Summary

A brief overview about the case study :-
After assessing the business problem a few insights were clear

- ✓ Supervised learning problem.
- ✓ Build logistic regression predictive model.
- ✓ Achieve 80% predictive rate of conversion from the model.

Based on above insights the following steps were carried out.

1. Data Understanding:
   The data had 9240 observations/ leads with total of 37 attributes. The attributes ranged from unique ID's assigned to each lead, sources & identification of potential leads, fields filled in forms such as country & city where they reside, email preferences, employment status, sectors in which they are employed, contact preferences, last activity status, website visit stats, whether or not the lead is converted a few to be listed.

2. Data Cleaning & Preparation:
   Though the data was partially clean, it needed more definite cleaning measures. At first, the columns with null value percentage greater than 20% were dropped. Later, based on intuitions few more columns were eliminated & also rows with null values since, it wouldn't add much weight to out model. Appropriate imputation techniques were used. Further, outlier variables were identified, based on which data was cropped.

3. EDA:
   Thorough exploratory data analysis was performed to retrieve hidden insights using plotting libraries.

4. Dummy Variable Creation:
   As we know we won't be able to use categorical features directly in our model, hence dummy variables for such features were created.

5. Preprocessing data for modelling:
   As this was done, next steps were to filter into independent & target variable, then split the data into train & test set & subsequently, standardize the numerical variables having varying ranges for which Min-Max Scaling technique was used.

6. Model building, metrics & results:
   At first feature selection was done using feature selection technique "Recursive feature elimination". 15 features were selected based on the same.
   Later, logistic regression model was built on train data. In combination with this multi-collinearity was measured using VIF.
   Recursively, models were built dropping few features with high p-value & high VIF till both were in acceptable range for all variables & thus model was finalized.
   Required metrics were calculated for the model.
   As optimal threshold was derived using metrics so as to achieve desired conversion rate.
   Later, predictions were made on test set, followed by calculating metrics for test set predictions.
   Observations on test data:
   1. Acc – 80%
   2. Sensitivity – 80%
   3. Specificity – 72%
   4. Influential variables:
      a) Last Notable Activity_Had a Phone Conversation
      b) Lead Origin_Lead Add Form.
      c) What is your current occupation_Working Professional.