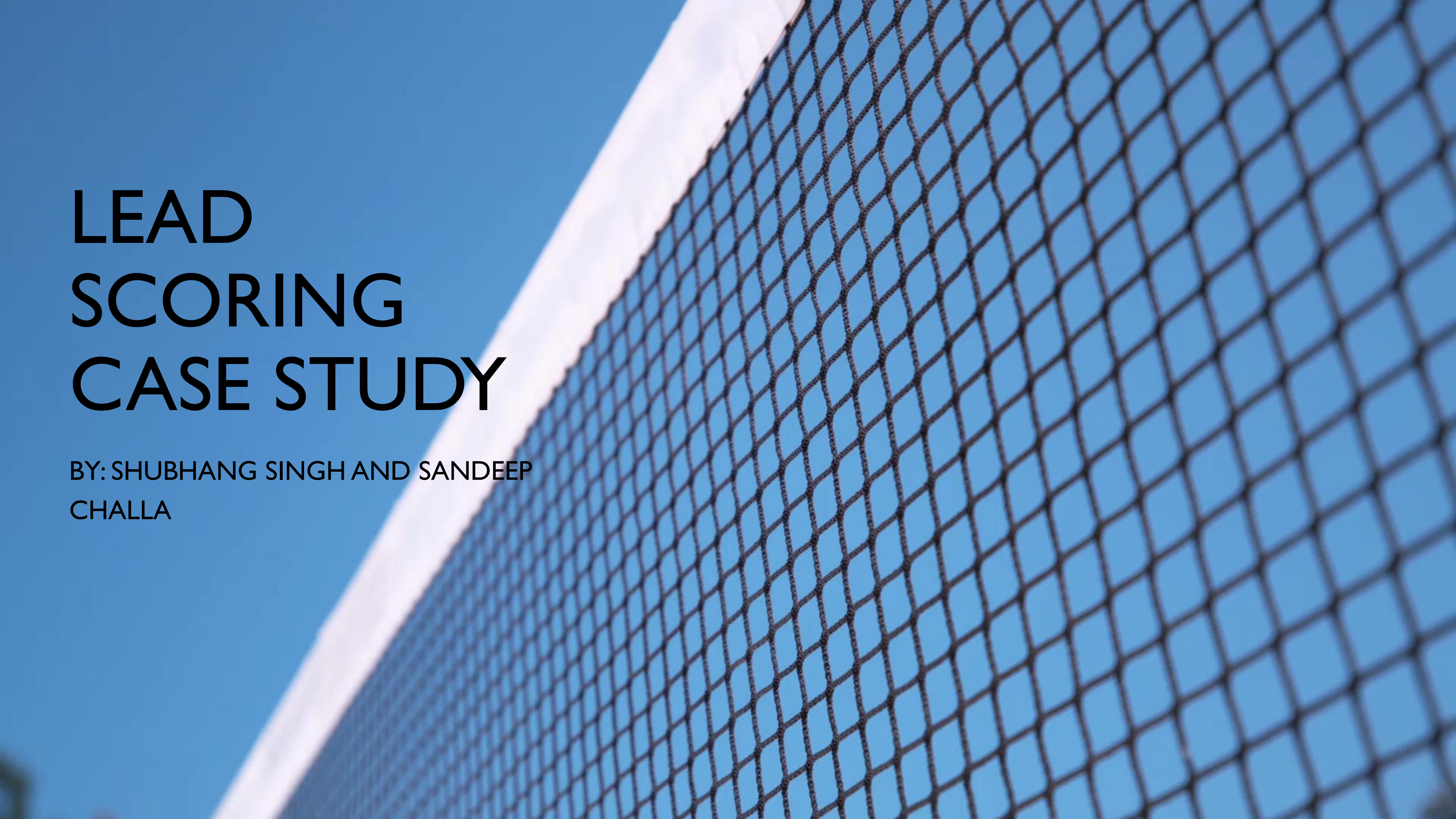


LEAD SCORING CASE STUDY

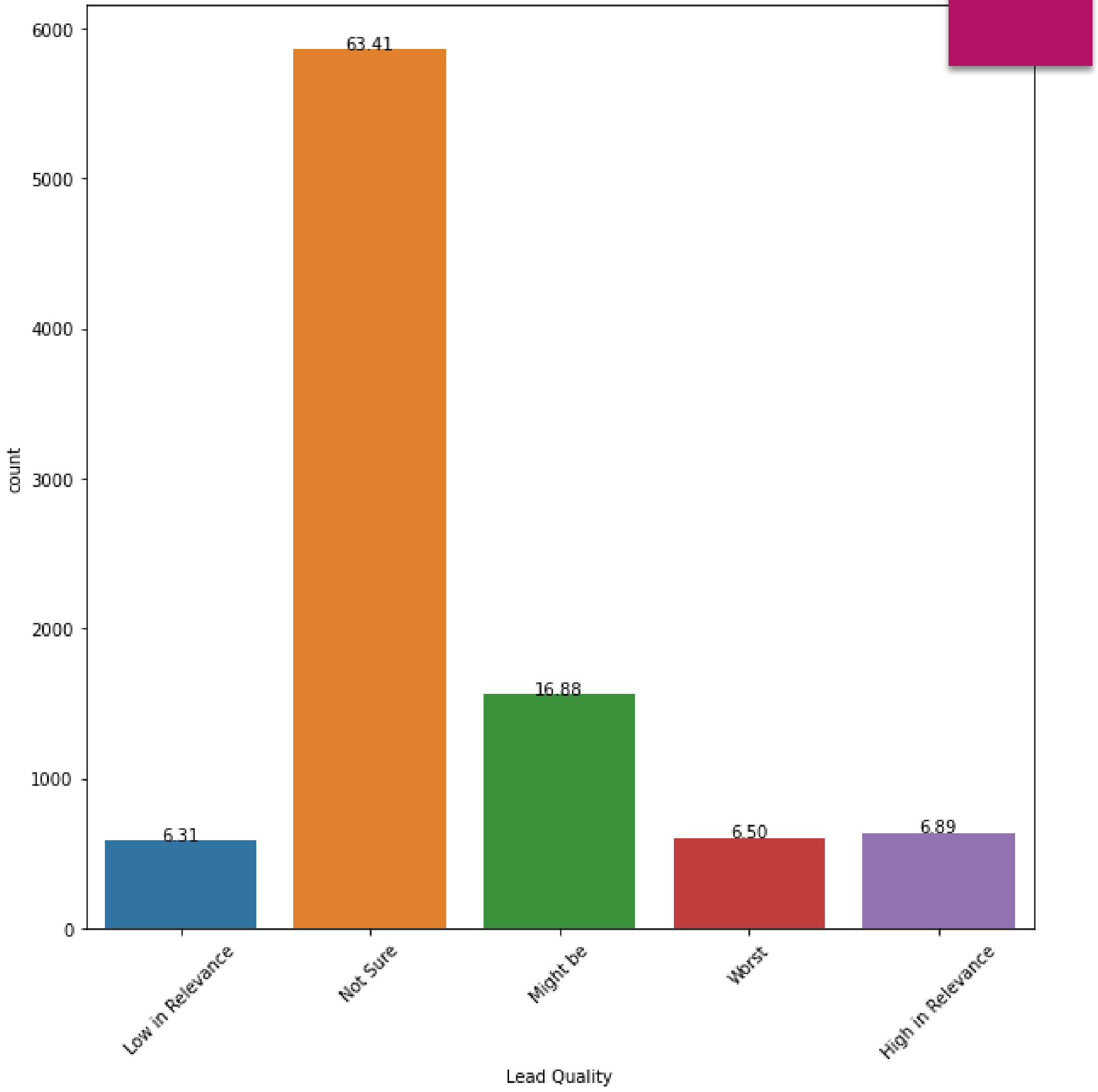
BY: SHUBHANG SINGH AND SANDEEP CHALLA

PROBLEM STATEMENT

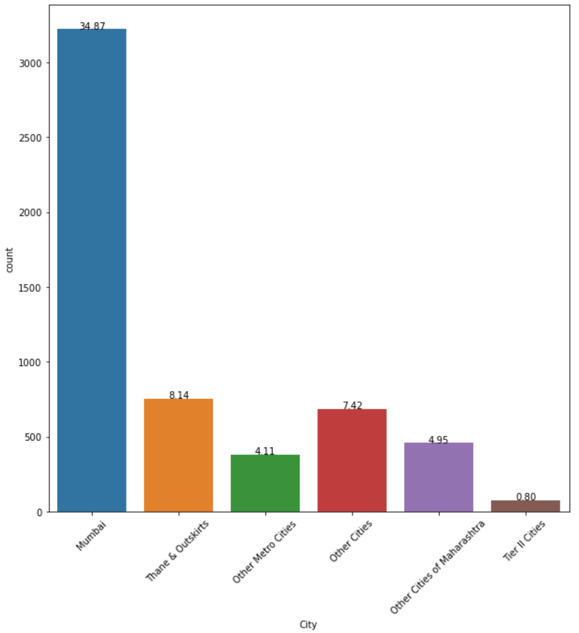
- Create a model in such a way that the customers with high lead score have higher conversion chance and low leads score have lower conversion chance. The ballpark of the target lead conversion rate is around 80%
