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 following are the steps used:

1. Data Reading and Understanding: Here we tried to get the look and feel of the data, we observed the following things

- Number of rows and columns
- Data types of each column
- Checking the first few rows how the data looks
- Checking how the data is spread.

2. Data Cleaning: Here we checked for discrepancies in the dataset

- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods

3. Data Visualization and Outliers Treatment:

- We performed univariate analysis on categorical columns to see which columns make more sense and removed those columns whose variance is nearly zero.
- We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- We performed univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
- We performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.

4. Feature Scaling: At this stage, our data was very clean and no outliers. We know that logistic regression takes the input parameters as numerical values. Hence, we converted all the categorical columns to numerical.

- Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.
- Columns which have more than two levels were converted to dummies using the pd.get_dummies function.

5. Train-Test split: Split the dataset into train and test datasets and scaled the dataset. b. After this, we plot a heatmap to check the correlations among the variables. c. Found some correlations and they were dropped. The conversion rate is 37.85.

6. Model Building: We have used Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain. RFE uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

7. Model Evaluation: A confusion matrix was made.

1. Train Data:
   - Confusion Matrix:

     ```
     [[3653,  252],
     [ 351, 2095]]
     ```
   - Accuracy: 90.5%
   - Sensitivity: 85.65%
   - Specificity: 93.54%
   - Precision: 89.26%
   - Recall: 85.65%
2. Test Data:
   - Confusion Matrix:

     ```
     [[1185,  549],
     [  71,  918]]
     ```
   - Accuracy: 77.23%
   - Sensitivity: 93%
   - Specificity: 68.34%
   - Precision: 63%
   - Recall: 93%

CONCLUSION:

Learning gathered are below:

i. Test set is having accuracy, recall/sensitivity in an acceptable range.

ii. In business terms, our model is having stability and accuracy with adaptive environment skills. This means it will adjust to the company's requirements and changes made in the coming future.

iii. This concludes that the model is in a stable state

iv. Top features for good conversion rate:

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Total Time Spent on Website