LEAD SCORING CASE STUDY

P ANAND RAO

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- 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.

Business Objective

X education wants to know most promising leads.

For that they want to build a Model which identifies the hot leads.

Deployment of the model for the future use.

Solution Methodology

- Data cleaning and data manipulation.
  a. Check and handle duplicate data.
  b. Check and handle NA values and missing values.
  c. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  d. Imputation of the values, if necessary.
  e. Check and handle outliers in data.
- EDA
  a. Univariate data analysis: value count, distribution of variable etc.
  b. 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 Manipulation

Total Number of Rows = 37, Total Number of Columns = 9240.

Single value features like “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” etc. have been dropped.

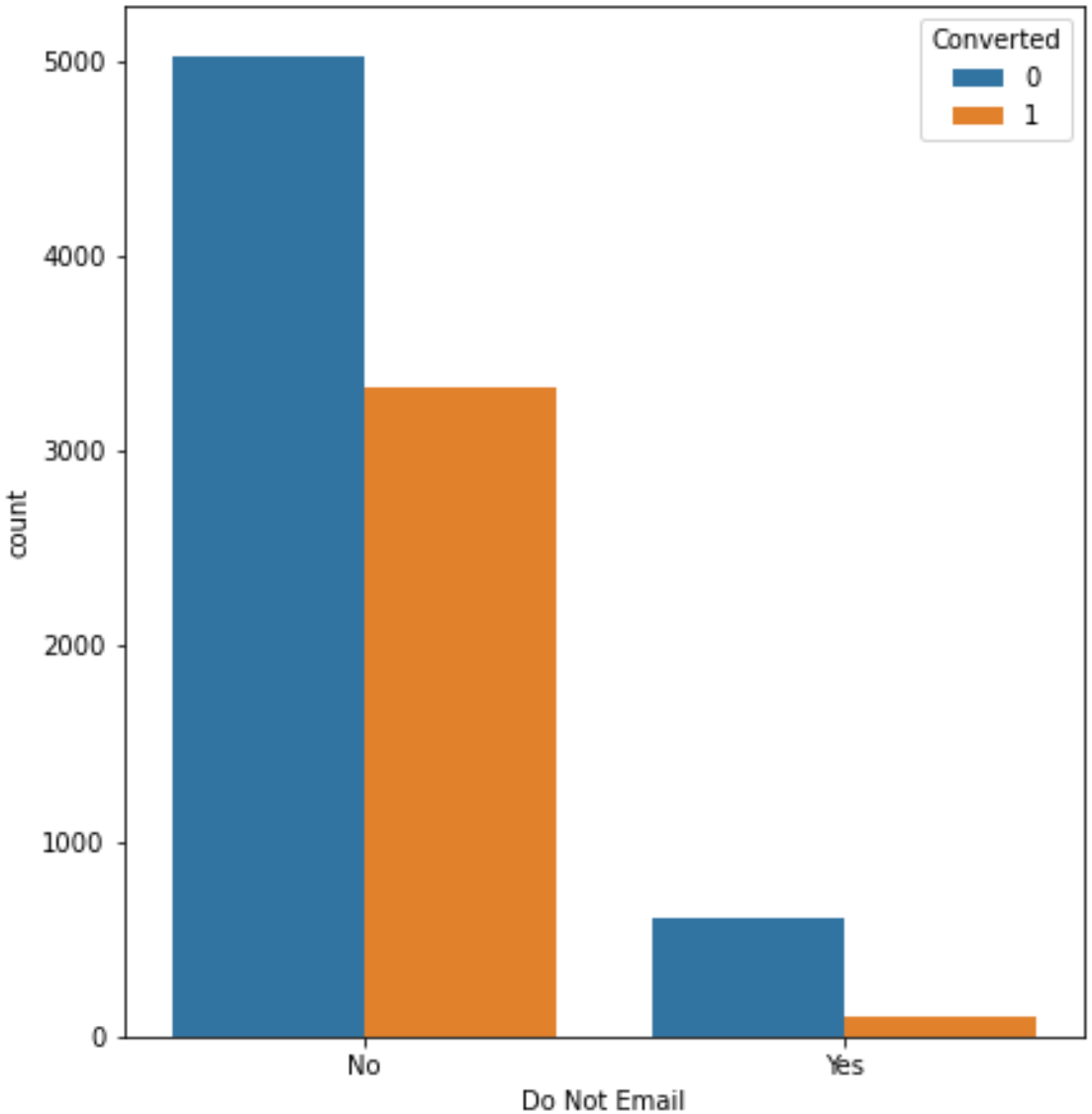
Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.

