Lead Score Case Study

Logistic Regression Assignment

SUBMITTED BY :

SONIA DHANDA

ABHISHEK VAGGAR

HEMANTH KUMAR

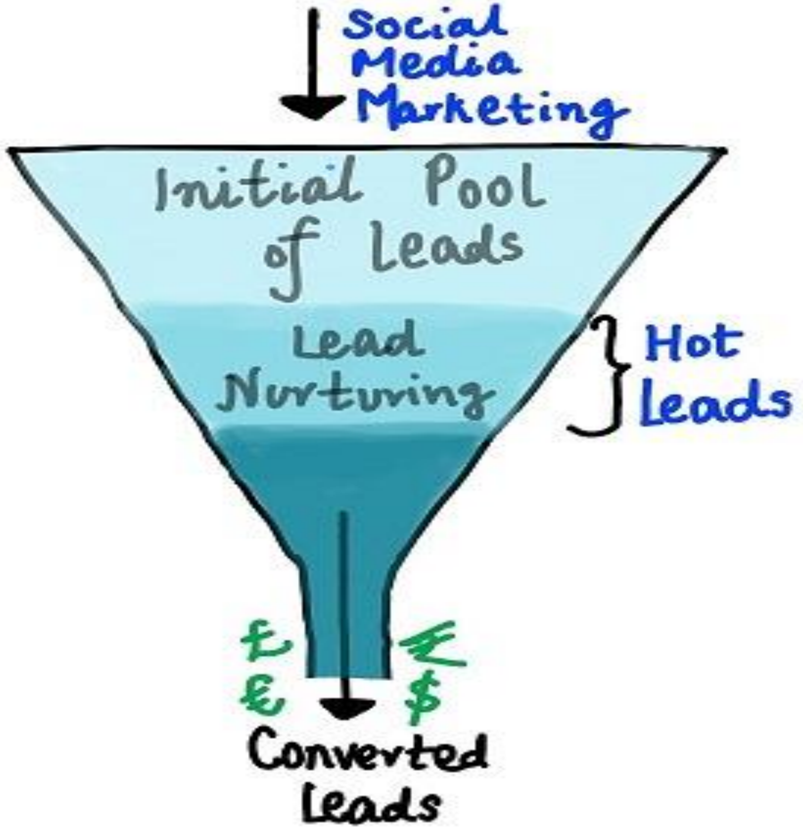
Problem Statement :

An education company named X Education sells online courses to industry professionals. 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.

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%.

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.

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



Objective :

1) 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.

2) The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data Understanding :

1)Data :
You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

2) Files given :

- one lead score data csv file

- data dictionary

- Problem statement

Solution Methodology :Steps Involved

1) Data Reading & Understanding

2) Data Cleaning – checking & handling null/select values

3) EDA – Univariate / Multivariate Analysis for Categorical Variables & Numerical Variables

4) Outlier Handling

5) Data Preparation for Logistic Regression – Dummy creation

6) Model Building – Test Train split

7) Feature Scaling of Continuous Variables

8) Feature Selection using RFE & Automated approach

9) Plotting the ROC Curve

10) Finding Optimal Cut-off point

11) Precision & Recall

12) Making Predictions on Test set

Observations :

1) There are 9240 rows and 37 columns in the data set. There are null values present in few columns that need to be handled

2) From the given data we can see that Prospect ID & Lead Number columns seems to have unique values so we can drop them as they are just a row identifier (indicative of the ID number of the Contacted People)

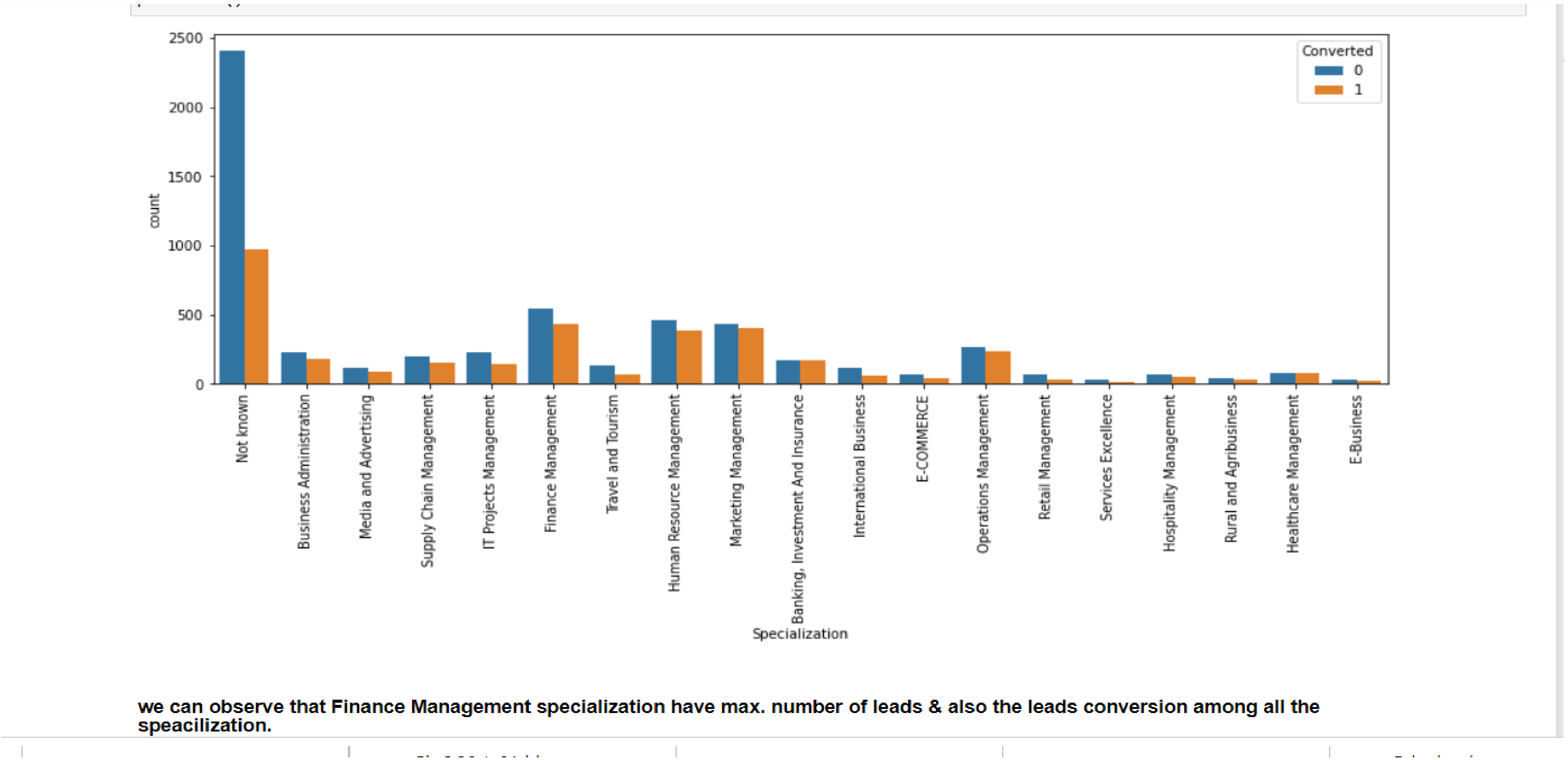
3) Dropping columns with more than 40% missing values

