Assignment Summary

After importing the dataset, we make some key observations by exploring it further. One prominent observation is the presence of ‘Select’ entries in various fields of our dataset. Another observation is that some attributes are dominated by a single entry. For example, 99% of entries in the Do not call column are yes. In our data cleaning stage, we have rectified these issues by replacing ‘select’ with null and dropping those aforementioned imbalanced columns. We also dropped columns that have more than 30% null values. Additionally, we consolidated less significant categorical variables as ‘Others’.

In the next stage, we start with the Exploratory Data Analysis. For our categorical attributes, we plotted Count Plots with the hue of our target variable, and Pie charts to get a better understanding of how our data is distributed. For numerical, we have created box plots with a similar intention. Some of the key observations we have made in EDA are:

1. Most of the leads are from India
2. Working professionals have a high ratio of converted leads
3. A large share of leads who do not convert spend less time on the website

The next order of business is preparing data for model creation. We start by creating dummy variables from our categorical columns. For this, we split the data into training and testing on a ratio of 70:30. We then use the min-max scaler to scale the numerical variables. Next, we use RFE to select the 15 best features. We train our model and check for p-values and VIFs. Here, we are satisfied that all p-values are under 0.05 and all our VIFs are under 5. Key Insights from the Model are as follows:

1. The higher the number of total visits made by the user, the higher the chance that the lead gets converted to a paying customer.
2. The more the number of pages a lead visits in a single visit, the lesser the chance of them converting.
3. The higher the time the lead spends on the website, the higher the chance that they convert to a paying customer

Next, We plot the ROC curve and the sensitivity-specificity curve to find the optimal cutoff. We determine that the optimal value for the cutoff is 0.36. We obtain the performance metrics on the training set as follows: accuracy = 0.802566, sensitivity = 0.802514, specificity = 0.802599, F1 Score = 0.756064.

Next, we perform prediction on the test set. We generated the confusion matrix and calculated the performance metrics like accuracy, sensitivity, specificity, precision, recall, and f1 score. Here are the values of the metrics we obtained.

1. The accuracy on the test set is 0.8084415584415584
2. The sensitivity on the test set is 0.8027397260273973
3. The specificity on the test set is 0.8121645796064401
4. The precision on the test set is 0.7361809045226131
5. The recall on the test set is 0.8027397260273973
6. the f1 score is 0.7680209698558322

Thus, we have met the CEO’s goal of having sensitivity and recall of more than 80%. Finally, we assign the score variable.