Lead Score Case Study Using Logistic Regression

SUBMITTED BY:

1. Gopinath Dhara
2. Rakesh Babu V Giraddi
3. Gowtham Venkata Sai Ram Maddala

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. 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Strategy

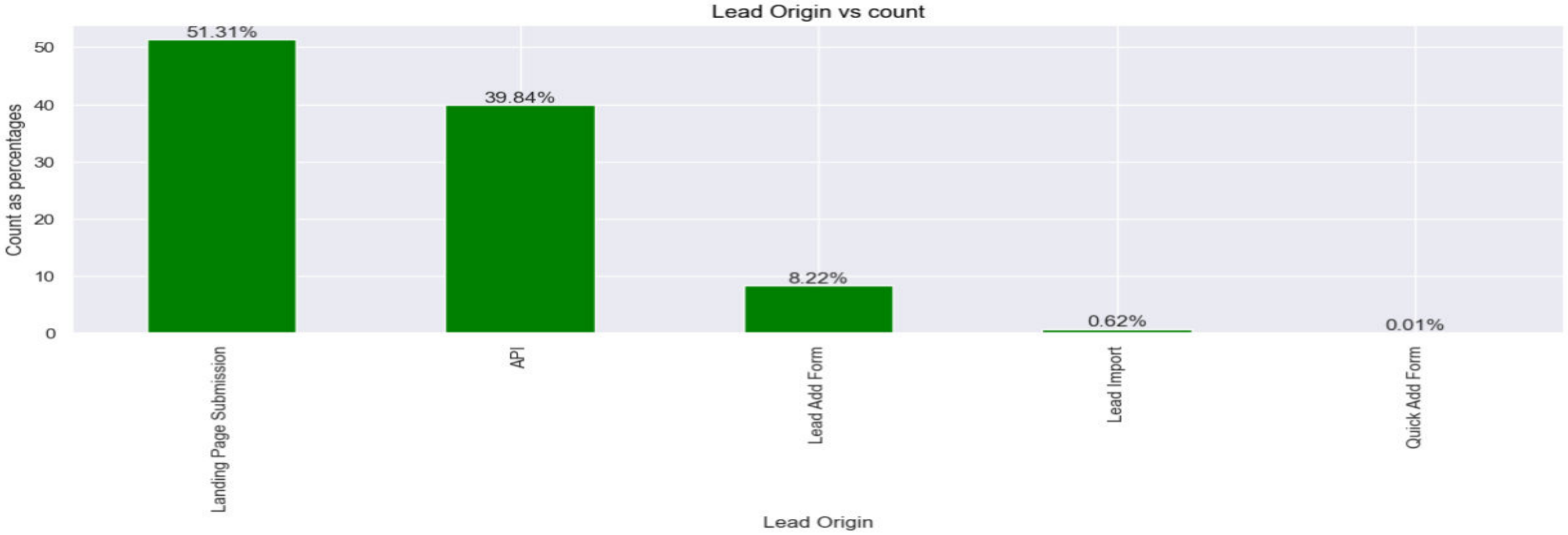
1. Reading and understanding the data
2. EDA or Data Visualization
3. Data Preparation
4. Splitting the Data into Training and Testing Sets, Rescaling
5. Building a logistic model
6. Prediction on Training data set using sensitivity and specificity approach
7. Model evaluation of the test data set using sensitivity and specificity approach
8. Prediction on Training data set using precision and recall approach
9. Model Evaluation of the test data set using precision and recall approach

1. Reading and understanding the data

a) Various libraries in python are invoked and the dataset available in csv format is analysed.
b) Analysed missing values and columns having null values more than 40% are dropped
c) Dropped unnecessary columns
d) Numerical columns having less percentage of missing values are replaced with their respective medians
e) Categorical columns having less percentage of missing values are replaced with their respective modes
f) Outliers are detected and handled

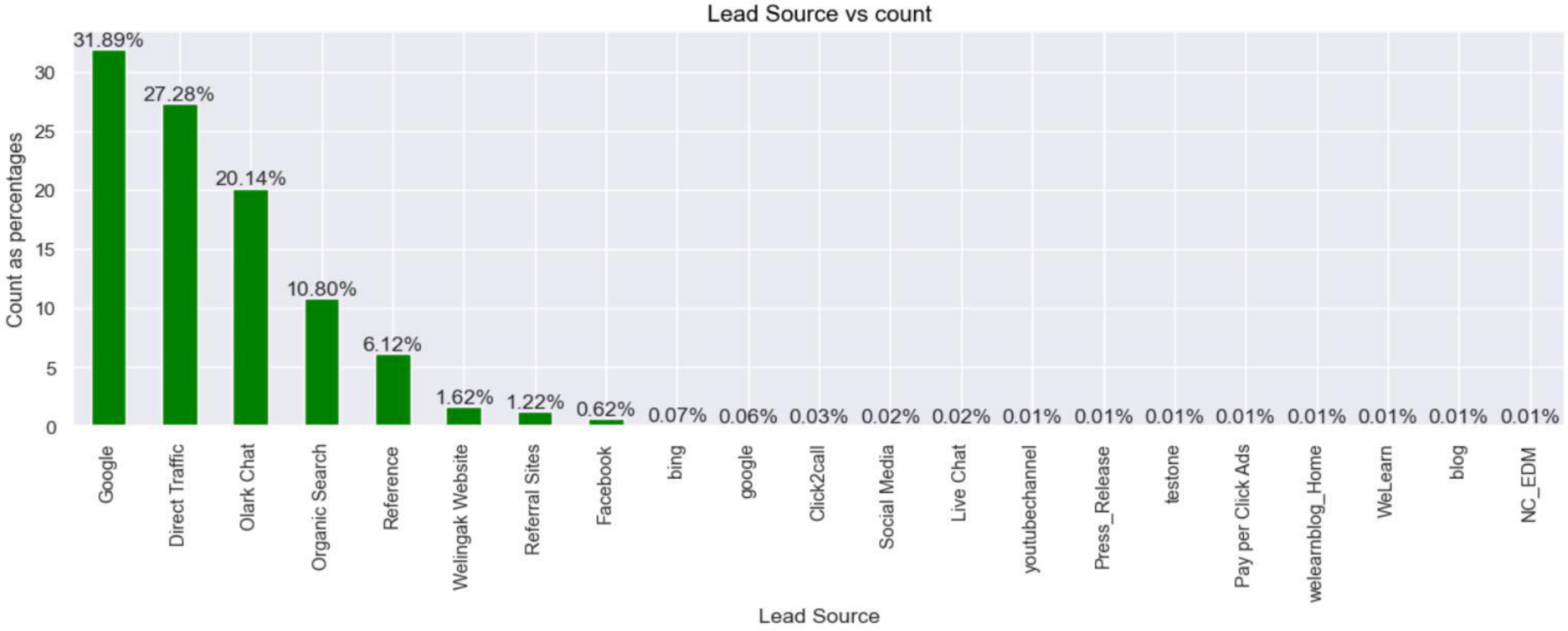
2. EDA or Data Visualisation – Overall inference

Univariate analysis using bar plot for categorical columns



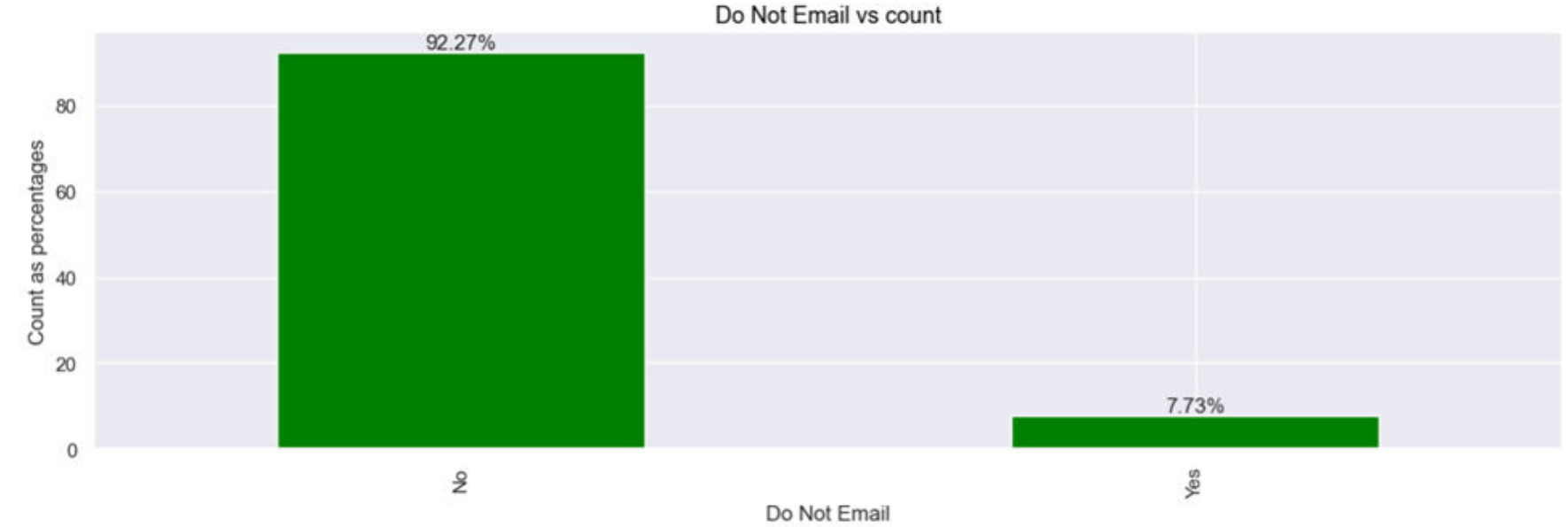
Lead origin :

Landing Page Submission has highest count followed by API.



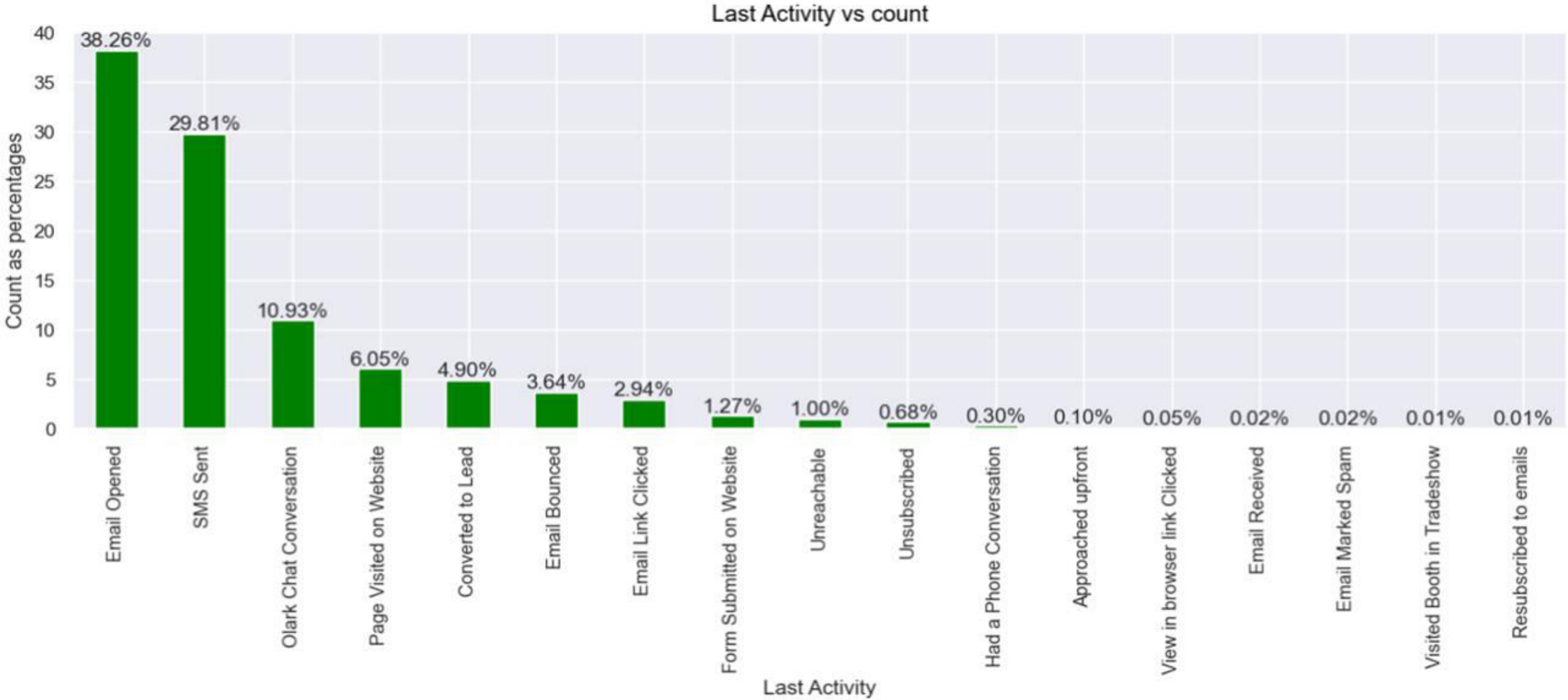
Lead source :

Most of the people came via Google followed by Direct Traffic source.

