LEAD SCORE CASE STUDY

-ANIKET PANDE

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- 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.
- 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.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

BUSINESS GOALS

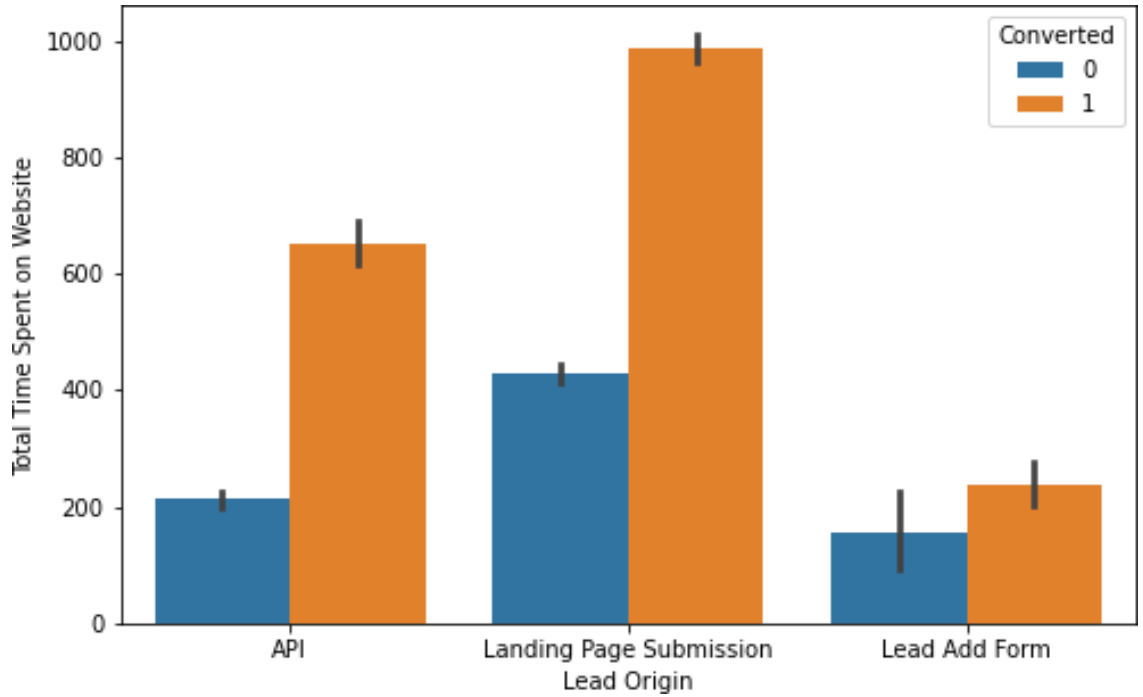
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- Building a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

STEPS INVOLVED

- Understanding the Business problem & then understanding the datasets provided.
- Data Sourcing & Data Inspection.
- Data Cleaning & Manipulation:
  1.Checking of Missing Values
  2.Imputing missing values
- EDA:
  1.Univariate Analysis
  2.Bivariate Analysis
- Creating dummy variable for some of the categorical v
- Splitting data set
- Feature Scaling
- Model Building:
  1.Feature Selection Using RFE
  2.Plotting the ROC Curve
  3.Finding Optimal Cutoff Point
  4.Precision and Recall
  5.Making predictions on the test set.