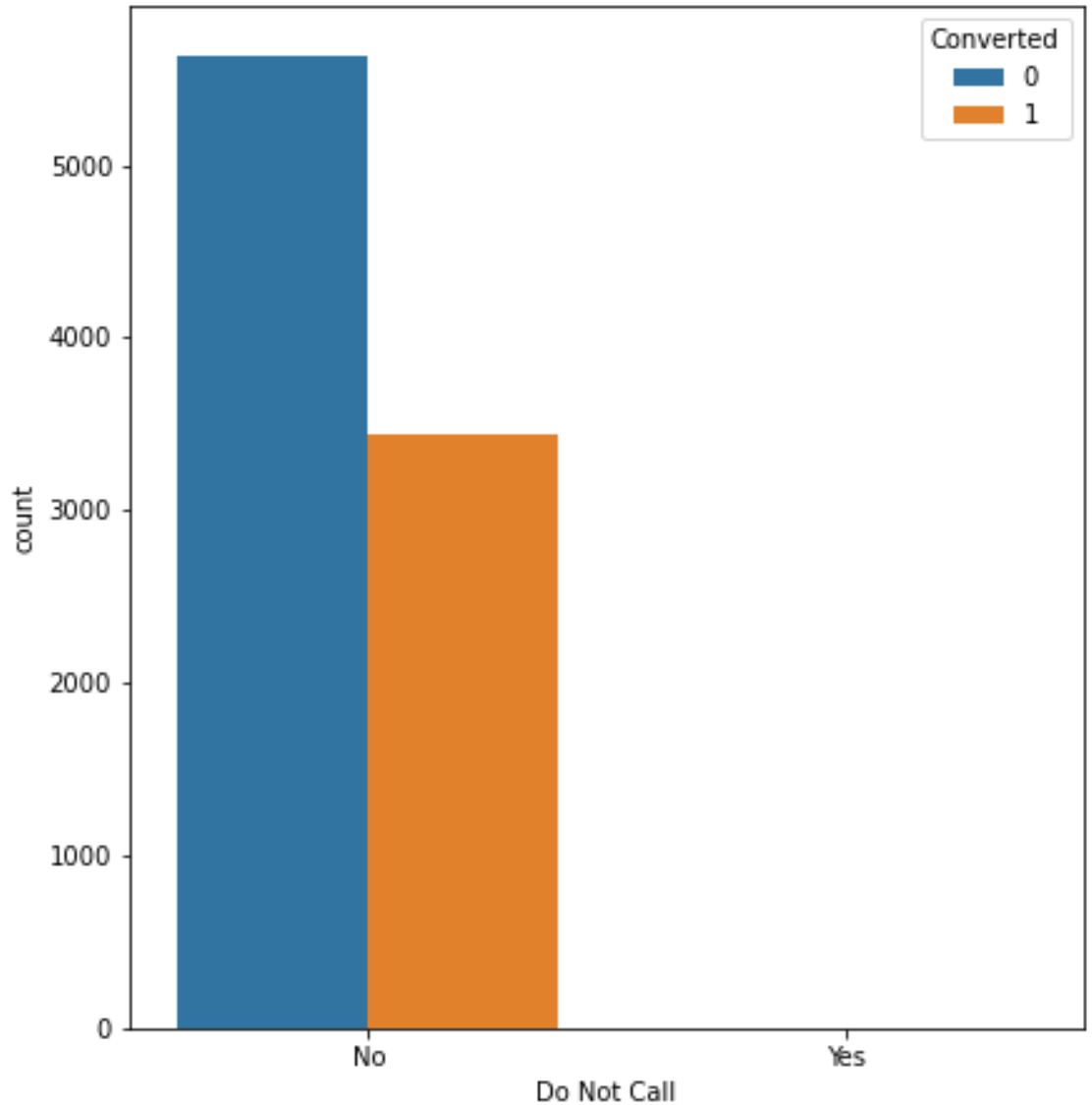
After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

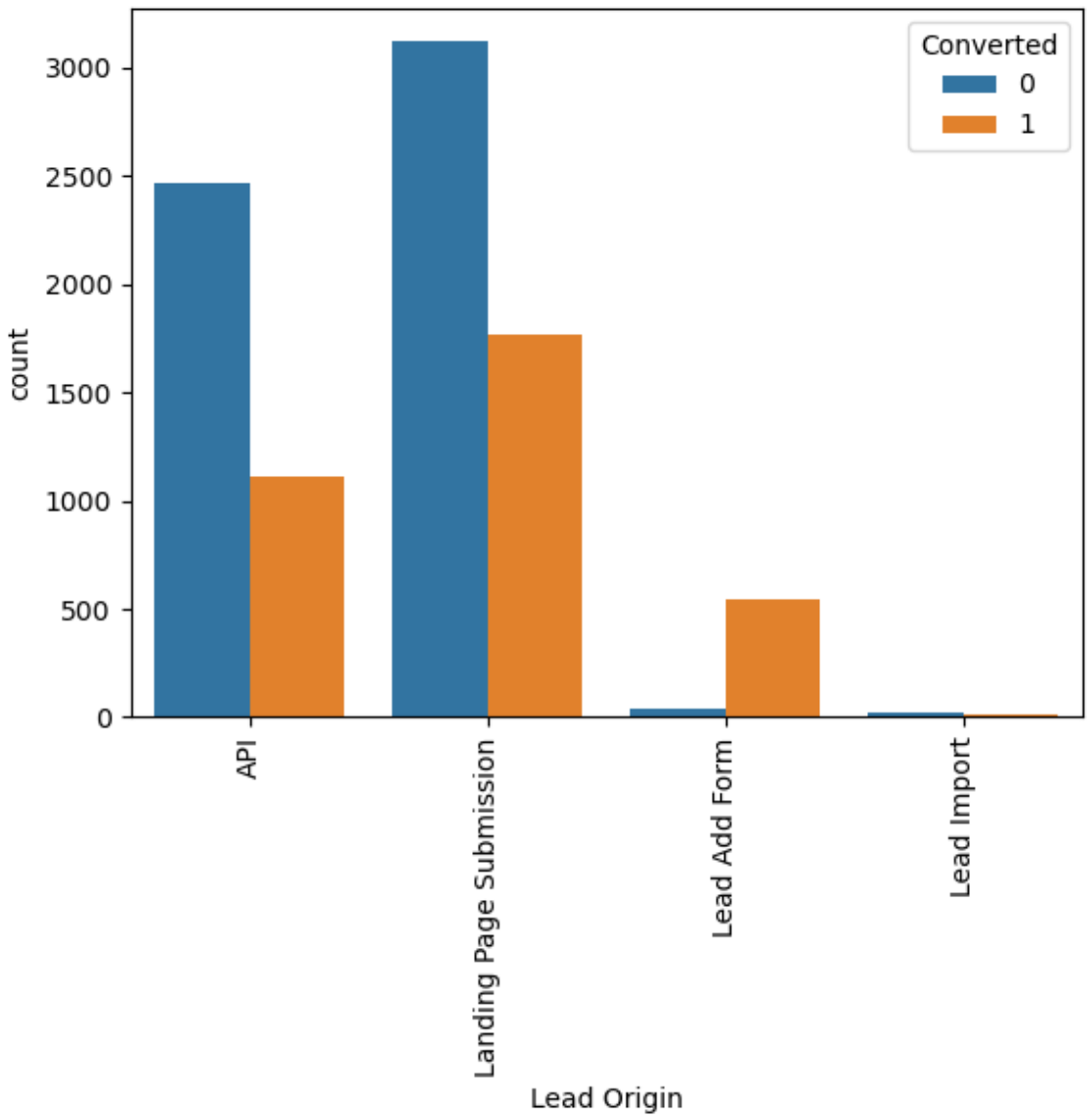
Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

