

# Lead Scoring Case Study Summary

## 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. **Cleaning data:** 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 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.
2. **EDA:** A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found
3. **Dummy Variables:** The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used StandardScaler.
4. **Train-Test split:** The split was done at 70% and 30% for train and test data respectively.
5. **Model Building:** Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
6. **Model Evaluation:** A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around accuracy -92%, sensitivity-85% and specificity-96% each.
7. **Prediction:** Prediction was done on the test data frame and with an optimum cut off as 0.27 with accuracy(80%), sensitivity(93%) and specificity of 73%.
8. **Precision:** This method was also used to recheck and a cut off of 0.41 was found with Precision around 67%.

It was found that the variables that mattered the most in the potential buyers:

Lead_Origin	Lead_Add_Form	0.879273
Lead_Source_Welingak	Website	3.173203
Last_Activity_SMS_Sent		1.903746
Tags_Busy		3.751529
Tags_Closed_by_Horizzon		8.710178
Tags_Lost_to_EINS		9.645545
Tags_Ringing		-1.727516

Tags_Will revert after reading the email	3.894197
Tags_switched off	-2.358825
Lead Quality_Not Sure	-3.510688
Lead Quality_Worst	-3.964046
Last Notable Activity_Modified	-1.734055
Last Notable Activity_Olark Chat Conversation	-1.306967