- Also, the model should be able to adjust if the company's requirement changes in near future.

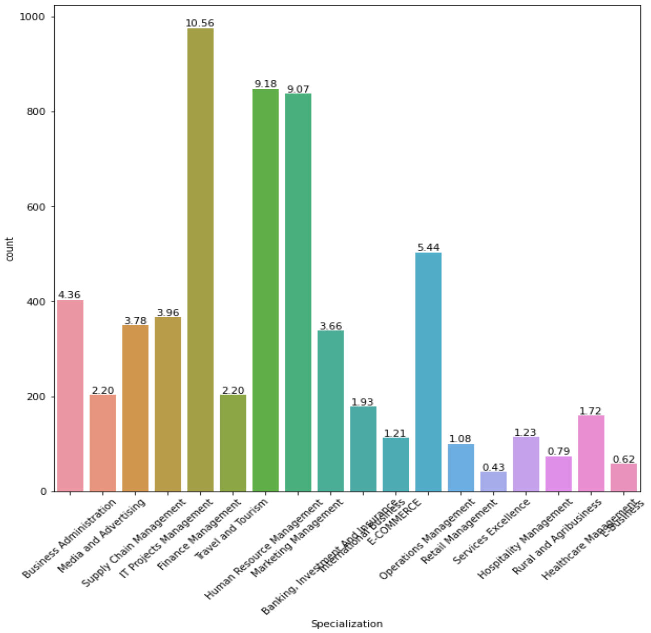
Approach of the analysis

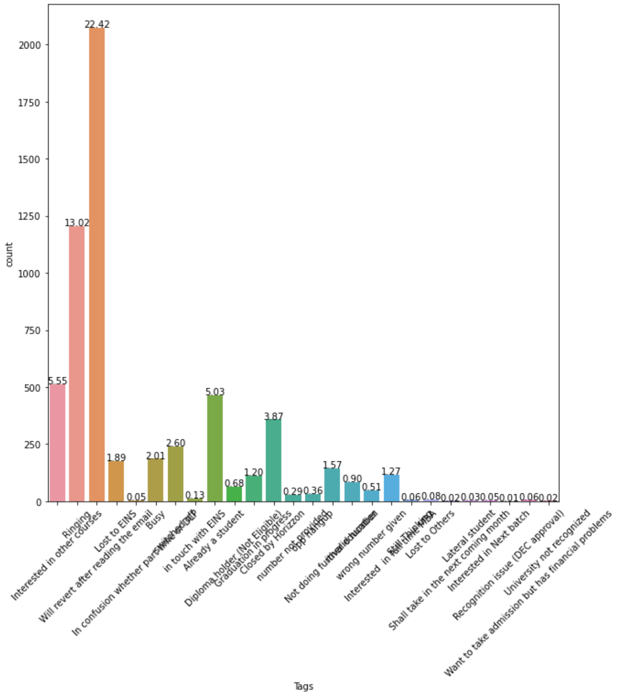
- We started our analysis by reading and understanding the data. Finding the numerical and categorical variables and also looking at the number of null values in each of them.
- Then we went on with data cleaning in which we dropped the columns having null percentage greater than 50%.

