SUMMARY

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

The following are the steps used:

Step1: Reading and Understanding of data

- Read the data and check its shape,features,datatypes etc

Step 2: Cleaning data

- Drop the variables with Unique values
- Option ‘select’ has to be replaced with a Null value since it did not give us much information.
- Columns with >35% null values are Dropped
- Few of the null values were changed to 'Not provided' or ‘Not specified’ so as to not lose much data,although they were later removed while making dummies.
- For some categorical variables impute with Mode

Step 3: EDA

- Imbalance in data checked and found only 38% leads are converted
  Performed Univariate and Bivariate analysis on categorical and numerical variables.
- Some categorical variables were irrelevant and they are dropped
- The numeric values with outliers are treated and plotted against Target Variable using boxplot and find their trends
- Heatmap was plotted to find correlation between the variables and no specific high correlation found among the variables

Step 4: Model Data Preparation

- Binary variables mapped to 0/1
  The dummy variables for categorical variables were created and later on the dummies with 'not provided' elements were removed.
- Train-Test split:The split was done at 70% and 30% for train and test data respectively.
- Standarization of Numeric variables done using StandardScaler

Step 5: Model Building

- RFE was done to attain the top 15 relevant variables.
- Later the rest of the variables were removed manually depending on the VIF values and p- value (The variables with VIF<5 and p-value <0.05 were kept).

Step 6: Model Evaluation

- Confusion Matrix was created and later optimum cutoff value found using ROC curve and Sensitivity-Specificity view and calculated following-
  Accuracy:77%,Sensitivity:83% and Specificity:73%
- As CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%,we proceed with Sensitivity-Specificity view instead of Precision-Recall view
- Lead score was assigned to data and conversion rate found to be above 80%

Step 7: Making Predictions on Test Data

- Scaling and prediction done on test data
- Found Sensitivity,Specificity and Accuracy to be similar to Train data
- Lead score was found and conversion rate calculated to be above 80%

It was found out that the variables that mattered the most in the potential buyers

- What is your current occupation_Working Professional
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website