M.S.E Project Report:

Exploratory data Analysis, Hypothesis Testing and Logistic Regression on Prediction and Analysis of UpGrad users/customers.

Abstract-
Buying Prediction

Submitted to –

Praxis Business School,Kolkata
Dr. Sayantani Roy Choudhury

Submitted by-

Aditya Pathania (A21002)
Shayantam Das (A21029)

RESEARCH OBJECTIVE

An education company named UpGrad 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 UpGrad 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.

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. In the middle stage, you need 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.

Our main Objective will cover:

- Factors that are most established to finally get the customer/lead converted to buy the course on the website.
- Analysis and Hypothesis if any of the data collection can lead to our research objective.
- Probability / chances of the Lead getting converted.
- Checking if the factors that are dependent /independent , having to predict the buying of the course by the lead.
- Working on the conversion ratio with the collected / provided dataset

The overall objective is to get the data and identity the factors that can lead to a customer buying the course , and to check or predict the nature of the customer related to buying the course , Identifying the customers/lead that can lead to getting converted with a certain probability , Hypothesis will include the factors that are dependent on getting the converted ratio more , as overall conversion ratio is at 30% we need to check if mean conversion is 30 or more by doing the hypothesis , as the company will need to get the lead conversion ratio increased

The research questions framed / given will have to be analysed and worked on related to the objective, the questions for the research are as follows .

RESEARCH QUESTIONS AND METHEDOLOGY

RESEARCH QUESTION:

As per the Objective Framed above , we would like to work on the leads that are more probable to be converted , the data provided by the company , has given the question as-

UpGrad Education wants to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score h have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

As per the research the Company wants us to build a model wherein the target lead conversion is subjected to 80 %, analysing the factors where in the customers are projected to convert building a model that will help to predict or analyse the important features which can cross the 80% limit .

METHODOLOGY

Following our research objective and question, the research will be done followed by points mentioned below :

- We need to get the data , that the company has provided i.e. we have secondary data for our research which is gathered by the company as per there business understanding.
- After gathering the data we need to understand the data , and the features that are provided in the data , and if they fulfil our research objective or not , if yes , then we will go ahead with the next step.
- This will then follow with the data exploration, we need to analyse the data related to univariate and bivariate analysis to understand the structure and the features of the data.
- If there are some steps that states that there are certain features that are highly related to each other so we need to do a bit of hypothesis to get the dependency check for the features , which may involve , mean of the features, chi square test or F test as well. We also need to check if the mean conversion collected on the dataset is even more than the 30% mark or not

- Once done we will start with model preparation , that will include , feature selection , feature engineering for the important or the relevant features , the main focus of the model building will work on the 90 % confidence , as the benchmark for the analysis is 80%
- Model valuation , we will need to evaluate our model , and pick a model that is best compared to a benchmark set for the analysis.
- Feature engineering will be done using the VIF , which states the features will be removed for which the P- value is not significant.
- The best model will be picked based on the classification report , based on the significant features, hypothesis inference done on the steps above.

The main task for the research is to get the good conversion rate , and work on such data that is relevant to pick the drivers for the conversion and lead to targeted conversion rate.

Data gathered by the company is according to the business understanding and online services or questioner provided to the customers while on the time of creation of the account.

The Detailed understanding and the analysis of the data is as follows.

DATA UNDERSTANDING

As per the data provided by the company contained multiple features that were gathered on the basis of the user's account or the user's activity on the website and weather the lead given to a sales agent leaded to the conversion or not , the main data will include the lead which were converted and which were not based on the account and the activity , the features provided are listed below , the list does not contain the fields with name or ID no's.

- Lead Origin- The features tells from where the sales got the lead from , type of the features is categorical , and it includes following categories.
    1) Landing page submission
    2) API
    3) Lead Add form
    4) Lead Import
    5) Quick Add form
- Lead Source – The feature tells what was the source of the lead , this is also a categorical and contains various categories like –

```
Google               2868
Direct Traffic       2543
Olark Chat           1755
Organic Search       1154
Reference             534
Welingak Website      142
Referral Sites        125
Facebook               55
bing                    6
google                  5
Click2call              4
Press_Release           2
Social Media            2
Live Chat               2
youtubechannel          1
testone                 1
Pay per Click Ads       1
welearnblog_Home        1
WeLearn                 1
blog                    1
NC_EDM                  1
Name: Lead Source, dtype: int64
```

- Do not email and Do not call- This feature is provided by the user account , and tells if the user wishes to be contacted via email /call or not, this is a binary category feature that includes
    1) Yes
    2) No
- Converted- This feature is basically our target variable , that contains the data if the lead was converted or not This also contains 2 categories
    1) Yes
    2) No
- Total Visits- This is a numerical feature that tells how many times did the user visited the website
- Total time spent on the website- This is also a numerical feature that tells the total time the user spent on the website.
- Page views per visit- This can be converted to categorical feature , but there are some entries that are in decimals ,which will be converted to singe category in data cleaning.
- Last activity- This is a category feature , that tells what was the last activity of the user
- Specialization – This feature tells the specialization of the user , which is divided in the category

Some features are also divided in the form of the questioner , which are :

How did you hear about upgrad education, what is your current occupation, what matters you the most while choosing the course .

- Asymmetrique activity index- This tells us about the user's activity , divided into 3 categories.
    1) Low
    2) Medium
    3) High
- Asymmetric activity Score- This is also a category variable , which has the score in categories ranging from 11 to 15
- Asymmetrique profile index and Score - The website algorithm score is given to each profile as per the subject matter expert , which is same as above activity
- Tags-This is also a category which consists on the responses given by the user on the sales rep conversation , weather the user contacted or what was the response of the user
- Lead quality – This tells the quality of the user , this tells the impression of the rep for the user and if the user can be converted or not , and how do the sales plans on contacting the user , the categories included are :

1) Might be
2) Not sure
3) High in relevance
4) Worst
5) Low in relevance

- Update me on supply chain content- This tells weather customer want update back or not
- I agree to pay through the cheque – This tells if the customer is willing to pay or not , this can be used to analyse if the customer is willing to pay and if user finally pays or not

There are also some features , that does not have any explanation but need to be analysed with the EDA which are – Search, Magazine, Newspaper Article, Through recommendations, Lead profile.

0	Prospect ID	9240 non-null	object
1	Lead Number	9240 non-null	int64
2	Lead Origin	9240 non-null	object
3	Lead Source	9204 non-null	object
4	Do Not Email	9240 non-null	object
5	Do Not Call	9240 non-null	object
6	Converted	9240 non-null	int64
7	TotalVisits	9103 non-null	float64
8	Total Time Spent on Website	9240 non-null	int64
9	Page Views Per Visit	9103 non-null	float64
10	Last Activity	9137 non-null	object
11	Country	6779 non-null	object
12	Specialization	7802 non-null	object
13	How did you hear about X Education	7033 non-null	object
14	What is your current occupation	6550 non-null	object
15	What matters most to you in choosing a course	6531 non-null	object
16	Search	9240 non-null	object
17	Magazine	9240 non-null	object
18	Newspaper Article	9240 non-null	object
19	X Education Forums	9240 non-null	object
20	Newspaper	9240 non-null	object
21	Digital Advertisement	9240 non-null	object
22	Through Recommendations	9240 non-null	object
23	Receive More Updates About Our Courses	9240 non-null	object
24	Tags	5887 non-null	object
25	Lead Quality	4473 non-null	object
26	Update me on Supply Chain Content	9240 non-null	object
27	Get updates on DM Content	9240 non-null	object
28	Lead Profile	6531 non-null	object
29	City	7820 non-null	object
30	Asymmetrique Activity Index	5022 non-null	object
31	Asymmetrique Profile Index	5022 non-null	object
32	Asymmetrique Activity Score	5022 non-null	float64
33	Asymmetrique Profile Score	5022 non-null	float64
34	I agree to pay the amount through cheque	9240 non-null	object
35	A free copy of Mastering The Interview	9240 non-null	object
36	Last Notable Activity	9240 non-null	object

DATA CLEANING AND UNIVARIATE ANALYSIS

The main part for the huge data we have is data cleaning , the first step includes checking null values .

