SUMMARY

This analysis is done for X Education and to find ways to improve its lead conversion rate. The company aims to achieve a target lead conversion rate of 80%. The main goal is to build an adaptable logistic regression model for company to target most potential leads and assigns lead scores to each lead – higher the lead score, more promising the lead is.

STEPS TAKEN:

1. Reading and Cleaning the Data:
    a. Reading & Understanding the Data
    b. Cleaning the Dataset: The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'others' so as to not lose much data and some were replaced with median values (Occupation, city)
2. EDA and Data Visualization:
    a. Univariate and Bivariate Analysis
3. Data Preparation for Model Building
    a. Model Pre-processing - Train Test Split
    b. Model Pre-processing - Feature Scaling
4. Model Building & Feature Selection
    a. Running Initial GLM Regression Model with all the features using Statsmodel
    b. Feature Selection using Recursive Feature Elimination (RFE)
    c. Model Building
5. Model Evaluation - Confusion Matrix, Accuracy, Sensitivity (Recall), Specificity, Precision, etc.
    a. Plotting the ROC Curve
    b. Finding the Optimal Cut-off Point
    c. Precision and Recall Making
6. Predictions on the Test Set

CONCLUSION:

With cut off at 0.35 and considering Precision-Recall trade off , difference between values on Test and train data is very less(less than 2%).Both Precision and Recall on both Test and Train data is around 80%

The following conclusion can be made from out model:

1. Lead Add Form have a conversion rate of 94. Landing Page Submission & API have a conversion rate of 36% and 31%
2. Welingak Website and Reference are the Sources which have the best conversion rates Google and Direct Traffic Sources generate the most leads out of all sources but their conversion rates are around 40% and 32% respectively.
3. Most people do not like to be called or emailed and those who do like to be called are more or less interested in the product and hence the conversation rate can be seen as 100%.
4. Those people spending more time on the website are more likely to convert on average.

5. Maximum leads are generated via SMS Sent and Email Opened categories and the conversion rates are 63% and 36% respectively.
6. Almost all of the specializations that are recognized by the company have a conversion rate between 30% - 50%.
7. Working Professional and Unemployed categories seems to bring in a lot of leads with a conversion rate of 92% and 43% respectively

Similar accuracy on train and test data's performance metrics suggests that the final model didn't overfit on the training data and is performing well. Higher sensitivity will ensure that the leads who are likely to get Convert are correctly predicted where as high Specificity will ensure that the leads that have a lower probability of getting converted are not selected. Depending on the business requirement, we can increase or decrease the probability threshold value which in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.