Lead Scoring Case Study Summary

Problem Statement:

The Education Company sells online courses to industry professionals. The Company needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company requests a model in which 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 has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step 1: Reading and Understanding Data:

Read and inspected the data with common statistical figures.

Step 2: Data Cleaning:

a. The first step is to clean the dataset we chose was to drop the variables having high portion values.
b. We dropped the columns having NULL values greater than 50% with further investigation for 'Lead Quality', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score'.
c. Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with 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 (google and Google).
d. Imbalanced fields are also checked and removed accordingly.

Step 3: Data Visualization

In this step, the target variable distribution as univariate analysis is being conducted along with some bivariate analysis using barplot and heatmap against other potential predictors.

Step 4: Data Transformation:

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

Step 5: Dummy Variables Creation:

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

Step 6: Test Train Split:

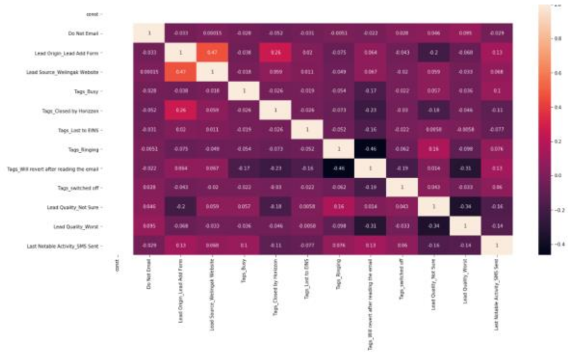
In this step, we divide the data set into test and train partitions with a proportion of 70%/30% records compared with original data input.

Step 7: Feature Rescaling:

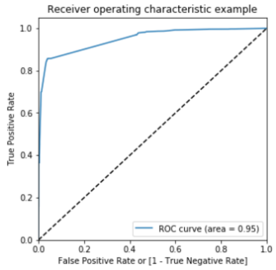
- We used Standard Scaling to scale the original numerical variables and then recheck the conversion rate to ensure completeness.

Step 8: Model Building:

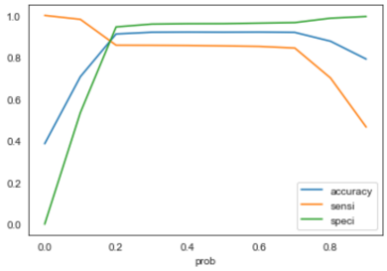
1. Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
2. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
3. Eventually, we arrived at the 12 (0 to 11 in VIF checking table) most significant variables after 4 iterations. The VIF's for these variables were also found to be reasonable with a maximum value less than 3.



4. For the final model, we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
5. We then plot the ROC curve for the features and the curve came out to be acceptable with an area coverage of 95% which further solidified the of the model.



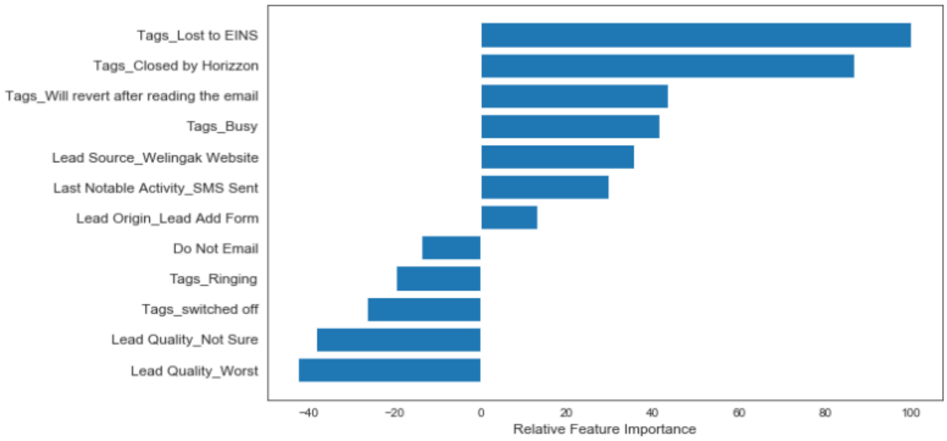
6. Then, checked if 92% cases are correctly predicted based on the converted column.
7. In the next step, we repeat the check on precision and recall with accuracy, sensitivity, and specificity for our final model on train set.
8. Next, with on the Precision and Recall trade-off, we got a cut off value of approximately 0.2.



9. Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and Accuracy to be 84.1%, 94.6%, and 90.1%, respectively.
10. Lastly, the Feature Importance is withdrawn to check on which variables to be meaningful for further actions.

Step 9: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of above 90% on the final predicted model which clearly meets the expectation of the CEO at 80%.
- Features which contribute more towards the probability of a lead getting converted both positively and negatively are illustrated as below.


Tags_Lost to EINS
Tags_Closed by Horizzon
Tags_Will revert after reading the email
Tags_Busy
Lead Source_Welingak Website
Last Notable Activity_SMS Sent
Lead Origin_Lead Add Form
Do Not Email
Tags_Ringing
Tags_switched off
Lead Quality_Not Sure
Lead Quality_Worst
Relative Feature Importance