# SUMMARY REPORT

The below are the steps that were followed in the analysis of the data using Logistic Regression.

- The dataset initially contained 9,240 rows and 37 columns. Several columns were found to impact the target variables, and numerous categorical variables were present, necessitating the creation of dummy variables.

- As the first step in our analysis, it was crucial to examine the null values. Upon inspection, we found a significant number of null values, which could potentially distort the dataset. To address this, rather than dropping individual rows, we chose to drop columns with more than 30% missing values.

- We also removed several unnecessary columns, such as "Receive More Updates About Our Courses," "Update Me on Supply Chain Content," "Get Updates on DM Content," "I Agree to Pay the Amount Through Cheque," and "Magazine." These columns were irrelevant to the analysis, so they were excluded from the dataset.

- Next, we visualized the “Select” values in the dataset. It was observed that the count of "Select" values was disproportionately high compared to other data points. To address this, we replaced these values with “Unknown.” Additionally, the "City" column only contained cities from Maharashtra, so we decided to drop the "City" column entirely.

- Upon reviewing the value counts across all columns, we noticed several columns where a single value overwhelmingly dominated. These included columns such as "Do Not Call," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," and "Through Recommendations." Since most of the data in these columns was consistently “No,” we decided to drop these columns, as they provided little value for the analysis.

- After cleaning and visualizing the dataset, we proceeded to analyse the categorical variables and conducted a bivariate analysis. The next step was to split the data into training and test sets, followed by scaling the features.

- For model building, the dataset contained a large number of variables. To streamline the process, we used Recursive Feature Elimination (RFE) to select a more manageable set of features from the pool of variables.

- Various models were then built and evaluated. After reviewing p-values and Variance Inflation Factors (VIFs), we selected the model with the best performance. We also calculated the Receiver Operating Characteristic (ROC) curve and chose the model with an area under the ROC curve (AUC) of 0.8. To fine-tune the model, we determined the optimal cutoff point by analysing the trade-offs between sensitivity and specificity, which turned out to be 0.43.

- In the final step, we identified the key factors driving the lead score: 'Lead Origin', 'Lead Source', 'Do Not Email', 'Total Time Spent on Website', 'Page Views Per Visit', and 'Last Notable Activity.' These factors were found to have the most significant impact on lead conversion.