4) Converting 'Select' values to NaN

5) Grouping values with low count

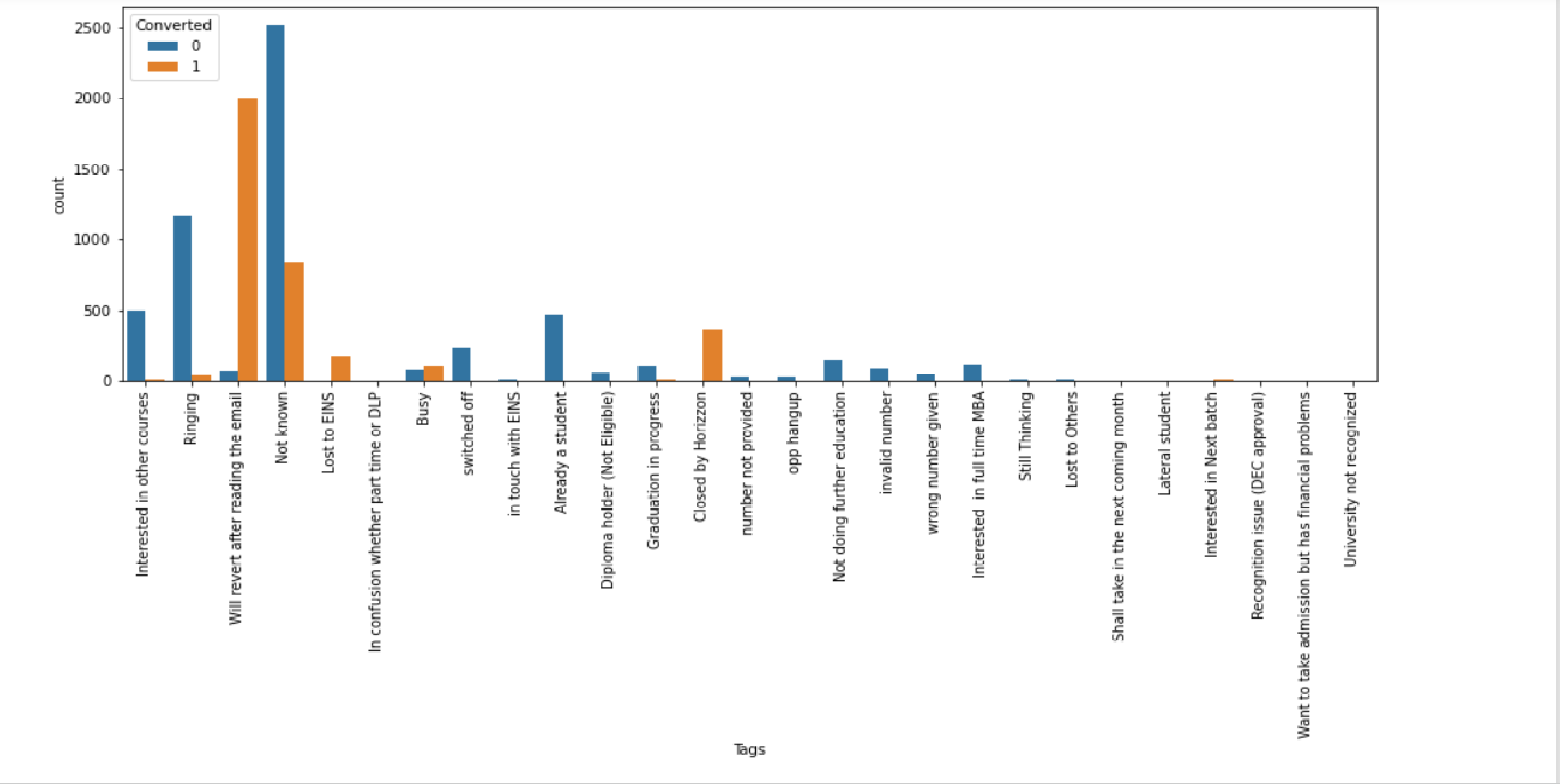
6) Exploratory Data Analysis – categorical & numerical variables

Plotting spread of specialization :



we can observe that Finance Management specialization have max. number of leads & also the leads conversion among all the speacilization.

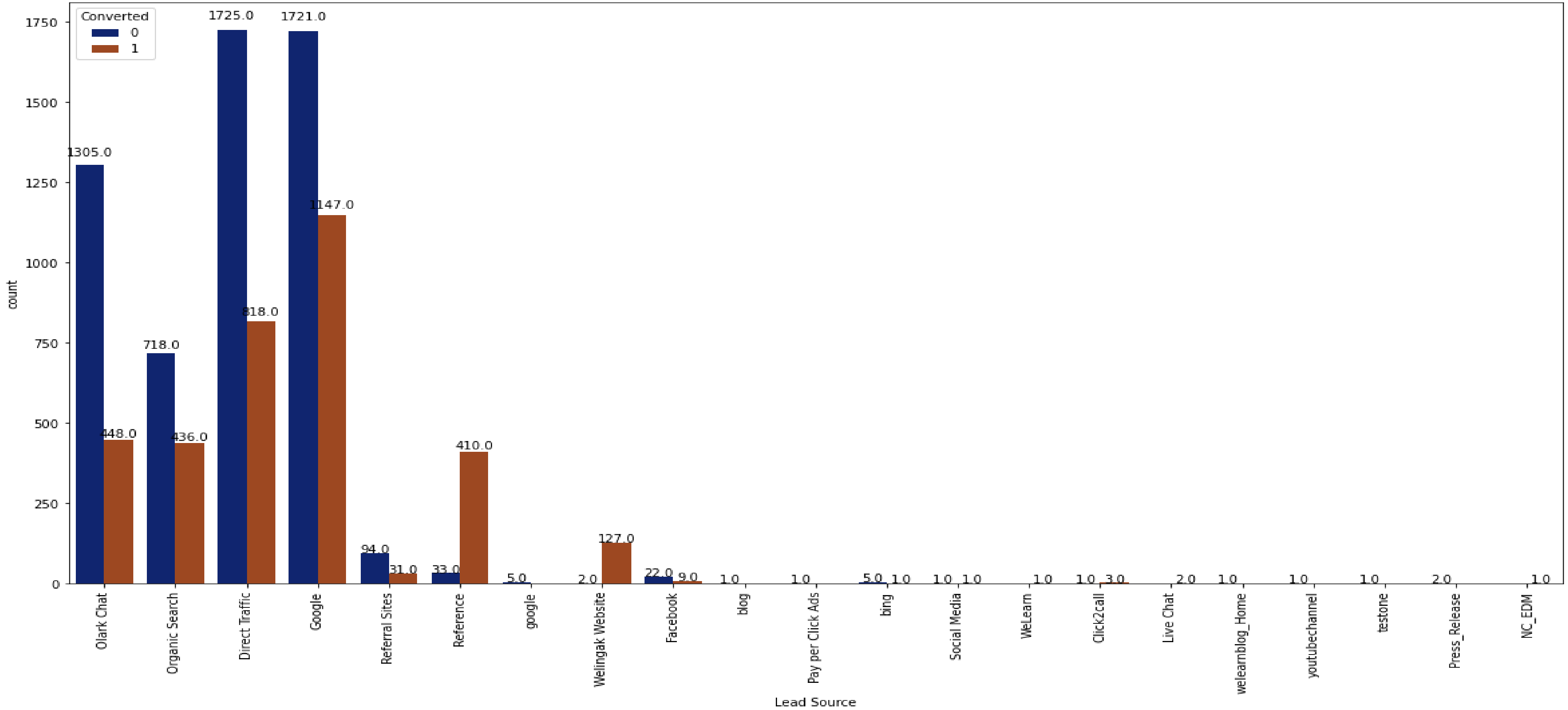
visualizing the tags column



Univariate Analysis

- creating function to plot categorical variables against converted(leads)

