Marketing Dataset Project

Introduction

This document outlines the project guidelines and dataset description for a comprehensive marketing analysis using various AI and data analytics techniques. The dataset contains 10,000 rows with customer-level information suitable for clustering, classification, regression, and other statistical analyses.

Dataset Description

The dataset includes the following variables:

- 1. CustomerID: Unique identifier for each customer.
- 2. Age: Age of the customer (18–70 years).
- 3. Gender: Gender of the customer (Male/Female).
- 4. Income: Annual income of the customer (USD 20,000–150,000).
- 5. SpendingScore: A score (1–100) indicating spending behavior, positively correlated with income and online shopping frequency.
- 6. OnlineShoppingFrequency: Number of times the customer shops online per month (1–20).
- 7. PreferredChannel: The channel the customer prefers (Online, In-Store, Both).
- 8. CustomerSatisfaction: Satisfaction score given by the customer (1–10).
- 9. PromotionalEmailsOpened: Number of promotional emails opened by the customer (0–50).
- 10. ProductReturns: Number of products returned by the customer (0–10).
- 11. LoyaltyProgram: Indicates if the customer is part of the loyalty program (Yes/No).
- 12. ConversionRate: Percentage of visits converted into purchases (0.01–0.30).

Project Guidelines

The project involves applying various AI and data analytics techniques to the dataset. The guidelines for each technique are as follows:

KMeans Clustering

Use the variables: SpendingScore, Income, and OnlineShoppingFrequency. Identify customer segments and analyze characteristics of each cluster.

ANOVA

Compare spending scores across different preferred shopping channels. Test for significant differences between groups.

Regression Analysis

Predict spending scores using age, income, and online shopping frequency as predictors. Interpret coefficients and assess model performance.

K-Nearest Neighbors (KNN)

Classify loyalty program membership (Yes/No) based on customer demographics and spending behavior. Optimize k for the best performance.

Naive Bayes

Use the Naive Bayes algorithm to classify loyalty program membership (Yes/No) based on variables such as age, income, spending score, and online shopping frequency. Evaluate the model's performance using accuracy, precision, recall, and F1 score.

Classification & Regression Trees (CART)

Build a decision tree to predict whether a customer belongs to the loyalty program. Analyze decision rules and tree structure.

Logistic Regression

Predict loyalty program membership using age, income, spending score, and shopping frequency. Assess accuracy and interpret model coefficients.

The purpose of AIDA was to build end to end models. How to build end to end models?

1. Start by creating clusters using KMeans.
2. For each cluster perform – Naïve Bayes, KNN, CART, Logistic Regression.

- Which model gives the highest accuracy for each clusters
- Can we first do classification and then do clustering?

Deliverables

1. A detailed report explaining methodologies, findings, and recommendations.
2. Visualizations for analysis.
3. Well-documented Python code.