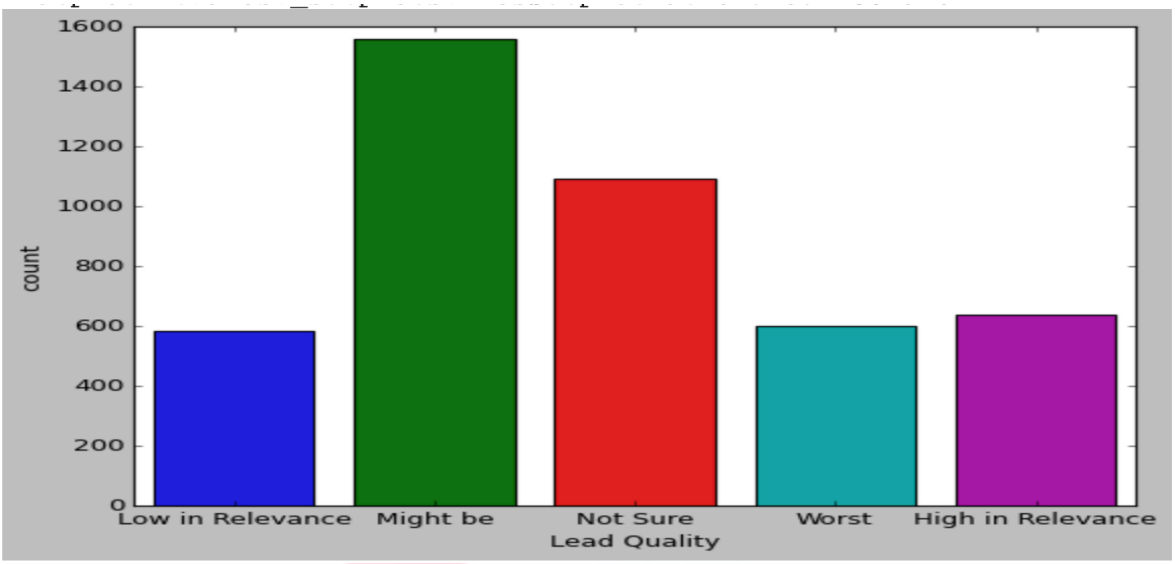
It was found that there were features that had more than 70% null values and others as well .

Prospect ID	0.00
Lead Number	0.00
Lead Origin	0.00
Lead Source	0.39
Do Not Email	0.00
Do Not Call	0.00
Converted	0.00
TotalVisits	1.48
Total Time Spent on Website	0.00
Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	36.58
How did you hear about X Education	78.46
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Search	0.00
Magazine	0.00
Newspaper Article	0.00
X Education Forums	0.00
Newspaper	0.00
Digital Advertisement	0.00
Through Recommendations	0.00
Receive More Updates About Our Courses	0.00
Tags	36.29
Lead Quality	51.59
Update me on Supply Chain Content	0.00
Get updates on DM Content	0.00
Lead Profile	74.19
City	39.71
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65
I agree to pay the amount through cheque	0.00
A free copy of Mastering The Interview	0.00
Last Notable Activity	0.00

Here we have 2 features that have null values more than 70% so we dropped the data for these features.

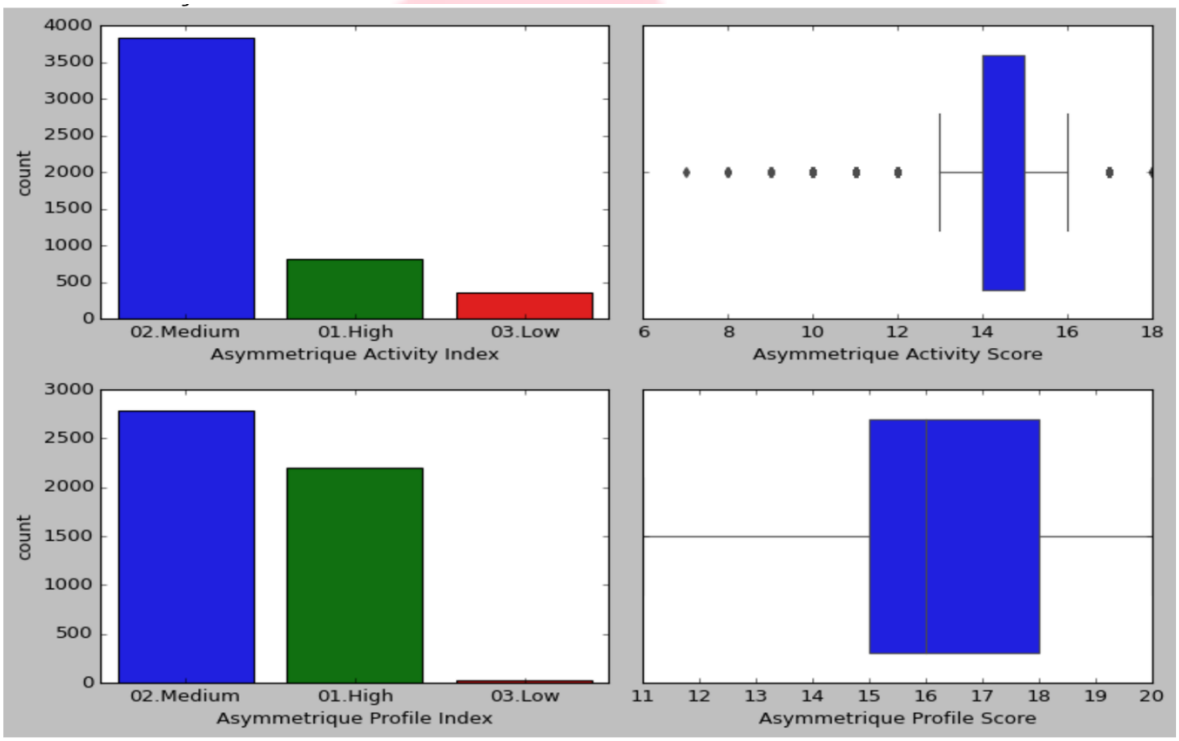
The other features that had the null values were treated based on univariate analysis, following are the results

- Lead Quality – We will bar plot the data for this feature



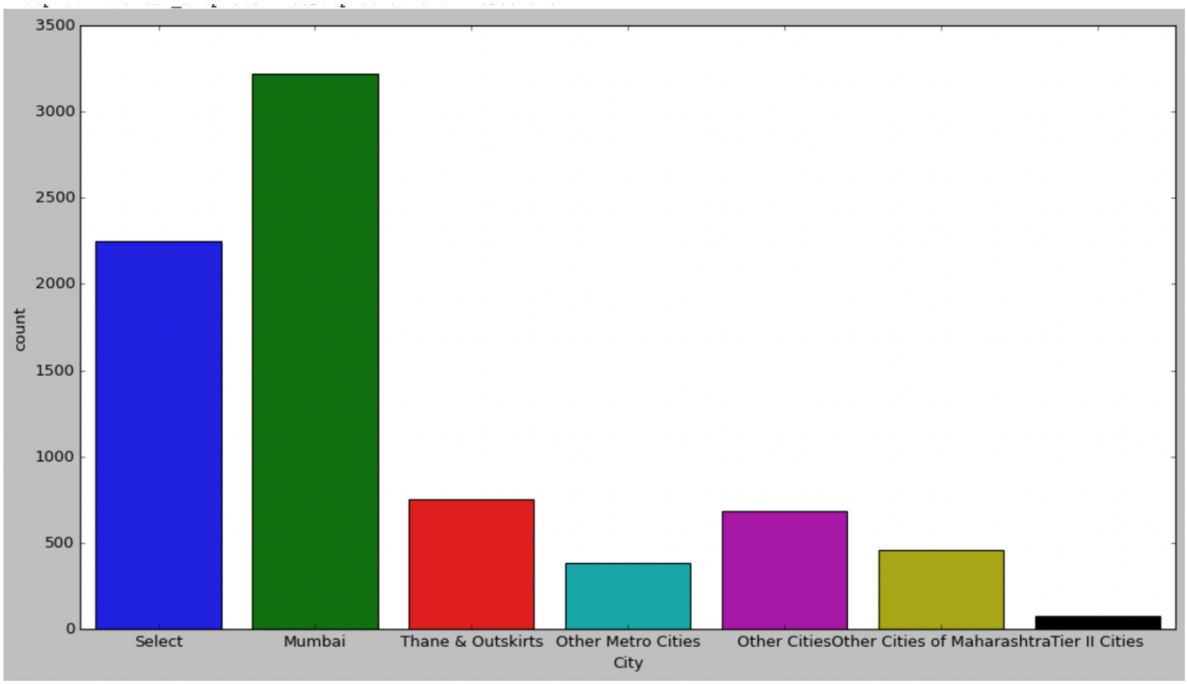
As Lead quality is based on the intuition of the rep , so if it is blank , it is best to replace it with Not sure .

- Asymmetrique Profile Index, Asymmetrique Profile Score, Asymmetrique Activity Index, Asymmetrique Activity Score.



