SUMMARY REPORT

Problem Statement:
X Education sought assistance in identifying the most promising leads with a high likelihood of conversion into paying customers. The goal was to build a lead scoring model that assigns a score to each lead, reflecting their conversion potential. The CEO provided a target lead conversion rate of around 80%.

Approach:
To address the problem, I followed a systematic approach consisting of four parts: data understanding and cleaning, exploratory data analysis, feature engineering, and model building and evaluation.

Part 1: Data Understanding and Cleaning
In this stage, I began by understanding the dataset's structure, including the variables and their meaning. The dataset contained information on leads and their interactions with X Education. I performed data cleaning tasks such as handling missing values, dropping irrelevant columns, and addressing data inconsistencies.

Part 2: Exploratory Data Analysis (EDA)
EDA involved gaining insights into the dataset through visualizations and statistical summaries. I examined the distribution of variables, identified patterns, and explored relationships between features and the target variable (conversion). Key findings from EDA included understanding the importance of certain features like total visits, time spent on the website, and page views in predicting conversion.

Part 3: Feature Engineering
Feature engineering focused on transforming and creating new features to improve the model's predictive power. I performed tasks such as encoding categorical variables, scaling numerical features, and handling outliers. Additionally, I split the dataset into training and testing sets for model training and evaluation.

Part 4: Model Building and Evaluation
In this phase, I selected the logistic regression algorithm to build the lead scoring model. Logistic regression is suitable for binary classification tasks and provides interpretability. I trained the model on the training set, fine-tuned its hyperparameters using cross-validation, and evaluated its performance on the test set.

The model achieved an accuracy of approximately 79.05% on the test set, close to the CEO's target of 80%. Evaluation metrics such as sensitivity, specificity, precision, and the precision-recall curve provided insights into the model's performance. The model showed promising results in correctly identifying converted leads, but there was room for improvement in recall.

Learnings:
Throughout this assignment, I gathered several key learnings:

1. Data Preprocessing: Data cleaning and preprocessing are critical steps in any data analysis project. Handling missing values, addressing inconsistencies, and transforming variables are essential for accurate modeling.

2. Exploratory Data Analysis: EDA is crucial for understanding the dataset, identifying patterns, and discovering relationships between variables. It helps in feature selection and provides insights for modeling decisions.

3. Feature Engineering: Transforming and creating relevant features can significantly impact model performance. Scaling numerical features, encoding categorical variables, and handling outliers contribute to building robust models.

4. Model Selection and Evaluation: Choosing an appropriate model and evaluating its performance using relevant metrics are key to achieving the desired outcome. Logistic regression, in this case, provided a balance between interpretability and predictive power.

5. Interpretability and Explainability: Logistic regression models offer interpretability, allowing us to understand the impact of features on the target variable. This helps in providing actionable insights and recommendations based on the model's coefficients.

6. Precision-Recall Trade-off: The precision-recall curve provides a useful visualization to understand the trade-off between precision and recall at different threshold values. It aids in selecting an optimal threshold based on the desired balance between precision and recall.

Conclusion:
In conclusion, this project involved developing a lead scoring model for X Education to identify potential customers with higher conversion