Lead Scoring Case Study Summary

Problem Statement:

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

Solution Summary:

Step1: Reading & Understanding Data

First, we read the data and understood all the variables of the data.

Step2: Data Cleaning

- There were some null values present in the data, where some were in the form of the 'Select' option.
- We converted the 'Select' option values to the null value.
- We removed the null value columns where null value percentage was more than 40%.
- Next, we removed the imbalanced and redundant variables from the data. Which includes imputing the missing values with median in the case of continuous variable and mode in the case of classification variables.
- Then we started identifying the outliers present in the data and then removed them from the data. Also, in one column the presented data variables were in different case (i.e. some in upper case and some were in lower case) so converted the data to same case for all the variables.
- All the variables related to sales team were removed to avoid any obscurity in the final solution.

Step3: Data Transformation

Then we transformed the data to binary variables i.e. 1 and 0.

Step4: Creating Dummy Variables

Then created the dummy variable for all the categorical variables, and removed all the duplicity from the data.

Step5: Train-Test Split

Next, we started with the Train-Test Split. We divided the data into test and train data with the proportion of 30 and 70 respectively.

Step6: Feature Rescaling

- We used the Standard Scaler to scale the numerical variable.
- Then we plotted the heatmap to see the correlation among the data.
- Then we dropped the variables that were highly correlated.

Step7: Model Building

- Using the recursive feature elimination (RFE) we selected the top 15 features.
- Then based on statistical information we tried looking at the P-values and coefficients and selected the most significant values. And dropped the insignificant values.
- Finally, we arrived at the 11 most significant variables. The VIF for these variables was satisfactory.
- For our final model we gone through the ideal probability cut off by checking the accuracy, sensitivity, and specificity.
- Then we plotted the ROC curve and by analyzing that it was giving very good results with the area coverage of 87%.
- After that the most important part was to check whether the 80% cases are correctly predicted or not.
- We checked the Precision and Recall with accuracy, sensitivity, and specificity.
- Based on the Precision and Recall trade-off, we got the break-off value of approximately 0.35.
- Then we implemented the learnings on the test model and calculated the conversion probability based on the sensitivity and specificity and found out the following values:
  - Accuracy – 0.78
  - Sensitivity –0.76
  - Specificity – 0.79

Step8: Conclusion

- The final predicted model clearly meets the expectations of the CEO, as the provided target of 80% lead conversion rate has been meat by the test set.
- Features which accord more towards the probability of lead generation are:
  - Lead Source_Welingak Website: 4.9341
  - What is your current occupation_Working Professional: 2.6840
  - Lead Source_Reference: 2.5365