Lead Score Case study

Group members:

Rahul Pal

Ayush nayak

Problem Statement

- Education company named X Education sells online courses to industry professionals.
- Company got lots of leads but lead conversion is poor, only about 30%.
- To make this conversion higher company wants to identify potential leads.
- If company identify potential leads sales team will focus more on communicating with the potential leads rather than call to everyone.

Solution

- By using leads data company need to build logistic regression model so that by applying that model education company can achieve potential leads.

Solution Methodology

- Data cleaning and data manipulation.

1. Check duplicity in data and then remove it when required.
2. Check for NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Imputation of the values, if necessary.

- EDA

1. Check and handle outliers when it required.
2. Univariate data analysis: value count, distribution of variable etc.
3. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

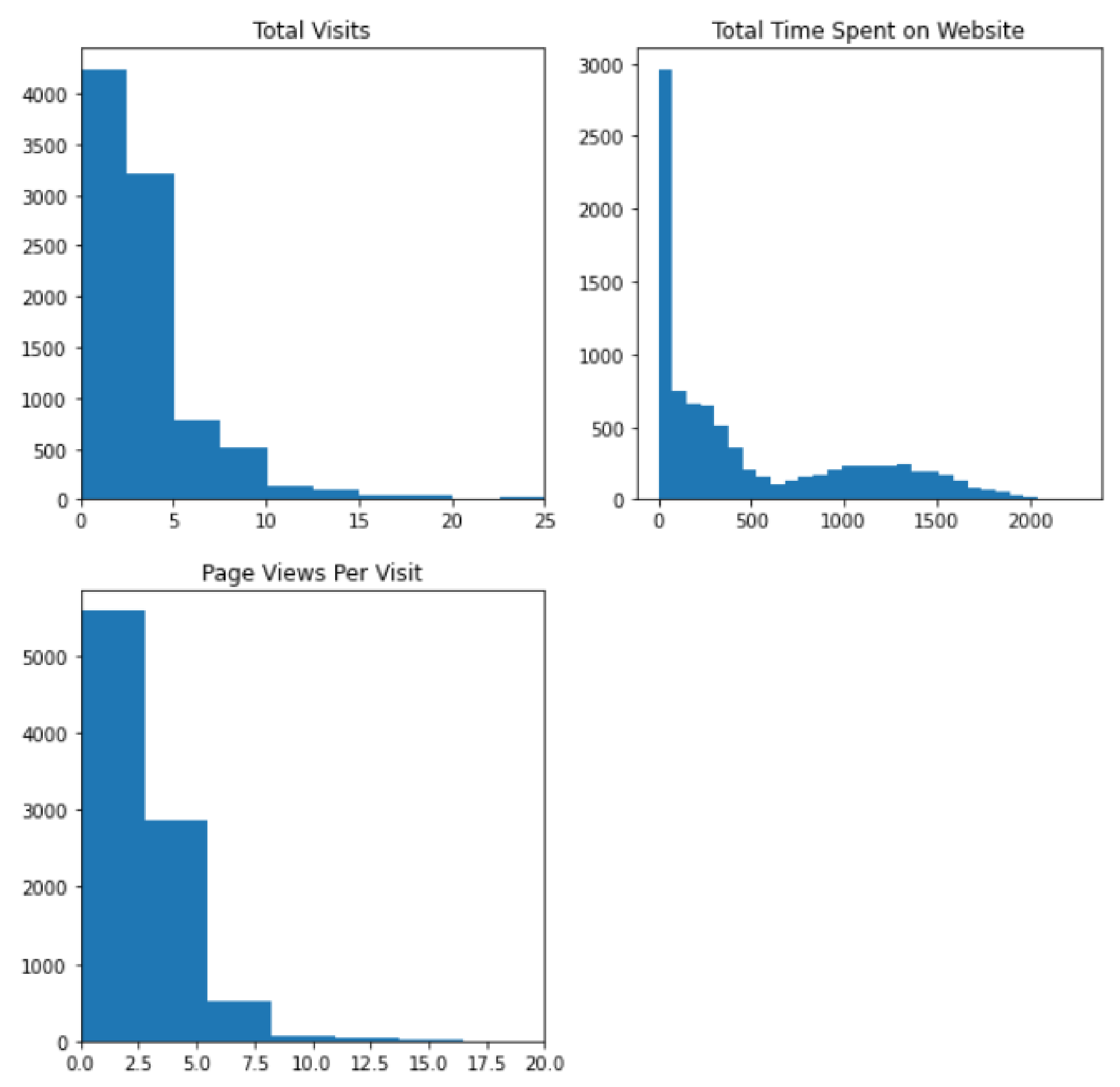
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

