Summary report:

- Problem statement: Finding the potential leads who turns out to be paying customer
- We import the libraries which are required and also will report other libraries as and when required

- Data Understanding:
  - We import the data and we notice that we have column called prospect ID which doesn't make sense in Data Analysis so we decide to drop it
  - Using different functions like describe, info, shape etc to get a brief data understanding

- Missing value treatment and Outlier Analysis:
  - Checking the percentage of missing values present in all the features i.e. columns in the data
  - Plotting a box plot for 'Total Visits' and 'Page Views Per Visit' and deciding weather we need to impute mean or median for the missing values in the mentioned column. Box plot clearly shows that we need to impute median for missing values
  - Deciding to drop some of the columns for the missing value percentage is above 40%
  - Checking the value count for the features present in the data. As it will give the idea about the variance in the data, if the data is skewed and if the missing values are present and if missing values are present, we will impute them with mode of the data. We do this for all the features
  - We recheck the missing value percentage and see its zero for all the columns after missing value treatment
  - Renaming some of the columns to shorter words to make it compact and better understanding
  - Performing outlier analysis for the numeric variables and observed that 'Total Visits' and 'Pages Viewed have outliers present and do a soft outlier treatment
  - After performing all the steps, we arrive at a final dataset on which we want to perform our machine learning model

- Data preparation:
  - Mapping all the yes's and no's present in the some of the columns to 0 and 1
  - For categorical variables with multiple levels, creating dummy features (one-hot encoded)
  - Putting the feature variables into X and the target variable into y
  - Performing a train-test split of the data
  - Scaling all the numeric data using standard scaler
  - Checking the converted rate of lead to the actual number of leads who turn out to be paying customers

- Model building:
  - Using RFE to decide the features that we want to use in the data and starting the RFE with 15 variables
  - Building logistic model and creating multiple iterations of it until we have p values less than 0.05 and VIF's less than 5 for the all the selected variables through RFE
  - Finally come to a model which satisfies the above criteria and we have 13 Features
  - Checking on sensitivity, specificity, accuracy and false positive rate if our model is performing correctly
  - Plotting a ROC curve to see the trade of between sensitivity and specificity
  - Creating columns with different probability cut-offs
  - Calculating accuracy, sensitivity and specificity for various probability cut-offs.
  - According to our requirement our model should have high sensitivity and according to this we select the criteria as 0.3 as cut-off
  - Just as a reference we check the recall and precision

- Applying model on test data and entire data set:
  - Performing just transform the test data set
  - Applying the logistic model on the test data set and checking the sensitivity, specificity and accuracy of the model
  - Storing the features in a separate variable and target feature in a separate variable
  - Performing just transform the entire data set
  - Applying the logistic model on the test data set and checking the sensitivity, specificity and accuracy of the model
  - The sensitivity, specificity and accuracy of the model is good
  - We filter the data to see all the probabilities greater than and equal to 0.8

  Finally, we have all the unique lead numbers whose probability of turning into actual paying customers is more than 80%