upGrad

CVRGU | DS | JRCP

UPGRAD Capstone Project

Name:- Vaibhav Kumar

Email:- 20010684@cgu-odisha.ac.in

upGrad
CVRGU | DS | JRCP

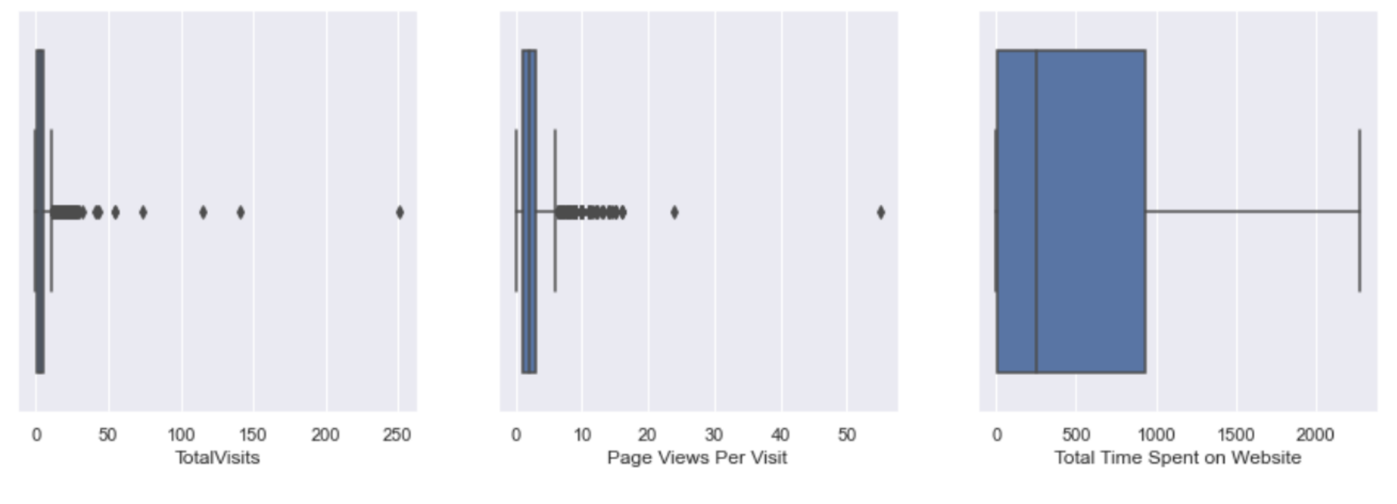
Problem Statement

- 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, search engines, and even social media sometimes.
- Once these people land on the website, they might browse the courses, fill out a form for the course, or watch some videos. When these people fill out a form with their email address or phone number, they are classified as leads.
- Now, although X Education gets a lot of leads, its lead-to-sale conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

upGrad
CVRGU | DS | JRCP

About The Data

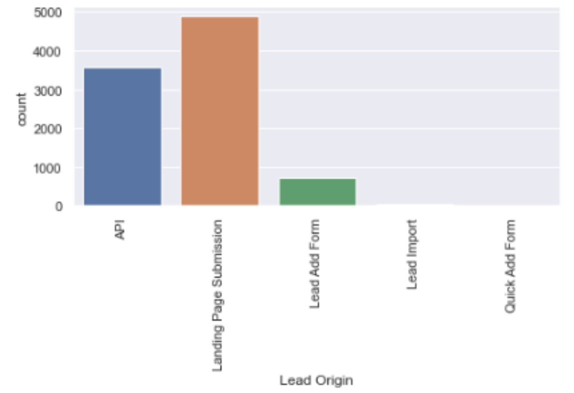
- The data set was partially clean except for a few null values and the option ‘Select’ has to replace with a null value since it did not give us much information.
- Dropped the high percentage of Null values more than 40%.
- Identified the Highly skewed columns and dropped them.
- Detected the Outliers