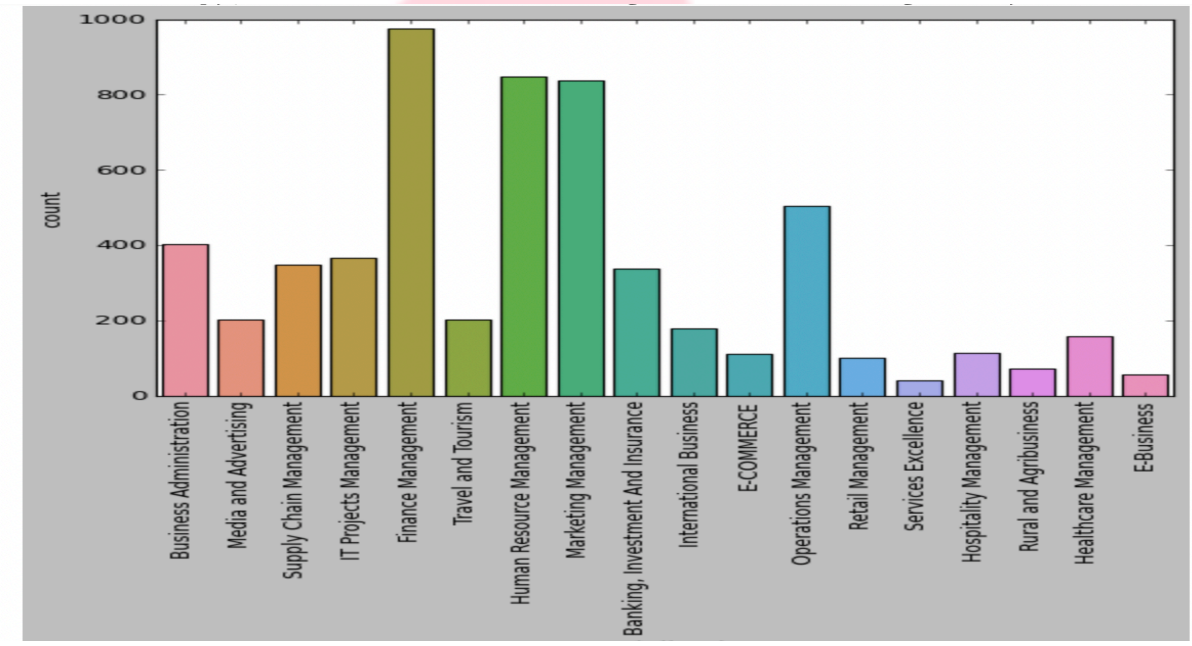
As we see that there is so much variation in the data , we will have to drop the features.

- City – Checking the leads are from which cities



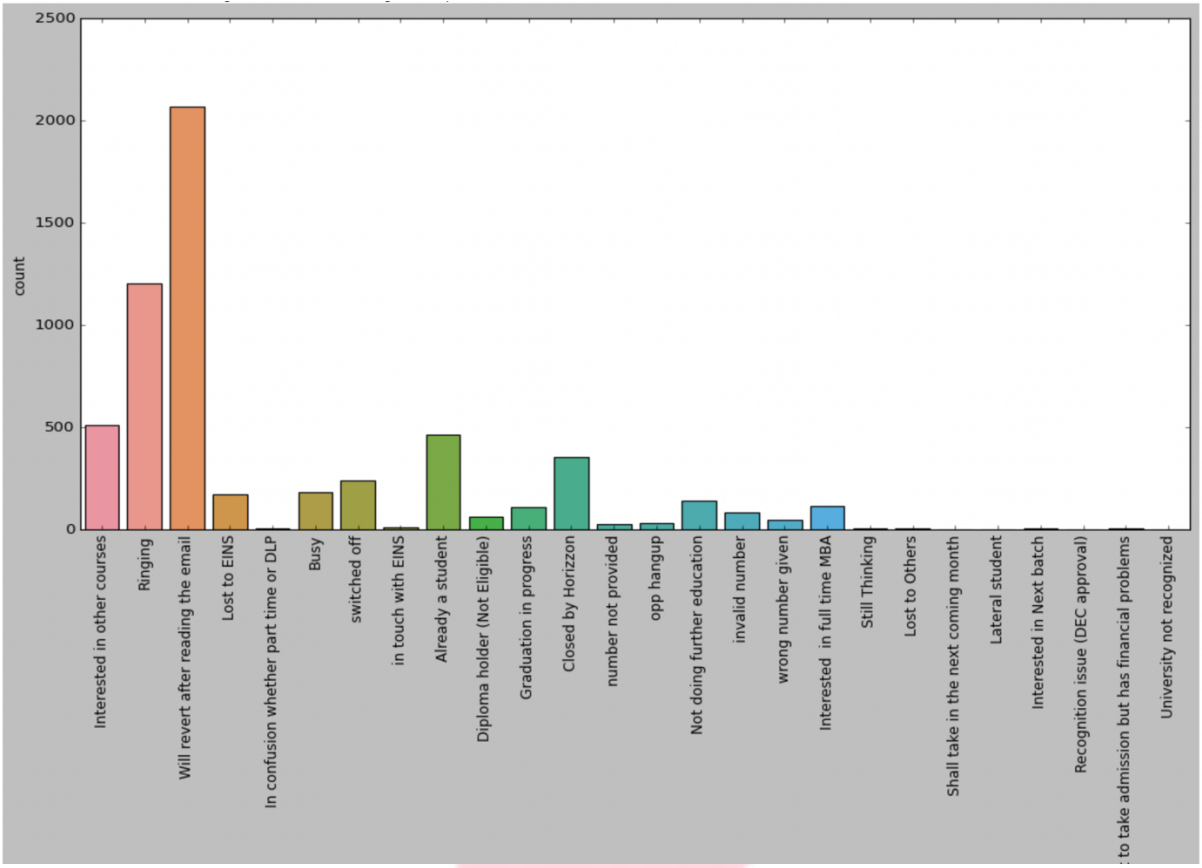
People from Mumbai are the most from the dataset , so we can replace the null with the Mumbai

- Specialization- We will check the specialization and replace the null values accordingly



It maybe that user has not entered any specialization , probably because the option might not be available , so we will replace the null with others as a new category

- Tags-



The null values here can be included here with the 'Will revert after reading the email'

- What matter you the most while choosing the course – as almost 99% of the values were 'better career prospects' so we can replace the null with that.

After all the replacement of the null values , others remaining null values are just 2 % , so we will drop the values.

The remaining features are –

```
Prospect ID                                                    0.0
Lead Number                                                    0.0
Lead Origin                                                    0.0
Lead Source                                                    0.0
Do Not Email                                                   0.0
Do Not Call                                                    0.0
Converted                                                      0.0
TotalVisits                                                    0.0
Total Time Spent on Website                                    0.0
Page Views Per Visit                                           0.0
Last Activity                                                  0.0
Country                                                        0.0
Specialization                                                 0.0
What is your current occupation                                0.0
What matters most to you in choosing a course                  0.0
Search                                                         0.0
Magazine                                                       0.0
Newspaper Article                                              0.0
X Education Forums                                             0.0
Newspaper                                                      0.0
Digital Advertisement                                          0.0
Through Recommendations                                        0.0
Receive More Updates About Our Courses                         0.0
Tags                                                           0.0
Lead Quality                                                   0.0
Update me on Supply Chain Content                              0.0
Get updates on DM Content                                      0.0
City                                                           0.0
I agree to pay the amount through cheque                       0.0
A free copy of Mastering The Interview                         0.0
Last Notable Activity                                          0.0
dtype: float64
```

EXPLORATORY DATA ANALYSIS

- Converted

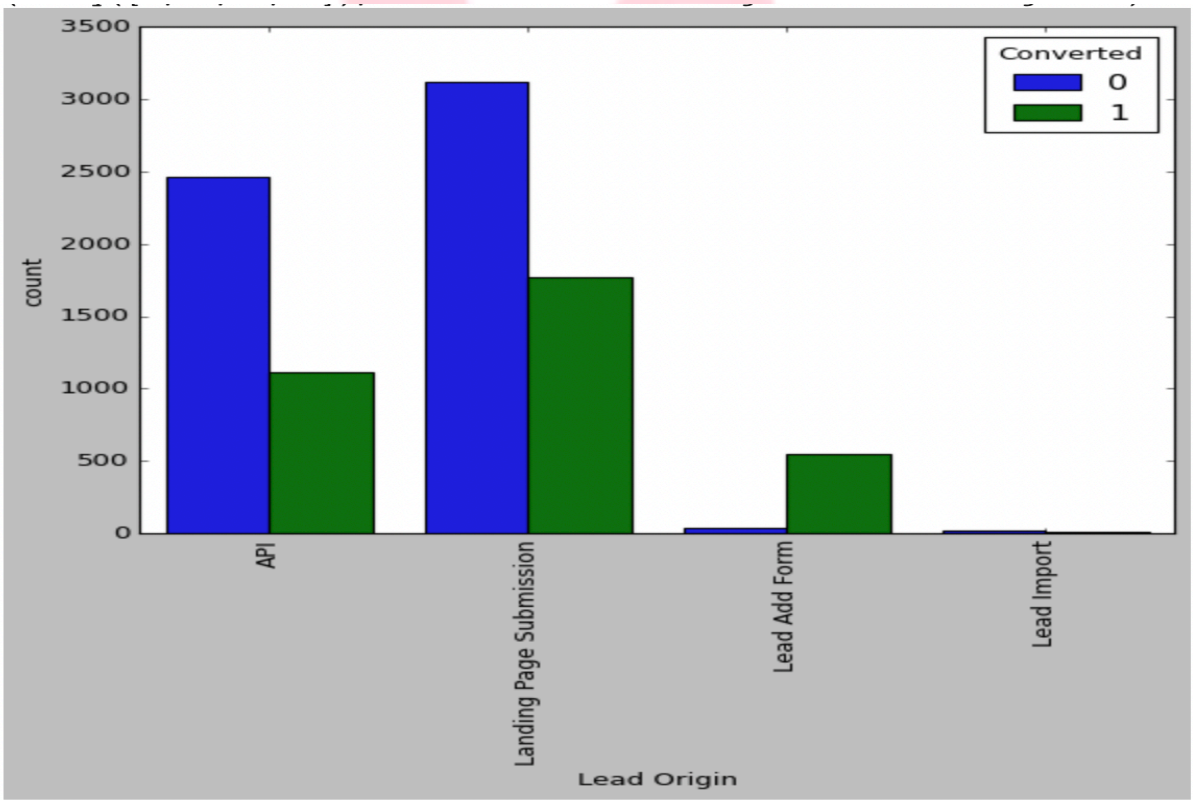
Ratio of the users that converted are

```
Converted = (sum(data['Converted'])/len(data['Converted'].index))*100
print(Converted,'%')
```

```
37.85541106458012 %
```