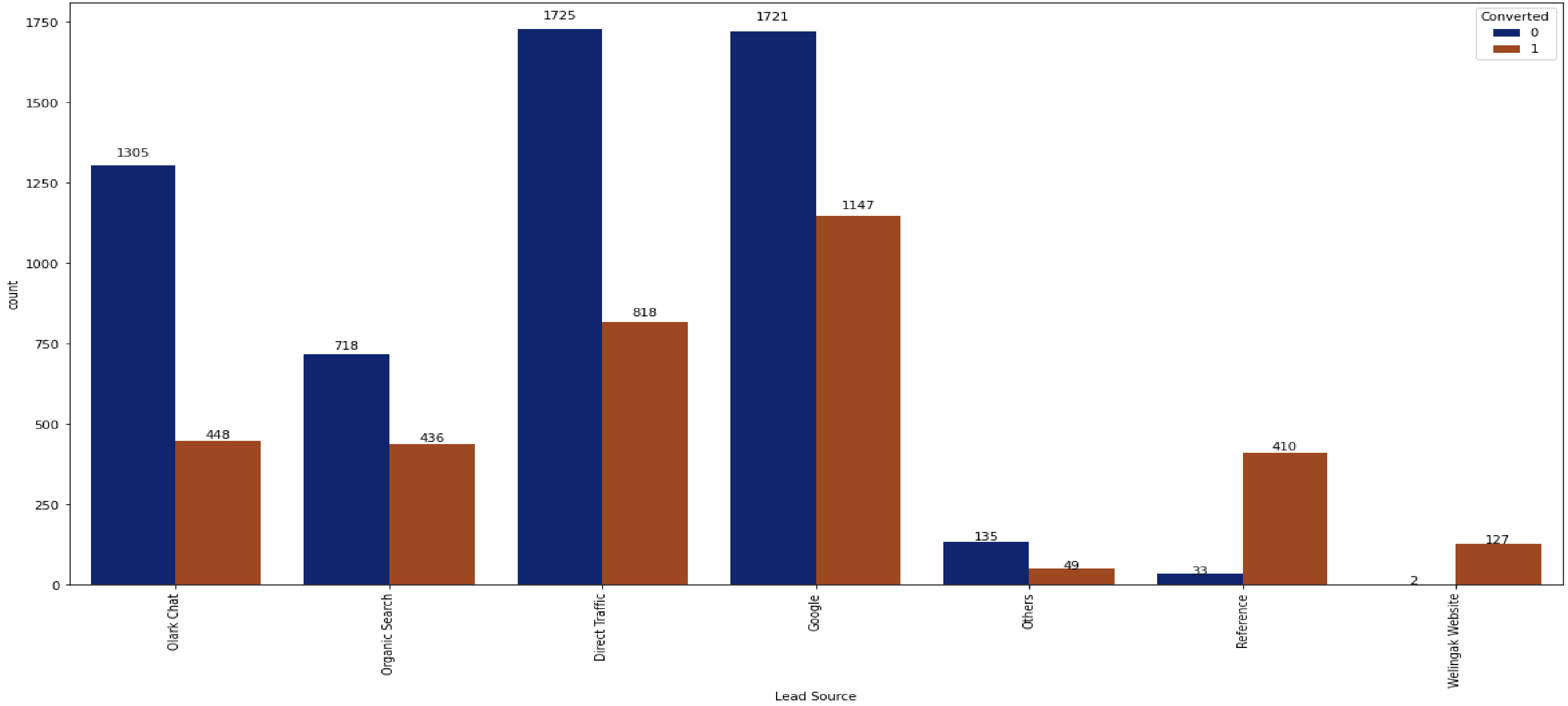
Lead Source Vs Converted



Lead Source Vs Converted (after grouping the low value count)

Inference:

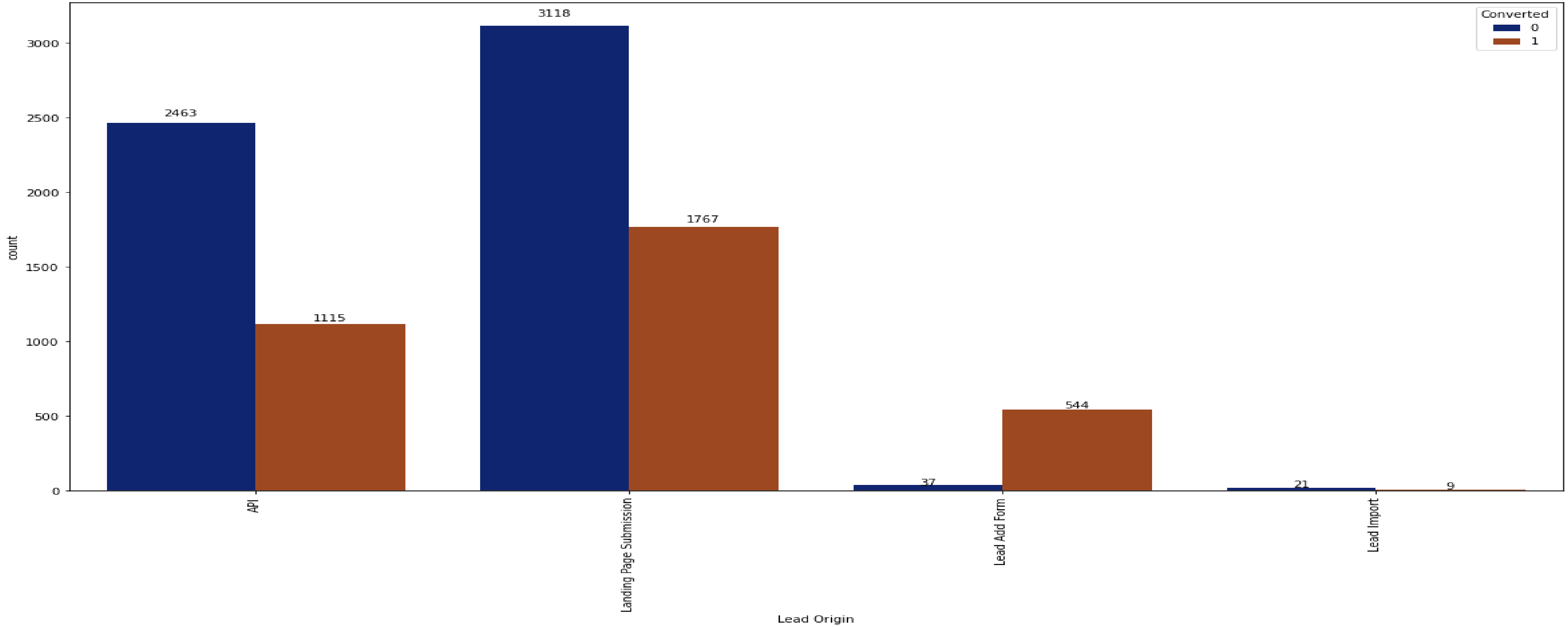
- Source of Lead via Google and Direct Traffic have higher number of leads coming but have poor conversion as compared to others.
- Leads coming via Reference - have the highest conversion rate.



Lead Origin Vs Converted

Inference:

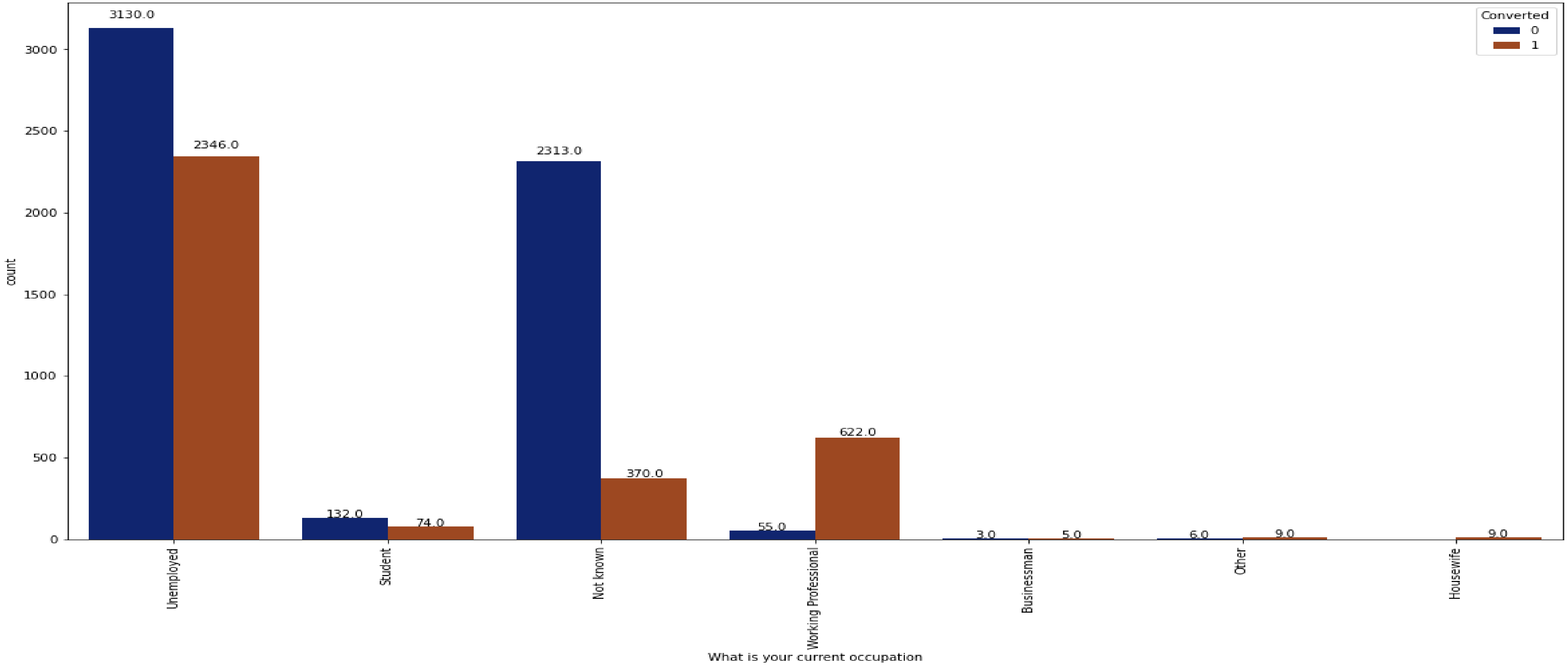
- Origin of Lead via 'Landing Page Submission' have higher number of leads coming but have poor conversion.
- Leads coming via 'Lead Add Form' - have the highest conversion rate.



