LEAD SCORE CASE STUDY

Submitted by :

Abhisek G. Mahapatra
Abhishek Chakraborty
RajeshKumar Palanivel
Shaun Steplon Aranha

PROBLEM STATEMENT

X Education sells online courses to industry professionals. 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%.

Business Goal:

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%.

APPROACH AND METHODOLOGY

- Get data source for analysis and understand the data
- Clean and prepare the data
- Exploratory Data Analysis
- Model Building
    - Feature Scaling
    - Splitting the data into Test and Train dataset.
    - Building a logistic Regression model and calculate Lead Score
- Model Evaluation
  Evaluating the model by using different metrics
    - Accuracy
    - Specificity and Sensitivity
    - Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.
- Making prediction on the test data set

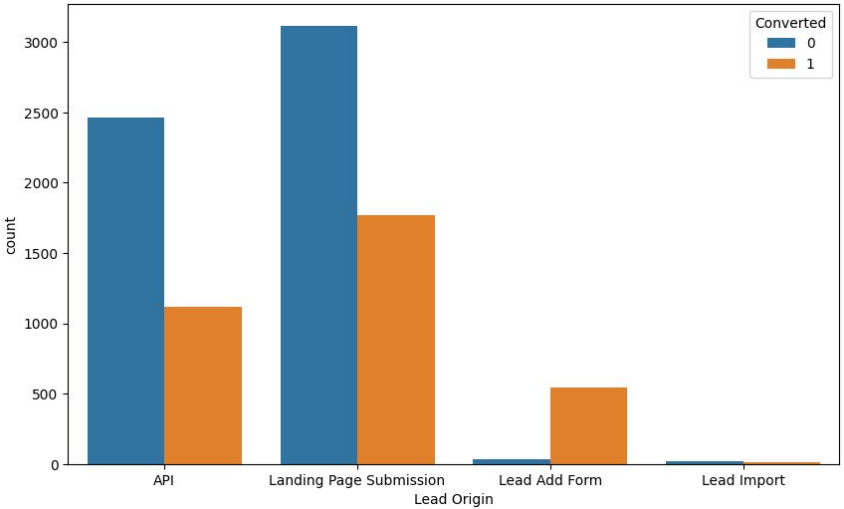
DATA CLEANING AND PREPARATION

Handled null values

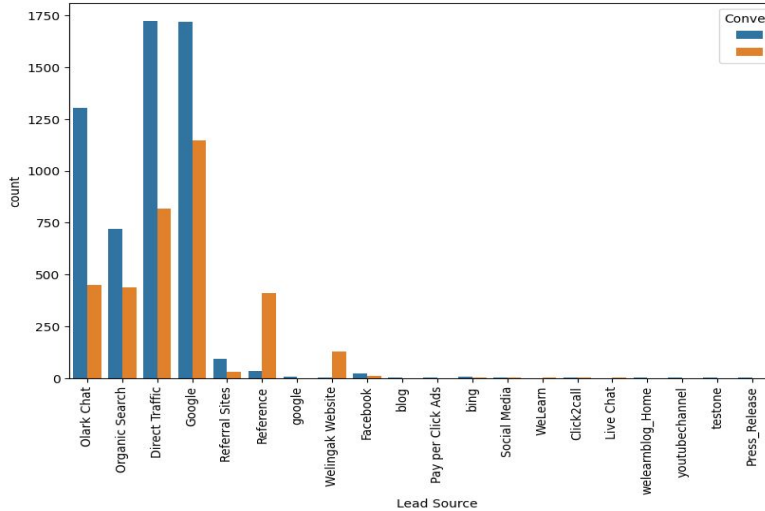
```
df.isnull().sum()
```

```
Lead Origin                                      0
Lead Source                                      0
Do Not Email                                     0
Do Not Call                                      0
Converted                                        0
TotalVisits                                      0
Total Time Spent on Website                      0
Page Views Per Visit                             0
Last Activity                                    0
Country                                          0
What is your current occupation                  0
What matters most to you in choosing a course    0
Search                                           0
Magazine                                         0
Newspaper Article                                0
X Education Forums                               0
Newspaper                                        0
Digital Advertisement                            0
Through Recommendations                          0
Receive More Updates About Our Courses           0
Update me on Supply Chain Content                0
Get updates on DM Content                        0
City                                             0
I agree to pay the amount through cheque         0
A free copy of Mastering The Interview           0
Last Notable Activity                            0
dtype: int64
```

EXPLORATORY DATA ANALYSIS



Lead origin



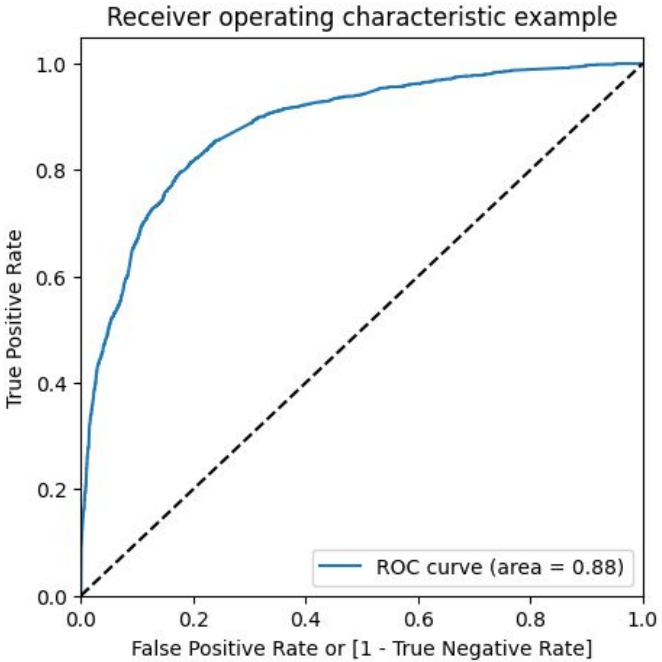
Lead source

MODEL EVALUATION (ON TRAINING DATASET)

