# LEAD SCORE CASE STUDY

Done by:-

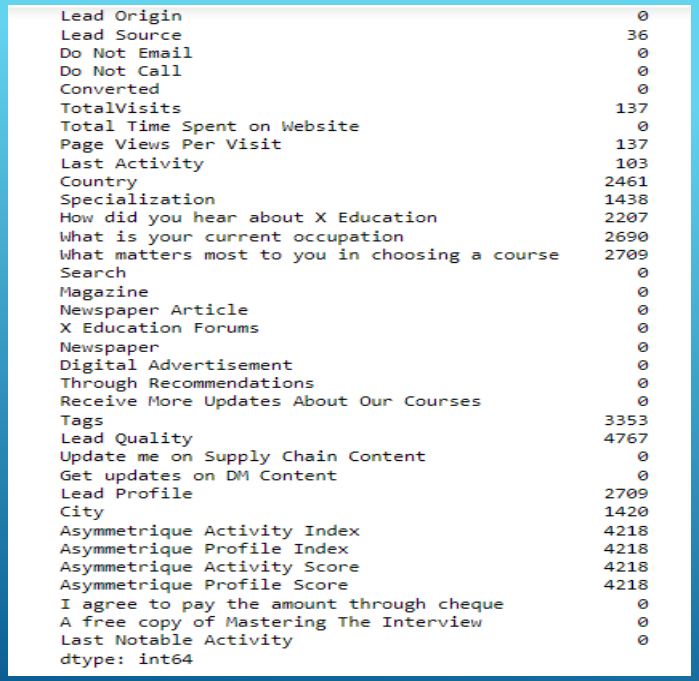
Subhangi Dogra

# PROBLEM STATEMENT:-

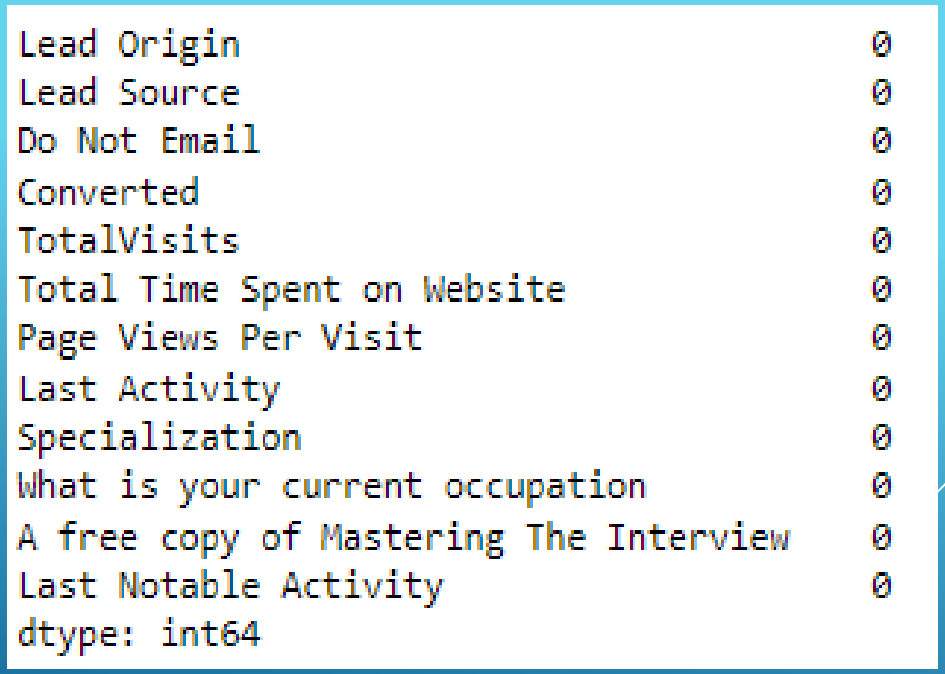
- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

