



# LEAD SCORING CASE STUDY

By KRUNAL NIRAVADEKAR



# 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.
- 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. 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. A typical lead conversion process can be represented using the following funnel:



## BUSINESS OBJECTIVE

- ▶ The company wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.



# OVERALL APPROACH

- 1. Reading and understanding the data.
- 2. Data Preparation.
- 3. EDA.
- 4. Preparing data for model building.
- 5. Model building.
- 6. Model Evaluation.
- 7. Making prediction using the data.



# Reading and understanding the data.

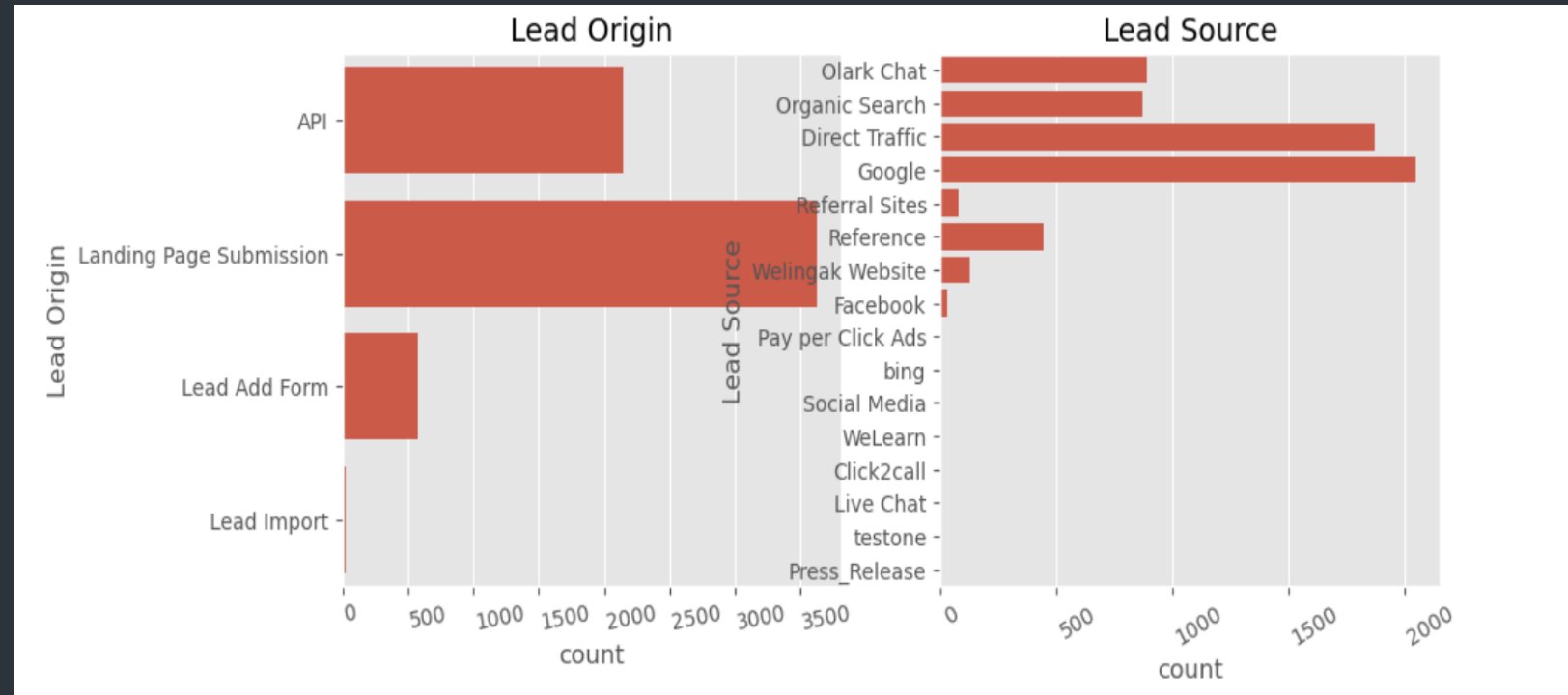
- The main dataset which is Leads.csv is used for the analysis.
- There are 9240 rows and 37 columns in the dataset which is scanned for further analysis.
- All the other necessary details of the data is removed which are type of data, shape etc