EDA

EXPLORATORY DATA ANALYSIS


Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Email


Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Call

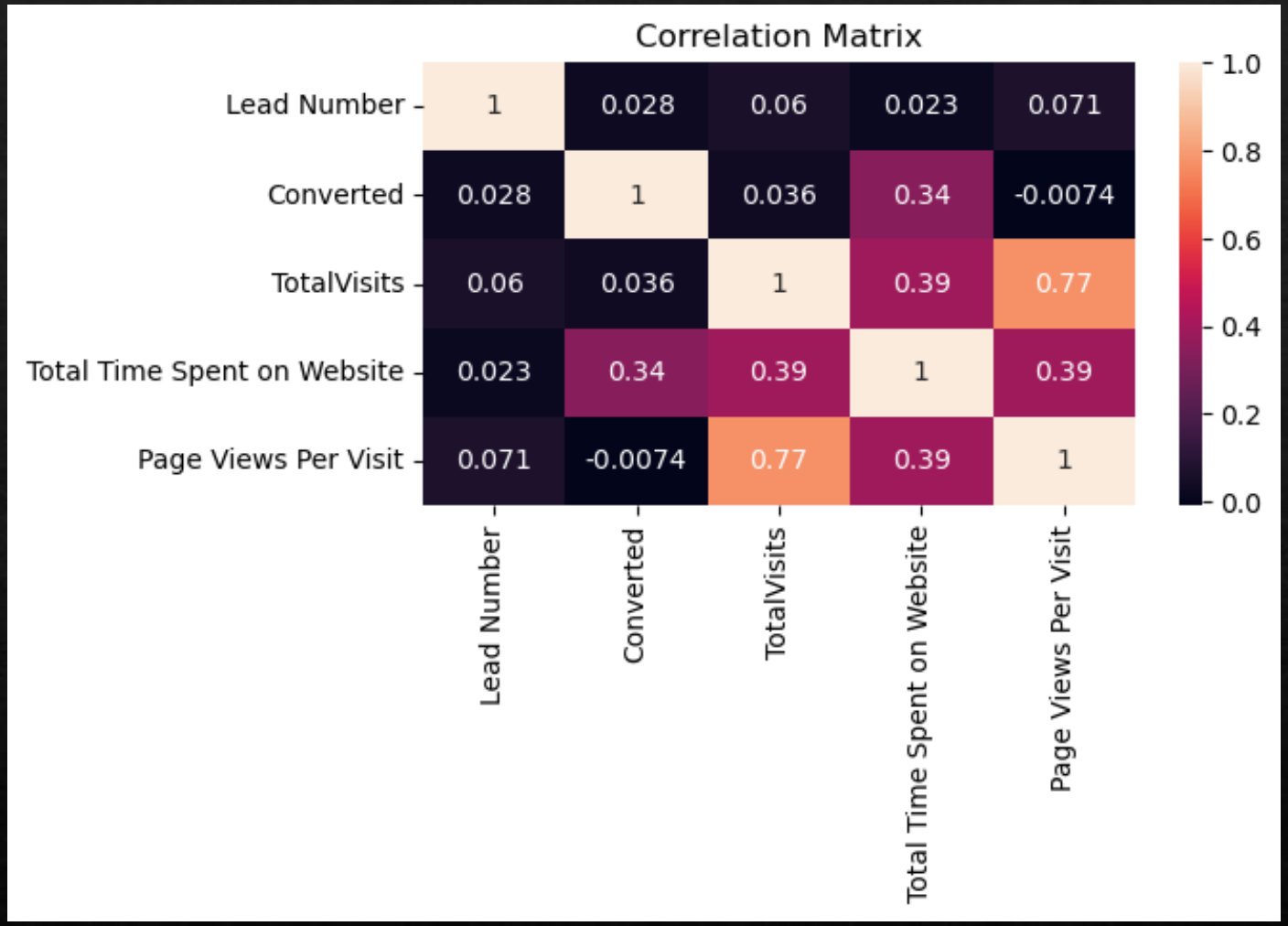


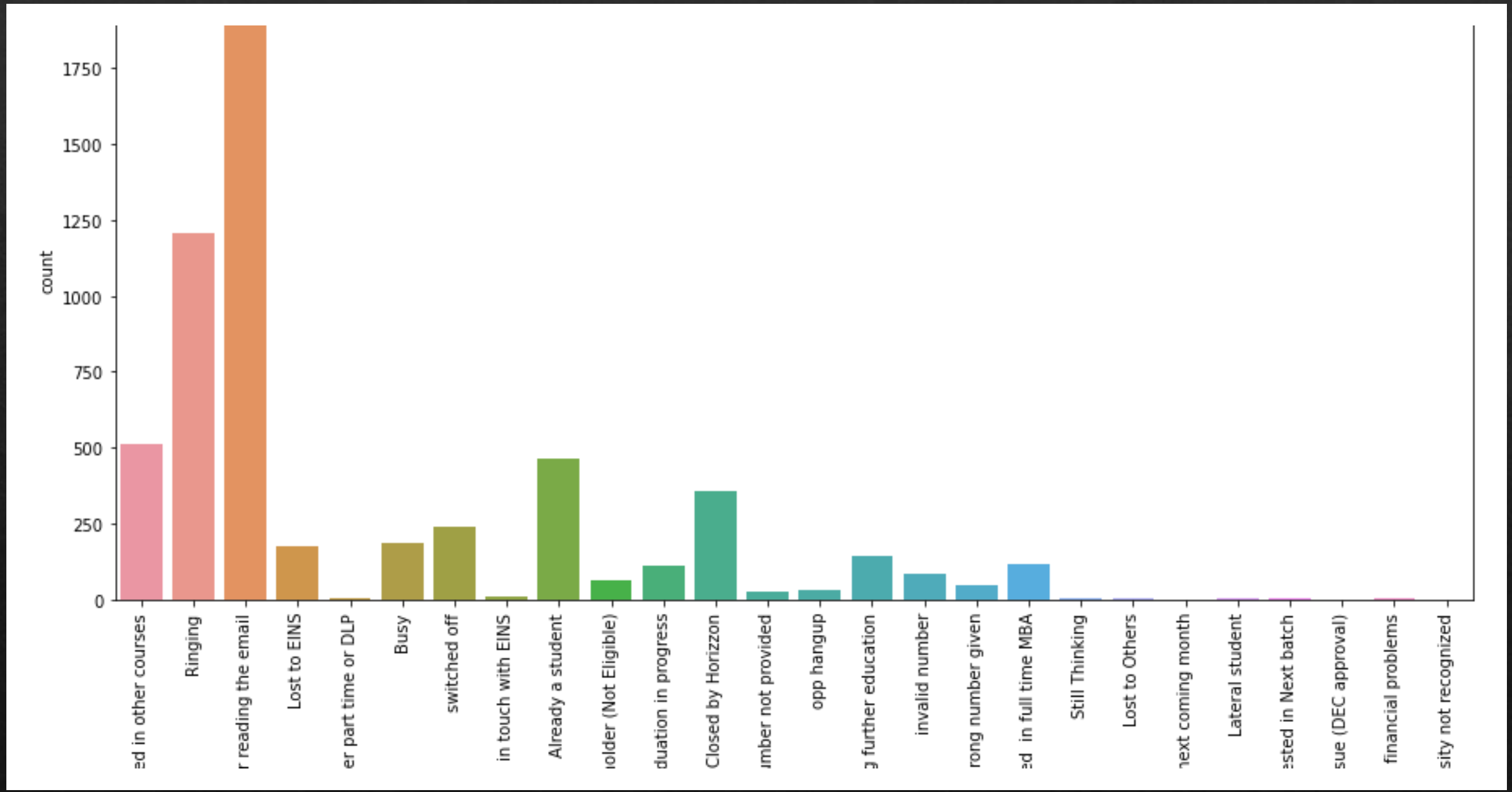
INFERENCES

- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