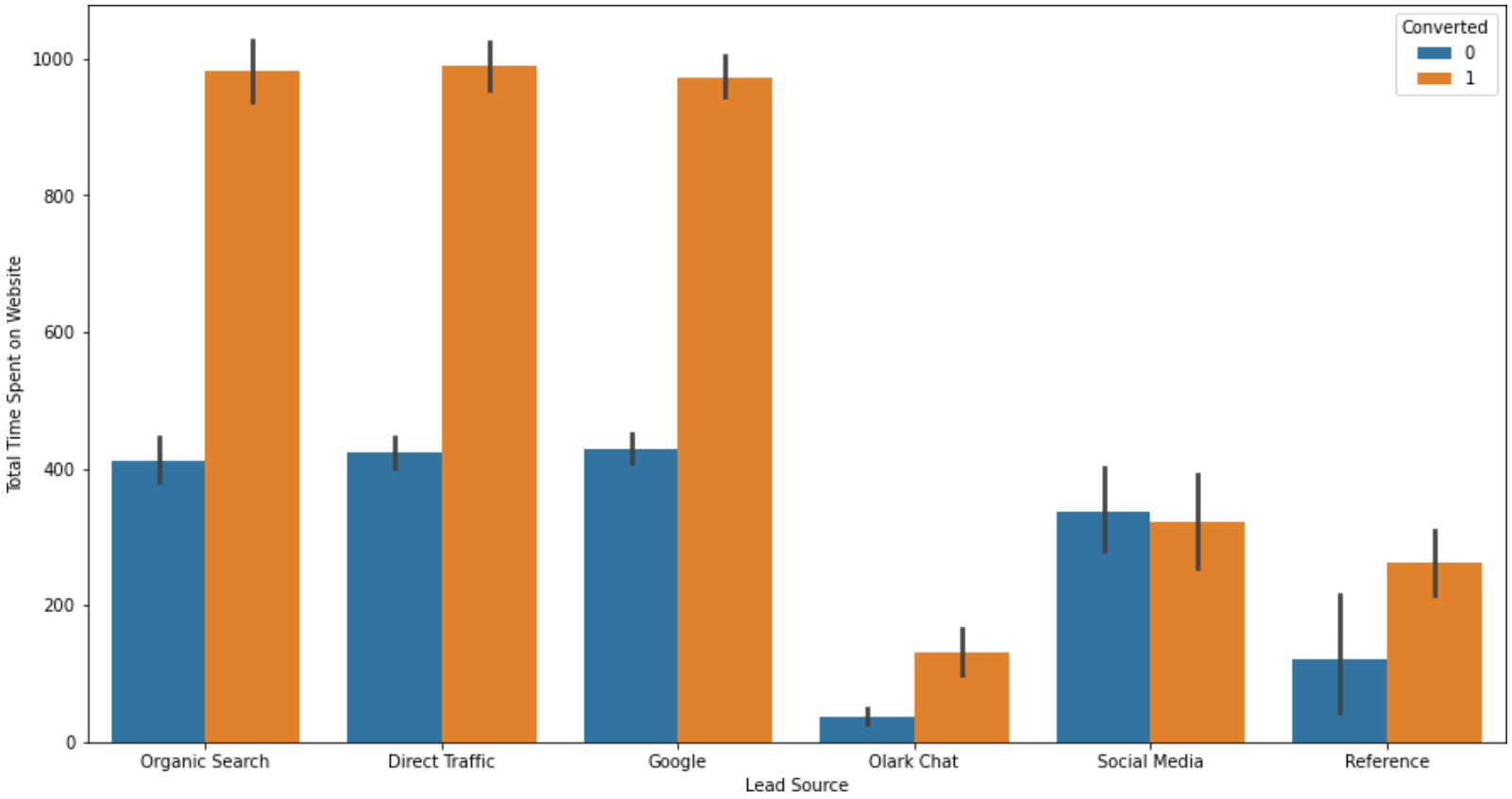
Multivariate Analysis

‘Lead Origin’ Column



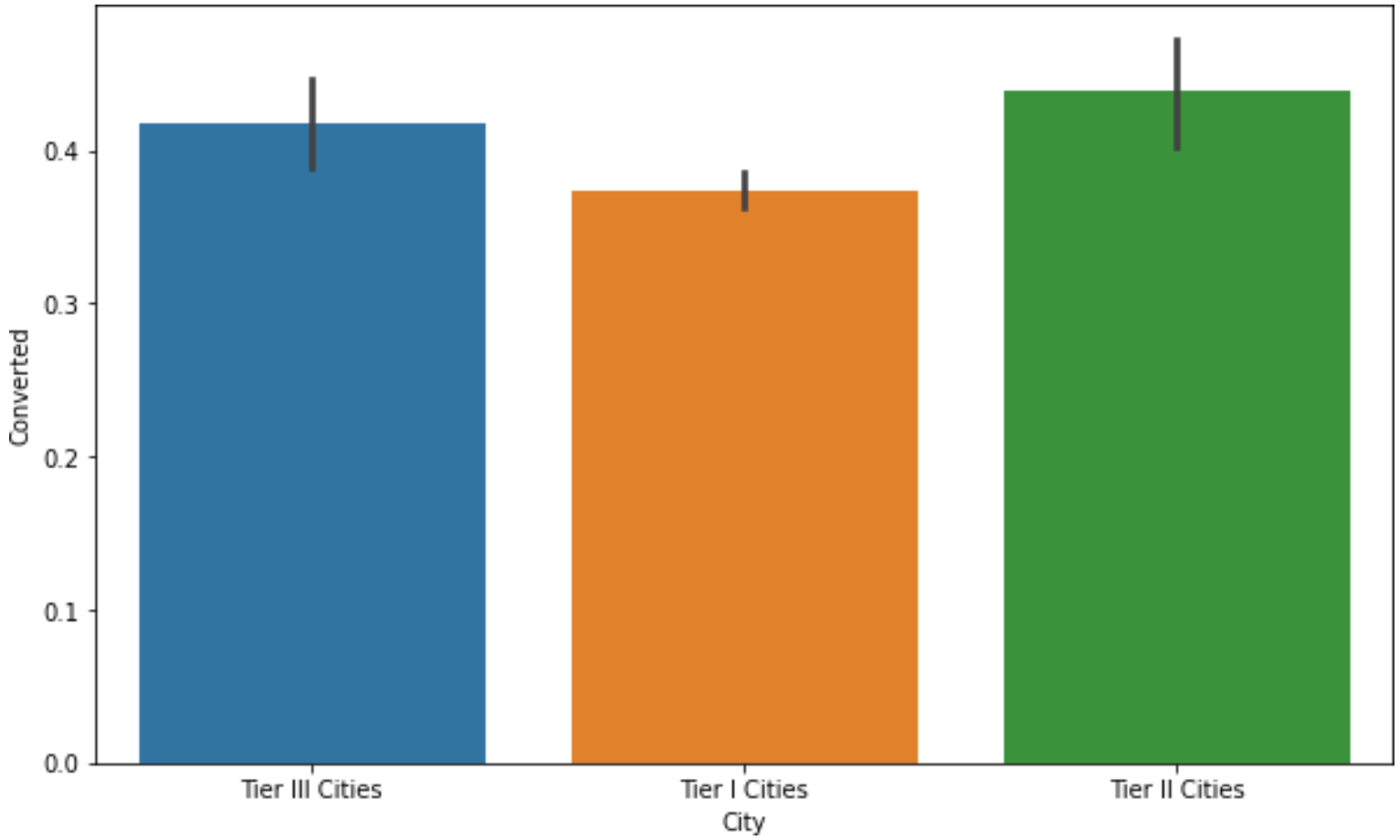
- In API and Landing Page Submission we can observe that it brings out highest number of leads as well as conversion.
- Lead add form has lower count of lead but it has higher conversion rate.
- So by this observation we have to generate more leads from Lead Add Form and improve lead conversion rate of API and Landing Page Submission origin in order to increase overall lead conversion rate

