LEAD SCORE CASE STUDY

By: Keshav, Suresh and Surabhi

PROBLEM STATEMENTS :

01

X education sells online courses to industry professional.

02

X education gets a lot of leads, their lead conversion rate is very low, for example if they acquire 100 leads in a day only 3o if them are converted.

03

To make this process more efficient the company wishes to identify the most potential leads, also known as ‘hot leads’.

04

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.

BUSINESS OBJECTIVE :

- For that they want to build a Model which identifies the hot leads.
- X education wants to know most promising leads.
- Deployment of the model for the future

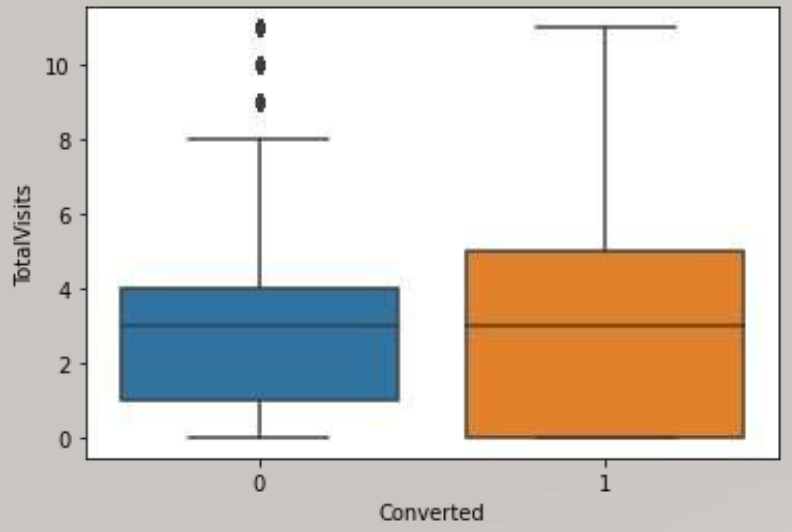
SOLUTION AND SUGGESTIONS :

- Data cleaning and manipulation
- Exploratory data analysis.
- Scaling and dummy variable treatment.
- Encoding the data.
- Classification method by Logistic regression.
- Validation of the model.
- Model presentation.
- Recommendations.

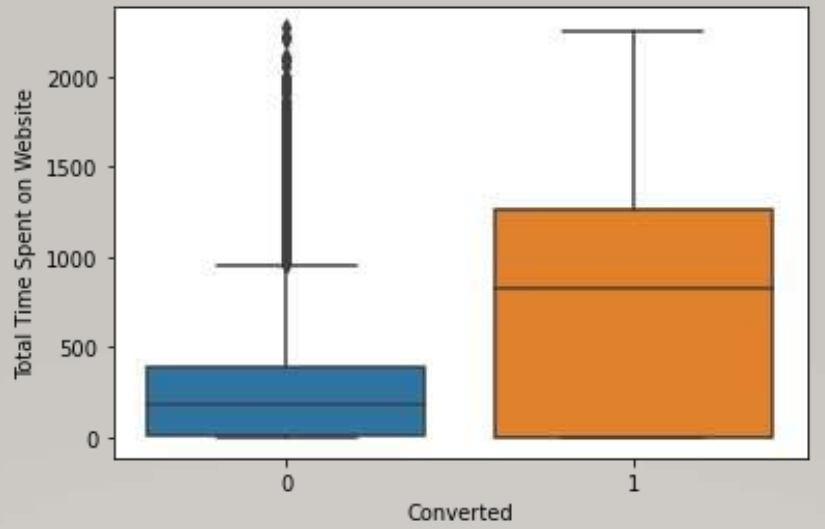
DATA MANIPULATION:

- Here in the given dataset many columns consist "Select" values hence will convert that into NaN values
- Now, identifying and Dropping the columns having more than 70% Null values to get better results
- Treating null values in each column one by one.
- Lead Quality: This will Indicate the quality of lead based on the data and intuition the employee who has been assigned to the lead.
- We can impute ‘Not Sure’ in the blank spaces in NaN safety, as Lead quality is based on the intuition of employee.
- We can drop few columns with 45% null values as it wont affect our overall analysis.
- Almost 60% of the data is Mumbai so we can impute Mumbai in the missing values.
- Creating a category "Others" for missing values.