# Data Preparation.

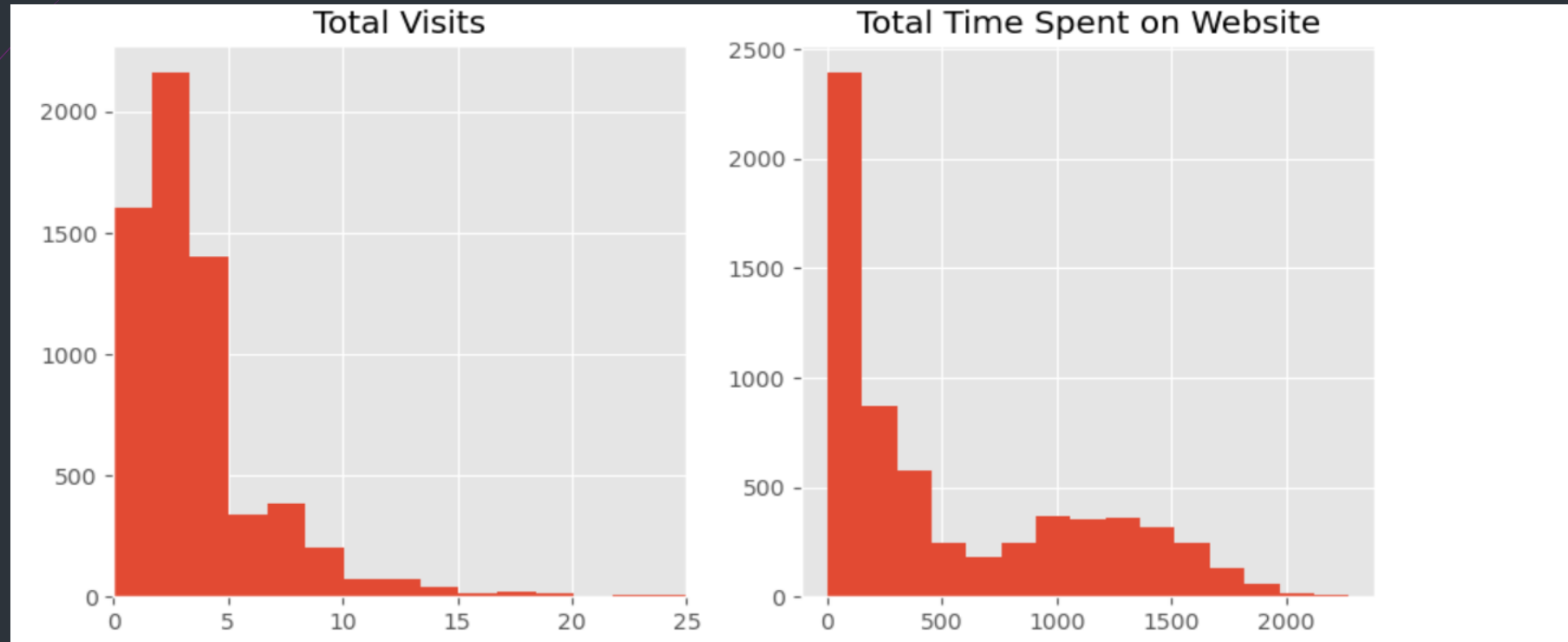
- Null values are observed from the dataset in order to remove unnecessary columns which are not required for analysis
- Columns having null values more than 3000 are dropped straight away.
- Unnecessary columns such as 'CITY' 'COUNTRY' and many other are drop for cleaning the data.
- Null values of certain important columns are removed rather than removing the entire columns for eg; 'What is your current occupation', 'Specialization', 'TotalVisits', 'Leads source'.

