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. The company requires you 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. The CEO, in particular has given a ballpark of the target lead conversion rate to be around 80% solution.

Summary:

Step-1: Reading and Understanding Data

Read and analyze the data.

Step-2: Data Cleaning:

We dropped the columns having NULL values greater than 70%. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step-3: Data Analysis:

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

Step-4: Creating Dummy Variables

We changed the binary variables into ‘0’ and ‘1’ and created dummy data for the categorical variables.

Step-5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step-6: Feature Rescaling

We used the Min Max Scaling to scale the original numerical variables.

Step-7: Feature selection using RFE:

Using the Recursive Feature Elimination we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values and vif values in order to select the most significant values that should be present and dropped the insignificant values. Finally, we arrived at the 16 most significant variables. We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0. Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model. We also calculated the ‘Sensitivity’ and the ‘Specificity’ matrices to understand how reliable the model is.

Step-8: Plotting the ROC Curve

We then tried plotting the ROC curve for the features and the curve came out be decent with an area coverage of 94% which further solidified the of the model.

Step-9: Finding the Optimal Cutoff Point

Then we plotted the probability graph for the ‘Accuracy’, ‘Sensitivity’, and ‘Specificity’ for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.32. Based on the new value we could observe that close to 80% values were rightly predicted by the model. We could also observe the new values of the ‘accuracy=87.61%, ‘sensitivity=88%’, ’specificity=87.35%’. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step-10: Computing the Precision and Recall metrics

We also found out the Precision and Recall metrics values came out to be 87.48% and 82.19% respectively on the train data set. Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

Step-11: Making Predictions on Test Set

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 87.66%; Sensitivity=88.12%; Specificity= 87.35%.

Step-12: Conclusion:

The lead score calculated in the train set of data shows the conversion rate of 86% which clearly meets the expectation of CEO. Features which contribute more towards the probability of a lead getting converted are:

1. Tags_Closed by Horizzon
2. Total Time Spent on Website
3. Lead Origin_Lead Add Form