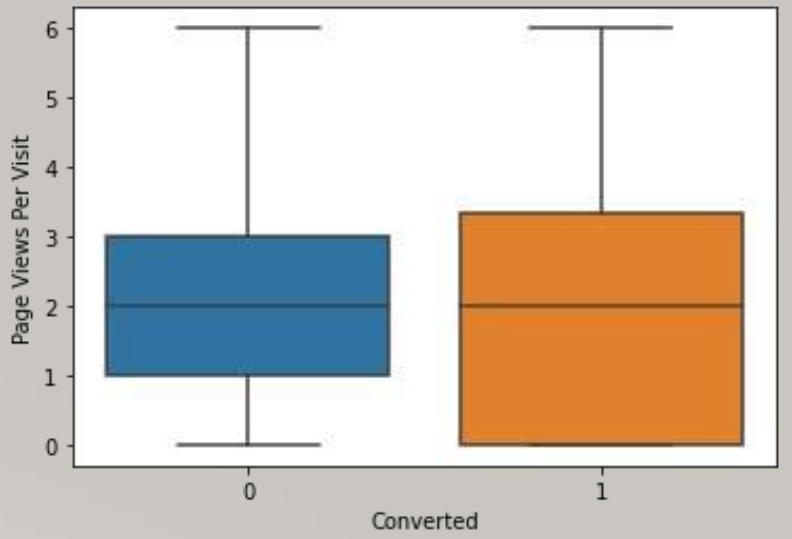
EXPLORATORY DATA ANALYSIS: Numerical Features



Median number of visits for both the converted as well as non-converted leads are more or less the same.So it seems that the total visits does not affect the conversion in general.