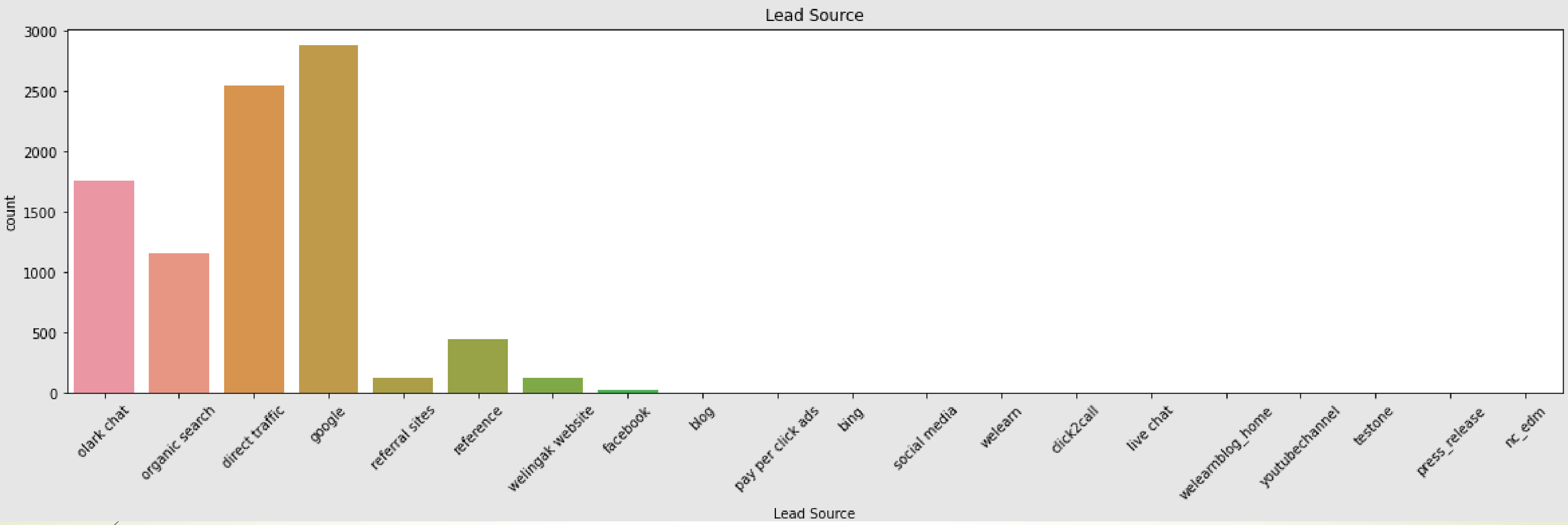
Data cleaning and manipulation

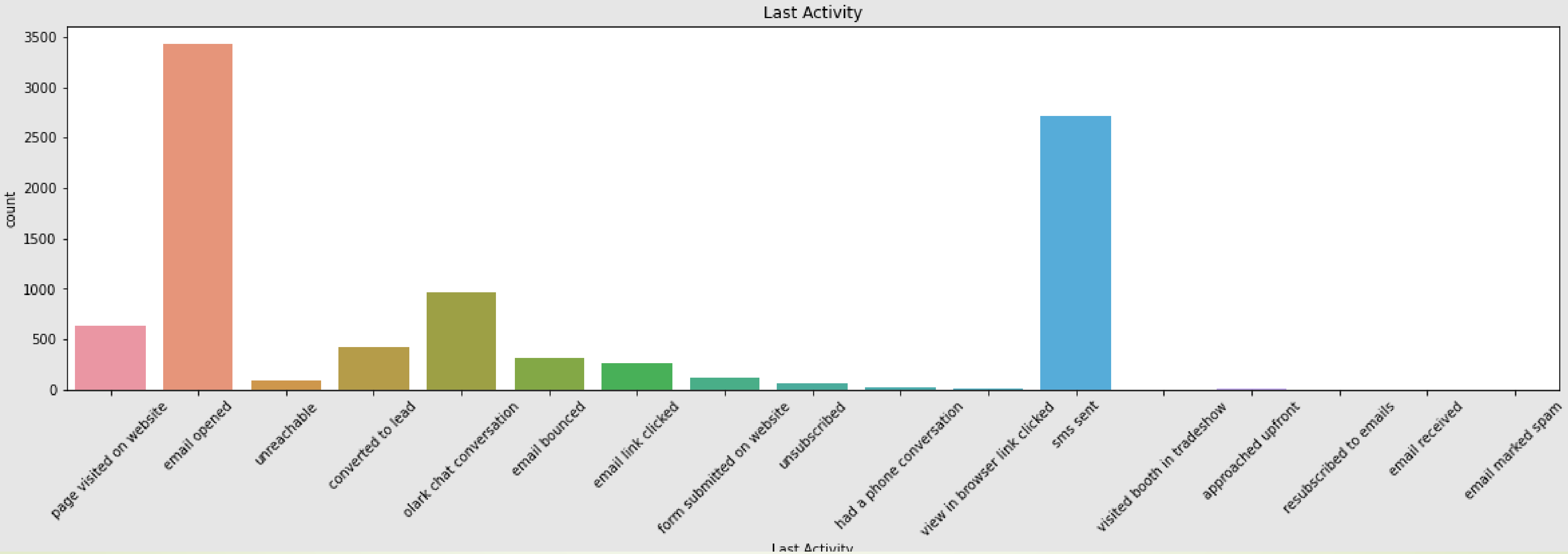
- We have 9240 rows and 37 columns.
- Drop those variable which contains unique values ('Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content','I agree to pay the amount through cheque‘)
- Drop columns('City', ‘Lead Number', 'Prospect ID‘) as they are not necessary for analysis.
- Dropping the columns ('Tags', 'Lead Quality', 'Asymmetrique Profile Index', 'Asymmetrique Activity Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score', 'Lead Profile', 'How did you hear about X Education') having more than 35% as missing.

