# Lead scoring case study

Participants

– Abhijeet Mishra

- Vismaya Murali

- Raghuveer Vempaty

-

# 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’. 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 of this assigment

there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom

We need to to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc. ) in order to get a higher lead conversion.

# Approach towards this assignment

We have been given data of leads which includes converted and non converted leads

Dataset consist of total 9240 rows and 37 columns

```
df_lead.shape
```

```
(9240, 37)
```

We understand most of the columns and categorical in nature and we have to do probability scoring hence chosen logistic regression to proceed forward

**to make a prediction about a categorical variable versus a continuous one**

We tried to understand data checking in case any duplicate values towards these leads

```
In [356]: #Checking for duplicay in data

In [357]: sum(df_lead.duplicated('Prospect ID'))
Out[357]: 0

In [358]: sum(df_lead.duplicated('Lead Number'))
Out[358]: 0
```

Then checked the null values at data

```
Country                                        26.634199
Specialization                                 36.580087
How did you hear about X Education             78.463203
What is your current occupation                29.112554
What matters most to you in choosing a course  29.318182
Search                                          0.000000
```

Dropped all the columns which is having high missing value – 35%

Most of the data having named select converted to null/nan values

Missing value imputations of numerical and categorical values

All categorical variables dummies value creating

Numerical values standardisation/scaling using MinMaxScaling

Features selections using RFE module

Manual features selections using Stats module checking P value and VIF score

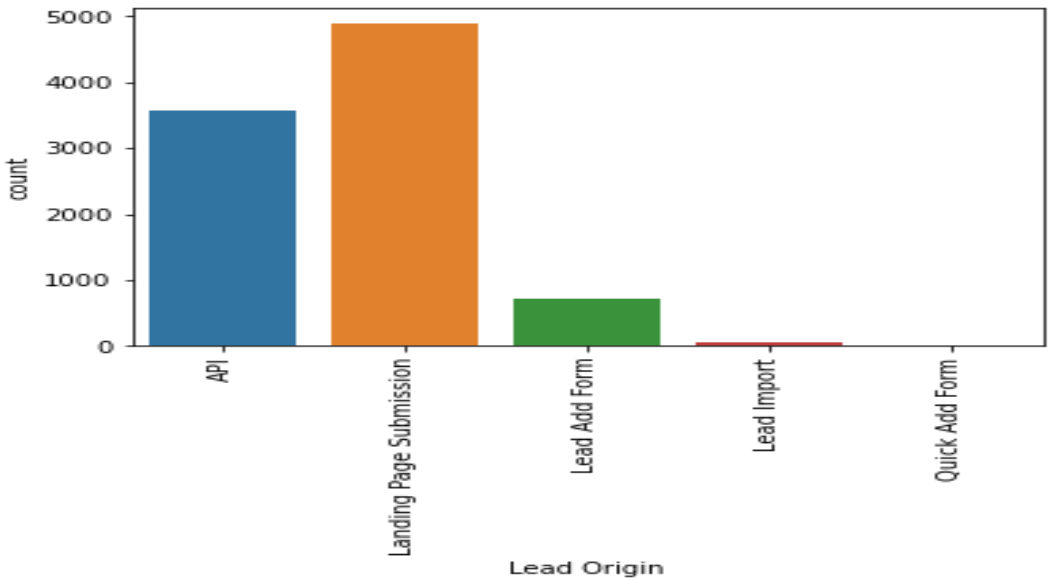
Evaluation of model using Confusion matrix

