Lead Scoring Case Study

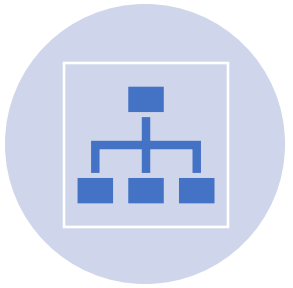
Problem Statement

An education company name X Education sells online courses. On any given day, many students or other persons who are interested in the courses land on their website and browse for course. The company markets its courses on several websites and search engines like Google.

X Education has provide a dataset to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model where the target lead conversion rate to be around 80%.

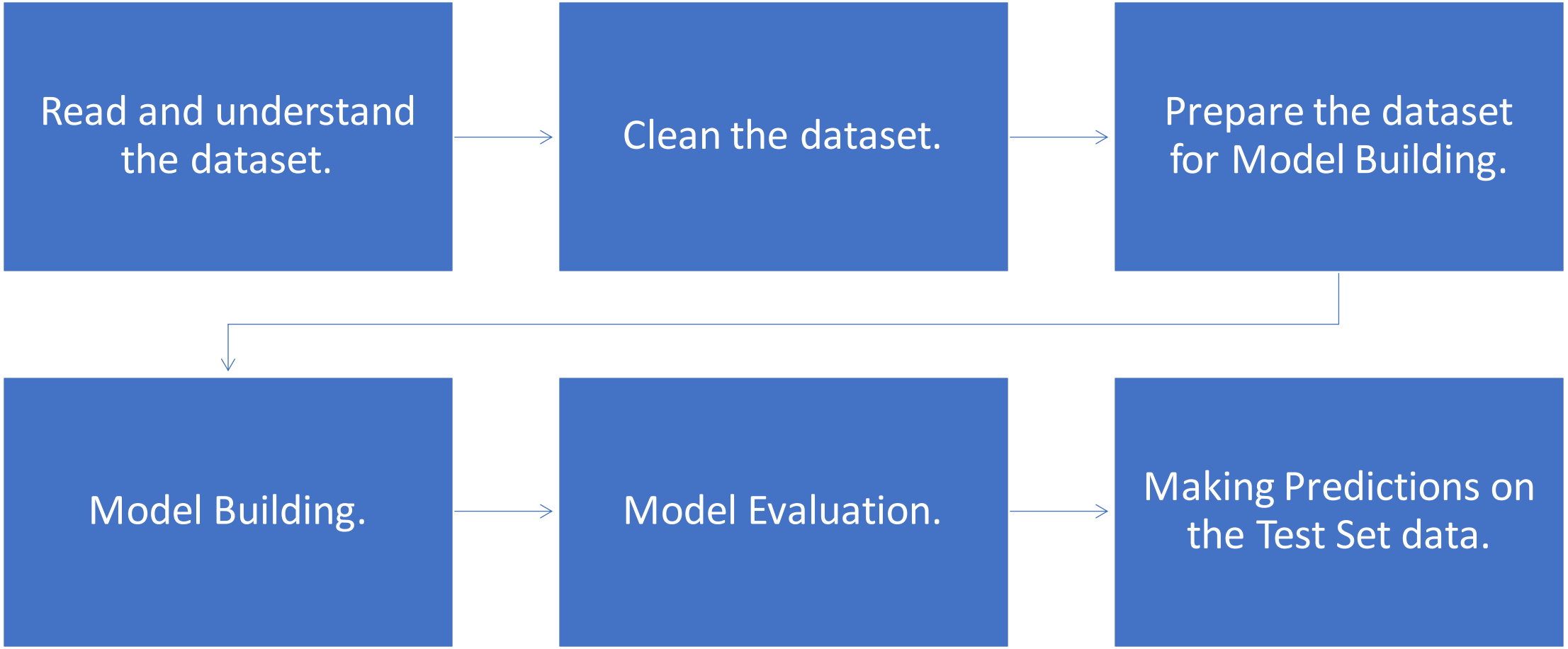
Goals Of the case study

To build a logistic regression model to where lead score assign between 0 and 100 to each of the leads which can be used by the company to get their target potential leads.



