Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal:

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 understand the dataset.

Step2: Data Cleaning:

- First step to clean the dataset we checked the Null values in provided dataset
- While reading the dataset we found that there are many variables have a level called 'Select' which means the leads did not choose any given option so we changed it to Null .
- We check the delicacy in provided dataset
- We dropped the columns having NULL values greater than 40%.
- Then we inspect the columns where Null values are exists.
- Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values and dropping imbalanced and redundant. Also, in one column was having identical label in different cases (first letter small and capital respectively). We fixed this issue by converting the label with first letter in small case to upper case.
- After imputing the columns, datatype changed and fix that issue too.
- Check the unique values and remove Prospect ID & Lead Number.

Step3: EDA on prepared dataset:

- Perform Multi Variate Analysis and Univariate Analysis on Categorical & Numerical variables.
- Drop variables which contains highly skewed columns and two labels but one is very High and another is very less number of labels.

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 80- 20% values.

Step6: Feature Rescaling:

- We used the Min Max Scaling to scale the original numerical variables.
- The, we plot the a heatmap to check the correlations among the variables.

Step7: Model Building:

- Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.
- 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.
- Finally, we arrived at the 13 most significant variables. The VIF's for these variables were also found to be good.
- For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 87% which further solidified the of the model.
- Then, checked if 79% cases are correctly predicted based on the converted column.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.
- Next, based on the Precision and Recall trade-off, we got a cut off value of approximately 0.4.
- 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 79.06%%; Sensitivity= 74.08%; Specificity= 82.16%.

Step 8: Conclusion:

- The lead score calculated the conversion rate of 74% on test set and 76% on train set which closely 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:
    - Total Visits
    - Total Time Spent on Website
    - Lead Origin_Lead Added Form