SUMMARY

Analysis is done for X Education and to find ways to get more industry professionals to join the 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.

Following steps have used for analysis:

1.CLEANING DATA:

- Dropped the columns that had single value
- Columns having Select value was replaced with Null value
- Dropped the ID column which do not contribute to lead conversion
- Dropped the columns that were created sales team
- Dropped the columns that have more than 40% null/missing values
- Dropped the columns for which data was highly skewed
- Imputed missing value with median for numeric columns and most occurring value for string columns
- For columns having more than 5 unique categorical values, less occurring categories were grouped into 'Other' category
- Outlier detection and treatment

2.EDA:

- A quick EDA was performed to check the our data
- Univariate/Bivariate/Multivariate analysis was done
- Many columns were having more than 5 unique categorical values. So less occurring categorical values were converted into Other

3.DUMMY VARIABLES:

- For all categorical columns dummy variable were created.

4.SCALING:

- Used standars scaler for numeric continuos variables

5.TRAIN-TEST SPLIT:

- Split was done at 80% and 20% for the train and test data respectively.

6.MODEL BUILDING:

- RFE was done to attain the top 15 relevant variables.

- The variables were removed manually depending on the VIF values and p-values.(the variable with VIF<5 and p-value<0.05 were kept).

7.MODEL EVALUTION:

- A confusion matrix was made.
- The optimum cut off value 0.35 (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 79.38% each.

8.PREDICTION:

- Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 78.73%.

9.PRECISION-CALL:

- This method was also used to recheck and a cut off of 0.41 was found with Precision around 70% and recall of around 79% on the test data frame.

10.CONCLUSION:

Below are the variables that contribute most in the probability of a lead getting converted

- Lead Origin for Lead Add Form
- Last Activity for SMS Sent
- Last Activity for Email Opened
- Last Activity for Other
- Total Time Spent on Website
- Lead Source for Olark Chat
- Last Activity for Page Visited on Website
- Lead Source for Google