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 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:

Step1: Reading and Understanding Data:

Read and inspected the data

Step2: Data Cleaning:

a. First step to clean the dataset we chose was to drop the variables having unique values.
b. Then, there were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to Null values.
c. We dropped the columns having NULL values greater than 30%.
d. Next, we 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. The outliers were identified and removed. Also, in one column was having identical label in different cases (first letter small and capital respectively). We fixed this issue by converting the label with first letter in small case to upper case.
e. All sales team generated variables were removed to avoid any ambiguity in final solution.

Step3: Data analysis:

- The conversion rate is more from leads who have spent more time on website.
- The converted leads have an average of more than 3 Total Visits.
- In case of Page Views Per Visit it is same for leads who have converted and who haven't.

Step4: Dummy Variables Creation:

a. We created dummy variables for the categorical variables.

b. Removed all the repeated and redundant variables.

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: Model Building:

a. Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.

b. 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.

c. Finally, we arrived at the 16 most significant variables. The VIF's for these variables were also found to be good.

d. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.

e. After, this model seems to be stable with significant p-values, and all the columns have VIF < 4. Hence, we shall go ahead with this model for further analysis.

f. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.

g. Next, based on the Precision and Recall trade-off, we got a cut off value of approximately 0.5.

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 81%; Sensitivity= 70%; Specificity= 88%.

Step7: Plotting ROC Curve: -

a. It Shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

b. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

c. The close the curve comes to the 45-degree of the ROC space, the less accurate the test.

Step8: Finding Optimal Cutoff Point: -

a. Optimal cutoff probability is that probability where we get balanced sensitivity and specificity.

b. It seems the 0.36 is the closest point (Probability) at which Accuracy, Sensitivity, and Specificity meet. Hence 0.36 is the optimum point to take it as a cutoff probability.

c. The Final prediction of conversions have a target of 80% conversion as per the X Education's CEO requirement.

d. This model has 80.7% of lead conversion rate predictability.

Conclusion:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity, and Specificity values of both trained set and test set are approximately closer
  - Test set - 81%, 81.6%, 80.4%
  - Train set - 81%, 81%, 82%
- The lead score calculated in the trained set of data shows the conversion rate on the final predicted model to be around 80%, which matches with the test set.
- Hence, this seems to be a good model.