This correlation matrix displays high correlation between 'Converted' and

- Total visits
- Page views per visit
- Time spend on website

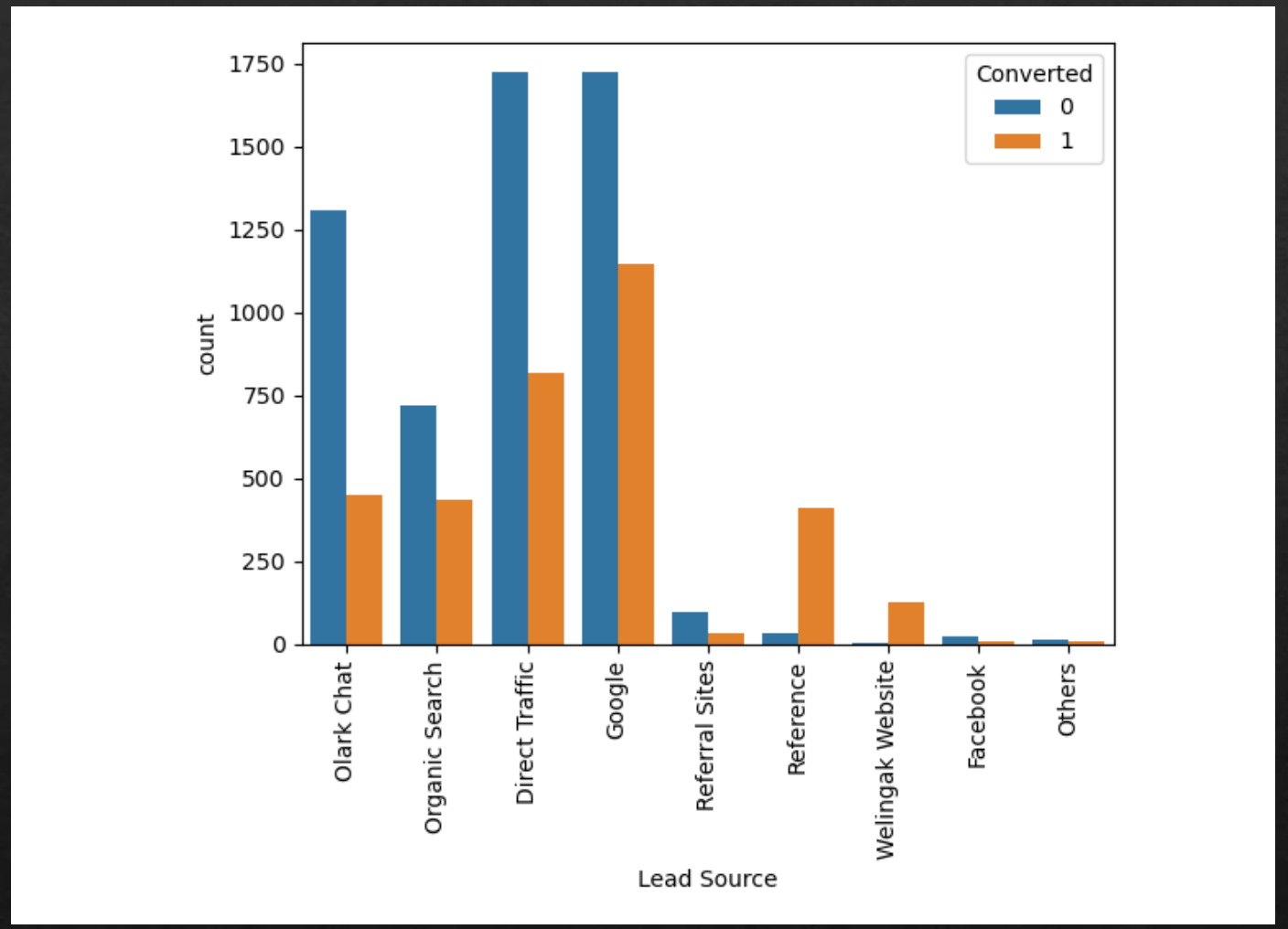


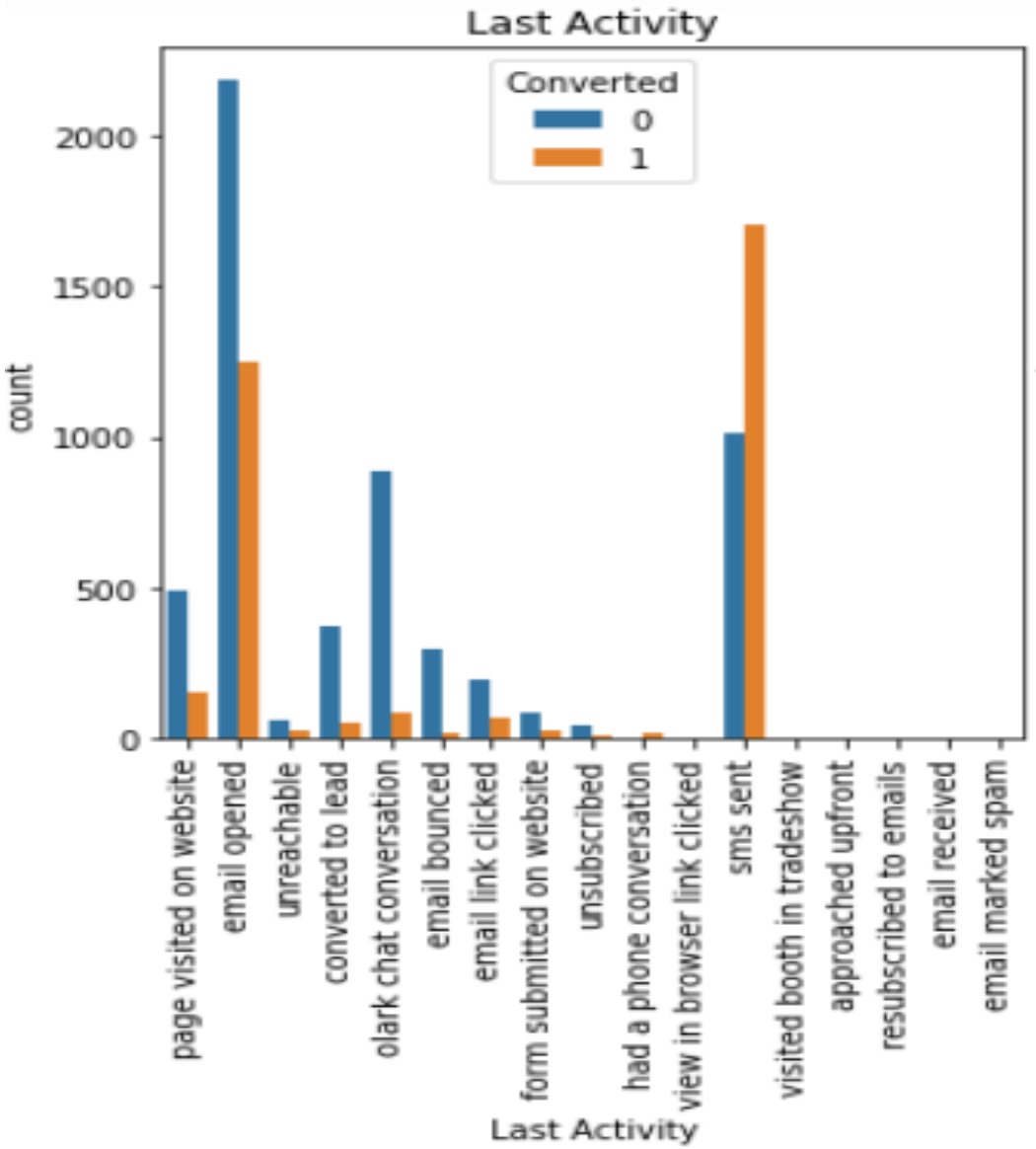

count
1750
1500
1250
1000
750
500
