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 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:

Step1: Reading and Understanding Data:

- Read and inspect the data using the pandas library in Python.

Step2: Data Cleaning:

- First step to clean the dataset was to drop the variables having unique values.
- Then, there were few columns with value 'Select' which means the leads did not choose any given option. 'Select' values were replaced with Null values.
- The columns having NULL values greater than 35% were dropped.
- Next, The imbalanced and redundant variables were removed. 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. The outliers were identified and removed. Also, one column was having identical label in different cases (first letter small and capital respectively). This issue was fixed by converting the label with the first letter in small case to upper case.
- All sales team generated variables were removed to avoid any ambiguity in final solution.

Step3: Data Transformation:

- Changed the binary variables into '0' and '1'.

Step4: Dummy Variables Creation:

- Dummy variables were created for the categorical variables.

- All the repeated and redundant variables were removed.

Step5: Test Train Split:

- The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling:

- The Min Max Scaling technique was used to scale the original numerical variables.
- Then, a heatmap was plotted to check the correlations among the variables.
- The highly correlated dummy variables were dropped.

Step7: Model Building:

- Using the Recursive Feature Elimination, 15 top important features were selected.
- Using the statistics generated, P-values are considered in order to select the most significant values that should be present and drop the insignificant values.
- Finally, 11 most significant variables were considered for the model. The VIF's for these variables were also found to be good.
- For the final model the optimal probability cut off is checked by finding points and checking the accuracy, sensitivity and specificity.
- The ROC curve is plotted for the features and the curve came out to be pretty decent with an area coverage of 86% which further solidified the model.
- It is observed that 80% of the cases are correctly predicted based on the converted column.
- The precision and recall with accuracy, sensitivity and specificity for our final model on the train set is checked.
- Next, Based on the Precision and Recall trade-off, a cut off value of approximately 0.3 was decided.
- Then, the learnings to the test model were implemented and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 77.52%%; Sensitivity= 83.01%; Specificity= 74.13%.

Step 8: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 83% on the final predicted model which clearly meets the expectation of the CEO who has given a ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of the model will help to select the most promising leads.
- Features which contribute more towards the probability of a lead getting converted are:
  i. Lead Origin_Lead Add Form
  ii. What is your current occupation_Working Professional
  iii.What is your current occupation_Unemployed