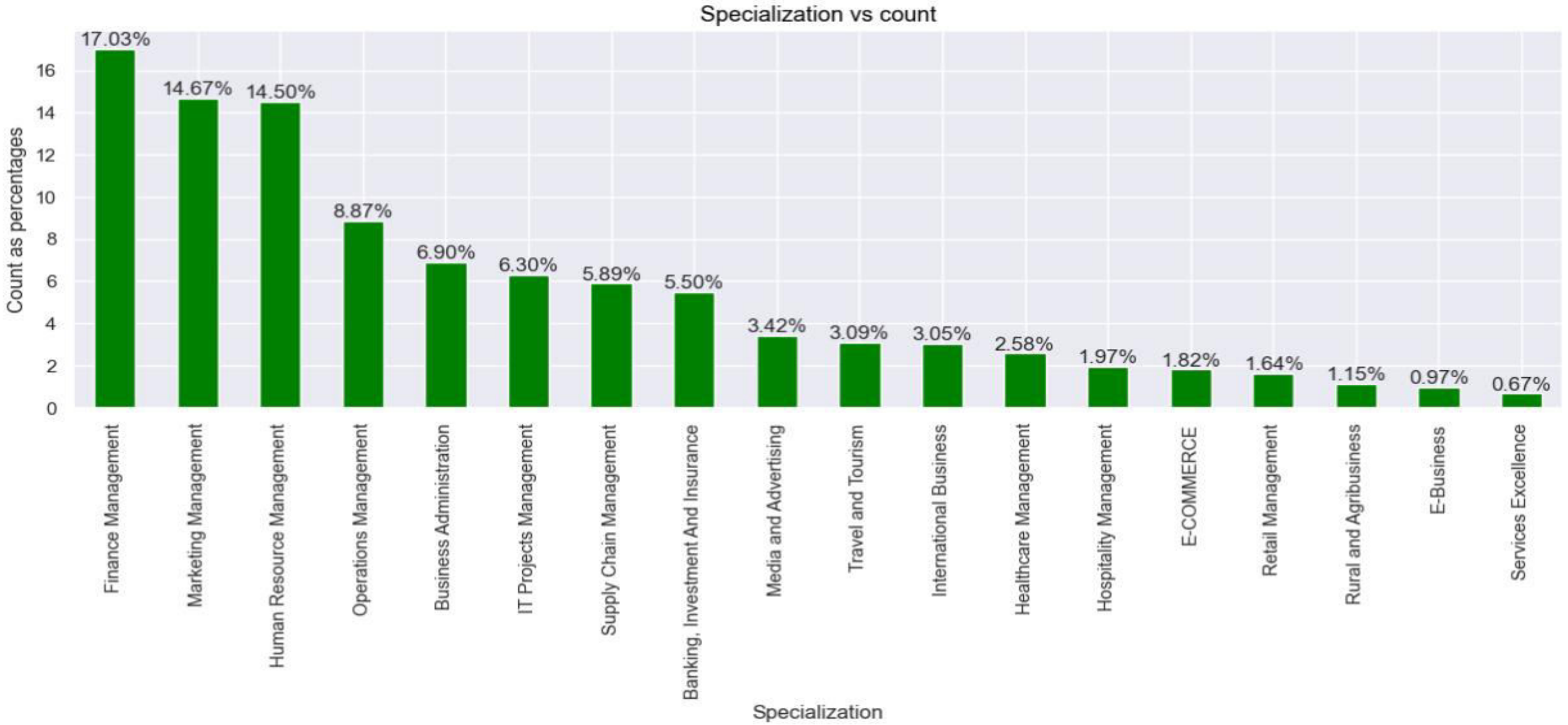


Don't Email & Don't Call:

Most of the people have chosen 'Don't Email' - No option 'Don't Call' - No option from this dataset.

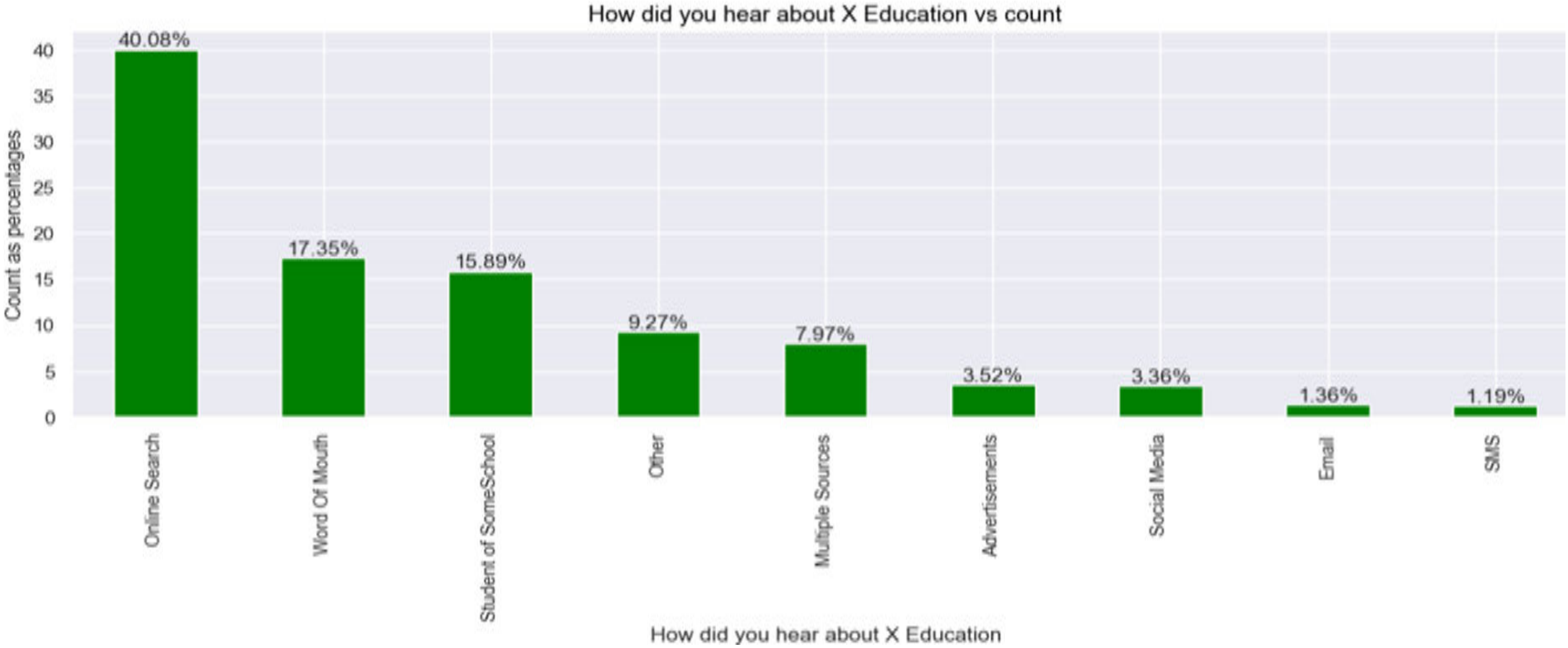


Last activity:
Email opened has highest count followed by SMS sent.

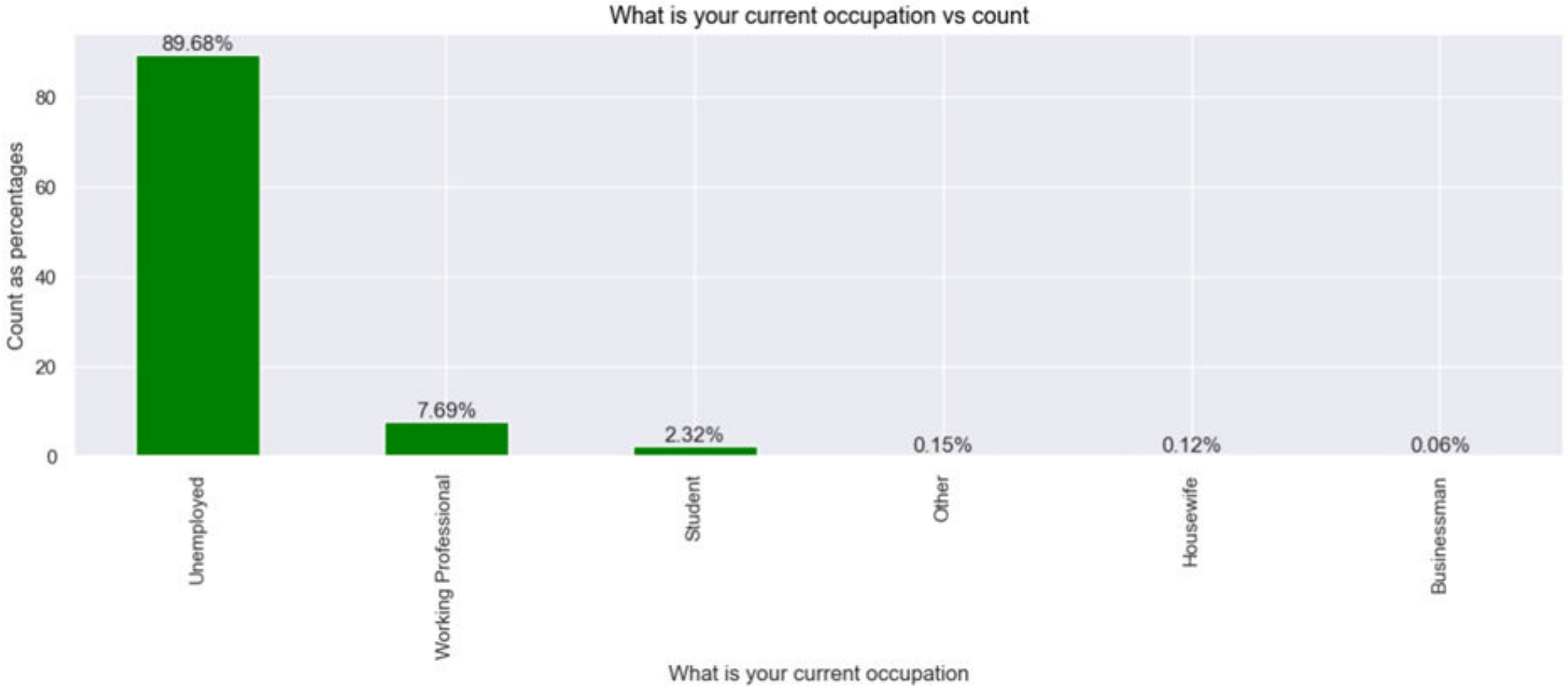


Specialization:

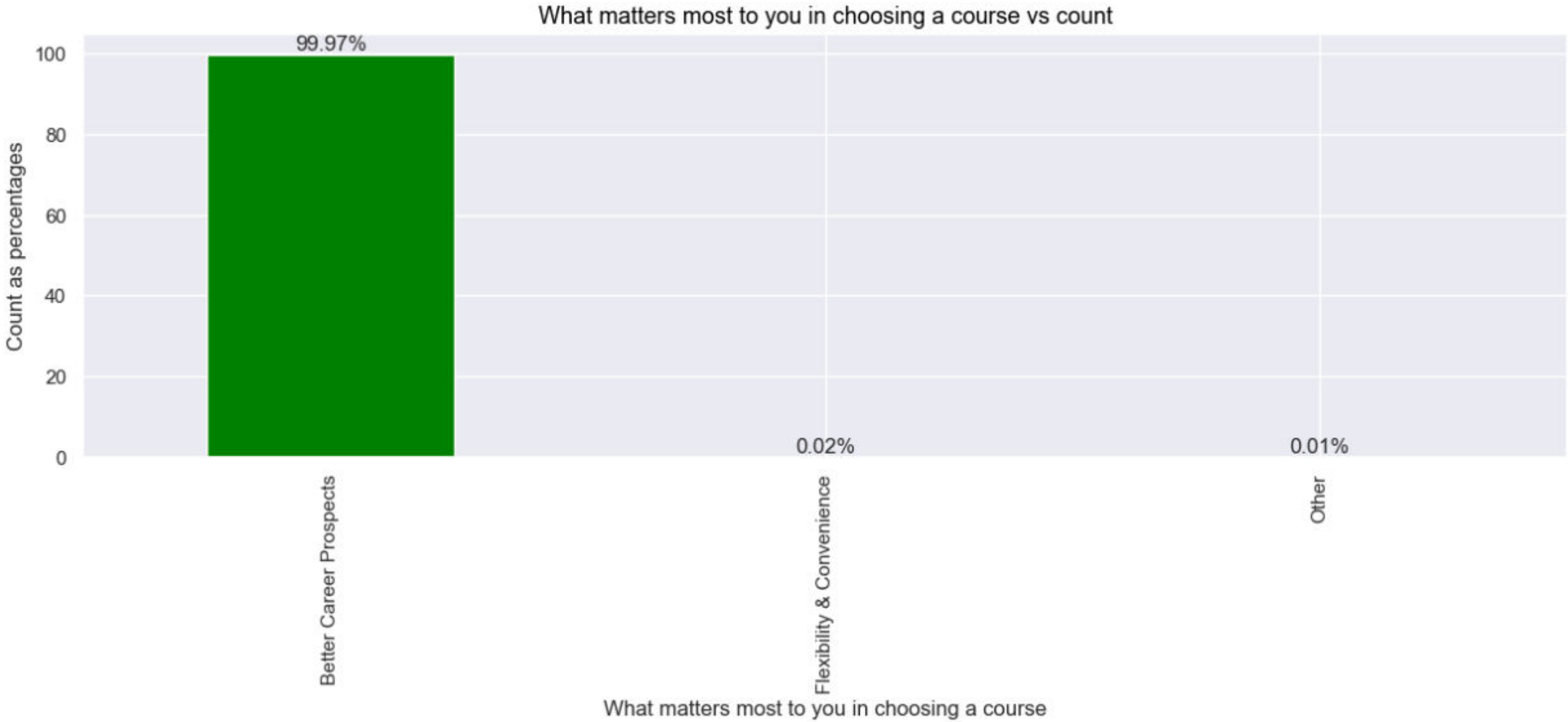
People having management profession in any genre are more likely to be a lead. Finance management has highest count followed by Marketing management, Human resource management.



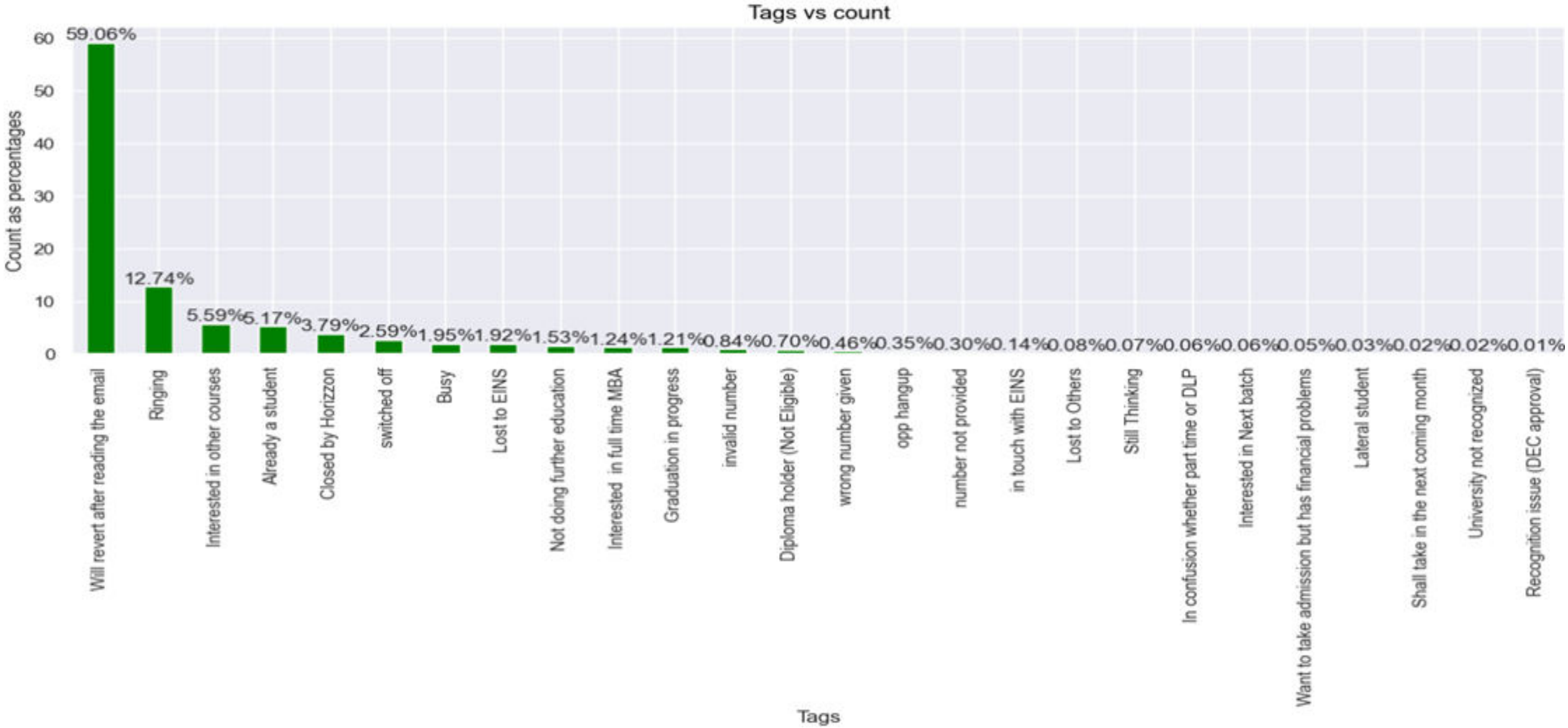
How did you hear about X Education:
Online search has highest count followed by word of mouth, student of some school.



What is your current occupation:
Unemployed has highest count followed by working professional.



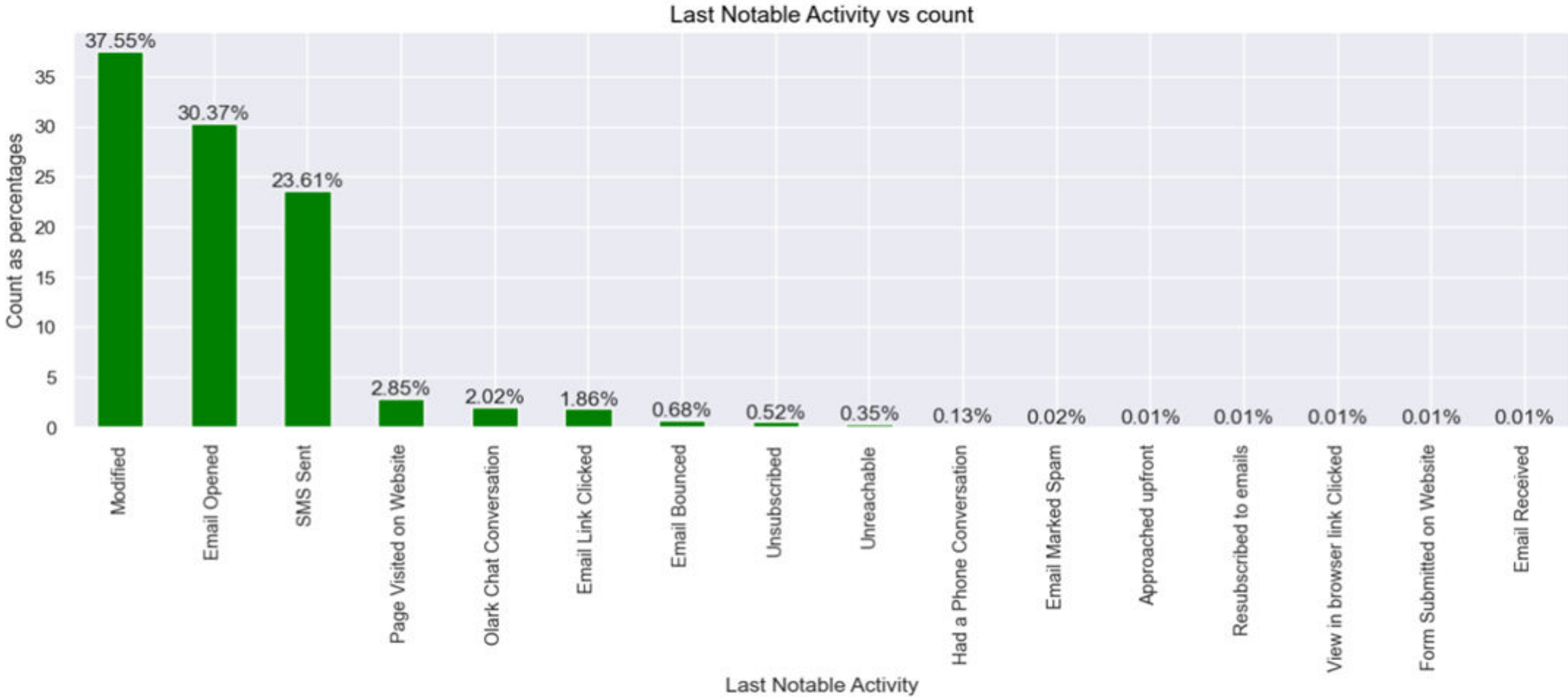
What matters most to you in choosing a course:
Better career prospects has the highest count.



Tags

Will revert after reading the email has highest count followed by Ringing.

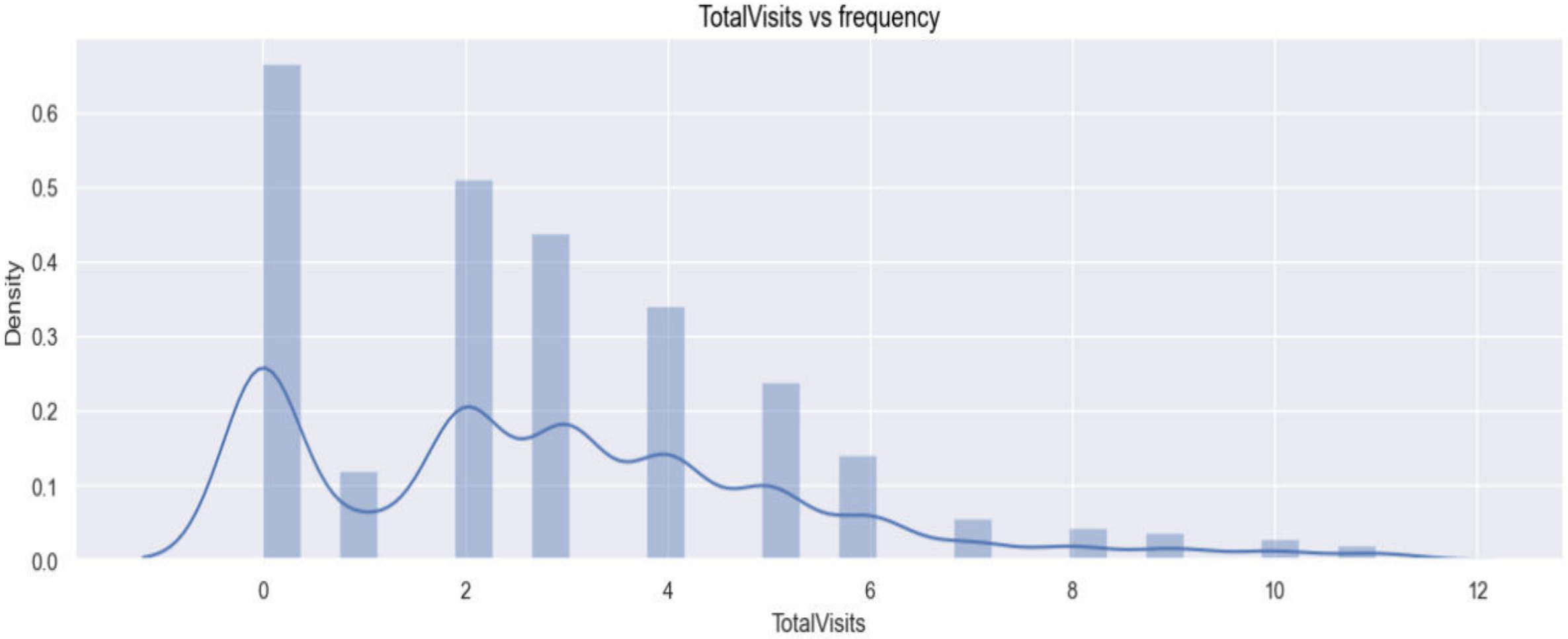
People asking for Better Career Prospects show highly positive response in conversion.



