Lead Scoring Case Study

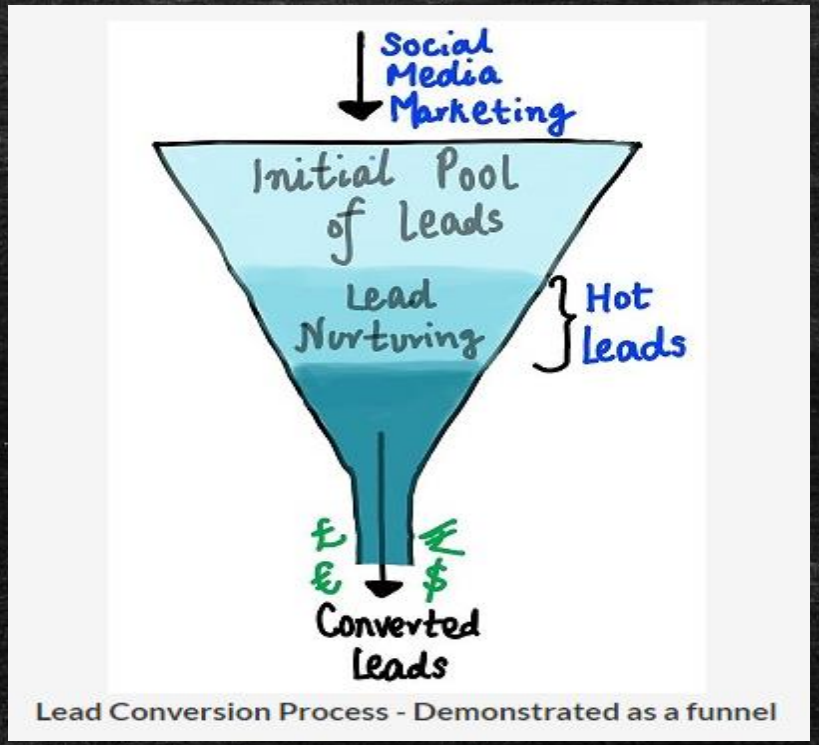
Authors: Ashutosh Kumar, Archit Bhandari

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.

Problem Statement

- A typical lead conversion process can be represented using the following funnel:



Lead Conversion Process - Demonstrated as a funnel

- As you can see, 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.

Problem Statement

- X Education has appointed you to help them 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 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%.

- **Business Goal:**

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Analysis Approach for the Problem in Hand

- We will be using the Logistic Regression Machine Learning algorithm to create model with higher conversion rate followed by recommendations and analysis.
- The overall approach will be as follows:

1. Data Preparation : Data loading, Data Inspection, Data Preparation as part of EDA.
2. Dividing the data into test and training data for our model.
3. Feature Scaling and Normalization of Data.
4. Model building using machine learning algorithm.
5. Model Evaluation based on various metrices.
6. Recommendations and Lead score assignment for the current data.

Data Inspection

- So there are 9240 data points(rows) in our dataset and 37 distinct features(columns).

```
print(main_lead_df.shape)
```

```
(9240, 37)
```

- Below are different columns and their data-type in our dataset:

```
Prospect ID                                      object
Lead Number                                       int64
Lead Origin                                      object
Lead Source                                      object
Do Not Email                                     object
Do Not Call                                      object
Converted                                         int64
TotalVisits                                     float64
Total Time Spent on Website                       int64
Page Views Per Visit                            float64
Last Activity                                    object
Country                                          object
Specialization                                   object
How did you hear about X Education               object
What is your current occupation                  object
What matters most to you in choosing a course    object
Search                                           object
Magazine                                         object
Newspaper Article                                object
X Education Forums                               object
Newspaper                                        object
Digital Advertisement                            object
Through Recommendations                          object
Receive More Updates About Our Courses           object
Tags                                             object
Lead Quality                                     object
Update me on Supply Chain Content                object
Get updates on DM Content                        object
Lead Profile                                     object
City                                             object
Asymmetrique Activity Index                      object
Asymmetrique Profile Index                       object
Asymmetrique Activity Score                     float64
Asymmetrique Profile Score                      float64
I agree to pay the amount through cheque         object
A free copy of Mastering The Interview           object
Last Notable Activity                            object
dtype: object
```

Data Inspection : Missing value analysis

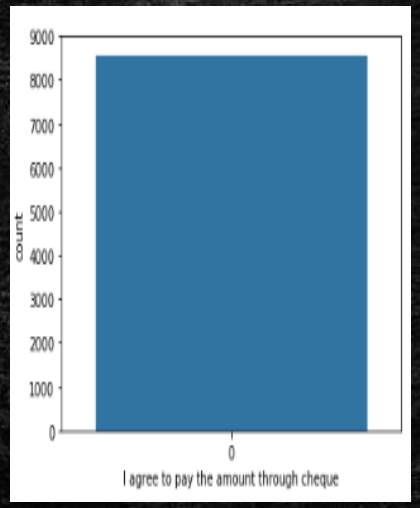
- So we have 17 columns out of 37 in which at least one of the value is missing:

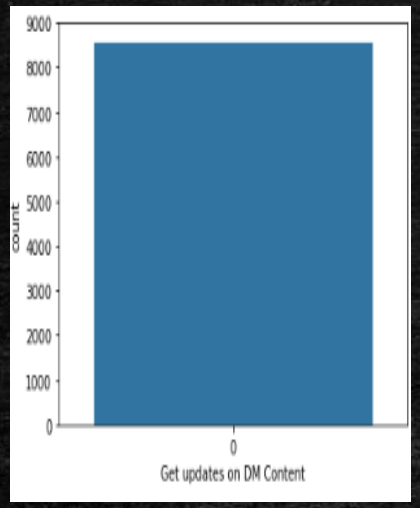
	Total	Percent
Lead Quality	4767	51.59
Asymmetrique Profile Score	4218	45.65
Asymmetrique Activity Score	4218	45.65
Asymmetrique Profile Index	4218	45.65
Asymmetrique Activity Index	4218	45.65
Tags	3353	36.29
What matters most to you in choosing a course	2709	29.32
Lead Profile	2709	29.32
What is your current occupation	2690	29.11
Country	2461	26.63
How did you hear about X Education	2207	23.89
Specialization	1438	15.56
City	1420	15.37
TotalVisits	137	1.48
Page Views Per Visit	137	1.48
Last Activity	103	1.11
Lead Source	36	0.39

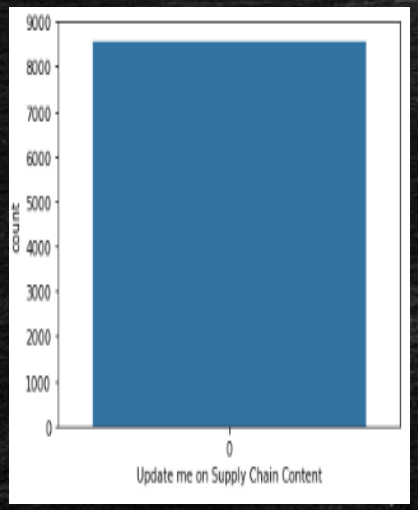
Data Preparation and Understanding

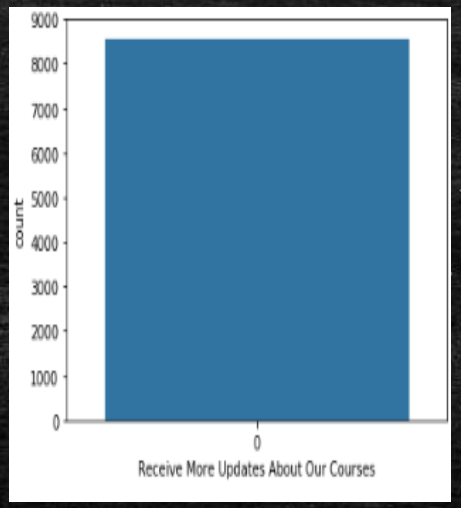
- Since all the values are 0 (No),no body seems to prefer it, so the below columns are of hardly any significance and can be dropped. These are binary columns which are having only 1 negative response in the dataset:

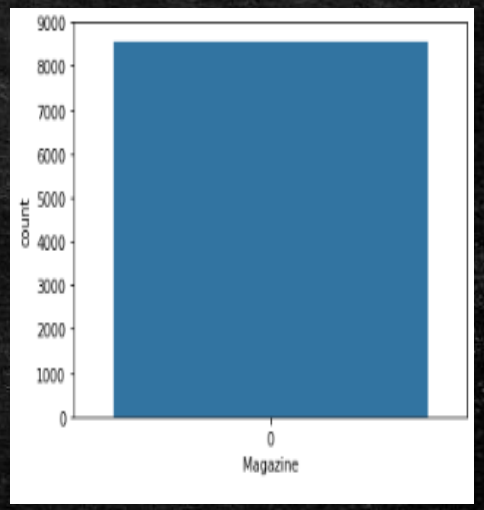
Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque







