Lead Scoring Case Study Summary

Submitted by -

Pratima Rathore &   
Apoorv Dubey



**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%

Solution Summary:

* **Step1 - Reading and Understanding Data**: Read and analyze the data.
* **Step2 - Data Cleaning:**
* We dropped the variables that had high percentage of NULL values in them.
* This step also included imputing the missing values as and where required with median /mode values in case of numerical variables and creation of new classification variables in case of categorical variables.
* The outliers were identified and treated
* Subsetting the data set into Converted and Non-converted for better analysis
* **Step 3- Data Visualization** :

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented.

* Univariant Analysis
* Bivariant Analysis
* Multivariant Analysis (Heatmap, pairplot)
* **Step 4 - Data Preparation :**
* Converting some binary variables (Yes/No) to 0/1
* Creating dummy variables for the categorical variables
* **Step 5 - Model Building:**
* Splitting Data into Training and Test Sets. step was to divide the data set into test and train sections with a proportion of 70-30% values.
* Rescaling the numerical variable. We used the Standard Scaling to scale the original numerical variables.
* Running Our First Training Model. using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.
* **Step 6- Feature selection using RFE:**
* Using the Recursive Feature Elimination we went ahead and selected the 15 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
* Finally, we arrived at the 12 most significant variables. The VIF’s for these variables were also found to be less than 5.
* 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.
* Finding the optimal Cutoff point : Then according to our business statement , to build model with 80% conversion , we came to conclusion that threshold should be – 0.36.
* Lead Score Calculation - Lead Score = 100 \* Conversion Probability
* Then we made model on test data. And got

Accuracy 0.806857

Sensitivity 0.788247

Specificity 0.818071

Precision 0.723048

Recall 0.788247