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:

Data Cleaning / Data Preprocessing:

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. Few of the null values were changed to ‘not provided’ so as to not lose much data. There are 'Select' values in many columns.It may be because the customer did not select any option from the list, hence it shows 'Select'.'Select' values are as good as NULL. So we can convert these values to null values. Dropping the columns with missing values. Retained 98% of the rows after cleaning the data .

Exploratory Data Analysis

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. Univariate Analysis and Bivariate Analysis. Plotting heat map to see the correlation.

Data Preparation

In data preparation three steps are follow:

1.Dummy Variables

The dummy variables were created and later on the dummies with ‘not provided’ elements were removed. Creating a dummy variable for the categorical variables and dropping the first one. After dummy creation five rows and 78 columns are remaining.

2.Train-Test split

The split was done at 70% and 30% for train and test data respectively. Splitting the data into train and test.

3. Scaling

In scaling check the Lead Conversion rate which is 38%.

Model Building

Firstly, RFE was done to attain the top 20 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). Assessing the model with StatsModels.

Creating Prediction

Making Prediction on the Train set and *Creating a dataframe with the actual Converted flag and the predicted probabilities.*

Model Evaluation

A confusion matrix was made. *chosen an arbitrary cut-off value of 0.5. We need to determine the best cut-off value and the below section deals with that. Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.*

Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

Prediction

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

Precision

Recall: This method was also used to recheck and a cut off was found with Precision around 80% and recall around 81% on the test data frame.

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

1. The total time spent on the Website.

2. Total number of visits.

3. When the lead source was:

- Google
- Direct traffic
- Organic search

- Welingkar website .

4. When the last activity was:

- SMS
- 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.