EDA

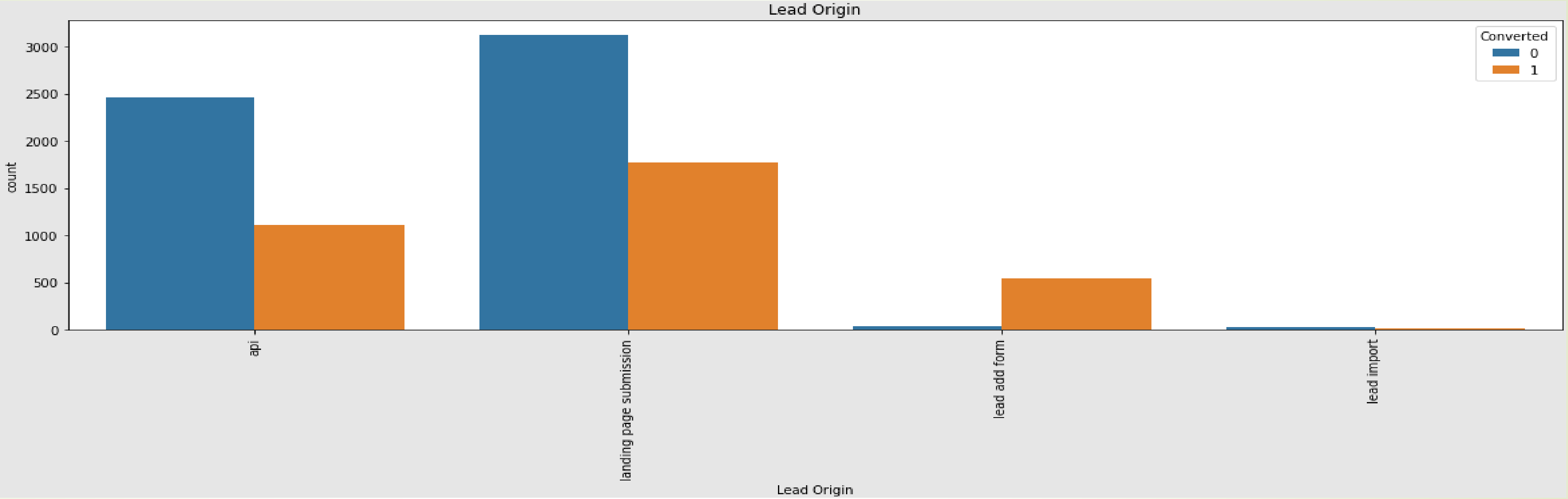
Univariate analysis

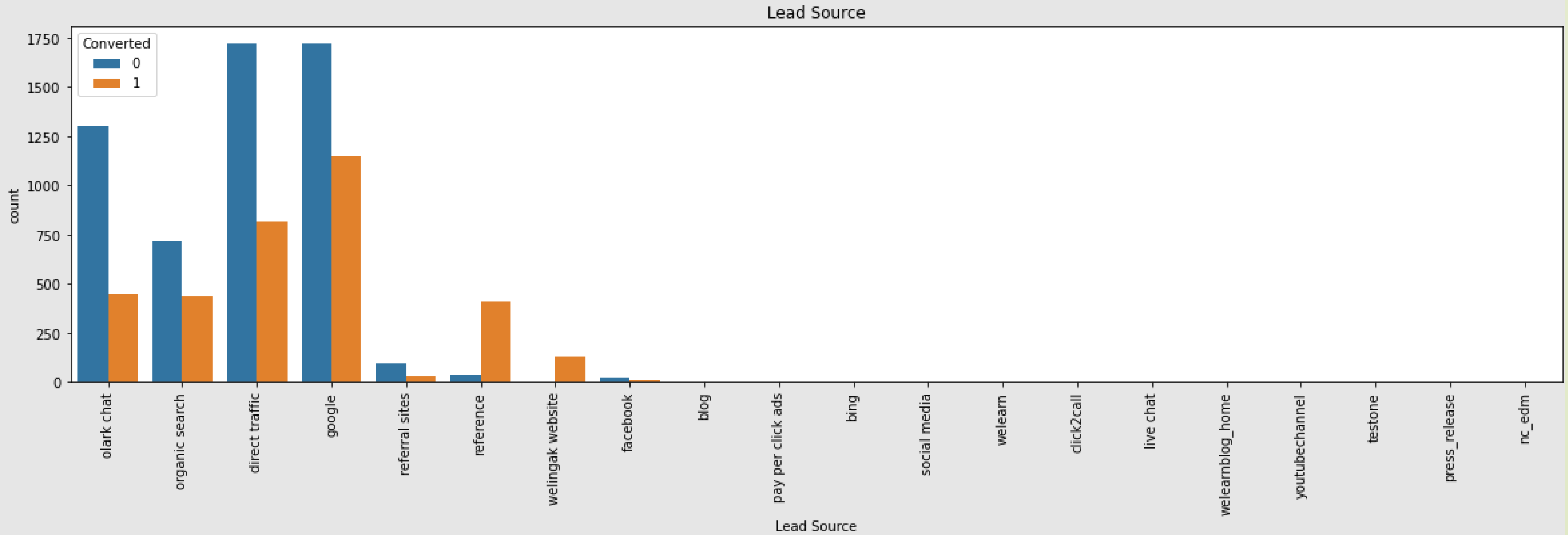