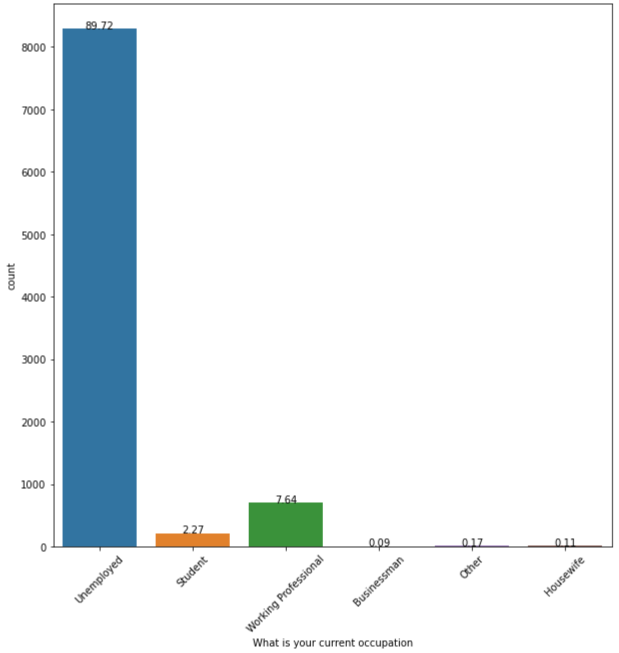


EDA: EXPLORATORY DATA ANALYSIS

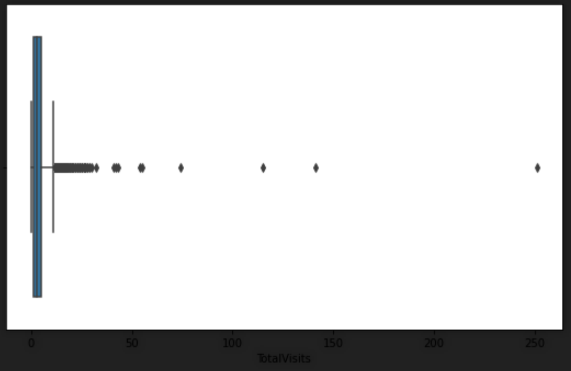


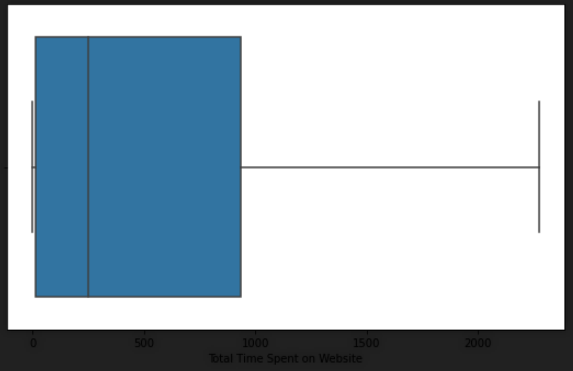


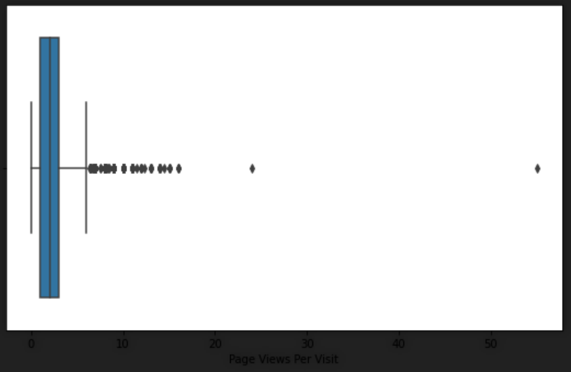




Outlier Treatment







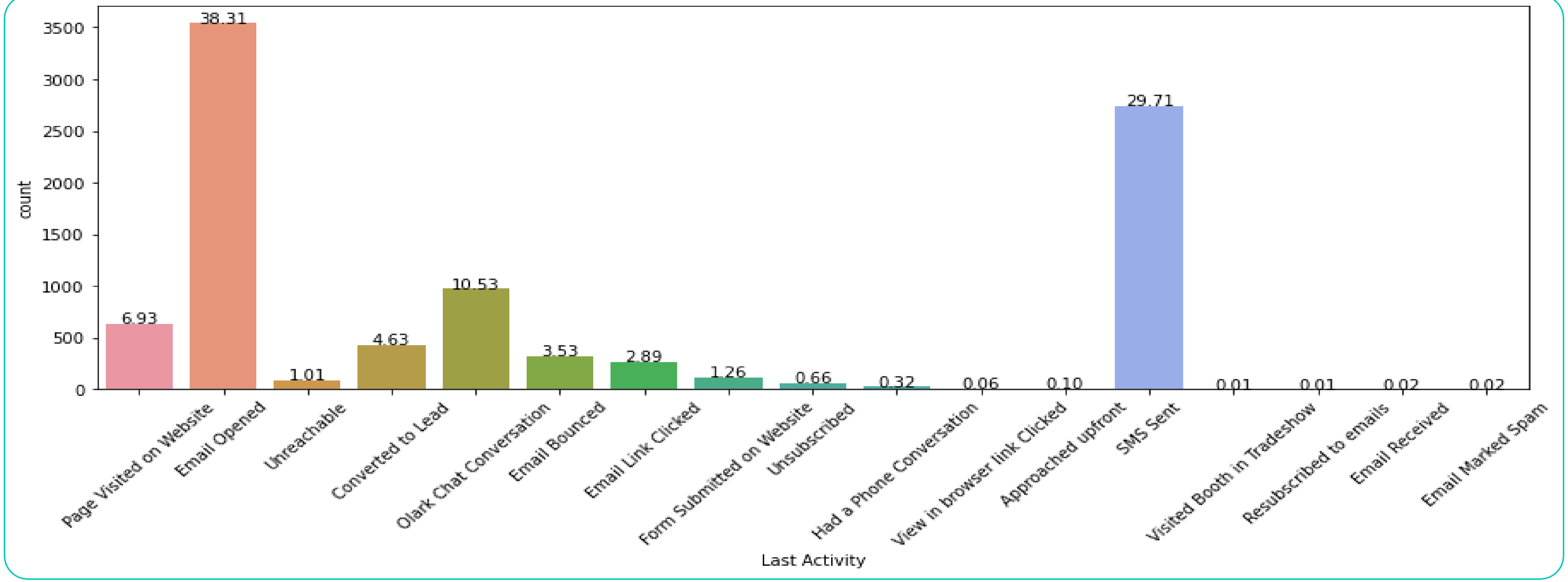
- There are lot of outliers in **TotalVisits** and **Page Views per Visit**.