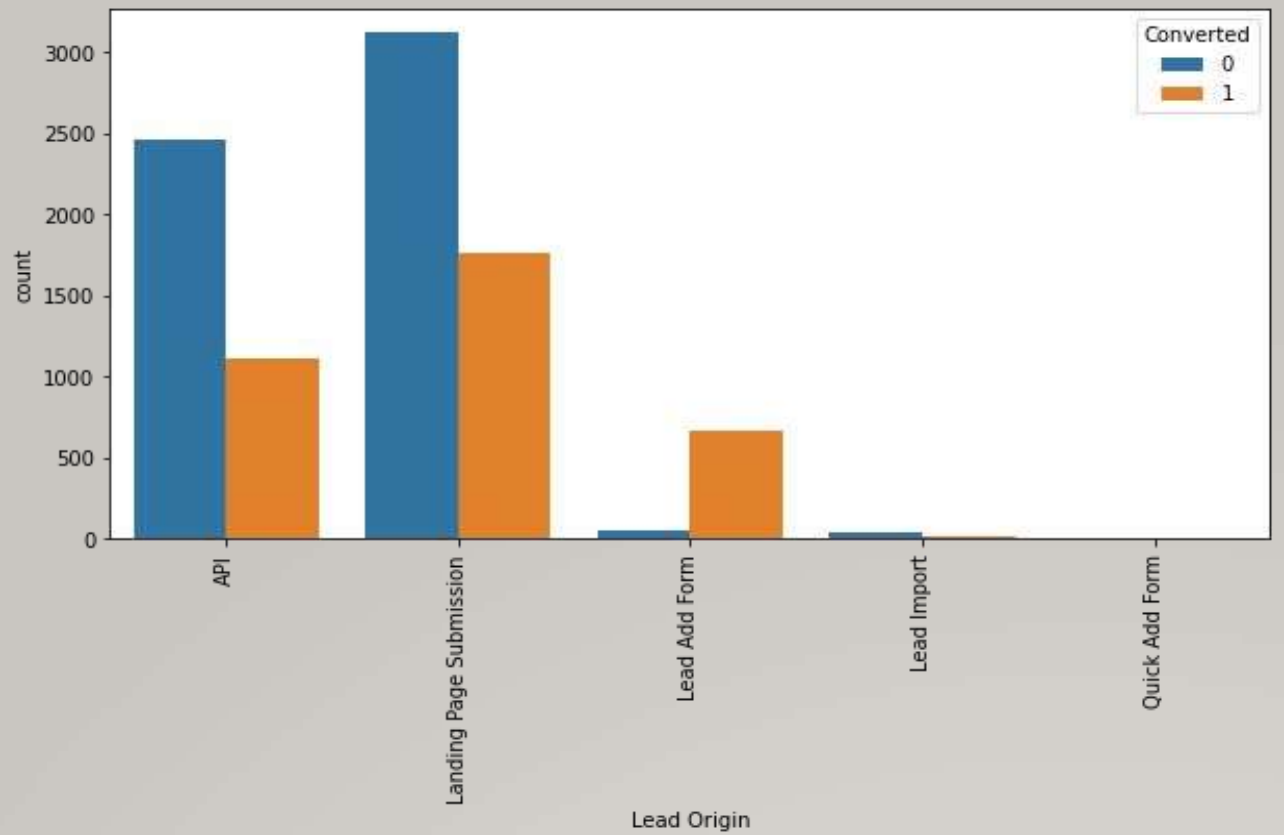


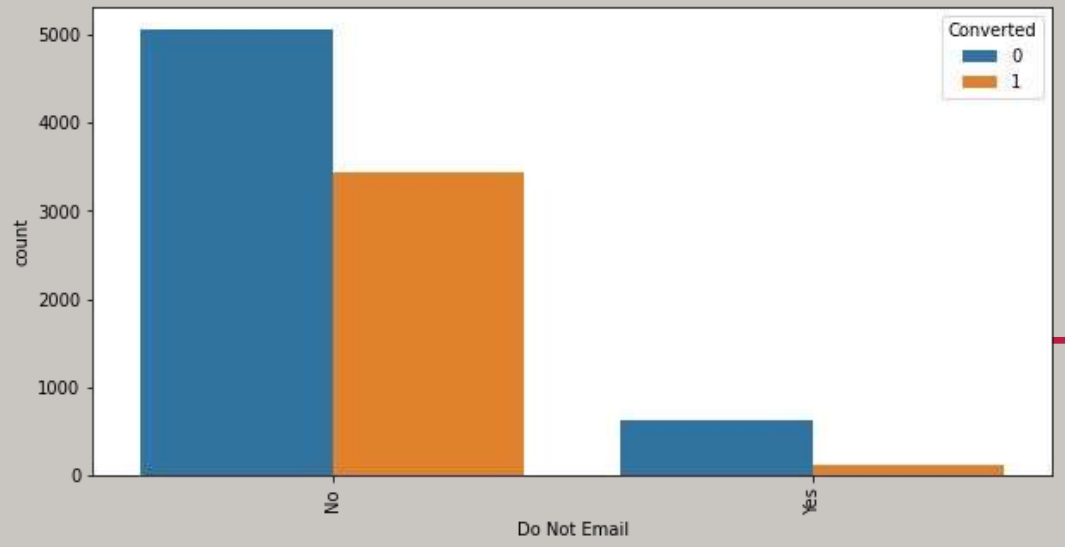
The total time spent on the website is clearly more for the converted leads. So it seems that the time spent is positively correlated to the conversion.



The average number of pages visited during the website views also don't seem to affect the conversion much.