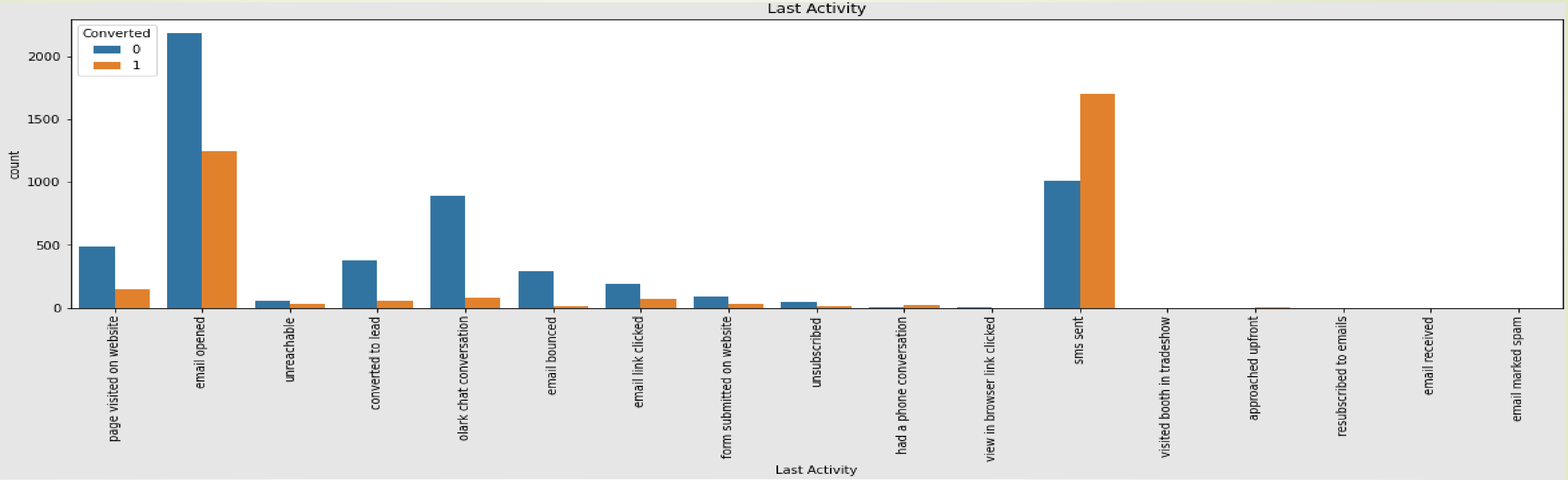

Lead Source
count
3000
2500
2000
1500
1000
500
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