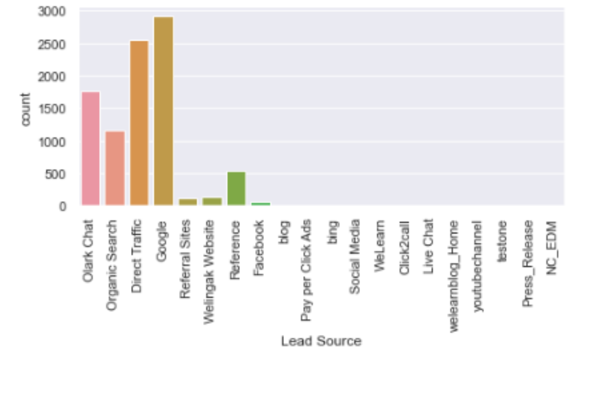


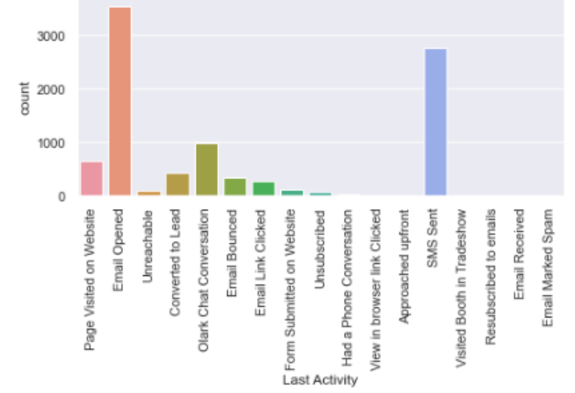
upGrad

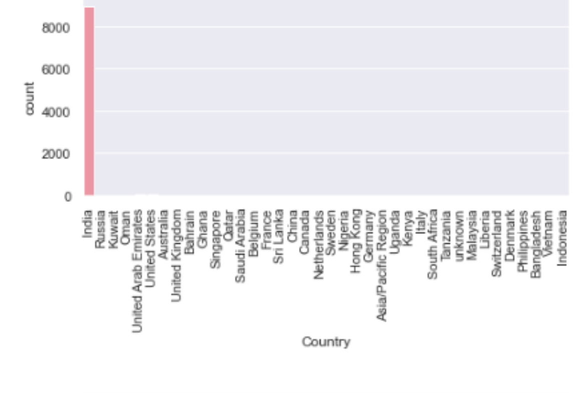
CVRGU | DS | JRCP

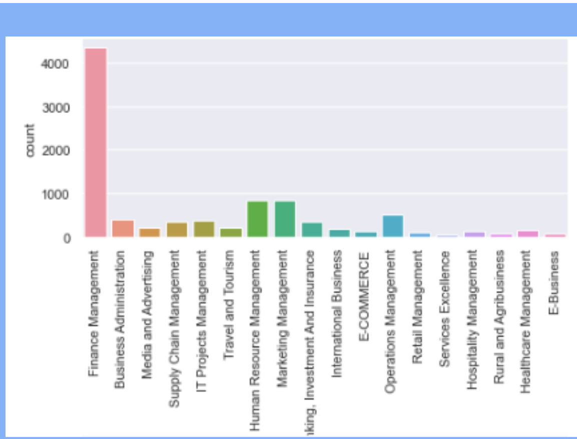
Data Distribution

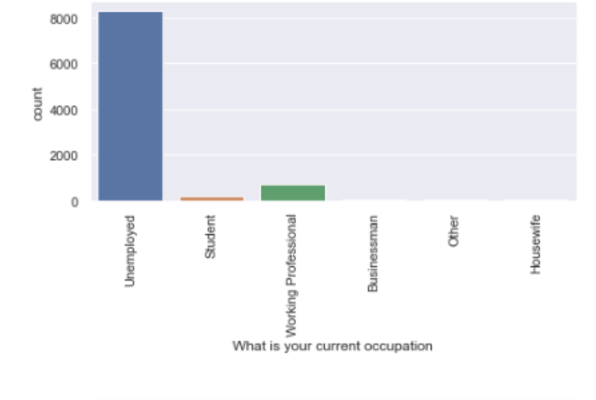