Current Occupation Vs Converted

Inference:

- Leads coming from 'Unemployed' category have higher number of leads coming but have poor conversion(~42%).
- Leads coming from 'Working Professional' - have the highest conversion rate(~92%).

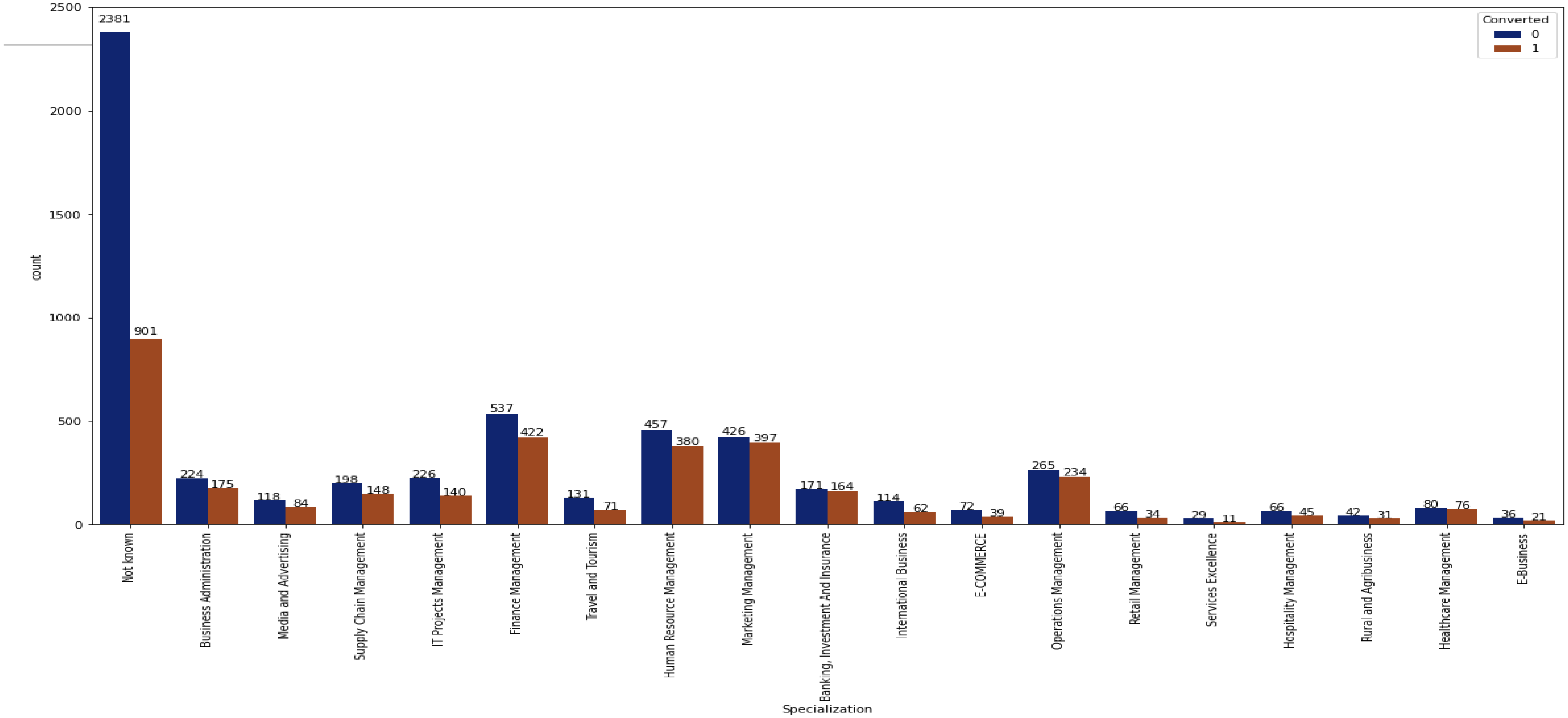


Specialization Vs Converted

Inference :

•Leads coming from customer having specialization in management field have good conversion rate.

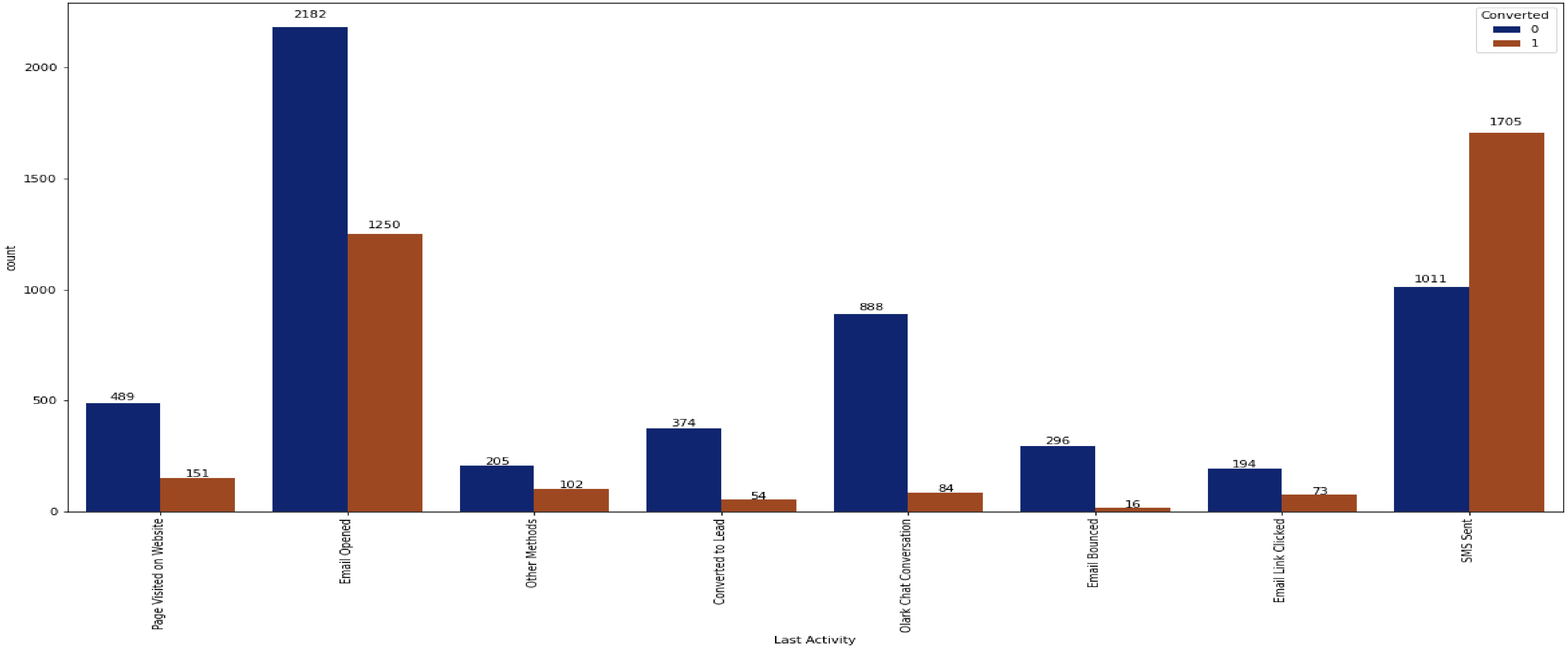
•Most of the customer's specialization is not known.