‘Lead Source' column



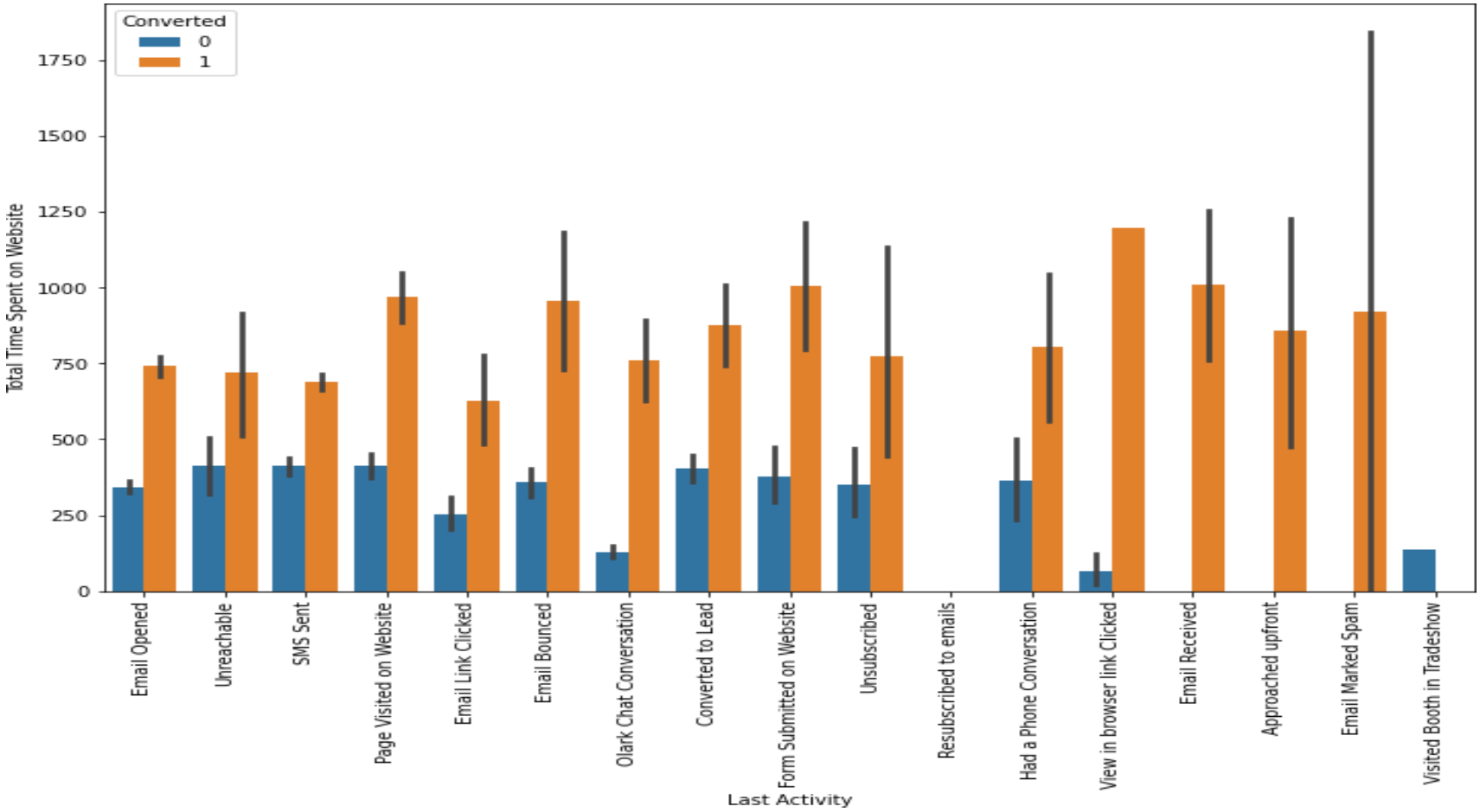
- Organic Search, Direct Traffic and Google has high conversion rate
- Reference and Olark Chat we can observe that it is generating maximum number of leads.
- Social Media have lowest conversion rate as compared to others.
- So by this observation we have to focus more on 'Social Media', 'Organic search', 'Direct traffic' and 'Google' leads to improve overall lead conversion rate.

‘CITY’ Column



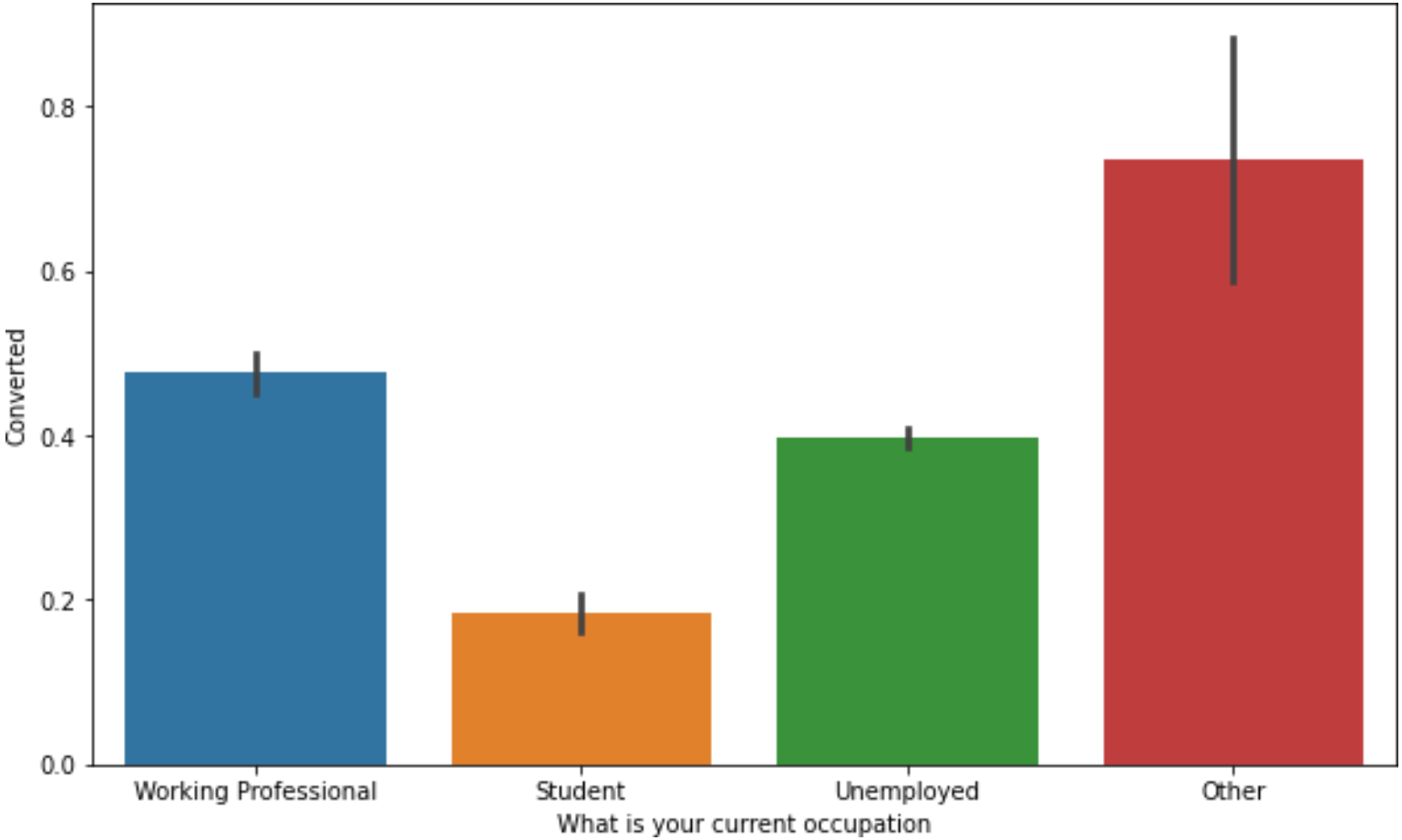
- **Tier I Cities, Tier II Cities, Tier III Cities are equally important for analysis.**

‘ Last Activity ‘ Column



- **Conversion rate of View in browser link clicked is Highest compared with others followed by Olark Chart Conversation.**

