SUMMARY REPORT

I. Introduction:

In this case study, we are helping an education company named X Education to identify the most potential leads who are likely to convert into paying customers. We have been provided with a dataset consisting of various attributes. We will be building a logistic regression model to assign a lead score to each of the leads. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted.

II. Data Cleaning:

We observe that some of the categorical variables have a level called 'Select', which must be handled because it is as good as a null value. We handled the missing values using techniques such as mode and dropping rows/columns. Specifically, we dropped the columns with missing values that are over 40%. With columns with moderate missing values:

- Impute missing values with mode.
- Create “Others” category for insignificant values.
- Drop other columns if the values fall into only one unique value.

III. Data Exploration:

We start by exploring the data to understand the distribution of different attributes and their impact on the target variable. We use various visualization techniques such as histograms, box plots, etc. to analyze the data. For instance, the univariate and bivariate analysis on Lead Origin helped us find out that to improve the lead conversion rate, X Education should focus on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

IV. Data Preparation:

We converted binary variables namely “Do Not Email” and “Do Not Call”. Other than that, we also handled dummy variables namely 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'City', and ‘Last Notable Activity’. After that, we splitted the data into training and testing sets, followed by rescaling the Features by normalization.

V. Model Building:

We split the data into training and testing sets and then build a logistic regression model using the training data. We use various evaluation metrics such as accuracy, sensitivity, and specificity to evaluate the performance of the model.

VI. Making Predictions:

We interpret the model coefficients to understand the impact of different features on the target variable. We use the ROC curve to optimize sensitivity as it shows the tradeoff between sensitivity and specificity, and it depicts how accurate the test is.

We also provide a list of recommended actions based on the lead score to help the sales team prioritize their efforts and improve the lead conversion rate.

VII. Conclusion:

In this case study, we helped X Education to identify the most potential leads by building a logistic regression model and assigning a lead score between 0 and 100 to each of the leads. We learned various data exploration, data preparation, model building, and model interpretation techniques during the process. We also learned the importance of handling missing values, converting categorical variables into dummy variables, and interpreting model coefficients. Overall, the logistic regression model is a simple yet effective technique for lead scoring and can be easily deployed in a web application.