To adjust the model if the company's requirement changes in the future so we will need to handle these as well.

Steps done though model building process



Dataset Information

The Company currently has a LEAD conversion rate of 30%

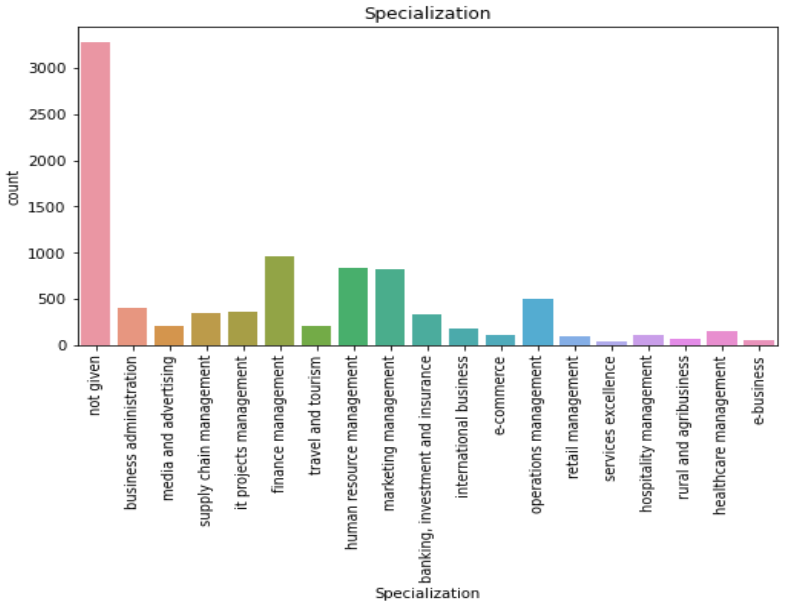
The Company has a target of LEAD conversion to 80%.

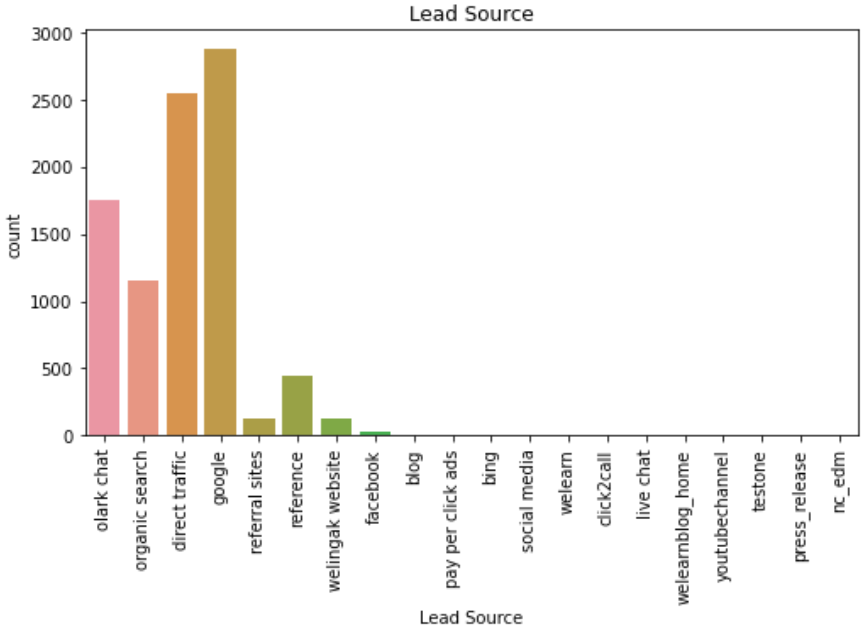
The data set has 9240 rows and 37 columns.

The majority of 30 columns hold object type data, 3 of then have integer type and rest 4 have float type data.

