Summary Report: Lead Scoring Assignment

1. Introduction:
   The aim of this assignment was to build a logistic regression model to assign lead scores to potential customers of X Education, an online education company. The lead scores would help identify hot leads with a higher probability of conversion, enabling the sales team to prioritize their efforts effectively. Additionally, the report explores strategies for aggressive lead conversion during internship periods and minimizing useless phone calls when sales targets are met ahead of schedule.

2. Methodology:
   a. Data understanding and exploration:
      I started by thoroughly understanding the provided dataset, which included various variables such as lead origin, lead source, website interactions, lead activities, and demographic information. This understanding was crucial for feature selection and modeling.
   b. Data preprocessing:
      I performed data preprocessing steps, including handling missing values, treating categorical variables, and removing irrelevant columns that would not contribute significantly to the lead scoring model. This process ensured clean and consistent data for analysis.
   c. Feature Engineering:
      To enhance the predictive power of the model, I conducted feature engineering by creating dummy variables for categorical columns and deriving new features from existing ones. This helped capture valuable insights and patterns in the data.
   d. Model development:
      - Logistic regression was chosen as the modeling technique due to its interpretability and effectiveness in binary classification problems.
      - The dataset was divided into training and testing sets, and the logistic regression model was trained on the training set.
   e. Model Evaluation:
      - The trained model was evaluated on the testing set using evaluation metrics such as accuracy, precision, recall, and F1-score to assess its performance in predicting lead conversions.
      - My first cut-off point is 0.5 has evaluation metrics: Accuracy 81%, Sensitivity 69%, Specificity: 89%. Specificity is good 89% but Sensitivity is 69%. It is because the cut-off was not optimum, so I analyzed ROC curve and calculate the AUC score to find the optimum cut-off point.

- The model achieved satisfactory results, with a high accuracy rate and reasonable precision and recall scores.

3. Result:
    a. Evaluation metrics:
    - Train data:
        - Accuracy: 81%
        - Sensitivity: 80%
        - Specificity: 81%
    - Test data:
        - Accuracy: 80%
        - Sensitivity: 80%
        - Specificity: 80%

    The model can predict the Conversion Rate well to help making calls with a higher lead conversion rate of 80%.

    b. Important features:
    - We should make calls to the leads coming from:
        - Lead Origin: Lead Add Form
        - What is your current occupation: Working Professional
        - Longer total time spent on Website
        - Last Activity: SMS Sent
    - We should not make calls to the leads coming from:
        - Last Notable Activity: Email Opened, Page Visited on Website, Modified, Email Link Clicked
        - Lead Source: Google, Organic Search, Referral Sites, Direct Traffic
        - Do Not Email: yes
        - Last Activity: Olark Chat Conversation

    c. Hot leads:
    There are 321 hot leads with score >= 90