LEAD SCORE CASE STUDY

- By- Mukesh Kumar (mk130819978@gmail.com) &
- Deepa Gupta (contactdeepagupta@gmail.com)

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. 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'

Goals

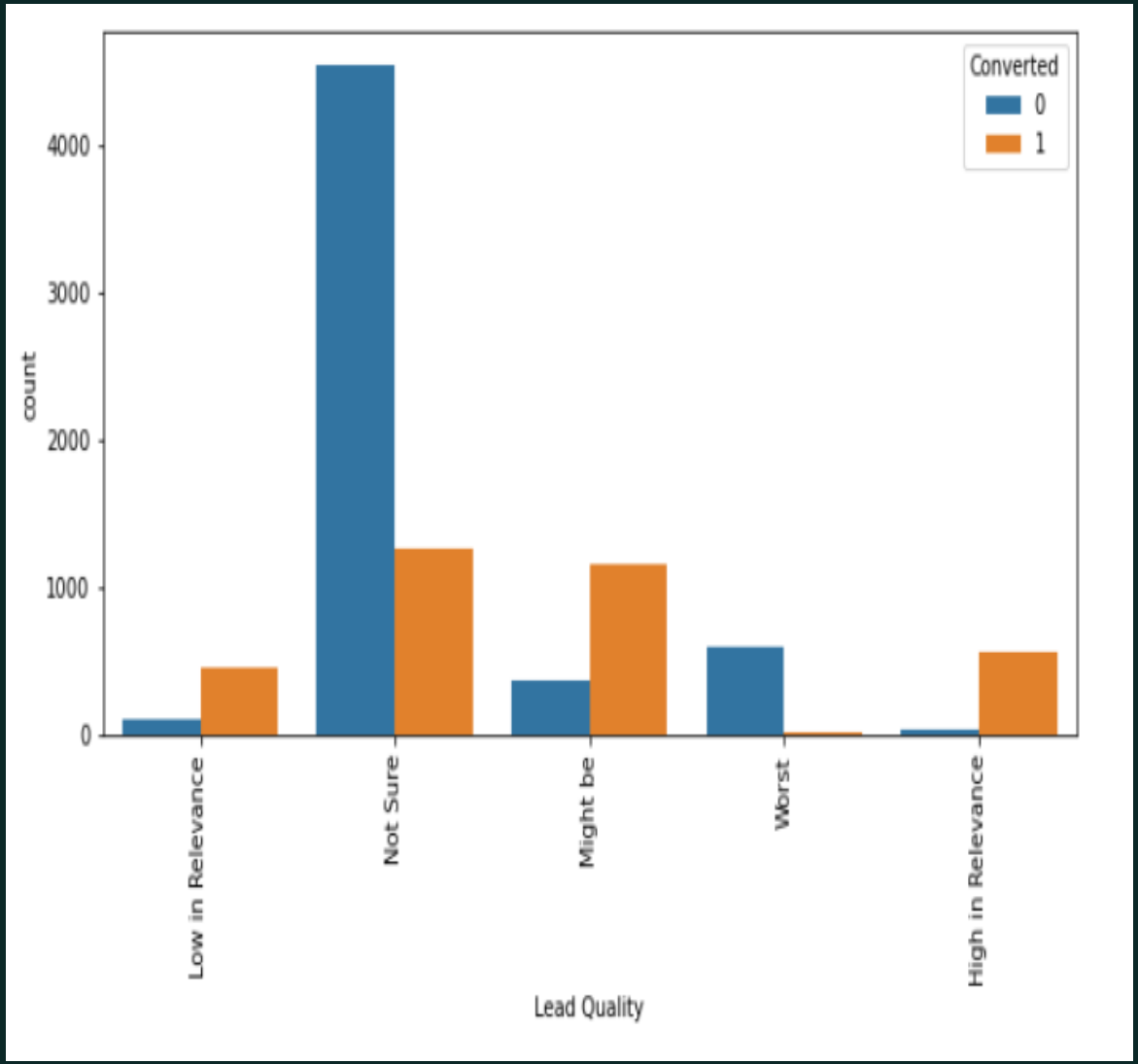
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

EDA (Exploratory Data Analysis)

There is one dependent/Target variable names as "Converted" having the Categorical value 0 and 1.

0-means lead was do not converted to join the organization.

1-means lead was ready to converted to join the organization.