# DATA CLEANING:-

- Dataset has 37 columns and 9240 rows.
- First we checked how many null/Na values are in the dataset.
- Many columns/features have a very large amount if nulls. Hence we drop the columns which have more than 3000 null values.
- The variables City, Lead Profile, How did you hear about X Education and country are not useful in our analysis hence we drop them.
- Many features have majority of No values in them and hence we drop them too.
- Also we remove the null columns of 'What is your current occupation','TotalVisits','Lead Source', 'Specialization'.
- We have converted Yes/No value to 1/0 of binary variables.



```
Lead Origin                                       0
Lead Source                                      36
Do Not Email                                      0
Do Not Call                                       0
Converted                                         0
TotalVisits                                     137
Total Time Spent on Website                       0
Page Views Per Visit                            137
Last Activity                                   103
Country                                        2461
Specialization                                 1438
How did you hear about X Education             2207
What is your current occupation                2690
What matters most to you in choosing a course  2709
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
Tags                                           3353
Lead Quality                                   4767
Update me on Supply Chain Content                 0
Get updates on DM Content                         0
Lead Profile                                   2709
City                                           1420
Asymmetrique Activity Index                    4218
Asymmetrique Profile Index                     4218
Asymmetrique Activity Score                    4218
Asymmetrique Profile Score                     4218
I agree to pay the amount through cheque          0
A free copy of Mastering The Interview            0
Last Notable Activity                             0
dtype: int64
```



```
Lead Origin                                       0
Lead Source                                       0
Do Not Email                                      0
Converted                                         0
TotalVisits                                       0
Total Time Spent on Website                       0
Page Views Per Visit                              0
Last Activity                                     0
Specialization                                    0
What is your current occupation                   0
A free copy of Mastering The Interview            0
Last Notable Activity                             0
dtype: int64
```

# DATA MODELLING:-

Dummy variables were created and extra columns were removed.

| | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | Lead Origin_Landing Page Submission | Lead Origin_Lead Add Form | Lead Origin_Lead Import | Lead Source_Direct Traffic | Lead Source_Facebook | Lead Source_Google | ... | Specialization_IT Projects Management | Sp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.0 | 0 | 0.0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | |
| 1 | 0 | 5.0 | 674 | 2.5 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | |
| 2 | 1 | 2.0 | 1532 | 2.0 | 1 | 0 | 0 | 1 | 0 | 0 | ... | 0 | |
| 3 | 0 | 1.0 | 305 | 1.0 | 1 | 0 | 0 | 1 | 0 | 0 | ... | 0 | |
| 4 | 1 | 2.0 | 1428 | 1.0 | 1 | 0 | 0 | 0 | 0 | 1 | ... | 0 | |

# TRAIN-TEST SPLIT:-

- Splitting the Data into Training Sets and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.

```
: #splitting Data
X_train, X_test,y_train, y_test = train_test_split(X,y,train_size = 0.7,test_size = 0.3, random_state=100)
```

# FEATURE SCALING:-

We have scaled the numeric features of the dataset using MinMaxScaler().

| | TotalVisits | Total Time Spent on Website | Page Views Per Visit |
|---|---|---|---|
| 8003 | 0.015936 | 0.029489 | 0.125 |
| 218 | 0.015936 | 0.082306 | 0.250 |
| 4171 | 0.023904 | 0.034331 | 0.375 |
| 4037 | 0.000000 | 0.000000 | 0.000 |
| 3660 | 0.000000 | 0.000000 | 0.000 |

# FEATURE SELECTION:-

After data cleaning the lead convert rate was found out to be 49.09 percent.

