SUMMARY

For X education company model building and prediction is done, so that company could find ways to convert its potential leads into numbers.

Leads are obtained through various ways on different platforms, so to convert those leads exact variables should be targeted.

The main business goal of case study is to make Lead conversion rate to 80% from the obtained leads, which is around 30% currently.

Approach:

From above problem description we conclude that the above problem is the classification problem, hence we choose logistic Regression to calculate the Lead rate.

Below are the steps followed to solve this problem.

Step1: Reading and Understanding Data:

- Number of rows and columns
- Data types of columns
- Checking details about each column in the dataset
- Checking how the data is spread.
- Checking for duplicates, if any.

Step 2: Data Cleaning:

- Checking for any change in data types
- Checking for null values and imputing them with appropriate methods
  - ✓ Dropping columns having null values more than 45%
  - ✓ Treating columns will null values less than 30%
  - ✓ Removing rows with null values in particular columns
  - ✓ Replacing values for some categorical columns.
  - ✓ mean imputation for numerical columns.
  - ✓ median imputation for numerical columns

Step 3: Data Visualization and Outliers Treatment:

- Univariate analysis on categorical column to see which columns makes more sense and removed columns whose variance is nearly zero.
- Univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
- Bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
- IQR method to treat the outliers in the data set.
- Plotted the correlation matrix to identify the columns which are correlated.

Step 4: Feature Scaling

Changed the binary variables into ‘0’ and ‘1’.

Step 5: Dummy Variables Creation:

Created dummy variables for the categorical variables.

Removed all the repeated and redundant variables.

Step 6: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70- 30% values.

Step 5: Model Building and Evaluation

- Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
- Using the statistics generated, we tried looking at the P-values in order to select the most significant values that should be present.
- Rest of the variables were removed depending on the VIF values and p-value.
- We then plot the ROC curve for the features and the curve came out be good with an area coverage of 96%.
- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
- Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.3.
- Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 89%; Sensitivity= 89%; Specificity= 90%.

Step 7: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 89% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.