Executive Summary

Analyzing the problem:

Our Hypothesis is that, with the help of the customer's profile, we can identify who's most likely to convert and who's not. Hence we're looking at a logistical regression approach to identify the most influential variable.

Data Cleaning and Preparation:

- After importing the data, unnecessary columns (Prospect ID and Lead ID as they are for identification) were dropped for better understanding
- Separating the categorical and numerical columns for later convenience in the EDA part.
- In a lot of categorical columns, the default option 'Select' was selected as they were unfilled. Hence we will be considering them as null values.
- Proceeding that, some columns with high null values (>30%) were dropped.
- Some of the categorical values have a single category dominating the entire feature. Which wouldn't be useful enough. So dropping those columns as well. (Country, What is your current occupation, What matters most to you in choosing the course)
- Dropping the remaining NaN rows since they are under 1.5% of the entire data.
- Identifying the columns with only one category as it won't give any contribution to the model. Thus dropping them.

EDA - Univariate Analysis of the entire data set:

- The numerical columns (Pages per View and Total Visits) were found to have some outliers in the data. It was well-treatable, and they were accordingly treated with quantile representations. There were still some outliers after the treatment but it was within the range of acceptable data and losing them would result in a huge loss of total data.
- The categorical variables (Last Activity, Lead Source, and Last Notable Activity) had a huge number of categories, therefore categories with very low contributions to the overall data were converted as 'Others' for better visualization.

EDA - Multivariate Analysis:

- All the columns were plotted against the conversion variable which revealed some insights as to some of the influential variables and a heatmap was plotted for the correlation matrix.

- Though the correlation matrix revealed a high correlation, These columns are required and can be evaluated using VIF later in the model, henceforth keeping them.

Model Building and evaluation:

- Train-Test splitting is done and necessary data preparation was done to the train and test split once again to make sure the correct data is fed to the model.
- Created dummy variables for the categorical features using one hot encoder
- Using RFE 16 features were selected and after which VIF is calculated using a user-defined formula.
- High VIF and High P-value features are removed until the final model arrives.
- After finalizing the VIF, F1 score, Precision, Recall, and other evaluation metrics, the model is finalized.
- The model's ROC curve, Precision vs Recall has been plotted to check for any abnormalities. Which was nill.
- The optimal threshold value for the regression model is found to be 0.40.
- The coefficients of the final model features were checked.
- Using the optimal threshold, the score data is predicted using a user-defined formula.
- It is then compared with the train data set afterward, predicting the data for the test data set as well.