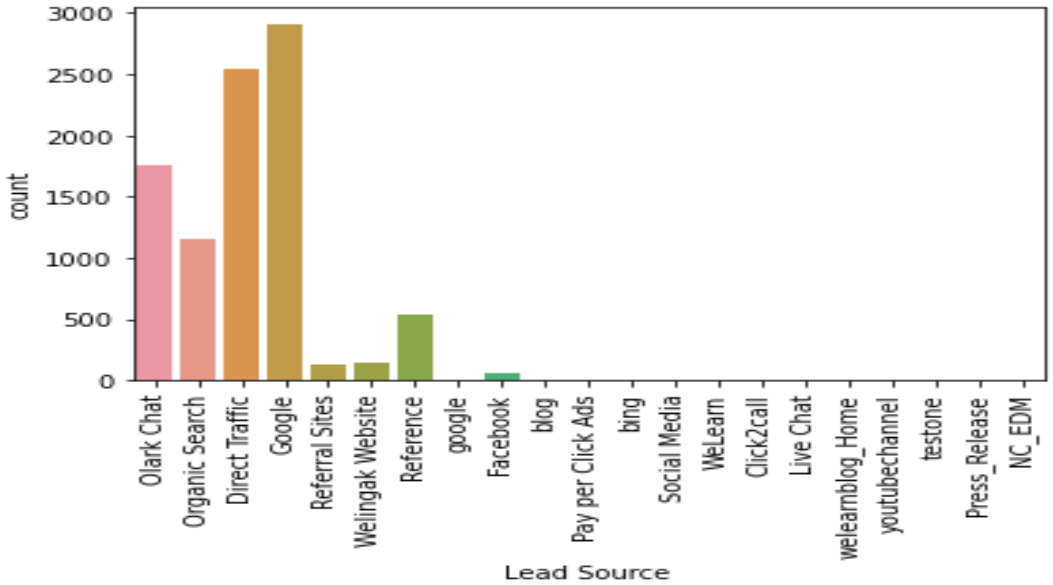
Checking Accuracy , Recall and Precession scores

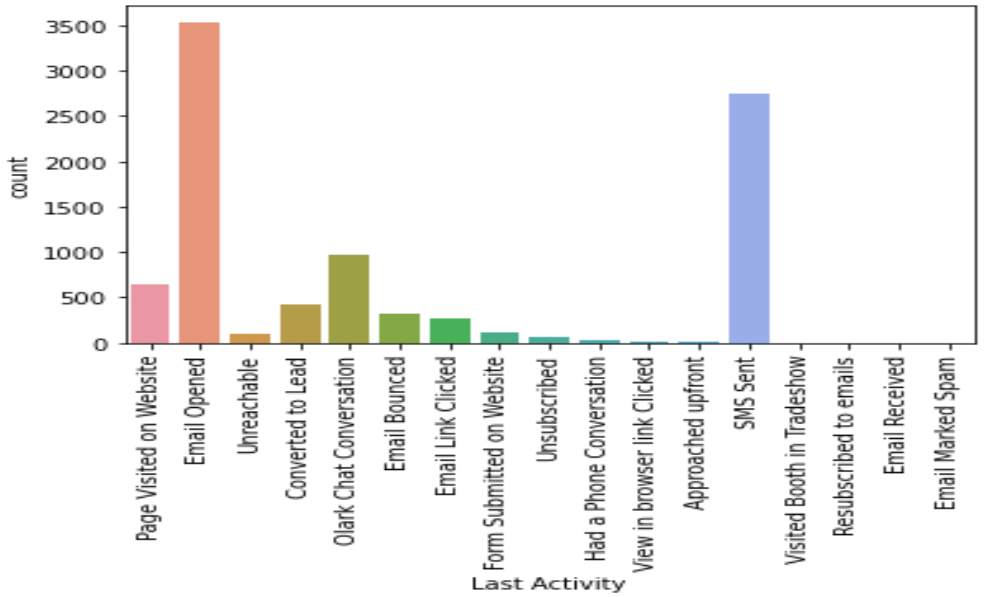
Checking ROC curve and finding optimal cutoff

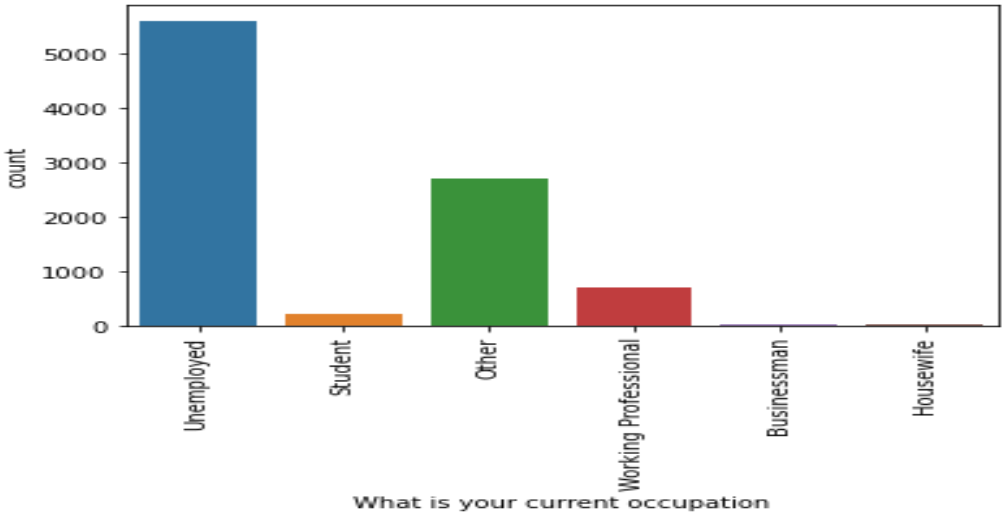
Making predictions of test data and comparing results with train data