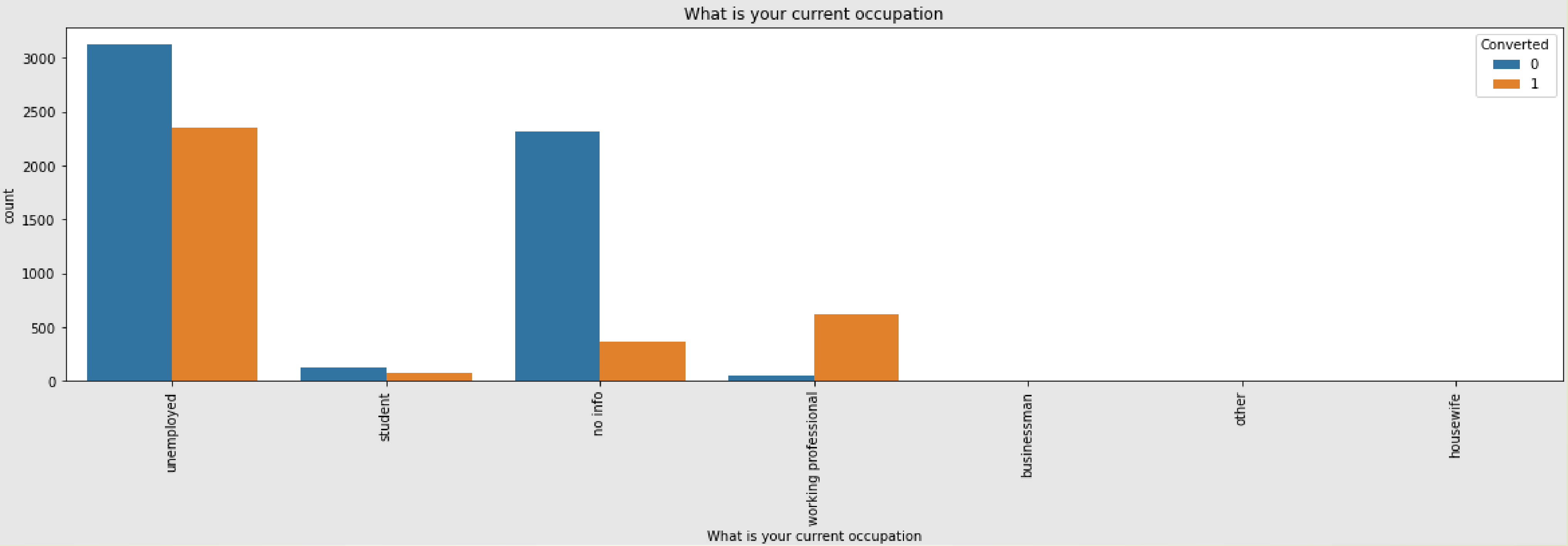

Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

Bivariate analysis






Last Activity
Converted
0
1
count
0
500
1000
1500
2000
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

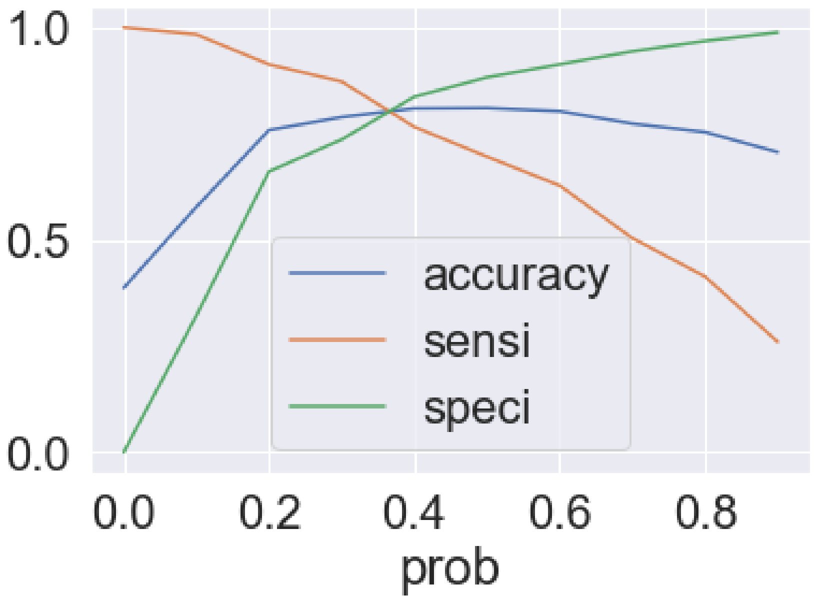

What is your current occupation
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
unemployed
student
no info
working professional
businessman
other
housewife
What is your current occupation