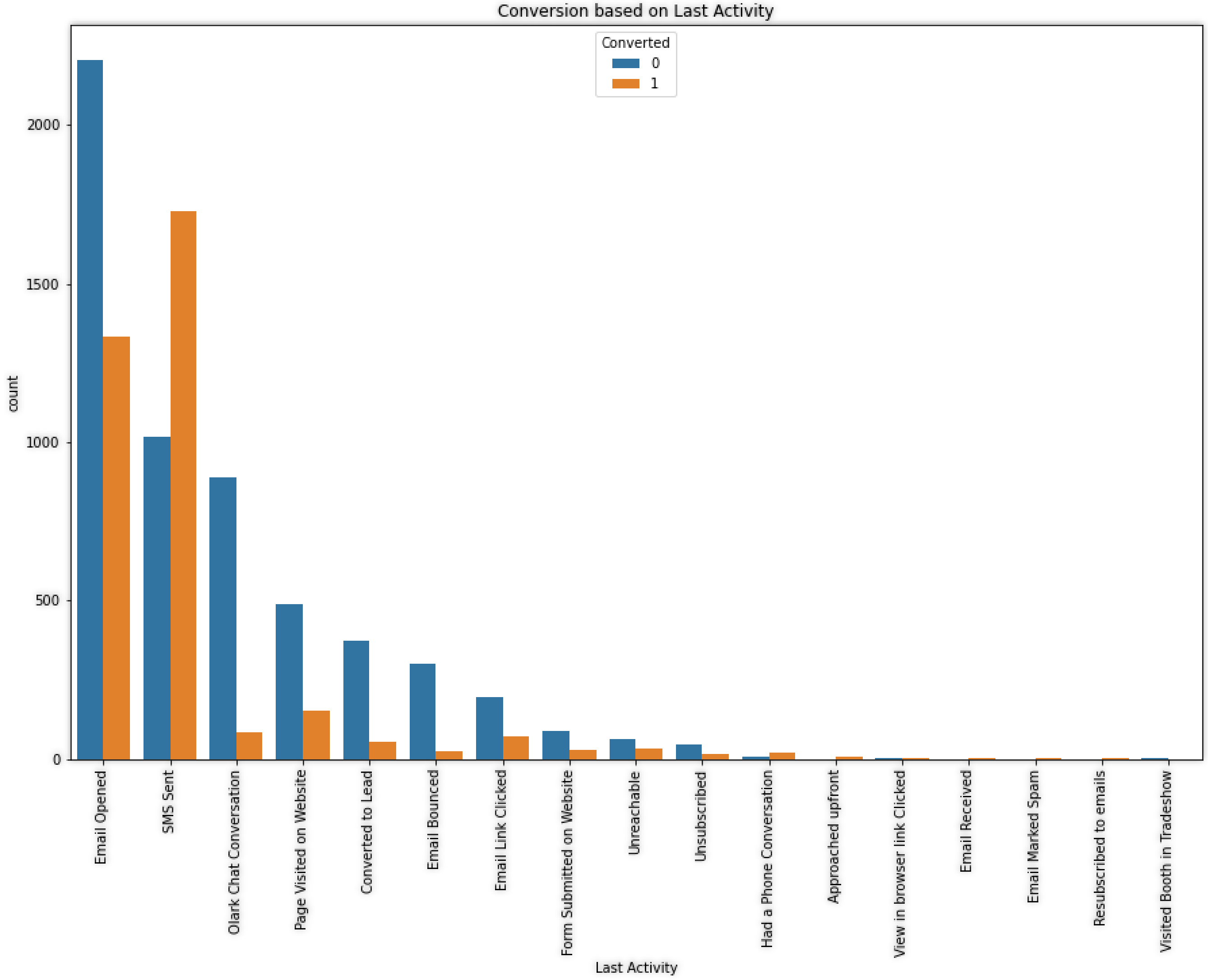
Univariate and Bivariate Analysis

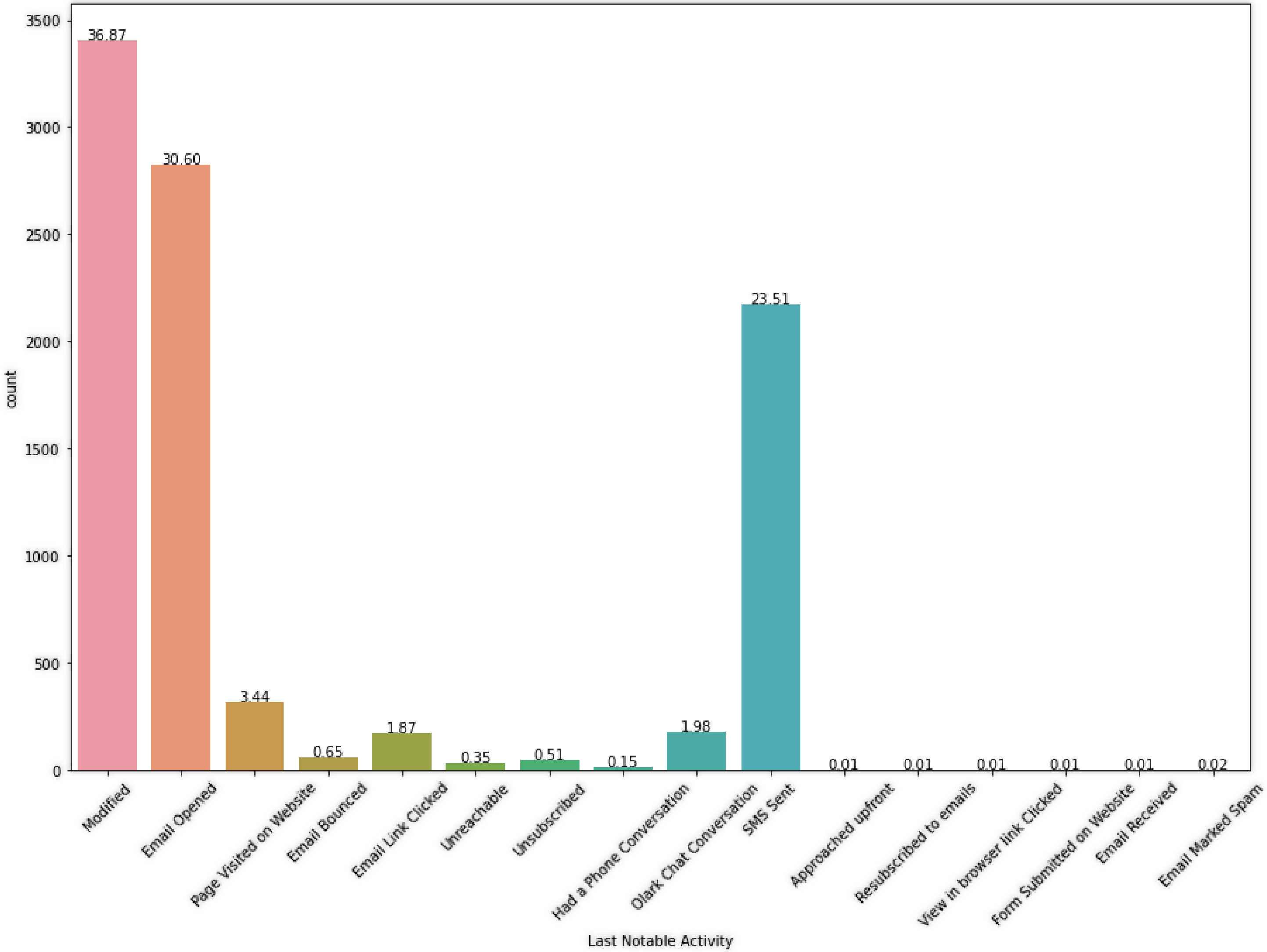
Observations:

1. Lead Add Form has the highest conversion rate at 92%
2. Quick Add Form has 100% conversion rate but it has only one entry, so not a reliable factor
3. API has the least amount of conversions.
4. Most student found X education via Google Search. But most leads were not converted to actual students.
5. Reference had the highest number of conversions at 92%
6. No conversions were made through Youtube channel, blog, press releases.