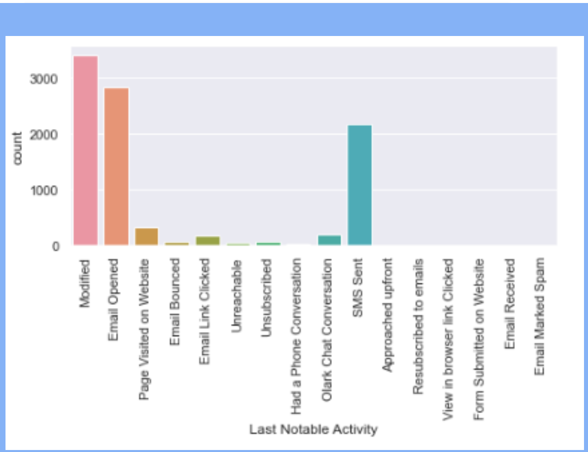










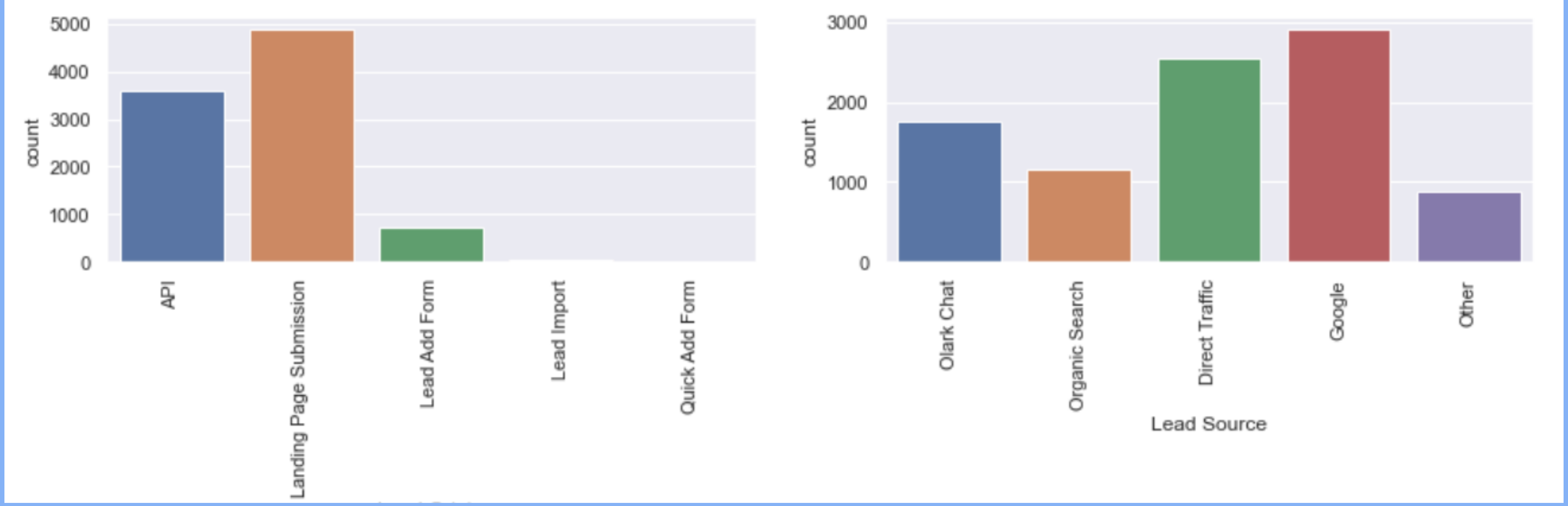


upGrad

CVRGU | DS | JRCP

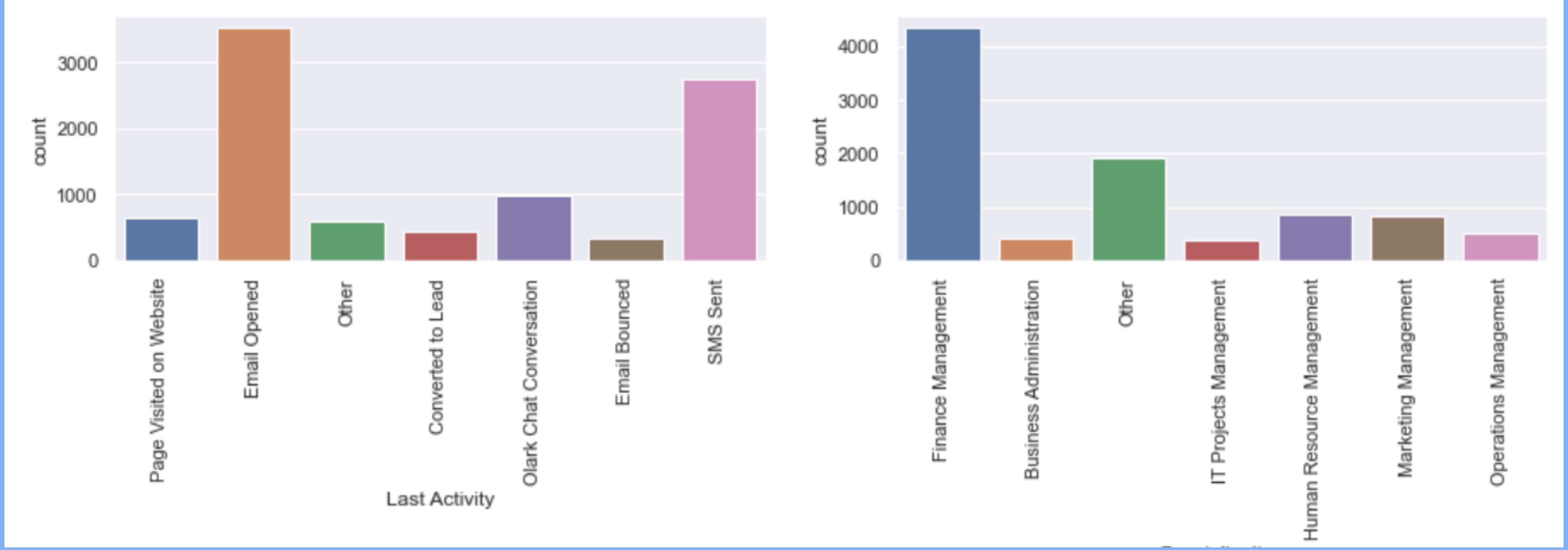
Exploratory Data Analysis

Univariate Analysis (Categorical)



upGrad

CVRGU | DS | JRCP

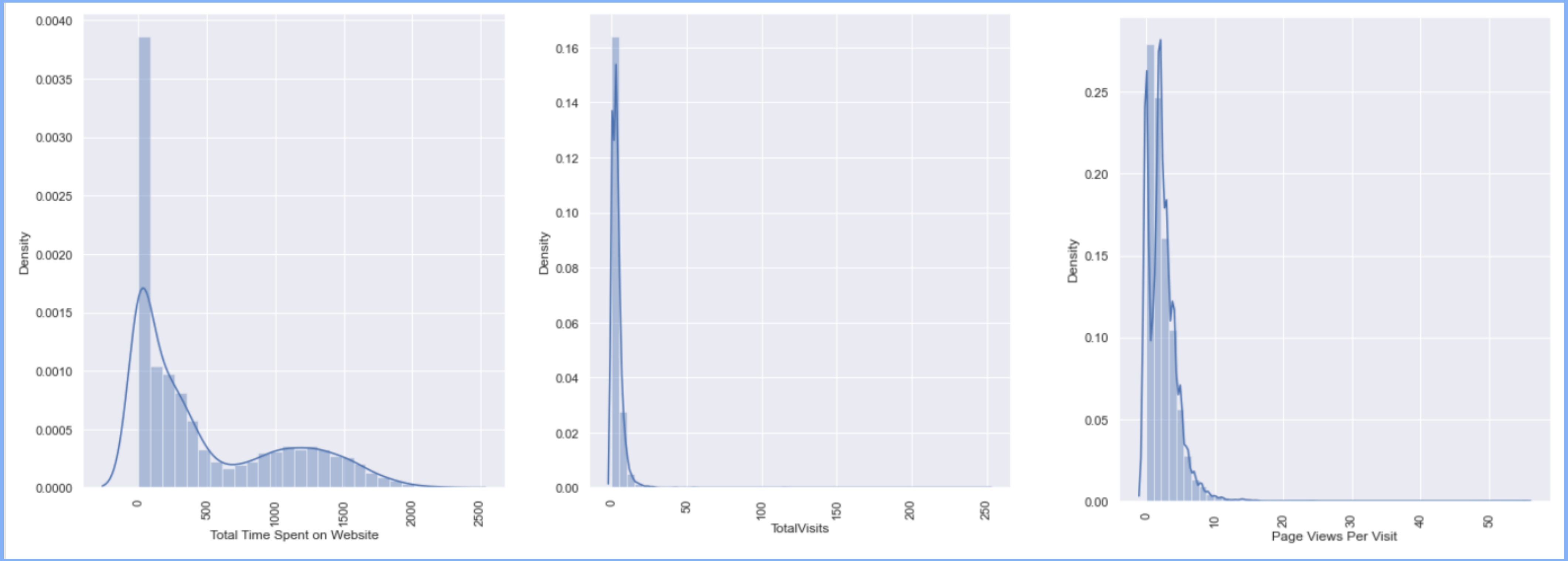


Insights :-

- In Lead Source Direct Traffic and Google are the two main source for Leads
- The Number of values is High in Email Opened and SMS Sent in Last Activity
- Most of the people chooses Finance Management Specialization rather than other Specialization

