Data Science-1 Assignment 2: Building Machine Learning Models

Summary Report

1. **Data Cleaning and Preparation**
    a. Importing Modules:
    - Pandas, NumPy, Seaborn, Matplotlib, Datetime, Sklearn
    - Modules are used to analyze and visualize data and produce defined metrics

    b. Data Reading and Merging:
        - To load all the 4 csv files and create a data frame to read the file.
        - Merging the 4 different data frames into a single data frame based on customer id and campaign date to perform analysis on entire dataset.

    c. Data Cleaning:
        - Identify different data types, check for duplicate, null and missing data.
        - Removing duplicates and assigning columns to appropriate data types
        - Filtering data targeting customers with a definite response: ['Yes', 'No']
        - Filling missing and null values of age (10.04%) and gender (5.42%) as these are a considerable number of values in our dataset.

2. **Feature Engineering**

- Understanding the relationship between different variables, graphs, derive insights and develop key business metrics to derive solutions and make business recommendations.

    a. Data Analysis:

    - Performing RFM Analysis including metrics like recency (last purchase date – current date), frequency (number of transactions), Lifetime Value (customer_total_transaction) based on customer id and transaction sum.
    - Calculated RFM score and derived a customer rfm score for better analysis

    b. Define Key Business Metrics:

    - Defining lifetime of customer (lifetime_customer) on purchase and join date.
    - Defining average customer spend (Customer_day_spend) based on total transaction and lifetime of customer.
    - Transforming Categorical Data using One-hot encoding for input values
    - Independent Variable: Age, site visits, emails opened, clicks, gender, product category, campaign response and promotion, recency, frequency
    - Output/Dependent Variable: value_customer column was calculated based on average customer spend and selected 70% quantile value ($58) as threshold to differentiate between 'High' value and 'Low' value customers.

c. Plots/Graphs:

- Analyzing campaign response plot: 'No' responses slightly higher than 'Yes'
- 'Heatmap plotted to understand correlation between the different variables.

3. **Model Building**
    a. Machine Learning Technique: Classification Models as our output is a binary
    b. Machine Learning Models:
    - 3 Classification Models: Logistic Regression, Decision Tree and Adaboost
    - The entire dataset had 89,989 values but only 10,000 unique customer id's hence the dataset was grouped on customer id to avoid duplicated values.
    - Using grid search to check for the best estimator for the desired model.
4. **Model Evaluation**
    a. Visualizations:
    - ROC Curve and Confusion Matrix was plotted to better realize the predicted values and understand the performance of the model.
    b. Scores:
    - Model accuracy, cross-validation score, mean cross-validation score and number of instances (0,1) was calculated.
    - Comparing predicted and original values (y_test /dependent variable).
    c. Best Model: **Adaboost** is the best model for classification
    - Accuracy: 93.6%, Precision (Class: 0,1): (0.96, 0.86)
    - Recall (Class: 0, 1): (0.97, 0.82), F1-Score (Class: 0, 1): (0.96, 0.84)
    - Cross-validation Mean Score: 92.96%
5. **Business Impact Analysis**

- Customer Segmentation: Targeted Marketing
    - By Targeting and identifying customers that are 'High' value with the help of promotional campaigns, product categories and response levels, EcomX can increase their revenue margins by increasing their average daily spend.
    - Introduce Tailored Offers, Bundled Discounts, Early Access, and Campaigns
    - Output Threshold value for 'high' value customers: $58 and Mean value of customer spend per day for 'high' value customers: $720
    - Consideration: $720 average customer spend/day by 'High' Value Customers
    - Consideration: 20,000 'High' Value Customers generate $14,400,000 revenue
    - Assuming a $50 increase for 15% of 'High' Value Customers (average spend)
    - Revenue Generated: 85% ($720 * 17000) + 15% ($770 * 3000) = $14,550,000
    - Total Profit: $150,000 ($14,550,000 - $14,400,000)
    - An increase in average daily spending of $50 for just 15% of high-value customers can lead to an additional $150,000 in revenue daily.