**:  
   Run preprocessing cells to handle missing values, create derived columns, and prepare the data.
3. **Visualize Data**:  
   Run EDA cells to understand the dataset's structure and trends using charts and graphs.
4. **Run Machine Learning Models**:  
   Execute the clustering and classification cells to generate and view results.
5. **Understand the Output**:
   * Cluster labels will be added to the dataset
   * Classification accuracy will be printed for each model
   * Visualizations will help interpret clusters and predictions

**Appendix D: Sample Source Code**

|  |
| --- |
|  |

[**GitHub link for SEM IV - Project's code file & DB**](https://github.com/chatterjeeso/MCA_4th_SEM_FINAL_PROJECT.git)

|  |  |
| --- | --- |
| **Appendix E: Glossary** |  |
|  |  |
|  |  |
|  |  |

13. References / Bibliography

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
2. Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
3. Dataset Source: *Marketing Campaign Data* – Provided for academic and educational use.
4. Official Documentation:
   * [Scikit-learn](https://scikit-learn.org/stable/)
   * [Pandas](https://pandas.pydata.org/)
   * [Seaborn](https://seaborn.pydata.org/)
   * [Matplotlib](https://matplotlib.org/)
5. Online tutorials and resources from:
   * Kaggle: [https://www.kaggle.com](https://www.kaggle.com/)

Towards Data Science: [https://towardsdatascience.com](https://towardsdatascience.com/)

**Thank You**