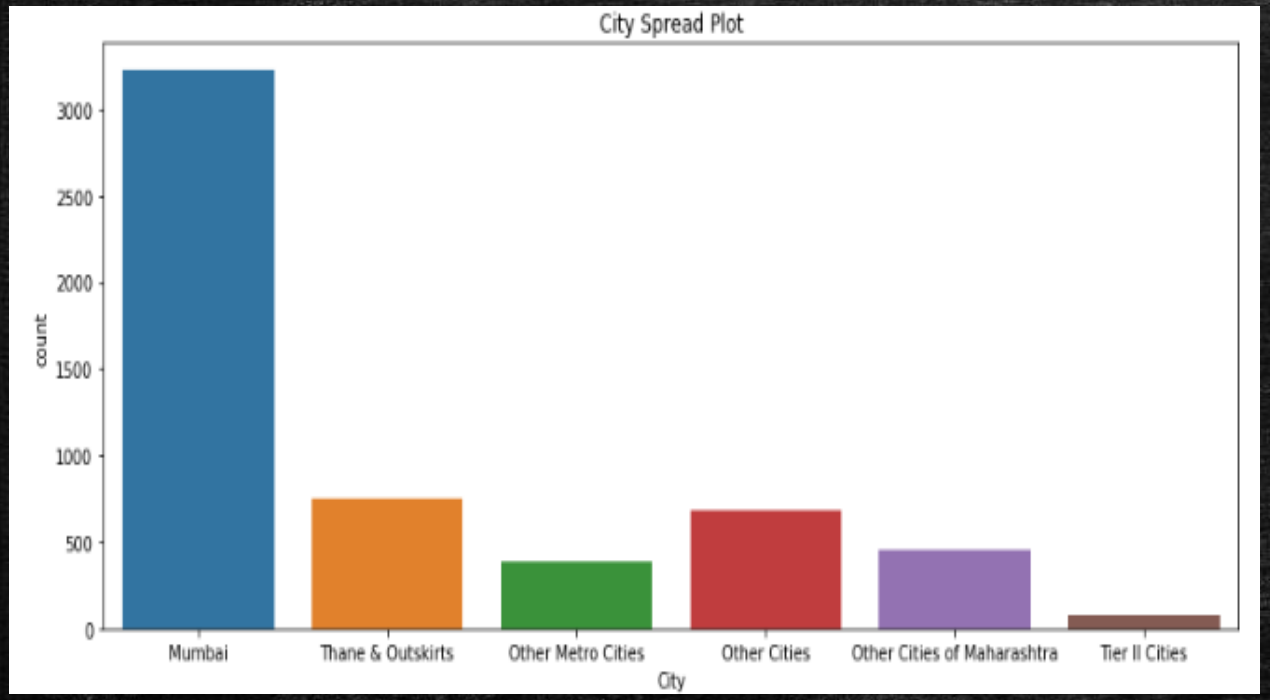


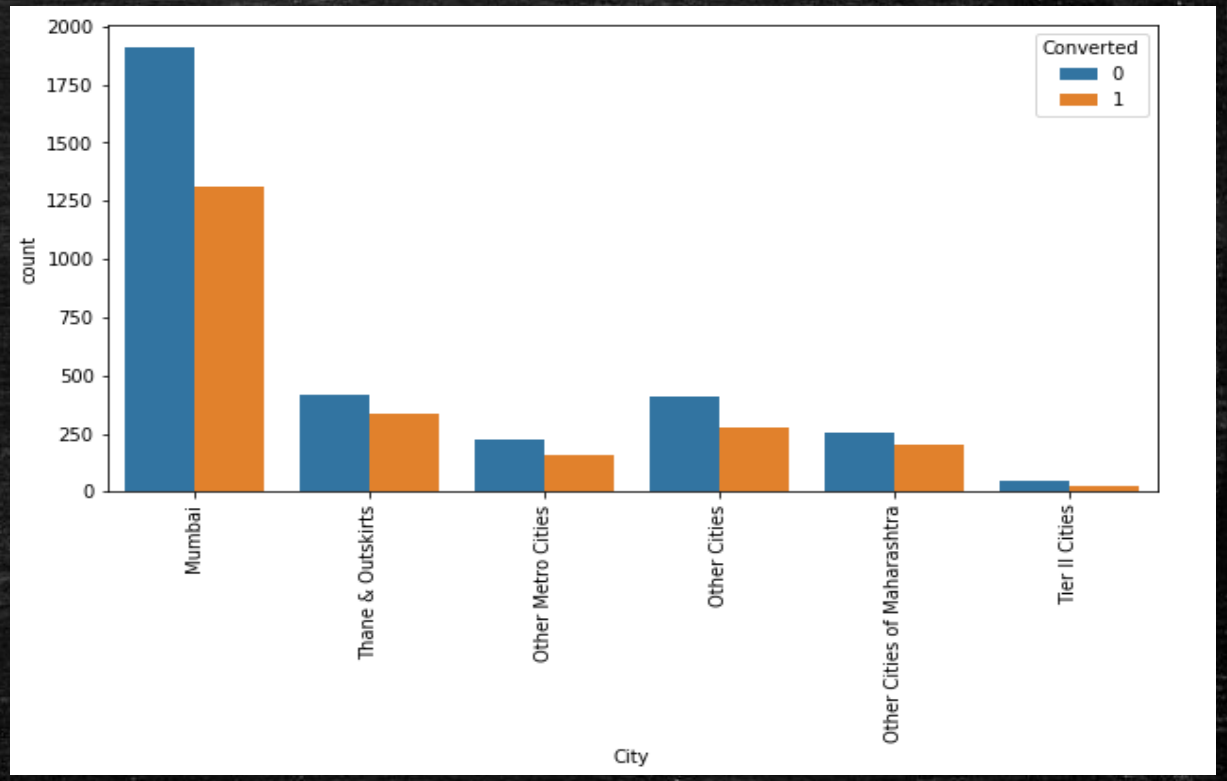


So we have 5 columns eliminated here as part of EDA itself

Univariate and Bivariate Analysis of Data

Variable selected : **City** : The city of the customer.

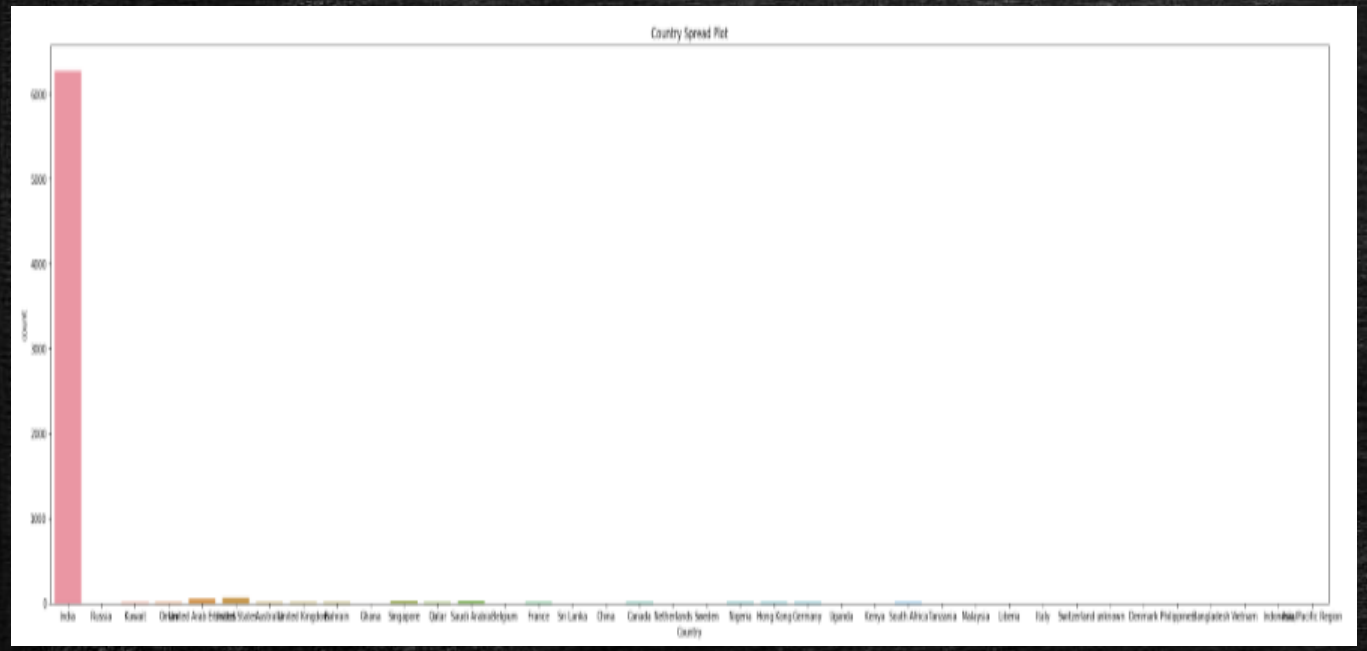


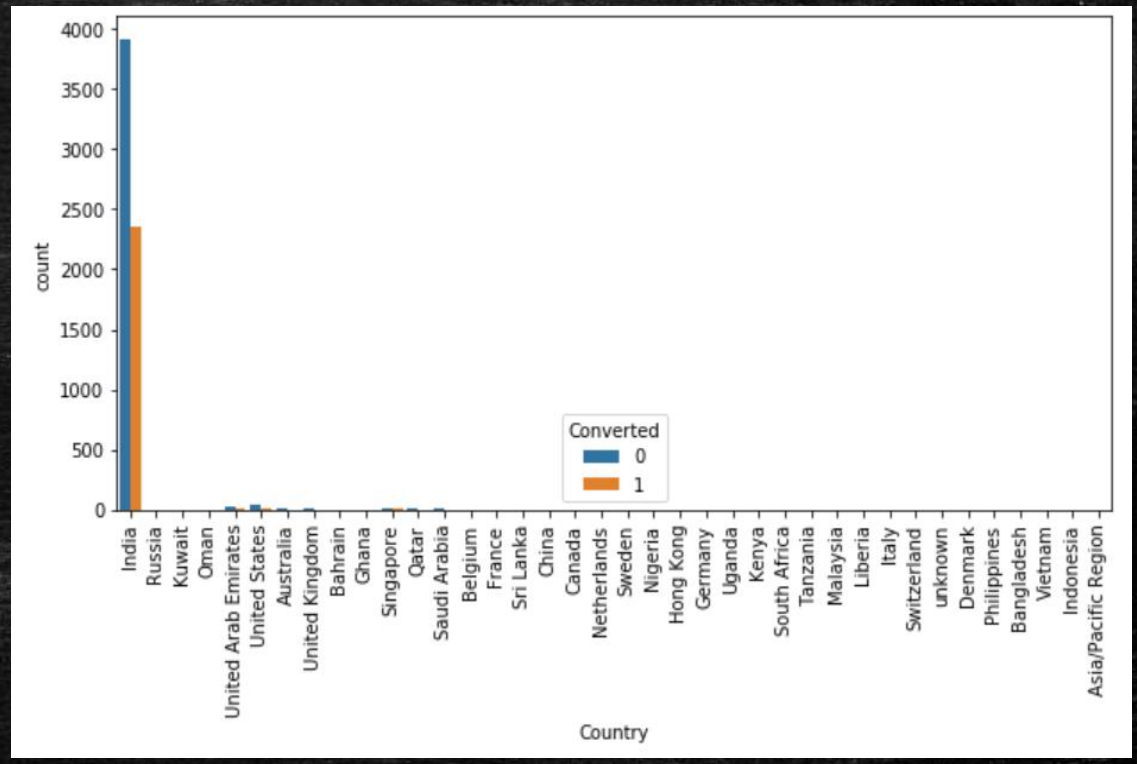


As we can see from above plots city Mumbai is highest for each column.
Also most conversions are from Mumbai city again with the conversion failure rate higher than success across all locations.

Univariate and Bivariate Analysis of Data

- Variable selected : **Country** : The country of the customer.

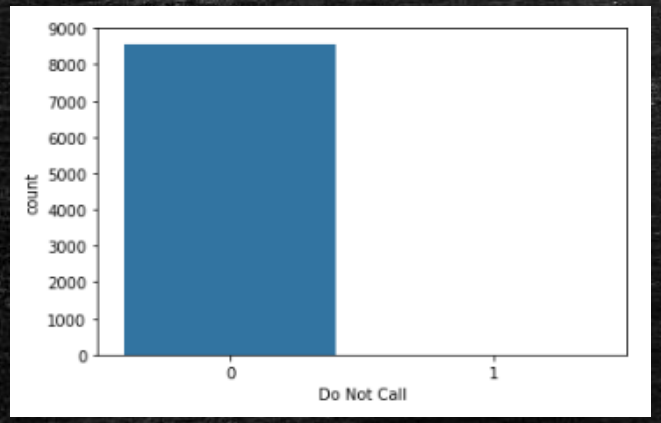




As we can see from above plots country India is highest for each column. But since we know from the business understanding that the location information is not going to impact our result/prediction as the output variable is for the online education portal , so lets drop them.

Univariate and Bivariate Analysis of Data

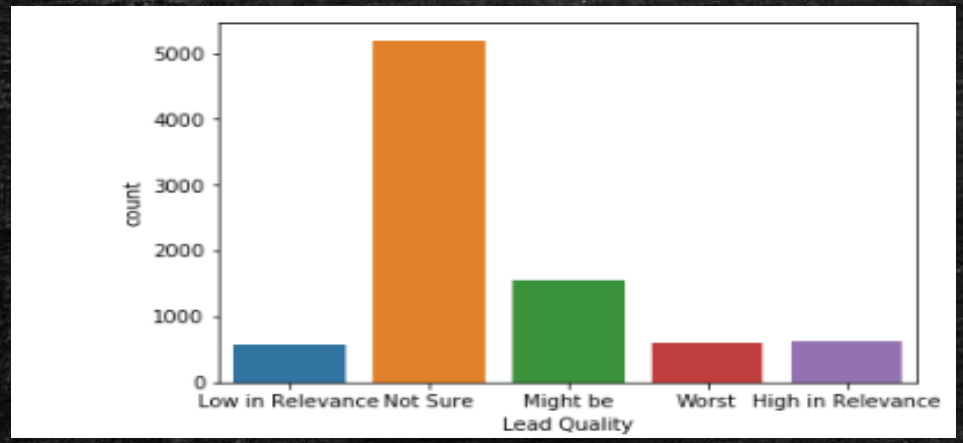
- Variable selected : **Do Not Call: An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.**



Since most of the values are o(No), so we can drop this column also.