250
0
ed in other courses
Ringing
r reading the email
Lost to EINS
er part time or DLP
Busy
switched off
in touch with EINS
Already a student
nolder (Not Eligible)
duation in progress
Closed by Horizzon
umber not provided
opp hangup
g further education
invalid number
rong number given
ed in full time MBA
Still Thinking
Lost to Others
next coming month
Lateral student
ested in Next batch
sue (DEC approval)
financial problems
sity not recognized

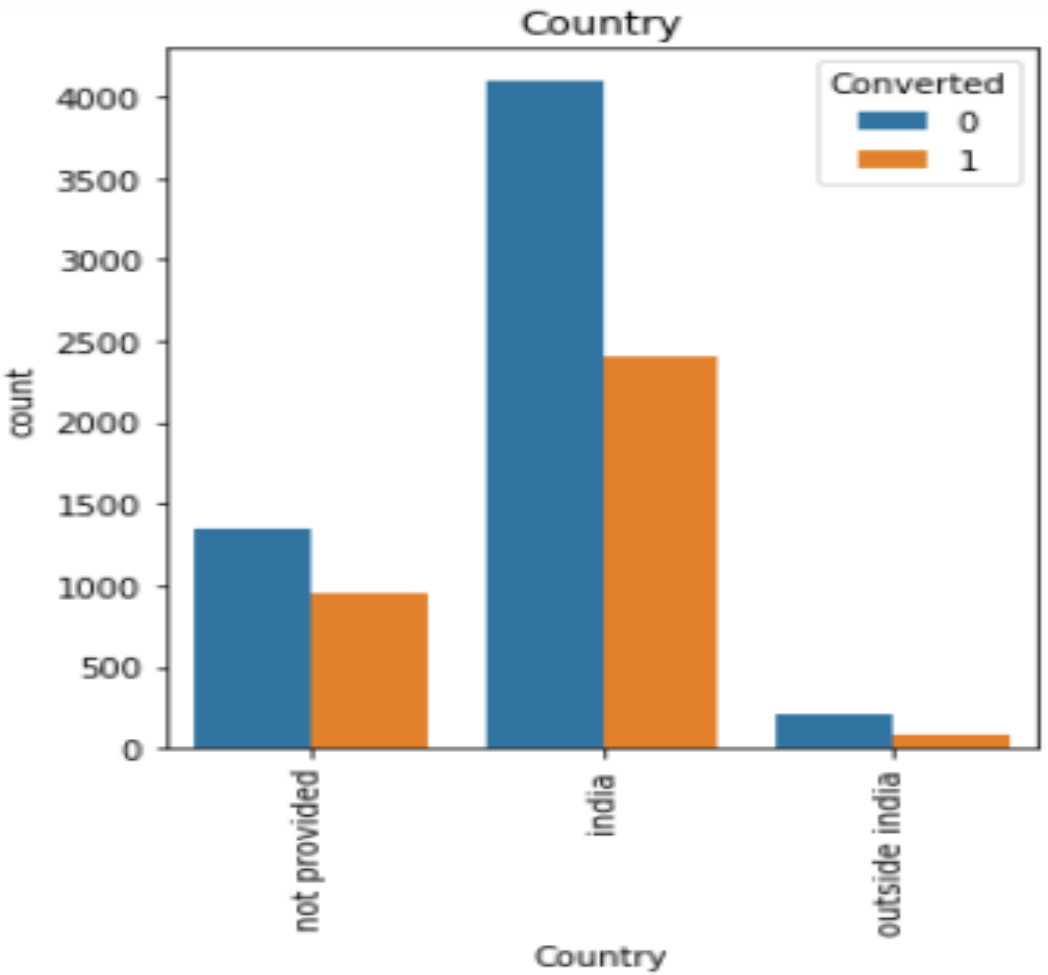
Maximum number of leads are generated by Google and Direct traffic.