Last notable activity

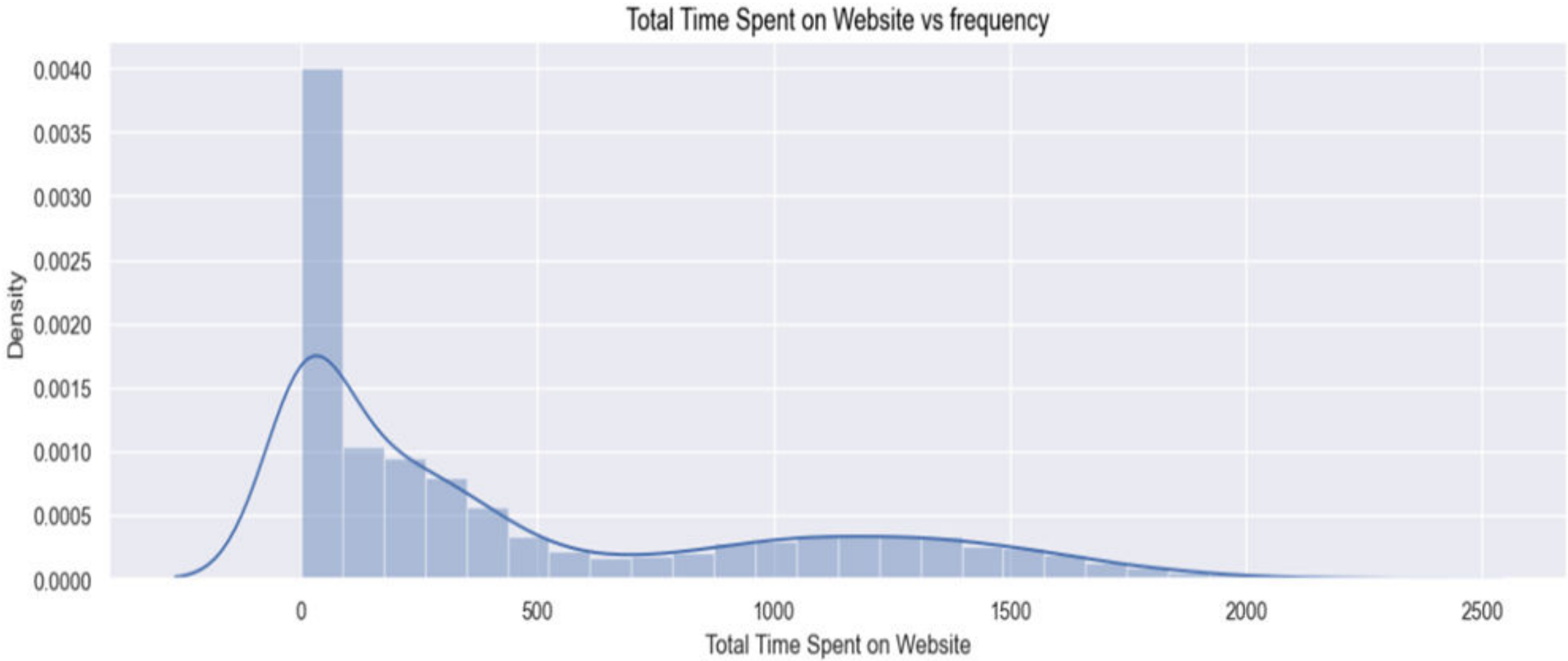
Modified has highest count followed by email opened and sms sent.

Univariate analysis using Histogram for Continuous columns



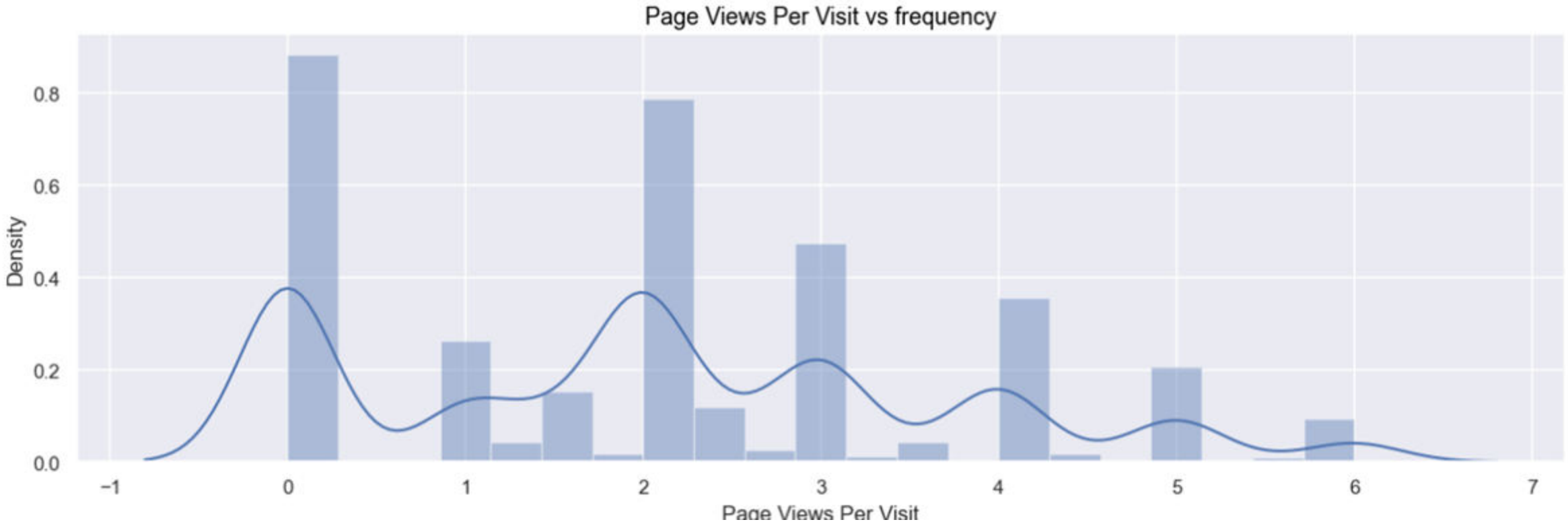
Total visits

Total visits is the highest in frequency 0 to 1 bucket followed by 1 to 2 bucket.



Total Time Spent on Website

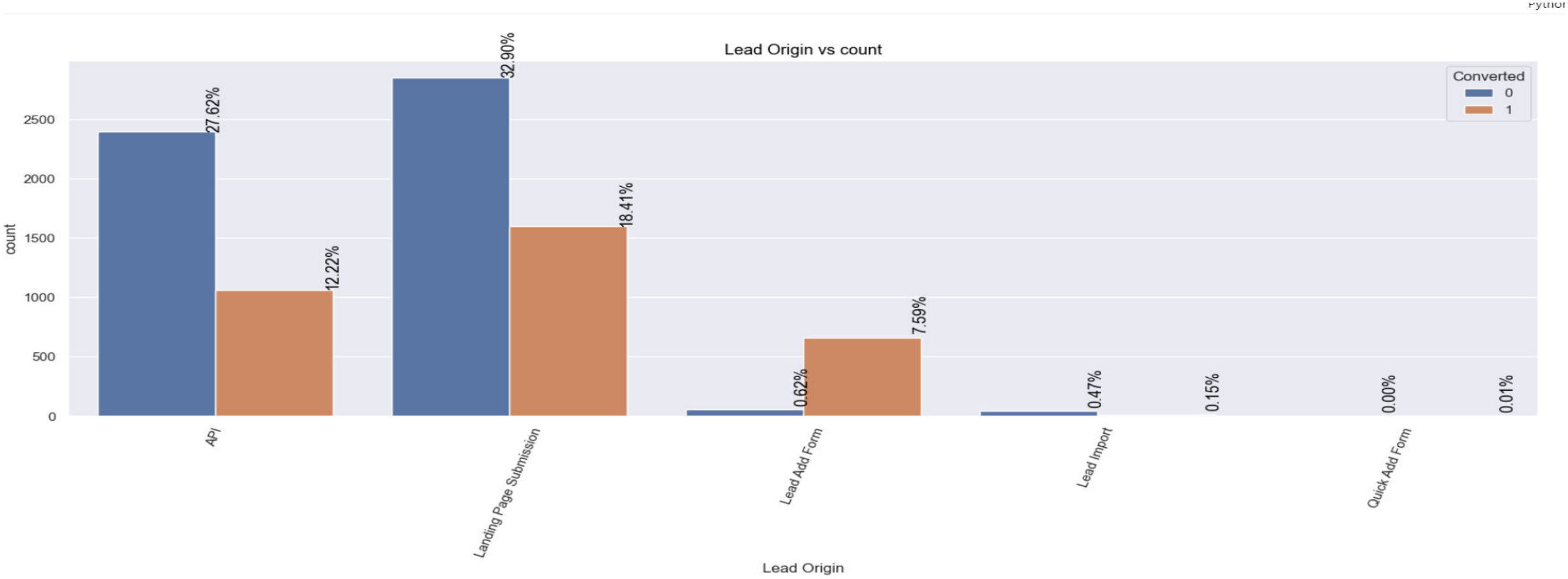
Total Time Spent on Website is the highest in frequency 0 to 227 bucket followed by 227 to 454 bucket.



Page Views Per Visit

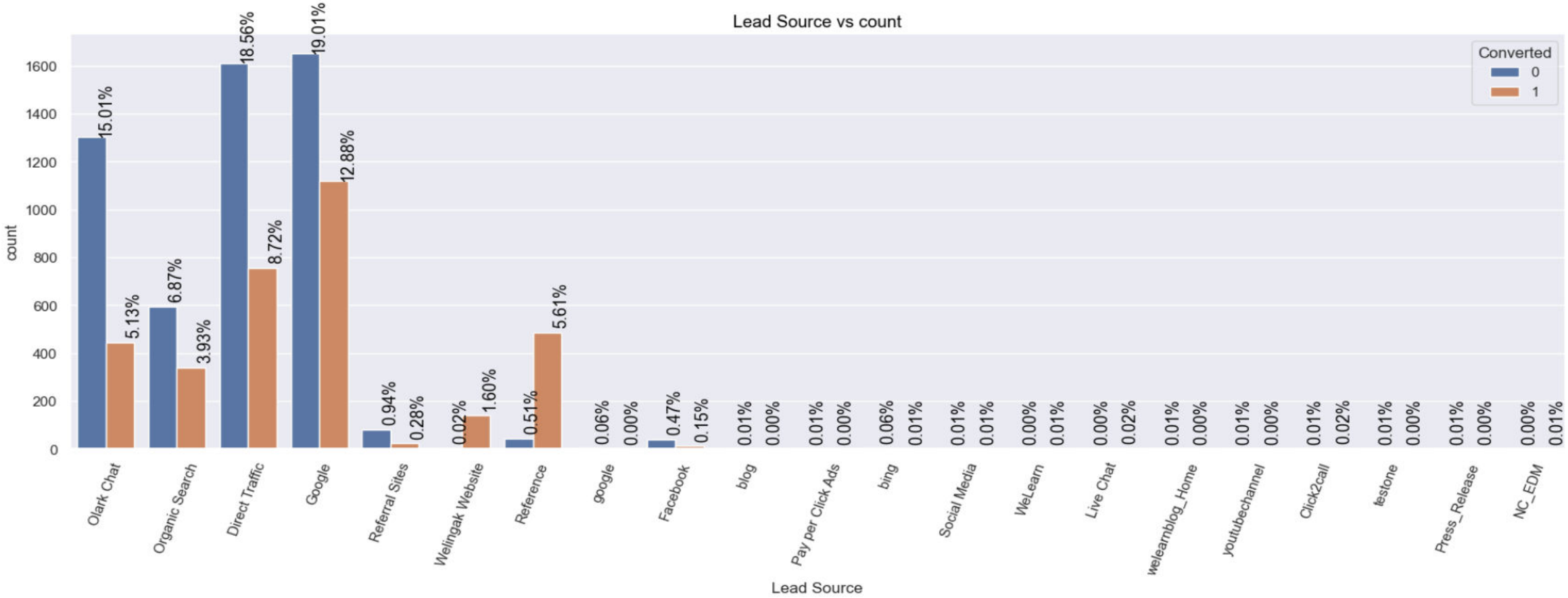
Page Views Per Visit is the highest in frequency 0 to 0.6 bucket followed by 1.8 to 2.4 bucket.