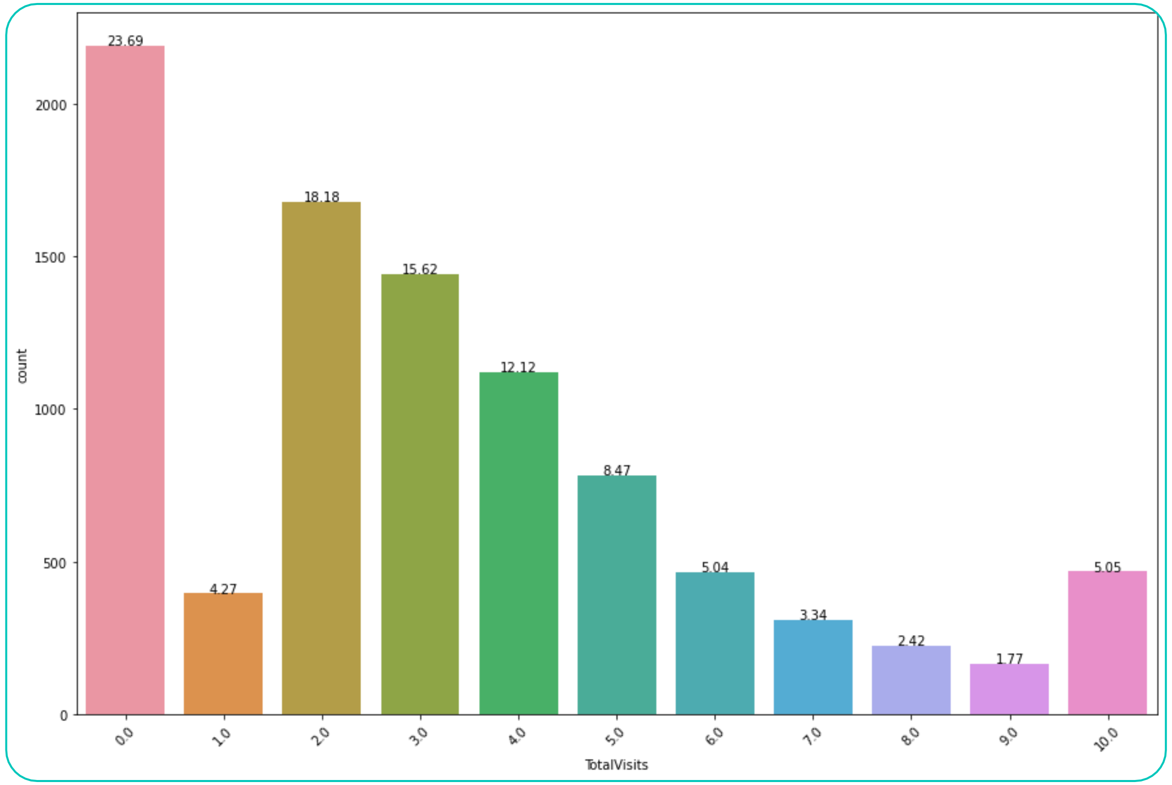
Bar Plots


3500
3000
2500
2000
1500
1000
500
0
count
6.93
38.31
1.01
4.63
10.53
3.53
2.89
1.26
0.66
0.32
0.06
0.10
29.71
0.01
0.01
0.02
0.02
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
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity


Conversion based on Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Email Opened
SMS Sent
Olark Chat Conversation
Page Visited on Website
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unreachable
Unsubscribed
Had a Phone Conversation
Approached upfront
View in browser link Clicked
Email Received
Email Marked Spam
Resubscribed to emails
Visited Booth in Tradeshow
Last Activity

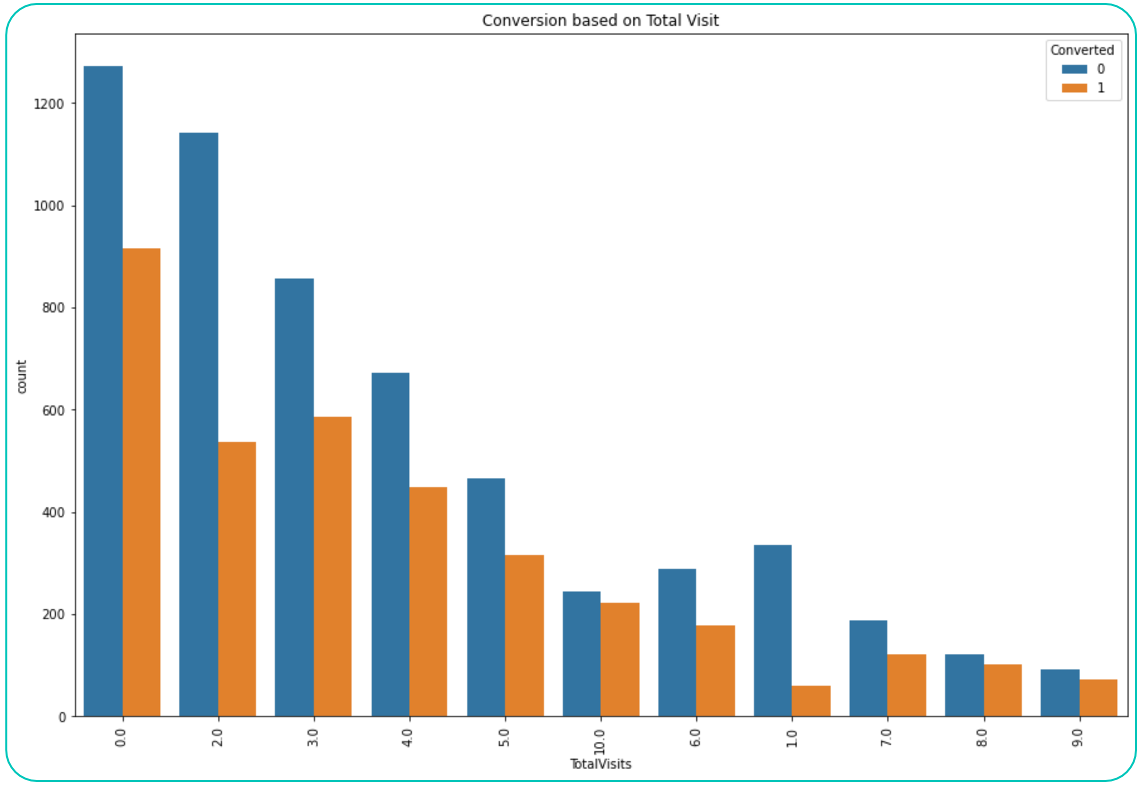

3500
3000
2500
2000
1500
1000
500
0
count
36.87
30.60
3.44
0.65
1.87
0.35
0.51
0.15
1.98
23.51
0.01
0.01
0.01
0.01
0.01
0.02
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