Univariate and Bivariate Analysis of Data

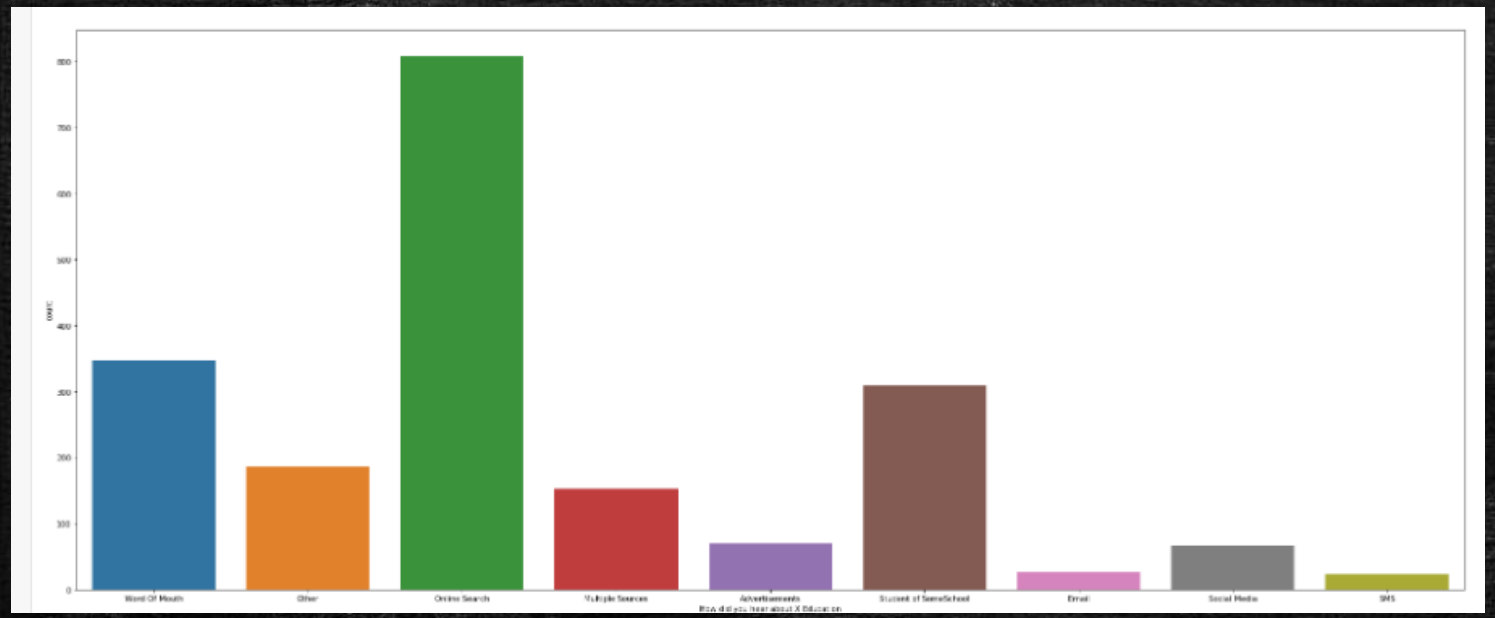
- Variable selected : **Lead Quality: Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead**

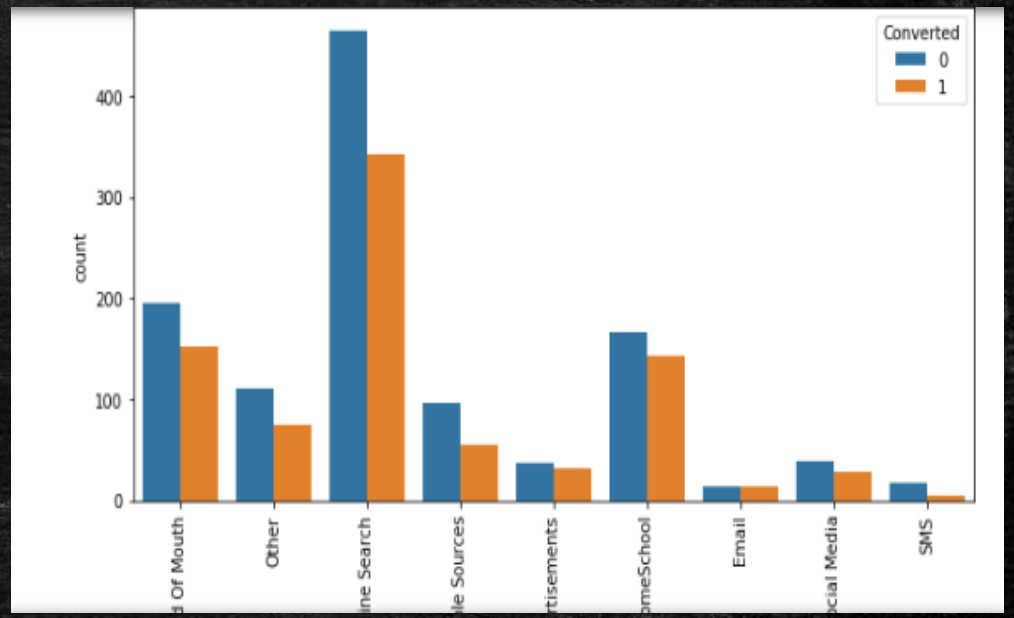


As most of the data point for this column is might be and Not sure and also since its based on the sales person intuition, this information can be dropped as its of hardly any statistical/business significance

Univariate and Bivariate Analysis of Data

- Variable selected :**How did you hear about X Education:** The source from which the customer heard about X Education.

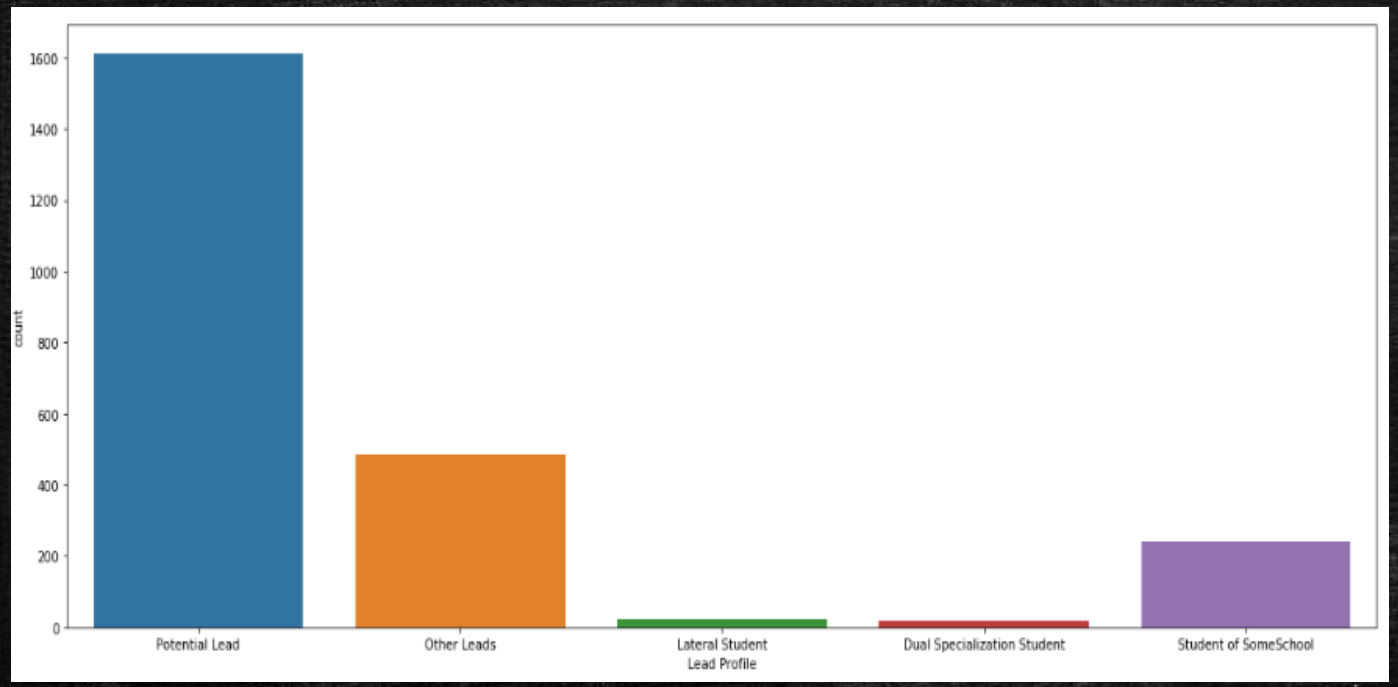


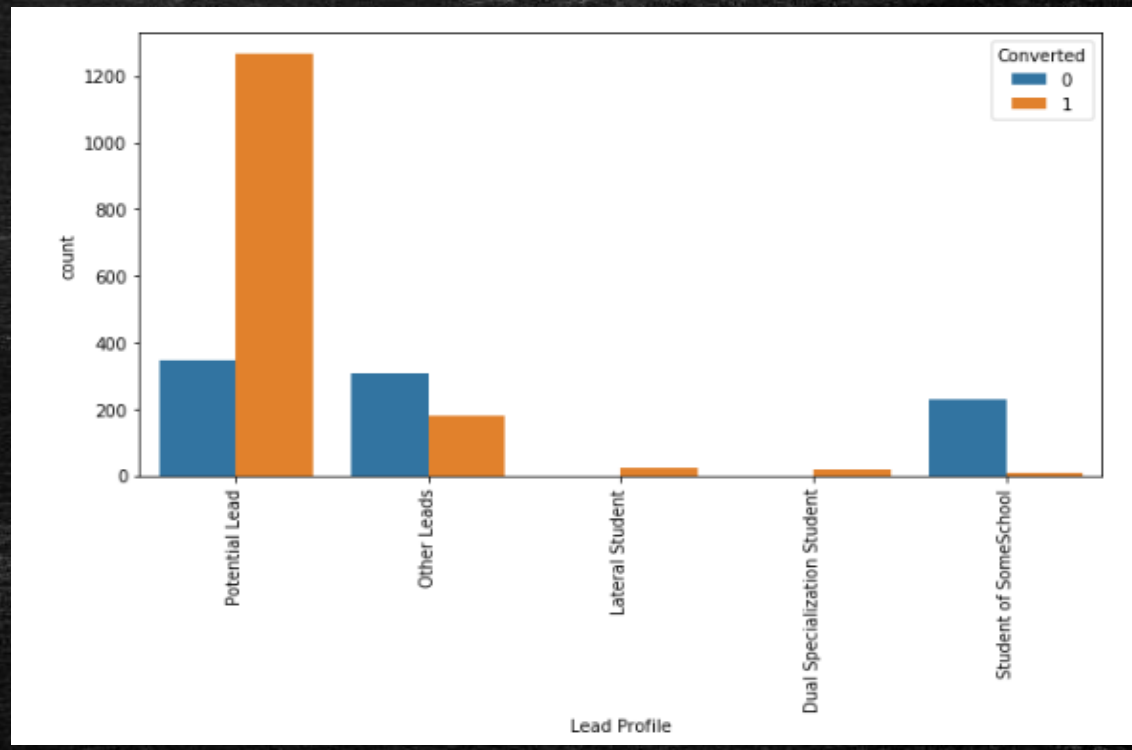


As around 77 % of the values are missing for this field and impact on the target variable is constant across the sources we can drop this variable.

Univariate and Bivariate Analysis of Data

- Variable selected :**Lead Profile:** A lead level assigned to each customer based on their profile.

