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.) **Data Cleaning:**

The data was partially clean except for a few null values and the option 'select' in some categorical variables had to be treated as a null value since it did not give us much information. We dropped quite a few columns because of a lot of missing values and also because of lack of variety in some categorical variables. And, for the columns with lesser number of null values, we dropped the rows containing the null values. *Ultimately, after all the data cleaning process we were able to retain 69% of the records.*

2.) **Data Preparation:**

Dummy Variables were created for the categorical variables and the original categorical variables were dropped. After that we did *train-test split (70-30) taking random_state=42.* We also scaled the train set using *MinMaxScalar*.

3.) **Model Building:**

Firstly, coarse tuning was done using *RFE* to attain the *top 15 relevant variables*. Later the rest of the variables were removed manually (fine tuning) using statsmodels depending on the VIF values and p-value *(The variables with VIF < 5 and p-value < 0.05 were kept)*. After 3 iterations, we were able to get all the statistical values under the above mentioned threshold.

4.) **Model Evaluation:**

A data frame (containing Converted and Predicted variables) was made to get the confusion matrix. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 78% each.

5.) **Prediction:**

Prediction was done on the test set and with an optimum cut off as 0.42 which gave us the accuracy, sensitivity and specificity of around 79-80% each.

6.) **Precision-recall:**

This method was also used to recheck and a cut off of 0.44 was found with Precision 80% and recall 77% on the test data frame.