Lead Scoring Case Study 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.Importing the required libraries :

The python libraries required for data analysis was imported.

2. Reading and understanding the data:

The data “Leads.csv” was imported to Jupyter notebook.

The data and the data dictionary was thoroughly inspected.

3.Data inspection and basic sanity check:

The information , description,shape of the data, null values and missing values in the data was checked.

4. Cleaning the data:

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. Select values were replaced with ‘nan’ so as to not lose much data. Columns with large number of missing values were dropped.

5. EDA:

Exploratory Anlysis 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 few outliers were found which should be treated. The variables were further subjected to univariate and bivariate analysis with respect to target variable.

6. Data Preparation and Features Selection:

Further sanity checks were made, dummy variables were created and the features were scaled using Standard scaling technique.

7. Train-Test split:

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

8. 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.

10. 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 .

11. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.37 with accuracy 77.3%, sensitivity 79.8% and specificity of 75.7%.

Precision – Recall:

This method was also used to recheck and Precision was around 68.5% and recall around 80.9% on the test data frame.

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

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. Olark chat

4. When the last activity was:

   a. SMS
   b. Olark chat conversation

5. When the lead origin is Lead add format.

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.