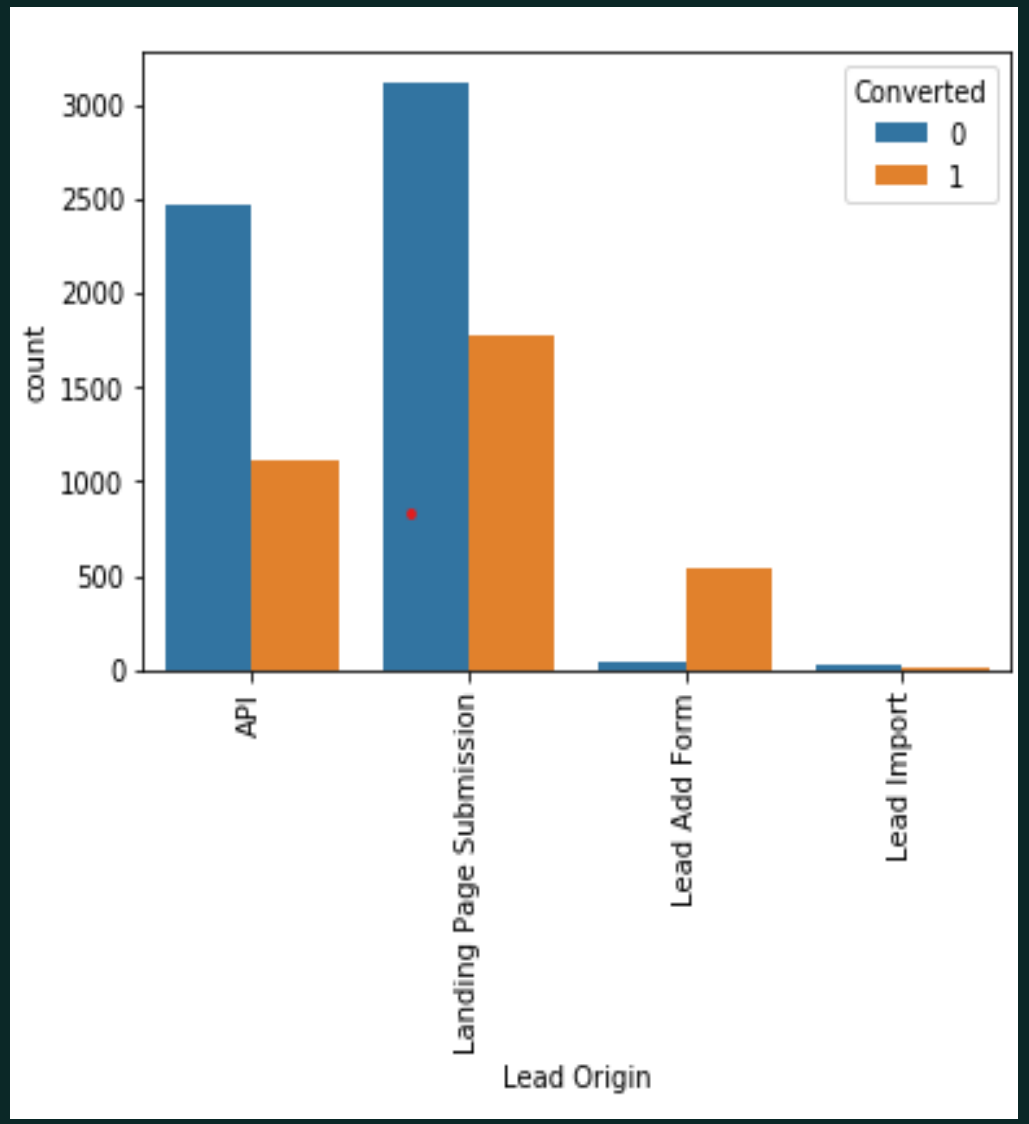
Target variable : Converted
Independent variable: Lead Quality

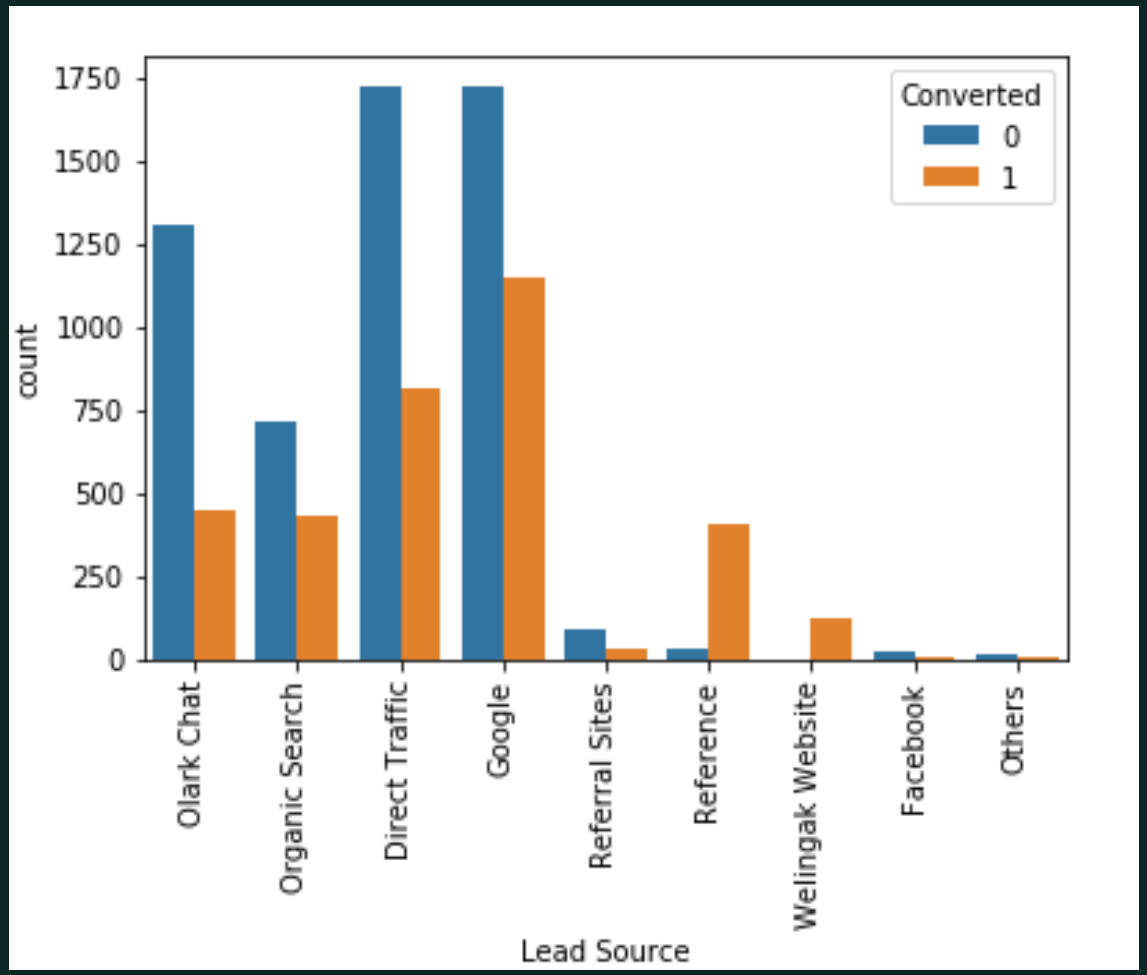
- **Inferences**
  - "Not Sure" people are high to say no but also giving 20% of "YES"
  - "Might be" people are high to say Yes as compare to no
  - "High in relevance" are also very high "yes" as compare to "NO"

