

Lead Scoring Case Study

This Case Study is done for an Education Company X to find optimal ways to get more industry Professional in joining their courses. The Company wants to select the most promising leads i.e. the leads that are most likely to convert into paying customers. Also the Company has given the target lead conversion needed to be around 80%.

Below are the Steps that we have proceeded for the assignment:

1. Data Cleaning:

- First step we have removed all the columns having missing values more than 33% since these columns will not be better features while building the model
- Then we have dropped 'City', 'Country' columns since these will not add much inference with the predictions
- There are few columns in which there is a level called 'Select' which means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are also considered to be missing values and hence dropped those columns
- There were few columns in which one level of class is almost present in all the records. Hence we dropped all those columns related with this since it will be of no use in further analysis
- Then we have retained 69% of the records from the Original records given after these steps

2. Data Transformation:

- Changed the multicategory labels into dummy variables and binary variables into '0' and '1'.

3. Data Preparation:

- The split was done at 70% and 30% for train and test data respectively
- Using MinMax Scaler , we have scaled all the numeric features from 0 to 1

4. Model Building:

- Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

5. Model Evaluation:

- By finding the Optimal Cutoff value using ROC curve, we found out that value as 0.42. Then we found out the accuracy, sensitivity, specificity seems to be around 80%

6. Prediction:

- By using the test set, we found out the predicted value by using cut off as 0.42. Then we found out the accuracy, sensitivity, specificity seems to be around 79%

7. Precision-Recall view:

- This method was also used to recheck and a cut off of 0.44 was found with Precision around 79% and recall around 77% on the test data frame.

Conclusion:

Features which contribute more towards the probability of a lead getting converted are

- TotalVisits
- Total Time Spent on Website

- When the Lead Origin was Lead Add Form
- When the Lead Source was
 1. Olark chat
 2. Welinkak Website
- When the Last Activity was
 1. Phone Conversation
 2. SMS sent

After obtaining the list of leads, we have to inform them about new courses, offers, services, job information and extension of higher studies to them

Conducting Surveys to the leads will help us to determine their intention in joining the online courses. This will help us in refining the approach better