Lead Score Case Study

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

X Education want 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%

The goal of the case study is as below:

Logistic Regression is a classification model which is used for making prediction when out variables are categorical.

As per the business requirement it a classification problem so we will be building a Logistic Regression model to assign a lead score between 0 and 1 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, i.e., is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

We followed below approach for the analysis:

1. Data Understanding:

We have leads dataset from the past with around 9000 records . This dataset consists of 37 columns such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

2. Data Cleaning:

a. Missing value treatment

We checked the number of missing values in each column and we observed there were few columns have missing values more than 35% which is too large. So, we dropped those columns.

b. Handling invalid values

Since it is an online course, we are not much concerned about customer's Country and City. So we decided to drop those columns.

We observed that there are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we identified the value counts of the level 'Select' in all the columns that it is present.

We dropped those columns which had count of Select level greater than 50% of the total value counts since it was redundant.

Also we decided to drop below columns since they had no variation in their levels.

- Newspaper
- Update me on Supply Chain Content
- Search
- Get updates on DM Content
- Do Not Call
- Receive More Updates About Our Courses
- Digital Advertisement
- Magazine
- I agree to pay the amount through cheque
- Through Recommendations
- Newspaper Article
- What matters most to you in choosing a course
- X Education Forums

3. Data Preparation:

a. The next step is to deal with the categorical columns present in the dataset. So we created dummy variables for each categorical columns.

b. As the part of data preparation, we split the data into Train and Test dataset in 70:30 ratio.

c. We re-scaled the continuous columns using MinMaxScaler() method which helped us to convert the column values in standard range.

We observed at the correlations of all columns using corr() method . Since the number of variables were pretty high, we decided to look at the table instead of plotting a heatmap.

4. Model creation:

a. We used RFE method (Recursive Feature Elimination) to get top 15 necessary columns.

b. We further optimized the model using manual feature elimination by observing p-values and VIFs

We finalized the model which satisfied below important conditions:

- p-values is less than 0.05
- VIFs should be less than 5 for all columns .

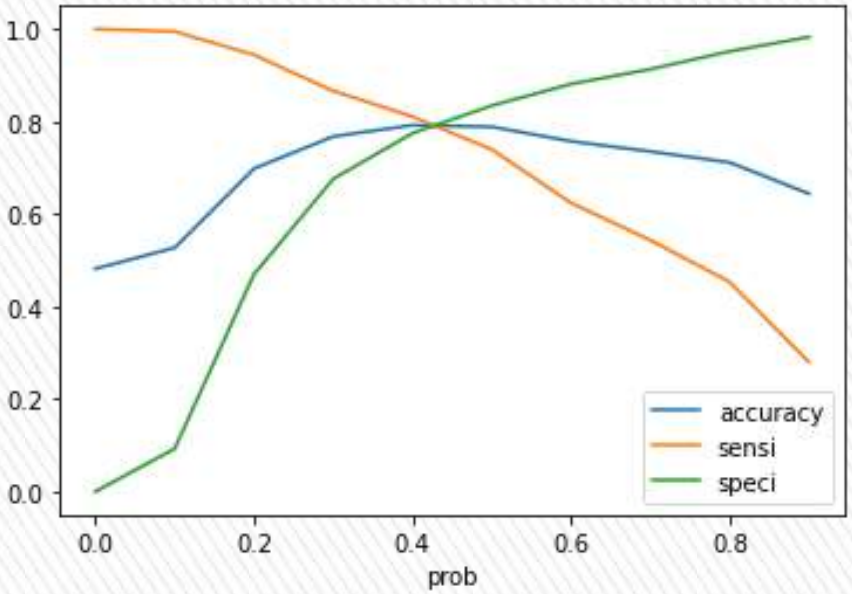
4. Model Evaluation:
We evaluated the final model by determining the optimal cut-off using Accuracy, Sensitivity, Specificity.

We created confusion matrix and derived the below:

- *For default cutoff 0.5 below were the values*

Accuracy: 0.78
Sensitivity: 0.73
Specificity: 0.83

We calculated Accuracy, Sensitivity, Specificity for cut-off values from 0 to 1 and plotted them as below:

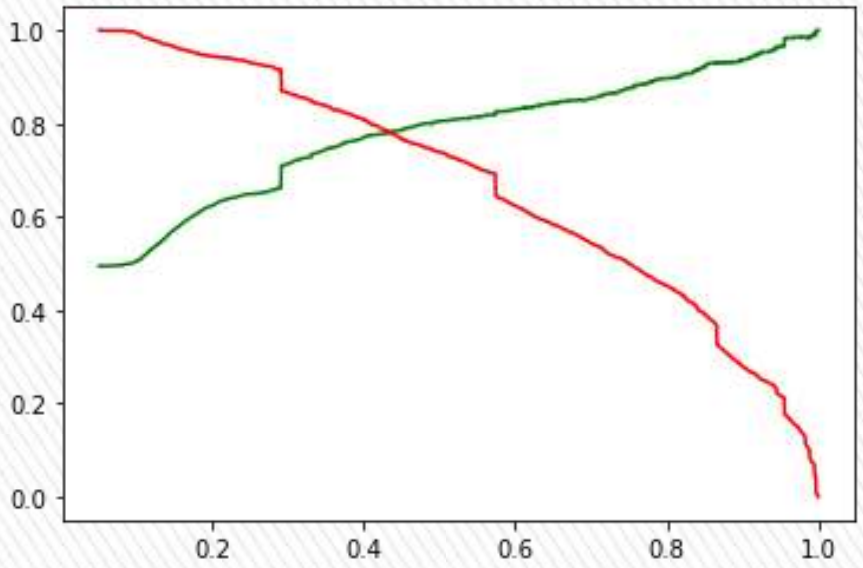


As we saw that around 0.42, we got the optimal values of the three metrics. So we choose 0.42 as optimal cut-off.

- *For optimal cutoff 0.42 below were the values*

Accuracy: 0.79
Sensitivity: 0.79
Specificity: 0.78

5. Making predictions on Test set:
We made predictions on Test set using optimal cut-off of 0.42.
We further calculated Precision – Recall tradeoff and plotted below graph:



As we observed from plot, the cut off is around 0.44. So we are finalizing this cut-off.

We made predictions on the Test Set based on final cutoff value (0.44)

Then we calculated Precision Recall values.

With the help of confusion matrix, we calculated below :

TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

And formule for ,
Precision = TP/(TP+FP) = 0.78
Recall = TP/(TP+FN)= 0.76

We also calculated FPR, TPR and F1 score as below:

FPR - False Positive Rate
FPR = FP/(TN+FP) = 0.19

False positive rate should be as minimum as possible.

TPR - True Positive Rate
TPR = Recall = Sensitivity should be as high as possible.
In our final test set, we have TPR almost of 77%.

F1 score :

As we know, an F1 score reaches its best value at 1 and worst value at 0.
F1 score = 2*((Precision * Recall)/(Precision + Recall)) = 0.77

So we observed that, F1 score is good for our final model.

Recommendations:

- Customers who have visited and spent more time on the website are potential Hot leads.
- The X education company should focus on customers who have reached to them via Welingak Website and Olark Chat.
- The company should reach out to customers who have provided their details using Lead add form option.
- In order to get good job opportunities, students or unemployed customers are always looking for high market demand skills like courses offered by the X education company. So these customers should be considered as potential Hot leads.
- Customers who had phone conversation with sales team and also enquired about the courses through SMS might be considered as Hot leads since these activities shows customers interest in courses.

THANK YOU!