SUMMARY

This report is regarding X Education to find the ways to get more leads to join for the courses. From the data we can analyse potential customers like who visits frequently, total time spent and how they reach the portal.

The steps are

1. Cleaning Data:

The data was partially cleaned and remaining values are replaced with ‘Not Specified’ not to lose mush data for further analysis. Next we replaced with dummy variable to extract the correct leads for further analysis.

2. EDA :

Now we perform EDA for analysing the data. From the analysis we identified that lot of elements in categorical variables are irrelevant and also numerical values are good with minimum outliers .

3. Dummy Variables:

We created dummy variables for the categorical values and also removed all the repeated and redundant variables.

4. Train-Test Split:

By the next step the data was split into train and test data set with the proportion of 70-30% respectively.

5. Feature Scaling :

a. Here we used Standard Scalar to scale the original numerical values.
b. Then, we plot the heatmap to check the correlation among the variables.
c. Dropped the highly correlated dummy variables.

6. Model Building:

Firstly , RFE (Recursive Feature Elimination) was done with top 15 relevant variables. Later the rest of the variables are removed depending on VIF(<4) and p-values (<0.05).

7. Model Evaluation:

Based on the initial assumptions , we derived Confusion Metrics. By using the values of Confusion Metrics we can derive ‘Accuracy’ , ‘Sensitivity’ and the ‘Specificity’ .From the values we can evaluate how reliable the model .

8. ROC Curve:

After plotting ROC Curve for the feature and curve came out be pretty decent with an area coverage of approximately 88% which further solidified the model.

9. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy , sensitivity and specificity of approximately 80%

10.Precision and Recall:

Train Data:

Accuracy : 75.54%

Sensitivity : 90.70%

Specificity : 61.66%

Recall : 90.70%

Test Data:

Accuracy: 73.78%

Sensitivity: 89.98%

Specificity: 66.26%

Recall: 89.37%

The variables which are most significant found to be

1. **Total time spends on the Website**
2. **Total number of visits.**
3. **When the lead source was:**
   a. **Google**
   b. **Direct Traffic**
   c. **Organic search**
   d. **Welingak website**
4. **When was the last activity was:**
   a. **SMS**
   b. **Olark chat conversation**
5. **When their current occupation is a working professional**.

By the above assumptions X Education company can grow with a very high chances to get most leads to be hot leads.