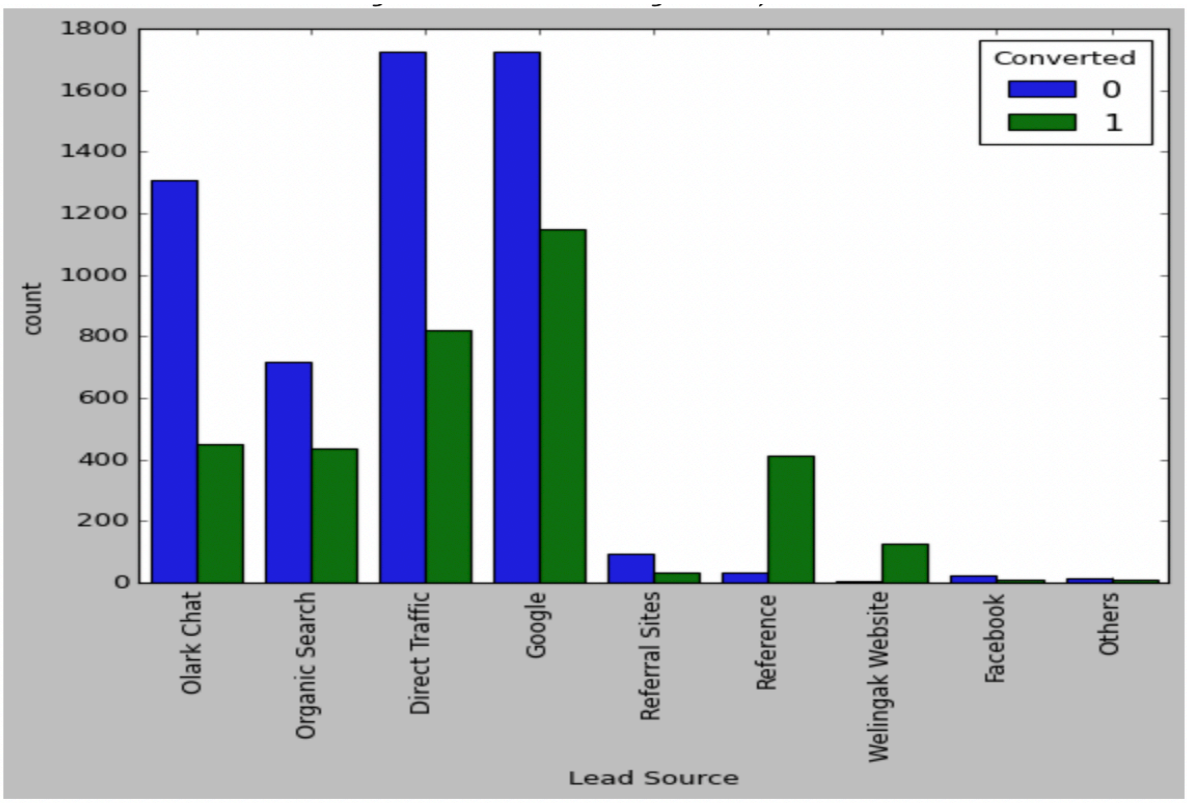
We see that our dataset has ratio of 37% of the users.

- Lead origin with converted



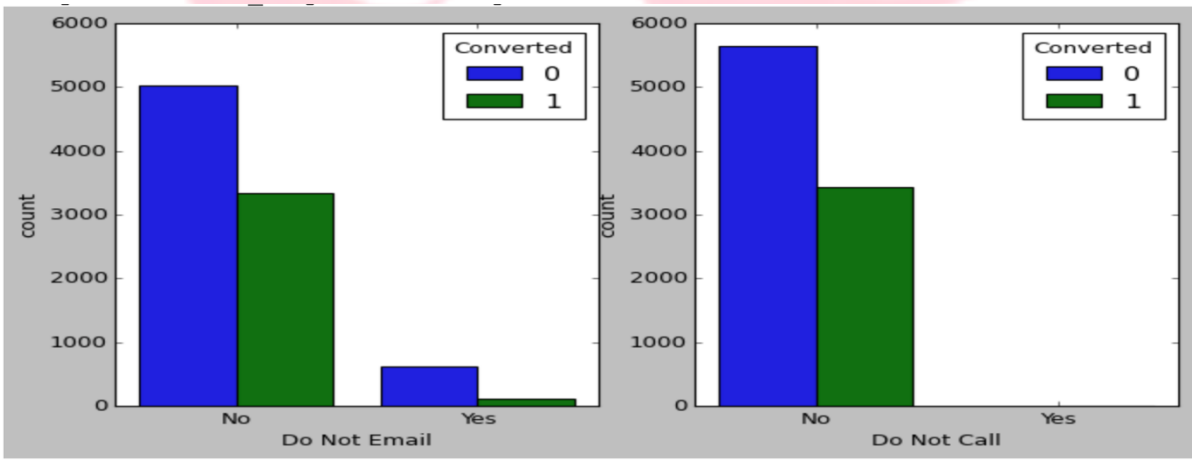
Inference-API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable. Lead Add Form has more than 90% conversion rate but count of lead are not very high. Lead Import are very less in count. To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

- Lead source – converted . as seen that there were many small entries which individually were not relevant , better if these were part of others, changing such features to others the plot looks like –



Inference-Google and Direct traffic generates maximum number of leads. 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.

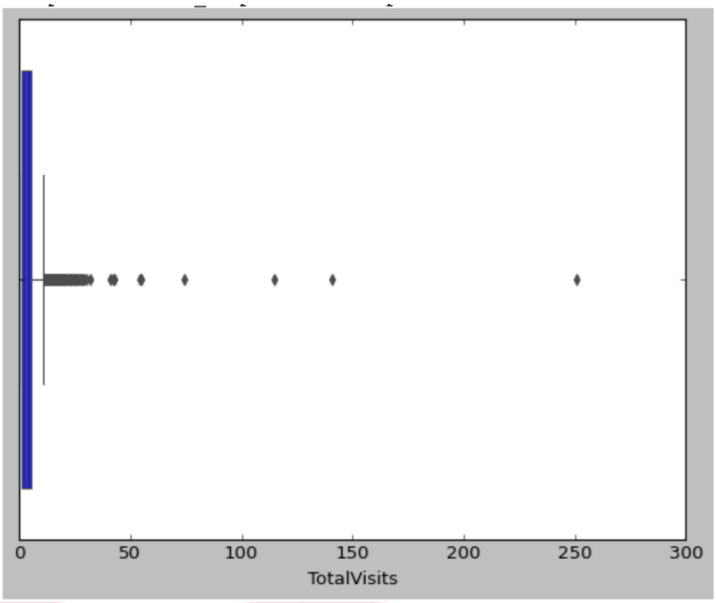
- Do not email and do not call –



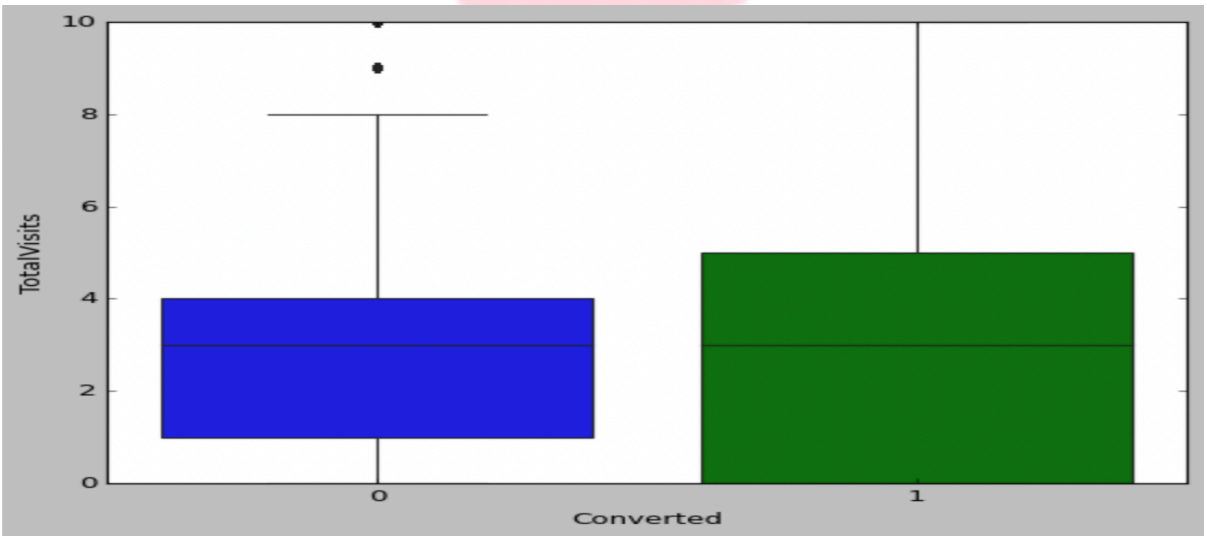
Inference- Looking at the plot we can say that when it comes to calling or email , the conversion efforts looks less , as a person opting for call /email is not getting converted with 50 % ratio