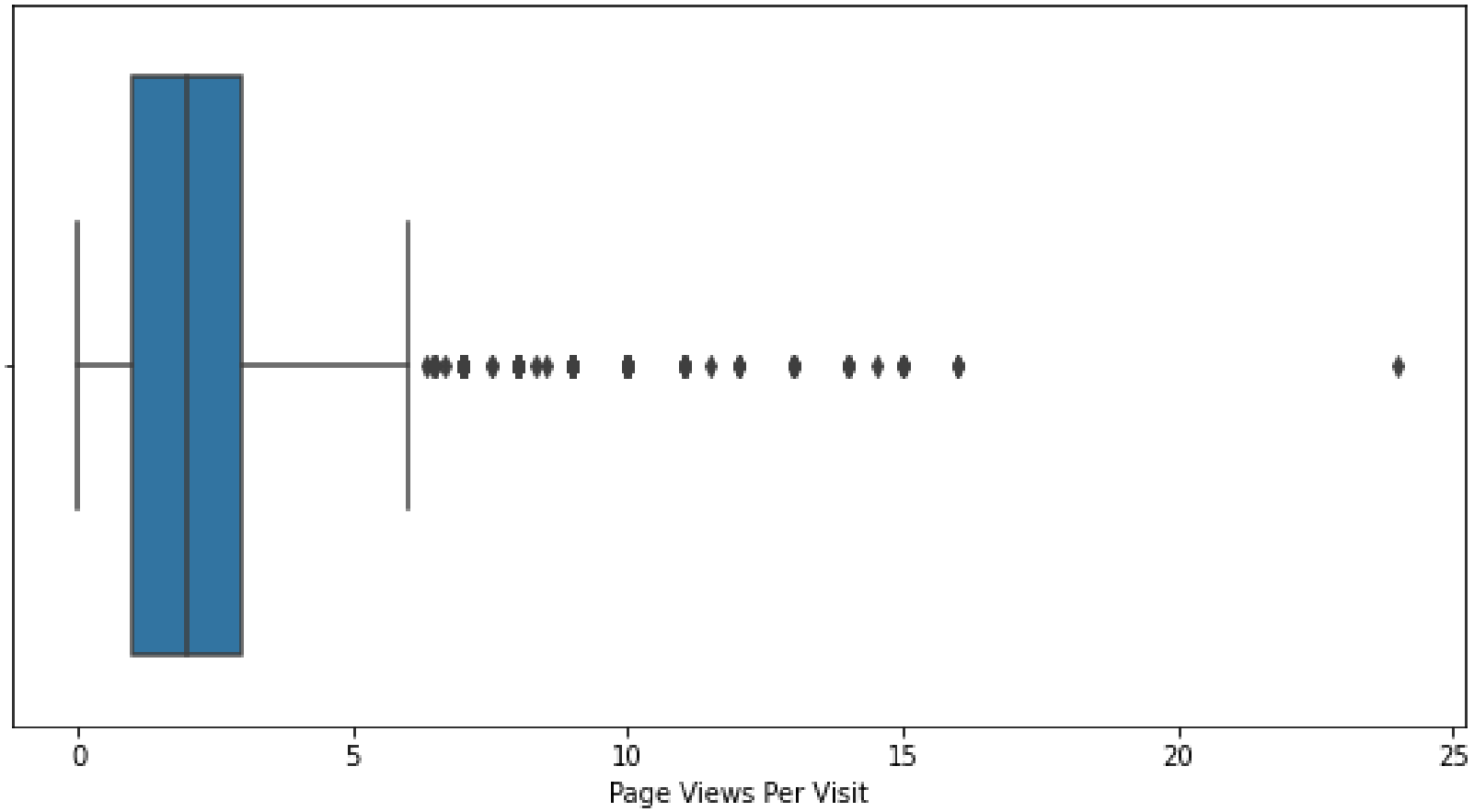
What Is Your Current Occupation



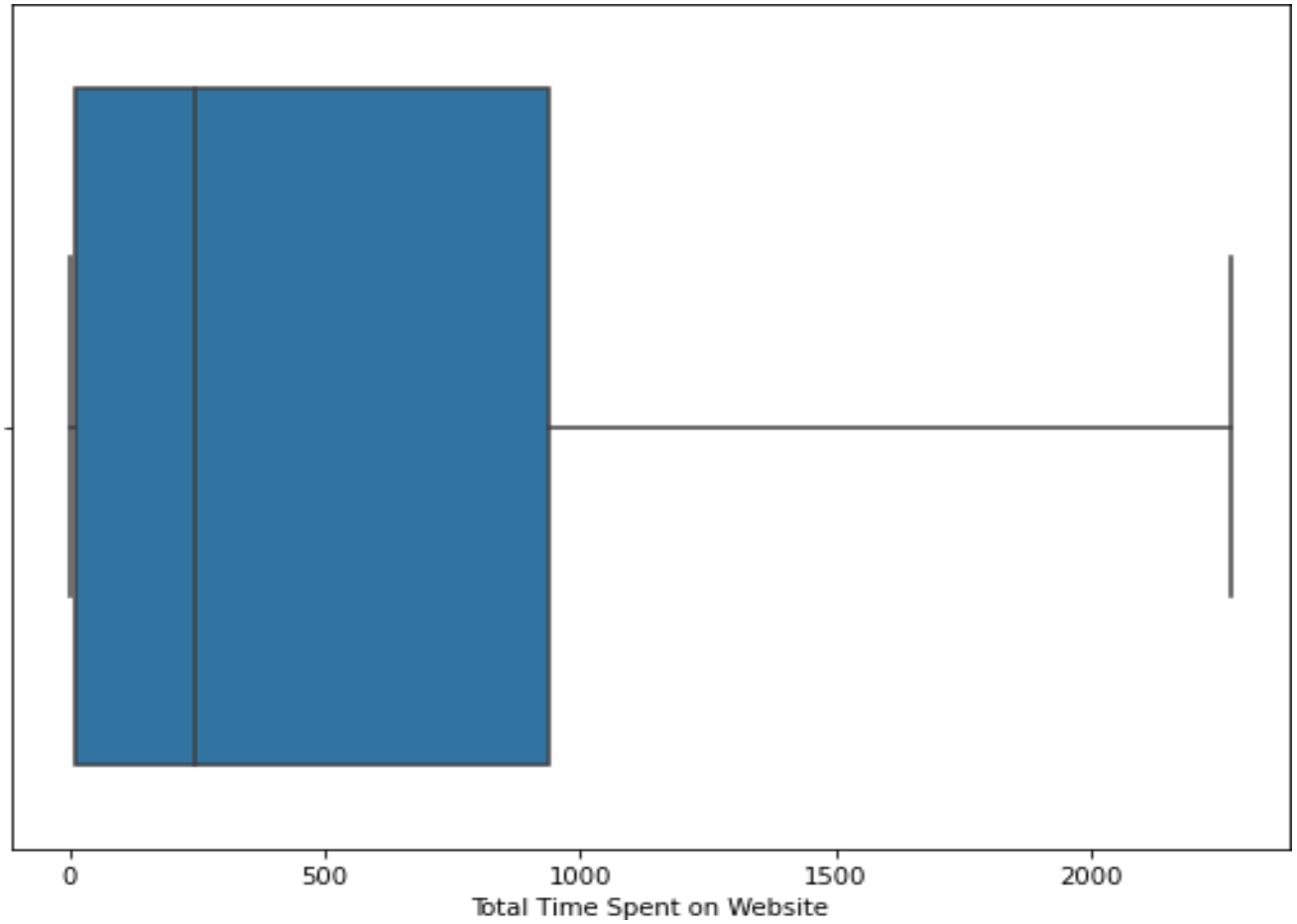
- **Other Category are highest in number followed by Working Professional .**