Target variable : Converted
Independent variable: Lead Origin

Inference

- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.





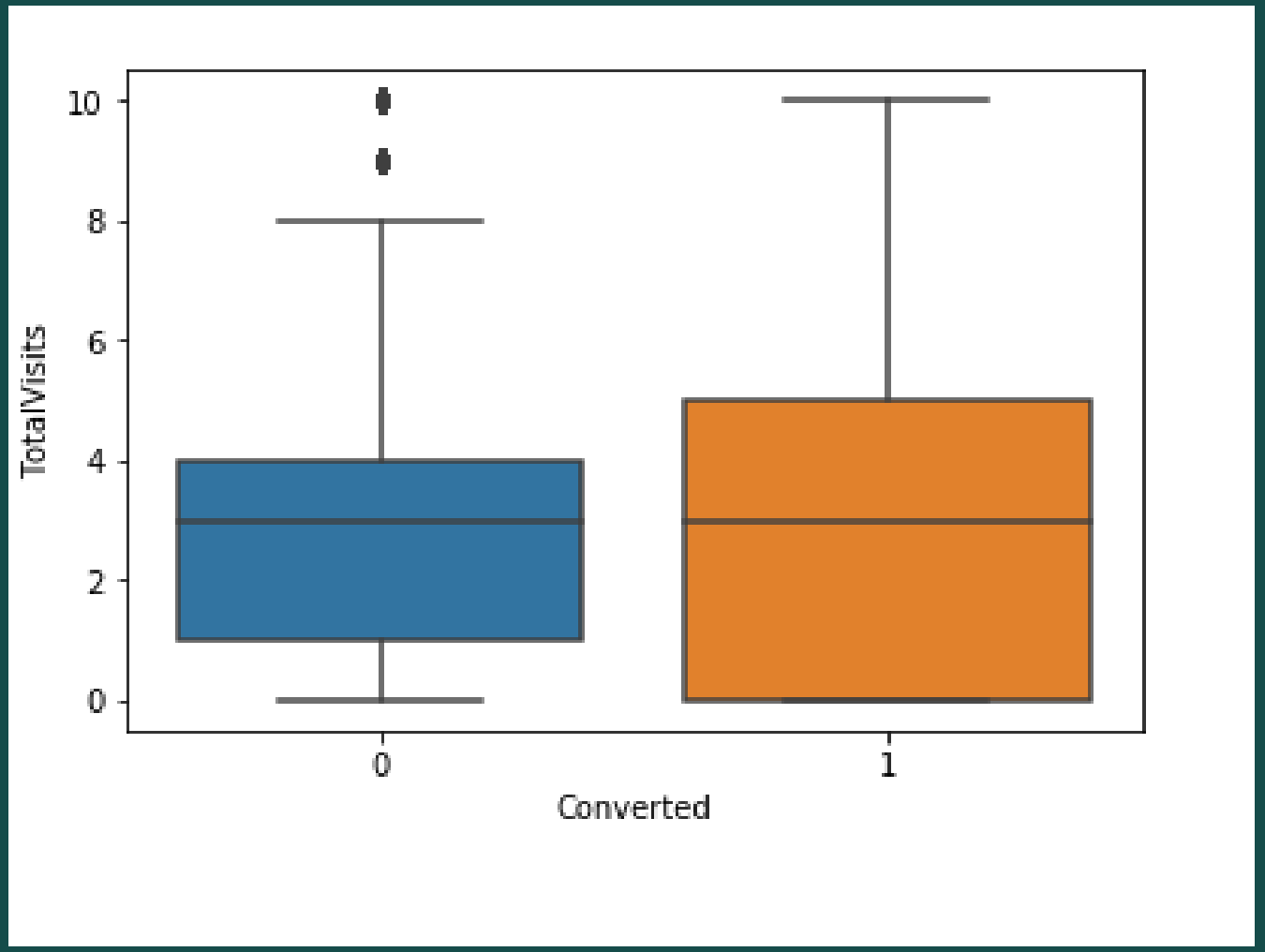
Target variable : Converted
Independent variable: Lead Source