- Total Visits-

```
count    9074.000000
mean        3.456028
std         4.858802
min         0.000000
5%          0.000000
25%         1.000000
50%         3.000000
75%         5.000000
90%         7.000000
95%        10.000000
99%        17.000000
max       251.000000
Name: TotalVisits, dtype: float64
```



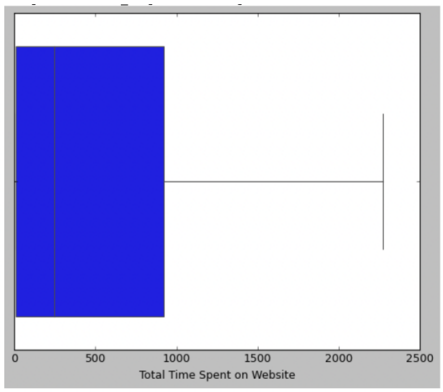
As there are lot of outliers, we will cap the outliers to 95% value for analysis.

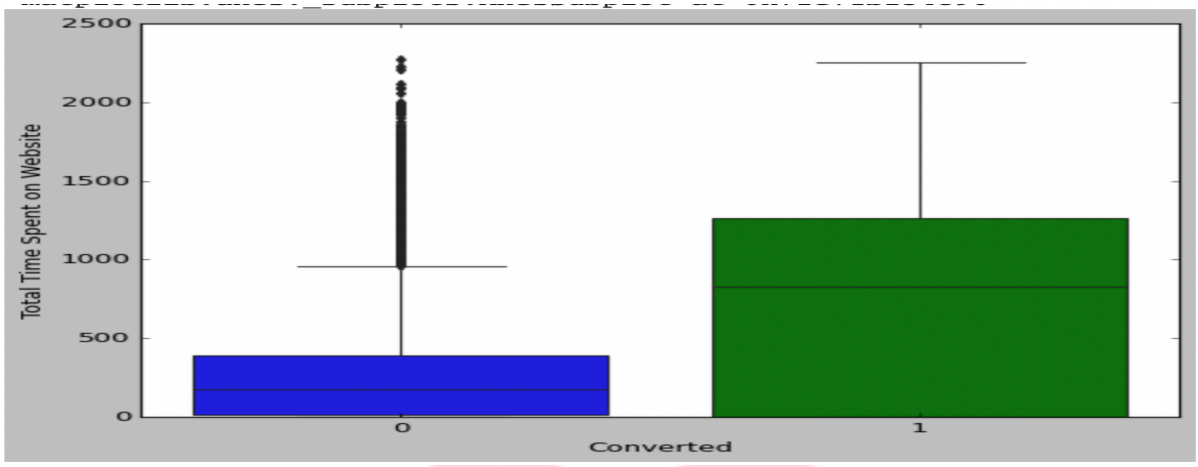


Inference-Median for converted and not converted leads are the same. Nothing conclusive can be said on the basis of Total Visits.

- Total time spend on the website – Analysing the total time for the website that the user spends

count	9074.000000
mean	482.887481
std	545.256560
min	0.000000
25%	11.000000
50%	246.000000
75%	922.750000
max	2272.000000

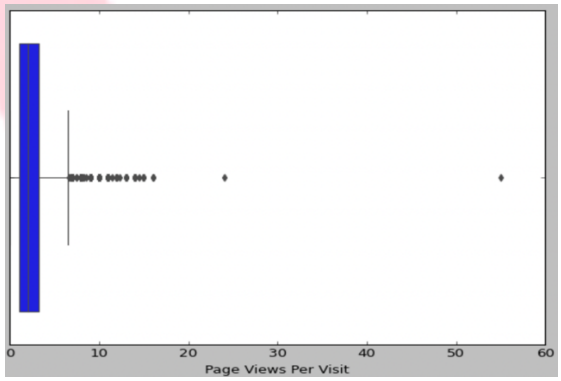




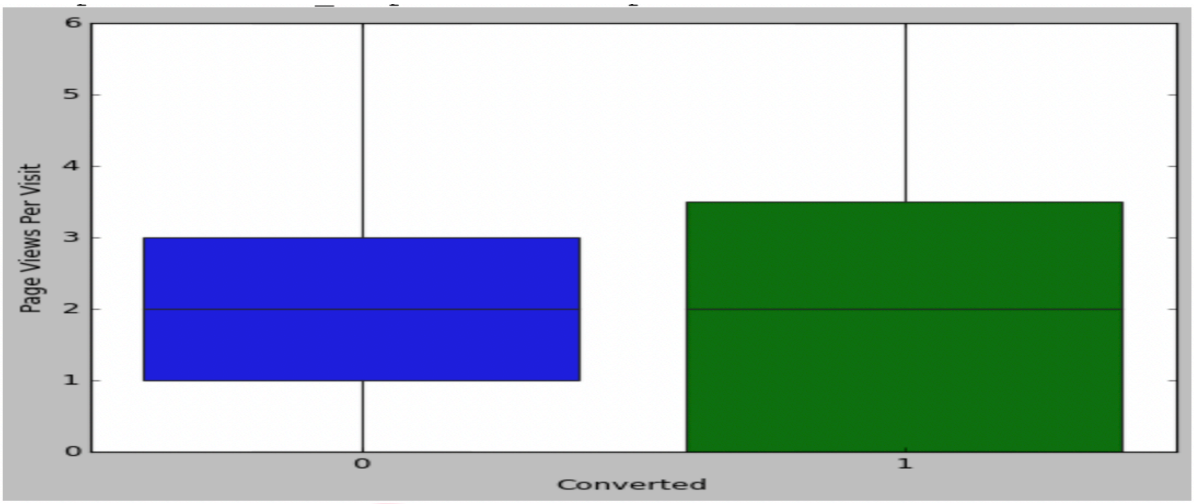
Inference-Leads spending more time on the website are more likely to be converted. Website should be made more engaging to make leads spend more time.

- Page Views per visit

count	9074.000000
mean	2.370151
std	2.160871
min	0.000000
25%	1.000000
50%	2.000000
75%	3.200000
max	55.000000



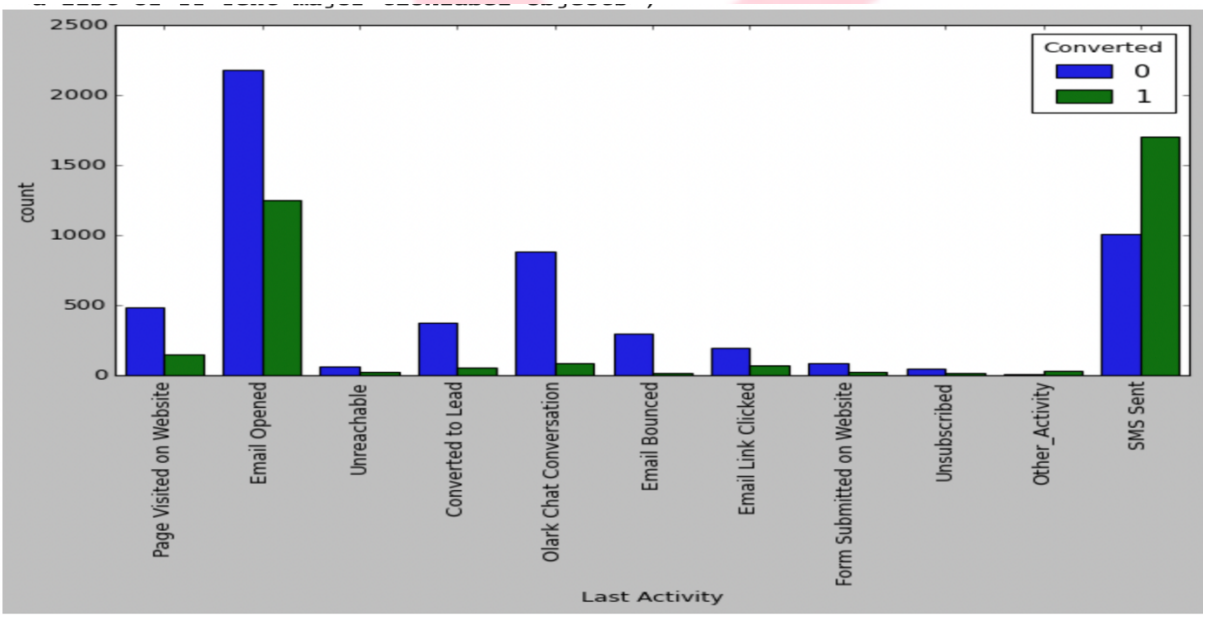
As there are many outliers in the data we will cap the data to 95%.



