# LEAD SCORING CASE STUDY

## Problem Statement:

An education company named X Education 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%.

Now, although 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.

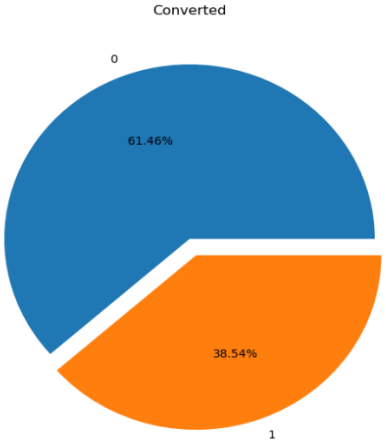
## Objective:

- To help the company to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- To come up with a model wherein it needs to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

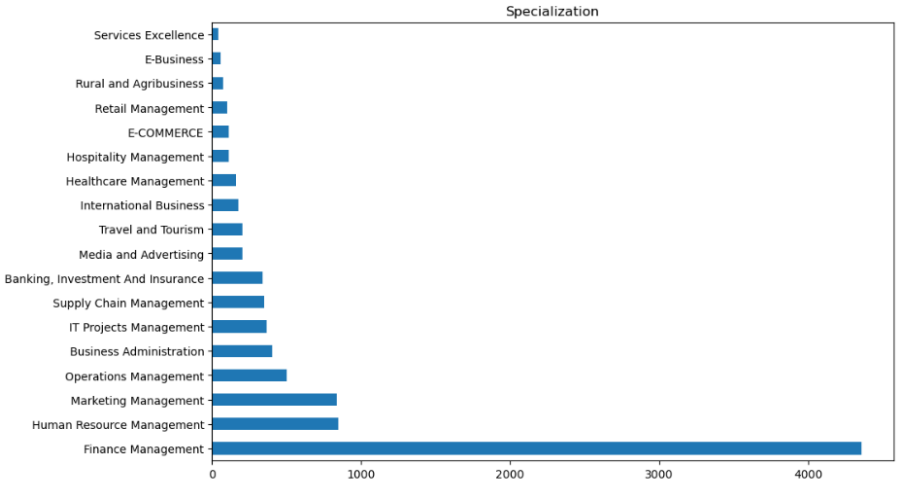
## Analysis:

Using data visualization and EDA, we have drawn few insights about the dataset.

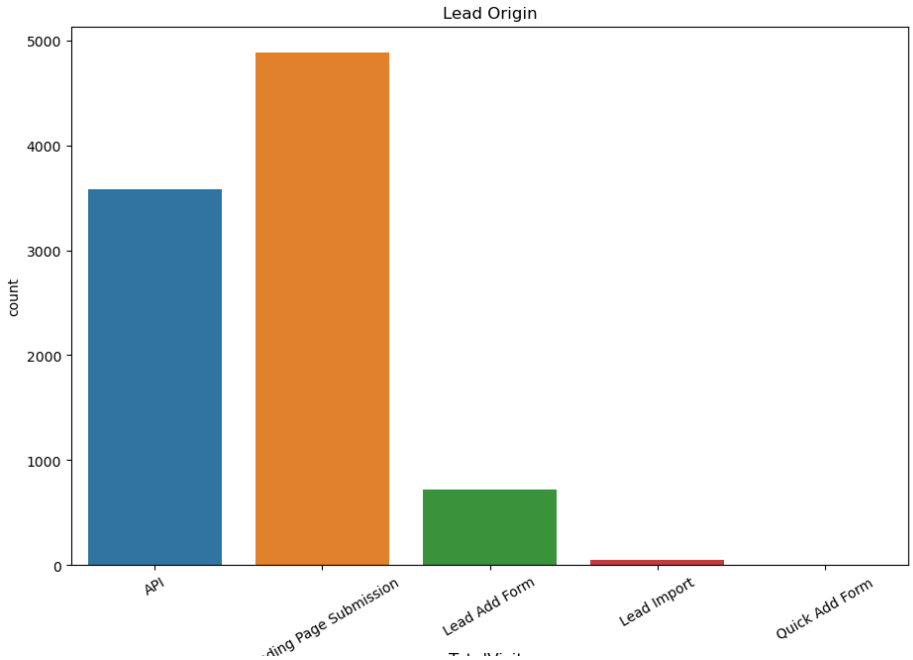
- In the dataset, we have 38.5% of leads who have converted and 61.5% of leads who haven't converted.



