Lead Score Case Study For X Education

Problem Statement:

An X Education requires assistance in identifying the most promising leads, which are the most likely to convert into paying customers. The company wants us to create a model in which we assign a lead score to each lead so that customers with a higher lead score have a higher conversion chance and customers with a lower lead score have a lower chance of conversion. The major objective is to achieve the target lead conversion rate to be around 80%.

Summary:

The following steps are the inference made by us after building up the logistic regression model step by step to achieve the stated objective. The output for each and every phase of the model ae:

Step1: Read the given problem statement and analyzed the data file thoroughly to proceed with the next step.

Step2: In the cleaning stage step, we removed the variables with a high percentage of NULL values. In this step itself we filled in the missing values with median values in the case of numerical variables and creating new classification variables in the case of categorical variables. Outliers were identified and eliminated.

Step3: In Data Analysis phase of our model, we began by performing an exploratory data analysis on the data set to get a sense of how the data is organized. In this step, approximately three variables were identified as having only one value in all rows. These variables have been dropped.

Step4: Created dummy data for the categorical variables.

Step5: In this Test Train Split phase, divide the data set into test and train sections with a proportion of 70-30% values.

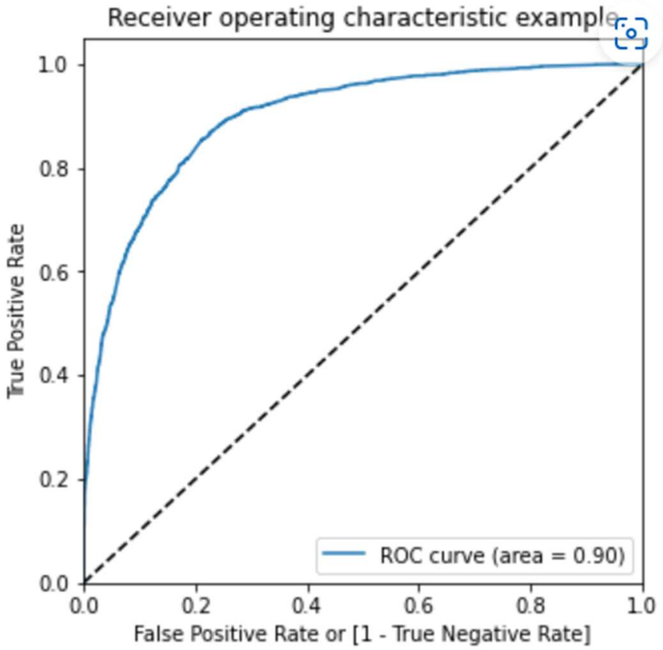
Step6: In Feature Rescaling step, implemented the Min Max Scaling to scale the original numerical variables. Then used the stats model we created our initial model, which gave us a complete statistical view of all the parameters of the model.

Step7: In Feature selection using RFE (Recursive Feature Elimination) step we employed this RFE algorithm and accomplished the following aspects:

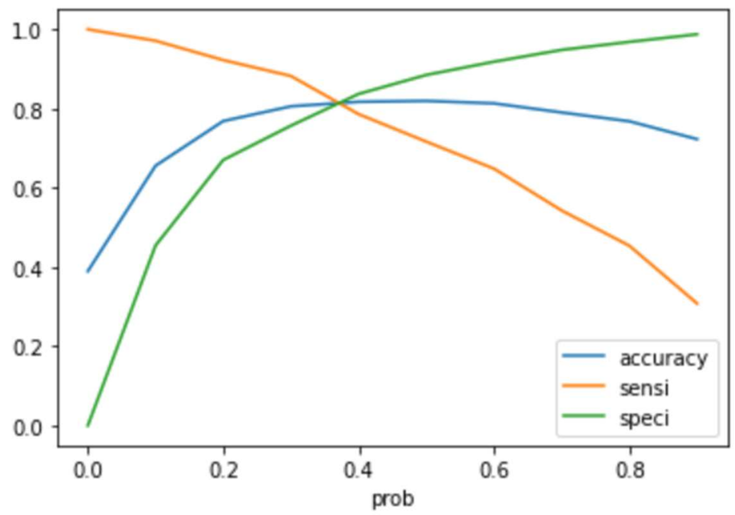
- Selected the 20 top important features.
- Using the generated statistics, we iteratively looked at the P-values to select the most significant values that should be present and dropped the insignificant values.
- Arrived at the 15 most significant variables by eliminating the 5 important features among them.

- The VIF's for these variables were also found to be good.
- Created the data frame having the converted probability values with an initial assumption that a probability value of more than 0.5 means 1 else 0.
- Based on the above assumptions, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
- Calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

Step8: In Plotting the ROC Curve phase, we plotted the ROC curve for the features and the curve came out be pretty good as shown in the screenshot below with an area coverage of 90% which further solidified the of the model.



Step9: In Finding the Optimal Cutoff Point step, plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values as shown in the screenshot below.



The intersecting point of the graphs was considered as the optimal probability cutoff point. From the curve above, 0.37 is the optimum point to take it as a cutoff probability. Based on the new value we could observe that close to 80% values were rightly predicted by the model. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step10: In Computing the Precision and Recall metrics step, found out the Precision and Recall metrics values to be 79% and 70.5% respectively on the train data set. Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

Step11: In Making Predictions on Test Set step, implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy values.

OUTCOME

The following are the inferences we could deduce from the case study:

- Implemented logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot which means it is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Some problems presented by the company which the model build was able to adjust even if the company's requirement changes in the future so that it can handle these as well.