Summary

The analysis is done for X education and to find ways to get more industrial professionals to join their courses. The basis data provided gave us a lot of information about how the customers visit the site, the time they spend there, how they reached the site and the conversion rate.

Solution Summary:

1) Reading and Understanding the data:
Read and analyze the data

2) Data Cleaning:
We dropped the variables which had high null values and imputing the missing values where required with median values in continuous variables and creation of new variables in categorical variables. Identified the outliers and removed.

3) Data Analysis:
Started with Exploratory Data Analysis of the dataset to get the feel of how data is oriented.

4) Data Preparation:
We created dummy variables for categorical variables and deleted the repeated variables.

5) Train-Test Split:
Divide the given dataset into train and test with ratio of 70:30 percent value.

6) Feature Scaling:
We used StandardScaler to scale the original continuous variables. Then we build the initial model using stats model which gives us statistical view of all the parameters of the model.

7) Feature Selection using RFE:
Using RFE we select top 15 important features and we looking at the p-values in order to select most significant values and dropped the insignificant values. Also checked VIF value and it found to be good. Also we derived the confusion metrics and calculated the overall accuracy of the model. Also we calculated sensitivity and specificity metrics to understand how reliable the model is.

8) Plotting ROC curve:
We plotted the ROC curve for the features and curve came out to be pretty good with area coverage of 89%.

9) Find Optimal Cutoff Point:
We plotted the graph for the Accuracy, Sensitivity, Specificity for different probability values. The cutoff point was found to be 0.37. We could also observed that the accuracy = 81%, sensitivity = 70%, specificity = 88%.
Also calculated the lead score and figure it out that the final predicted variables approximately gave a target lead prediction of 81%.

10) Computing the Precision and Recall metrics:
We found the precision and recall values came out to be 79% and 70% respectively on train dataset. Based on precision and recall tradeoff we get a cutoff value of approximately 0.42.

11) Making prediction on Test set:
We predicted on test set and calculated the conversion probability based on the sensitivity and specificity metrics and found out to be accuracy = 81%, sensitivity = 81%, specificity = 80%.