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.

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.
   Features having 3000+ null values were removed.
   Features not useful were removed.
   Null rows of some features were removed as these features were important for further analysis and cannot be removed.
   Yes/No variable were converted to 1/0.
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.
3. Dummy Variables: The dummy variables were created. For numeric values we used the MinMaxScaler.
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 79.04% 78.82%, and 79.23% respectively for train data.

7. Prediction: Prediction was done on the test data frame and with an optimum cut off as 0.45 with accuracy, sensitivity and specificity of 78.97%, 78.05% and 79.81% respectively.
8. Precision – Recall: This method was also used to recheck and a cut off of 0.42 was found with Precision of 78.49% and recall of 77.29% on the test data frame.

What mattered most in converting a lead is:-

1. The total visits a person does on the website.
2. The total time he sends on the website.
3. The origin of the lead was Form.
4. Last interaction with the lead was a phone conversation.
5. The lead source was:-

1. Welingak Website
2. Olark Chat