upGrad
CVRGU | DS | JRCP

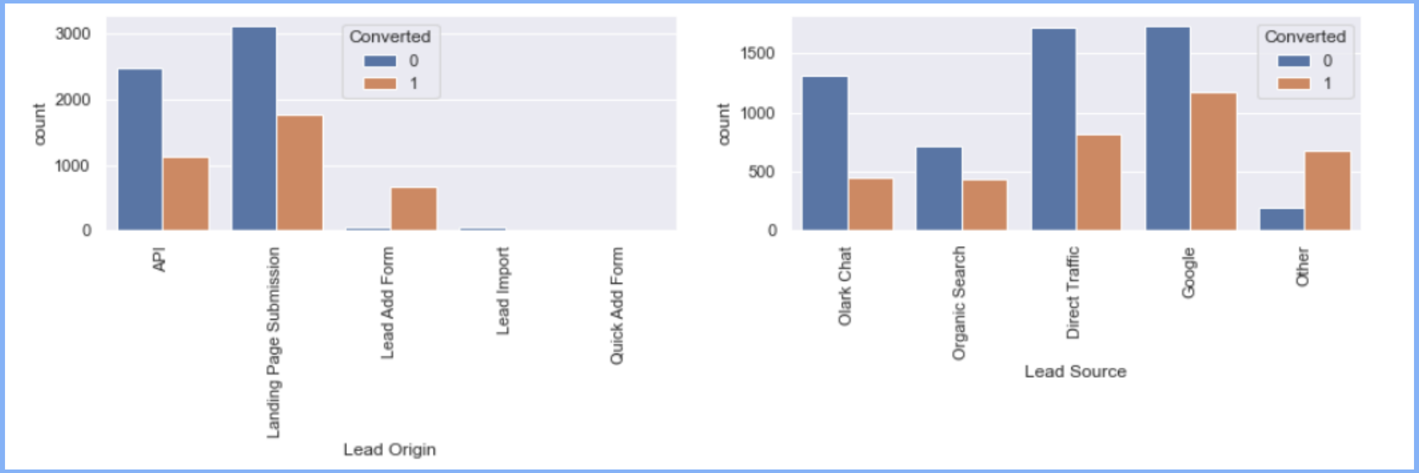
Univariate Analysis(Contenious)

