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 inspect the data.

Step2: Data Cleaning:

- The initial step taken to clean the dataset was to eliminate any variables that had singular or unique values.
- Subsequently, some columns contained the value 'Select', indicating that the leads did not select any of the options provided, and we replaced those values with null values.
- Columns that had null values exceeding 35% were eliminated.
- The subsequent step involved removing imbalanced and redundant variables. This process also involved filling in missing values, as needed, with median values for numerical variables and creating new classification variables for categorical variables. Any outliers were identified and removed. Additionally, we addressed an issue in a column that contained labels in both uppercase and

lowercase letters by converting the first letter of the lowercase label to uppercase.
- To prevent any confusion in the final solution, we eliminated all variables that were generated by the sales team.

Step3: Data Transformation:

Changed the binary variables into ‘0’ and ‘1’

Step4: Dummy Variables Creation:

- We created dummy variables for the categorical variables.
- 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:

- To scale the original numerical variables, we utilized the Min Max Scaling technique.
- Subsequently, we generated a heatmap to examine the correlations between the variables.
- Dropped the highly correlated dummy variables.

Step7: Model Building:

- We employed the Recursive Feature Elimination method to choose the top 6 important features.
- We utilized the generated statistics to recursively examine the P-values and determine which values were significant and which were not. Based on this analysis, we chose to retain the significant values and drop the insignificant ones.

- Ultimately, we identified the 5 most significant variables. Upon analysis, we determined that the Variance Inflation Factors (VIFs) for these variables were satisfactory.
- In order to develop our final model, we assessed the optimal probability cutoff by testing various points and evaluating their accuracy, sensitivity, and specificity.
- Subsequently, we generated an ROC curve to visualize the performance of the features. The curve indicated a strong model with an area under the curve of 86%, further bolstering our confidence in its effectiveness.
- Next, we verified whether the model accurately predicted 80% of the cases based on the converted column.
- We evaluated the precision and recall of our final model on the training set, taking into consideration the accuracy, sensitivity, and specificity.
- Using the Precision and Recall trade-off analysis, we determined an approximate cut-off value of 0.3.
- After applying the learned techniques to the test model, we calculated the conversion probability using the Sensitivity and Specificity metrics. The resulting accuracy value was 77%, with a Sensitivity of 77.44% and a Specificity of 78%.

Step 8: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 83% on
- the final predicted model which clearly meets the expectation of CEO has given a
- ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.
- Features which contribute more towards the probability of a lead getting converted are:
    - Lead Origin_Lead Add Form
    - What is your current occupation_Working Professional
    - Total Time Spent on Website