Univariate Analysis

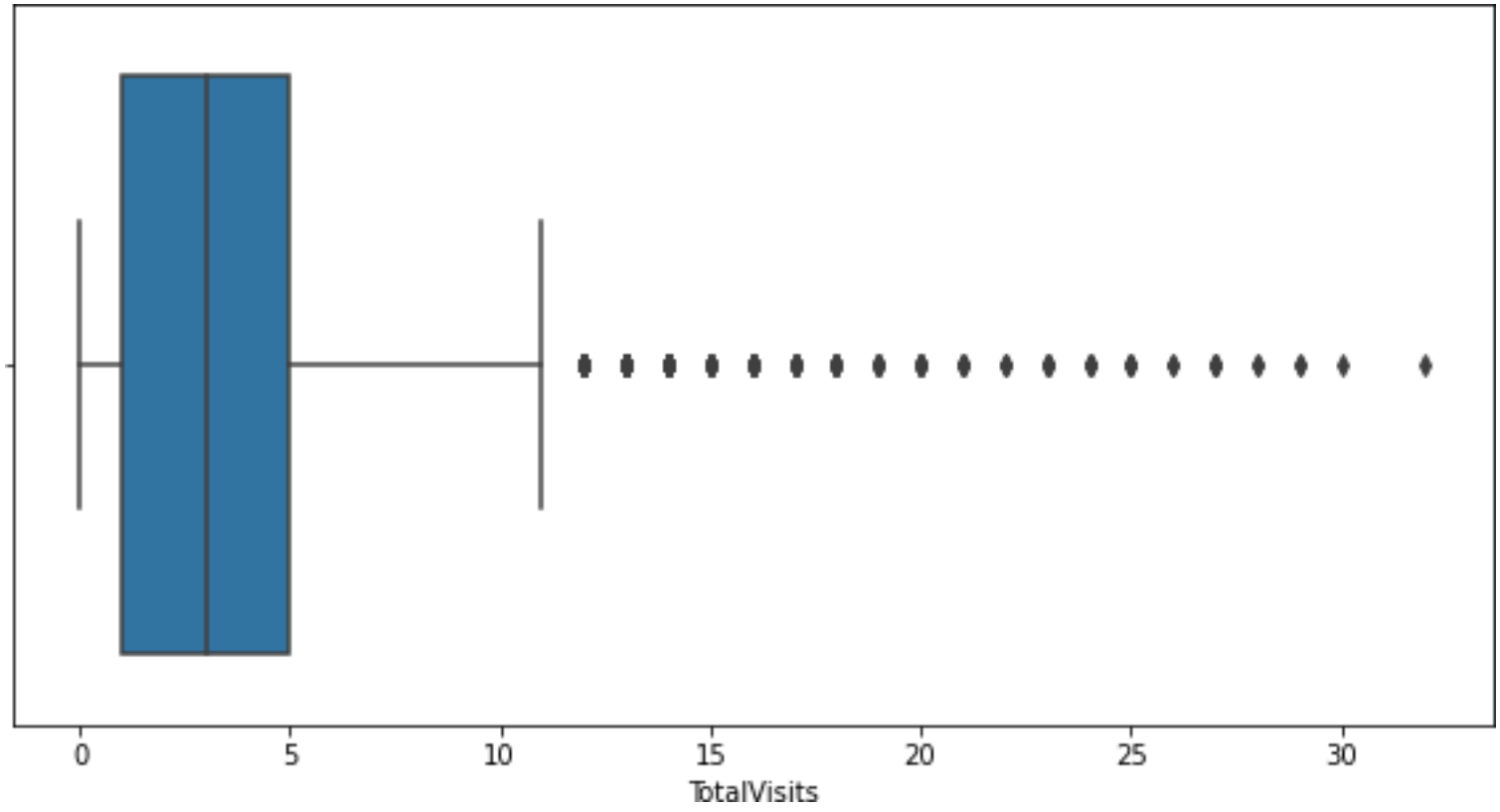
Page Views Per Visit



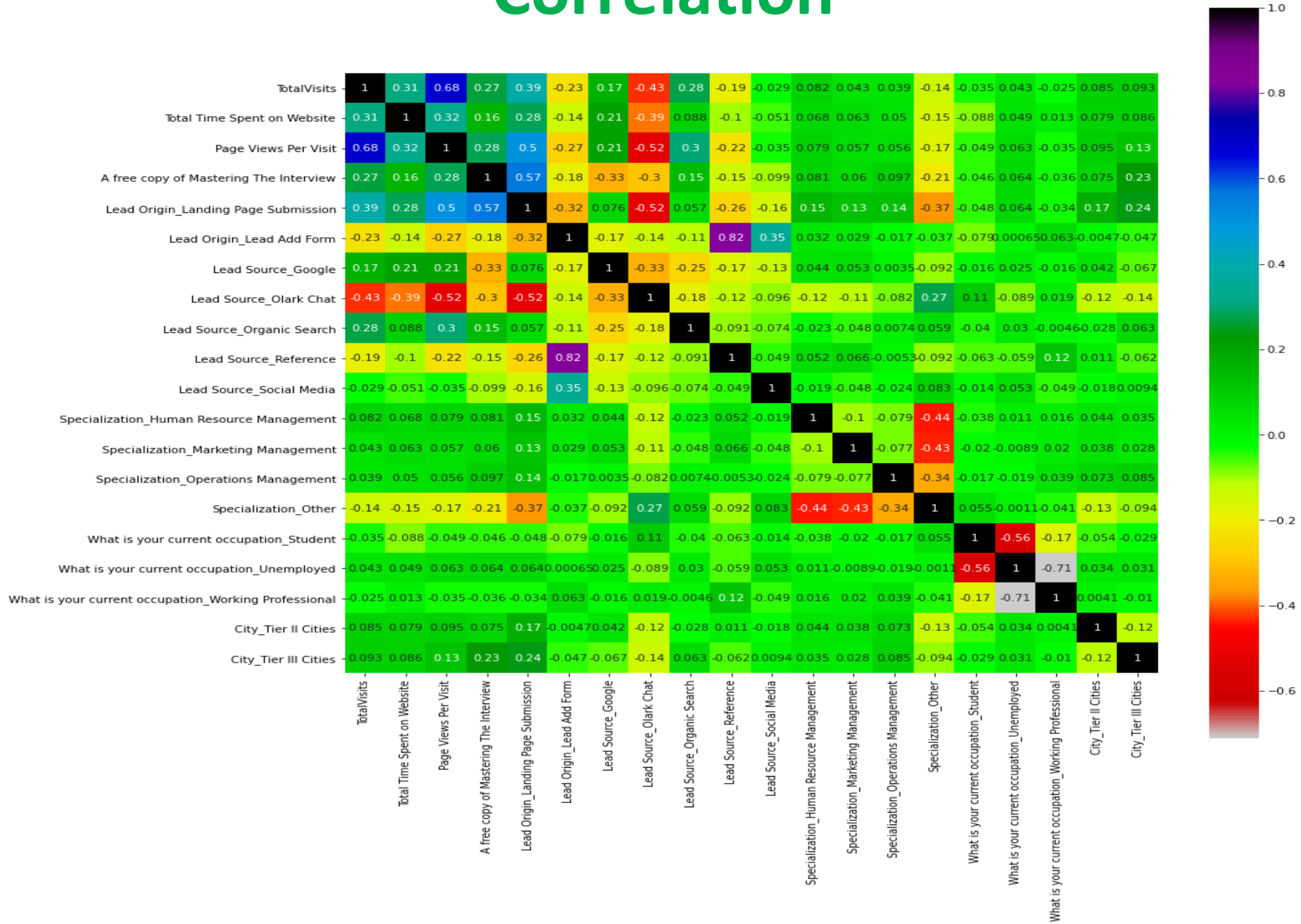
Total Time Spend On Website



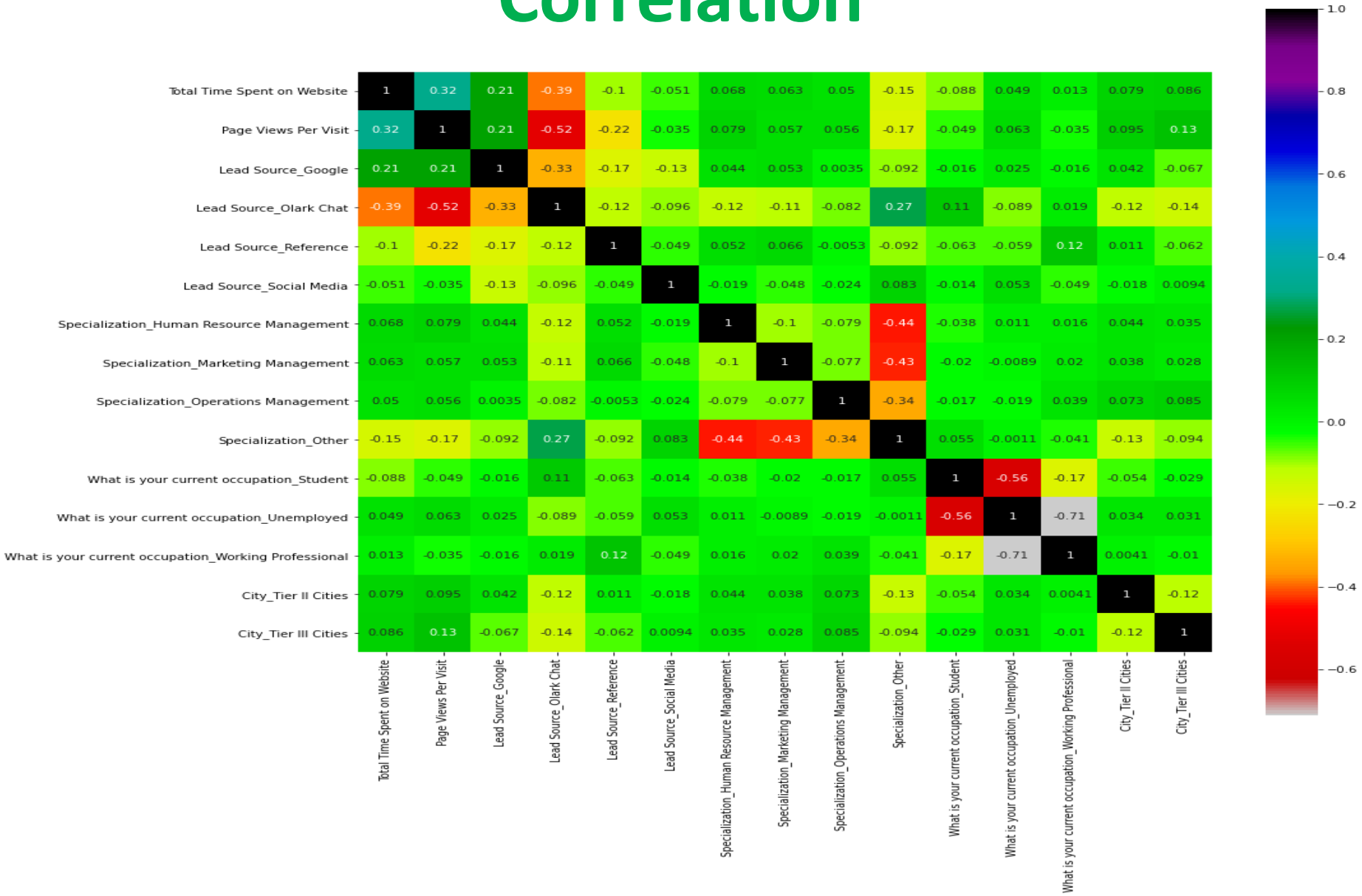
Total Visits



Correlation



Correlation



❑ **Heat map After Removing Multicollinearity**

Final Model Visualization

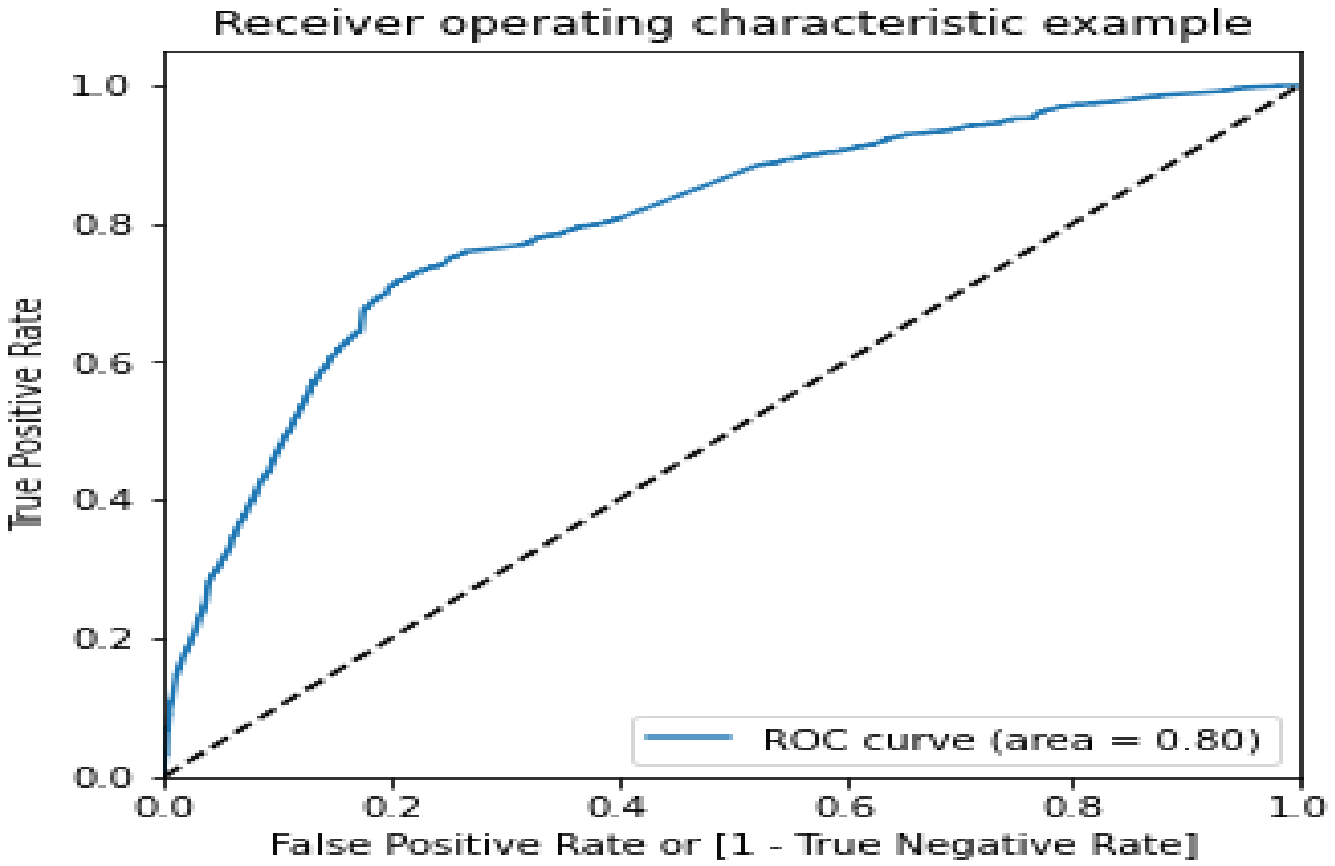
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6461
Model:	GLM	**Df Residuals:**	6450
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3379.9
Date:	Wed, 13 Oct 2021	**Deviance:**	6759.7
