LEAD SCORE CASE STUDY SUMMARY

PROBLEM STATEMENT:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

SOLUTION:

The following are the steps used to get to a solution:

1.Reading and understanding the data:

The data is first read in python, the shape, number of missing values, types of variables etc is understood.

2.Data cleaning:

The null values percentage of each column is checked and the column with highest null val. percentage is removed, missing values are imputed with various median or mean or requisite entity.

3.Exploratary Data Analysis:

The data is plotted into graphical representation and graphical analysis of the data helps a get an overview of the data.

4.Dummy variable creation:

Dummy variables were created for Categorical variables which helped understand the data more.

5.Train-Test split:

The data was split into two parts Train and test with a 70-30% ratio.

6.Scaling of data:

The Scaling helped us convert all the units of the data into a common scale.

7.Feature Selection using RFE:

Selection of columns with important features were done contributing to the model analysis. A lookout for better significant model was done by checking P values and VIF values. Features with high P and VIF were eliminated and we got 11 significant features that contribute to the analysis.

8. Confusion Matrix:

Confusion Matrix was created and accuracy, sensitivity and specificity was found out

9. ROC curve:

Area under ROC curve was 0.8

10. optimal cut-off point:

- The sensitivity:0.73
- The Specificity:0.81
- The Precision:0.71
- The recall:0.73
- The f1 score:0.72

Optimal cut-off point is between 0.30 to 0.45

11. Prediction on test model:

The above info helped us made prediction on the test set we got accuracy of 0.76. The hot lead IDs were found out and focussing on these IDs were important for the organization.