# Univariate analysis

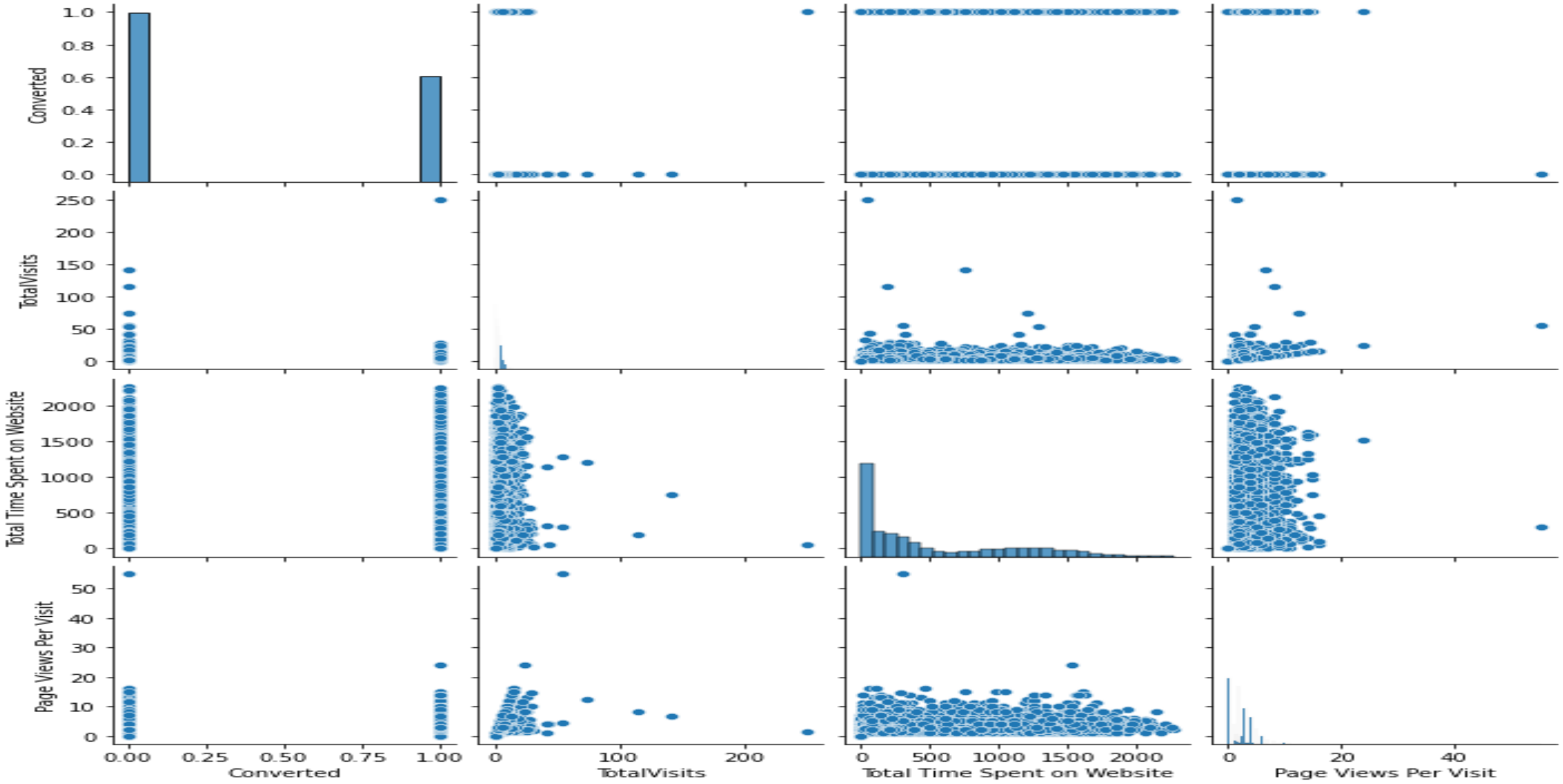


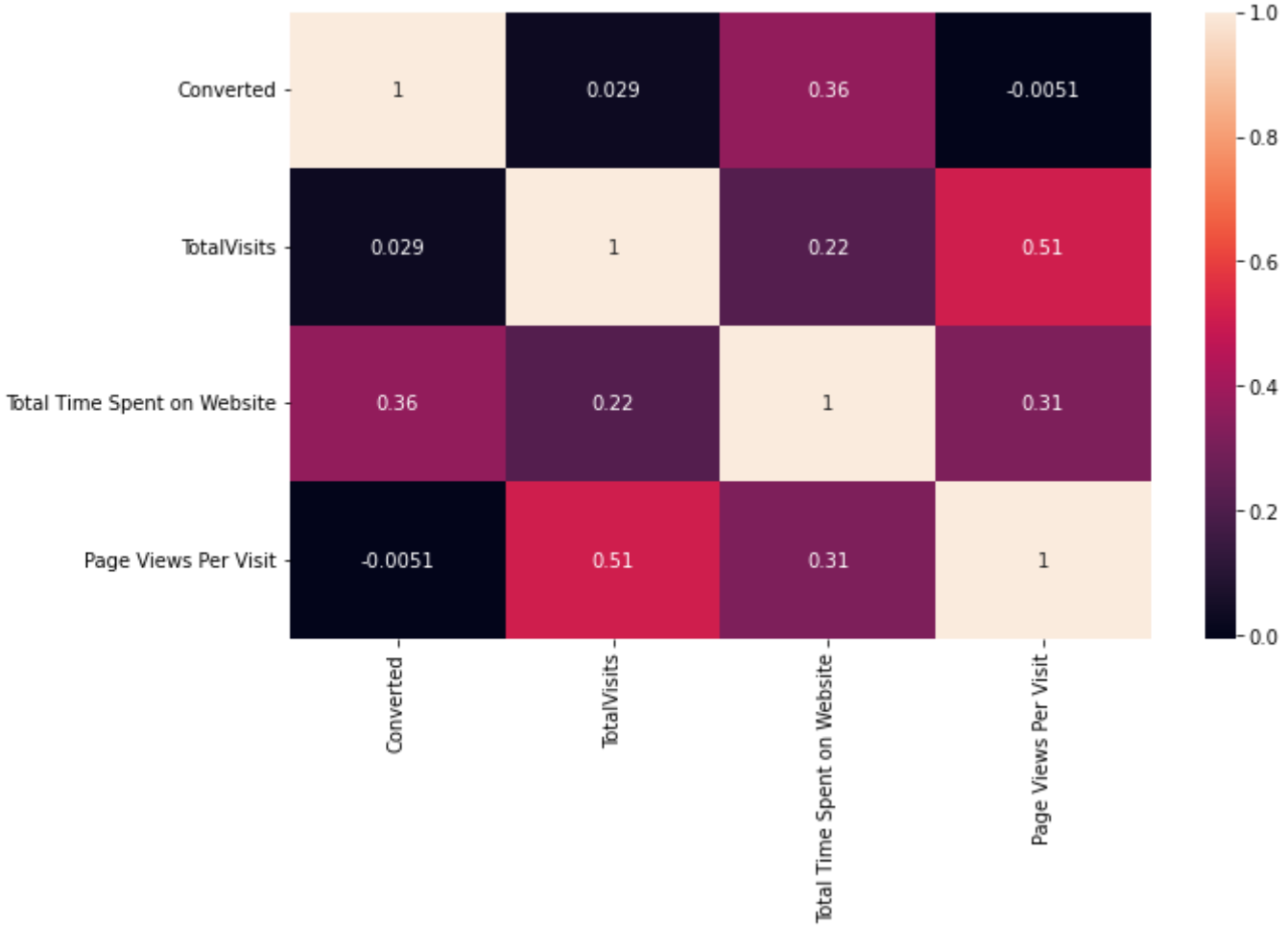



# Univariate analysis



# Heatmap


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.029
0.36
-0.0051
0.029
1
0.22
0.51
0.36
0.22
1
0.31
-0.0051
0.51
0.31
1
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0

**Below are the points made based on the observation during the analysis:**

The main target variables that have impacted the potential leads:

1.Google

2.Direct traffic

3.Organic Search

The last activity that affected the leads were:

1.Opened emails

2.SMS

3.Olark Chat Conversation

1.The total time spent on the website and the total number of visits had their share of impact

2.Working professions contributed to the lead

3.Final Model (res) res = logm4.fit()

4.The cut off probability is 0.35

5.More than 0.35 were converted as lead

6.Less than 0.35 will not be converted as lead