Time:	20:21:37	**Pearson chi2:**	6.60e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3075	0.114	-11.434	0.000	-1.532	-1.083
Total Time Spent on Website	4.3239	0.147	29.464	0.000	4.036	4.612
Page Views Per Visit	-0.8754	0.441	-1.983	0.047	-1.741	-0.010
Lead Source_Google	0.2869	0.071	4.054	0.000	0.148	0.426
Lead Source_Olark Chat	0.8277	0.109	7.579	0.000	0.614	1.042
Lead Source_Reference	4.0007	0.216	18.514	0.000	3.577	4.424
Lead Source_Social Media	1.6501	0.153	10.820	0.000	1.351	1.949
Specialization_Other	-0.3328	0.066	-5.058	0.000	-0.462	-0.204
What is your current occupation_Student	-1.2779	0.126	-10.180	0.000	-1.524	-1.032
What is your current occupation_Unemployed	-0.3242	0.077	-4.184	0.000	-0.476	-0.172
City_Tier III Cities	0.2085	0.091	2.287	0.022	0.030	0.387

	Features	VIF
8	What is your current occupation_Unemployed	3.40
6	Specialization_Other	2.81
1	Page Views Per Visit	2.79
0	Total Time Spent on Website	2.03
3	Lead Source_Olark Chat	1.82
2	Lead Source_Google	1.66
7	What is your current occupation_Student	1.42
9	City_Tier III Cities	1.17
5	Lead Source_Social Media	1.11
4	Lead Source_Reference	1.09

