Lead Scoring Case Study - Summary

MOUNICA NAGULAPALLI | BHARATHNAN KUMARAN |SNEHA PADALA

Dec 2023 - Batch : DS C57 June 2023

Problem Statement:

X Education sells online courses to the industry professionals. X Education needs help in selecting the most promising leads i.e, the leads that are mostly likely to convert into the paying customers.

The company needs a model wherein a lead score is assigned to each of the leads such that the customers with the higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Solution Summary:

1.**Cleaning Data:**

The data was partially clean except for a few null values and the select had to be replaced with a null value since it did not give us much information. Few of the null values were changed so that much of the data has not been lost. Although they were later removed while making dummies. Since there were many from India and few from outside India, the elements were changed to 'India'.

2.**EDA:**

A quick EDA has done to check the condition of the data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers found.

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 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and the 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 the find the accuracy, sensitivity and specificity.

7.**Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.46 with below accuracy, sensitivity and specificity.

Accuracy: 0.78

sensitivity: 0.79

Specitivity: 0.78

8.**Precision- Recall:**

This method was also used to recheck and a cut off of 0.45 with below accuracy, sensitivity and specificity.

* Accuracy: 78 (No major difference between Initial model and recall)

* Sensitivity: 79 (No major difference between Initial model and recall)

* Specificity: 77 (No major difference between Initial model and recall)