Last Activity Vs Converted

Inference:

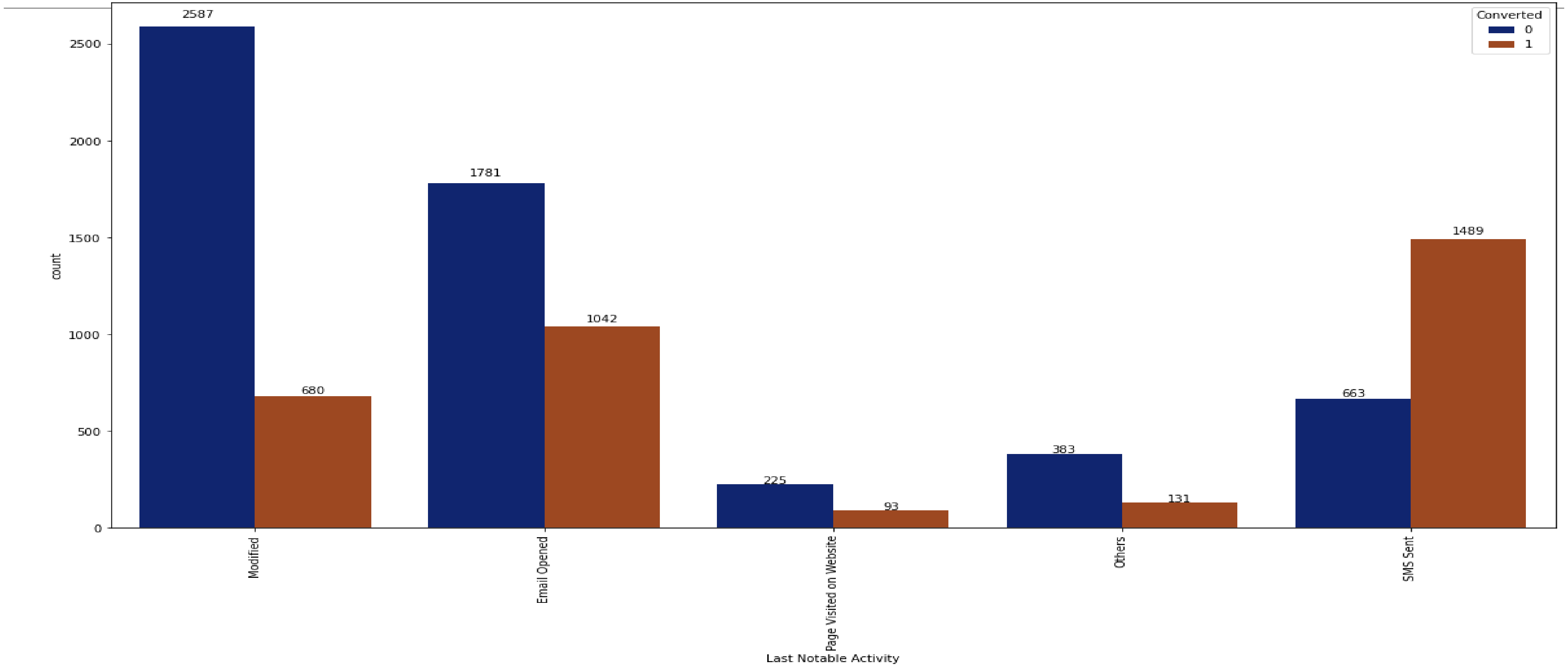
•Last activity performed by the customer as 'SMS Sent' have the higher conversion rate (~63%).



Last Notable Activity Vs Converted

Inference:

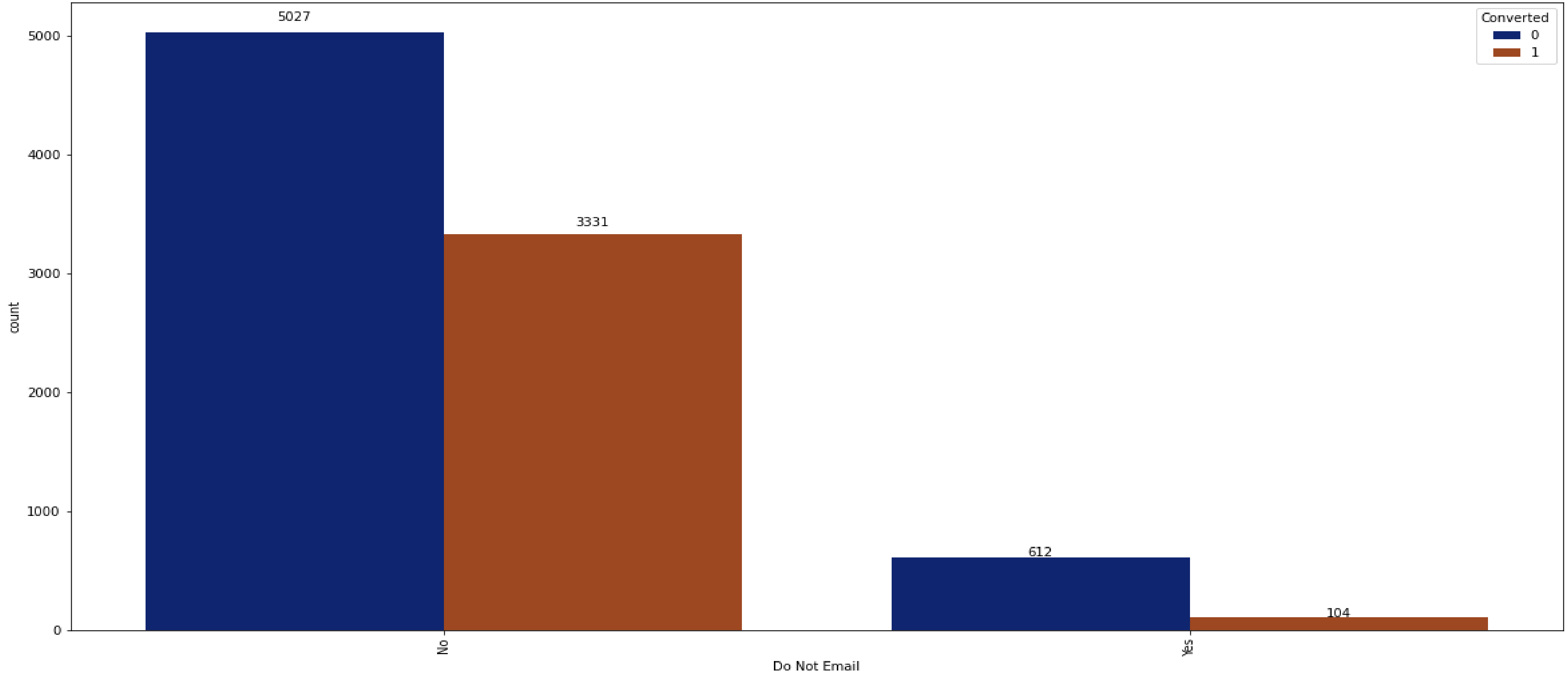
•The last notable activity performed by the student as SMS Sent have the higher conversion rate (~69%).