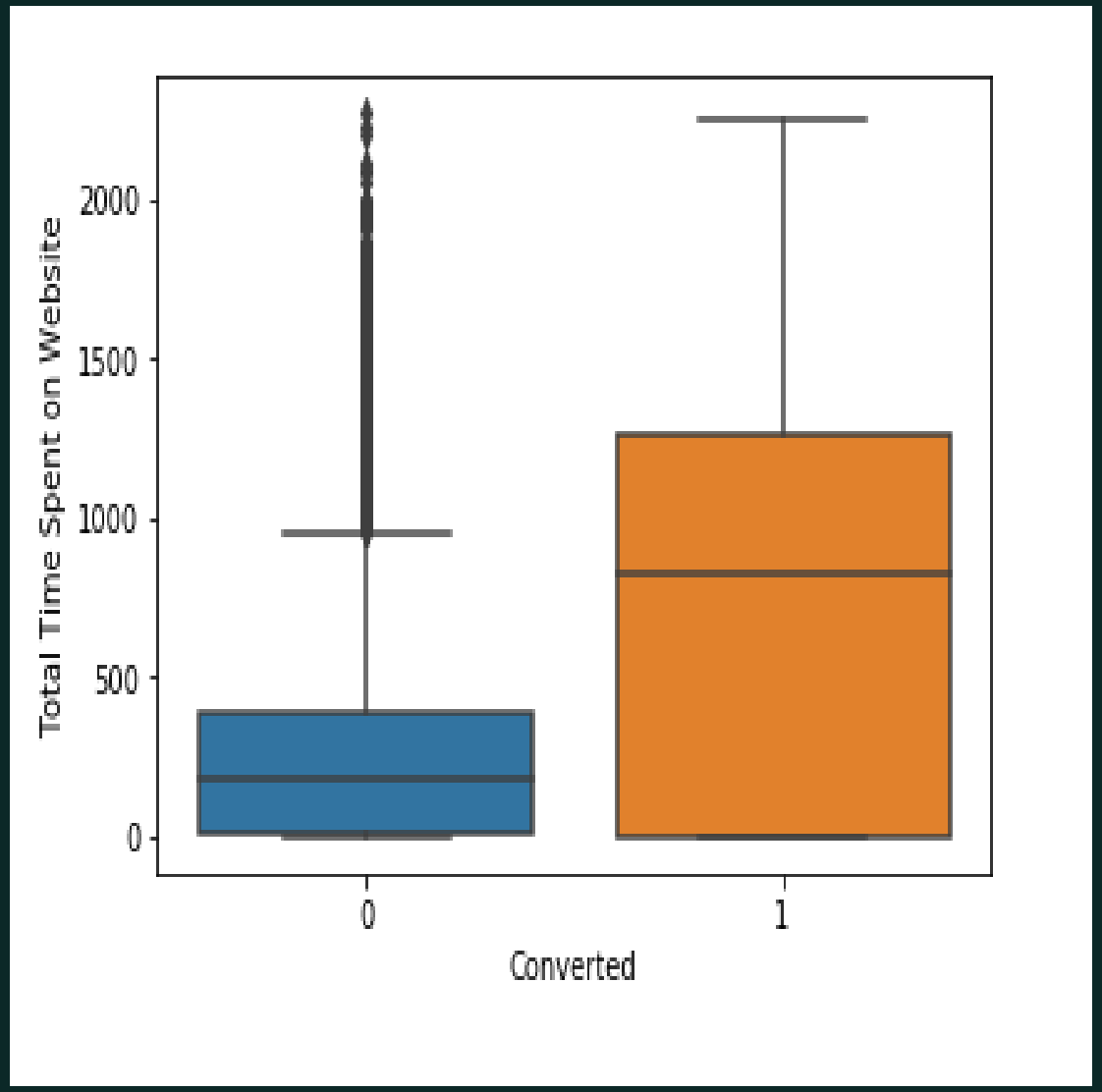
- Inference
- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.

Target variable : Converted
Independent variable: Total Visits

Inference

- Median for converted and not converted leads are the same.

