Leads score 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 the base data:
    - The data was half 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% null values are removed as we cannot impute this minute null values.
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 seem good and no outliers were found. Some columns were dropped if they skewed towards converted or not converted.
3. Dummy Variables:
    - The dummy variables were created for all categorical values. For numeric values we used the Scalar Transformation.
4. Train-Test split:
    - The split was done at 70% and 30% for train and test data respectively with randomness = 70.
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). Totally 3 models are created out of which 3rd model is selected wit accuracy of 84.8%.
    - Optimum cutoff was calculated at point where sensitivity, accuray, specificity are maximum.
    - It came out to be 0.35..
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 84% each.
7. Prediction:
    - Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 85%.
8. Precision – Recall:
    - This method was also used to recheck and a cut off of 0.41 was found.

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

- const
- Do Not Email
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Reference
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- What is your current occupation_Working Professional
- What matters most to you in choosing a course_Other
- Lead Quality_Low in Relevance
- Lead Quality_Might be
- Lead Quality_Not Sure
- Lead Quality_Worst
- Last Notable Activity_Modified