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 70%.

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

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:

Feature Rescaling: a. We used the standard scaler to scale the original numerical variable b. Dropped the highly correlated dummy variables.

Step7:

Model Building: a. Using the Recursive Feature Elimination(RFE), 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 14 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. We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 96% which further solidified the of the model. f. Then, checked if 80% cases are correctly predicted based on the converted column. g. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set. h. Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.27. i. 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 83%%; Sensitivity= 94.5%; Specificity= 76.4%.

Step 8:

Conclusion: Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert).

2)Our final Logistic Regression Model is built with 14 features.

3)The final model has Sensitivity of 0.928, this means the model is able to predict 92% customers out of all the converted customers, (Positive conversion) correctly.

4)The final model has Precision of 0.68, this means 68% of predicted hot leads are True Hot Leads.