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 needs a model wherein you a lead score is assigned to each of the leads such
that the customers with higher lead score have a higher conversion chance and the Customers with lower lead scores 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: Reading and Understanding Data.
Step 2: Data Cleaning.

a. First step is to clean the dataset. In this step we dropped the columns with least information, and imputed values with median or mean based on the type of variable.
b. In the second step, we have a few columns with value “Select” which means one doesn’t choose any option. We changed those values to null value.
c. In third step, dropped the columns having null values greater than 40%.
d. In fourth step, we have removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables.

Step 3: **Data Transformation.**

a. Changed the binary variables into ‘0’ and ‘1’

Step 4: Creating Dummy Variables.

a. We created dummy variables for the categorical variables.
b. Removed all the repeated and redundant variables.

Step 5: Test Train Split.

a. The next step was to divide the data set into train- 70% and test- 30%.

Step 6: Feature Rescaling.

a. We have used the Min Max Scaling technique to scale the original numerical variables.
b. Followed by plotting heatmap to check the correlations among the variables.

Step 7: Model Building.

a. Using the **Recursive Feature Elimination-RFE method**, selected the top 15 important features.
b. Then recursively tried looking at the p-values by creating different models in order to select the most significant values that should be present and dropped the insignificant values.
c. Finally, we arrived at the 8 most significant variables. The VIF's for these variables was also good and less than 4.
d. Model-8 is our final model and on that model we calculated the accuracy, sensitivity and specificity.
e. Plotted the ROC curve for the features and the curve came out be pretty decent with an area coverage of 84%.
f. Verified the precision and recall with accuracy, sensitivity and specificity for our final model on train datasets.
g. Now based on Accuracy, Sensitivity, Specificity, we got the optimal values of the three metrics at 0.3, so we chose 0.3 as our new cut off now.
h. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 79.21%, Sensitivity= 64.87%; Specificity= 87.56%.

Step 8: Conclusion

- With the current cut off as 0.3 we have Precision around 72% and Recall around 79%
- The lead score calculated in the test set of data shows that the Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model.
- Features which contribute more towards the probability of a lead getting converted are:
    - Lead Source_Reference
    - What is your current occupation_Student
    - Total Time Spent on Website