Summary

Education is the most important part of our life.Today there is such an easy way to get an education that we do not need to go anywhere to get an education. To take education, we can get education from the teacher just sitting at home from the online mood.In today's time, facilities like the internet are available in all the homes. Online education is proving to be very effective in the time of Corona. Nowadays online education is becoming very prevalent everywhere whether it is village or city.

In the changing environment, there have been many changes in technology and its use is also big. Many changes have also been seen in the way of taking education due to technology. Today, the teaching-related material used in online education can be sent from one place to another through technology online.No matter where we are in the world, we can get the learning material delivered to another place in no time. Like any link, any video related to education, any file. All these types make online education even more creative.

Such company is X Education which sells online courses to industry professionals

For the given situation we were asked to create a model which gives the clarity about approaching the highest probable customers who can purchase their courses.

The analysis for the given assignment was done by following the below steps :

1.Loading the dataset :

The given dataset was imported in the jupyter notebook and was loaded to visualise the given data.

2.Inspecting the dataset :

The dataset was further inspected with the entries available, null values present etc. and changes were made accordingly.

3.EDA :

Tha basic EDA was performed to infer few important business based insights from the available variables which might be dropped further.

4.Cleaning :

The dataset is properly cleaned and thee required variables were dropped to have a proper availability of the resources to build a model ahead.

5.Creating the dummy variables :

Dummy variables were created according to the requirement and were also scaled where necessary.

6.Test-Train Split :

The dataset was then splitted into the training and testing datasets with a 7:3 ratio or 70% & 30% distribution respectively.

7.Model Building :

Finally on the relevant 15 variables that were obtained through RFE, with logistic regression we moved ahead for preparing the model. The models were summarised, then were further refitted by looking at the P values and VIF'sof the variables. The final model was obtained for the fifth time was finalised where all the P values were less than (<)0.05 and VIF's were less than(<)5 .

8.Model Evaluation :

To evaluate the model a confusion matrix was introduced whereafter we used **ROC** curve to obtain the optimum cut off, being **0.42**. After which the prediction was changed by keeping the same as the threshold value for the cut off. All the factors Accuracy, Sensitivity and Specificity revolved around the given value of 80%.

9.Prediction :

The prediction was done on the test set with 0.42 cutoff and Accuracy, Sensitivity and Specificity being nearly 80%.

10.Precision-Recall score :

Finally precision score and recall score were calculated where we acknowledged that we've got a higher recall score than the precision score. As we already know, a higher precision means that an algorithm returns more relevant results than irrelevant ones, and high recall means that an algorithm returns most of the relevant results (whether or not irrelevant ones are also returned). So, the obtained scores for precision and recall i.e **71.3%** and **83.5%** were good for the model prepared for the company.

11.Lead Score :

Finally the feature of lead score was provided for the company in the model to identify the most probable converting customers.

Few of the conclusions that we drew from the model were :

1. The Accuracy, Precision and Recall score we got from test set are in an aceptable range.
2. The recall score is higher than the precision score which is a good indicator for the model that we prepared.
3. Looking at the above metrics we can say that the model has an ability to adjust with the company's future requirements in coming years.

The most important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

1. Total Time Spent on Website
2. Lead Origin Lead Add Form
3. Last Activity_Had a Phone Conversation

Thus this is how the problem was approached by us and the model was prepared accordingly keeping the future aspects in mind.