Column name 'Prospect ID' one and only column holds all unique values.

More than 10 columns have null values more than 10%.

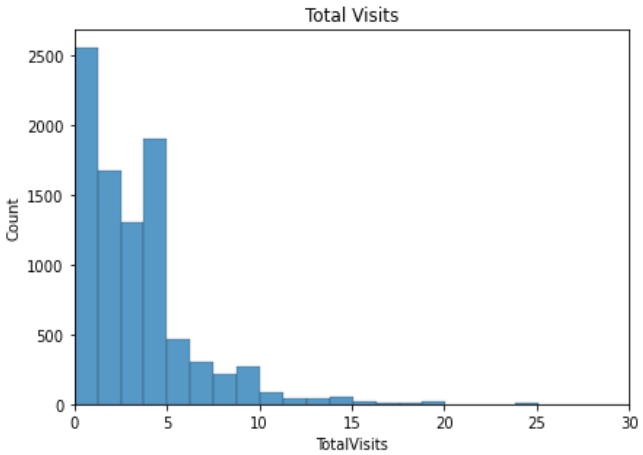


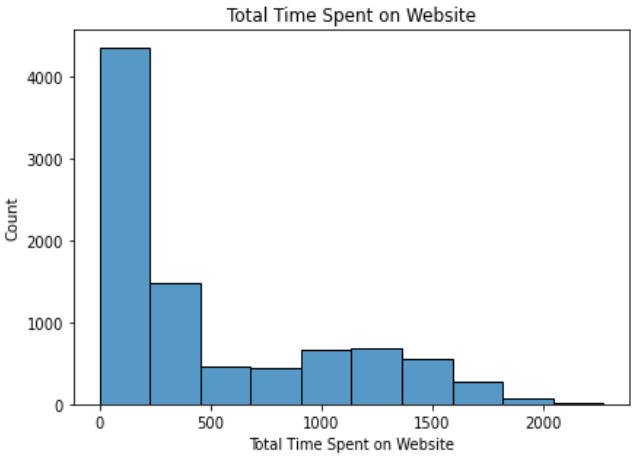


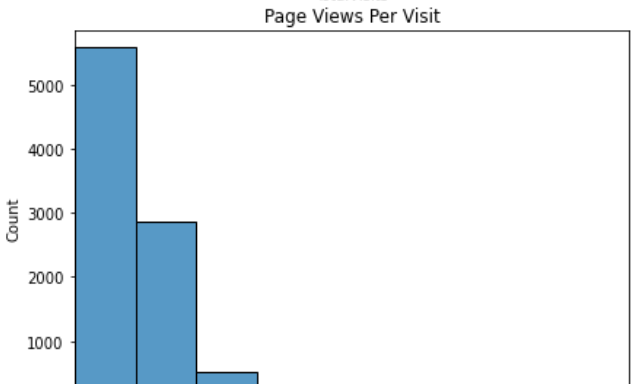
Data Visualization

- Columns like 'Country', 'Lead Source', 'Specialization', 'Last notable activity' etc. Holds high skewed data.

