LEAD SCORING 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.

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%.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step 1: Reading the data:

Reading and inspecting the dataset.

Step 2: Data Cleaning:

a. In the Data Cleaning step, we started with dropping the unique values.
b. Then, there were multiple columns the having value ‘Select’ which could mean that the leads did not choose any option. We changed them to Nan values.
c. Next, we dropped the columns having Null values less than 45%.
d. Then we checked the imbalance and redundant columns in which the missing values were required to be imputed as and where with Median in case of numerical variables and creation of new classification variables for categorical ones.
e. A few values in one column were having identical label with different class (first letter of the value small and capital respectively). We fixed this issue by replacing it with one with capital.
f. Outliers were removed and the variables generated by the sales team were removed to avoid any ambiguity.

Step 3: Data Transformation:

In this step, we changed the binary variables into 0 and 1.

Step 4: Data Preparation:

a. Train-Test Split, i.e., splitting the data for modelling 70-30 proportions.
b. Feature Rescaling using MinMax Scaler for numerical values.

Step 5: Model Building:

a. Using Recursive Feature Elimination, we selected the top 15 variables.
b. Looking at the p-values and we dropped the insignificant variables and selected the most significant ones.
c. Also, after checking the VIF valued of the final model we kept the 10 most significant variables.
d. We checked the Optimal Probability Cut-off by finding points and checking the accuracy, sensitivity and specificity.
e. Then, we plotted the ROC curve for the features and the curve came out to be decent.
f. 80% cases are correctly predicted, we then checked the Precision and Recall with Accuracy, Sensitivity and Specificity for our final model on the train set.
g. Next, on the basis of the Precision and Recall, we got the cut-off value of 0.2 approx.
h. After implementing the learning to the test data, we calculated the conversion probability based on the Sensitivity and Specificity metrics.
i. We found the values as Accuracy – 86.00%, Sensitivity – 86.33% and Specificity -85.79%.

Step 6: Conclusion:

- As per the lead score calculated in the test set of data shows the conversion rate around 83% on the final predicted model which clearly meets the expectations of the CEO.
- Good value of Sensitivity of our model will help to select the most promising leads.
- Features which contributed more toward the probability:
    a. 1.Lead Origin_Lead Add Form
    b. 2. What is your current occupation_Student
    c. 3. What is your current occupation_Working Professional