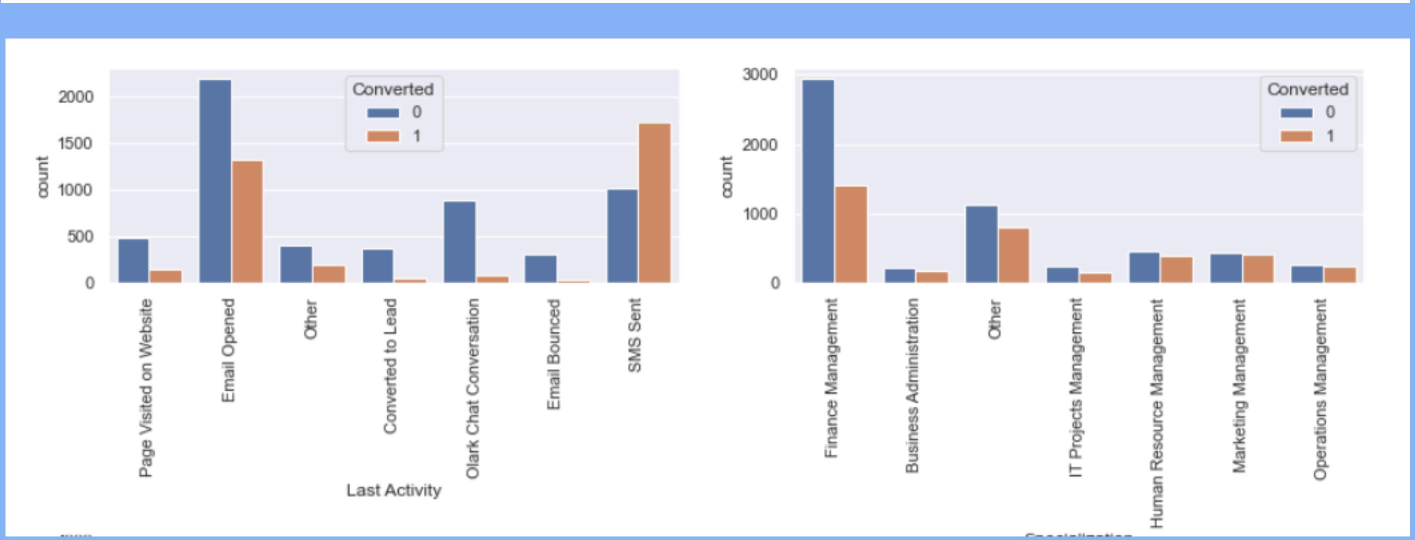


Insights :-
None of the Continueous Variables are in Normal distribution
Presence of Outliers in Total Visits and Page Views Per Visit
In total visits more values is between 0-50 and page views per visits 0-20

