SUMMARY

The data we have is that of X-company as to how their lead conversion has been done. We were given about 9000 data points to work and build a logreg prediction model to understand how to target users. The idea was to increase the conversion rate from 30% as it is right now to 80% or higher by predicting hot leads.

Here are the steps we followed:

1. Cleaning data:
   Here we found that there were a few columns (features) that had significant missing values – these were removed. Also ther were cases where we had to replace value such as "Select" as it did not provide much information. We collated some of the minor values under a common header such as "Others" so that we don't loose data and at the same time we are able to categorize them into logical buckets for logreg model building. We also at times substituted missing values with majority values as that of the case of "India" being used as the value for missing values in country.
2. EDA:
   Here we focussed on further analysis of data. Univariate and bi-variate analysis was done. We couldn't find any significant multi-collinearity. We handled a few outliers also here.
3. Dummy Variables:
   We focussed on getting dummy variables for the categorical values. We avoided values we felt were irrelevant.
   In this stage we also used Standard Scaler to scale the numeric values.
4. Data Split and Model Building:
   Here we split the data into 70:30 to train and test the data. We used RFE to identify 15 top variables and then we successively built multiple models by eliminating initially those with high p values. Thereafter we analysed the VIF values and further modified the algo.
5. Model Evaluation and Prediction
   In this step we analysed the model by building the confusion matrix and plotting the ROC curves and probability distribution curves to understand the cutoff as ~ 0.35.
   We then used the test data to check the predictions. Here again we checked the values for specificity, accuracy, etc.
   We verified if the values we received were consistent with the values we received initially at the model evaluation step with the train data set. We got coherent values.