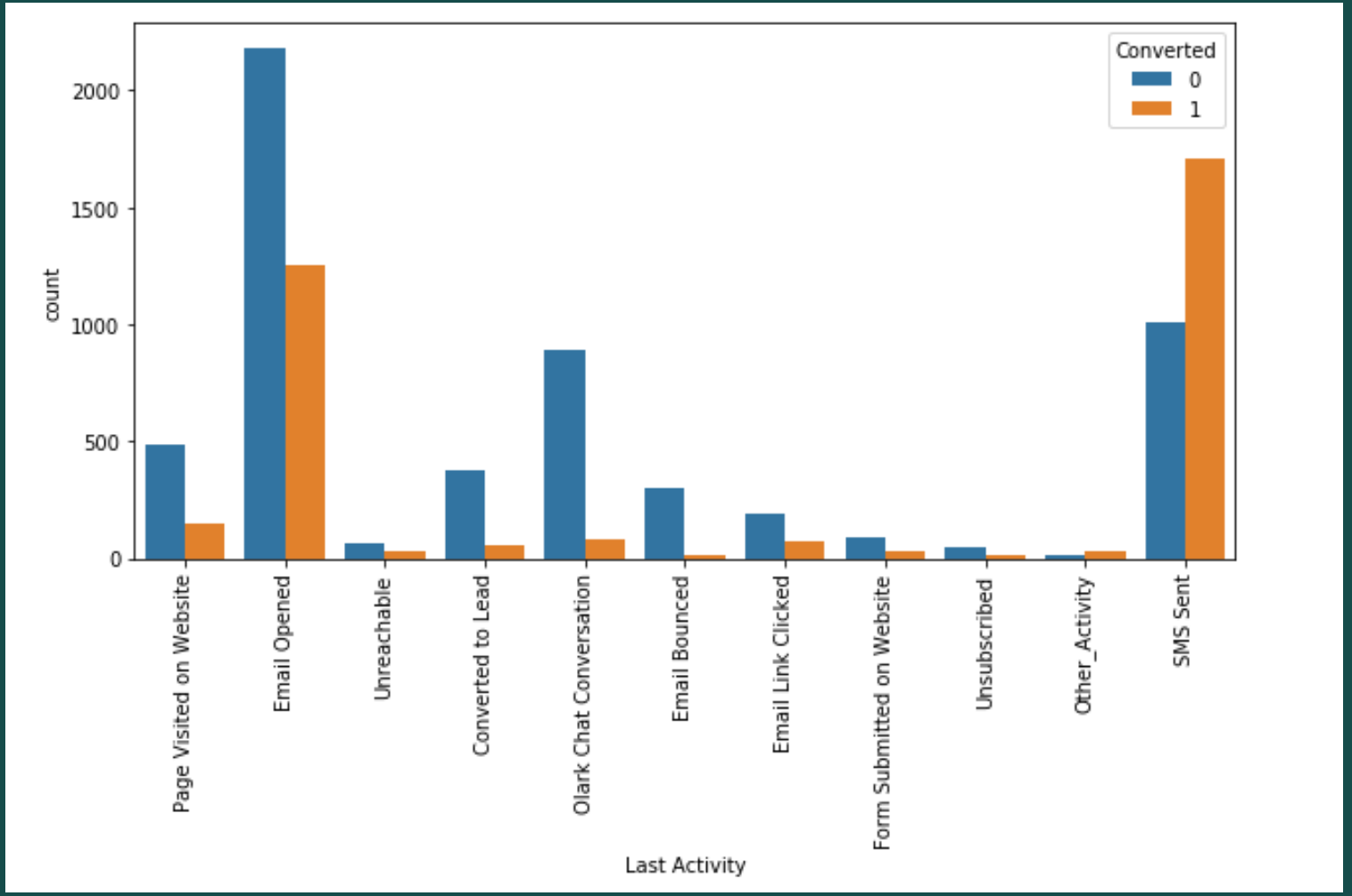




Target variable : Converted
Independent variable: Total Time Spent on Website.

- Inference
- Leads spending more time on the weblsites are more likely to be converted.
- **Website should be made more engaging to make leads spend more time.**

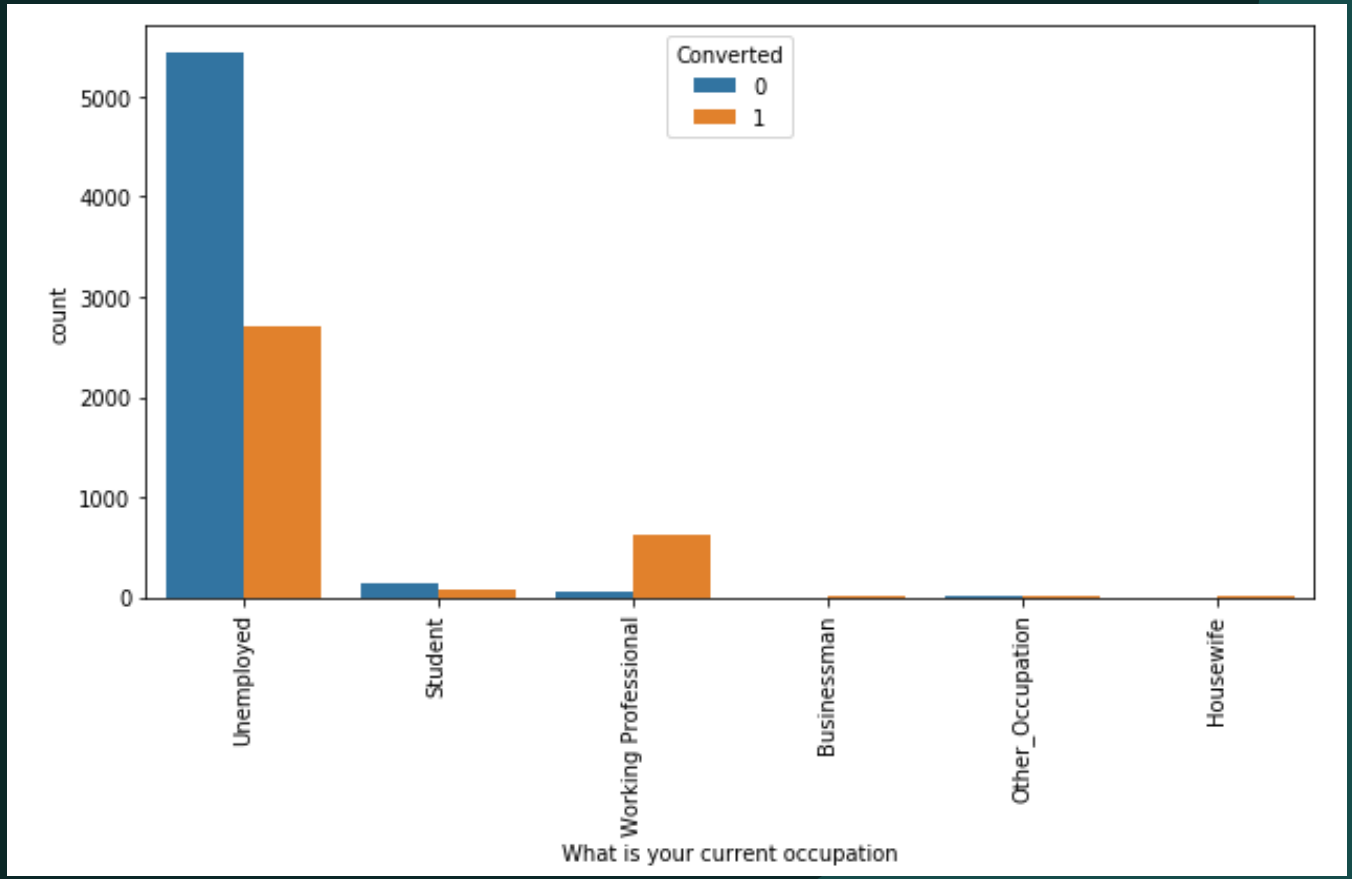
Target variable : Converted
Independent variable: Last Activity



