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:

- **Cleaning data:**
  • Columns such as 'Tags', 'Lead Quality', 'Asymmetrique Activity Index','Asymmetrique Profile Index', 'Asymmetrique Activity Score',Asymmetrique Profile Score', 'City' , 'Country’ are dropped as they have more than 35% missing values.
  • Columns which plays no role for our analysis as most of the data points have only one value are also dropped.
  • For columns "What is your current occupation“, “TotalVisits”, “Lead Source” and “Specialization” null value rows have been dropped

- **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 seems good and no outliers were found.

- **Data prepration:**
  • Created dummy variables for all categorical variables.
  • Converted binary variables such as “Do Not Email” to 0/1

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

- **Feature scaling:**
  Scaling the feature "TotalVisits", "Total Time Spent on Website", "Page Views Per Visit“ using MinMaxScalar

- **Model Building:**
  • Using RFE method to select features and running it with 20 variables.
  • Assessing the models with Statsmodels.
  • Columns containing p – values > 0.05 and VIF > 5 are insignificant. Hence the aim is to drop the columns with the mentioned criteria.

- **Model Evaluation:**
  Model evaluated on the train dataset have an accuracy of 79% with sensitivity and specificity as 74% and 83% respectively

- Making predictions on test datasets:

Accuracy for the test dataset is found to be 78%.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
   a. Google
   b. Direct traffic
   c. Organic search
   d. Welingak website
4. When the last activity was:
   a. SMS
   b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

X-----X-----X------X