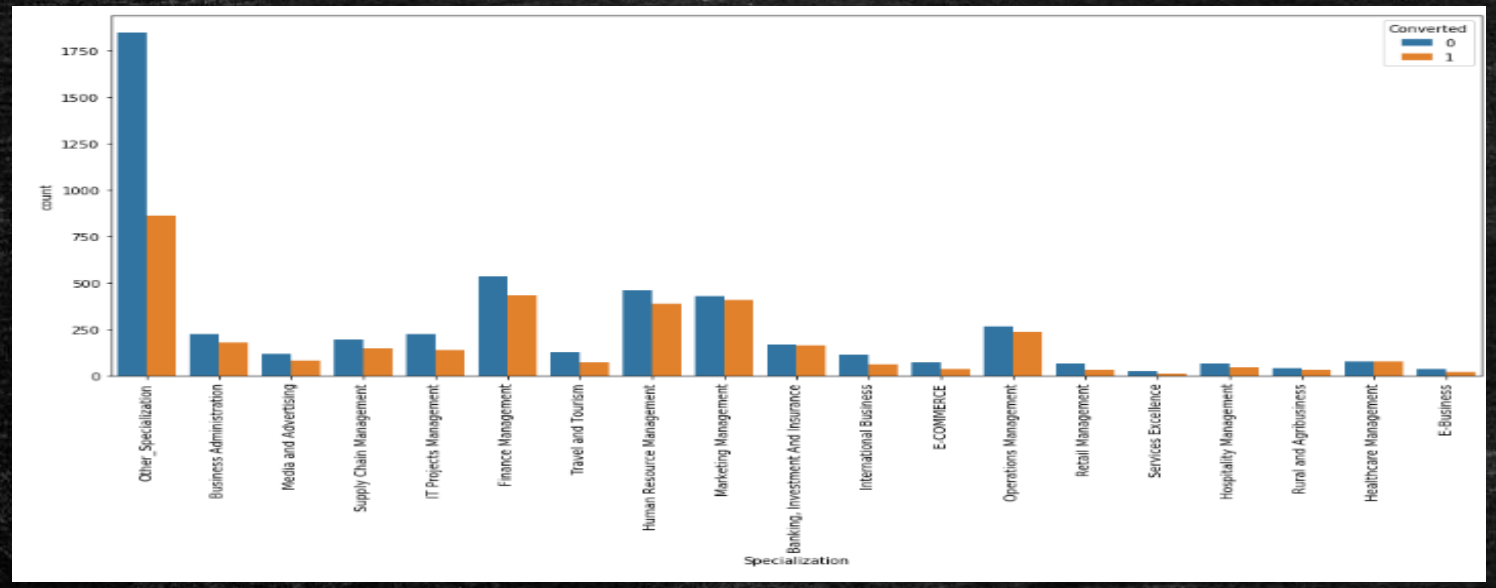


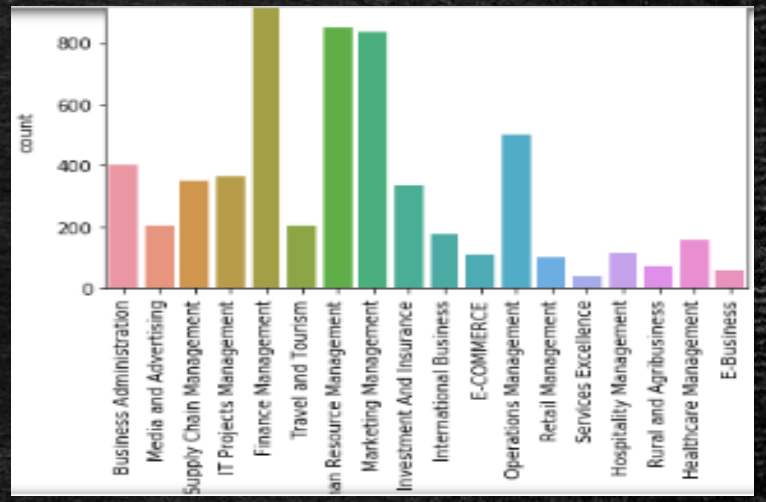


We can see conversion ratio is higher for the "Potential Lead" but we can drop this column as we have around 75 percent of missing values.

Univariate and Bivariate Analysis of Data

- Variable selected :**Specialization**: The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.

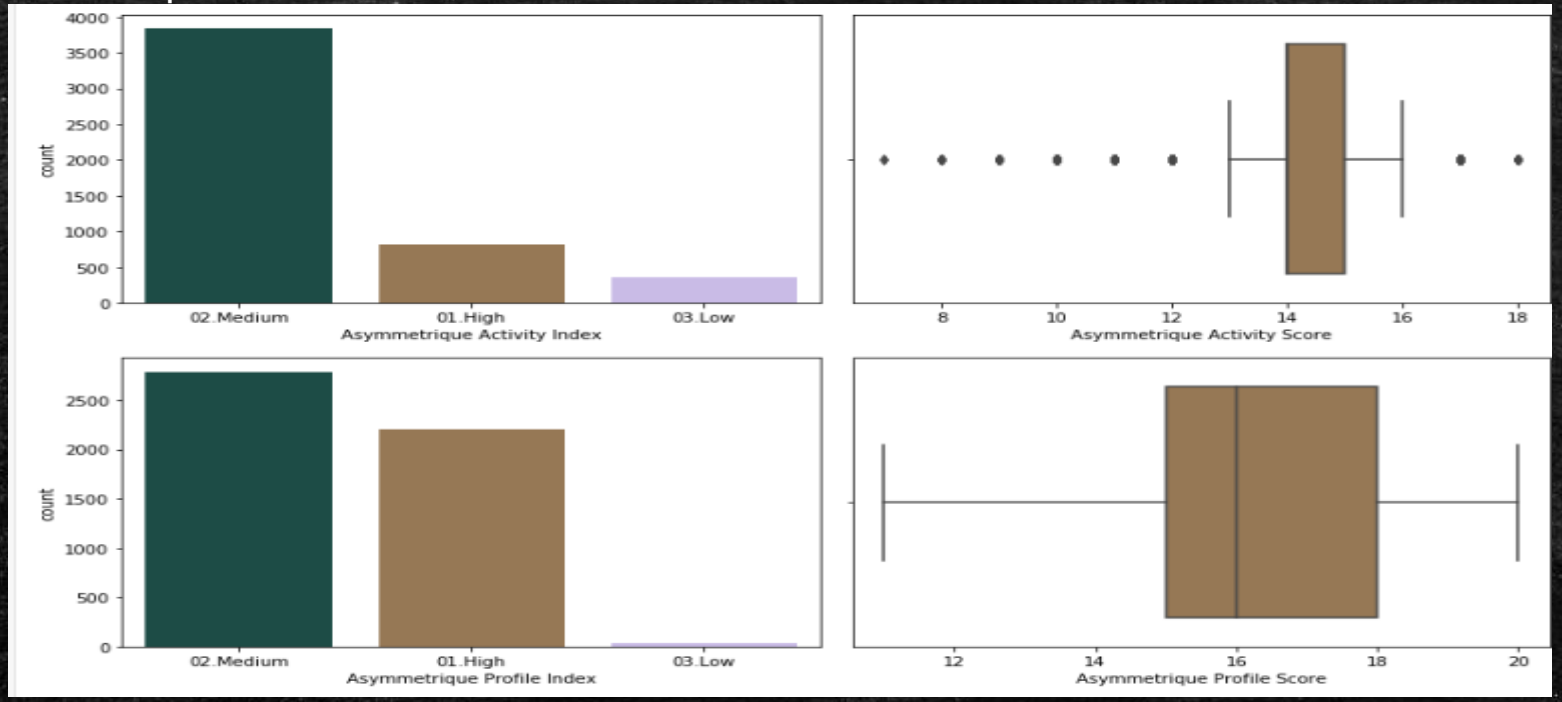




So Finance management is top specialization followed by other.
We know that from business point of view this is an important feature and we have around 31 % missing values.
The missing values could be because either the candidate is a student or his choice of profession was not available in the options available on the portal.
Lets create a new option for the missing values called Others where it is missing and continue with other variables.

Univariate and Bivariate Analysis of Data

Variable selected :**Asymmetrique Activity Index, Asymmetrique Activity Score, Asymmetrique Profile Index, Asymmetrique Profile Score** : An index and score assigned to each customer based on their activity and their profile.

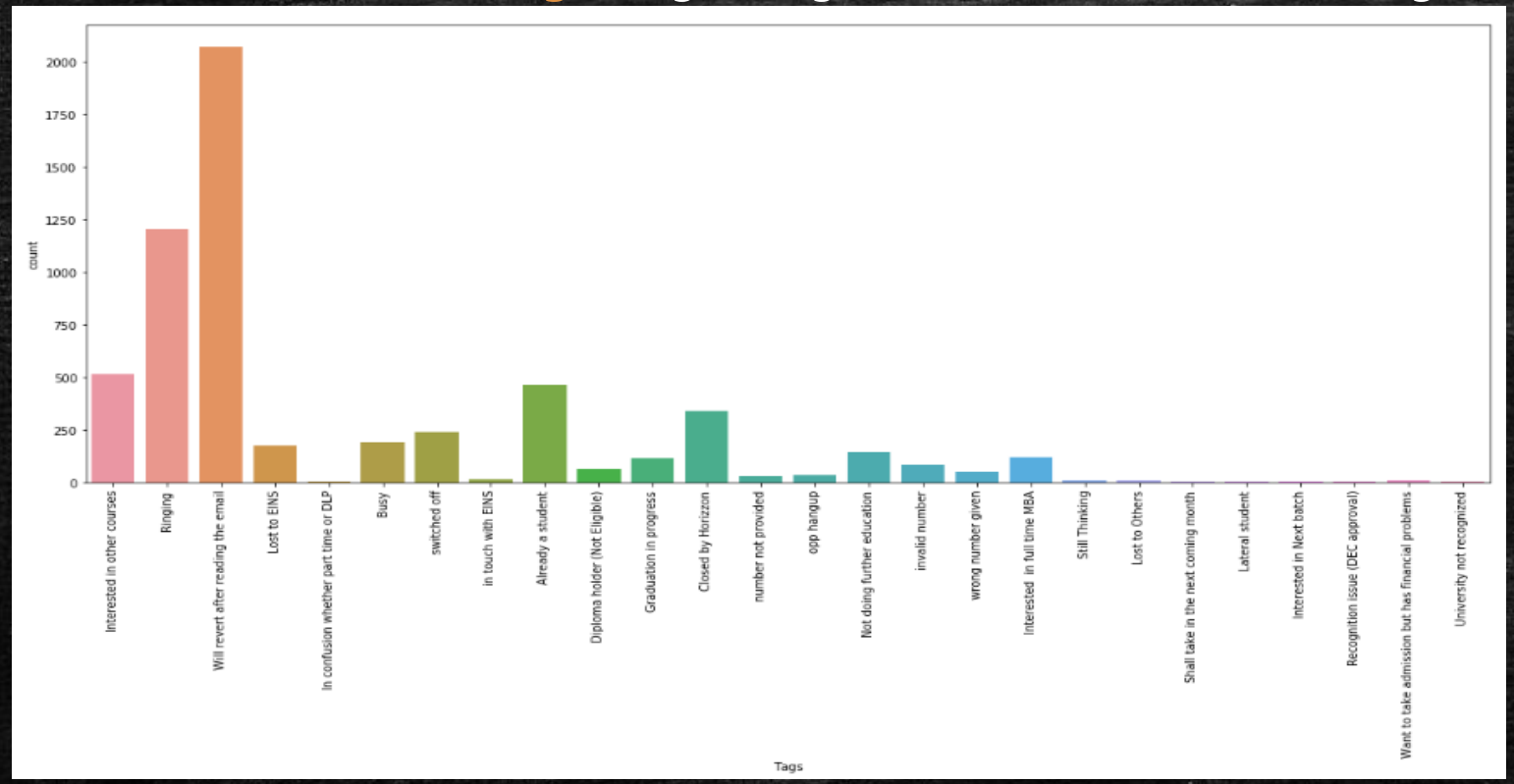


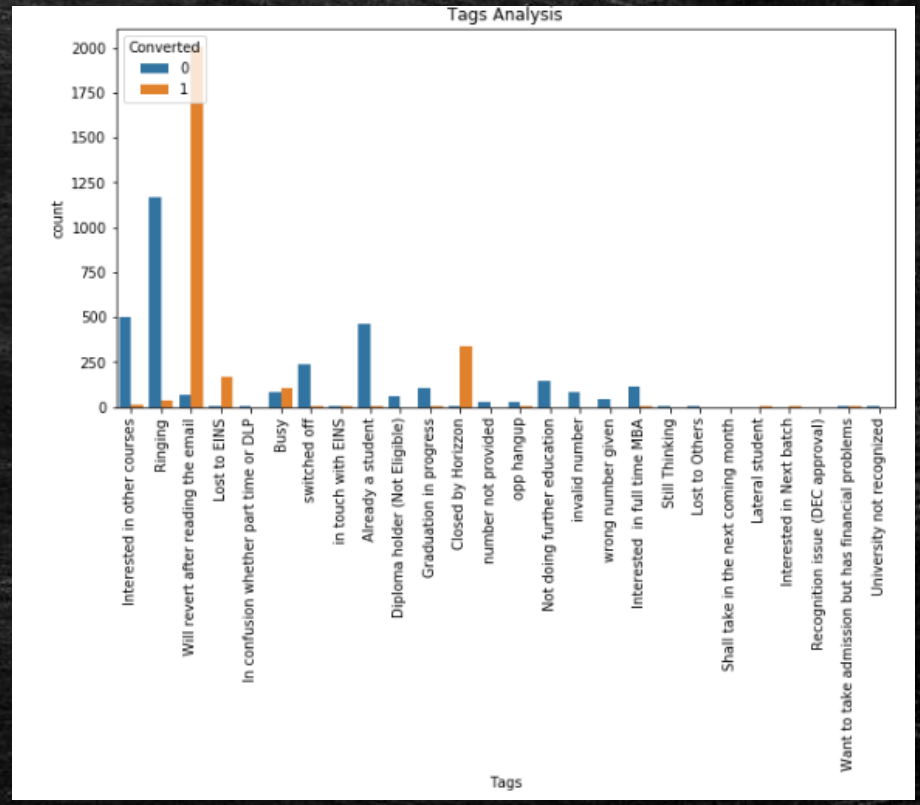
Around 42 percent of values are missing for these 4 features and the feature does not seems to be very reliable as there are many outliers , hence too much variation in the data.