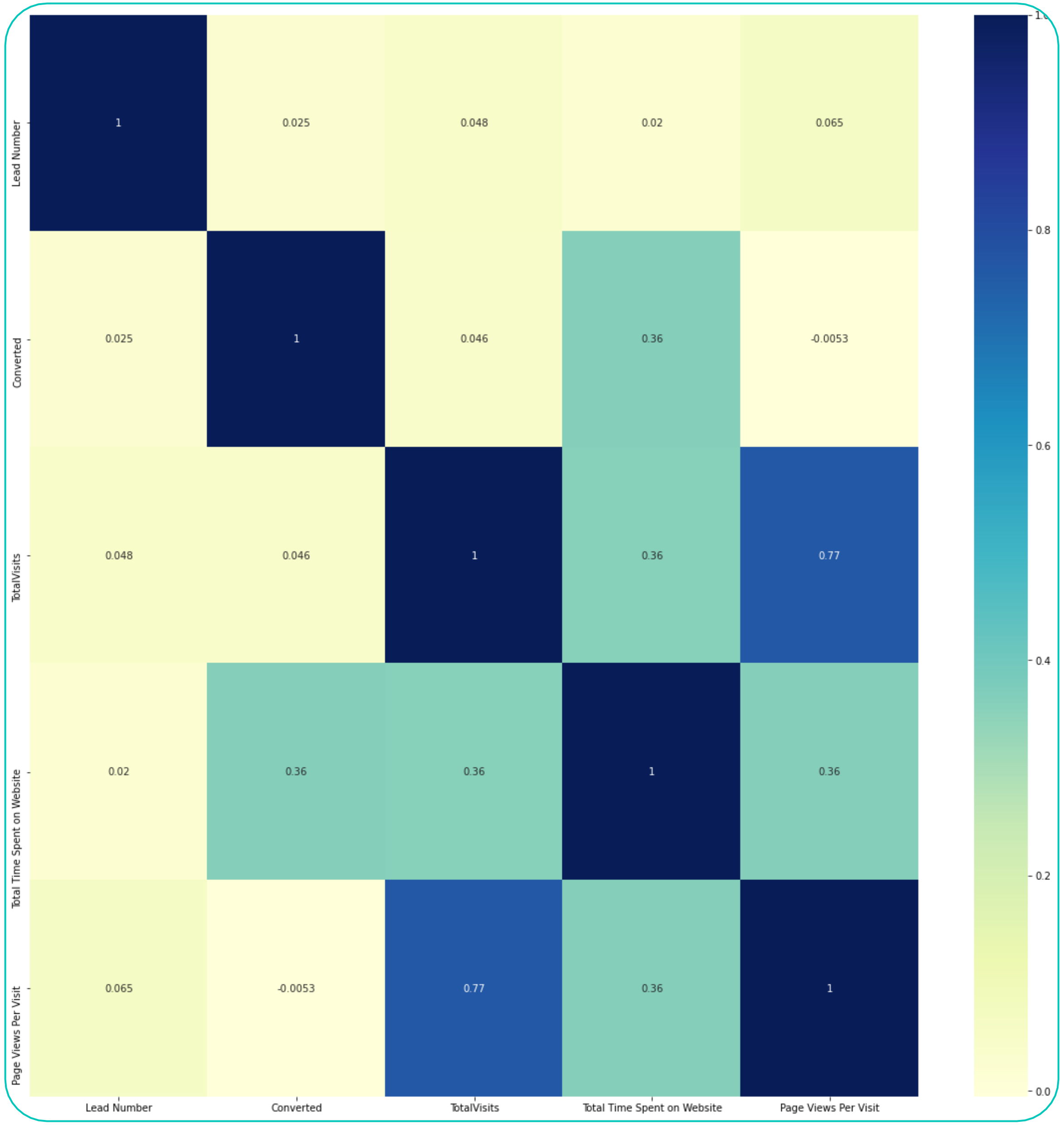
Continuous Variable Bar Plots



For Total Visit

Total Visits vs Conversion based on Total Visits





Checking Corelation b/w Variables

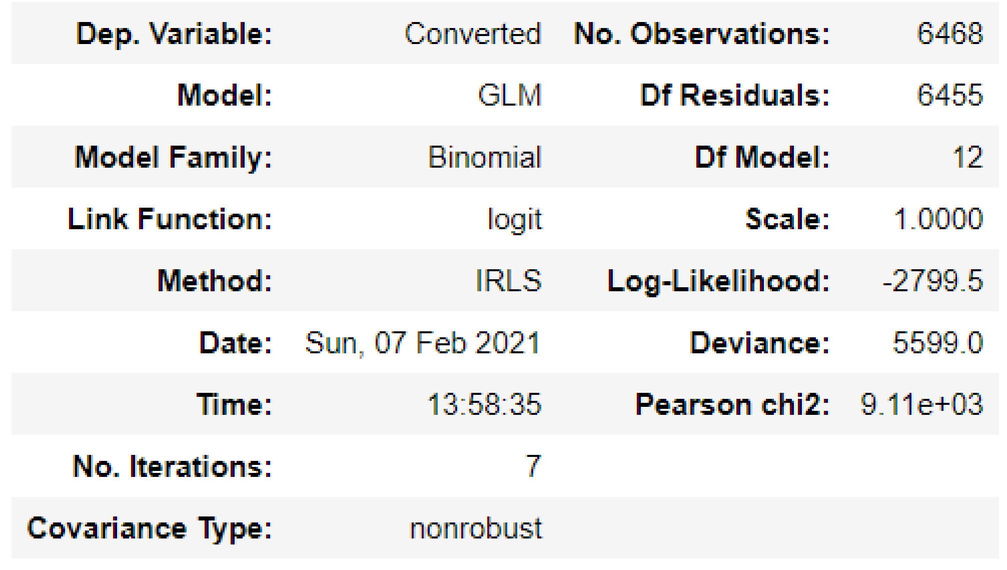
Data Preparation

- Creating dummy variables.
- Performing Train-Test Split
- Scaling

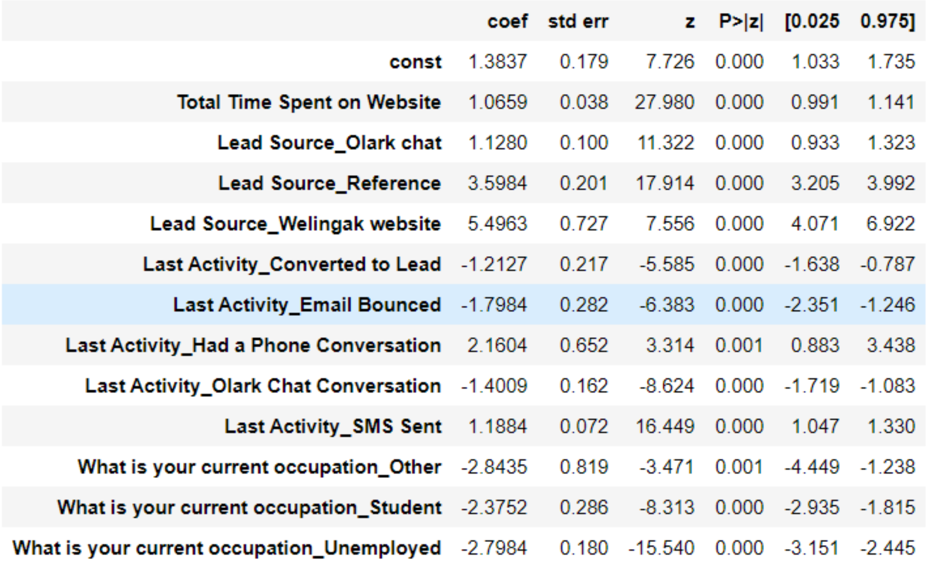
Modelling

- Creating a Logistic Regression Model.
- Feature Selection using RFE
- Assessing the model with StatsModels api
- Recalibrating the model again and again till we reach a point where all the variables are significant and the VIF values are less than 5 to reduce multi-colliearity.