Data Visualization



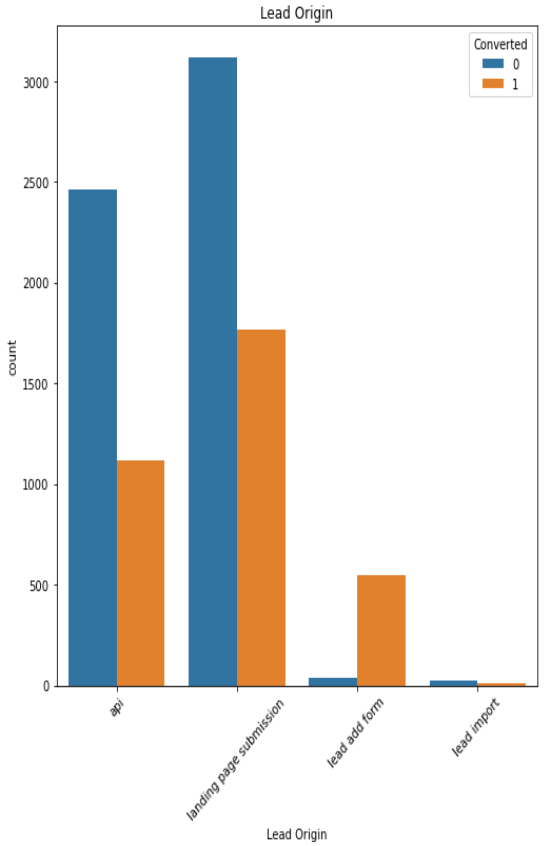


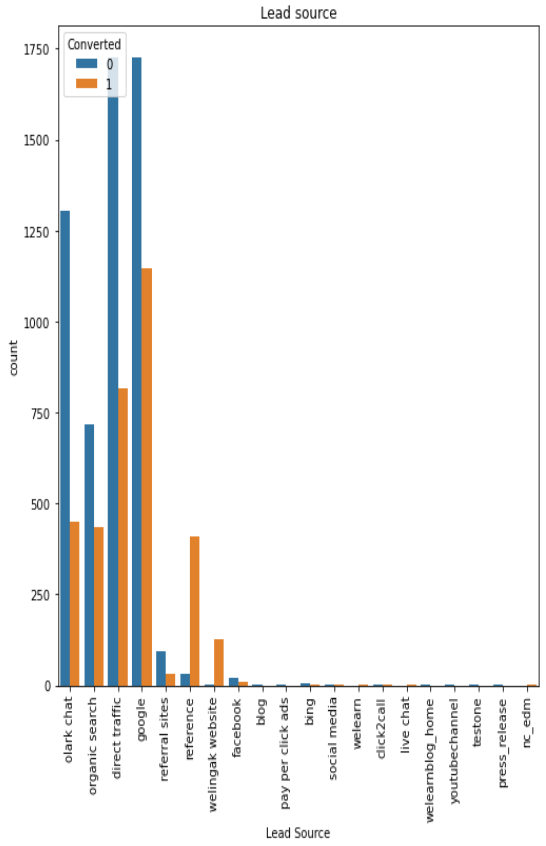


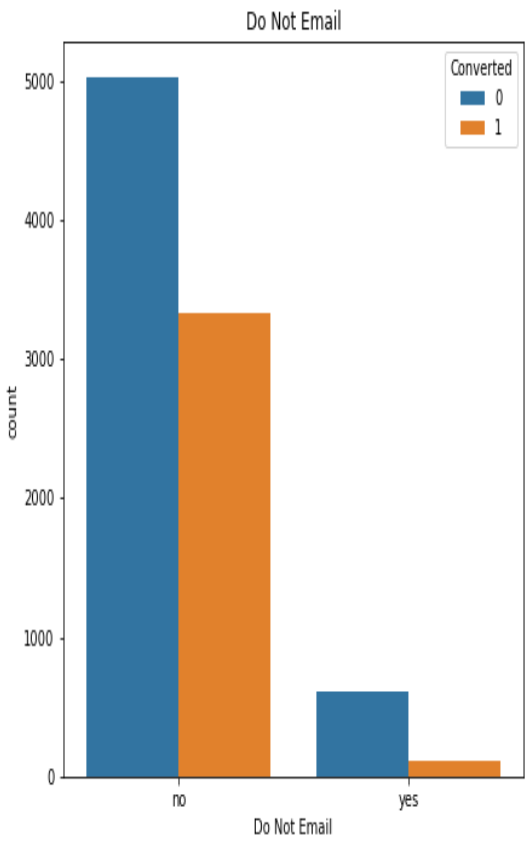
- 'Total Visits', 'Total time spent on website' are the main source of conversation.