Bivariate analysis w.r.t Target variable (categorical column vs categorical column)



Lead Origin

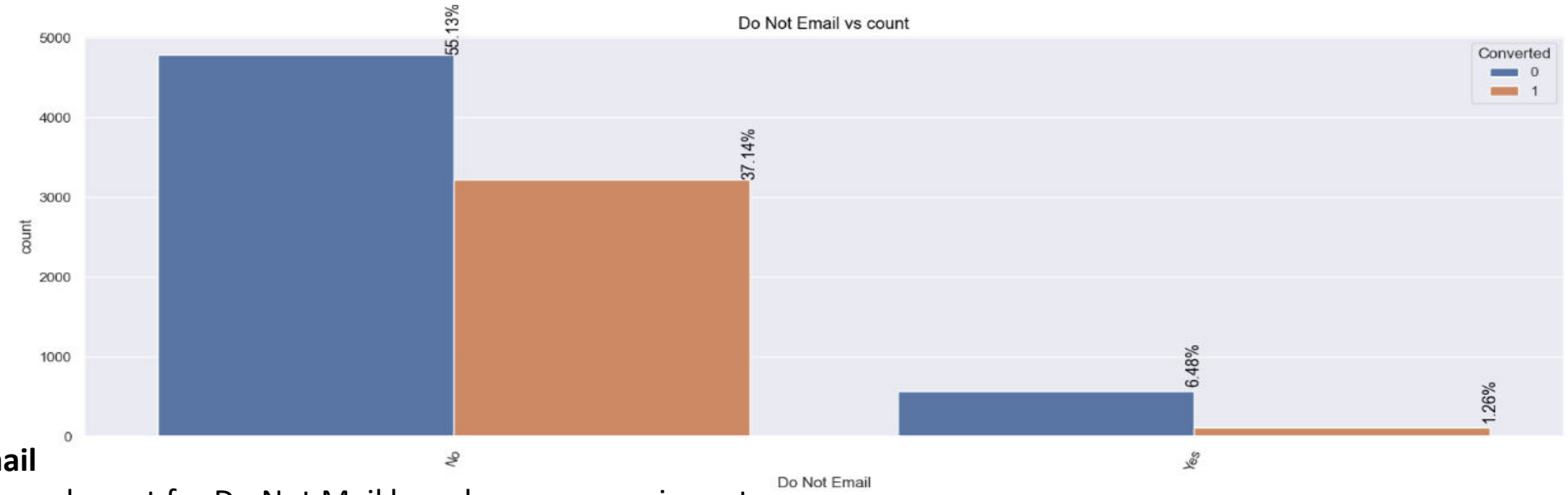
1. Customers who were identified as Lead from Landing Page submission, constitute the majority of the leads.
2. Customers originating from Lead Add Form have high probability of conversion. These Customers are very few in number.
3. Lead Import has the least conversion rate. Customers from Lead Import are very few in number.



Lead Source

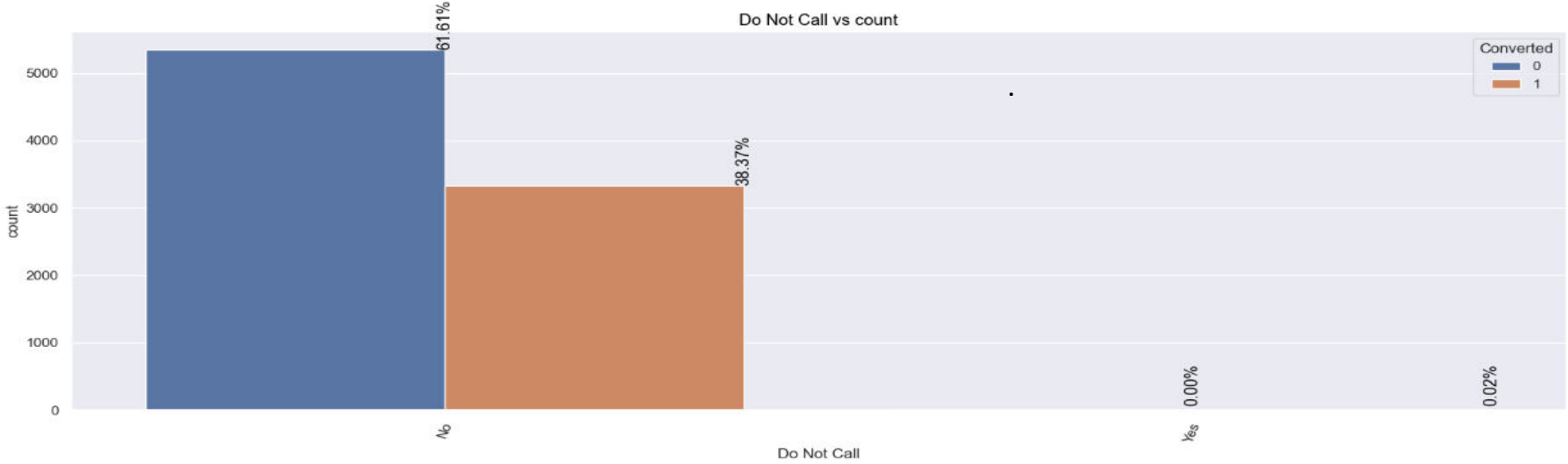
Majority source of the lead is Google & Direct Traffic.

Leads with source Reference has maximum probability of conversion



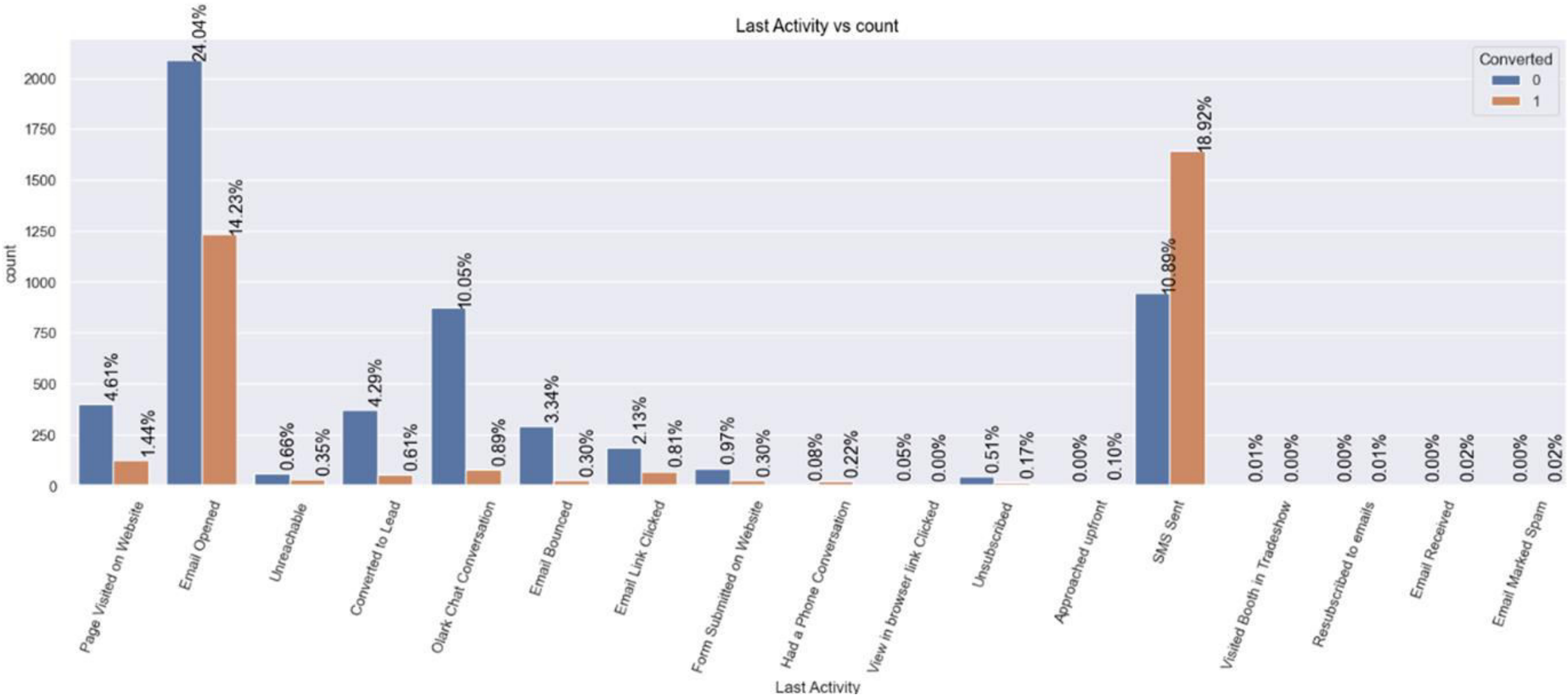
Do Not Email

1) Customers who opt for Do Not Mail have lower conversion rate.
2) Customers who do not opt for Do Not Mail have higher conversion rate. These constitute the majority of the leads.



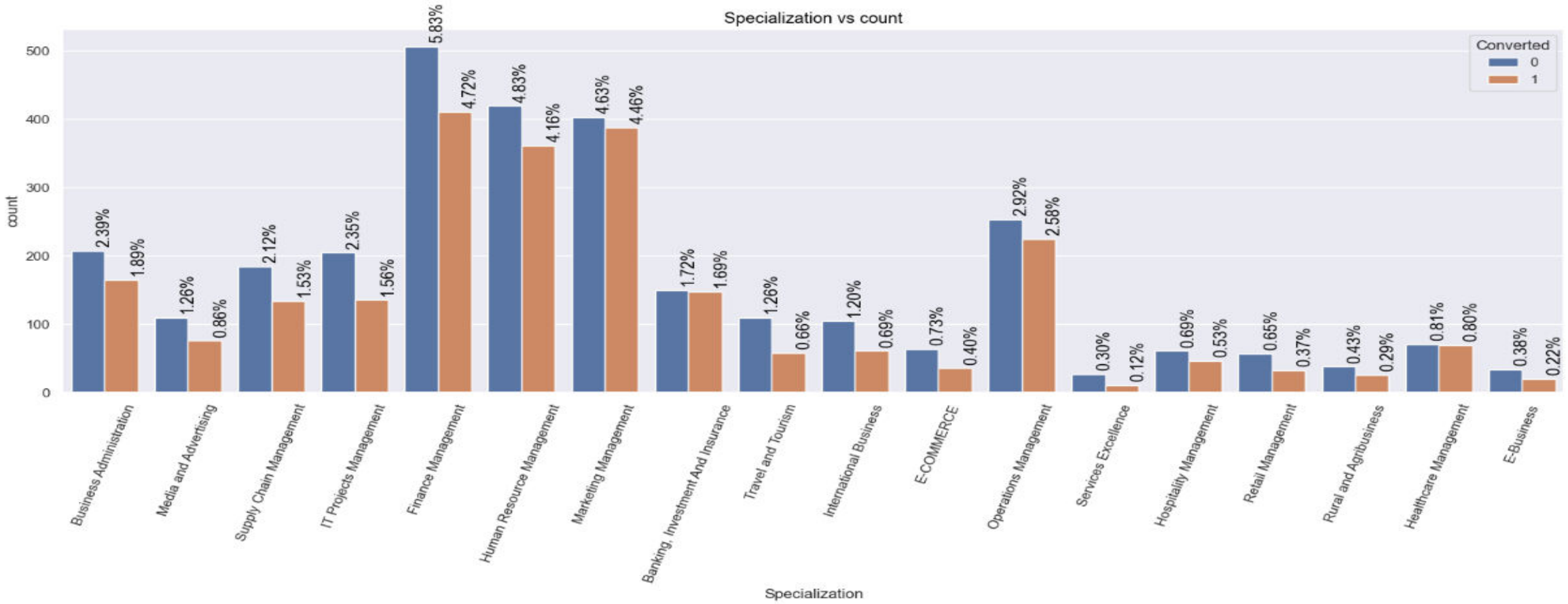
Do Not Call

Customers who do not opt for Do Not call have Higher conversion rate. These constitute the majority of the leads.