So we will go ahead and drop these 4 columns.

Univariate and Bivariate Analysis of Data

Variable selected : **Tags:** Tags assigned to customers indicating the current status of the lead.

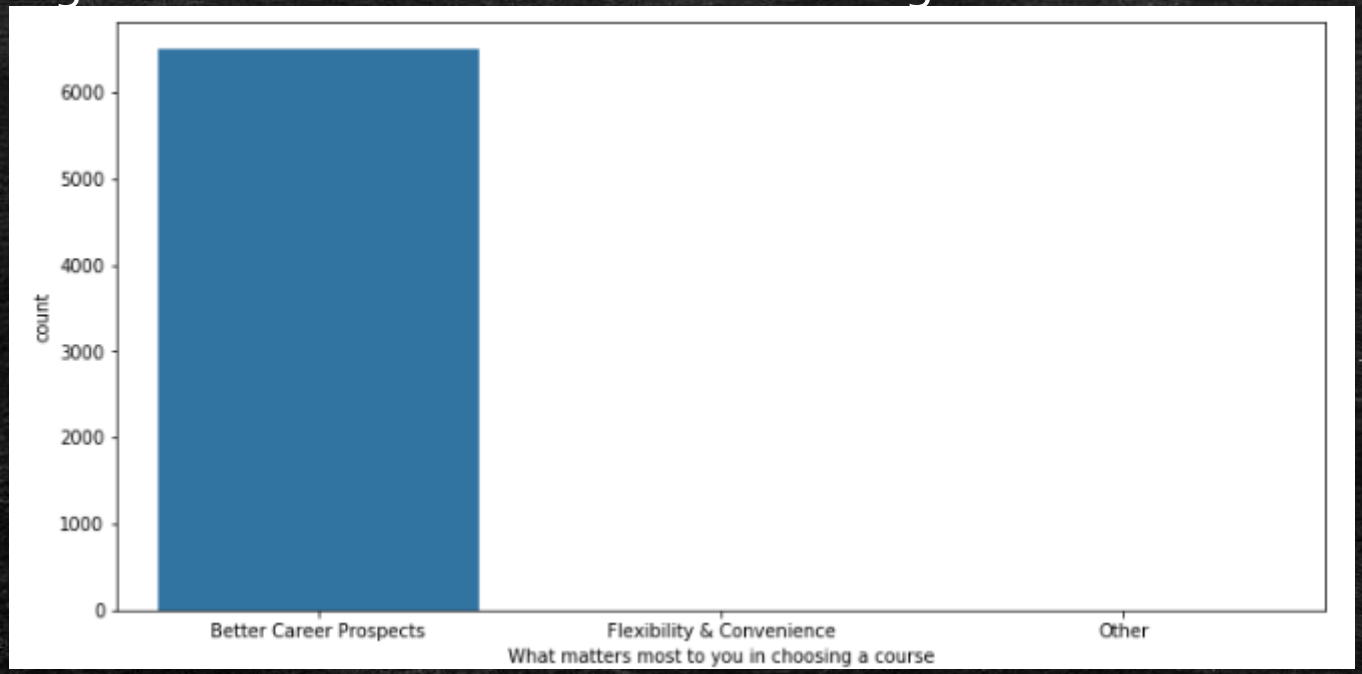




So although around 32 percent of values are missing, since its important feature for our problem we can not drop it.
Also most number and ratio of conversion is from "Will revert after reading the email" tag.
Also we can see that this categorical column has around 2071 value out of 5864 data points, so we will impute this with mode.

Univariate and Bivariate Analysis of Data

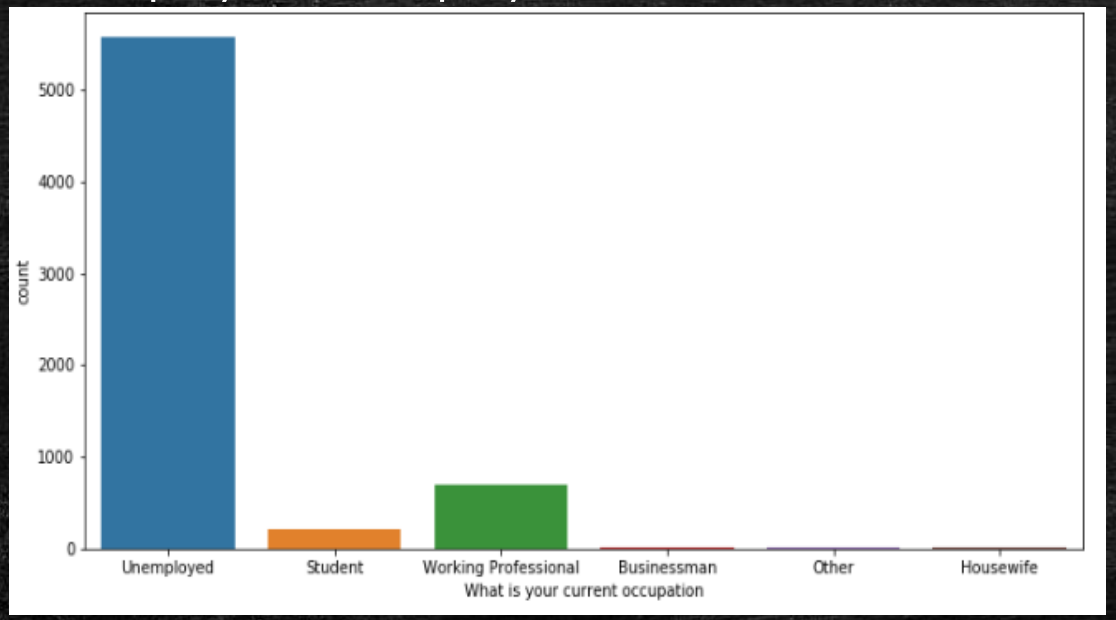
Variable selected :**What matters most to you in choosing a course** : An option selected by the customer indicating what is their main motto behind doing this course.

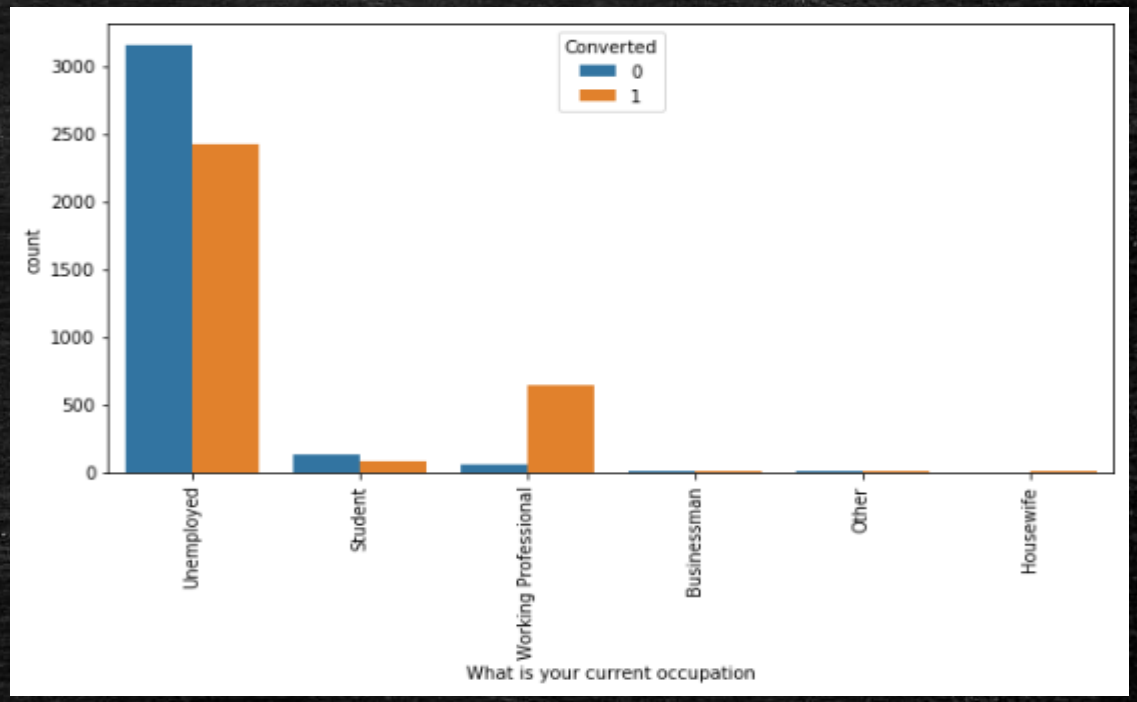


Around 75 % of the data is having Better Career Prospects as What matters most to you in choosing a course.
But around 24 % of data is missing for the column, so there might be some factor which was not in option while applicant was entering details, we will impute the same with Others

Univariate and Bivariate Analysis of Data

Variable selected :**What is your current occupation:** Indicates whether the customer is a student, unemployed or employed.

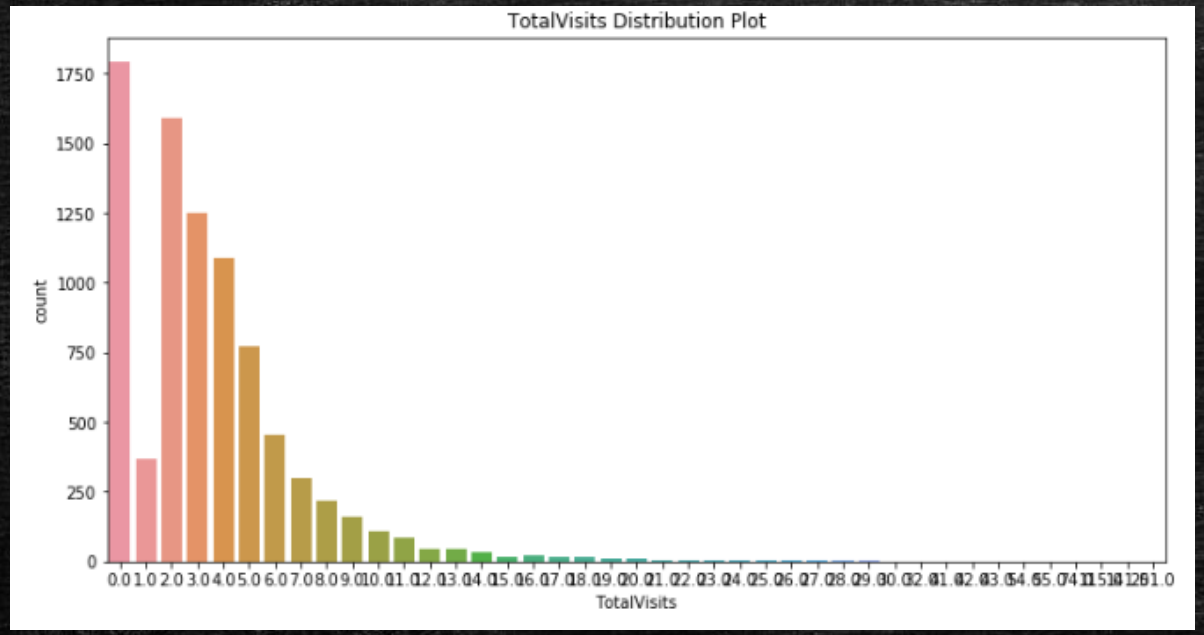


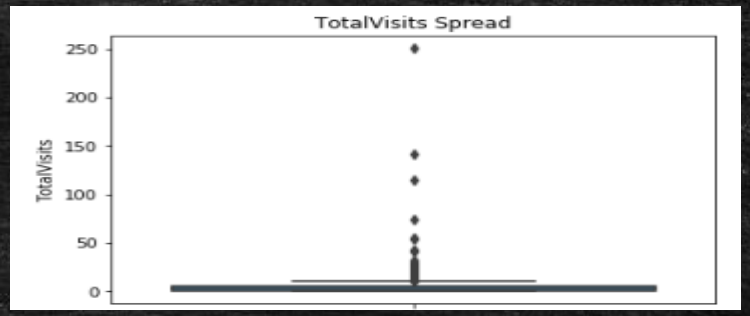


Most of the applicants(more than 80%) with our data are Unemployed and we have around 24 % of data missing here. Also the highest conversion rate is for the working professionals.