Inference

- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost 60%.b

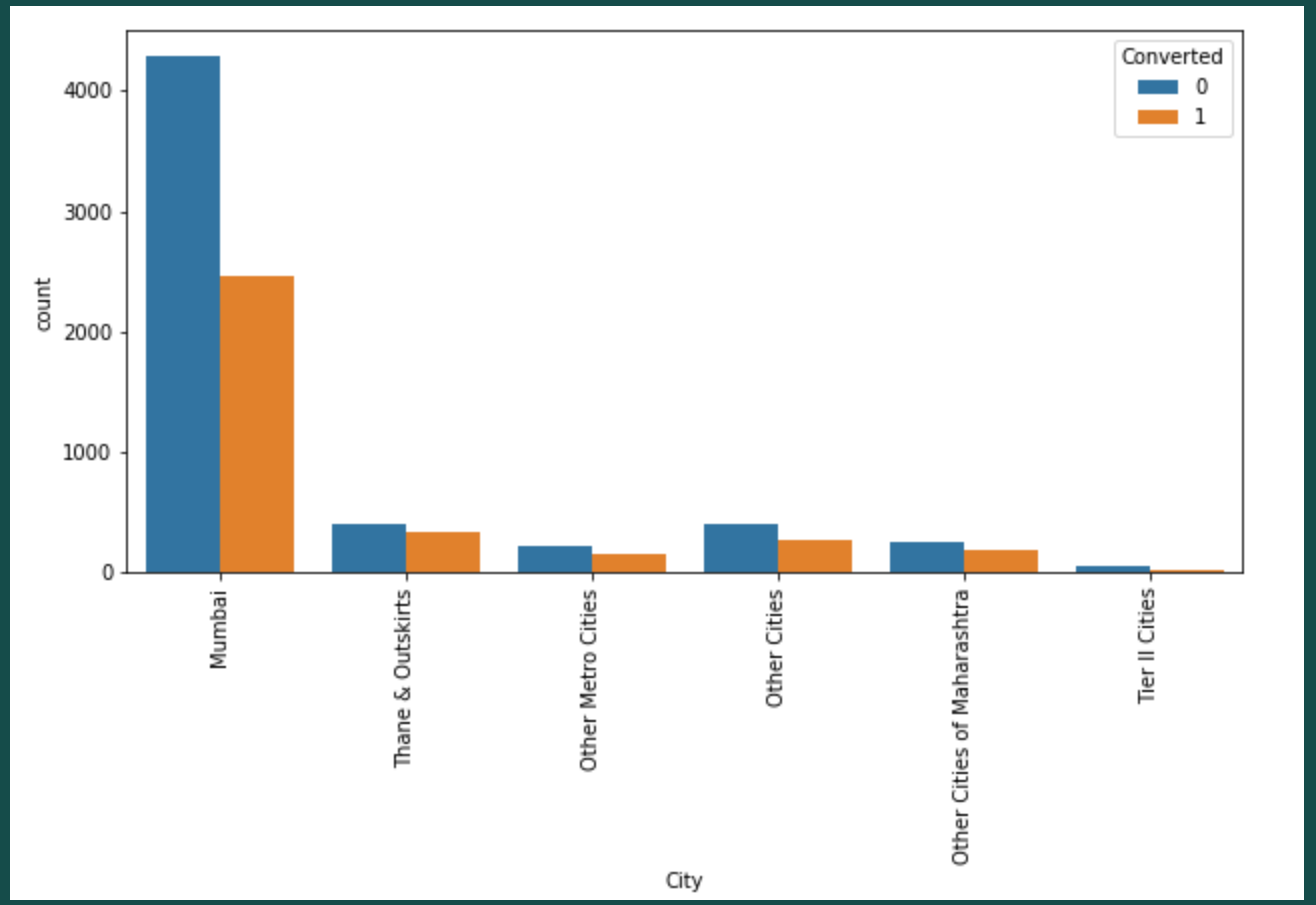
Target variable : Converted
Independent variable: What is your current occupation