upGrad
CVRGU | DS | JRCP

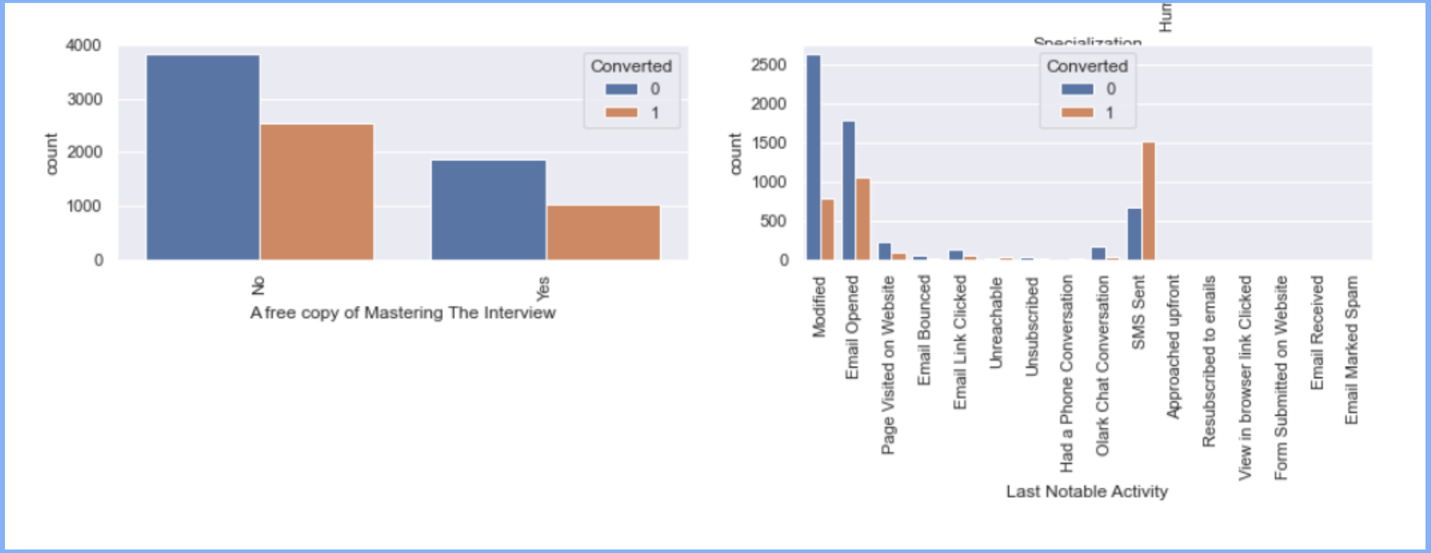
Bivariate Analysis





upGrad

CVRGU | DS | JRCP



Insights :-

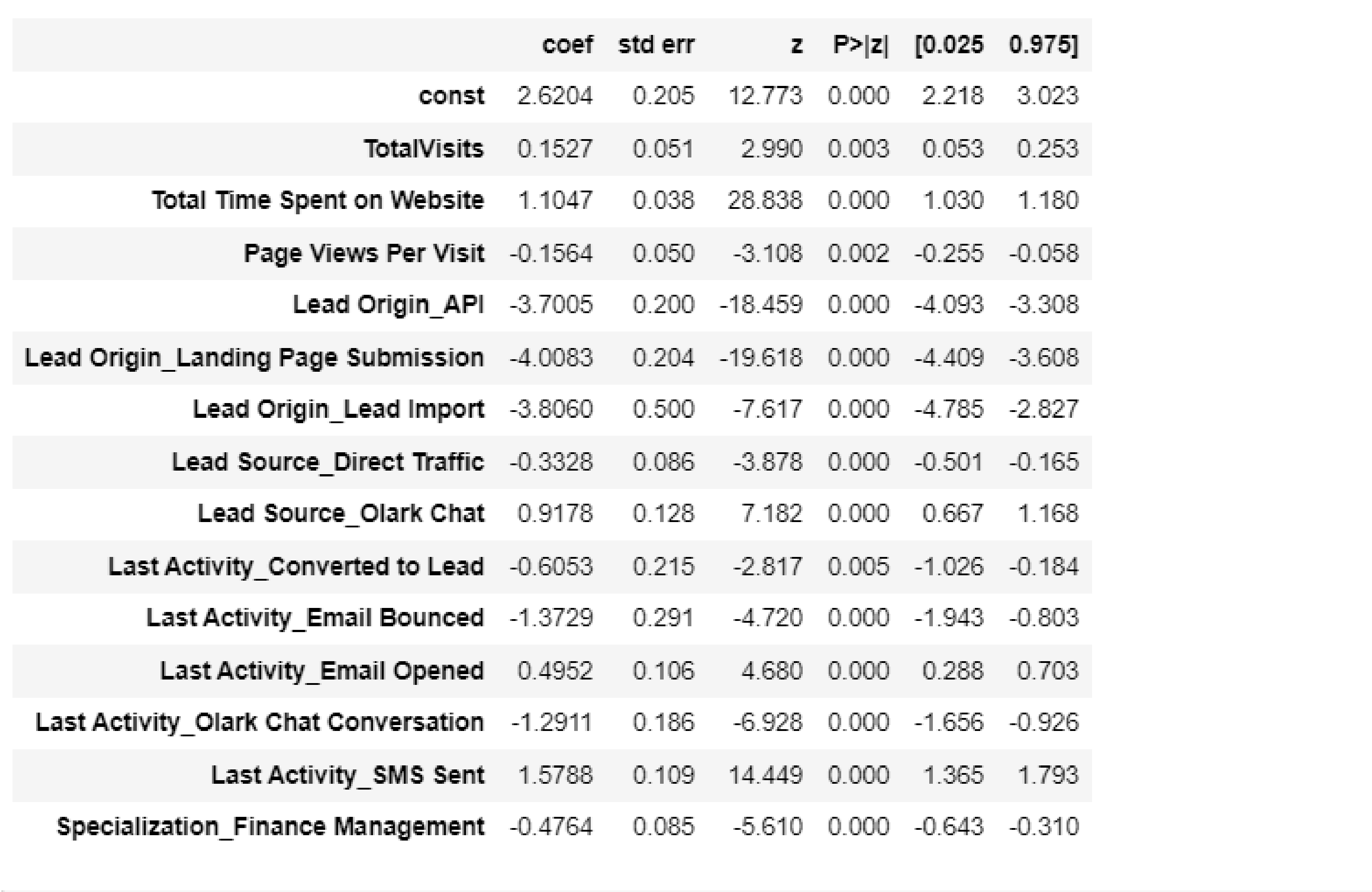
In Lead Source The number of Hot leads is higher in Direct Traffic and Google less in Other Category

In Last Activity the number of Hot leads is higher in SMS and in EMAIL cold leads is higher than hot leads.

In Last Notable Activity it's mostly same as Last Activity.

