Problem Statement:

An ‘X’ Education Company, offers online courses to industry professionals, gets many people landing on their website from the advertisements posted on various platforms and search engines like Google. The interested people provide their contact details to the company. These people are called as leads and are followed by the sales team to convert them, i.e. sell them the course. The conversion to lead ratio is currently 30%. The company wants to improve this ratio by following up only the potential leads that would most probably be interested in buying the course and filter out other. As a Data Scientist, we have to build a Machine Learning model that helps the company in achieving this and reducing the irrelevant follow ups by the sales team.

Approach:

1. We need to read, understand and prepare the given data before putting it through a ML algorithm for prediction. Starting with importing the required libraries and reading the csv file. The given data has 37 columns and 9240 records.
2. Dropping the columns that are not required such as unique value columns like ‘Prospect ID’ and ‘Lead Number’, columns created by the company like score columns. Assigning a null value for the non-responsive fields that have the value ‘Select’, which means no response added. Removing the columns having more than 50% null values. Removing the records having more than 5 null values. Imputing the columns that are left with appropriate central tendency.
3. After performing the data cleansing task we are left with 28 columns and 9191 rows. The percentage of records retained is about 99.47 % which is a good number.
4. Performed univariate analysis of numerical columns to understand how they are distributed. Checked the number of occurrences of values in categorical columns to understand the skewness and removed the columns with very high skewed values. After removing highly skewed columns we are left with 12 columns.
5. Created boxplot of numerical columns for outlier analysis. Found some outliers. Increased the bin size to include 99% of the values to remove most of the outliers. Assigned binary values to Yes/No columns. Also performed bucketing to keep most significant values in categorical variables and treat remaining as ‘Others’. Created dummy variables wherever possible.
6. Created a new dataframe by removing the target variable ‘Converted’ and ‘Lead Number’. Split the data in train and test to build and test the model efficiency. Created collinearity heat map to removed highly collinear variables.
7. Started building model using Recursive elimination method with top 13 features. Checked the p-value and VIF to remove the variables that are statistically insignificant. Iterated this until there were no insignificant variables left. Created a dataframe with actual converted flag and predicted probabilities of conversion. Defined a cut-off to set flag for predicted column based on the probability scores.
8. Created confusion matrix and calculated sensitivity, accuracy and specificity to understand how good our model is. Created the ROC curve to get the best cut-off value which is defined by the intersection point of sensitivity, accuracy and specificity. Also calculated F1 score to again get the measure of accuracy of the model.

Conclusion:

The final model has 11 features with 78% accuracy and 82% sensitivity.

The top 3 variables affecting the lead conversion are:

Lead Source, Last Notable Action and Last Action

The top 3 dummy variables affecting the lead conversion are:

LS_Direct Traffic, LS_Organic Search and LS_Google