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:

Step 1: Reading and Understanding Data:

Read and analyze the data.

Step 2: 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 40%.
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.

Step 3: Data Transformation:

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

Step 4: Dummy Variables Creation:

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

Step 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.

Step 6: Feature Rescaling:

a. We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.
b. Dropped the highly correlated dummy variables.

Step 7: 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 11 most significant variables. The VIF's for these variables were also found to be good.
d. Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
e. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
f. We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 88% which further solidified the of the model.

Step 8: Finding the Optimal Cutoff Point

a. Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.38
b. Then, checked if 80% cases are correctly predicted based on the converted column.

Step 9: Computing the Precision and Recall metrics

a. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
b. Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.41.
c. We could also observe the new values of the 'accuracy=80.1%, 'sensitivity=79.1%', 'specificity=80.8%'

Step11: Making Predictions on Test Set

a. 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 81.6%, Specificity= 84.4%, Precision=73.4% Recall= 76.7%

Step 12: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 81% on the final predicted model 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.
- Our model is having stability an accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in future.

- Features which contribute more towards the probability of a lead getting converted are:
    i. Total Visits
    ii. Total Time Spent on Website
    iii. Page Views Per Visit

- Variables have very lower chance to get converted are:
    i. Lead Origin API
    ii. Lead Origin Landing Page Submission
    iii. Lead Origin Lead Import
    iv. Last Activity Email Bounced