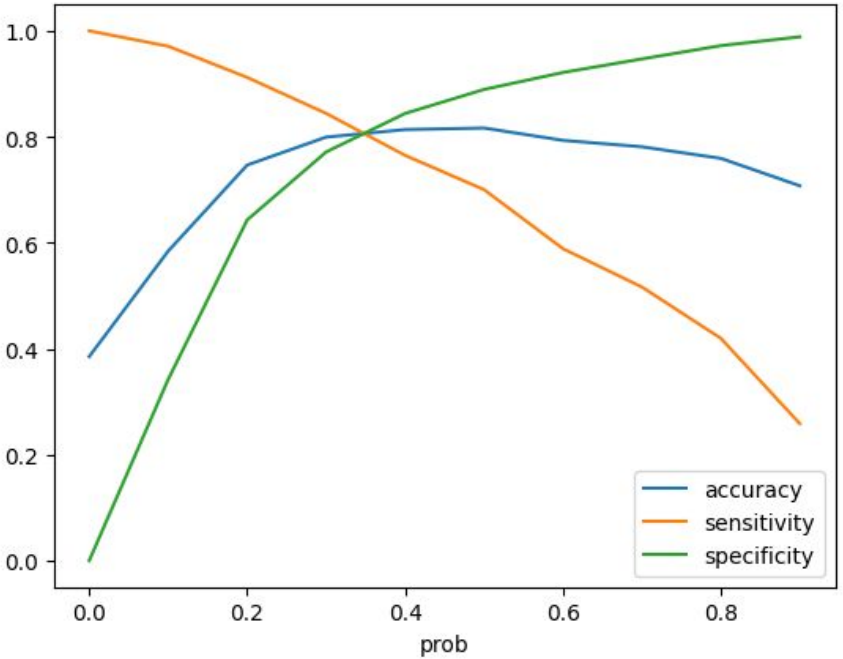
```
[ ]  #Choosing arbitrary cut-off probability point of 0.5 to get predicted labels
     y_train_pred_final['Predicted_label'] = y_train_pred_final.conv_prob.map(lambda x:1 if x>0.5 else 0)
     y_train_pred_final.head()
```

	Converted	conv_prob	Prospect ID	Predicted_label
0	0	0.199037	3009	0
1	0	0.340572	1012	0
2	0	0.357910	9226	0
3	1	0.872211	4750	1
4	1	0.809460	7987	1

MODEL EVALUATION (ON TRAINING DATASET) ROC CURVE

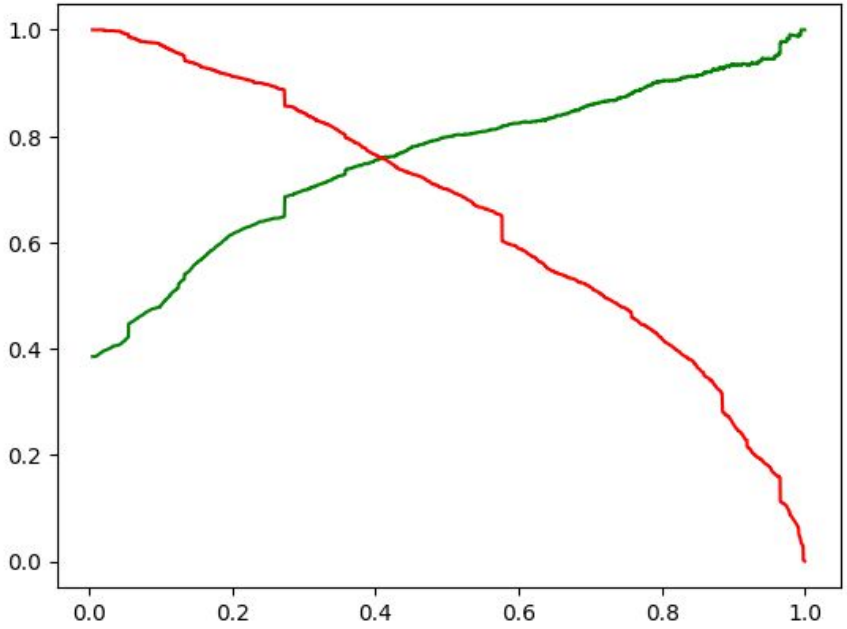


PLOT OF ACCURACY SENSITIVITY SPECIFICITY



PRECISION & RECALL

Trade Off between precision and recall



MODEL EVALUATION (ON TEST DATASET)

Train Data:

Accuracy : 80.9 %

Sensitivity : 80.9 %

Specificity : 80.8 %

Test Data:

Accuracy : 80.3 %

Sensitivity : 79.4 %

Specificity : 80.8 %

CONCLUSIONS

While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

Accuracy, Sensitivity and Specificity values of test set are around 80.3 % , 79.4 % and 80.8 % which are approximately closer to the respective values calculated using trained set.

Also the lead score calculated shows the conversion rate on the final predicted model is around 80 %

The top 3 variables that contribute for lead getting converted in the model are

- Lead Origin_Lead Add Form
- Last Activity_Others
- Last Notable Activity_Unreachable

Hence overall this model seems to be acceptable.