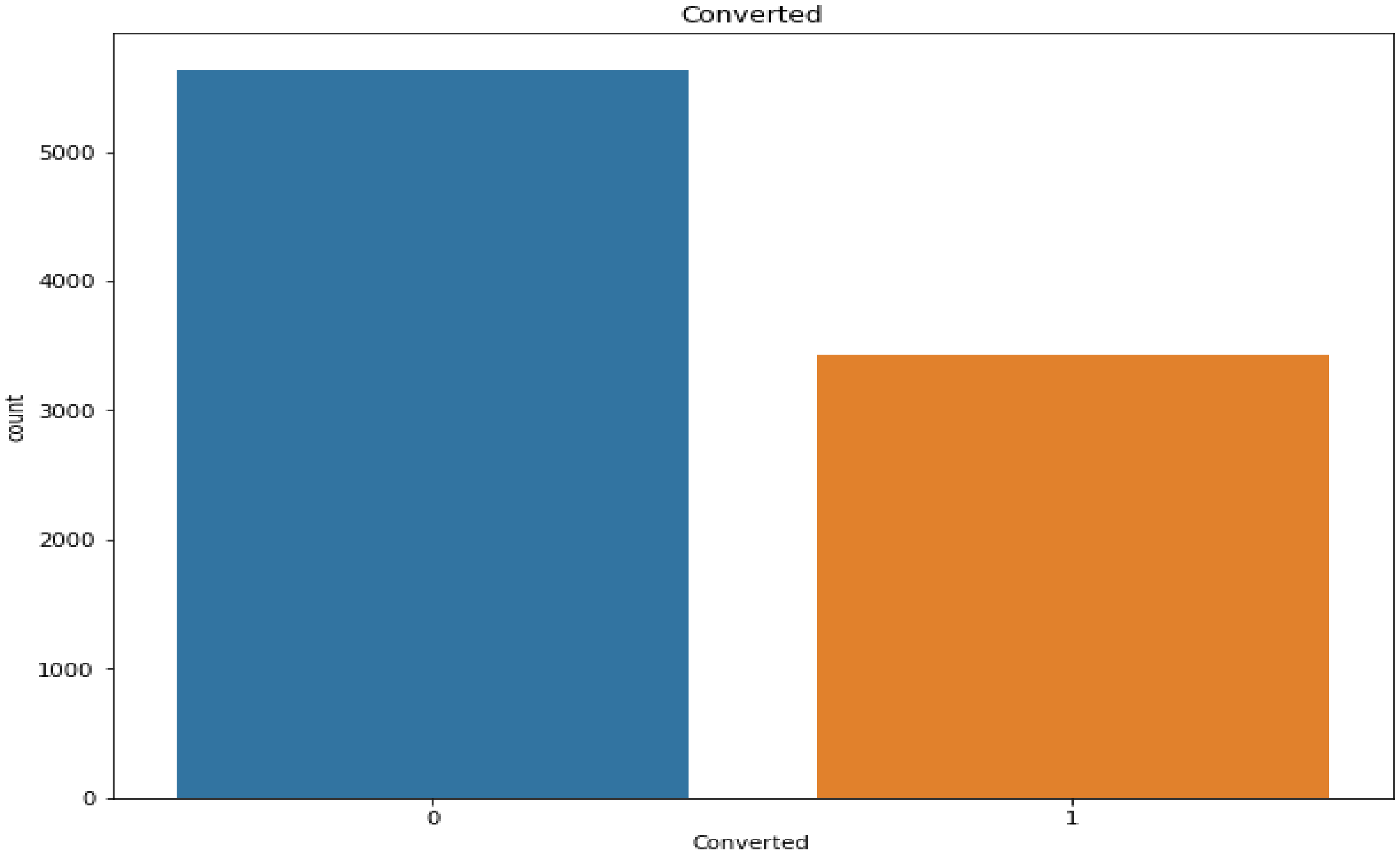
Do Not Email Vs Converted

Inference :

•Its interesting to see that customers who are not willing to be emailed about the course have conversion rate of ~40% where as for the customers who want to be emailed about the same have conversion rate of only ~15%.