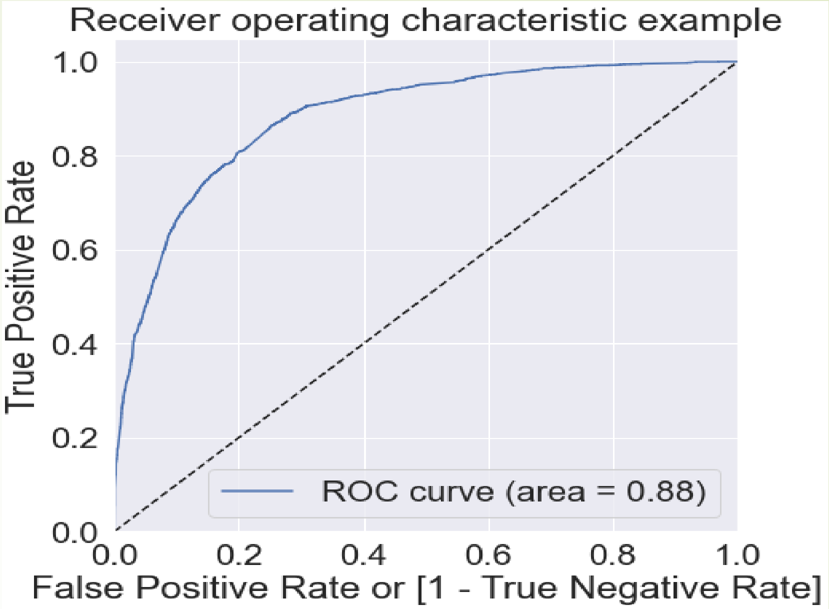
Data Conversion

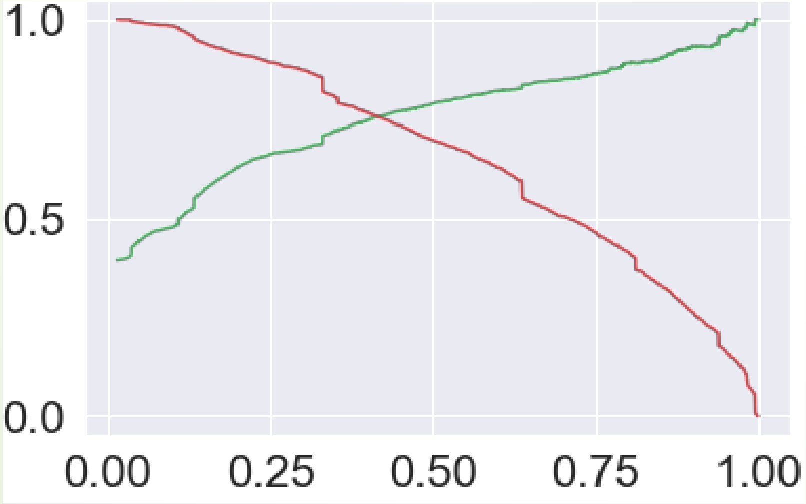
- Numerical variables are Normalised.
- Dummy variables are created for object type variables.
- We have 9074 rows and 81 columns for analysis.

Model Building

- Splitting data in train and test set, 70% train and 30% test.
- Use RFE for feature selection.
- Running RFE with 15 variables as output.
- Building Model by removing the variable whose p- value>0.05 and VIF >5.
- Predictions on test data set.
- Overall accuracy attain on cut off value 0.35 is 80.7%.







- We have balanced sensitivity and specificity at optimal cut off value 0.35.
- We have balanced precision and recall value at cut off value 0.41.

Conclusion

The variables that mattered the most for the potential buyers are:

1.) Total time spend on website.

2.) Total number of visits.

3.) When lead source was:

a) Welingak website

b) Olark chat

c) Organic research

d) Google

4.) When last activity was:

a) SMS sent

b) Olark chat conversation

5.) When the lead origin is ‘lead add form’.

6.) When their current occupation is working professional.

By keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.