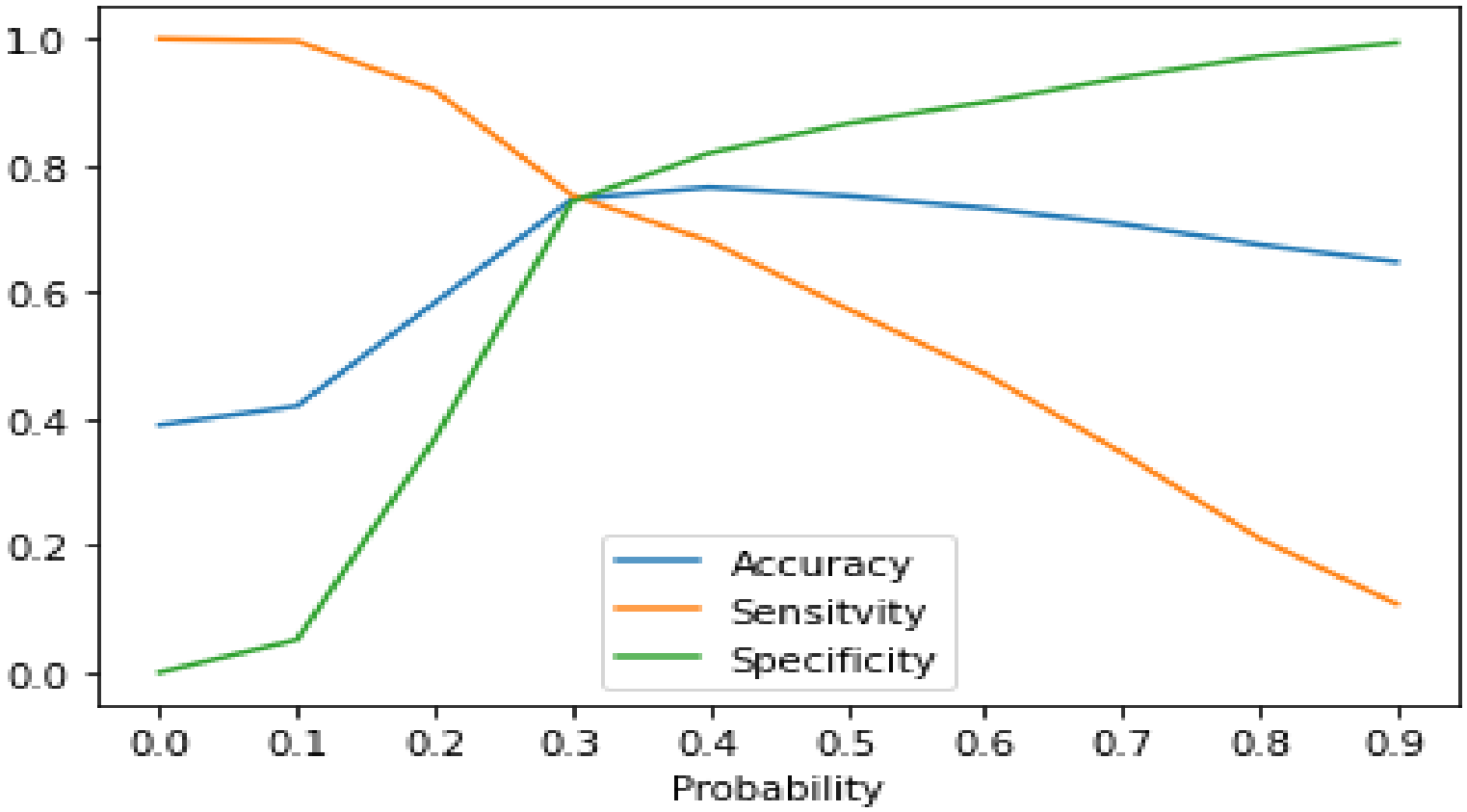
Plotting the ROC Curve



1.The curve is closer to the left side of the border than to the right side hence our model is having great accuracy.

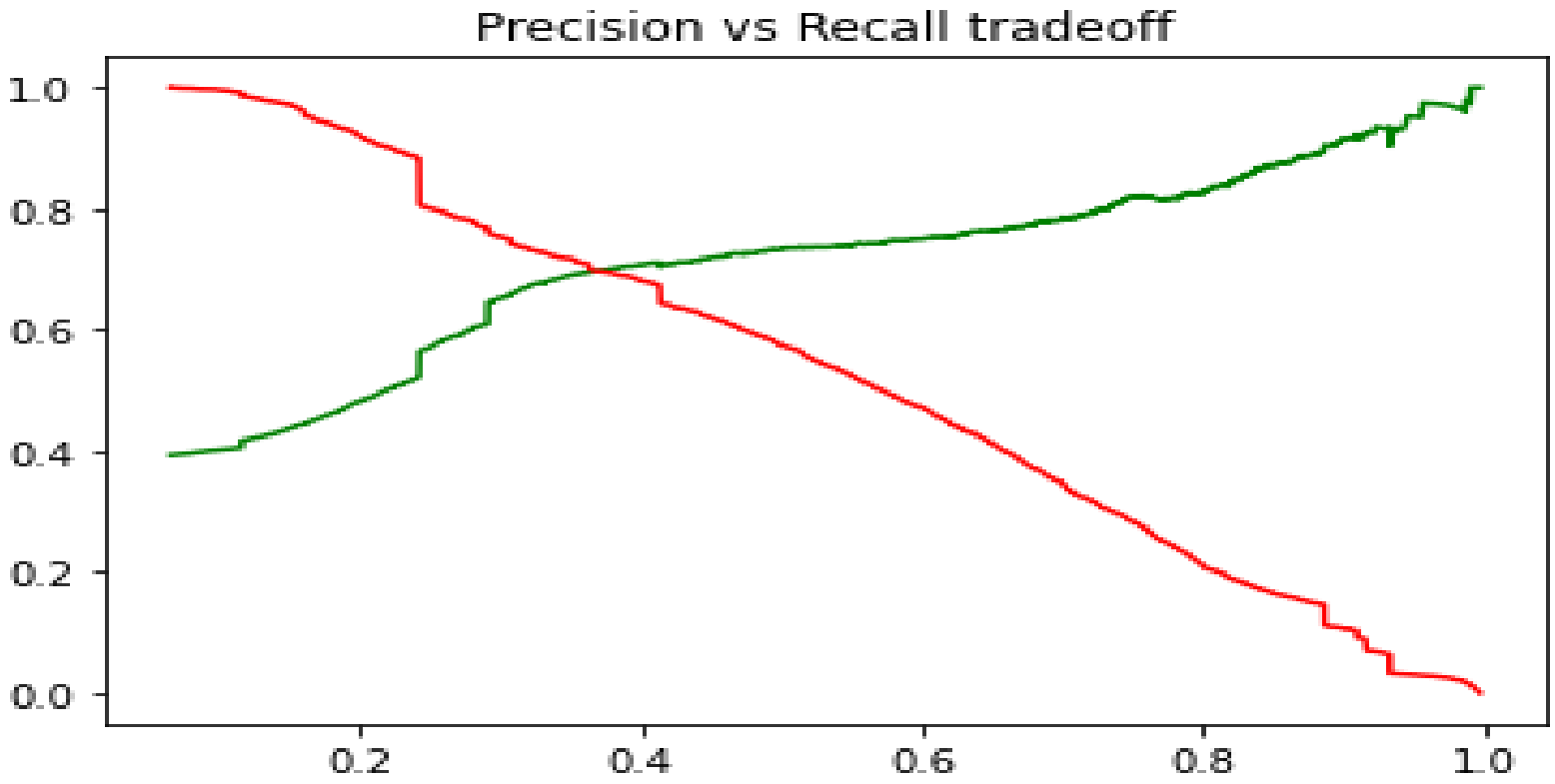
2. The area under the curve is 80% of the total area.

Optimal Cutoff Point



- From above plot we can observe probability threshold is nearly 0.3.
- Considering 0.3 to tradeoff sensitivity against accuracy.

Precision and Recall Tradeoff



- From above plot we can observe there is tradeoff between Precision and recall and optimum point is nearly at 0.4

Observations

- Comparison for the values obtained in train & test:

- **Train Data:**
  1. Accuracy: 72.65%
  2. Sensitivity: 76.40%
  3. Specificity: 70.24%

- **Test Data:**
  1. Accuracy: 77.65%
  2. Sensitivity: 68.49%
  3. Specificity: 83.14%

Conclusion

- Important features we got during training of model which is responsible for good conversion rate are as follows
  1] Total Time Spend On website
  2] Lead Source_Reference
  3] Lead Source_ Social Media

- We got the Recall value greater than Precision Value also it is acceptable for Business aspect.
- Sensitivity, Specificity, Accuracy we got from test data set when compared with train data set it is in acceptable range.
- The model has ability to adjust as per company requirement and it will give good result.

THANK YOU