Last Activity

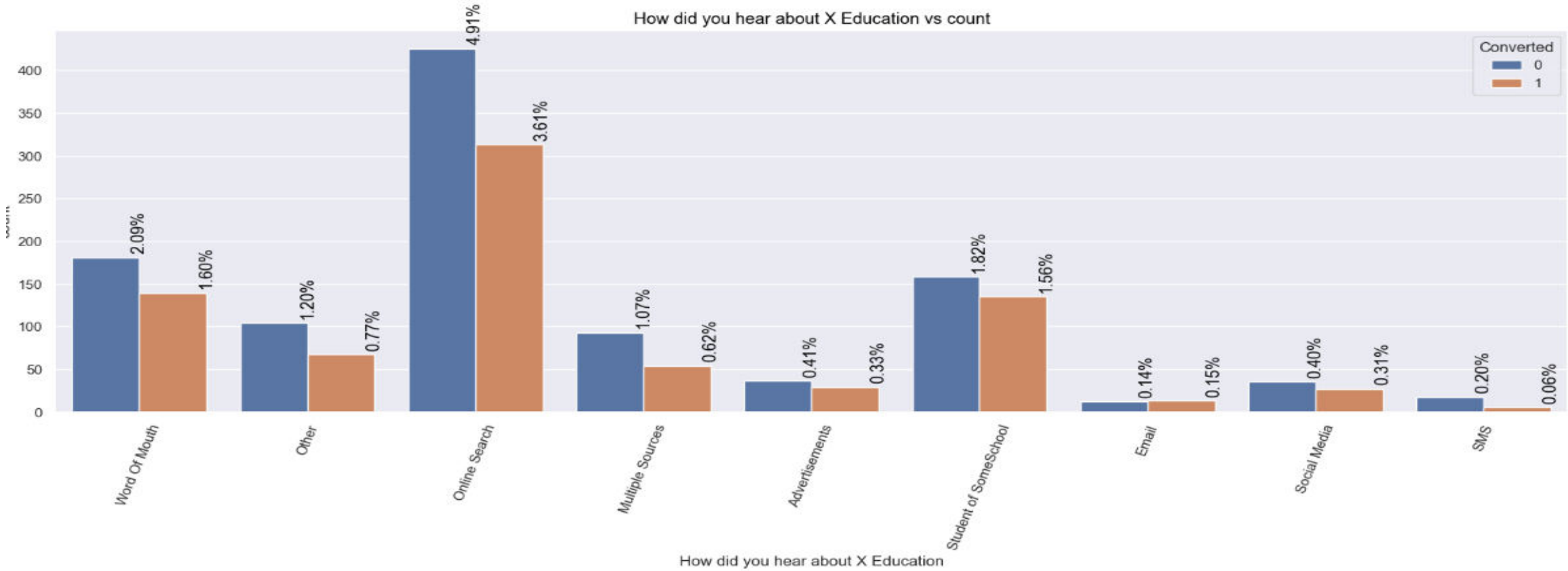
1) Customers who last activity was SMS Sent have higher conversion rate.
2) Customers who last activity was Email Opened constitute majority of the customers



Specialization

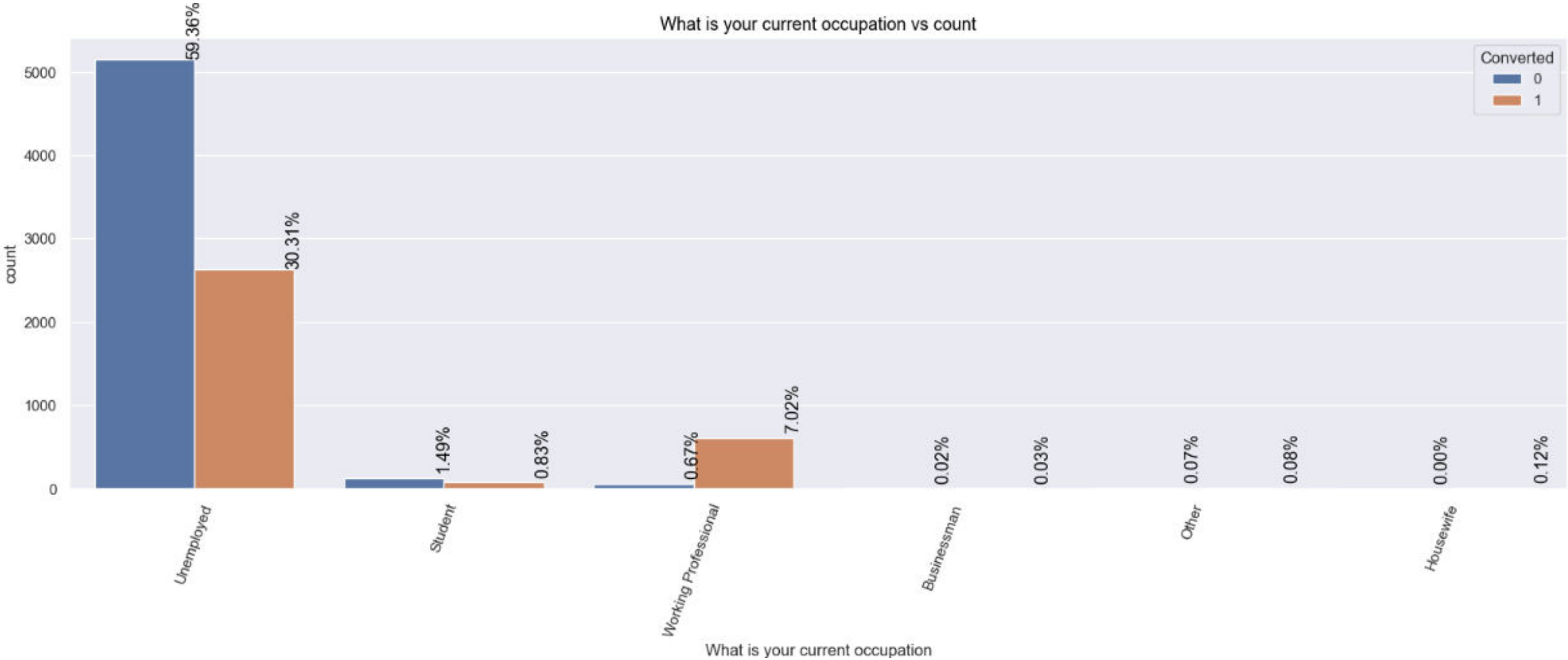
1.) Management professions like Finance, HR, Marketing and Operations have very good count of conversion compared to other specializations.

2.) Leads with specialization as Rural & Agribusiness, Services excellence, E-business have least probability of conversion.



How did you hear about X education

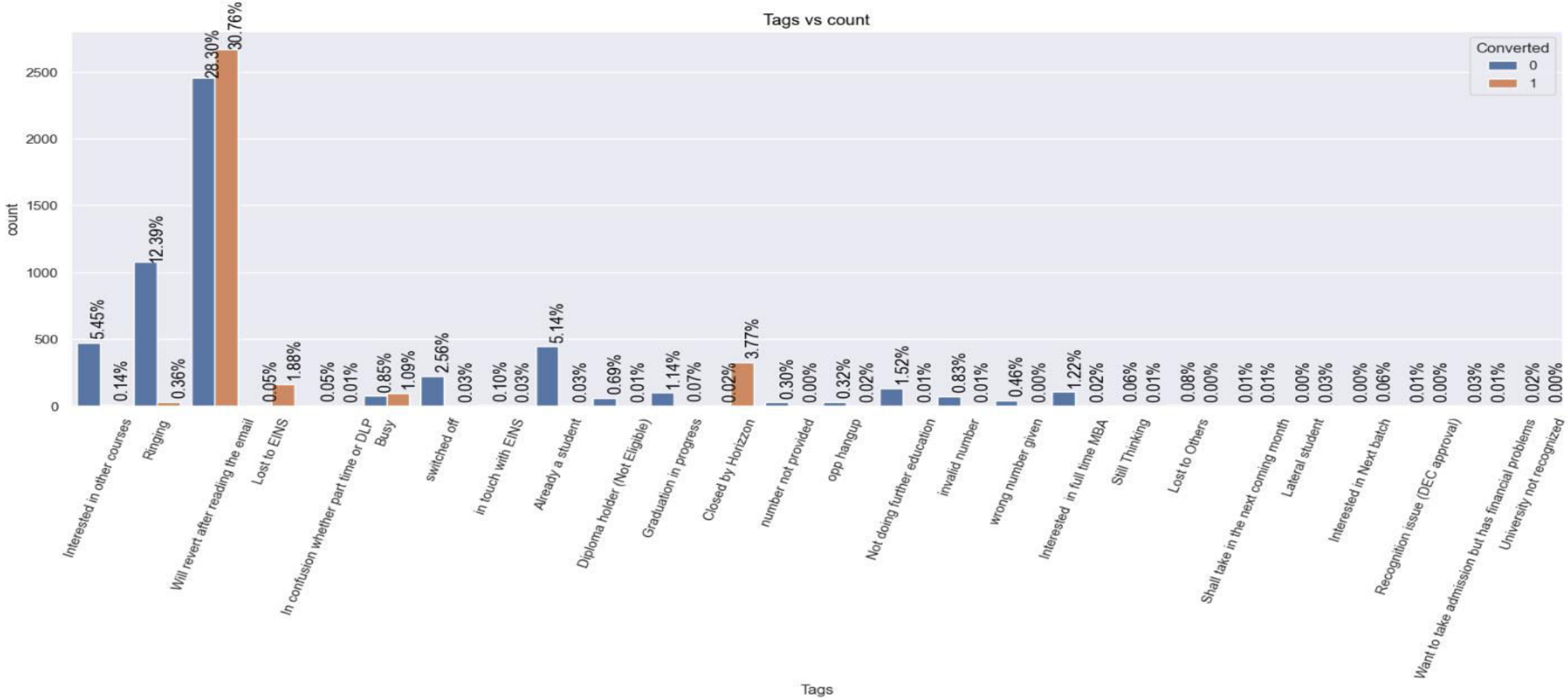
Maximum Leads are from online search. Minimum leads are from SMS and Email



What is your current occupation

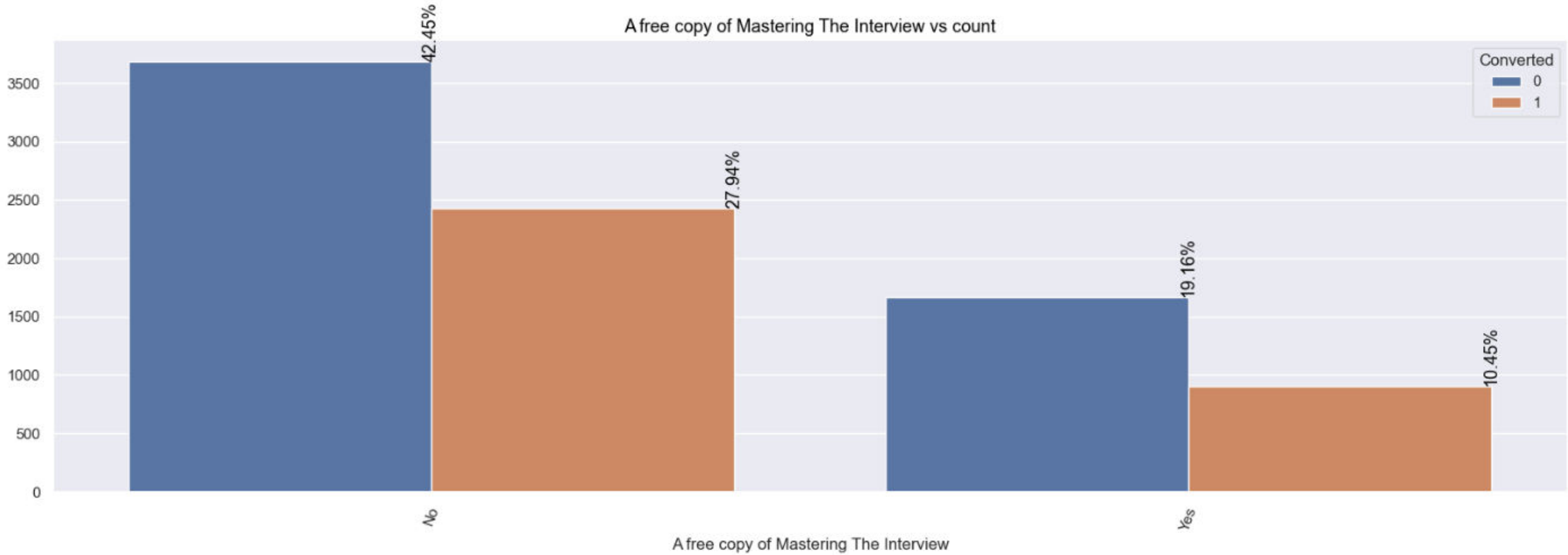
1.) Maximum Leads have occupation as Unemployed.