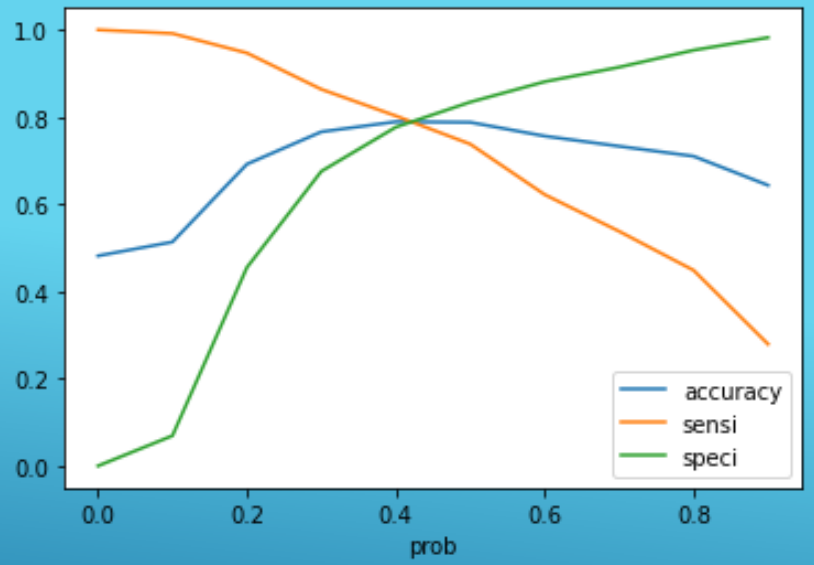
It is neither perfectly balanced nor heavily imbalanced.

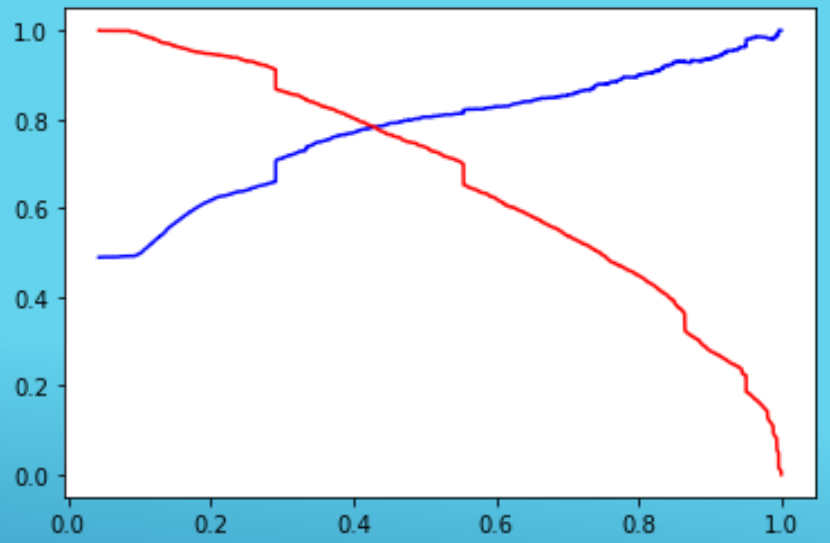
The following variables (below image) were selected for our analysis depending on their correlation with the target variable.

```
['TotalVisits', 'Total Time Spent on Website',
 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
 'Lead Source_Reference', 'Lead Source_Welingak Website',
 'Last Activity_Email Bounced', 'Last Activity_Had a Phone Conversation',
 'Last Activity_SMS Sent', 'What is your current occupation_Housewife',
 'What is your current occupation_Student',
 'What is your current occupation_Unemployed',
 'What is your current occupation_Working Professional',
 'Last Notable Activity_Had a Phone Conversation',
 'Last Notable Activity_Unreachable'],
```

After Feature selection, the model is created and manually the features are removed until we get a good model .

Features are removed if their p-value is greater than 0.05 and/or their VIF is greater than 5.



Overall accuracy of the model comes to be 79.02 percent.

Overall precision of the model comes to be 78.70 percent.

Overall Recall of the model comes to be 77.07 percent.

# CONCLUSION:-

What mattered most in converting a lead is:-

The total visits a person does on the website.

The total time he sends on the website.

The origin of the lead was Form.

Last interaction with the lead was a phone conversation.

The lead source was:-

1. Welingak Website
2. Olark Chat

Taking all of this in consideration, X Education can drastically increase their Lead Conversion rate.