Summary Report

My approach to this Case Study involved the following steps:

- Inspecting and understanding the data

- Data Cleaning and EDA including:
  - Handling the 'Select' level that is present in some categorical variables
  - Handling missing values
  - Checking the number of unique categories in categorical columns
  - Dropping extremely skewed features

- Data Preparation including;
  - Converting some binary variables (Yes/No) to 0/1
  - Creating dummies for other categorical variables
  - Performing train - test split
  - Feature Scaling

- Model building and Evaluation
- Making predictions on the test set
- Calculating Lead Score
- Determining Feature Importance

Business Recommendations

- When looking at the feature Lead Source:
  - Majority of generated leads come from Google and Direct traffic, while the least number of leads originate from Others
  - Welingak website exhibits a very high conversion rate, it is advisable to maximize leads from this website
  - Prioritizing Olark chat, Organic search, Direct traffic, and Google leads may result in increased lead conversion

- Management specialization is important as managers are likely to convert favorably.

- From EDA, most of the successfully converted leads come from the Unemployed, special attention should be given to this group.

- There are good chances of a working professional to sign up for a course.

- It appears that the categories of "Housewives," "Businessman," "Student," and "Other" are not easily converted to enroll in the course.

- When looking at the feature Last Activity,SMS Sent has the highest success conversion rate followed by Email Opened.

- The probability of lead conversion tends to increase as the values of the following features increase:
  - Lead Source_Welingak Website
  - Lead Origin_Lead Add Form
  - Lead Origin_Landing Page Submission
  - What is your current occupation_Working professional
  - Lead Source_Olark Chat
  - Lead Quality_High in Relevance
  - Total Time Spent on Website

- The probability of lead conversion tends to decrease as the values of the following features decrease:
  - Lead Quality_Worst
  - Lead Quality_Not Sure
  - Last Notable Activity_Email Link Clicked
  - Last Notable Activity_Modified
  - Last Notable Activity_Page Visited on Website
  - Last Notable Activity_Olark Chat Conversation
  - Last Notable Activity_Email Opened
  - Last Activity_Email Bounced

- ➢ Do Not Email

- ❖ It is important to note that, based on the business requirements, we have the flexibility to adjust the probability threshold value. Modifying this threshold value allows us to control the trade-off between sensitivity and specificity in the model. Increasing the threshold will decrease sensitivity but increase specificity, while decreasing the threshold will have the opposite effect, increasing sensitivity but decreasing specificity. This adjustment allows us to tailor the model's behavior to align with specific business needs and priorities.

- ❖ A high sensitivity value ensures that most leads who are likely to convert are correctly predicted as such. On the other hand, a high specificity value ensures that leads with borderline conversion probabilities are not falsely selected. In other words, high sensitivity focuses on minimizing false negatives (leads who should have been identified as likely to convert but were missed), while high specificity aims to minimize false positives (leads who are incorrectly identified as likely to convert). Balancing these two measures is crucial to achieve an optimal prediction outcome for lead conversion.