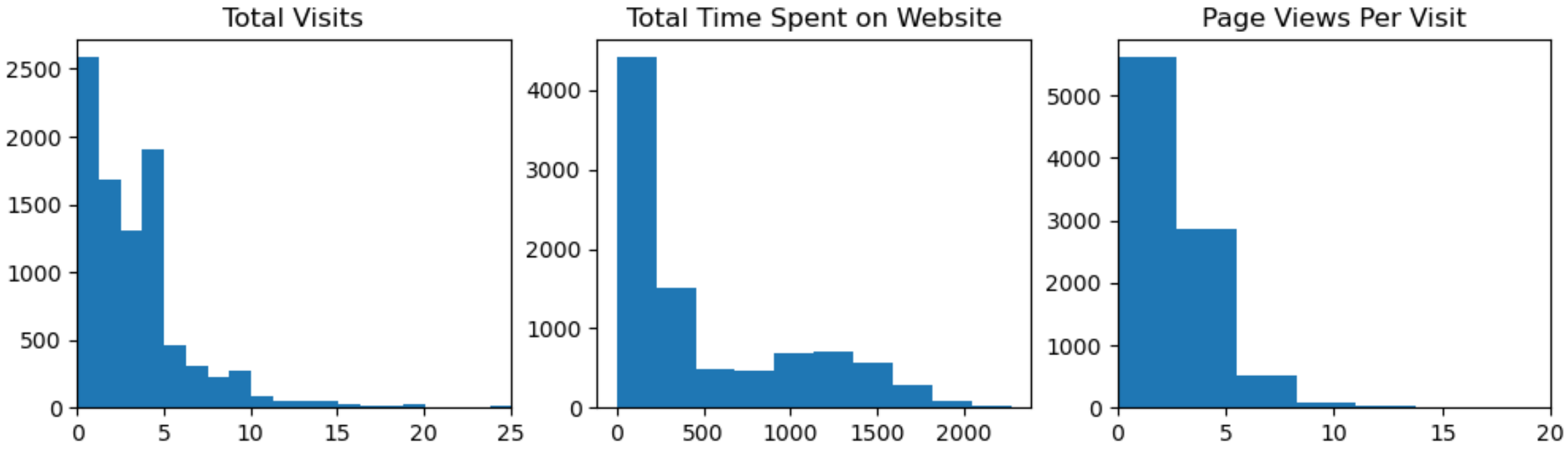
Conversion Rate of reference leads and leads through welingak website is high.

To improve overall lead conversion rate, focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.




Last Activity
Converted
0
1
count
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


Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

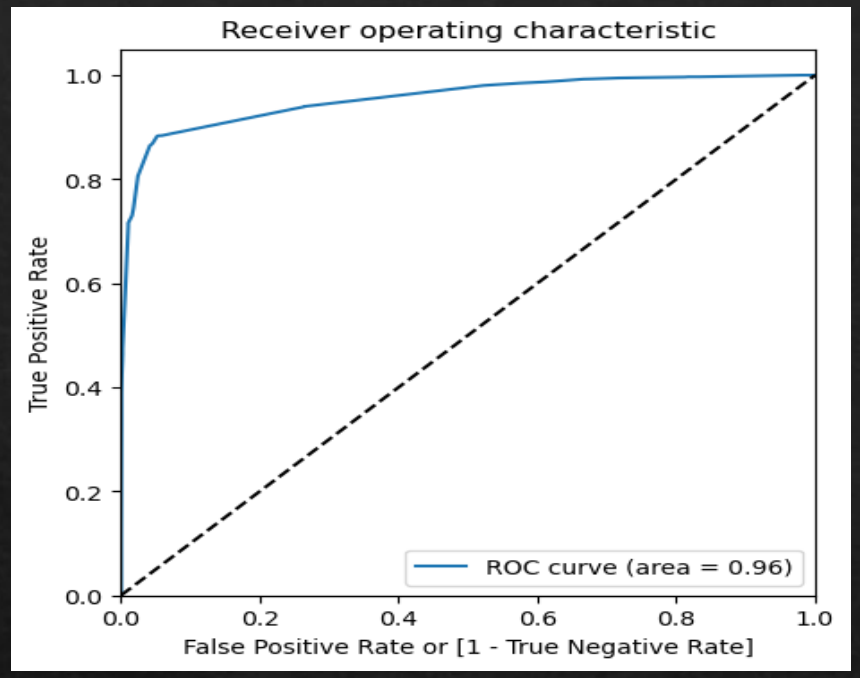

Total Visits
2500
2000
1500
1000
500
0
0
5
10
15
20
25
Total Time Spent on Website
4000
3000
2000
1000
0
0
500
1000
1500
2000
Page Views Per Visit
5000
4000
3000
2000
1000
0
0
5
10
15
20

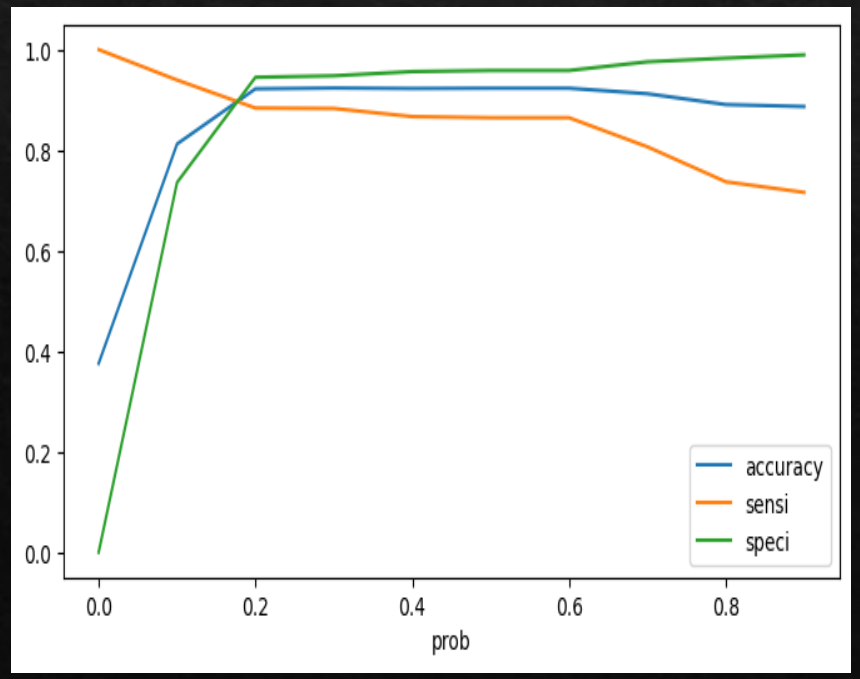
Data Conversion

- Numerical Variables are Normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis after dummy variable creation: 8474
- Total Columns for Analysis after dummy variable creation: 79