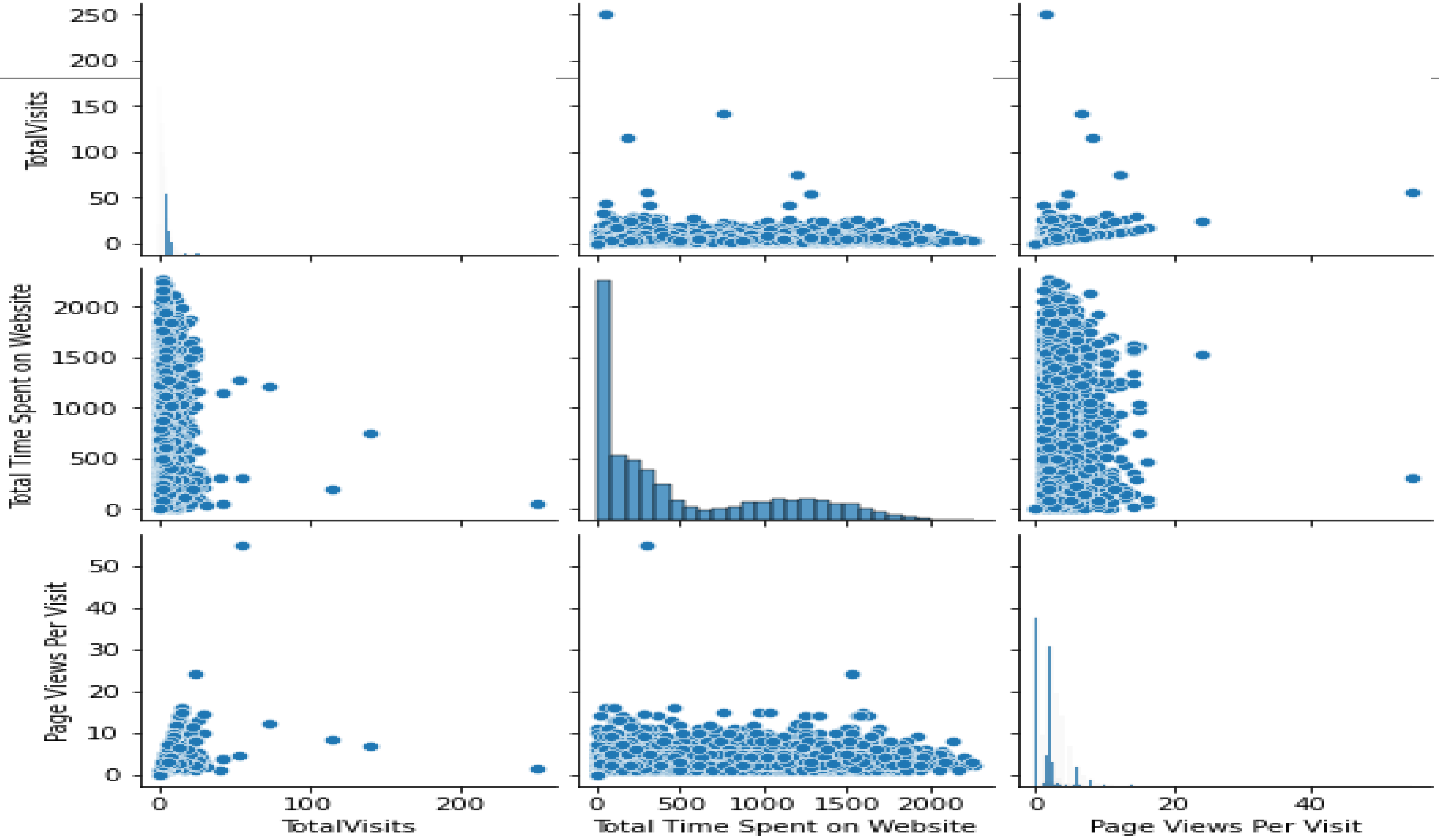


Present Lead Conversion Rate is ~38%



Checking numerical variables

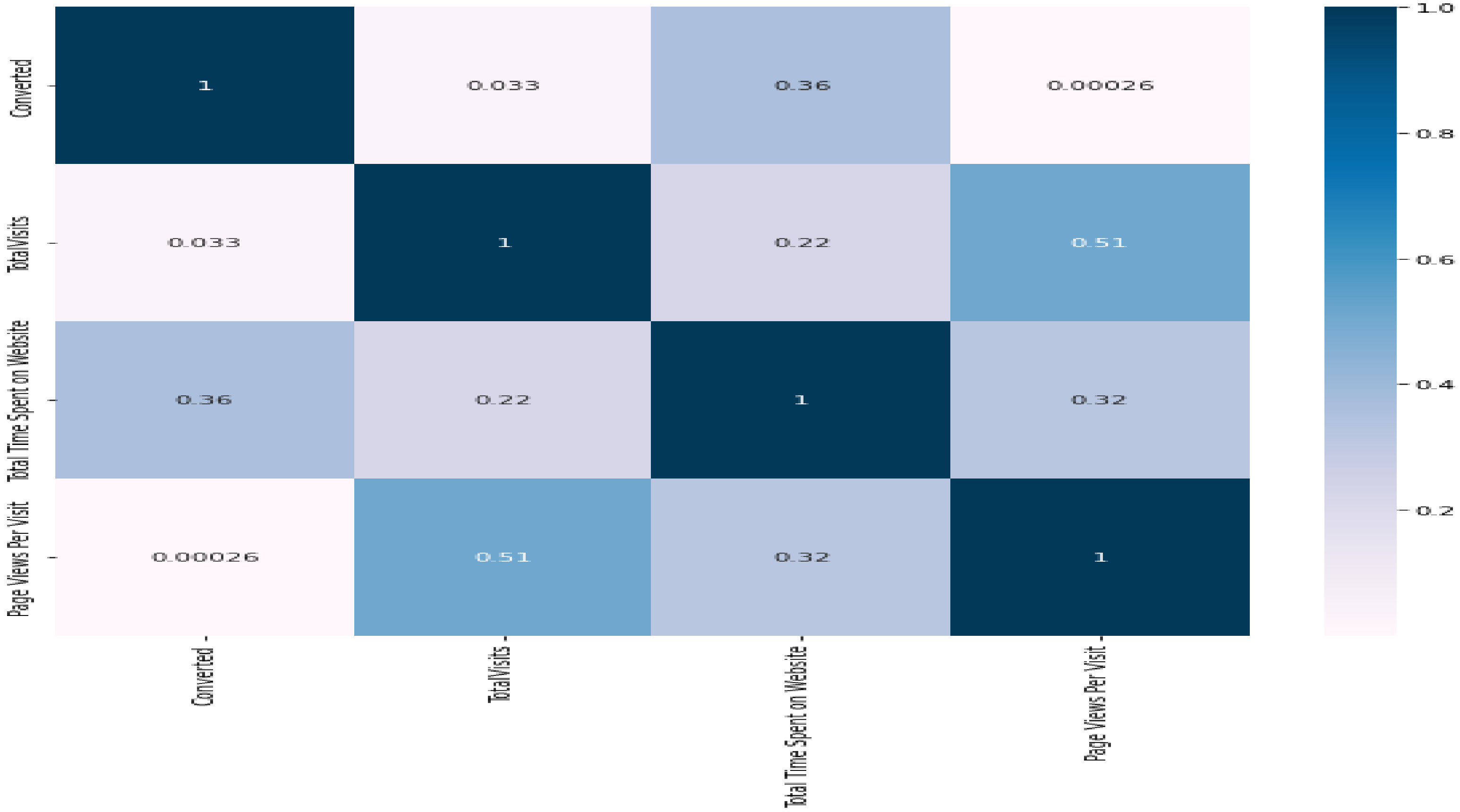
(pair plot to understand numerical variables)



Corelation among numerical variables

Inference :

•There is highest correlation seen between TotalVisits & Page Views Per Visit (0.51)

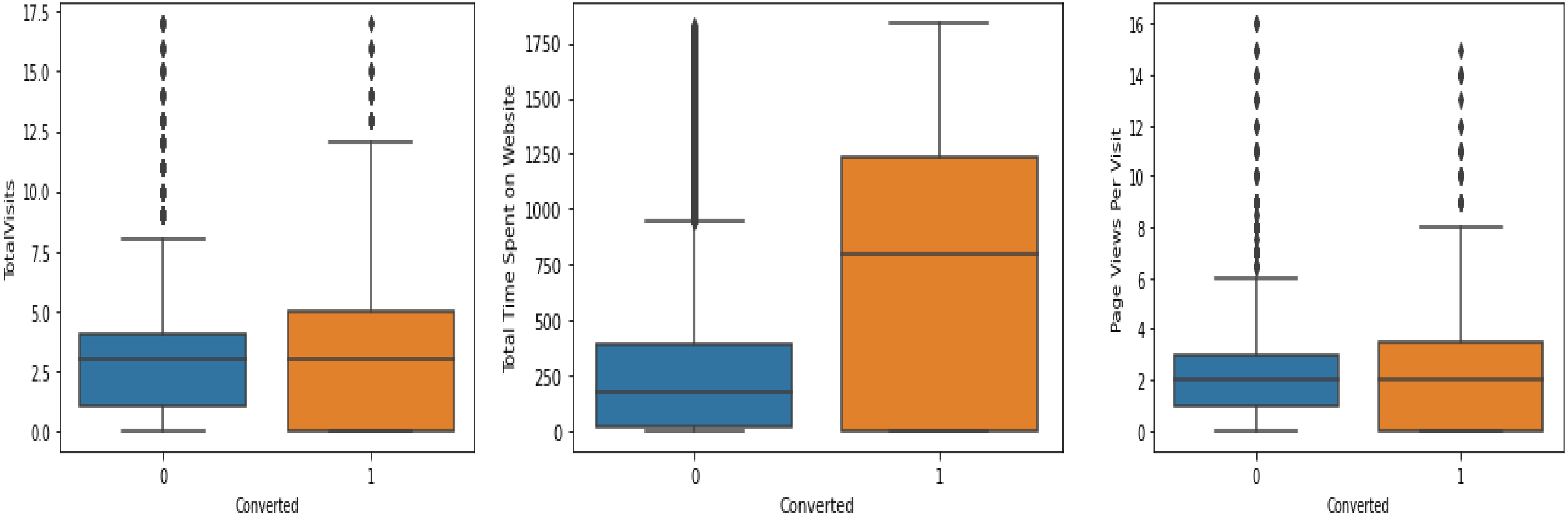


Boxplot to check outliers among numerical variables

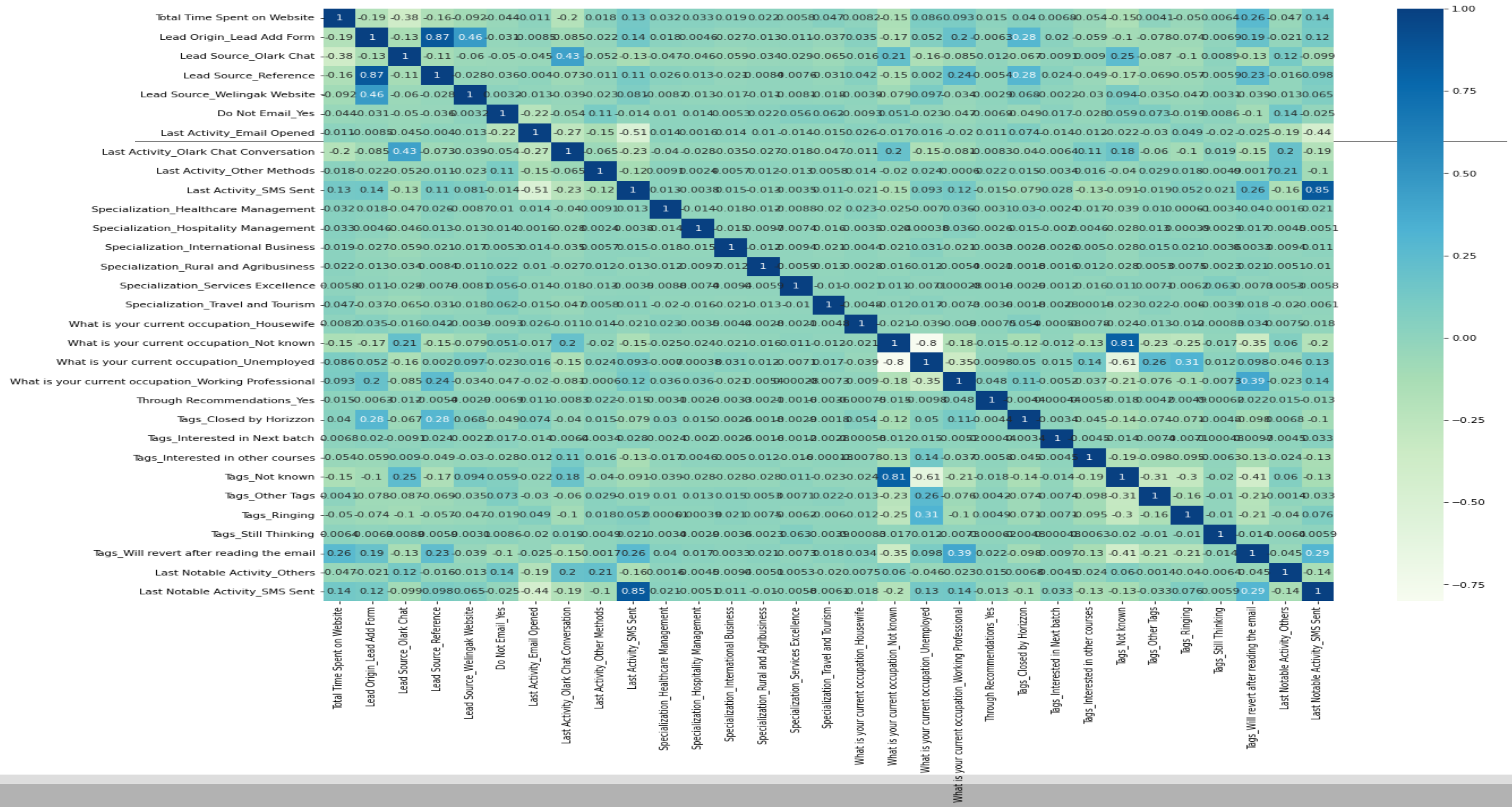
Inference :

