Summary

LEAD SCORING CASE STUDY

Introduction - 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 rates.

Steps –

1. Data Reading and Cleaning: 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 non-provided so as to not lose much data.

They were later removed while making dummies. Since there were many from India

and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

2. EDA: 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.

The numeric values seem good and no outliers were found.

In the data many columns with null values so instead of dropping the we replaced it.

Noticed that, India is most occurring country so replace missing values with India.

3. Dummy Variables: The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

4. Train-Test split: The split was done at 70% and 30% for train and test data respectively.

5. Model Building: Firstly, RFE was done to attain the top 13 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. 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.

Sensitivity = 67.17651922281934

Sensitivity identifies all 1's Here its 67% which is ok.

Specificity = 88.86921325575511

Specificity is for negative values which is 88% is good.

7. Prediction: Accuracy is 80%. So reasonable model. As in this case, we are interested in identifying customers who would get converted. Here, the values are quite good. i.e., With the current cut off as 0.5 we have around 80% accuracy, sensitivity of around 67% and specificity of around 88%.

8. These are the variables which can be considered by the company in finding leads which can be converted.

```

1.Total Time Spent on Website,

2.Lead Origin_Lead Add Form,

3.Lead Source_Direct Traffic,

4.Lead Source_Organic Search,

5.Lead Source_Referral Sites,

6.Lead Source_Welingak Website,

7.Do Not Email_Yes,

8.Last Activity_Converted to Lead,

9.Last Activity_Olark Chat Conversation,

1.Last Activity_SMS Sent,

11.What is your current occupation_Working Professional,

12.What is your current occupation_not specified,

13.Last Notable Activity_Had a Phone Conversation,

14.Last Notable Activity_Unreachable
```