SUMMARY

The analysis is done for X Education to get more industry professional to join their online courses. To make the process more efficient, company wishes to identify the potential leads i.e. 'HOT leads' so that lead conversion rate should go up by focusing only on potential leads. The basic data provided to us gave lot of ideas about potential customer and conversion rate.

The following steps used to build the model to identify the hot leads and achieve conversion rate of around 80%: -

1. **Data Understanding and data cleaning:** -
   The data file has been understood with the help of data dictionary. The data has been checked for duplicate and unique values and option 'Select' has been replaced with NaN value as these values are as good as missing values. Columns having one/unique value and >40% null values have been dropped. Few of the columns (like Specialization, City, Tags etc) having null values have been changed to 'not provided' so as to lose not much data. As many of them belongs to INDIA, the country column changed to 'India',' outside India' and 'not provided'.

2. **EDA: -**
   Univariate analysis and Bivariate analysis have been done for categorical & numerical features it is found that most of the categorical features are irreverent.
   Outliers treatment have been performed for numeric features.

3. **Dummy Variable creation: -**
   Dummy variable has been creating for categorical columns and dummies with 'not provided' has been dropped.

4. **Train-Test Split: -**
   The split has been done at 70% and 30% for train and test data respectively.

5. **Feature Scaling: -**
   Feature scaling has been performed for numerical columns using min-max scaler.

6. **Model Building: -**
   RFE was done to get the top 15 features. Later features have been removed manually to maintain the VIF<5.0 and p-value<0.05.

7. **Model Prediction: -**
   Prediction has been done on train data frame with cut-off as 0.41.

8. **Model Evaluation: -**

   Confusion matrix has been created and using ROC curve, optimum cut-off value has been used to find the Accuracy, Sensitivity and Specificity which came to be around 81%.

9. **Precision –Recall: -**

   This method has been used to recheck and a cut-off of 0.41 has been found with precision 73% and recall around 78% on train data frame.

10. **Prediction on test data frame: -**

    Prediction has been done on test data frame and with an optimum cut-off as 0.36 with Accuracy of 79% , Sensitivity of 80 % and Specificity of around 79%.

11. **Conclusion: -**

    The Model seems to predict the conversion rate well with Accuracy, Sensitivity and Specificity values around 79%, 80% and 79.2% respectively; this is close to the respective values calculated during Train set.

    Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%.

    Features responsible for good conversion rates are: -

    a. Total Time spent on website

    b. Total number of visits

    c. When lead source was: - Google, Direct Traffic, Olark chat and Organic search (in the same order)

    d. When last activity was: - SMS, Email opened

    e. When the current occupation is - Working Professional

    f. Lead origin: - Landing Page Submission and Lead Add Format bring higher number of leads as well as conversion.