Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once the leads are acquired, employees from the sales team start making calls, writing emails, etc. so that leads get convert but currently lead conversion rate at X education is around 30%.

X Education has appointed us to help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

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

Solution Summary:

Step 1: Reading and Understanding the Data

We will Read the data and inspect the data.

Step 2: Data Cleaning

a. Firstly, we clean the dataset by dropping the variables with Unique data.
b. Then we drop the columns with having NULL values greater than the 45%.
c. Then we drop those rows who's having the Null values but its count is low as we have already drop lots of columns.
d. We have also removed the sales generated variables to avoid ambiguity in final solution.

Step 3: Data Preparation

a. Firstly we changed the binary variables into "0" or "1".
b. Then we Create the dummy variable for the categorical variables and also removed the original columns and the repeated columns.

Step 4: Test – Train Split

a. We will divide the data into two parts, First is the Train part and the other one is Test part i.e. 70% of data as the Train part and the other 30% of data will be the test part.

Step 5: Feature Re-scalling

a. We will scale the numerical variables with the help of Min Max Scalling.
b. Then we also check the Correlation among the variables and will drop the highly correlated dummy variable

Step 7: Model Building

a. Firstly we use the Recursive Features Elimination to select the top 15 important features.
b. Then with the help of Stats model, we recursively tried looking the P-Values in order to select the most significant values that are present and drops the insignificant values.
c. Finally, we arrive at the 8 most significant variables and the VIF's for these variables are also good.
d. Then for final model we find the optimal cutoff point and also check the metrics like Accuracy, sensitivity and Specificity.
e. Then we plot the ROC curve for the features and curve will comes as an ideal curve with having area coverage of 86%.
f. Then we check the Conversion percentage based on the converted values and it comes to be around 80%.
g. Then Based on the Precision and Recall, we found out that the Optimal Cutoff values is 0.4.
h. Then we find the Precision and Recall with accuracy, sensitivity and specificity for final model
i. Then on the basis of the model we will implement the learnings to the test set and found the metrics values to be Accuracy 78%, Sensitivity 79% and Specificity 77%.

Step 8: Conclusion

a. Lead Score is calculated on the Test set of data shows the conversion rate of 80% on the Final predicted model which is the exact same expectations of the CEO of the company.
b. Good values of the Sensitivity of our model will help to select the most promising leads.
c. Important features which contribute the probability of lead getting converted are:
    a. Lead Origin_Lead Add Form
    b. What is your current occupation_Working Professional
    c. Lead Source_Olark Chat