Lead Socre Summary :

Our problem statement :

A company names X education help provide various online courses to industry professionals. Company uses various advertisement methods to get target users for their platform and their sales is dependent on the same. In order to shoot the sales, the responsible department reaches out to potential buyers of the course. However the conversion rate is low.
We need to optimize the approach to reach the highly like converting audience and analyse how they can improve the sales.

For this assignment we were give data that provides detail on the source of customer visit, time spent of course, form fill, contact method opted for example main or sms, occupation they are in. All these factors derive the conversion rate.

Steps we took during our LogisticRegression model building.

1. Read the data
   We got access to the sample data set given and understood the attributes provided. It includes using the data dictionary and going over some records in the data set.
2. Cleaning Data
   We saw there are many 'Select' values which meant data was Not Specified by the potential applicant. Therefore replacing such values by null.
   We analysed all attributes null percentage to derive the impact of a column in our model building. Any row with more than 45% null values didn't add to our analysis and therefore we dropped such columns
   We also analysed some categorical fields and grouped the smaller data group into one single group. For eg. countries to India and outside India
   We also replaced some fields with majority values in the data group.
   Redundant/duplicate field groups were clubbed to one as well. for eg Google and google.
3. Exploratory Data Analysis
   We did checks on the data types for categorical and numerical fields.
   Additionaly we came across fields which were single value dominant like 'I agree to pay through cheque' Such fields didn't add any insights to our analysis, therfore dropped them.

We also capped the outliers in numerical data set. for example 'TotalVisits'

4. Dummy Variables
   Next we created dummy variables for all categorical fields and also did binary field mapping for Yes/No values
5. Train and Test Data
   We split the given data set into 70-30 for train and test data set for the model respectively.
6. Model building on Train Data
   We performed RFE model building to fetch top 20 variables.
   Susequently removing fields with higher p value and VIF and rebuilding the model on the remaining fields. This was done to achive p value below 0.05 and VIF value below 5 for all variables.
   The optinum cut off value using ROC curve was established.
   Lastly after above conditions, we achived accuracy, specificity and sensitivity to be around 80% which is well accepted.
7. Model on Test data
   We calculated the conversion value based of the cut off 0.38
   Resulting accuracy, sensitivity and specificity to be same around 79-80%