Lead Scoring Case Study

UpGrad-IIITB EPGDS C50
(Data Science Program - November 2022)
by
Abhishek Shubhrant

Problem Statement

X Education sells online courses to industry professionals. 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%.

Data Cleaning & Preparation

Step 1: Reading and Understanding Data:
Read and inspected the data.

Step 2: Data Cleaning:
a. The variables with unique values were removed as the first stage in cleaning the dataset we selected.

b. Following that, a few columns had the value "Select," indicating that the leads had not selected any of the available options. These values were modified to be Null values.

c. We eliminated the columns with NULL values that exceeded 40%.

d. After that, we eliminated the duplicate and unbalanced variables. In addition, when necessary, missing values were imputed using median values for numerical variables and new categorization variables were created for categorical data. The outliers were found and eliminated. Additionally, one column contained the same label in both small and capital letters for the initial letter, respectively. We capitalized the label to address this problem.

Data Cleaning & Preparation

Step 3: Data transformation:
Binary variables were converted to '0' and '1'.

Step 4: Dummy Variable Creation
We created dummy variables for the categorical variables (and eliminated all the repeated variables).

Model Building (1/2)

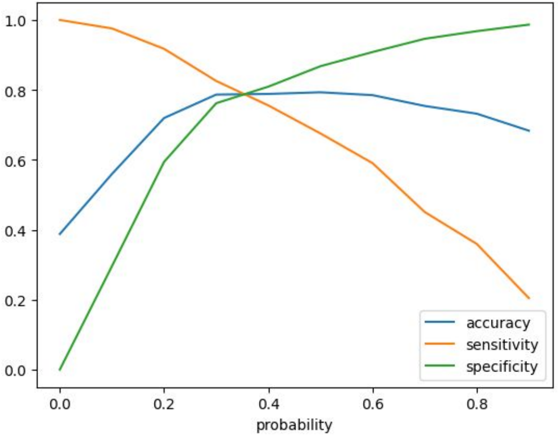
Step 1: Test Train Split
The data set was split into test and train sections with a 70–30% value ratio as the next stage.

Step 2: Rescaling Features
a. To scale the initial numerical variables, we employed the Min Max Scaling.

Step 3: Model Building
a. We chose the 15 most crucial characteristics using the Recursive Feature Elimination.
b. With the help of the statistics produced, we recursively tried examining the p-values in an effort to pick the most significant values that should be present and eliminate the unimportant values.
c. Finally, we identified the 13 characteristics that were the most important. These variables' VIFs were also found to be good.
d. To test the accuracy, sensitivity, and specificity of our final model, we found points and assessed the best probability cutoff.
e. The ROC curve for the characteristics was then plotted, and it showed a reasonably good shape with an area coverage of 86%, further supporting the model.

Model Building (2/2)

f. The conversion probability was then determined based on the Sensitivity and Specificity metrics, and we found that the accuracy value was 79%; Sensitivity was 68%; and Specificity was 87%. Next, we applied the learnings to the test model. (Cut-off assumed was 0.5)



1. Cut-off of 0.38 was selected basis the Accuracy, sensitivity & specificity trade-off curve.
2. The accuracy value was then found 79%; Sensitivity was 77%; and Specificity was 80%. Next, we applied the learnings to the test model.

Conclusion

1. Prediction on the test data: The accuracy value was then found 81%, Sensitivity was 71% and Specificity was 82%.
2. Top three variables in your model which contribute most towards the probability of a lead getting converted?
   TotalVisits
   Total Time Spent on Website
   Lead Origin_Lead Add Form