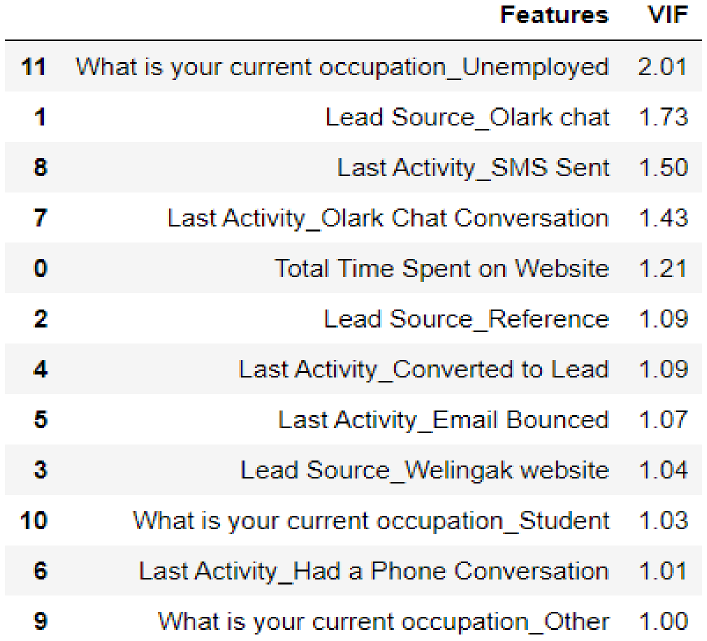
Final Model Visualization with VIF



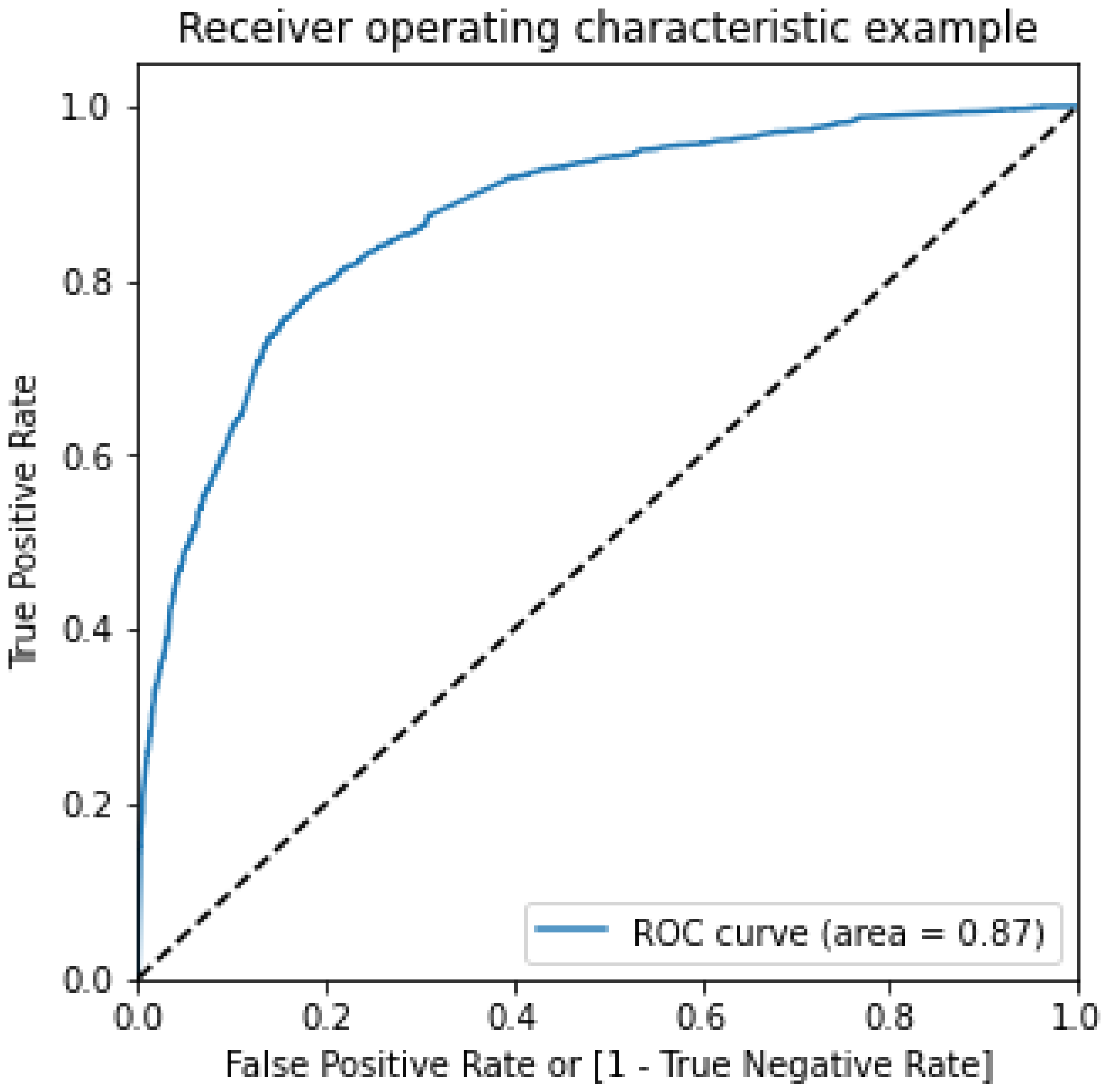
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6455
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2799.5
Date:	Sun, 07 Feb 2021	**Deviance:**	5599.0
Time:	13:58:35	**Pearson chi2:**	9.11e+03
No. Iterations:	7		
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	1.3837	0.179	7.726	0.000	1.033	1.735
Total Time Spent on Website	1.0659	0.038	27.980	0.000	0.991	1.141
Lead Source_Olark chat	1.1280	0.100	11.322	0.000	0.933	1.323
Lead Source_Reference	3.5984	0.201	17.914	0.000	3.205	3.992
Lead Source_Welingak website	5.4963	0.727	7.556	0.000	4.071	6.922
Last Activity_Converted to Lead	-1.2127	0.217	-5.585	0.000	-1.638	-0.787
Last Activity_Email Bounced	-1.7984	0.282	-6.383	0.000	-2.351	-1.246
Last Activity_Had a Phone Conversation	2.1604	0.652	3.314	0.001	0.883	3.438
Last Activity_Olark Chat Conversation	-1.4009	0.162	-8.624	0.000	-1.719	-1.083
Last Activity_SMS Sent	1.1884	0.072	16.449	0.000	1.047	1.330
What is your current occupation_Other	-2.8435	0.819	-3.471	0.001	-4.449	-1.238
What is your current occupation_Student	-2.3752	0.286	-8.313	0.000	-2.935	-1.815
What is your current occupation_Unemployed	-2.7984	0.180	-15.540	0.000	-3.151	-2.445



	Features	VIF
11	What is your current occupation_Unemployed	2.01
1	Lead Source_Olark chat	1.73
8	Last Activity_SMS Sent	1.50
7	Last Activity_Olark Chat Conversation	1.43
0	Total Time Spent on Website	1.21
2	Lead Source_Reference	1.09
4	Last Activity_Converted to Lead	1.09
5	Last Activity_Email Bounced	1.07
3	Lead Source_Welingak website	1.04
10	What is your current occupation_Student	1.03
6	Last Activity_Had a Phone Conversation	1.01
9	What is your current occupation_Other	1.00



