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.

Following steps are used:

1. Understanding the data: Load, read and analyse basic data.
2. Cleaning the data: The data is partially cleaned except for a few null values and option select is also replaced with a null value as it was not having enough information. Some Null values are replaced with 'not provided' so as we do not lose much data. However, the same at later stage is removed while making dummies. Also, categorise data as from 'India" , 'Outside India' & 'Not provided'.

3. Data Modelling (EDA) : While doing Exploratory data analysis , it was found that a lot of elements in categorical variables are irrelevant. The numeric value seems good with no outliers.

4. Dummy Variables: Then created the dummy variables for categorical values.

5. Train test split: In this step the data was split into train and test data set with a probability of 70-30% respectively.

6. Feature scaling : Used MinMaxScaler to scale the original numerical values.

7. Model Building : We have used Recursive Feature Elimination to get the top 15 most relevant variables. Now depending on the VIF values and p-value the rest of the variables are removed.

8. Model Evaluation: Taking an initial assumption that a probability of more than 0.5 means 1 else 0, we created the data frame having the converted probability values. Based on the initial assumption we derived the Confusion Metrics. We further Calculated the overall Accuracy , Sensitivity and Specificity of the model matrices to evaluate how reliable the model is.

9. Plotting the ROC curve: Further plotted the ROC curve for the features and the curve came out pretty good with an area coverage of approximately 88% which further solidified the mode.

10. Prediction : It is done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of approx.. 80%.

11. Precision Recall: This method was also used to recheck and a cut off of 0.41 was found with Precision around 78% and recall around 69% on the test data frame. It was found that the variables that the most significant are (in descending order):
    --- Total time spent on the website.
    --- Total count/ number of visits on website.
    --- What was the lead source of visits i.e. Google / Direct traffic / Organic search / Welingak website
    --- Time of Last activity
    ----Occupation etc.

Keeping the above mentioned points in the mind X Education can increase all the potential buyers to change their mind and buy courses.