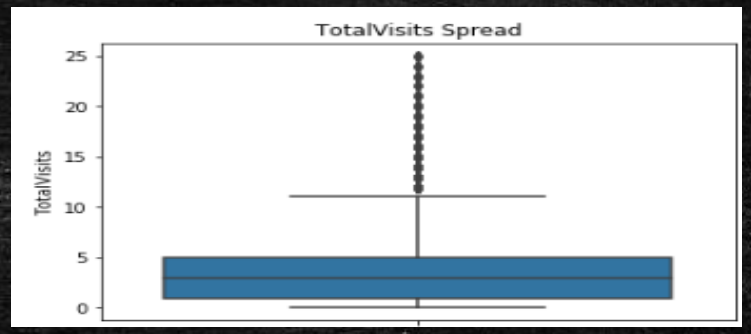
Univariate and Bivariate Analysis of Data

Variable selected :**TotalVisits:** The total number of visits made by the customer on the website.





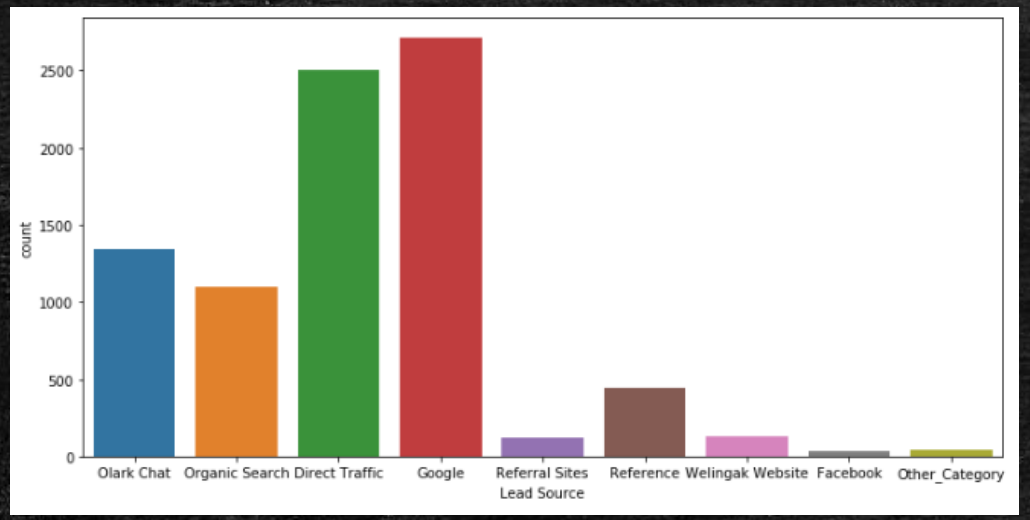
Re-checking the data after outlier treatment:

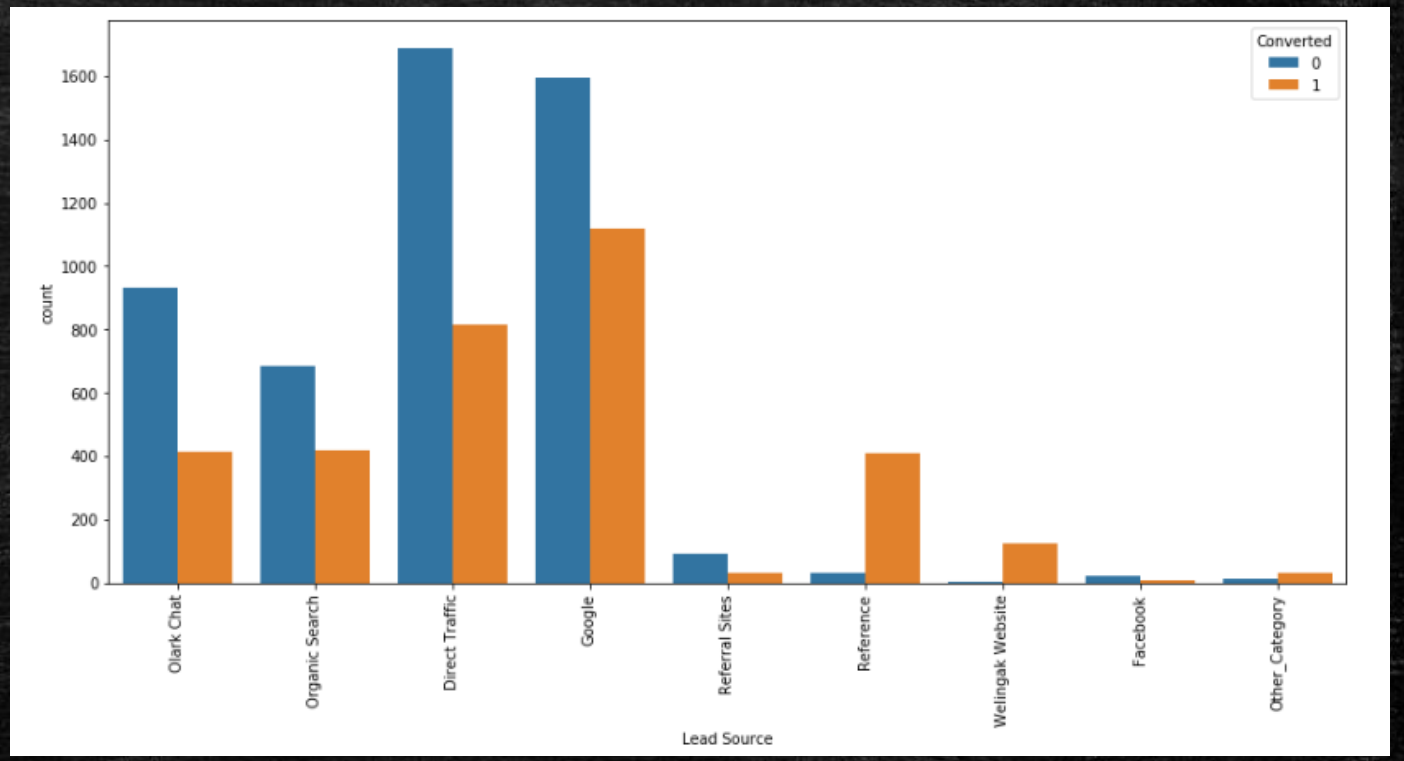


As we know that we have around 1.32 % missing values for this field and since the distribution looks fairly normal and it's a numerical variable , we will replace the same with mean.

Univariate and Bivariate Analysis of Data

Variable selected :**Lead Source:** The source of the lead. Includes Google, Organic Search, Olark Chat, etc.

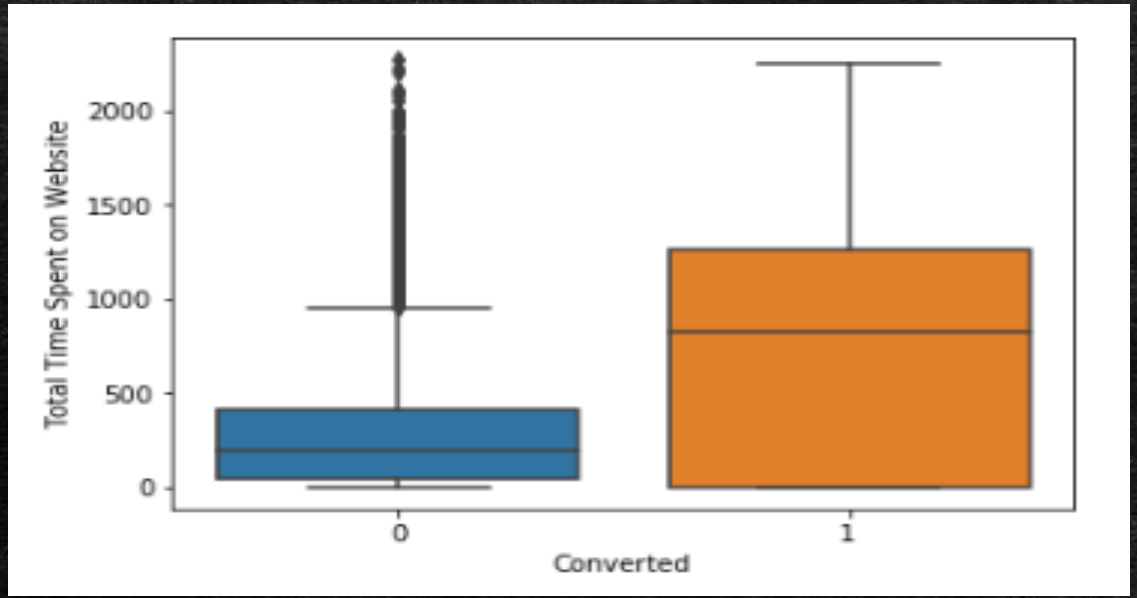




So the highest conversion is from Google with highest conversion ratio obtained from Reference and Welingak Website sources.

Univariate and Bivariate Analysis of Data

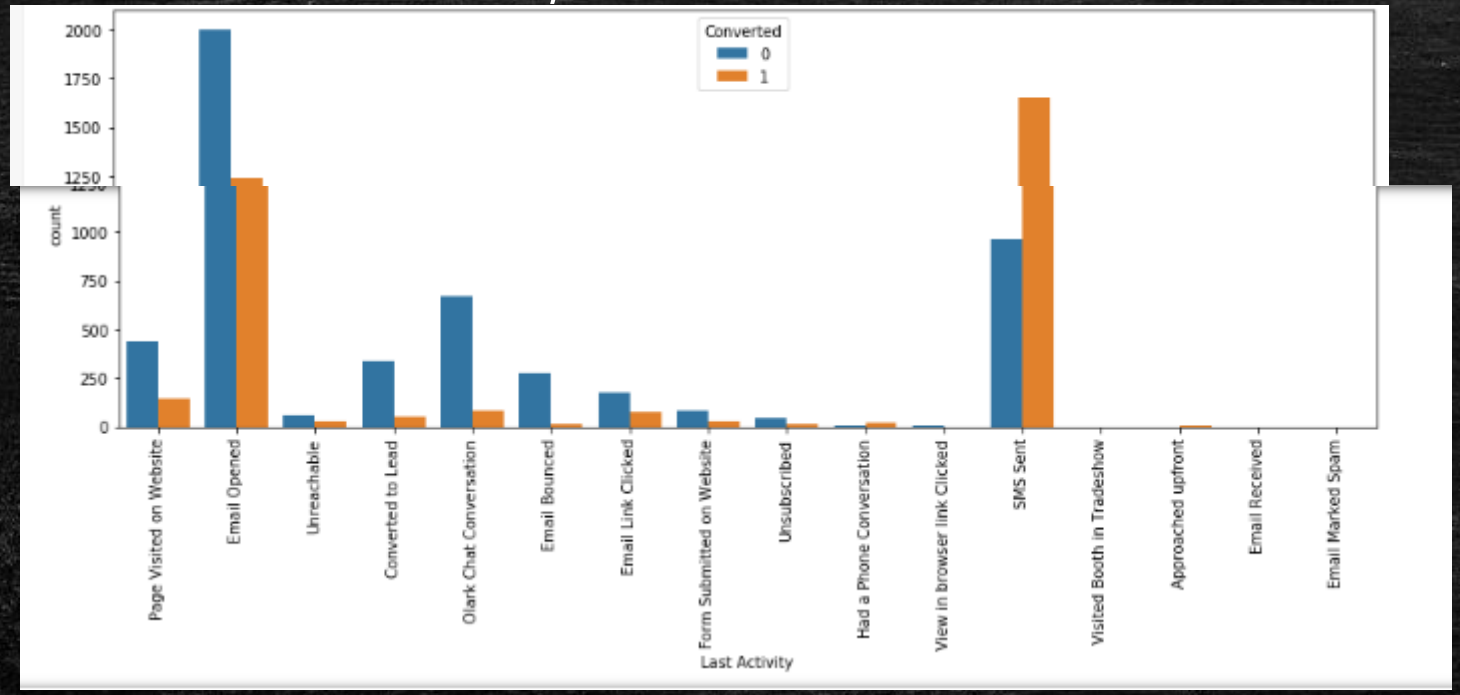
Variable selected : **Total Time Spent on Website**: The total time spent by the customer on the website.