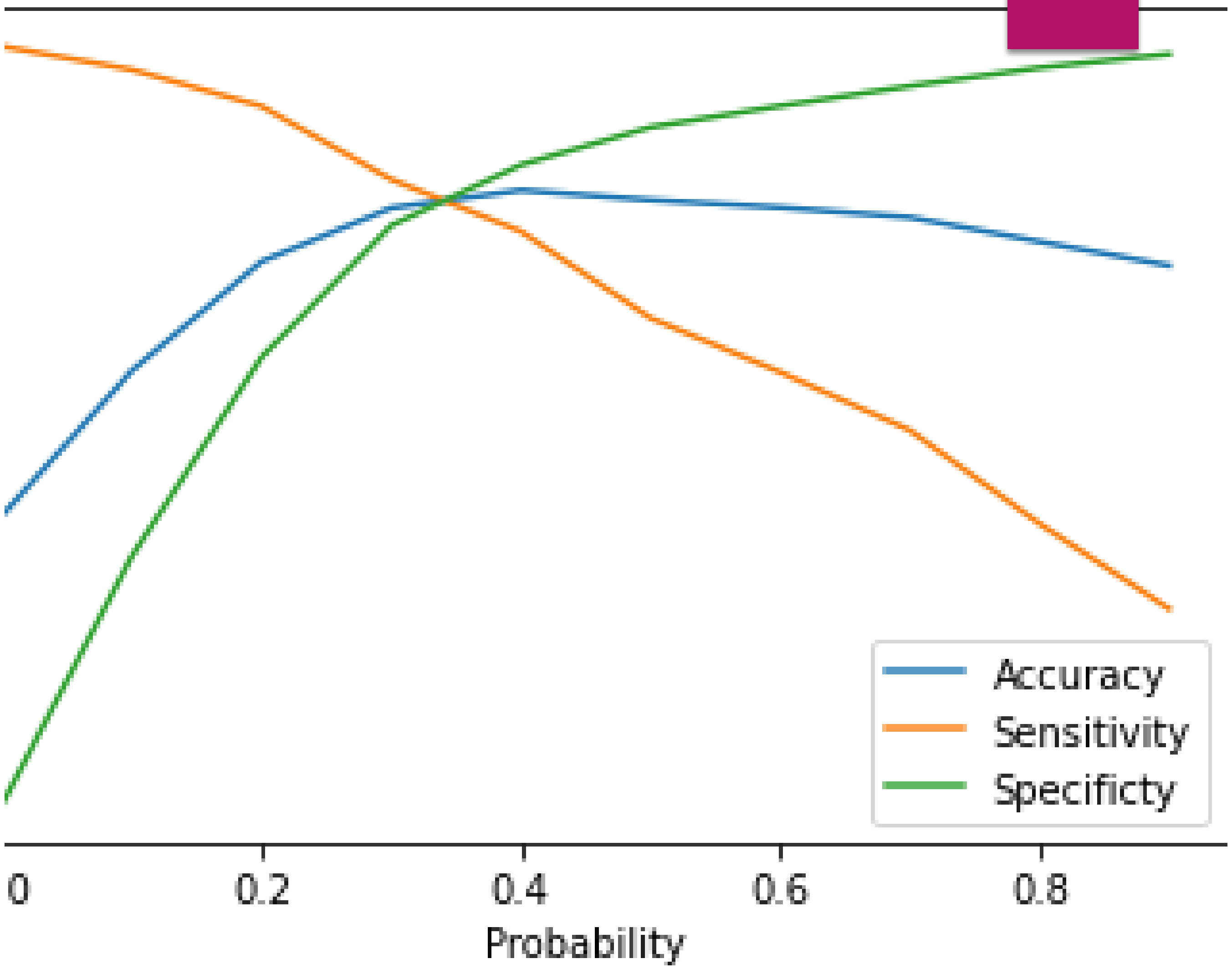
Plotting the ROC curve

Observation:
From the ROC curve we can say that the model will be able to provide us with a good result overall.

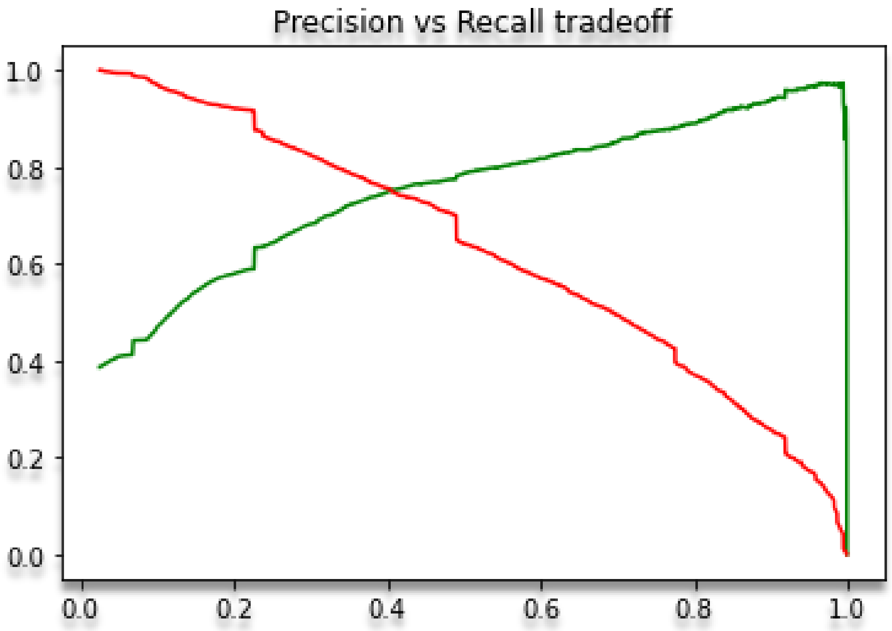
Finding Optimal Cut off point

- Observation:

From the curve we can see that the optimal cut off is at 0.33. This is the point where all the parameters are equally b alnaced.



Precision and Recall Curve



- From the scores we note that our model has a good overall relevancy, defined by Precision at 70% and a great return of relevant results, defined by Recall at 80%.
- The precision vs recall tradeoff value from the graph is at 0.4
- Comparing it with the F1 score we can say that our model us fairly accurate.

Final Model Reporting and Equation

- **log odds = 1.3837 +(1.0659 Total Time Spent on Website) + (1.1280 Lead Source_Olark chat) + (3.5984 Lead Source_Reference) + (5.4963 Lead Source_Welingak website) + (-1.2127 Last Activity_Converted to Lead) + (-1.7984 Last Activity_Email Bounced) + (2.1604 Last Activity_Had a Phone Conversation) + (-1.4009 Last Activity_Olark Chat Conversation) + (1.1884 Last Activity_SMS Sent)+(-2.8435 What is your current occupation_Other)+(-2.3752 What is your current occupation_Student)+(-2.7984 * What is your current occupation_Unemployed)**

Insights

- Hot Leads are identified as Customers having lead score of 33 or above.
- Sales Team of the company should first focus on the “Hot Leads”.
- Higher the Lead score, higher the chances of conversion of “Hot Leads” into “Paying Customers”
- The “Cold Leads should be focused after the Sales Team is done with the “Hot Leads”.