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 will have a higher conversion chance and the customers with lower lead score will 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:

1. Reading and Understanding the Data

- Read and inspected the data.

2. Data Cleaning

- First step to clean the dataset we chose was to drop the variables having unique values.
- Then there were few columns with value 'Select' which means the leads did not chose any given option. We changed those values to **NULL** values.
- We dropped the columns having **NULL** values greater than 35%.
- Next, we removed the imbalanced and redundant variables. This also included imputing the missing values 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, in one column was having identical label in different cases. We fixed this issue by converting the label with first letter in small case to upper case.
- All sales team generated variables were removed to avoid any ambiguity in final solution.

3. Data Transformation

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

4. Dummy Variables Creation

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

5. Test Train Split

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

6. Feature Rescaling

- We used the Min-Max Scaling to scale the original numerical variables.
- We plotted the heatmap to check the correlations among variables.
- Dropped the highly correlated dummy variables.

7. Model Building

- Using the RFE, we went ahead and selected the top 15 important features.
- Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
- Finally, we arrived at the 11 most significant variables. The VIF's for these variables were also found to be good.
- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- We then plot the ROC curve for the features and the curve came out to be pretty decent with an area coverage of 86% which further solidified the model.
- We checked if 80% cases are correctly predicted based on the converted column.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
- Next, based on the precision and the recall trade off, we got a cut off value of approximately 0.3
- 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 77.52%
- The sensitivity was found out to be 83.01% and specificity was 74.13%.

8. Conclusion

- The lead score calculated in the test set of data shows the conversion rate of 83% on the final predicted which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.