Lead Score Case Study

Problem Statement

An X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score 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%.

Process Summary

1) **Reading and Understanding of Data**
    - There are 9240 rows and 37 columns
2) **Data Cleaning**
    - There are a few columns with “Select” which means that the leads did not choose any given option. So, we changed those values to Null values.
    - We found there are NULL values for many columns. So, we kept a threshold of above 40% and dropped all those columns which have NULL values above 40%.
    - We removed redundant variables and columns with a unique value of "1" as they are not useful.
    - We dropped some more columns where only one value is predominant in it.
    - We corrected the values in some variables and grouped some of the values of the columns as “Others” (the ones with less significance).
3) **EDA**
    - Categorical Variables – Univariate and Bivariate analyses were done and found that many of the variables do not add value to the analysis.
    - Numerical Variables - Univariate and Bivariate analyses were done for the same and they seem to be good for analysis.
4) **Dummy Variables**
    - Created dummy variables for all categorical variables.
5) **Dataset split into Train and Test**
    - The dataset was split into 70% train and 30% test.
    - Missing value treatment was done for X_train.
6) **Feature Scaling**
    - Using MinMaxScaler, all the numeric variables were re-scaled.
7) **Model Building**
    - Used a Hybrid approach of both automatic and manual to find the features of the model. Using Recursive Feature Elimination (RFE), we selected the 15 top important features.
    - Using a manual approach, with the statistics generated, selected the most significant values and dropped the insignificant ones looking at the p-values and VIF values where the variables with VIF < 5 and p-value < 0.05 were retained.
    - Checked the optimal probabilities and the cutoff points on the train set.
8) **Model Evaluation and Prediction on Test dataset**
    - Checked the Accuracy, Specificity, and Sensitivity using a Confusion matrix for the default cutoff point for train dataset.

- Plotted the ROC curve and found that the area under the curve was 0.96 which is pretty good.
- Plotted the accuracy sensitivity and specificity for various probabilities and got an optimal cutoff point (0.3).
- For the train dataset, with a 0.3 cutoff point, we got Accuracy, Sensitivity, and Specificity around and above 91%.
- Used the same results to predict the test dataset and got Accuracy (91%), Sensitivity (92.5%), and Specificity (91%).

9) Precision – Recall

- Using this method, found a new cutoff point of 0.41 which gave
    - Accuracy is 91.9%, sensitivity is 89.2% and specificity is 93.5% on the train dataset.
    - Accuracy is 91.6%, sensitivity is 90.4% and specificity is 92.4% on the test dataset.
    - And, precision is 88%, and recall is 90% on the test dataset.

10) Conclusion and Findings

- The lead score calculated on the test dataset showed 90.41% which clearly satisfies the business need for the target lead conversion rate to be around 80%.
- Features that contribute more towards the probability of a lead getting converted are:
    - Tags_Closed by Horizzon
    - Tags_Lost to EINS
    - Lead Source_Welingak Website
    - Tags_Will revert after reading the email.
- There are some negatively affected variables also which we need to consider and avoid considering them.
    - Tags_Ringing
    - Tags_switched off
    - Tags_invalid number