Lead Scoring 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. Company wants to identify the potential leads so that sales team can focus more on these leads rather than calling all the leads, thus increasing the chances of conversion.

Objective

The aim of this case study is to identify the driving factors and converted/non converted leads and derive a logistic model which can predict the 'HOT' leads

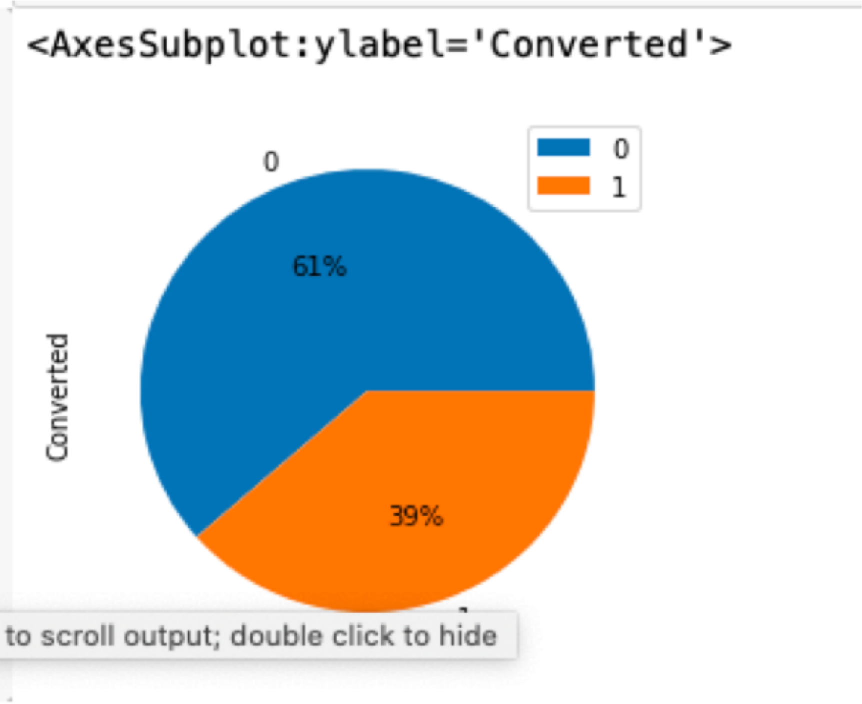
Approach

The steps that we will follow are as below:

- Reading and interpreting data
- Data Cleaning
- Preparing the data for modelling
  - Handling binary categorical columns
  - Creating dummy columns for other categorical columns
  - train-test data
  - rescaling continuous data
- Training the model
- Predicting and evaluating Model on the test data (metrics test)

Approach

Imbalance % only 39% of the leads have been converted



```
<AxesSubplot:ylabel='Converted'>
```

EDA

- We observe value 'Select' in some columns('Specialization','How did you hear about X Education','Lead Profile','City')which is nothing but Null. So we replaced them with NULL
- Columns with >40% missing values have been dropped

- Also columns 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque' have all the records with value 'No', hence can't be used as a driving factor and have been dropped.

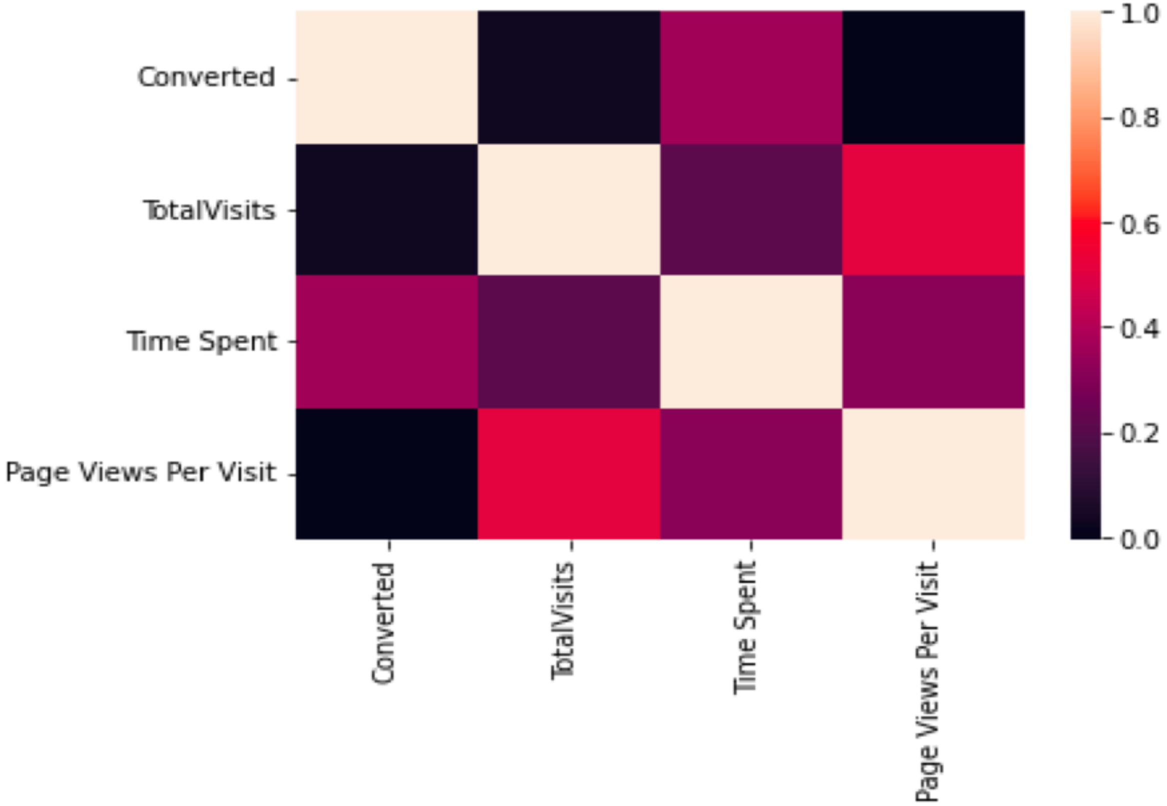
- Missing values in 'Country' have been replaced with India where city is an Indian city