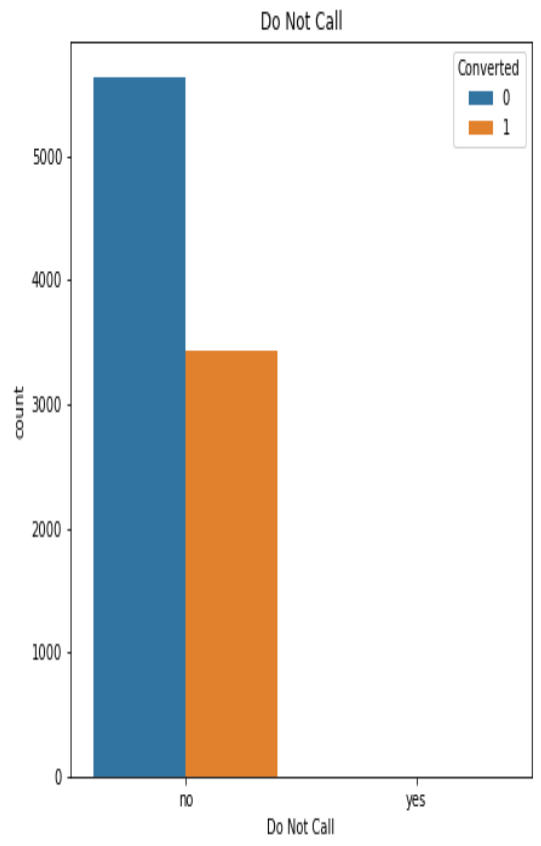
Data Visualization

- 'Lead Origin', 'Lead Source', 'Do not Email' and 'Do not Call' highly effects the lead conversation.