# EXPLORATORY DATA ANALYSIS



- Above charts shows various lead origin and its sources.

# EXPLORATORY DATA ANALYSIS



- Above charts shows Total visits and Total time spent on websites by individuals according to datasets.





## Preparing data for model building.

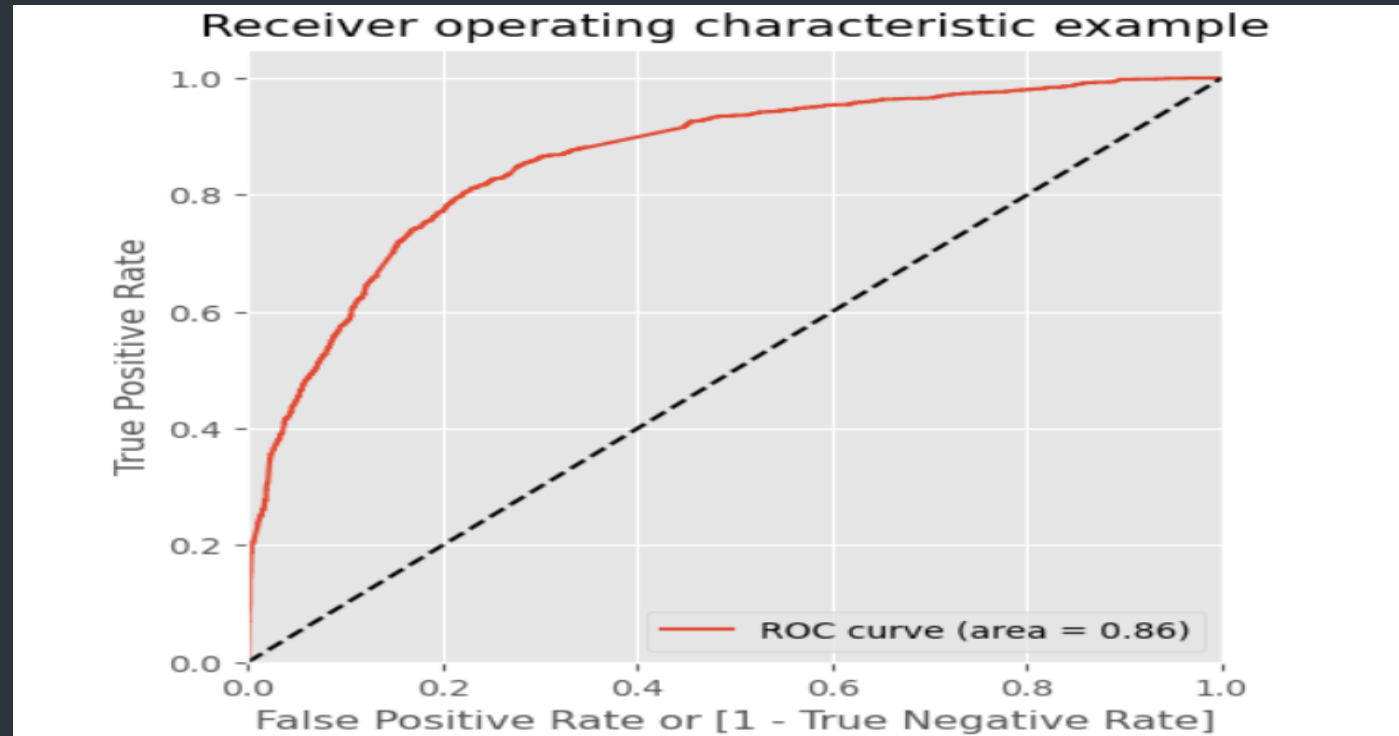
- Dummy variables are created in data for variables containing object datatypes.
- Special dummy variable is created for 'SPECIALIZATION' column.
- The data is divided into train and test for further model building.
- Plotted a correlation heatmap for better understanding.