EDA

- Majority of the leads are from India. Rest of the countries we have clubbed together As 'Others' so as to reduce the dummy columns later.

Correlation between Numeric Variables

- There is some correlation between 'Time Spent' and 'Converted'



Final Logistic Model Parameters

const	0.319789
Do Not Email	-0.121590
Time Spent	0.182821
Lead Origin_Lead Add Form	0.520396
Lead Source_Olark Chat	0.187147
Lead Source_Welingak Website	0.178706
Last Activity_Had a Phone Conversation	0.137343
Last Activity_Olark Chat Conversation	-0.078582
Last Activity_Resubscribed to emails	0.809135
Last Activity_SMS Sent	0.198661
Specialization_Hospitality Management	-0.102859
Occupation_Working Professional	0.318921
Reason_Other	-0.154207
Last Notable Activity_Had a Phone Conversation	0.301803
Last Notable Activity_Modified	-0.108790
Last Notable Activity_Unreachable	0.292718

Confusion Matrix

Actual	Predicted	
	Not Converted	Converted
Not Converted	3572	430
Converted	783	1683

Accuracy

```
metrics.accuracy_score(y_train_pred_final.Converted,y_train_pred_final.Predicted)
```

```
0.8124613481756339
```

Sensitivity & Specificity

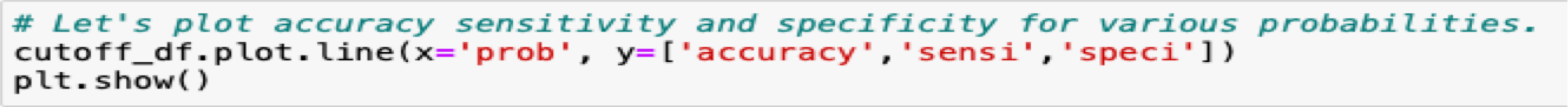
TP / float(TP+FN) = 0.6824817518248175

TN / float(TN+FP) = 0.8925537231384307

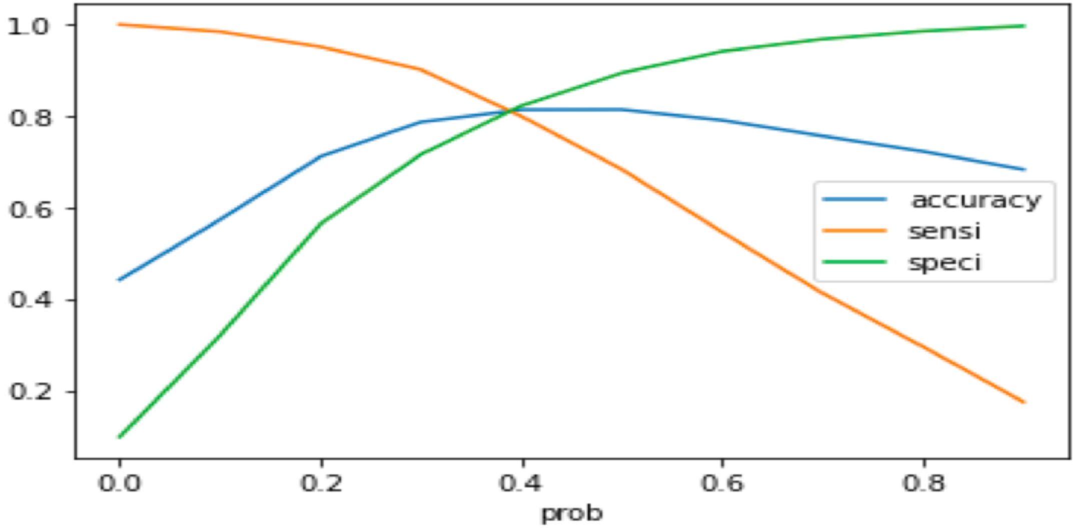
Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

Optimal cutoff probability

- Balanced sensitivity and specificity. From the curve below, 0.4 is the optimum point to take it as a cutoff probability.



```
# Let's plot accuracy sensitivity and specificity for various probabilities.
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```



Precision and Recall

TP/(FP+TP)= 0.7329608938547486

TP/(FN+TP) = 0.7980535279805353

Making predictions on the test set

Actual	Predicted	
	Not Converted	Converted
Not Converted	1383	294
Converted	222	873

Accuracy = 0.8138528138528138

Sensitivity = 0.7972602739726027

Specificity = 0.8246869409660107

Conclusion

Sales team should focus more on the leads who have resubscribed to the email. This is a very strong point as they might be reconsidering taking the admission.

Also since company provides industrial courses which are meant for industrial professionals, they should target working professionals rather than students.

They should ignore ‘Do Not Email’ leads as this shows less interest and also not waste time in entertaining chat enquiries as it may not be the optimum use of the time.