Leads spending more time on the website are more likely to be converted. Website should be made more interactive and engaging for our leads.

Univariate and Bivariate Analysis of Data

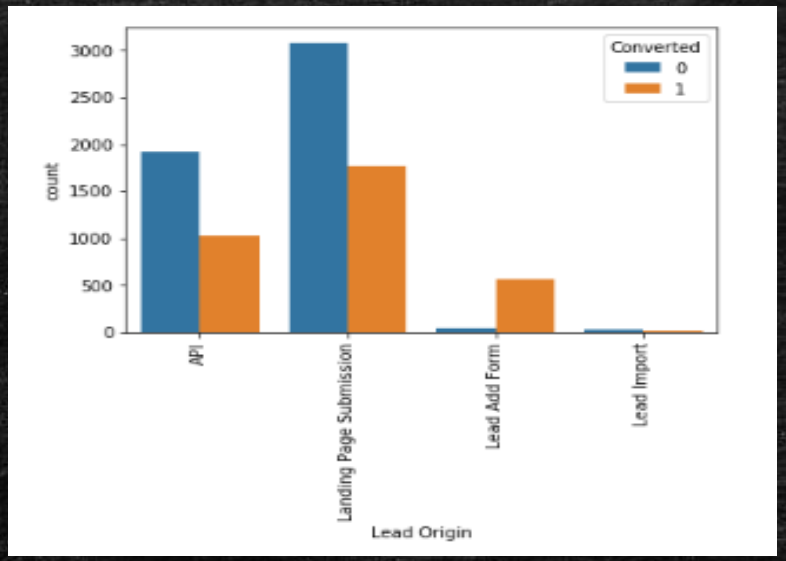
Variable selected :**Last Activity:** Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.



Most of the leads have email opened as the last activity Conversion rate is highest for the SMS sent category.

Univariate and Bivariate Analysis of Data

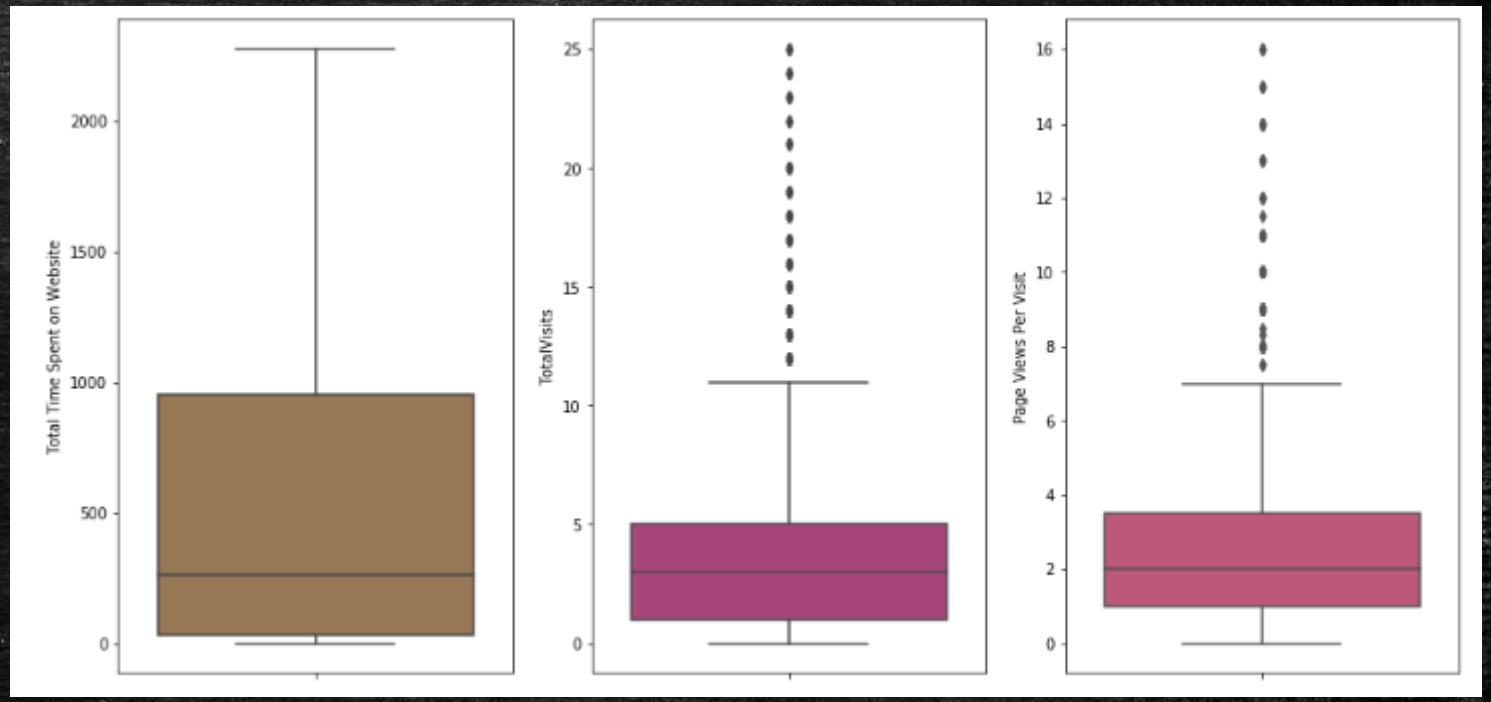
Variable selected :**Lead Origin:** The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission,etc.



Lead Add Form has highest conversion rate but highest number of leads comes from Landing Page Submission.

Outlier Analysis of the Data Set:

	Total Time Spent on Website	TotalVisits	Page Views Per Visit
count	8433.000000	8433.000000	8433.000000
mean	503.034863	3.473853	2.483103
std	550.261144	3.309653	2.067539
min	0.000000	0.000000	0.000000
25%	32.000000	1.000000	1.000000
50%	262.000000	3.000000	2.000000
75%	954.000000	5.000000	3.500000
90%	1390.000000	7.000000	5.000000
95%	1569.000000	10.000000	6.000000
99%	1844.680000	16.000000	9.000000
max	2272.000000	25.000000	16.000000



So there are few outliers present in the dataset for numeric data , which we are handling through IQR.

Splitting Data into Training and Test Sets and Feature Scaling

```
# Splitting the data into train and test:
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

```
# Scale the three numeric features present in our dataset:
# We will use standard scaler

scaler = StandardScaler()

X_train[['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']] = scaler.fit_transform(X_train[['TotalVisit

X_train.head()
```

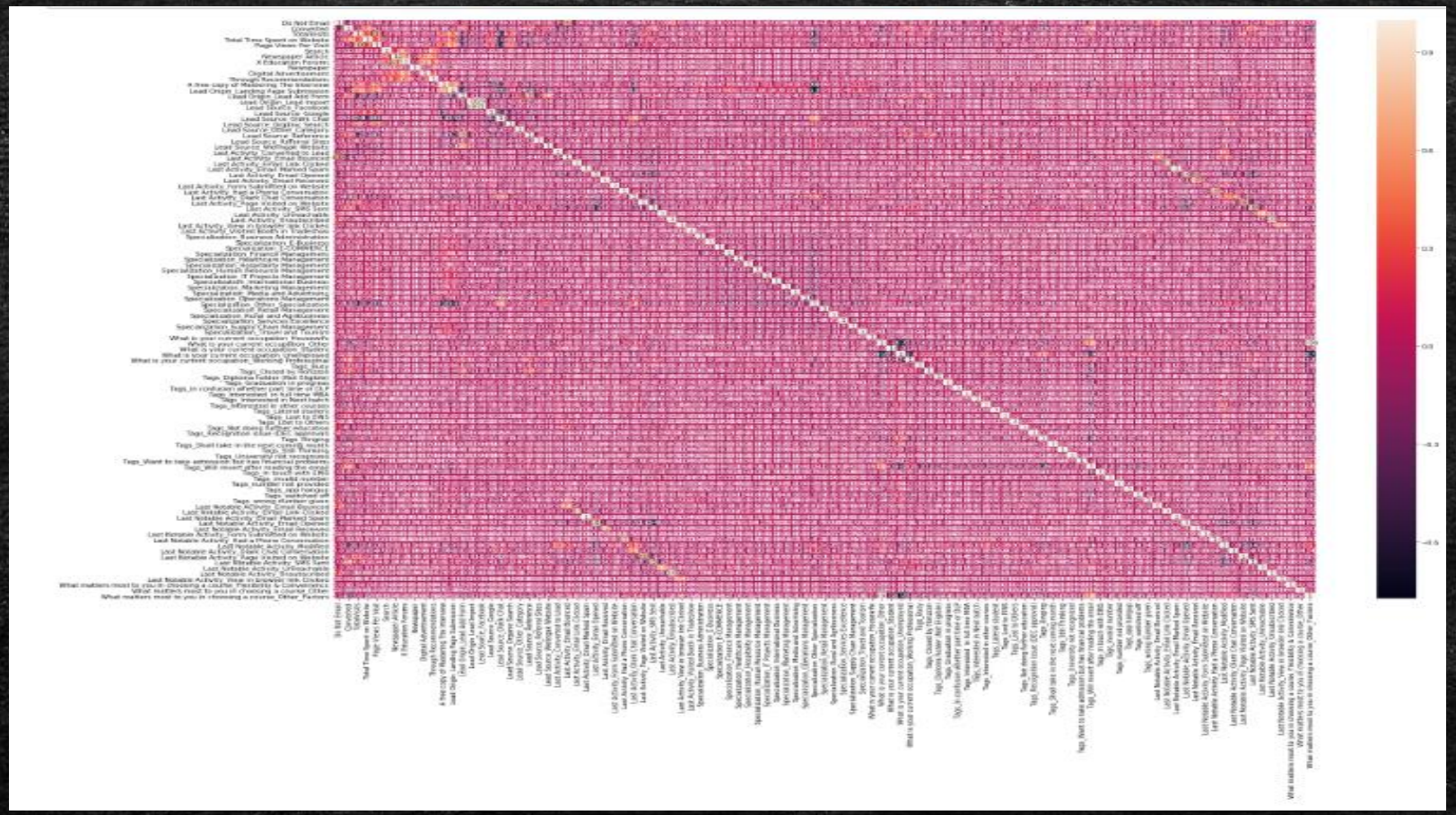
Checking the current conversion rate:

```
56]: conversion = (sum(main_lead_df['Converted'])/len(main_lead_df['Converted'].index))*100
     conversion

56]: 39.962053836120006
```

So as per the dataset after data preparation the current conversion rate is around 40%

Checking the initial correlation:



Nothing much can be interpreted from the heatmap as there are many variables present.

Feature selection using RFE:

- We have 104 features currently and we want to get top few features for our model:

```
logreg = LogisticRegression()

from sklearn.feature_selection import RFE
rfe = RFE(logreg, 15)             # running RFE with 15 variables as output
rfe = rfe.fit(X_train, y_train)
```

