Lead Scoring Case Study Summary

Problem Statement:

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The data provided has a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate

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 inspect 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. These values were changed to NULL.
c. We dropped the columns having NULL values greater than 35%.
d. Next, we removed redundant variables and imputed the missing values 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.
e. All sales team generated variables were removed to avoid any ambiguity in the 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 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 Min Max Scaling to scale the original numerical variables.
b. Then, we plot a heatmap to check the correlations among the variables and the highly correlated variables were dropped.

Step7: Model Building:

a. Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
b. Using the statistics, we recursively looked at the P-values in order to select the most significant values that should be present and dropped the insignificant values. We arrived at the 12 most significant variables with high VIFs.
c. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
d. The ROC curve was plotted for all features and it came out with a decent area coverage of 94%.
e. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on the train set.
f. Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.3.
g. 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 89%; Sensitivity= 89%; Specificity= 87%.

Step 8: Conclusion:

i. Working professionals should be the main target for the higher conversion rate along with the below highlighted attributes.
ii. The top variables contributing towards hot lead are- Tags (Closed by Horizon or Lost to EINS), Total time spent on website and the source of the lead is 'Add form' on landing page. Hence the company should focus more on such leads.
iii. X Education should spend less time on leads where the emails were bounced or customers whose phones are not available, as they are more likely cold leads.