Summary

In order to find out the “Hot leads”, Lead scoring was conducted using a logistic regression model to adhere to business criteria and limits. Although, there are many leads at first, very few of them end up becoming paying clients.

First we started with reading and inspecting the given data set, then moved to data cleaning where the first step to clean the dataset we chose was to drop the variables having unique values. 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. We dropped the columns having NULL values greater than 40%. 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, Google), we fixed this issue by converting the label with first letter in small case to upper case. 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 Min Max Scaling to scale the original numerical variables.

b. The, we plot the a **heatmap** to check the correlations among the variables and dropped the highly correlated dummy variables.

Step7: 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 8 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 excellent with an area coverage of 93% 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 iterative method, we got a cut off value of approximately **0.3.**
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 85.7%; Sensitivity of 86.2%; and Specificity of 85.5%.

Step 8: Conclusion:

• The lead score calculated in the test set of data shows the conversion rate of 85% on the final predicted model which clearly meets the expectation of CEO (80%)

• Good value of sensitivity of our model will help to select the most promising leads.

• Top Features which contribute more towards the probability of a lead getting converted are:

1) Total visits
2) Tags_Closed by Horizzon
3) Total Time Spent on Website
4) Tags_ Lost to EINS
5) Lead Origin_Lead Add Form