```
rfe.support_
```

```
array([False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False,  True,  True, False, False,
       False, False, False,  True, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False,  True,  True,  True, False,
       False, False, False, False, False,  True,  True, False,  True,
       False,  True, False, False, False, False,  True, False, False,
       False, False,  True,  True, False, False, False, False, False,
       False, False, False, False, False,  True, False, False, False,
       False, False,  True])
```

Following 15 features were selected by RFF:

```
['Last Activity_Converted to Lead', 'Last Activity_Email Bounced', 'Last Activity_Olark Chat Conversation', 'What is your current occupation_Working
Professional', 'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Lateral student', 'Tags_Lost to EINS', 'Tags_Not doing further education', 'Tags_Ringing',
'Tags_Will revert after reading the email', 'Tags_switched off', 'Tags_wrong number given', 'Last Notable Activity_SMS Sent',
'What matters most to you in choosing a course_Other_Factors']
```

Model Building: Collinearity check

- After iterative model building for 5 times to minimize the p-value and VIF. We are eliminating the variables in each iteration of the new model building.

Model Building: Final Model Interpretation

Generalized Linear Model Regression Results

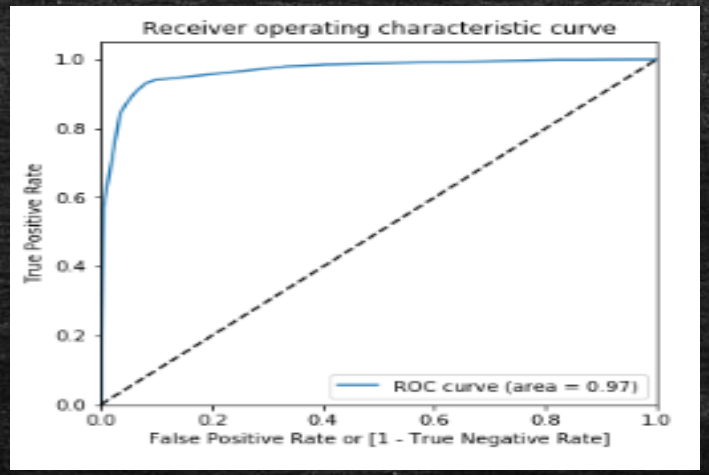
Dep. Variable:	Converted	No. Observations:	5903
Model:	GLM	Df Residuals:	5890
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1262.9
Date:	Mon, 26 Aug 2019	Deviance:	2525.8
Time:	14:01:39	Pearson chi2:	1.12e+04
No. Iterations:	7	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.4708	0.179	-19.350	0.000	-3.822	-3.119
Last Activity_Converted to Lead	-1.4452	0.300	-4.820	0.000	-2.033	-0.858
Last Activity_Email Bounced	-2.6793	0.417	-6.427	0.000	-3.496	-1.862
Last Activity_Olark Chat Conversation	-1.8792	0.217	-8.654	0.000	-2.305	-1.454
What is your current occupation_Working Professional	1.5701	0.295	5.327	0.000	0.992	2.148
Tags_Busy	2.7424	0.271	10.107	0.000	2.211	3.274
Tags_Closed by Horizzon	8.1946	0.737	11.112	0.000	6.749	9.640
Tags_Lost to EINS	10.1310	0.657	15.410	0.000	8.842	11.419
Tags_Ringing	-1.1238	0.280	-4.009	0.000	-1.673	-0.574
Tags_Will revert after reading the email	5.3471	0.196	27.272	0.000	4.963	5.731
Tags_switched off	-1.7187	0.617	-2.784	0.005	-2.929	-0.509
Last Notable Activity_SMS Sent	2.0440	0.137	14.942	0.000	1.776	2.312
What matters most to you in choosing a course_Other_Factors	-3.9167	0.128	-30.555	0.000	-4.168	-3.665

	Features	VIF
4	Tags_Busy	1.04
1	Last Activity_Email Bounced	1.03
5	Tags_Closed by Horizzon	1.03
9	Tags_switched off	1.03
6	Tags_Lost to EINS	1.02
3	What is your current occupation_Working Profes...	0.60
10	Last Notable Activity_SMS Sent	0.21
8	Tags_Will revert after reading the email	0.20
11	What matters most to you in choosing a course_...	0.20
2	Last Activity_Olark Chat Conversation	0.06
7	Tags_Ringing	0.06
0	Last Activity_Converted to Lead	0.05

As the VIF value is under control (<2) and correlation is <0.50 so the problem of multi-collinearity is not there in model. This is also verified from the correlation matrix and heatmap.
Also p-value is under 0.05 , so we can take this model further.

Model Evaluation : ROC Curve Analysis



An ROC curve demonstrates several things:

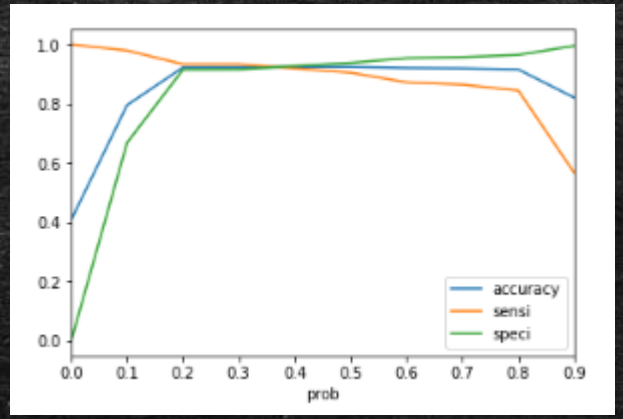
•It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

•The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

•The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

- This ROC curve suggests that our model is good as the area under the curve is 0.97 and its closer to 1.

Model Evaluation : Finding cutoff Probability



- From the curve, 0.3 is the optimum point to take it as a cutoff probability.

```
y_train_pred_final['final_predicted'] = y_train_pred_final.Converted_prob.map( lambda x: 1 if x > 0.3 else 0)

y_train_pred_final.head()
```

	Converted	Converted_prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted
0	0	0.115027	441	0	1	1	0	0	0	0	0	0	0	0	0
1	1	0.998156	693	1	1	1	1	1	1	1	1	1	1	1	1
2	1	0.867186	5820	1	1	1	1	1	1	1	1	1	1	0	1
3	0	0.072390	7656	0	1	0	0	0	0	0	0	0	0	0	0
4	1	0.867186	2223	1	1	1	1	1	1	1	1	1	1	0	1

Model Evaluation :Assigning Lead Score, Precision and recall tradeoff

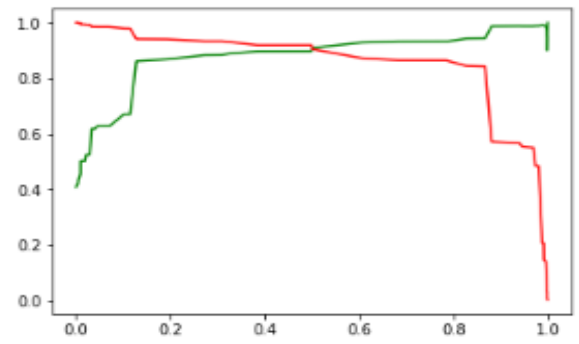
```
y_train_pred_final['Lead_Score'] = y_train_pred_final.Converted_prob.map( lambda x: round(x*100))

y_train_pred_final.head()
```

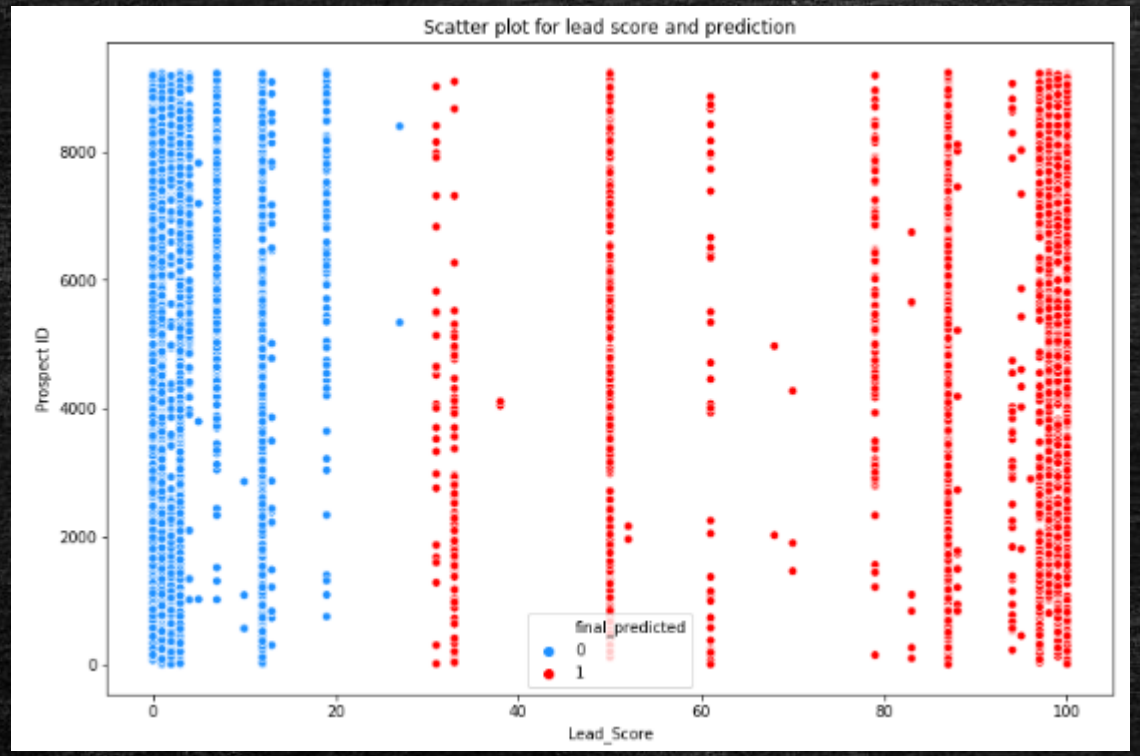
	Converted	Converted_prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead_Score
0	0	0.115027	441	0	1	1	0	0	0	0	0	0	0	0	0	12
1	1	0.998156	693	1	1	1	1	1	1	1	1	1	1	1	1	100
2	1	0.867186	5820	1	1	1	1	1	1	1	1	1	1	0	1	87
3	0	0.072390	7656	0	1	0	0	0	0	0	0	0	0	0	0	7
4	1	0.867186	2223	1	1	1	1	1	1	1	1	1	1	0	1	87

```
p, r, thresholds = precision_recall_curve(y_train_pred_final.Converted, y_train_pred_final.Converted_prob)
```

```
plt.plot(thresholds, p[:-1], "g-")
plt.plot(thresholds, r[:-1], "r-")
plt.show()
```



Recommendation and Summary:



	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score	Lead_Type
0	4523	1	0.980552	1	98	Hot
1	7562	0	0.115027	0	12	Cold
2	3097	1	0.995891	1	100	Hot
3	5968	1	0.500913	1	50	Hot
4	4030	1	0.867186	1	87	Hot

- So we have lead score assigned now for each prospect lead for all the eligible data points in our data-set.

Recommendation and Summary:

Comparing the test and train metrics:

```
The train accuracy is : 0.922582 and the test accuracy is : 0.927668
The train sensitivity is : 0.933167 and the test sensitivity is : 0.933167
The train specificity is : 0.915283 and the test specificity is : 0.915283
The train recall is : 0.904525 and the test recall is : 0.933167
The train precision is : 0.909053 and the test precision is : 0.883648
```

- Below are the different feature variables with the respective weights as coefficient value selected for our model:

	coef
const	-3.4708
Last Activity_Converted to Lead	-1.4452
Last Activity_Email Bounced	-2.6793
Last Activity_Olark Chat Conversation	-1.8792
What is your current occupation_Working Professional	1.5701
Tags_Busy	2.7424
Tags_Closed by Horizzon	8.1946
Tags_Lost to EINS	10.1310
Tags_Ringing	-1.1238
Tags_Will revert after reading the email	5.3471
Tags_switched off	-1.7187
Last Notable Activity_SMS Sent	2.0440
What matters most to you in choosing a course_Other_Factors	-3.9167

Also the model is good as the performance of model is better on the test data than the train data.