2.) Very few leads are Students



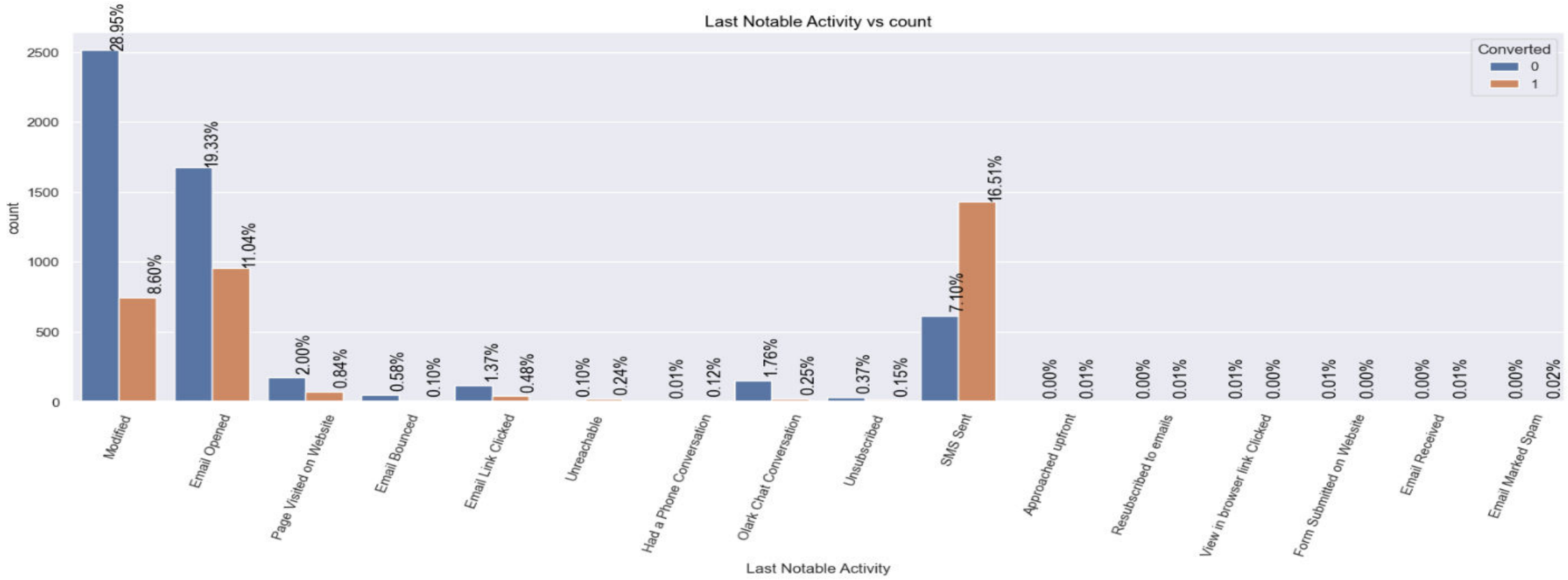
- **Tags**

  More focus shall be given on the leads as will revert after reading the mail as these are potential leads and have higher rate of conversion.



A Free Copy Of Mastering the Interview

Customer didn't demand for a free copy of Mastering the Interview have good count of conversion

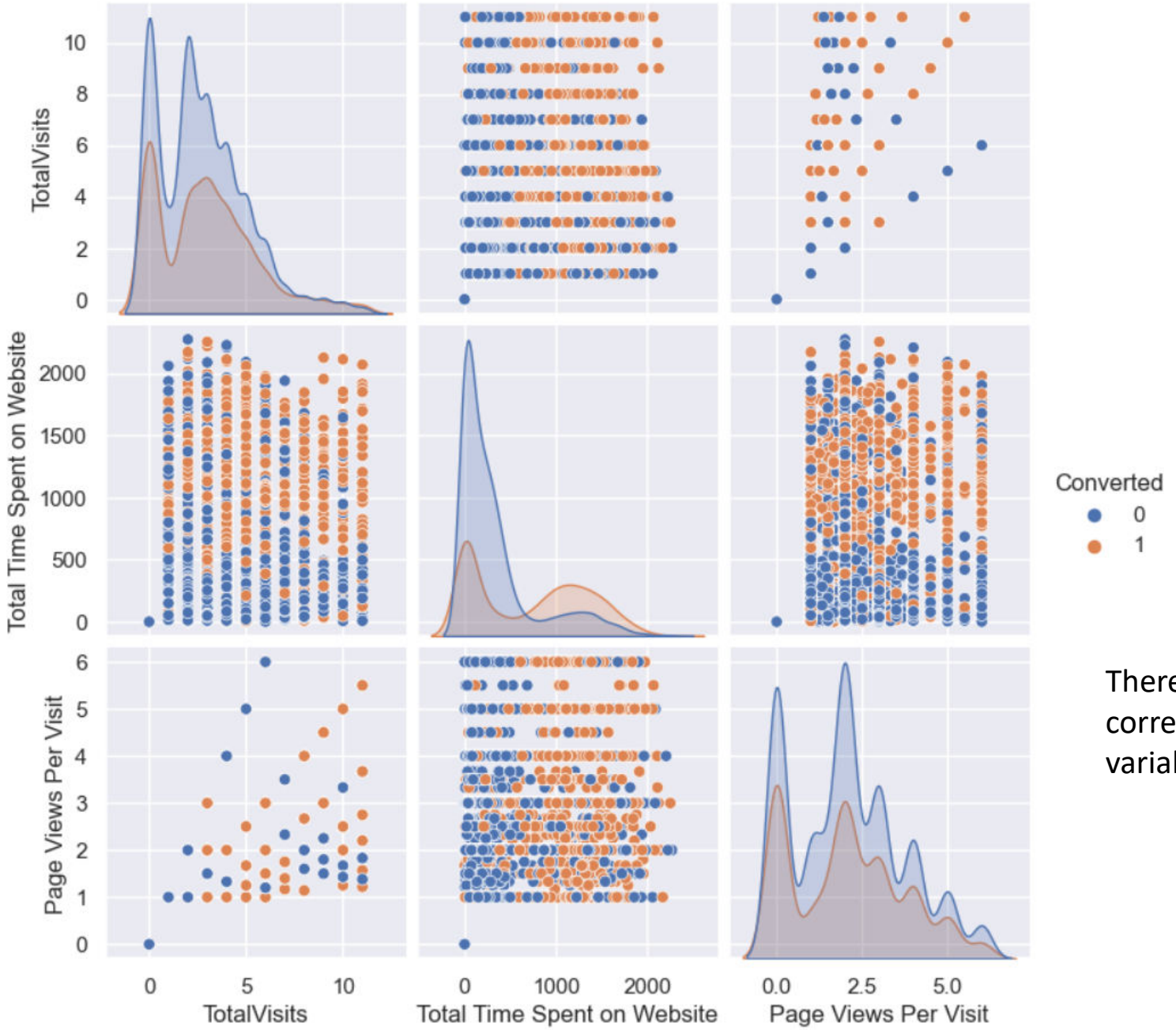


Last Notable Activity

1.) customers whose last notable activity was Modified are more in number.

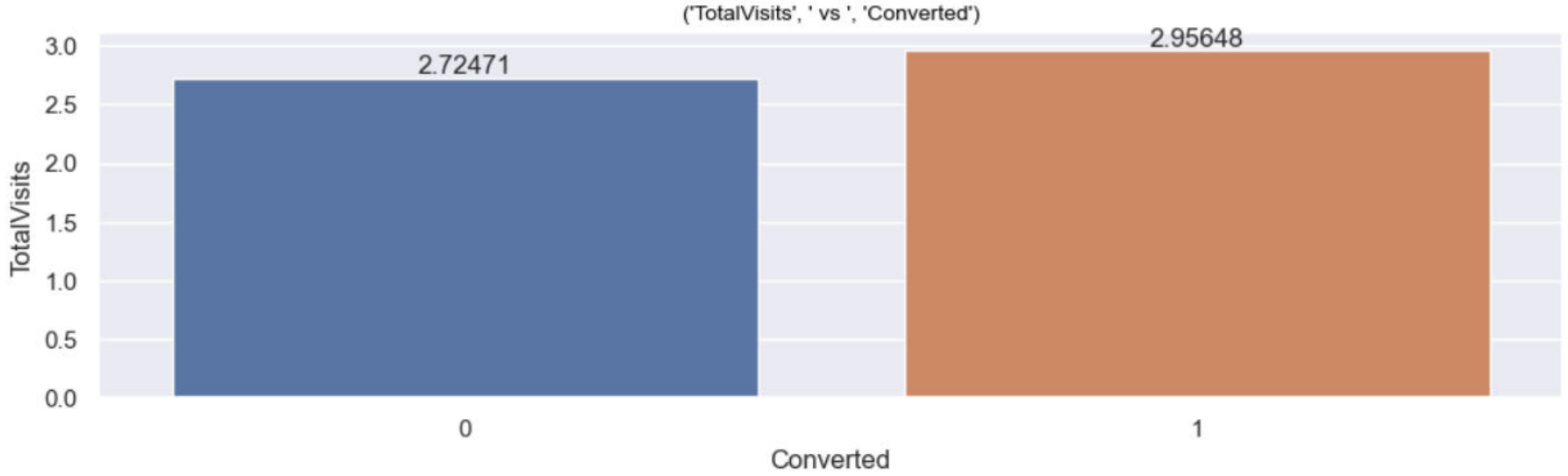
2.) Customers whose last notable activity was SMS Sent have higher probability of conversion

Bivariate analysis w.r.t Target variable (numerical column vs numerical column)

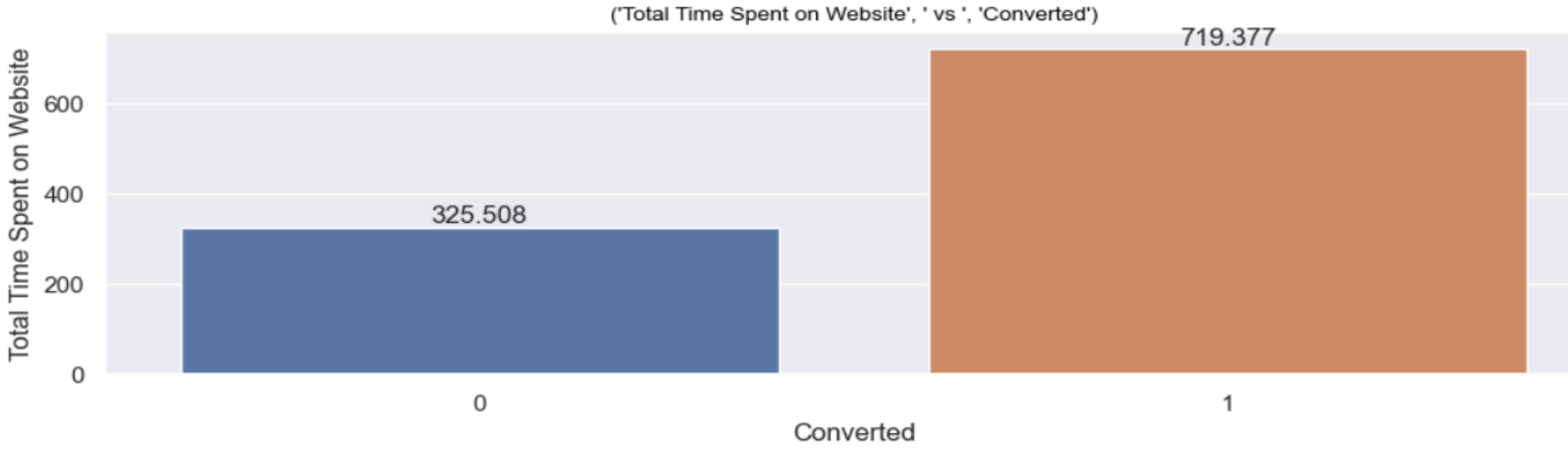


There is a strong positive correlation between the TotalVisits variable and Page Views Per Visit

Bivariate analysis w.r.t Target variable (numerical column vs categorical column)

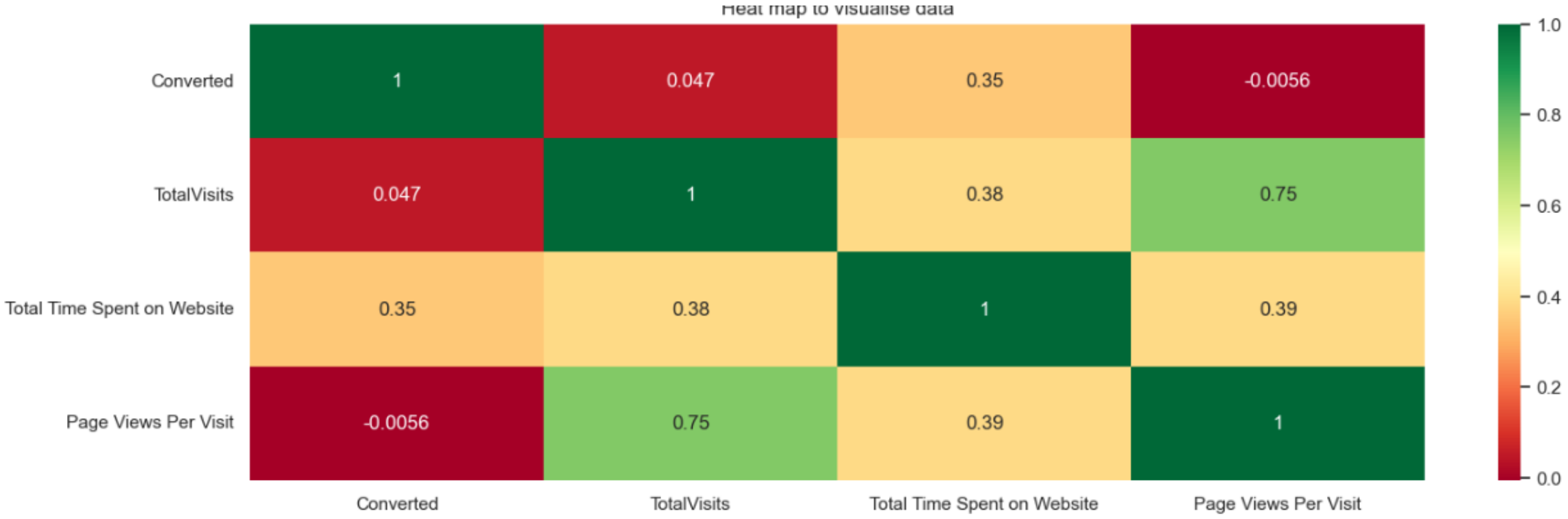


Converted leads have the highest Total visits.