Inference-Median for converted and unconverted leads is the same. Nothing can be said specifically for lead conversion from Page Views Per Visit.

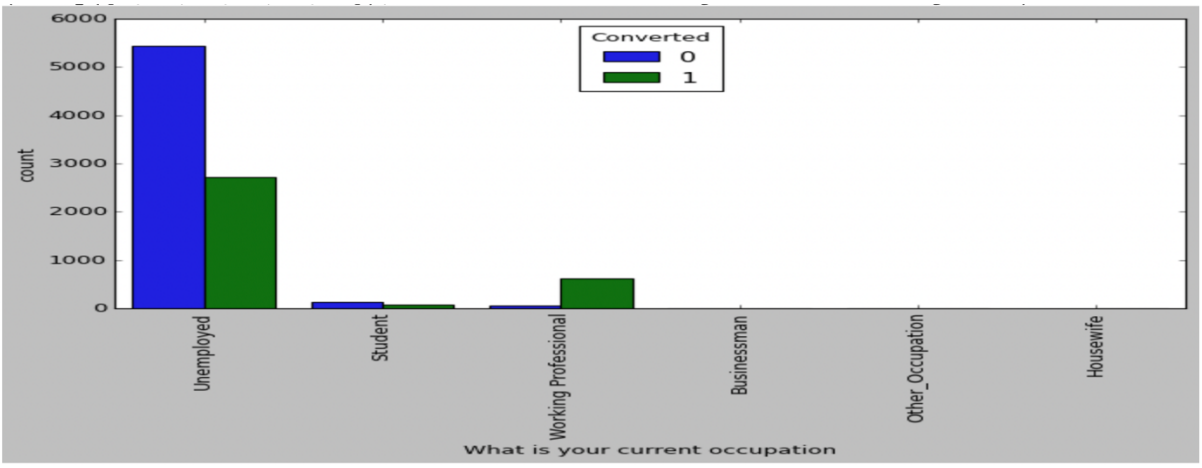
- Last activity

There are some features that are not relevant individually so combining them to others.



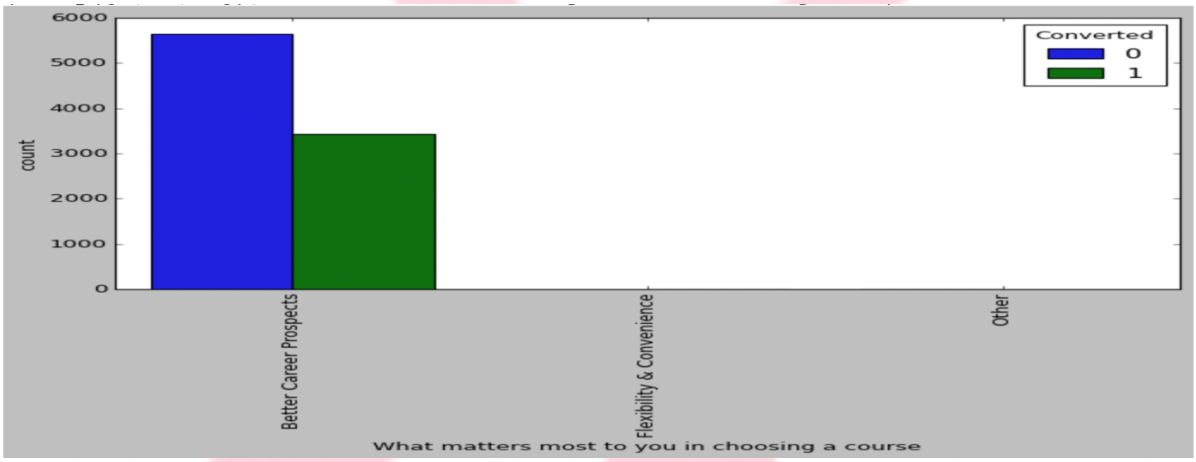
Inference-Most of the lead have their Email opened as their last activity. Conversion rate for leads with last activity as SMS Sent is almost 60%.

- Occupation – Checking the plot for occupation and the one's that converted



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.

- What matters most to you in choosing a course



As discussed above , better career prospect has most converted ratio.

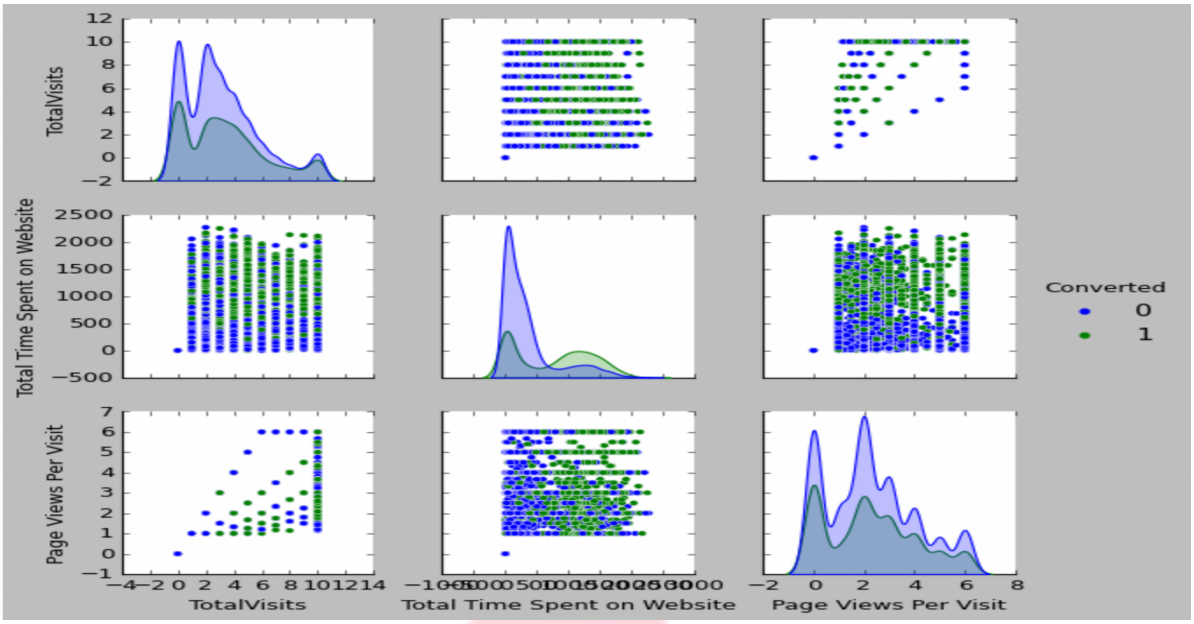
Inference from Univariate- also covered the topics such as Search, digital advertisement , Newspaper etc, Most entries are 'No'. No Inference can be drawn with this parameter.
Results Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we can drop them for further analysis.
Columns remaining are-