Model Building

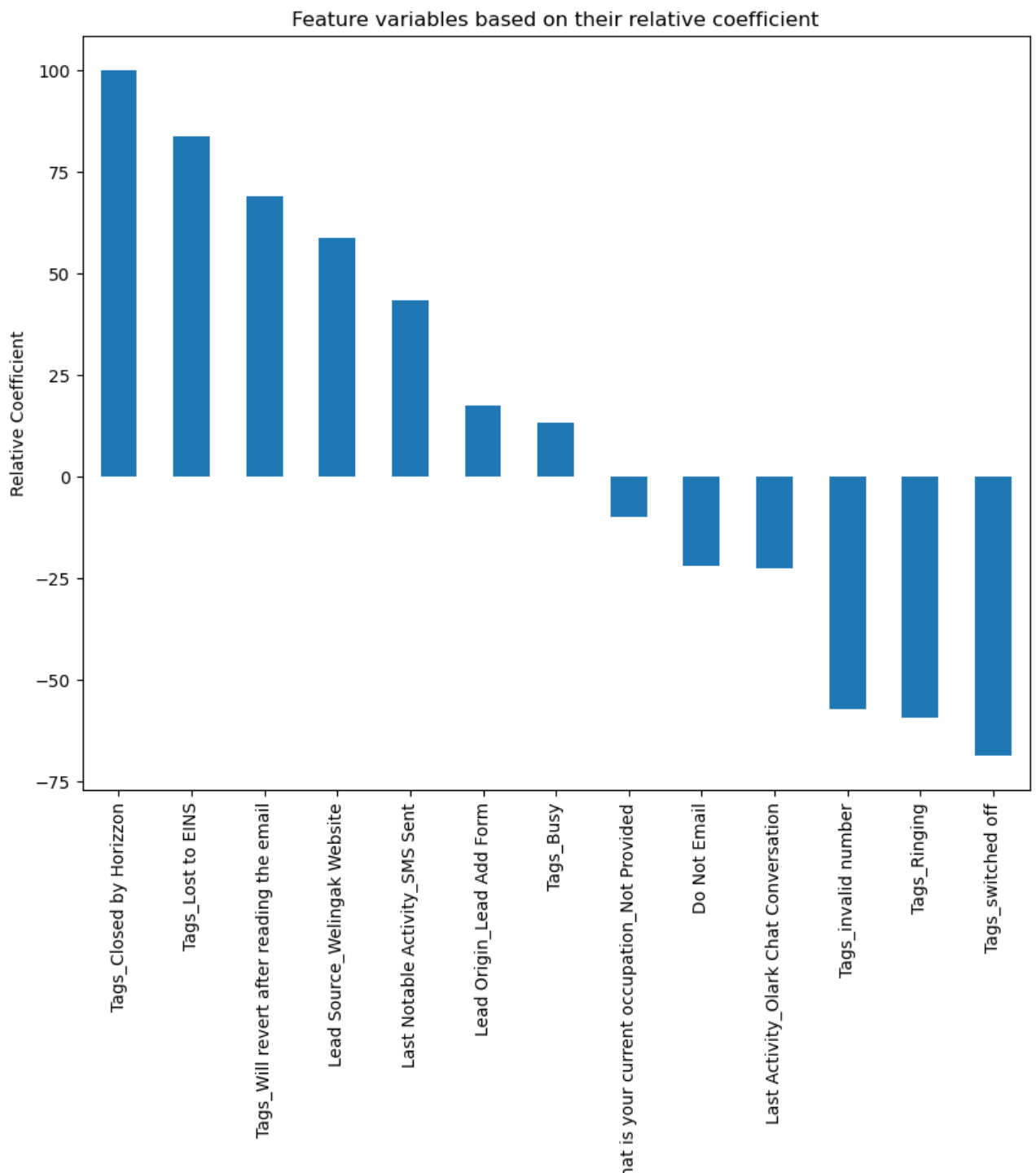
- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5.
- Predictions on test data set.
- Overall accuracy 92%.





ROC Curve

- Finding Optimal Cut off Point
- Optimal cut off probability is that Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.2



IMPORTANT FEATUTRES

Variables that have high lead conversion Probability

- Tags – Closed by Horizon, Lost to EINS, Will
- revert after reading the email
- Lead Source – Welingak Website
- Last Notable Activity – SMS Sent

Variables that need improvement in converting Quality Lead

- Tags – Invalid Number
- Tags – Ringing
- Tags – Switched off
- Last activity olark chat conversation

CONCLUSION

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welniak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional. 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.

RECOMMENDATION

- Focus more on customers who spend a lot of time on the company's website as their conversion rate is high as per EDA.
- Providing special offers to customers who are highly interested and are seen visiting back to the website.
- Leads who have Tags such as 'Ringing', 'Switched Off', 'Invalid Number' can be avoided as the probability of them converting is very low.
- Follow up through calls and emails with high conversion probability leads is suggested.

THANK YOU