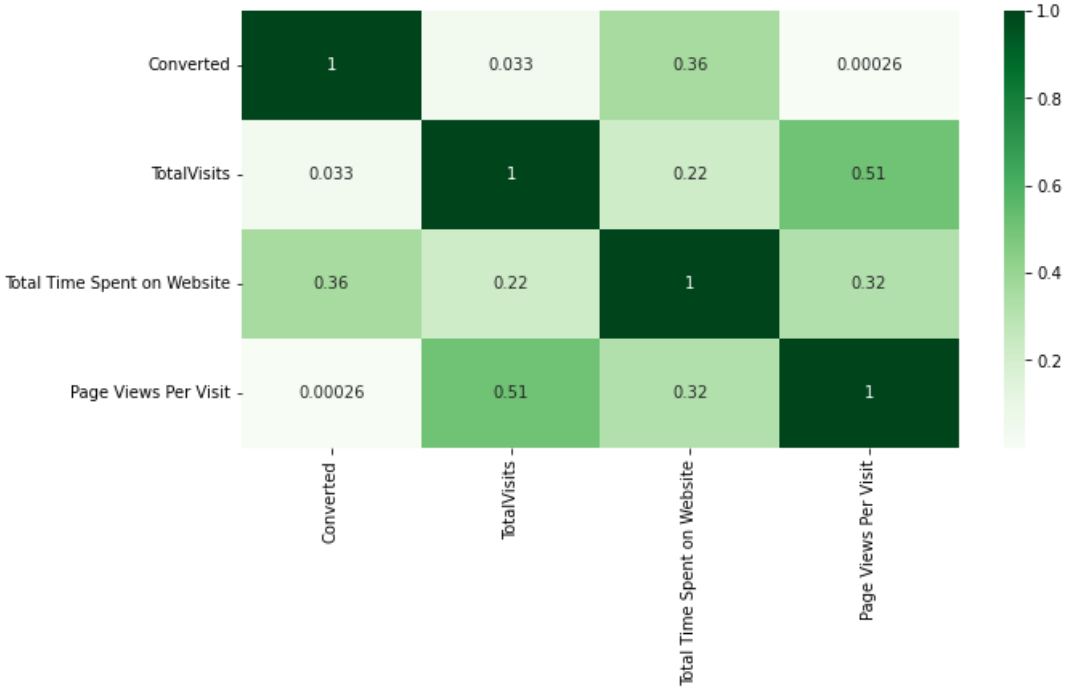






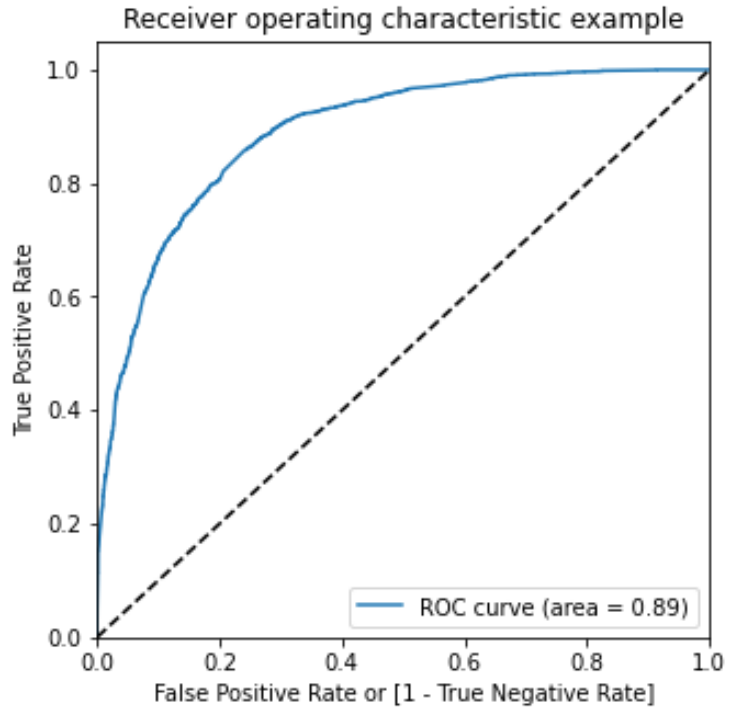
Data Visualization

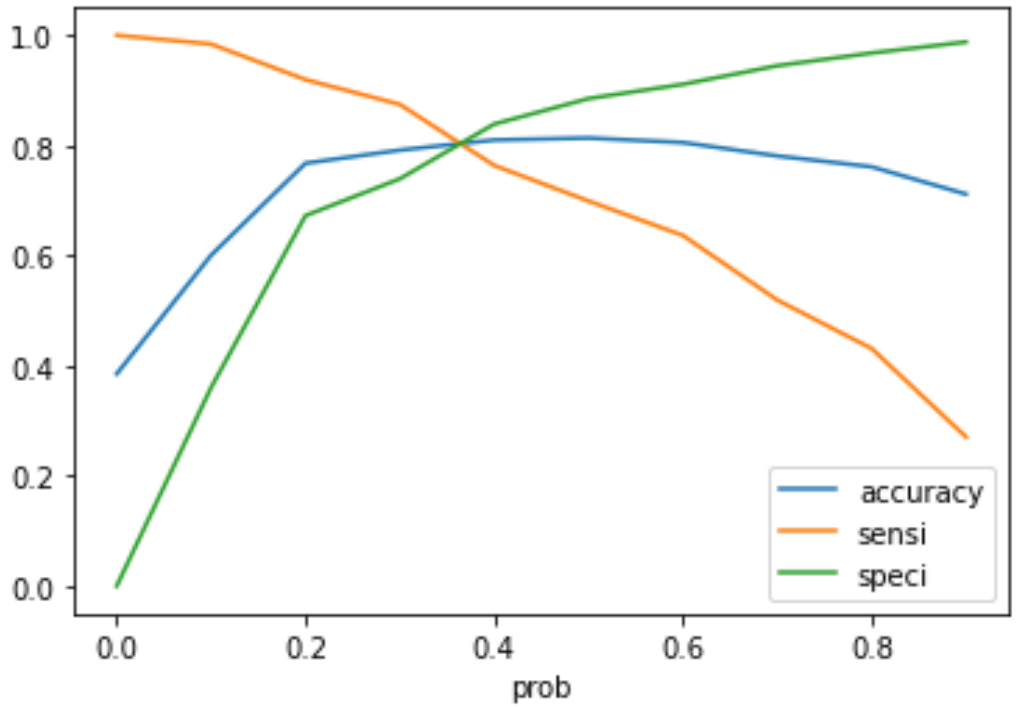
- 'Converted', 'Total Visits', 'Total Time Spent on Websites' and 'Page views per visit' co-related with each other.



