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 important 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 specified' or 'missing_city' so as to not lose much data.

2. Data Preparation: A quick EDA was done to check the condition of our data. It was found few elements in the categorical variables were irrelevant. For categorical variables where to many unique values are present, 'Others' is used to group the values which are <0.5% . To handle outliers we have used capping to 99th Percentile. For numeric values we used the StandardScaler.

3. Train-Test split: The split was done at 70% and 30% for train and test data respectively.

4. Model Building using RFE: Logistic regression is used and 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 < 0.005 were kept). Overall Accuracy of the model is calculated.

5. 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 90% and 94% respectively.

6. Prediction: Prediction was done on the test data frame and with an optimum cut off as 0.25 with accuracy, sensitivity and specificity of 93%.

7. Precision – Recall: This method was also used to recheck and a cut off of 0.42 was found with Precision and recall around 90% on the test data frame.

Overall accuracy was around 93%

It was found that the variables that mattered the most are (In descending order):(Positives)

1. Tags lost to EINS. 2. Tags closed by Horizzon. 3. Tags_Will revert after reading the email 4. Tags_Not_Specified 5. Lead Source_Welingak Website 6. Tags_Busy. 7. Last Activity_SMS Sent.

(Negatives)

8. Last Notable Activity_Modified 9. Tags_Ringing 10. Tags_switched off 11. What matters most to you in choosing a course_Crse_Not_Specified

X education company needs to focus on following key aspects to improve overall conversion rate:

- Focus on the top 3 tags which are very positive for business.
- Focus on working professional & unemployed who have high conversion rate.
- Focus on rewards for Referrals as the conversion rate is high.