Lead Scoring Case Study Summary

X Education gets a lot of leads, its lead conversion rate is very poor at around 30%. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have higher conversion chance. CEO’s target for lead conversion rate is around 80%.

1. Reading & Understanding the Data

First, we read the Dataset and Inspect the Data.

2. Data Cleaning

a) First step to clean the dataset we dropped the columns having NULL values greater than 35%.
b) Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required. The outliers were identified and removed.
c) Then we chose to drop the variables having unique values.

3. Data Preparation

a) Changed the binary variables into ‘0’ and ‘1’.
b) We created dummy variables for the categorical variables.
c) Then, Removed all the repeated and redundant variables.

4. Train Test Split

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

5. Feature Scaling

a) We used the Min Max Scaling to scale the original numerical variables.
b) The, we plot the a heatmap to check the correlations among the variables.
c) Dropped the highly correlated dummy variables.

6. Model Building

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

b) Then, we recursively tried looking at the P-values to select the most significant values that should be present and dropped the insignificant values.
c) Then, we arrived at the 11 most significant variables.
d) Then we tried looking at the VIF values to select the most significant values that should be present and dropped the insignificant values.
e) Finally, we arrived at the 10 most significant variables.
f) For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity, and specificity.
g) We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 84% which further solidified the of the model.
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 77.36%; Sensitivity= 76.97%; Specificity= 77.73%.

7. Model Evaluation

Evaluation involved creating a confusion matrix and selecting a cutoff point of 0.39 based on accuracy, sensitivity, and specificity plots, yielding metrics around 80%. Sensitivity-specificity view was chosen over precision-recall for the final predictions to align with the CEO's goal.

8. Conclusion

a) While we have checked both Sensitivity-Specificity, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
b) Accuracy, Sensitivity and Specificity values of test set are around 77.36%, 76.97% and 77.73% which are approximately closer to the respective values calculated using trained set.
c) Hence overall this model seems to be good.
d) Features which contribute more towards the probability of a lead getting converted are:

   i. Lead Origin_Lead Add Form
   ii. What is your current occupation_Working Professional
   iii. Total Time Spent on Website