•Median of both Converted and Not Converted for 'Page Views Per Visit' & 'TotalVisits are almost same.

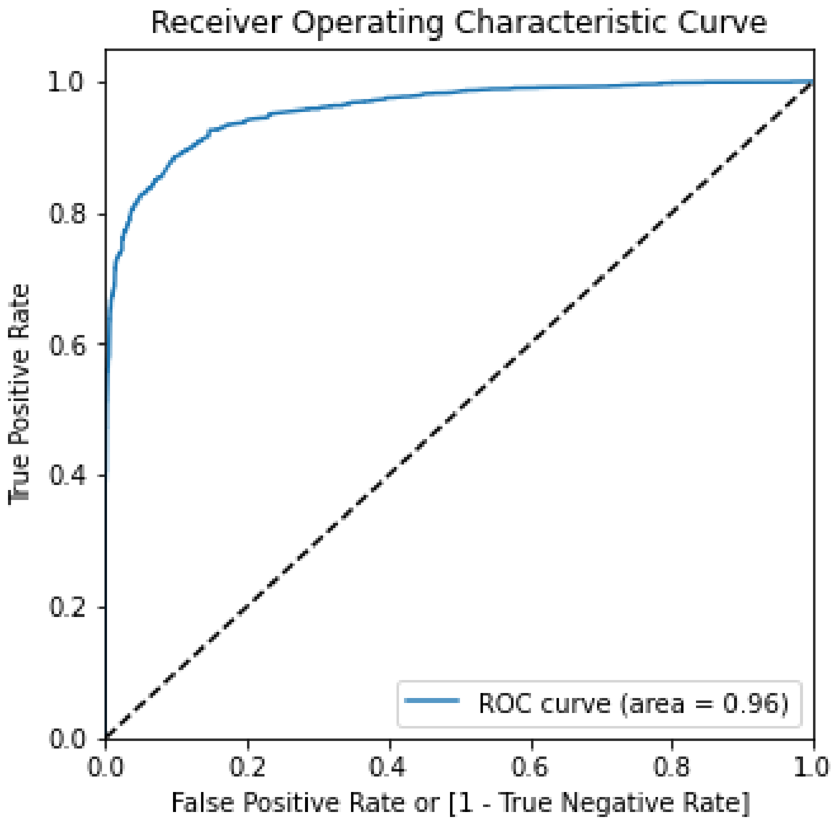
•Leads spending more time on the website tend to be converted.



Checking correlation of features selected by RFE with target column

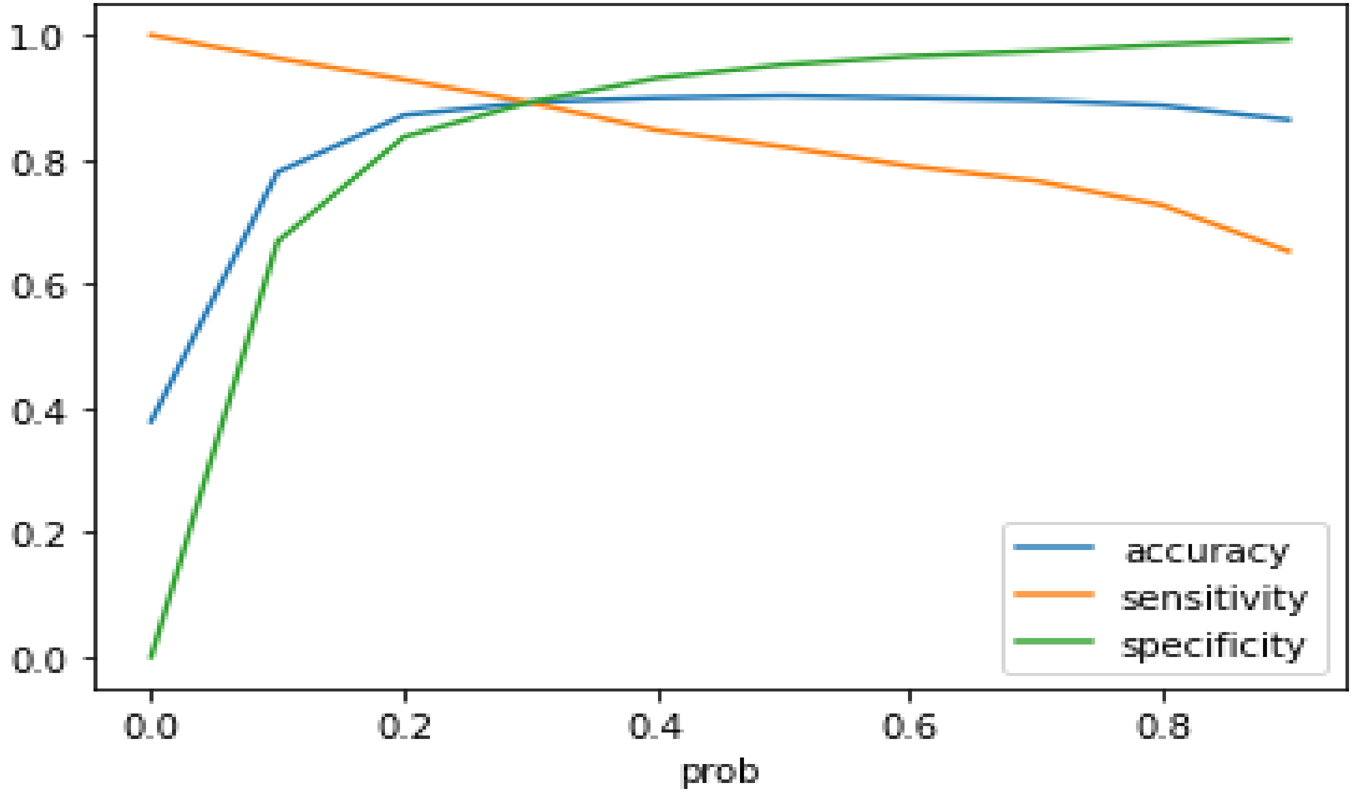


ROC Curve


Receiver Operating Characteristic Curve
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.96)

To find optimal point - plotting accuracy sensitivity and specificity for various probabilities

As we can observe that 0.3 is the optimal point



Conclusions :

As we can observe that, we have achieved an overall accuracy of about 0.87 on our Logistic Regression model. That is, there is 87% chance that our predicted leads will be converted. This meets the CEO's target of at least 80% lead conversion.

Final Observation:

Let us compare the values obtained for Train & Test:

Train Data:-

Accuracy : 89.19%

Sensitivity : 89.15%

Specificity : 89.22%

Test Data:-

Accuracy : 87.12%

Sensitivity : 89.36%

Specificity : 85.82%

Significant variables to predict the lead conversion are : -

- Tags_Closed by Horizon
- Tags Ringing
- Tags Will revert after reading the email
- Tags Interested in other courses
- Last Notable Activity SMS Sent
- Total Time Spent on Website
- Last Activity Olark Chat Conversation
- Lead Origin Landing Page Submission

Thank You