Summary Report

Objective

The primary objective of this assignment was to perform Exploratory Data Analysis (EDA), clean the dataset, handle missing values, build a logistic regression model to predict the conversion status of leads, and evaluate the model's performance.

Data Understanding and Cleaning

The first step involved understanding the dataset provided. The dataset consisted of various features such as Lead Origin, Lead Source, Do Not Email, TotalVisits, Total Time Spent on Website, and more. The target variable was Converted, indicating whether a lead was converted into a customer.

Data Inspection

The initial inspection revealed some important characteristics of the data, such as the presence of missing values and categorical features. A summary of the dataset was generated to identify columns with missing values and the overall data distribution. The columns with the most missing values were Asymmetries Activity Score and Asymmetries Profile Score.

Handling Missing Values

To handle missing values, the following strategies were employed:

Numeric Columns: Missing values were filled with the mean of the respective columns to ensure that these columns could be used in the model without introducing biases.

Categorical Columns: Missing values were filled using the most frequent value or a placeholder ('None') to maintain consistency.

Dummy Variables

Categorical variables needed to be converted into numerical values for the logistic regression model. This was achieved by creating dummy variables. The pd.get_dummies function was used to convert categorical columns into a series of binary variables.

Data Preparation

After handling missing values and converting categorical features, the dataset was split into training and testing sets using a 70-30 split. This division ensured that the model could be trained and then evaluated on unseen data.

Splitting the Data

The features (X) and the target variable (y) were separated. The dataset was split into X_train, X_test, y_train, and y_test to facilitate model training and evaluation.

Model Building

A logistic regression model was chosen due to its efficiency and suitability for binary classification tasks. The following steps were followed:

Model Instantiation: A logistic regression model was instantiated with a maximum iteration limit set to 1000 to ensure convergence.

Model Training: The model was trained using the training data (X_train and y_train).

Model Evaluation

The trained model was evaluated using the test data (X_test and y_test). The evaluation metrics included accuracy, confusion matrix, and classification report:

Accuracy: The overall correctness of the model.

Confusion Matrix: Provided insights into true positives, true negatives, false positives, and false negatives.

Learnings and Insights

This assignment provided several valuable learnings:

Data Cleaning and Preparation: Understanding the importance of handling missing values and converting categorical data into numerical format was crucial. These steps are foundational to building a reliable model.

Handling Errors: Encountering and resolving errors, such as ensuring all data is numeric for VIF calculation, reinforces the importance of data preprocessing.