Converted leads have the highest Total time spent on website.

Multivariate analysis



1. There is strong positive correlation 0.75 between the TotalVisits and Page Views Per Visit
2. There is positive correlation 0.38 between the TotalVisits and Total Time Spent on Website
3. There is positive correlation 0.39 between the Total Time Spent on Website and Page Views Per Visit

3. Data Preparation

A. Creating the dummy variables- For categorical variables with multiple levels, created dummy features (one-hot encoded)
B. Converted some binary variables (Yes/No) to 0/1

4. Splitting the Data into Training and Testing Sets, Rescaling

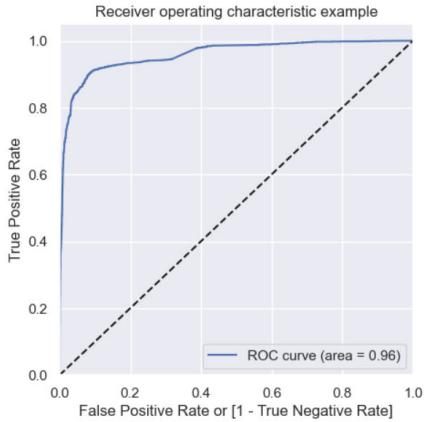
- Feature variables are assigned to X data frame
- Response variable ‘Converted’ is assigned to y data frame
- Splitted the data into train and test using train_test_split() function with 70% as training and 30% as test data
- Feature Scaling is done using MinMaxScaler() function

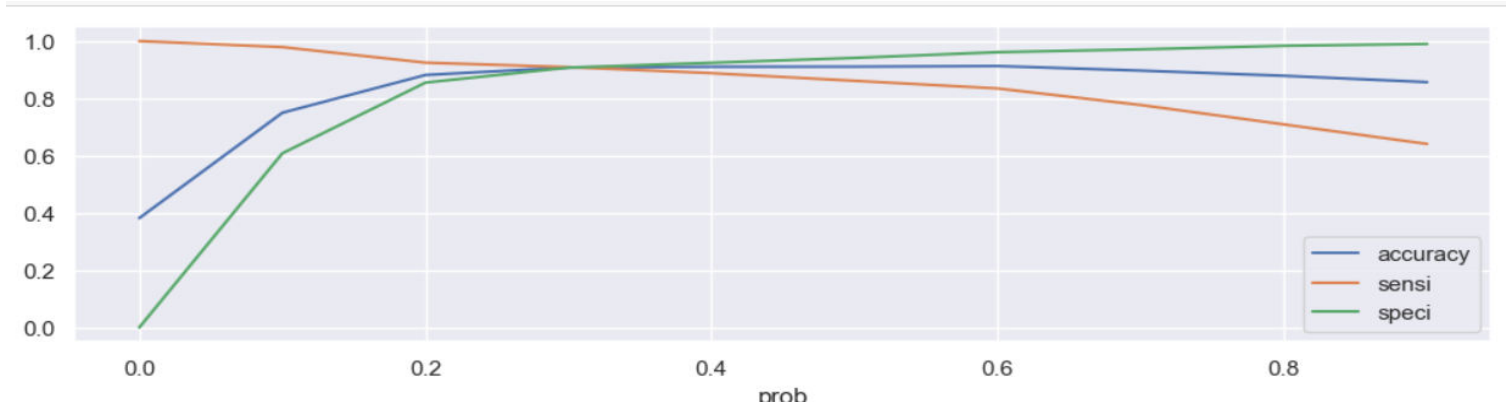
5. Logistic Model Building

a) Imported sklearn library and the model is built using Recursive Feature Elimination, a feature selection technique to determine the most important features (variables) for a predictive model.
b) 20 number of features were specified.
c) Model was trained with the training data using GLM and based on summary values – P value and VIF(variance inflation factor). We are keeping features having P values less than 0.05 and VIF values less than 5. some of features were dropped one after the other to arrive at the final model because they have P values higher than 0.05.
d) Predicted the values on train data set and a data frame was formed

6 : Sensitivity & specificity approach on training data set

- Plotted the ROC curve and based on Sensitivity & Specificity tradeoff found the optimum **cutoff point to be 0.3**





- Based on optimum cutoff value we predict on training data set.
- Confusion matrix was found to be [3408, 348]

  [211, 2108]

- **<u>Observation for Train Data Set:</u>**

A. **Accuracy 90.79%**
B. **Sensitivity 90.90%**
C. **Specificity 90.73%**

7: Making Predictions and model evaluation using sensitivity and specificity on test data set

1. Now test data is scaled and transformed
2. Prediction was done on test data set
3. Confusion matrix was found to be [1442, 149],

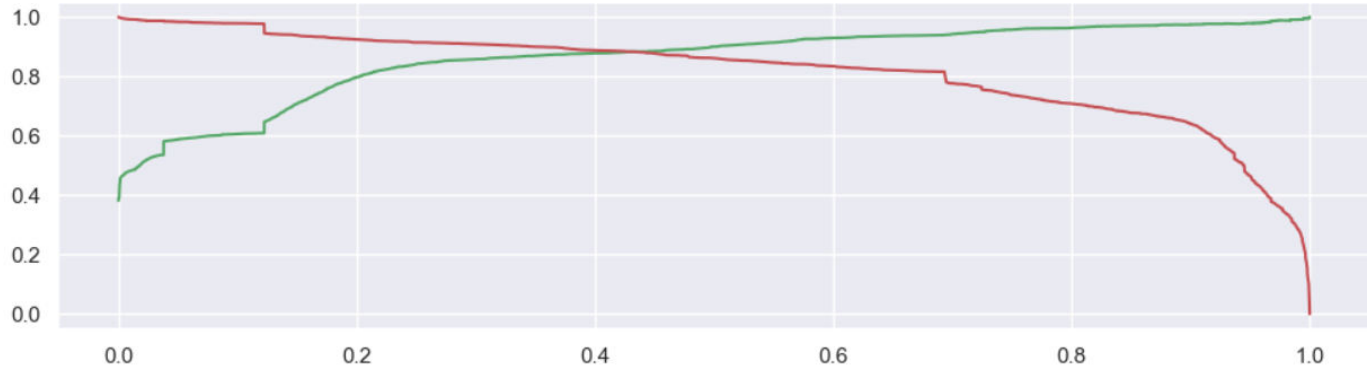
   [99, 914]

Observation for Test Data Set:

A. **Accuracy 90.47%**
B. **Sensitivity 90.22%**
C. **Specificity 90.63%**

8: Precision & recall approach on training data set

- precision_recall_curve was employed to get this curve to find the optimal cutoff.



- **0.42** is used as optimal cutoff
- **<u>Observation for Training Data Set:</u>**

A. **Accuracy 91.04%**
B. **Precision 88.07%**
C. **Recall 88.52%**

9: Making Predictions on the Test Set using precision and recall

- Making predictions on the test data set using **0.42** as the optimal cutoff
- **Observation for Test Data Set:**

A. **Accuracy 90.62%**
B. **Precision 88.25%**
C. **Recall 87.56%**

Evaluation Metrics of the model

Sensitivity,Specificity approach:

Train Data Set:

- Accuracy 90.79%
- Sensitivity 90.90%
- Specificity 90.73%

Test Data Set:

- Accuracy 90.47%
- Sensitivity 90.22%
- Specificity 90.63%

Precision,Recall approach:

Training Data Set:

- Accuracy 91.04%
- Precision 88.07%
- Recall 88.52%

Test Data Set:

- Accuracy 90.62%
- Precision 88.25%
- Recall 87.56%

Conclusion : Univariate analysis for categorical columns

Lead origin :

- Landing Page Submission has highest count followed by API.

Lead source :

- Most of the people came via Google followed by Direct Traffic source.

Don't Email & Don't Call:

- Most of the people have chosen 'Don't Email' - No option 'Don't Call' - No option from this dataset.

Last activity:

- Email opened has highest count followed by SMS sent.

Specialization:

- People having management profession in any genre are more likely to be a lead. Finance management has highest count followed by Marketing management, Human resource management.

Conclusion : Univariate analysis for categorical columns

How did you hear about X Education:

- Online search has highest count followed by word of mouth, student of some school.

What is your current occupation:

- Unemployed has highest count followed by working professional.

What matters most to you in choosing a course:

- Better career prospects has the highest count.

Tags

- Will revert after reading the email has highest count followed by Ringing.

Lead profile

- Potential lead has highest count followed by other leads and student of some school.

Last notable activity

- Modified has highest count followed by email opened and sms sent.

Conclusion : Univariate analysis for numerical columns

Total visits

- Total visits is the highest in frequency 0 to 1 bucket followed by 1 to 2 bucket.

Total Time Spent on Website

- Total Time Spent on Website is the highest in frequency 0 to 227 bucket followed by 227 to 454 bucket.

Page Views Per Visit

- Page Views Per Visit is the highest in frequency 0 to 0.6 bucket followed by 1.8 to 2.4 bucket.

Conclusion - Bivariate Analysis for categorical columns

Lead Origin

a) Customers who were identified as Lead from Landing Page submission, constitute the majority of the leads.

b) Customers originating from Lead Add Form have high probability of conversion. These Customers are very few in number.

c) Lead Import has the least conversion rate. Customers from Lead Import are very few in number.

Lead Source

a) Majority source of the lead is Google & Direct Traffic.

b) Leads with source Reference has maximum probability of conversion.

Do Not Email

a) Customers who opt for Do Not Mail have lower conversion rate.

b) Customers who do not opt for Do Not Mail have higher conversion rate. These constitute the majority of the leads.

Conclusion - Bivariate Analysis for categorical columns

Do Not Call

a) Customers who do not opt for Do Not call have Higher conversion rate. These constitute the majority of the leads.

Last Activity

a) Customers who last activity was SMS Sent have higher conversion rate.

b) Customers who last activity was Email Opened constitute majority of the customers.

Specialization

a) Management professions like Finance, HR, Marketing and Operations have very good count of conversion compared to other specializations.

b) Leads with specialization as Rural & Agribusiness, Services excellence, E-business have least probabilty of conversion.

How did you hear about X education

a) Maximum Leads are from online search.

b) Minimum leads are from SMS and Email.

Conclusion - Bivariate Analysis for categorical columns

What is your current occupation

a) Maximum Leads have occupation as Unemployed.

b) Very few leads are Students

What matters most to you in choosing a career

a) People asking for Better Career Prospects show highly positive response in conversion.

Newspaper , Digital Advertisement, through recommendations, Search Customers who have seen the add of the education company in any form, are very few in number. Nothing meaningful insight can be concluded from the plot that will impove the overall lead conversion Rate.

Tags

a) More focus shall be given on the leads as will revert after reading the mail as these are potential leads and have higher rate of conversion.

Conclusion - Bivariate Analysis for categorical columns

A Free Copy Of Mastering the Interview

a) Customer didn't demand for a free copy of Mastering the Interview have good count of conversion.

Last Notable Activity

a) Customers whose last notable activity was Modified are more in number.

b) Customers whose last notable activity was SMS Sent have higher probability of conversion.

Bivariate Analysis - numerical column

a) there is strong positive correlation 0.75 between the TotalVisits and Page Views Per Visit

b) there is positive correlation 0.38 between the TotalVisits and Total Time Spent on Website

c) there is positive correlation 0.39 between the Total Time Spent on Website and Page Views Per Visit

Conclusion

Bivariate Analysis - numerical column

a) Converted leads have the highest Total time spent on website.

b) Converted leads have the highest Total visits.

Multivariate Analysis

a) There is strong positive correlation 0.75 between the TotalVisits and Page Views Per Visit

b) There is positive correlation 0.38 between the TotalVisits and Total Time Spent on Website

c) There is positive correlation 0.39 between the Total Time Spent on Website and Page Views Per Visit

Finally

I. Sensitivity was calculated in the test set of data which is more than 90% for final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.

II. Good value of sensitivity and recall of our model will help to select the most promising leads.

III. Top 3 features which contribute more towards the probability of a lead getting converted are a) Total Time Spent on Website b) Lead Origin_Lead Add Form c) Lead Source_Welingak Website