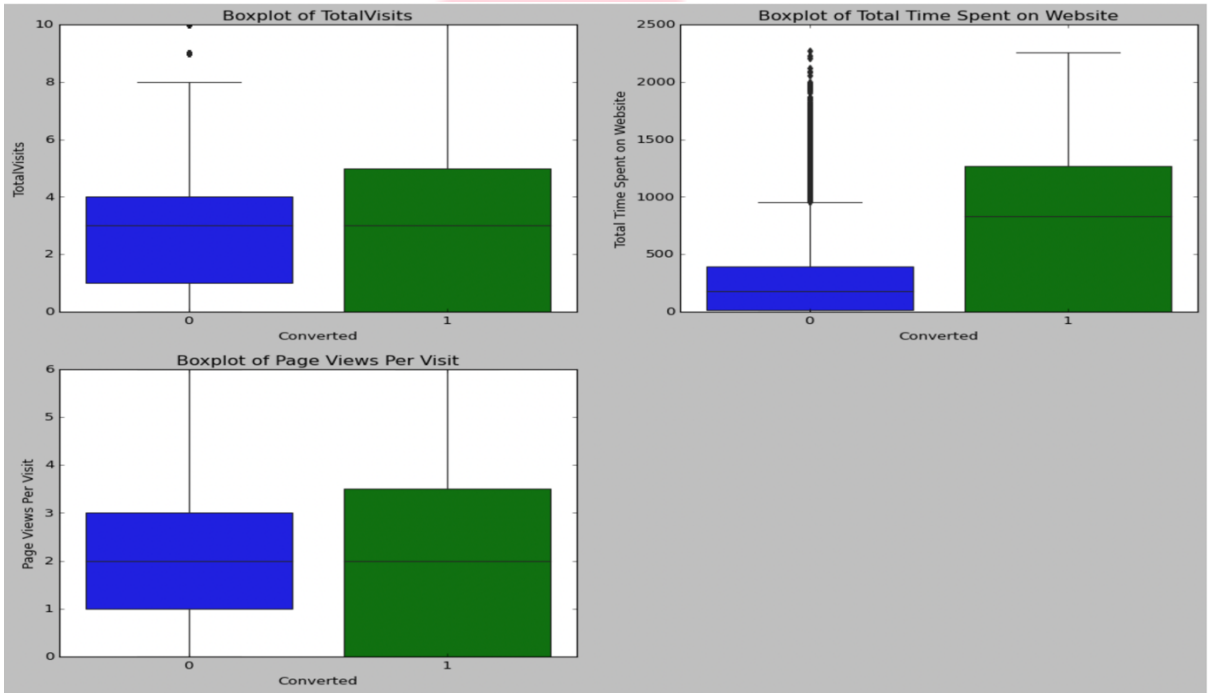
```
Index(['Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call',
       'Converted', 'TotalVisits', 'Total Time Spent on Website',
       'Page Views Per Visit', 'Last Activity', 'Specialization',
       'What is your current occupation', 'Tags', 'Lead Quality', 'City',
       'Last Notable Activity'],
      dtype='object')
```

MULTIVARIATE AND HYPOTHESIS TESTING

NUMERICAL:

- Total visits
- Total time spent on website
- Page view per visit
  - Converted





Inference -The Total Time Spent on the website by the successfully converted customers is higher than the non-converters. There is a wide range of dispersion seen for successfully converted customers.

HYPOTHESIS-

Chi-square testing

1) Tags vs What is your current occupation-

   Null Hypothesis- The variables are independent.

   Alternate Hypothesis- The variables are dependent.

```
# Chi-square test of independence.
c, p, dof, expected = chi2_contingency(contigency)
# Print the p-value
print(p,c,dof)
```

```
1.1746378948685731e-172 1187.9670694457257 125
```

Chi square statistic= 1187.9670694457257 with a P-value 1.1746378948685731e-172- that means the value is significant with degrees of freedom 125

Critical value= 152.093

```
#find Chi-Square critical value
scipy.stats.chi2.ppf(1-.05, df=125)
```

```
152.0938756919578
```

Inference- we reject the null Hypothesis and say there is a relationship between variables also chic square sts> critical value
That means the variables are dependent

2) Specialization vs Lead quality-

   Null Hypothesis- The variables are independent.

   Alternate Hypothesis- The variables are dependent.

```
# Chi-square test of independence.
c, p, dof, expected = chi2_contingency(contigency1)
# Print the p-value
print(p,c,dof)
```

```
9.340394139002688e-153 947.2308708250157 72
```

```
#find Chi-Square critical value
scipy.stats.chi2.ppf(1-.05, df=324)
```

```
366.9769672012343
```

Chi square statistic= 947.231 with a P-value 9.340394e-153- that means the value is significant with degrees of freedom 324

Inference-we reject the null Hypothesis and say there is a relationship between variables also chic square statistic> critical value

T-Test –Two tailed

3) Converted vs total time spent-

   Null Hypothesis – leads with converted status yes have equal mean total time spend as that of leads with converted status as no.

   Alternate- leads with converted status yes does not have equal mean total time spend as that of leads with converted status as no.

Test statistic- 84.295360

p-value for 2 tailed test nearly equal to 0

```
Test statistic is 84.295360
p-value for two tailed test is 0.000000
Conclusion n Since p-value(=0.000000) < alpha(=0.05) we reject the null hypothesis
```

Inference- we reject the null Hypothesis and say that the leads with converted status yes does not have equal mean amount of total time spent as that with leads that have converted status as no.

Anova – one way

4) Page view per visit vs total visits

   Null hypothesis- Variance is same for page view per visit and total visits

   Alternate hypothesis- Variance is not same for page view per visit and total visits

F statistic= 711.13 – High value

P-value= 1.02374e-153

```
stats.f_oneway(a,b)
```

```
F_onewayResult(statistic=711.1349623282032, pvalue=1.0274268487873719e-153)
```

Inference-As the p value is significant , we reject the null and say that Variance is not same for page view per visit and total visits

5) Total time spend on website vs total visits

Null Hypothesis- Variance is same for total time spend on website and total visits

Alternate Hypothesis- Variance is not same for total time spend on website and total visits.

F statistic- 7022.911

P-value-0.0

```
stats.f_oneway(a,b)
```

```
F_onewayResult(statistic=7022.911192999592, pvalue=0.0)
```

Inference- We reject the null and say that variance for these both features is not same.

6) Total time spend on website vs page view per visit

Null Hypothesis- Variance is same for total time spend on website and page view per visit.

Alternate Hypothesis- Variance is not same for total time spend on website and page view per visit.

```
stats.f_oneway(a,b)
```

```
F_onewayResult(statistic=7050.156327619726, pvalue=0.0)
```

Inference-

As the P-value is nearly 0 so we reject the null and say that variance for these both features is not same.

As we covered the hypothesis we got an overall view of the features we want and feature selection part as well

We will now work on model building which will include feature selection and feature engineering as well , the models built will keep on increasing the accuracy and we keep minimizing the features as per the feature selection part.

LOGISTIC REGRESSION

- **Label Encoding**

The features provided in the dataset have been transformed to encoded labels , the target variable , 'converted' has been encoded to 0 and 1 , linear modelling , other features like Do not email , and do not call has been transformed as well

```
binary_cat_vars = ["Do Not Email","Converted"]

def binary_map(x):
    return x.map({"Yes" : 1, "No" : 0})


df[binary_cat_vars] = df[binary_cat_vars].apply(binary_map)
```

Rest categorical variables that showed the impact for more than 80 % conversion has been transformed in dummies following EDA :

```
dummy = pd.get_dummies(data = df[df.select_dtypes("object").columns])
dummy.drop(["Lead Origin_API", "Lead Source_Others", "Last Activity_Other_Activity",
            "Specialization_Other_Specialization"], axis = 1, inplace = True)
df = pd.concat([df,dummy], axis = 1)
df.drop(df.select_dtypes("object").columns, axis = 1, inplace = True)
df.head()
```

Feature selection and feature engineering:

Feature were selected on the basis of the high conversion rate to start with the project , after converting to dummies there are total of 104 columns :

```
Index(['Do Not Email', 'TotalVisits', 'Total Time Spent on Website',
       'Page Views Per Visit', 'Lead Origin_Landing Page Submission',
       'Lead Origin_Lead Add Form', 'Lead Origin_Lead Import',
       'Lead Source_Olark Chat', 'Lead Source_Welingak Website',
       'Last Activity_Email Bounced', 'Last Activity_Olark Chat Conversation',
       'Last Activity_SMS Sent', 'Specialization_E-COMMERCE',
       'What is your current occupation_Student',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional',
       'Tags_Already a student', 'Tags_Busy', 'Tags_Closed by Horizzon',
       'Tags_Diploma holder (Not Eligible)', 'Tags_Graduation in progress',
       'Tags_Interested  in full time MBA', 'Tags_Interested in other courses',
       'Tags_Lateral student', 'Tags_Lost to EINS', 'Tags_Lost to Others',
       'Tags_Not doing further education', 'Tags_Ringing',
       'Tags_Still Thinking',
       'Tags_Want to take admission but has financial problems',
       'Tags_Will revert after reading the email', 'Tags_in touch with EINS',
       'Tags_invalid number', 'Tags_number not provided', 'Tags_opp hangup',
       'Tags_switched off', 'Tags_wrong number given',
       'Lead Quality_High in Relevance', 'Lead Quality_Low in Relevance',
       'Lead Quality_Might be', 'Lead Quality_Not Sure', 'Lead Quality_Worst',
       'City_Tier II Cities', 'Last Notable Activity_Email Bounced',
       'Last Notable Activity_Email Link Clicked',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_Modified',
       'Last Notable Activity_Olark Chat Conversation',
       'Last Notable Activity_Page Visited on Website',
       'Last Notable Activity_SMS Sent', 'Last Notable Activity_Unreachable',
       'Last Notable Activity_Unsubscribed'],
      dtype='object')
```

After Using RFE(Recursive feature elimination) these were the columns that were remaining