Inference

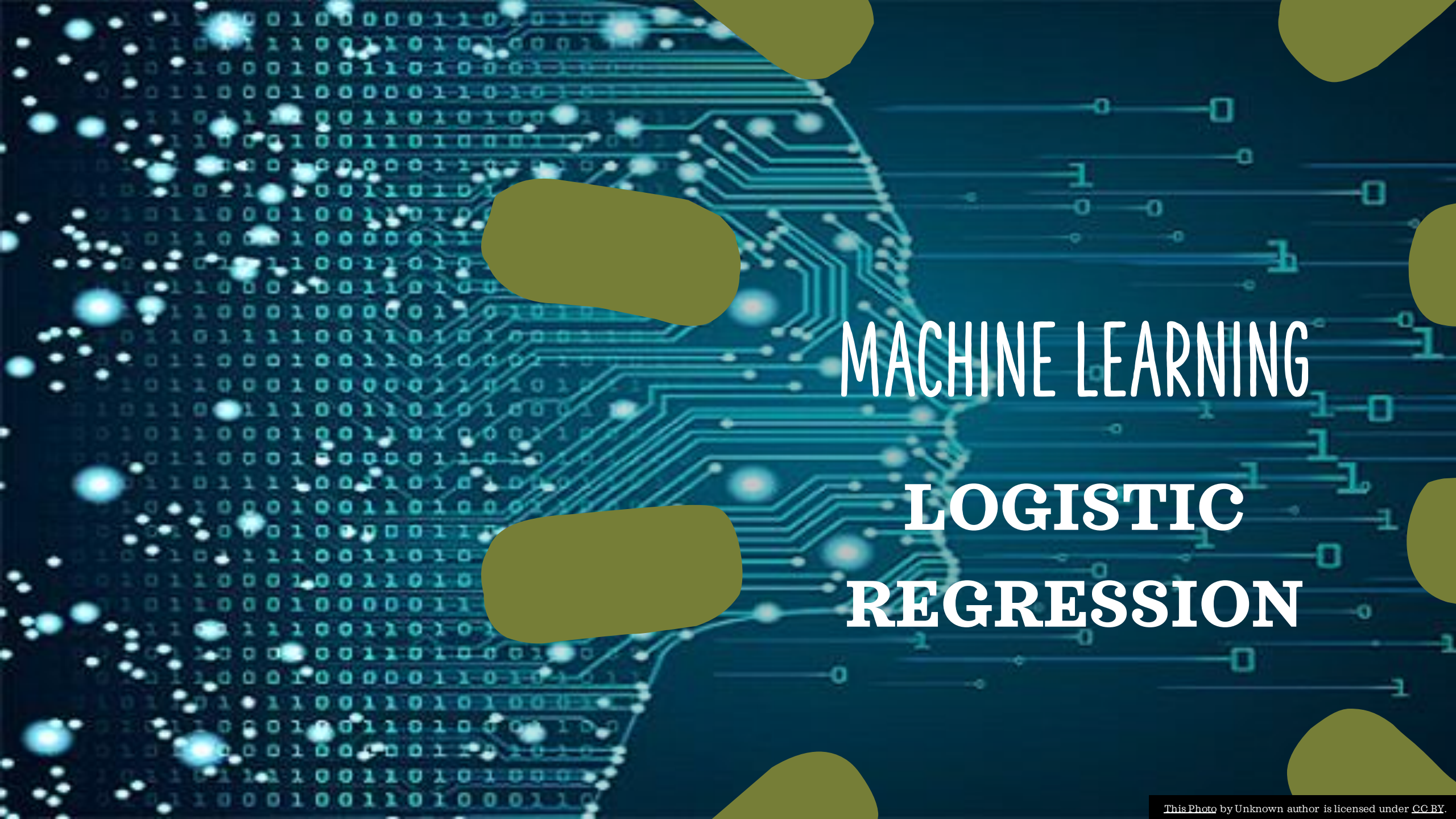
- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in numbers but has around 30-35% conversion rate.

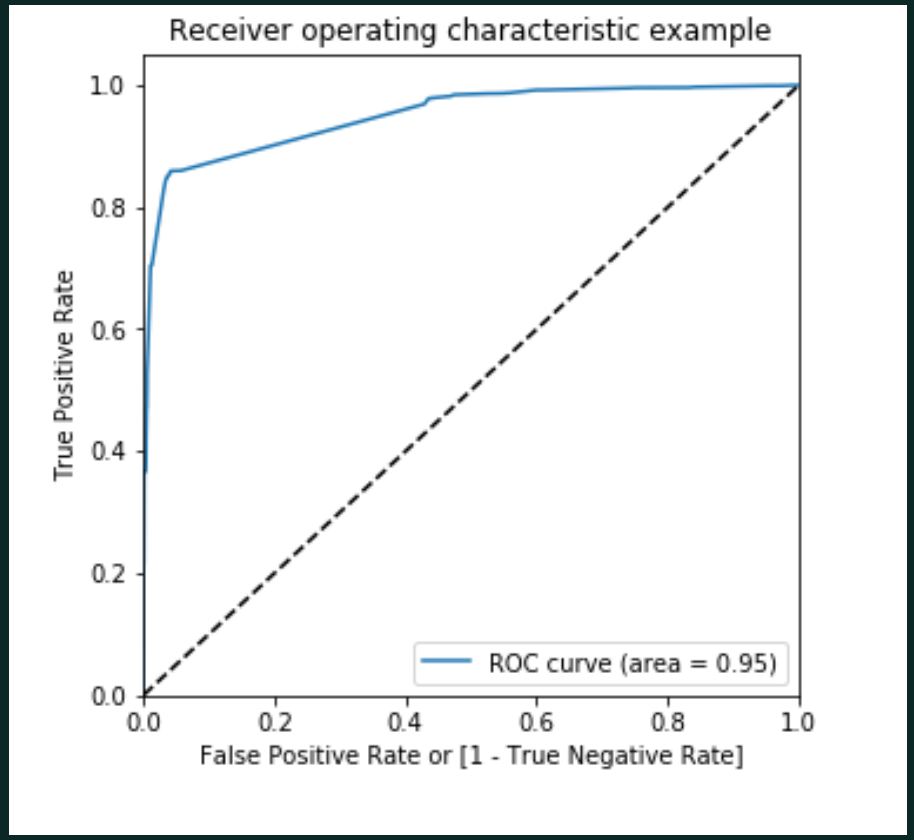
Target variable : Converted
Independent variable: City



Inference

- Most leads are from Mumbai with around 30% conversion rate.