upGrad
CVRGU | DS | JRCP

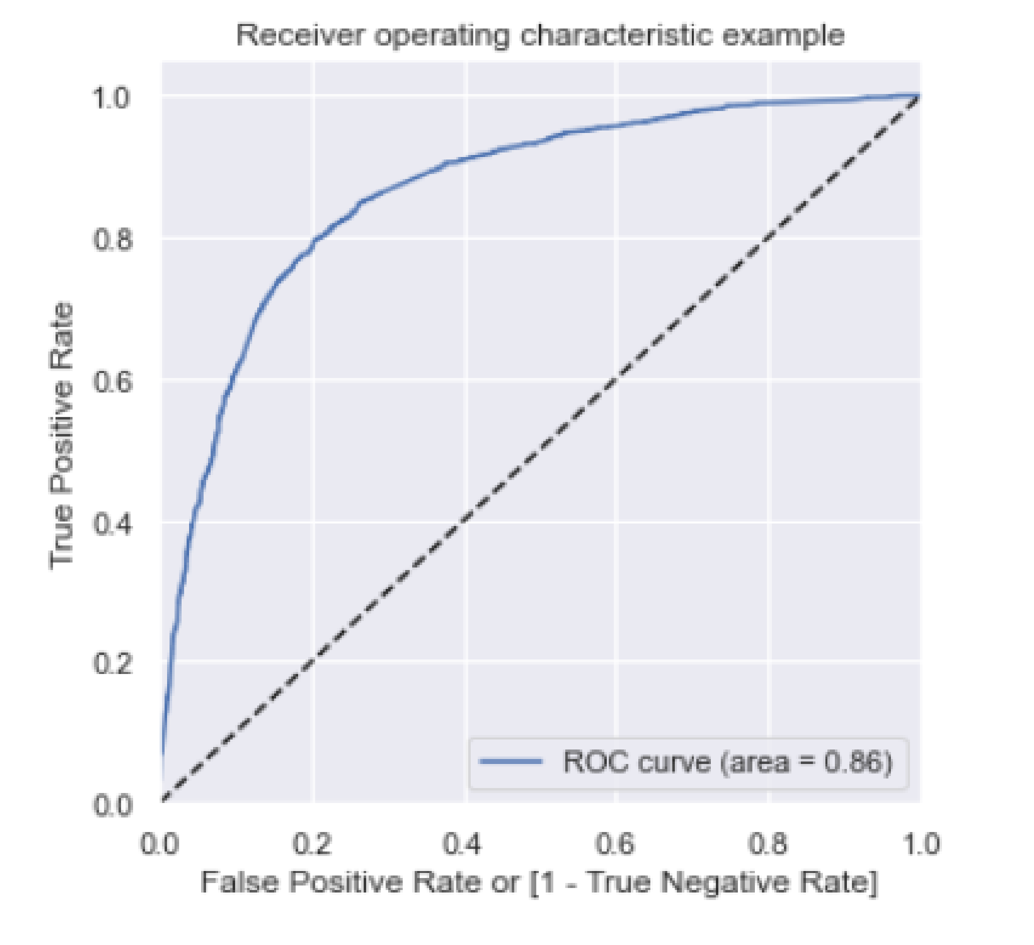
Final Model



	coef	std err	z	P>|z|	[0.025	0.975]
const	2.6204	0.205	12.773	0.000	2.218	3.023
TotalVisits	0.1527	0.051	2.990	0.003	0.053	0.253
Total Time Spent on Website	1.1047	0.038	28.838	0.000	1.030	1.180
Page Views Per Visit	-0.1564	0.050	-3.108	0.002	-0.255	-0.058
Lead Origin_API	-3.7005	0.200	-18.459	0.000	-4.093	-3.308
Lead Origin_Landing Page Submission	-4.0083	0.204	-19.618	0.000	-4.409	-3.608
Lead Origin_Lead Import	-3.8060	0.500	-7.617	0.000	-4.785	-2.827
Lead Source_Direct Traffic	-0.3328	0.086	-3.878	0.000	-0.501	-0.165
Lead Source_Olark Chat	0.9178	0.128	7.182	0.000	0.667	1.168
Last Activity_Converted to Lead	-0.6053	0.215	-2.817	0.005	-1.026	-0.184
Last Activity_Email Bounced	-1.3729	0.291	-4.720	0.000	-1.943	-0.803
Last Activity_Email Opened	0.4952	0.106	4.680	0.000	0.288	0.703
Last Activity_Olark Chat Conversation	-1.2911	0.186	-6.928	0.000	-1.656	-0.926
Last Activity_SMS Sent	1.5788	0.109	14.449	0.000	1.365	1.793
Specialization_Finance Management	-0.4764	0.085	-5.610	0.000	-0.643	-0.310

upGrad
CVRGU | DS | JRCP

Final Model



ROC Curve

upGrad
CVRGU | DS | JRCP

Conclusion

We have noted that the variables that important the most in the potential buyers are:

- The total time spend on the Website.
- Total number of visits.
- When the lead source was: -
  Olark Chat
- When the last activity was: -
  SMS
  Olark chat conversation