# Lead Scoring Case Study- Summary

By:
Hayagreeva Sudarshan Sundareswaran
Sanskruti Babar
Sangamitra Senthil Kumar

The analysis was done on the Lead Scoring Case Study to find ways to get more industrial professionals to join their courses or programs. The basic data given to us was about the customers visiting the X Education website, the time they have spent there, their choices, how they found out about the site, and the conversion rate.

In the analysis that was done, the following steps were taken:

1. Data Quality Checks and Data Cleaning:

   After the data was imported and read, it was checked for any null, duplicated, or incorrect values. There were no duplicates present, but there were null values, that were handled.

2. EDA

   The data we had was visualized to check the condition of it. We found that most of the categorical values were irrelevant to the data and hence were dropped from the data set. There were some outliers in the data set, but they were handled.

3. Dummy Creation:

   Dummy variables were created for the categorical values present in the data set.

4. Train_Test_Split and Scaling:

   The data set split was done at % for train and test respectively. It was then scaled using the StandardScaler() function.

5. Model Building

First of all, RFE was done to obtain the top relevant variables. Then the rest of the irrelevant variables were removed manually, depending on their VIF and p-values. The variables with VIF<5 and p-values<0.5 were kept.

6. Probability Prediction:

The prediction probability was done on the test data set with 0.28 as the optimum cutoff with accuracy, sensitivity, and specificity of 80%.

7. Creating Confusion Matrix:

A confusion matrix was made to find the optimum trade-off value using the ROC curve.

8. Plotting ROC curve:

An ROC curve demonstrates several things: It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows to the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

9. Precision and Recall:

This method was used to recheck and the trade-off found was 0.33 at 79 and 85%

10. Lead Scoring:

After running the model on the Test Data these are the figures we obtained:

Accuracy = 81.10%, Sensitivity = 77.08%, Specificity = 83.54%

After the final analysis, the following can be recommended to X Education:

- The Leads that have a higher 'Lead Score' should be focused on more for a better conversion rate.

- Encouraging existing converted leads for referrals by providing some kind of reward for doing so.
- The unemployed category should be focused on more.
- Students should be focused less, as they have a lower conversion rate.