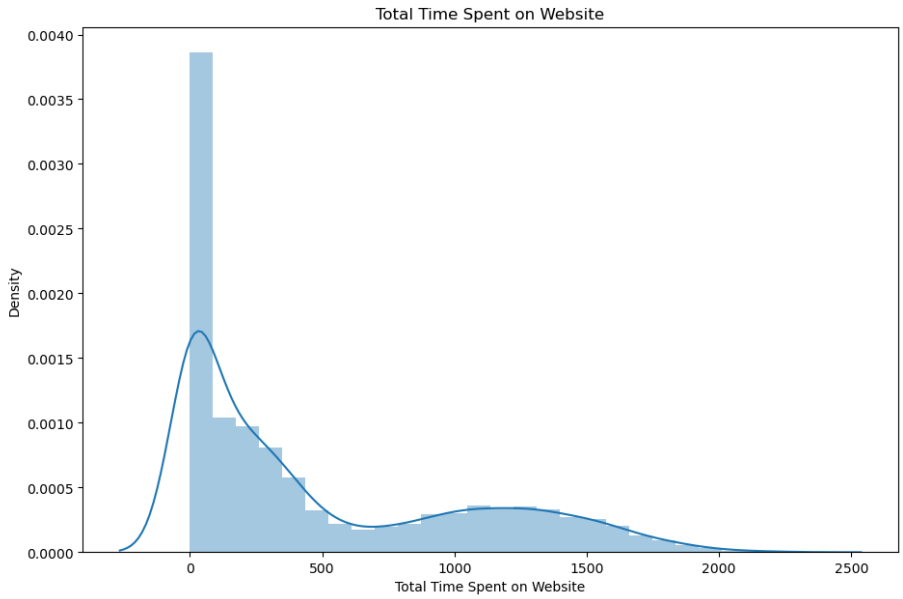
- For most of the features, it is observed there is skewness in the data.



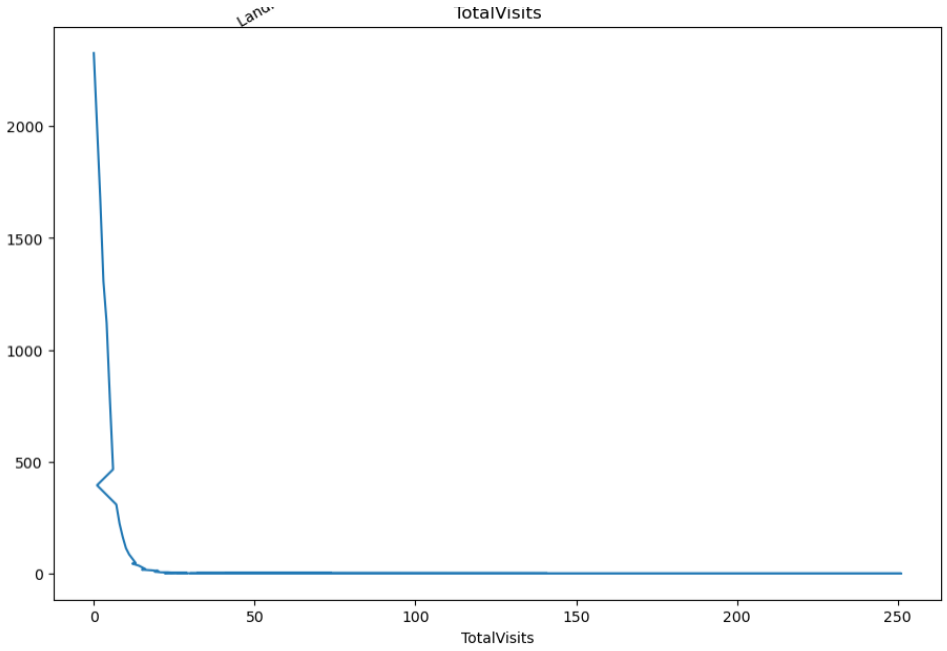

People who worked in Finance Management specialization are found to be appeared the most in the leads.



