Summary

The analysis is done for an education company, X Education, to find out ways whereby more industry professionals join their courses. The basic data outlined information on the potential customers who visit the site, the amount of time they spend there, how they reach the site and the conversion rate.

The goal of the case study was to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which could be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

The following steps were used:

1. ***Cleaning of 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 provide much insight. They were later removed while making dummy variables.

2. ***EDA***: 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 numerical values seemed good. No outliers were found.

3. ***Dummy Variables***: Dummy variables were created, those with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

4. ***Train-Test Split***: Train and Test data was split at 70% and 30% respectively.

5. ***Model Building***: 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 (The variables with VIF < 5 and p-value < 0.05 were kept).

6. ***Model Evaluation***: A confusion matrix was made, post which the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which was around 80% each.

7. ***Prediction***: Done on the test data frame with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. ***Precision*** – ***Recall***: Was used to recheck and a cut off of 0.41 was found with Precision around 73% and Recall around 75% on the test data frame.

Lead scoring case study was done using logistic regression model to meet the constraints as per business requirements.

There were a lot of leads in the initial stage but only a few of them were converted into paying customers. The most numbers of leads are from INDIA with Mumbai being the highest number.

Most of the leads have Specialization from Finance Management. Leads from HR, Finance & marketing management specializations have high probability to convert.

Improvement in customer engagement through email & calls will help to convert leads. Emails as well as sending SMS has a high probability to convert.

With the help of the above inputs provided, X Education can flourish as they have a very high chance to convert almost all potential buyers to change their mind and buy courses which would ultimately benefit them.