Summary

Problem Statement:

The problem statement was to build a Logistic Regression Classifier to help X Education company identify hot leads through various diversified parameters and classify whether a lead will potentially turn up and buy the course they are offering.

Leads dataset was provided to use containing 9240 data points and 37 different parameters (columns). The data had some missing values and some unwanted columns so data needed to be prepared first before proceeding ahead with building the machine learning model.

The following steps are followed to achieve the objective of this case study:

1. **Loading and Understanding the data**: The data was cleaned of all unwanted columns and columns with lot of missing values and in the columns with minimal missing values, the rows containing missing values were dropped.
2. **EDA**: then we come to EDA. Here we do univariate and bivariate analysis. In univariate analysis we check the count-plot for categorical variables and histogram plots for continuous variables. Then in bivariate analysis, we see the conversion rate of different parameters with "converted" column.
3. **Dummy variables creation**: then the categorical variables having more than 3 categories are split by using dummy variables and the original columns are dropped for better analysis.
4. **Splitting data**: now the data is split into training and testing data sets in 70:30 ratio and the data is now ready for model building.
5. **Model Building**: since it is required to give a score between 0 and 10 to each lead based on his probability of conversion, we are going with Logistic regression model. Then based on a certain threshold obtained from accuracy-sensitivity-specificity plot which was found to be 0.35, we can categorize the lead as converted (1) or not converted (0). During building this model, feature selection is done using REF and some columns having p-values > 0.05 and VIF values >5 is dropped for better accuracy.
6. **Model Evaluation:** now the model is to be evaluated by checking the ROC curve to see the area under the curve is high or not. Based on the ROC curve, if the True Positive Rate is high and False Positive Rate is low, then the model built is good. We also found the values of Accuracy, sensitivity and specificity of the model which were all found to be close to 80% which is quiet acceptable.
7. **Predictions of Test set:** now the model is fed the test set and the entire testing of accuracy, sensitivity and specificity were done and were found to be close to 80% again. The precision and recall values of the model on test set are verified based on precision-recall tradeoff curve, we found that a cutoff of 0.41 might be a better value.

8. **Evaluation of model performance on test set:** based on the new cutoff, we found that the model performed with close to 80% accuracy again. So the model predicting the lead conversion with 80% accuracy.

Conclusions:

Based on our observations the following three parameters were found to effect the conversion rate of hot leads more than others.

- Total time spent on the website
- Total number of visits
- When lead source was Google.