Lead Score Assignment

Summary Report

As we have seen from the problem statement the target variable (Converted) is a categorical variable, so we induced that this a classification type problem that we are dealing with and hence we could proceed with getting the resolution by using Logistic Regression model.

We have started off with understanding the data and cleaning it. Then we divided 70% of the data in training set and remaining 30% into the test set. The training set will be used to train the model to predict the conversion of a lead. Post this we have performed scaling on the train set and continued with the building a model. Once the apt model was ready whose coefficients p value was less than 0.05 and Variance Inflation Factor less than 3, we proceeded with considering **0.5 as a threshold value** for the conversion probability of a lead obtained by the model and checking the **leads that were predicted to be converted** with respect to the **leads that actually converted.** That is, any lead with a probability higher than 0.5 was considered to be converted and vice versa. Then we plotted the confusion matrix using the actual and predicted converted variables to check the true positives, true negatives, false positives, and false negatives. We also found the accuracy of the model to be 81.5%. But by using the values obtained from the confusion matrix we found the sensitivity to be 69.70% and the specificity to be 89.9%. Thus, telling us that the cut off chosen should be changed in order to obtain better sensitivity. To get a better sense of the model we plot the ROC curve, which is x-> Sensitivity(True Positive Rate) vs y-> 1-Specificity(False Positive Rate). ROC curve illustrates the diagnostic ability of the binomial model. A good ROC curve is the one that touches the upper left corner of the graph, so higher the area under the curve the better the model constructed is. In our case, we observe that the ROC curve area = 0.88.

We then proceed further with finding the optimal cut off point. For this we will create columns with different probability cutoffs, after which we will calculate accuracy, sensitivity and specificity for various probability cutoffs that was calculated. Now we will plot a graph with accuracy, sensitivity, and specificity (on y axis) for various probabilities (on x axis). The optimal cut-off point exists where the values of accuracy, sensitivity, and specificity are decent and almost equal. At the cut-off of 0.3, the metric values are 0.795764, 0.832928 and 0.772864 respectively. This is the optimal value of threshold that we can have.

After finding the optimal cut off, we have re-calculated the final predicted value for the lead's conversion. Accuracy of the model now is 79.5%, whereas sensitivity has improved to be 83.2%.

Finally, we use this model constructed and 0.3 cut off value on the test set that we had prepared earlier, we found the accuracy to be 80.6%, **sensitivity to be 83.4%** and specificity to be 78.7%. The training set sensitivity and the test set sensitivity are close enough to each other. We calculated the lead score by using the probability of the conversion of the leads. Finally, we have arranged the leads in the descending order of their lead scores, i.e., hot leads are displayed on the top of the list. Hence, we can proceed with presenting the model constructed to the CEO of X Education.