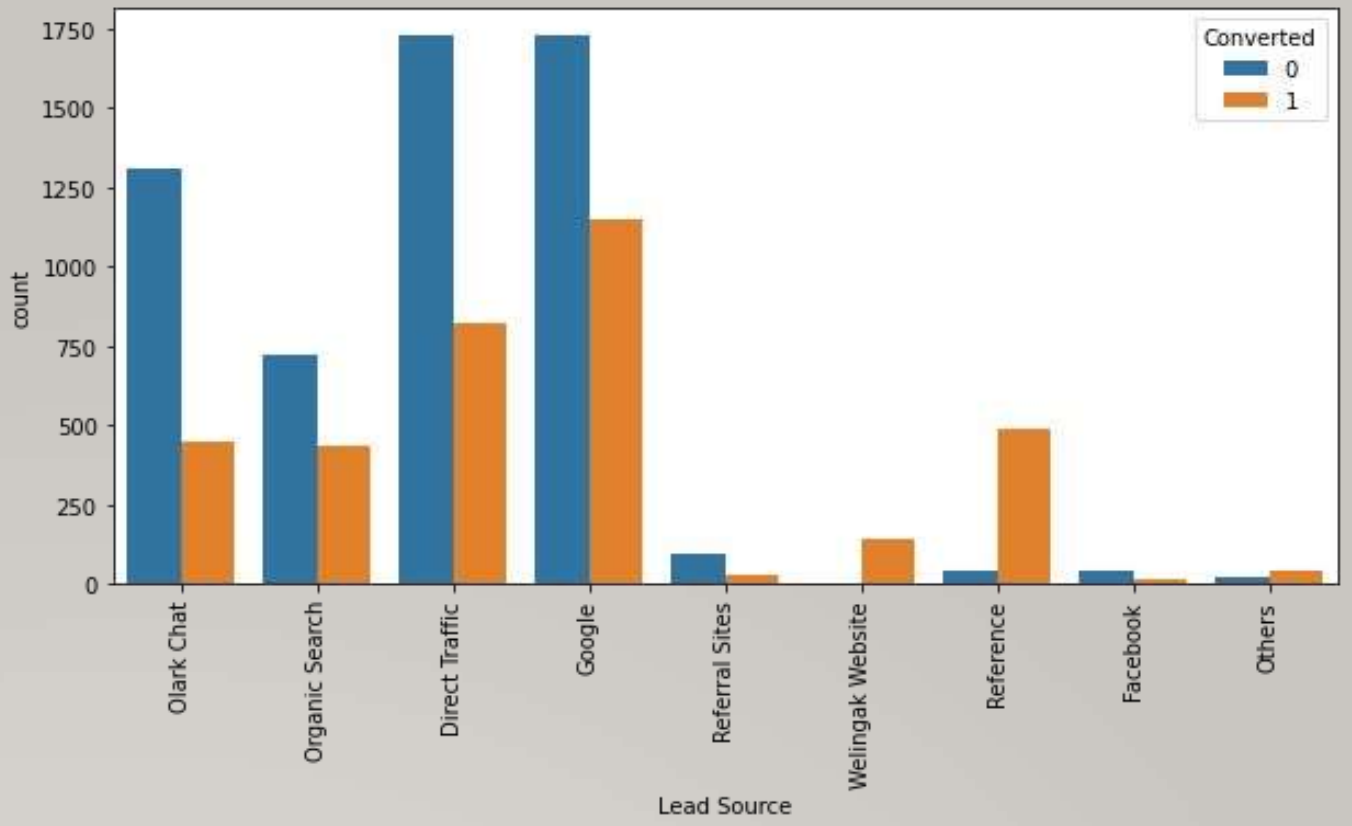
Categorical Features:

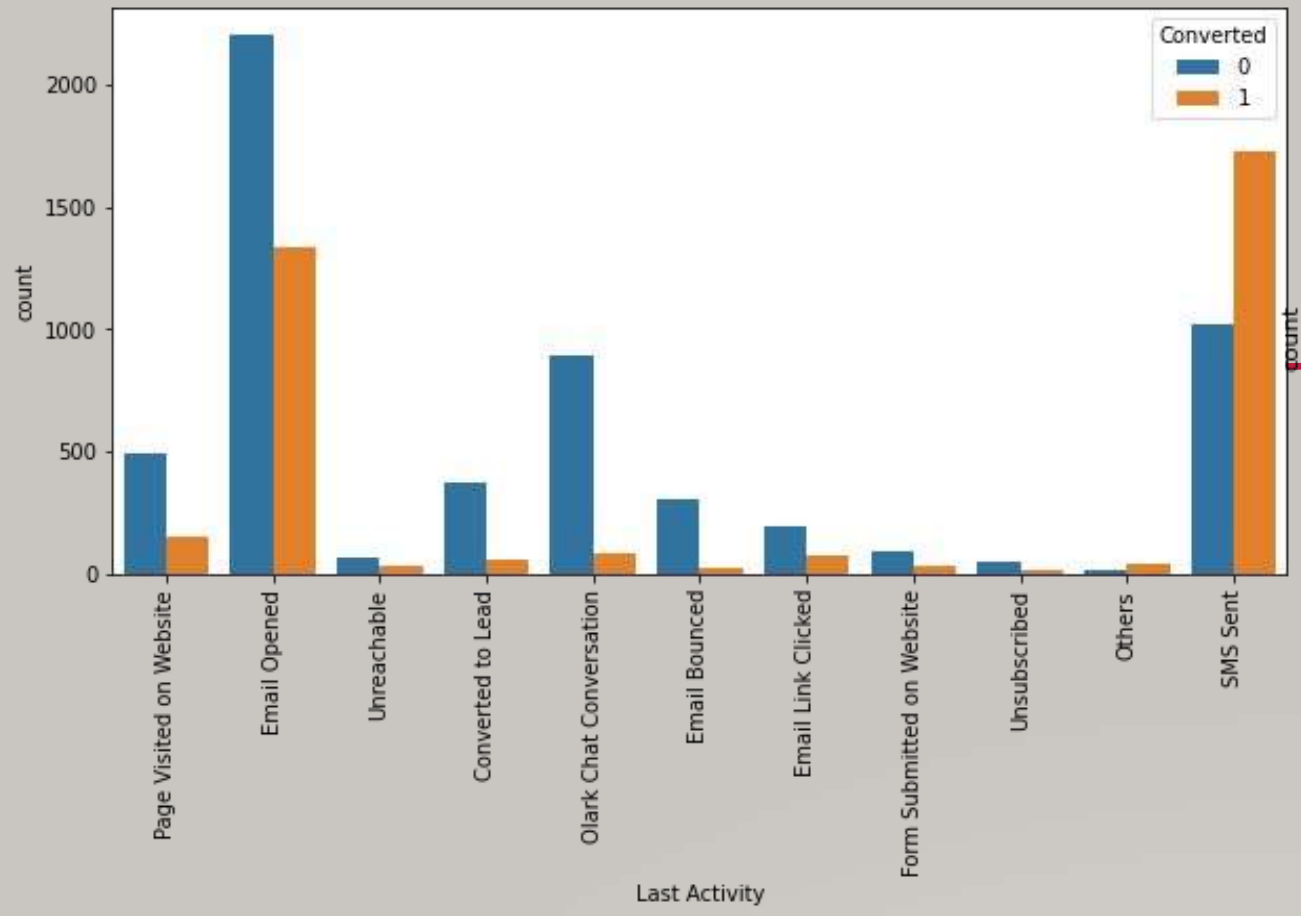


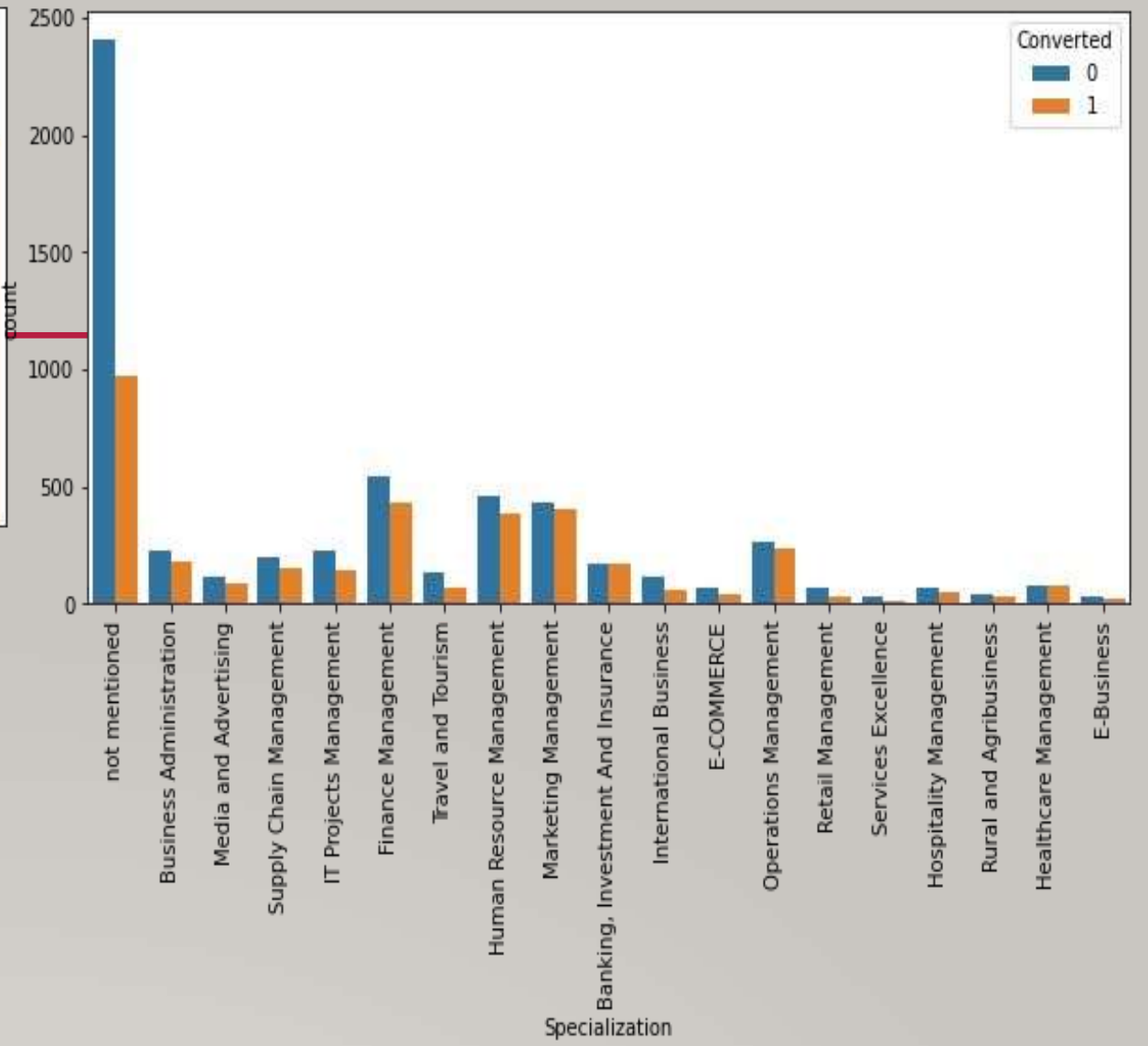


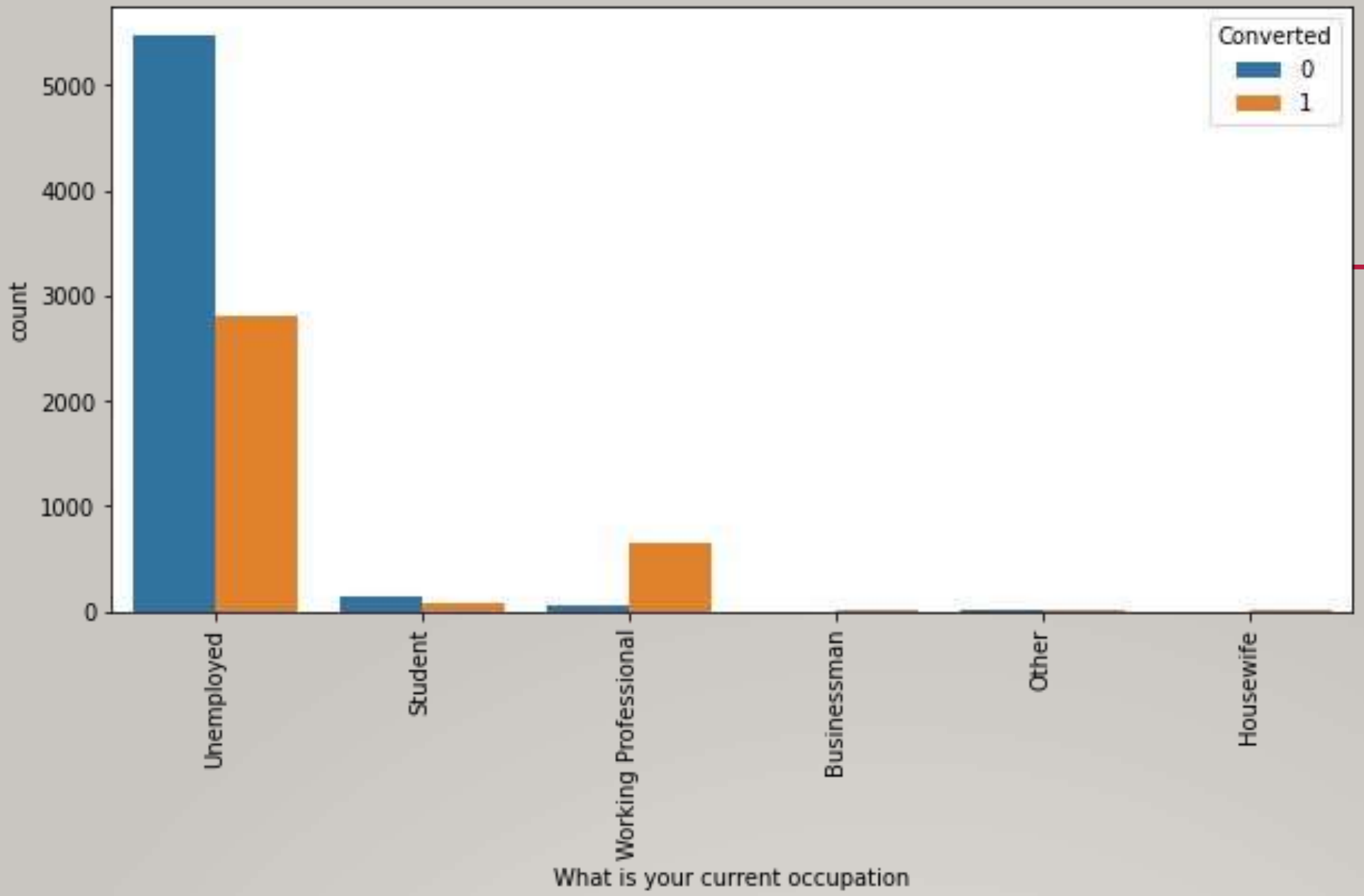
Nearly all the leads opted not to be mailed about the course. It does not seem to affect the conversions much.