# Model building.

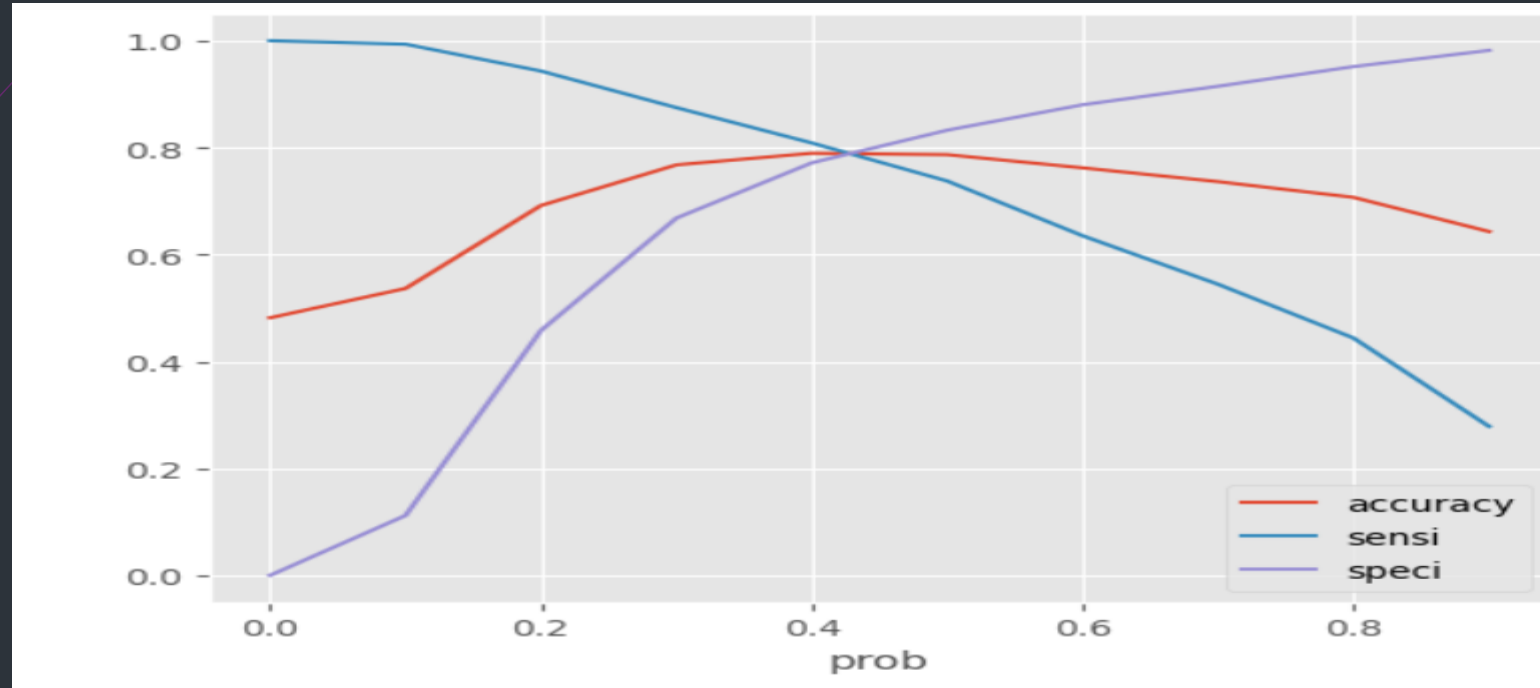
- Build a logistic regression model.
- Build a RFE using 15 variable model for the analysis.
- All the columns having high p values and VIF are removed and we have rebuilt the model

# Model Evaluation



- ROC curve covered an area of 0.86

# Model Evaluation



- ▶ We got 0.42 as the cutoff optimal value
- ▶ **We got 79.07% Accuracy on the test data**



# Conclusion

- According to our analysis X Education company should focus more on the following variable
  1. Total Visits
  2. Total Time Spent on Website
  3. What is your current occupation\_Unemployed
  4. Last Activity\_SMS Sent