Lead Scoring Case Study Summary

Problem Statement:

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

The company is looking for a model where we 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 customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution Summary:

Step 1: Import the data into notebook and understanding it.

1) Data Cleaning and Preparation:

- We have checked for no of missing values in each column and dropped columns having missing values greater 3000(30% of the ~9000 columns).
- There are a few columns with value 'Select' which means the leads did not chose any given option. We applied countplot on the these columns and as can be seen that the levels of "Lead Profile" and "How did you hear about X Education" have a lot of rows which have the value Select which is of no use to the analysis so we dropped them.
- We noticed that, when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque. Since practically all of the values for these variables are No, it's best that we drop these columns as they won't help with our analysis.
- We have dropped rows in few columns where the null value is less. After the data cleaning we are left with 70% of data which seems good enough for our analysis.

2) Dummy Variable Creation:

- We found the categorical variables from the data and created dummy variables for these variables.
- For the specialization variable we created dummy variables specifically as it has 'select' data and we have dropped that level explicitly.

3) Test Train Split:

This Step to split the data into test dataset and train dataset with a 3:7 ratio giving 30% of data to train.

4) Feature Scaling:

We used MinMaxScaler() to scale the numerical variables. And created a heatmap to look at the correlations at he variables.

Step 2: Model Building:

a. Using the Recursive Feature Elimination (RFE), we went ahead with 15 selected features.
b. Using the statistics generated, we tried looking into the P-Values in order to select the most significant values that should be present and we have dropped the insignificant value and left with 14 features.

Step 3: Model Evaluation:

a. For out final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.
b. We have plotted the ROC curve of the features and the curve came out to be pretty decent with area coverage of 87%.
c. We checked the precision and recall with accuracy, sensitivity and specificity for final model on train dataset.
d. Based on precision recall trade off we choose 0.43 as our cutoff.
e. We have calculated the conversion probability based on the Sensitivity and specificity metrics and found out the accuracy value to be 79.3%, Sensitivity ti be 78.5 and Specificity to be 80.1 which supports our cut off point is good to go.

Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 79.3 on the final predicted model and it is around the expectation of CED has given which is around 80%.
- Good Value of sensitivity of our model will help in securing most promising leads.