ROC CURVE

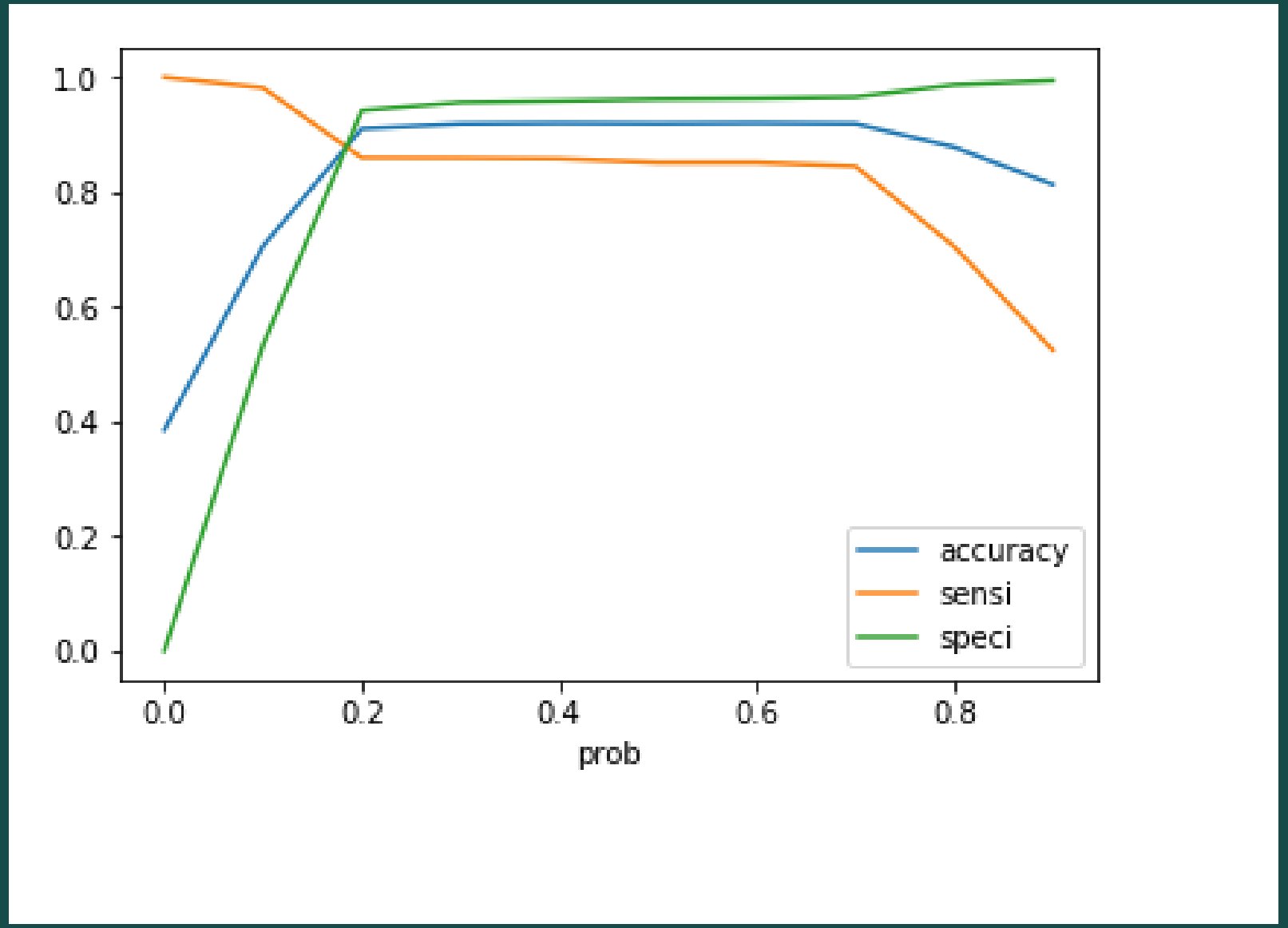
- Inference:
  - Our ROC curve looks so good and bend toward y axis.
  - So this means that our model is giving good predictive values.

OPTIMAL CUT-OFF

Let's plot accuracy sensitivity and specificity for various probabilities.

Inference:

- *From the graph we saw that the optimal cut off will be* **0.2**



Similarity in the results of Train and Test data set

TRAIN DATA SET

- Accuracy: **0.90**
- Specificity: **0.94**
- Sensitivity: **0.85**

TEST DATA SET

- Accuracy: **0.90**
- Specificity:**0.94**
- Senstivity:**0.84**

Top variables from the data are:

- *Do Not Email*
- *Lead Origin_Lead Add Form*
- *Lead Source_Welingak Website*
- *What is your current occupation_Unemployed*
- *Tags_Busy*
- *Tags_Closed by Horizzon*
- *Tags_Lost to EINS*
- *Tags_Ringing*
- *Tags_Will revert after reading the email*
- *Tags_switched off*
- *Lead Quality_Not Sure*
- *Lead Quality_Worst*
- *Last Notable Activity_SMS Sent*