Irrespective of the fact whether a lead gets converted or not, Google is the source of most of the leads. However, it is the reference that plays somewhat important role in the conversion since in this case, the number of conversions are higher than non-conversions.

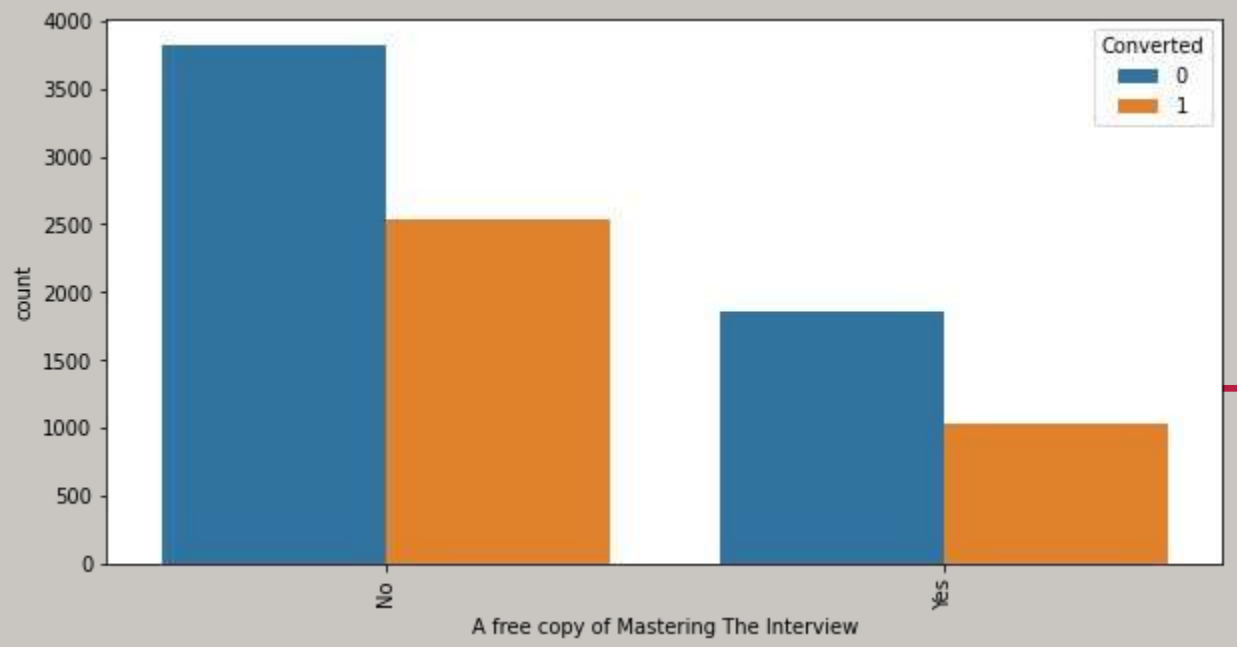


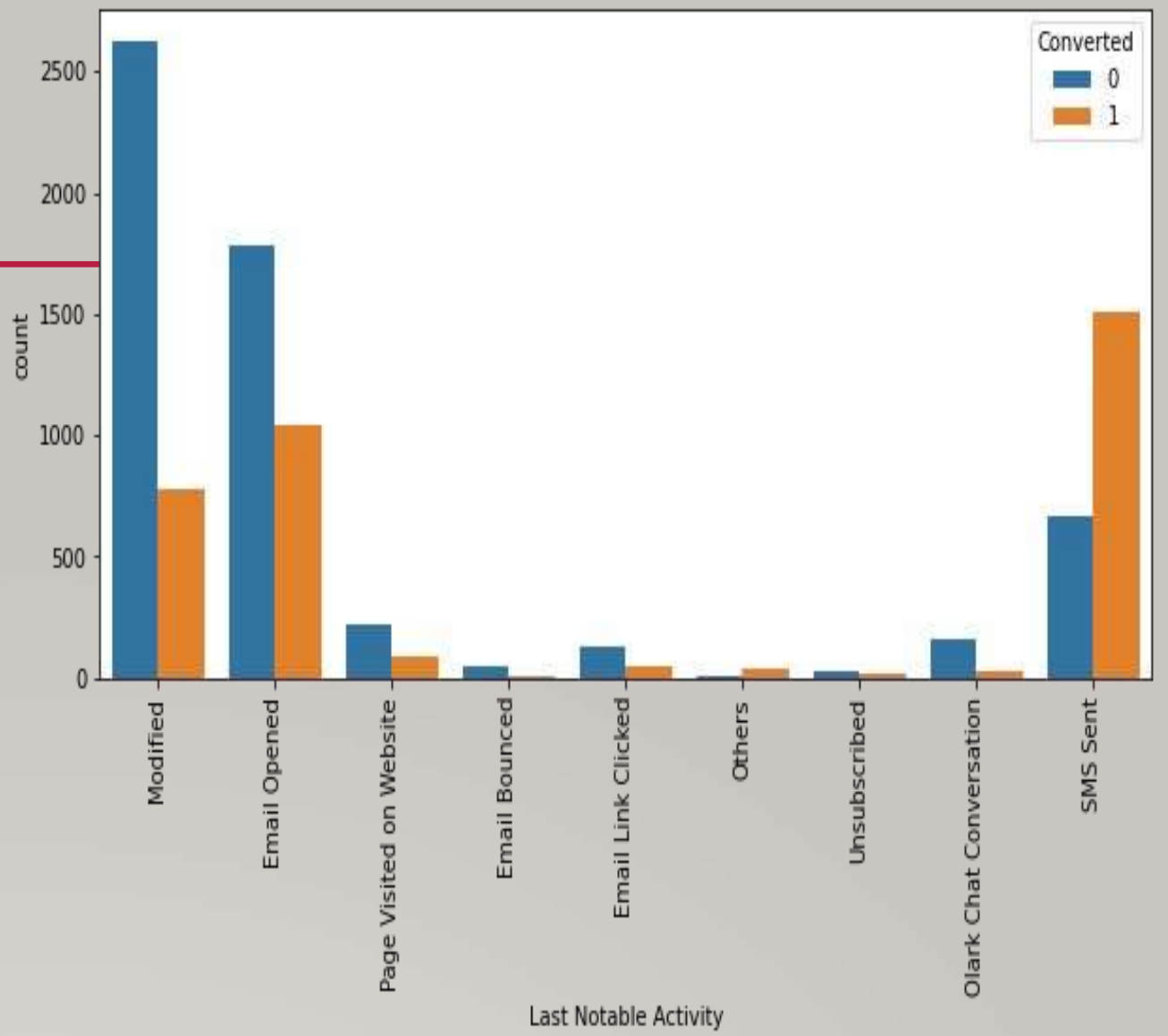

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Others
SMS Sent
Last Activity


Converted
0
1
count
2500
2000
1500
1000
500
0
not mentioned
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization



The working professionals show the highest degree of conversion. However, the unemployed are largest among the leads.





- **Most of the customers don't want a free copy of matering the interview.**
- **The last notable activity for most of the converted leads is SMS sent.Email opened and modified are also some significant activity but mostly done by both the categories of leads.**

MODEL BUILDING:

- Firstly, Split the Data into Training and Testing Sets.
- Now, applying the first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection Running RFE with 15 variables as output.
- Now, we build Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5. Followed by this, Predictions on test data set to be done.
- Model accuracy is almost 91%.

Recommendations :

Factors that affect lead conversion :

- Occupation (working professionals) : It was seen that working professionals were mostly covered. Hence the company should focus more on this vertical.
- Time spent on website. We should concentrate more on the time spent on the website as it directly affects the lead conversion. When the lead origin is Lead add format, the lead conversion is very high and should be focussed upon.
- The following sources are needed to be focussed:

a. Google

b. Direct traffic

c. Organic search