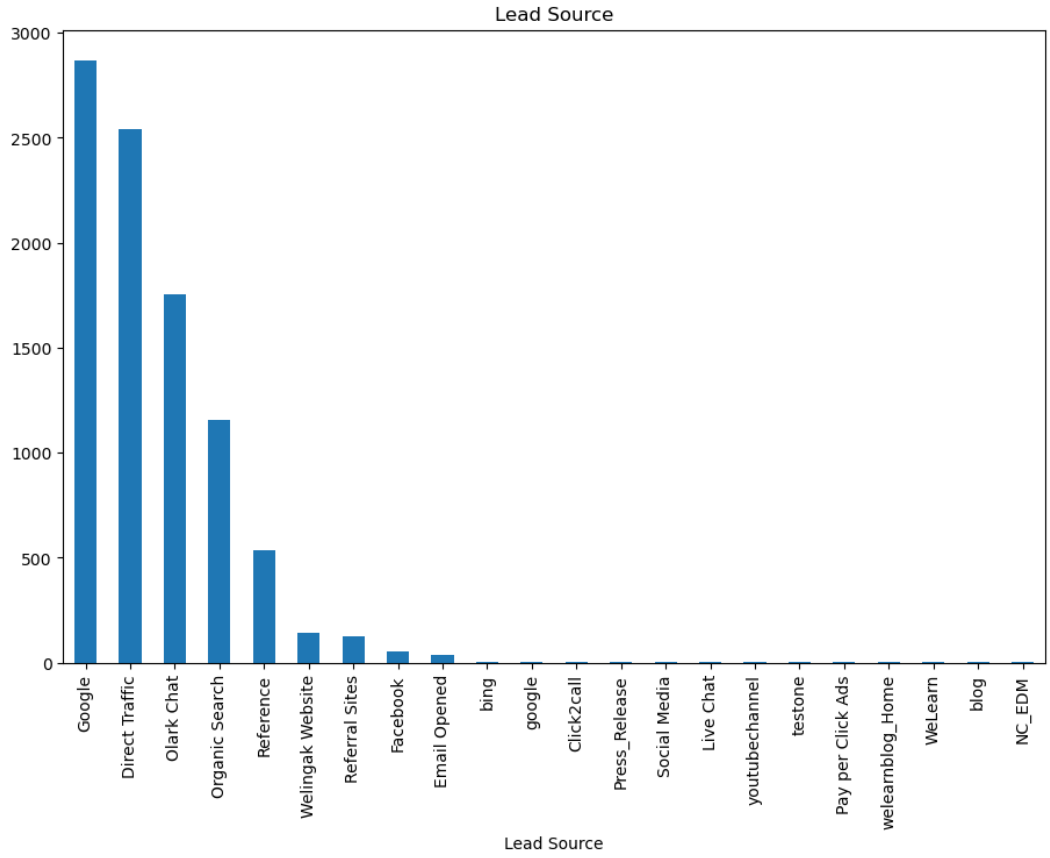

People who have classified as lead are most through Landing Page Submission.



