Lead Scoring Case Study

Summary Report

Prepared By

Soumya Prakash Parida & Ashok Mohapatra

Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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 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%.

Approach

We have used ***Logistic Regression*** to determine and assign the probability score against each customer based on the available data and choose the candidates which have better probability of joining the course.

The lead dataset is split into Train and Test datasets and once the model is built on the Train dataset, it is used on the Entire dataset to verify the accuracy, sensitivity and specificity of the model, which in turn indicates how good the prediction model is.

Model Performance

The final model had the below performance parameters:

Accuracy: 92%

Specificity: 87%

Sensitivity: 95%

Also, the ROC curve resulted with an AUC of 0.95

Learnings:

- There were too many variables in the existing dataset, and a lot of the categorical values had more than 10 categories. So, doing EDA on them and getting meaningful dummies were a challenge.
- Also, determining and eliminating null values were important. Several columns had single value with higher frequency which made them insignificant. Since this was tedious to do manually, we wrote methods to do these automatically.

[See Next Page]

- Some categories like Tags/Lead Source etc had values with very low frequency. Probably a better way to handle these would be to mark them as “Other” instead of having their own dummy.
- Also, some categorical values were too lengthy, these could be abbreviated for easy reference and plotting. We did not try this but used rotation for plotting.

- Due to correlation, 4-5 variables that were found significance during EDA did not figure in the list after RFE. This created additional inspection steps for us to validate that the RFE process was OK, and the ignored columns did not make a significant change to the performance.
- We validated this by running multiple iterations of the model with and without those variables manually and checking the generate coefficients. The variables were found to have insignificant coefficients and high p value instead.

- Finally determining the Optimal Cutoff point was a challenge. Since we needed to make sure we get high conversion rate without increasing the burden of calling all leads. We went with 0.15 probability as cutoff.
- However as noted in the presentation, a 0.1 probability can used too, if resources are available for contacting the leads or if the sales has to be a bit more aggressive. Even with 0.1 probability almost 50% of the cold leads were being filtered out saving resources who can focus on the hot leads.