Do Not Email	-0.9942	0.249	-3.992	0.000	-1.482	-0.506
TotalVisits	1.0153	0.247	4.118	0.000	0.532	1.498
Total Time Spent on Website	4.5471	0.212	21.494	0.000	4.132	4.962
Page Views Per Visit	-0.8580	0.260	-3.305	0.001	-1.367	-0.349
Lead Origin_Landing Page Submission	-0.5035	0.114	-4.403	0.000	-0.728	-0.279
Lead Origin_Lead Add Form	1.7292	0.360	4.803	0.000	1.024	2.435
Lead Origin_Lead Import	0.8361	0.689	1.214	0.225	-0.514	2.186
Lead Source_Olark Chat	0.8640	0.160	5.384	0.000	0.549	1.179
Lead Source_Welingak Website	3.5722	0.799	4.473	0.000	2.007	5.137
Last Activity_Email Bounced	-0.8869	0.611	-1.452	0.146	-2.084	0.310
Last Activity_Olark Chat Conversation	-0.7406	0.225	-3.290	0.001	-1.182	-0.299
Last Activity_SMS Sent	0.9044	0.187	4.833	0.000	0.538	1.271
Specialization_E-COMMERCE	0.6471	0.389	1.665	0.096	-0.115	1.409
What is your current occupation_Student	-1.0177	0.979	-1.040	0.298	-2.936	0.900
What is your current occupation_Unemployed	-1.2257	0.888	-1.380	0.167	-2.966	0.515
What is your current occupation_Working Professional	0.2217	0.924	0.240	0.810	-1.590	2.033
Tags_Already a student	-2.2275	1.126	-1.978	0.048	-4.435	-0.020
Tags_Busy	2.0679	0.940	2.201	0.028	0.226	3.910
Tags_Closed by Horizzon	6.7569	1.192	5.668	0.000	4.421	9.093
Tags_Diploma holder (Not Eligible)	-2.5396	1.418	-1.792	0.073	-5.318	0.239
Tags_Graduation in progress	-1.1962	1.048	-1.142	0.254	-3.250	0.857
Tags_Interested in full time MBA	-2.0131	1.163	-1.731	0.083	-4.292	0.266
Tags_Interested in other courses	-2.1624	0.973	-2.221	0.026	-4.070	-0.254
Tags_Lateral student	16.9574	3701.704	0.005	0.996	-7238.250	7272.164
Tags_Lost to EINS	7.1094	1.063	6.686	0.000	5.025	9.194
Tags_Lost to Others	-26.8921	1.7e+05	-0.000	1.000	-3.32e+05	3.32e+05
Tags_Not doing further education	-3.1491	1.471	-2.140	0.032	-6.033	-0.265
Tags_Ringing	-2.9746	0.943	-3.153	0.002	-4.824	-1.126
Tags_Still Thinking	-1.6991	1.683	-1.010	0.313	-4.998	1.599
Tags_Want to take admission but has financial problems	-1.7690	1.393	-1.270	0.204	-4.499	0.961

As we see here "Tags_Lateral_Student" has P-value of 0.96 and is not significant , so we dropped the feature .

On the basis of Variation Inflation factor we had 4 models created step wise and analysed the VIF created with that , As the Number of columns were more we had a lot of non linearity and multi-collinearity which was analysed by the VIF as well

	Features	VIF
38	Lead Quality_Not Sure	432.86
37	Lead Quality_Might be	114.15
28	Tags_Will revert after reading the email	89.11
39	Lead Quality_Worst	46.17
35	Lead Quality_High in Relevance	45.07
25	Tags_Ringing	42.80
36	Lead Quality_Low in Relevance	40.87
14	What is your current occupation_Unemployed	26.28
15	What is your current occupation_Working Profes...	20.81
22	Tags_Interested in other courses	20.43
16	Tags_Already a student	19.40
18	Tags_Closed by Horizzon	12.91
33	Tags_switched off	10.47
17	Tags_Busy	8.35

To solve the effect by using Generalised linear model , so improve the linearity , P value and VIF for the features

After 7 Iterations we got our model , which is significant in terms of P value and VIF factor as well

Model:	GLM	AIC:	4105.7739
Link Function:	logit	BIC:	-78466.1928
Dependent Variable:	Converted	Log-Likelihood:	-2035.9
Date:	2022-02-26 10:16	LL-Null:	-6019.3
No. Observations:	9074	Deviance:	4071.8
Df Model:	16	Pearson chi2:	3.09e+04
Df Residuals:	9057	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-3.2119	0.2073	-15.4956	0.0000	-3.6181	-2.8056
Do Not Email	-1.2222	0.1832	-6.6728	0.0000	-1.5812	-0.8632
Total Time Spent on Website	3.7172	0.1723	21.5762	0.0000	3.3796	4.0549
Lead Origin_Lead Add Form	1.9226	0.3190	6.0277	0.0000	1.2975	2.5478
Lead Source_Welingak Website	3.4912	0.7931	4.4017	0.0000	1.9366	5.0457
What is your current occupation_Working Professional	1.3574	0.2566	5.2901	0.0000	0.8545	1.8603
Tags_Busy	3.7833	0.2774	13.6387	0.0000	3.2396	4.3270
Tags_Closed by Horizzon	8.0483	0.7629	10.5491	0.0000	6.5530	9.5436
Tags_Lost to EINS	8.5780	0.5548	15.4618	0.0000	7.4906	9.6653
Tags_Ringing	-1.2900	0.2850	-4.5271	0.0000	-1.8485	-0.7315
Tags_Will revert after reading the email	3.8391	0.2025	18.9620	0.0000	3.4422	4.2359
Tags_invalid number	-1.9089	1.1526	-1.6561	0.0977	-4.1680	0.3502
Tags_switched off	-2.3259	0.5786	-4.0203	0.0001	-3.4599	-1.1920
Lead Quality_High in Relevance	1.1978	0.2824	4.2416	0.0000	0.6443	1.7513
Lead Quality_Not Sure	-2.8841	0.1167	-24.7132	0.0000	-3.1128	-2.6553
Lead Quality_Worst	-3.2006	0.6406	-4.9964	0.0000	-4.4562	-1.9451
Last Notable Activity_SMS Sent	2.5824	0.1048	24.6324	0.0000	2.3769	2.7879

Inferences-

As per looking at the AIC for a large dataset , the Fit is good for this Model

Also the BIC criteria is very low , that means the model is a good fit

The constant and the other features are significant for the model

Tags_invalid Number is kept as the model can be used in the 90% confidence interval as well due to the research requirement

GLM fit better than the Logit model

Train Data

- Accuracy: 94%

Sensitivity-Specificity Approach

- Sensitivity: 93%
- Specificity: 94%

Precision-Recall Approach

- Precision: 92%
- Recall: 92%

Test Data

- Accuracy: 94%

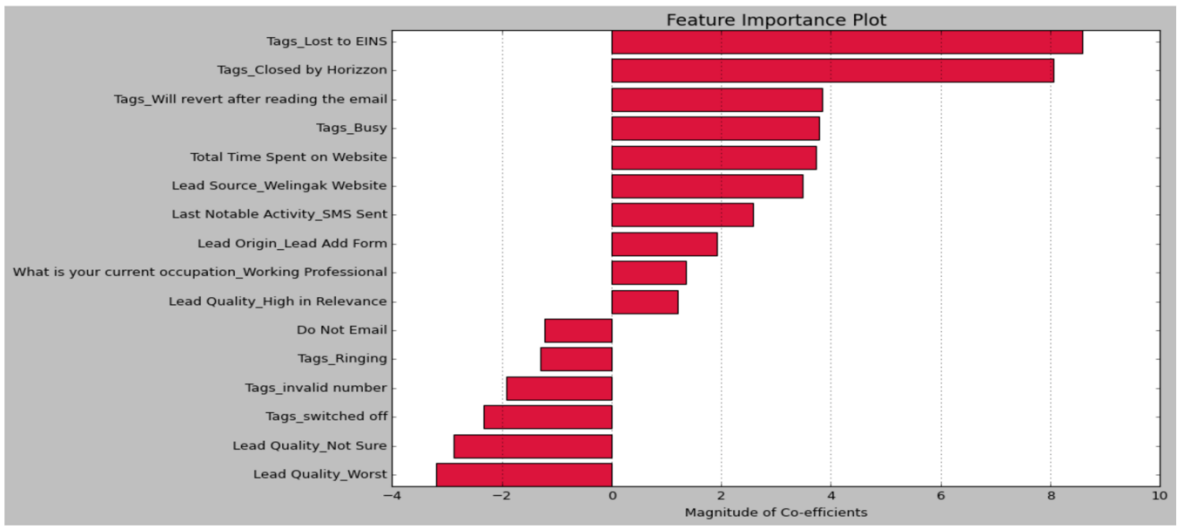
Sensitivity-Specificity Approach

- Sensitivity: 93%
- Specificity: 94%

Precision-Recall Approach

- Precision: 91%
- Recall: 91%

Feature Importance



Important variables as per the model:

1. Tags
    - Busy
    - Closed by Horizzon
    - Lost to EINS
    - Not Specified
    - Ringing
    - switched off
2. Total Time Spent on Website
3. Lead Source
    - Welingak Website
4. What is your current occupation
    - Student
    - Working Professional
5. Page Views Per Visit