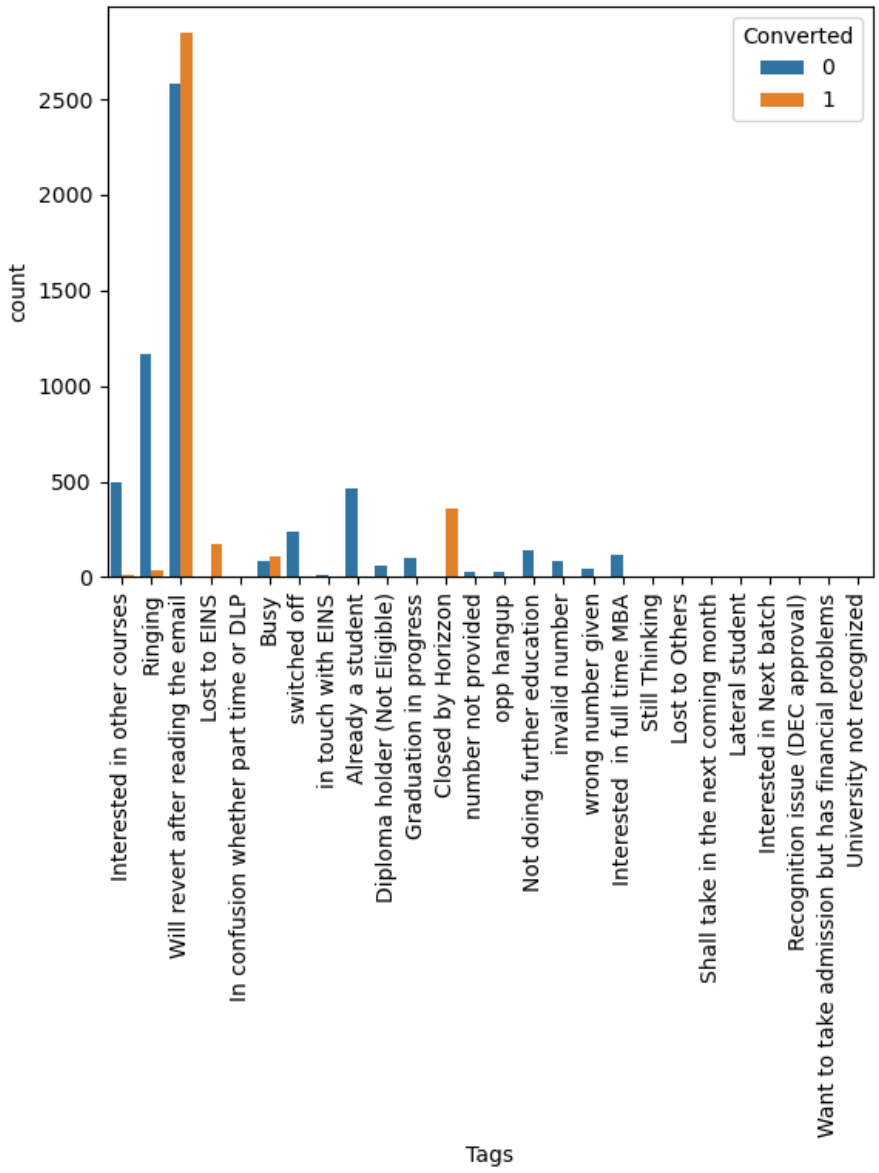

We observed there is positive skewness in the data and the people who have maximum time less in number.




The maximum number of people who visited their website are very less in number.




It is through Google most of the people are classified as leads and next comes Direct Traffic.




Most of the leads have responded saying ‘Will revert after the reading the email’ and its them who also have high conversion rate.

## Identifying the predictors which are driving high for lead conversion:

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -1.5853 | 0.262 | -6.057 | 0.000 | -2.098 | -1.072 |
| **Total Time Spent on Website** | 0.9841 | 0.041 | 24.028 | 0.000 | 0.904 | 1.064 |
| **Lead Origin_Lead Add Form** | 3.2119 | 0.217 | 14.814 | 0.000 | 2.787 | 3.637 |
| **Last Activity_Olark Chat Conversation** | -0.5817 | 0.173 | -3.353 | 0.001 | -0.922 | -0.242 |
| **Last Activity_SMS Sent** | 0.8990 | 0.161 | 5.588 | 0.000 | 0.584 | 1.214 |
| **What is your current occupation_Unemployed** | -0.6573 | 0.257 | -2.556 | 0.011 | -1.161 | -0.153 |
| **What is your current occupation_Working Professional** | 1.8255 | 0.329 | 5.549 | 0.000 | 1.181 | 2.470 |
| **Tags_Closed by Horizzon** | 7.1621 | 1.020 | 7.023 | 0.000 | 5.163 | 9.161 |
| **Tags_Ringing** | -2.8576 | 0.243 | -11.748 | 0.000 | -3.334 | -2.381 |
| **Tags_Will revert after reading the email** | 1.8043 | 0.100 | 18.066 | 0.000 | 1.609 | 2.000 |
| **Last Notable Activity_Modified** | -1.0228 | 0.106 | -9.669 | 0.000 | -1.230 | -0.815 |
| **Last Notable Activity_SMS Sent** | 1.0547 | 0.191 | 5.533 | 0.000 | 0.681 | 1.428 |

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional
- Tags_Closed by Horizzon
- Tags_Ringing
- Tags_Will revert after reading the email
- Last Notable Activity_Modified
- Last Notable Activity_SMS Sent

## Model Metrics:

Confusion metrics representation:

| Actual/Predicted | Not Converted | Converted |
|---|---|---|
| Not Converted | True Negative | False Positives |
| Converted | False Negatives | True Positives |

## Train dataset:

Confusion matrix:

| Actual/Predicted | Not Converted | Converted |
|---|---|---|
| Not Converted | 3268 | 734 |
| Converted | 353 | 2113 |

Accuracy: 87%

Precision: 83%

Recall: 82%

F1 score: 82%

ROC AUC score: 86%

## Test dataset:

Confusion matrix:

| Actual/Predicted | Not Converted | Converted |
|---|---|---|
| Not Converted | 1493 | 184 |
| Converted | 198 | 897 |

Accuracy: 86%

Precision: 83%

Recall: 82%

F1 score: 82%

ROC AUC score: 85%

With the above metrics, we can conclude that the model we have built seems to be doing good. Thus, this model would help the X Education company to identify the most potential leads.