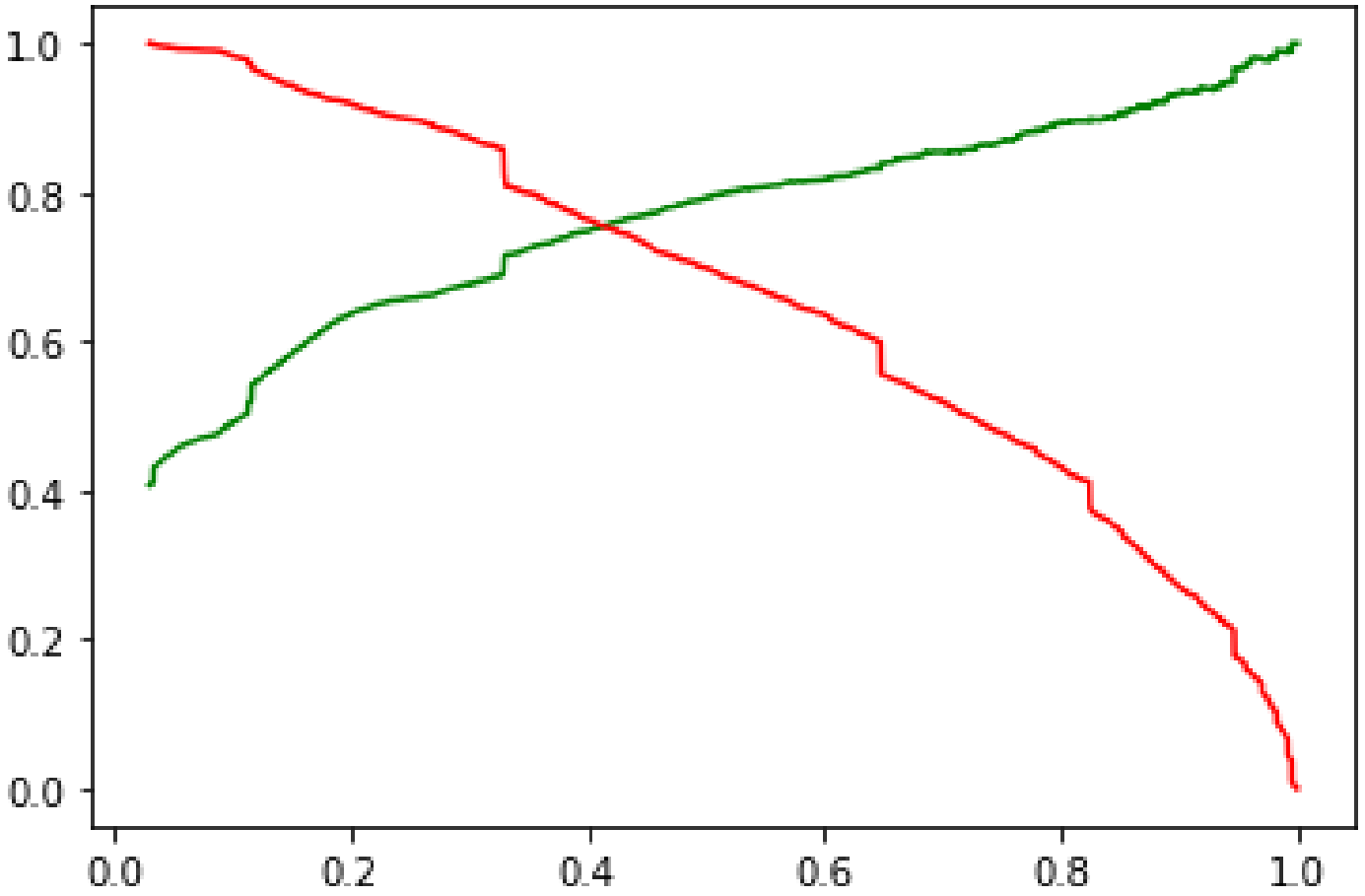
Data Evolution

- The Roc curve in being generated.
- From the curve get 0.37 is the optimum point to take it as a cutoff probability.



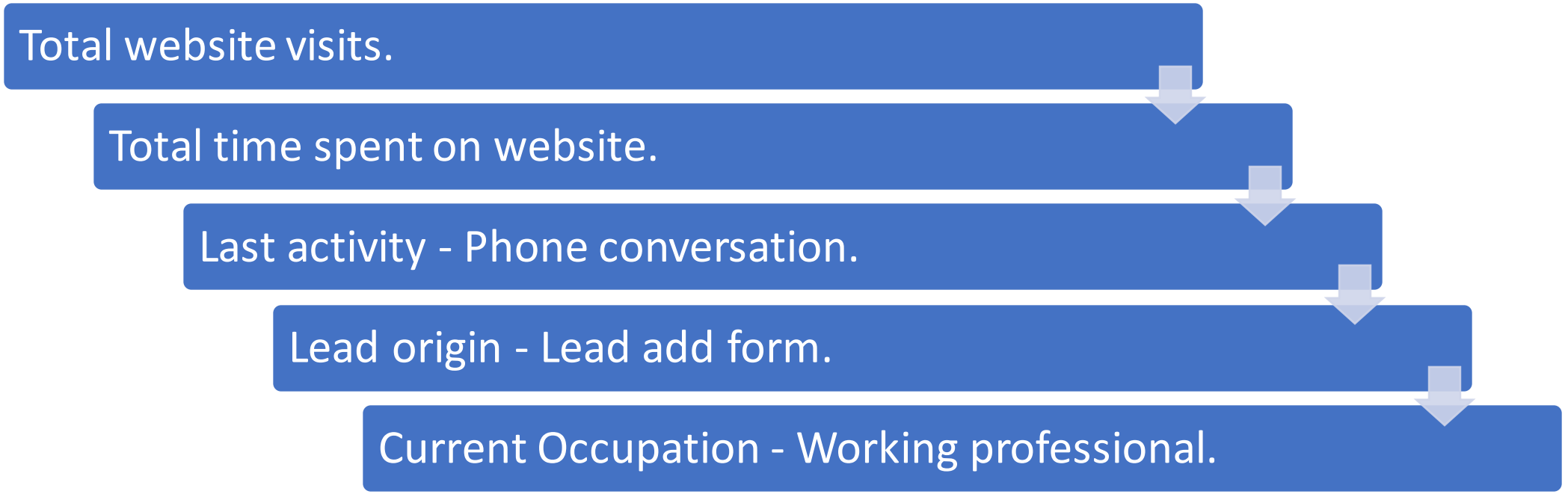


Data Evolution

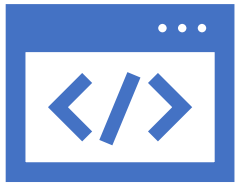


- So from the Optimum Cutoff Point of 0.36 we get -
- Accuracy of 80%
- Sensitivity of 79%
- Specificity of 81%
- So with cutoff of 0.41 we get -
- Accuracy of 82%
- Precision of 74%
- Recall 76%

Most reasons behind high leads



Business Recommendations



Making websites more informative and attractive to increase the time spent on the website.

Mostly target working professionals, who want to switch their career field or want to grow their career more.

Provide some mentoring or industry expert guidance sessions to make the conversation rate much higher.

- ***Thank You.***