Summary:

The overall approach taken to resolve this problem can be broken down into below 11 points:

1) Understanding the problem statement - We went through the excel sheet and understood majority attributes mentioned in the lead csv file.
2) Analyzing and handling data anomalies – We fixed the missing and null values. We also ensured the values which are having select in the attributes are handled by assigning them as unknown. We handled the outliers by capping the max values to 0.95 to ensure that the outliers don't interfere and impact the model and output
3) Univariate and Bivariate analysis (EDA) - Analyzed the target variables against different input variables using bivariate analysis and plots and charts to identify the attributes which are impacting the lead conversion
4) Dummy variable creation - To ensure that the categorical variables are converted to numerical values, we created dummies for all the categorical attributes.
5) Test Train Split - We then split the data into train and test data – 70/30 ratio.
6) Scaling -We then performed the scaling using standard scaler for the variables other than dummy or yes no variables- to bring everyone on the same scale
7) Feature Selection - The number of columns was around 69 and thus to select only the most relevant attributes we used RFE method to ensure we do not use many attributes for modelling. We selected the top 20 features for our model building using RFE
8) Model building - During this phase – we built the model using GLM and re-iterated the model by dropping the columns which have high p value(>0.05) or columns which have high VIF (greater than 5)
9) Model Evaluation – We evaluated the model on the basis of confusion matrix , accuracy, specificity, sensitivity and TPR and FPR. We also plotted the ROC and identified the Area under the curve which was as high as 0.95. This showed us that the overall model looks good.
10) Finding the optimal cutoff – We identified the optimal cutoff point by plotting Sensitivity, Specificity on the graph and the optimal came out to be 0.25
11) We then ran the model on test data and got the final accuracy as 89% - showing that the model does not only have good performance on train data but also on the test data.

Learnings – We tried 2 approaches for this problem statement – 1) Using label encoders and the second one using Dummy variables. We understand that using Dummies has increased our overall model accuracy , sensitivity and specificity – and thus helped us in predicting better.