

Lead Score Case Study


Logistic Regression

Presented by
Pallabita Ghosh



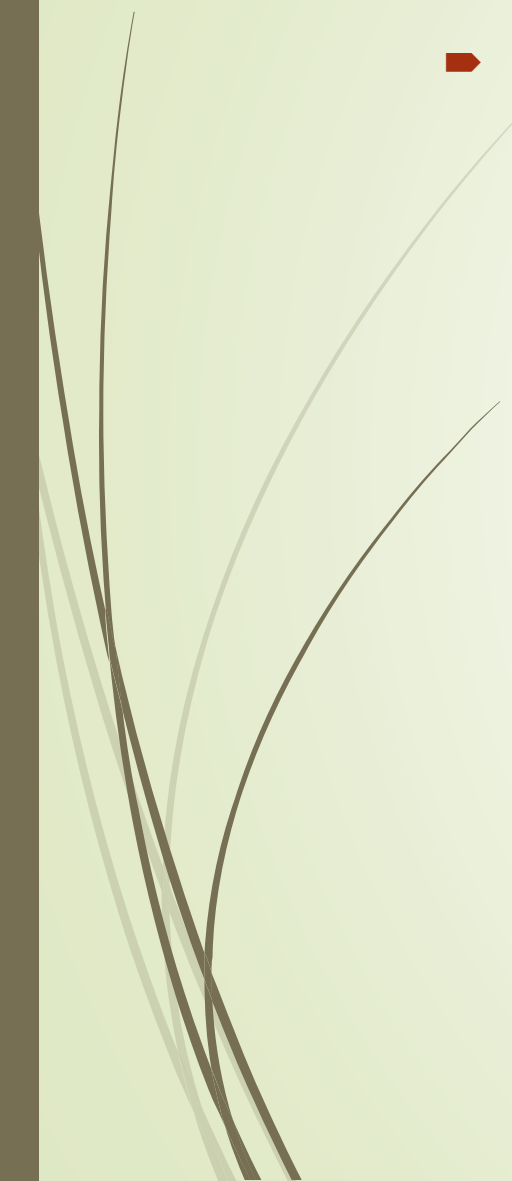


Problem Statement

- An education company name X sells online courses to industry professionals. So people who are interested for their courses can land their website and browse.
 - Company's main agenda is to increase leads through various platform and different possible ways who can purchase the courses.
 - Though company is getting lots of leads through call, email, recommendation, social platform, web browsing but very few people are interested to pay for their courses. This typical lead conversion rate is around 30%.
 - So making this process more efficient, company wants to focus more on potential leads or hot leads, sales team will concentrate more on them make phone calls with them.
- 



Goal of the Case Study

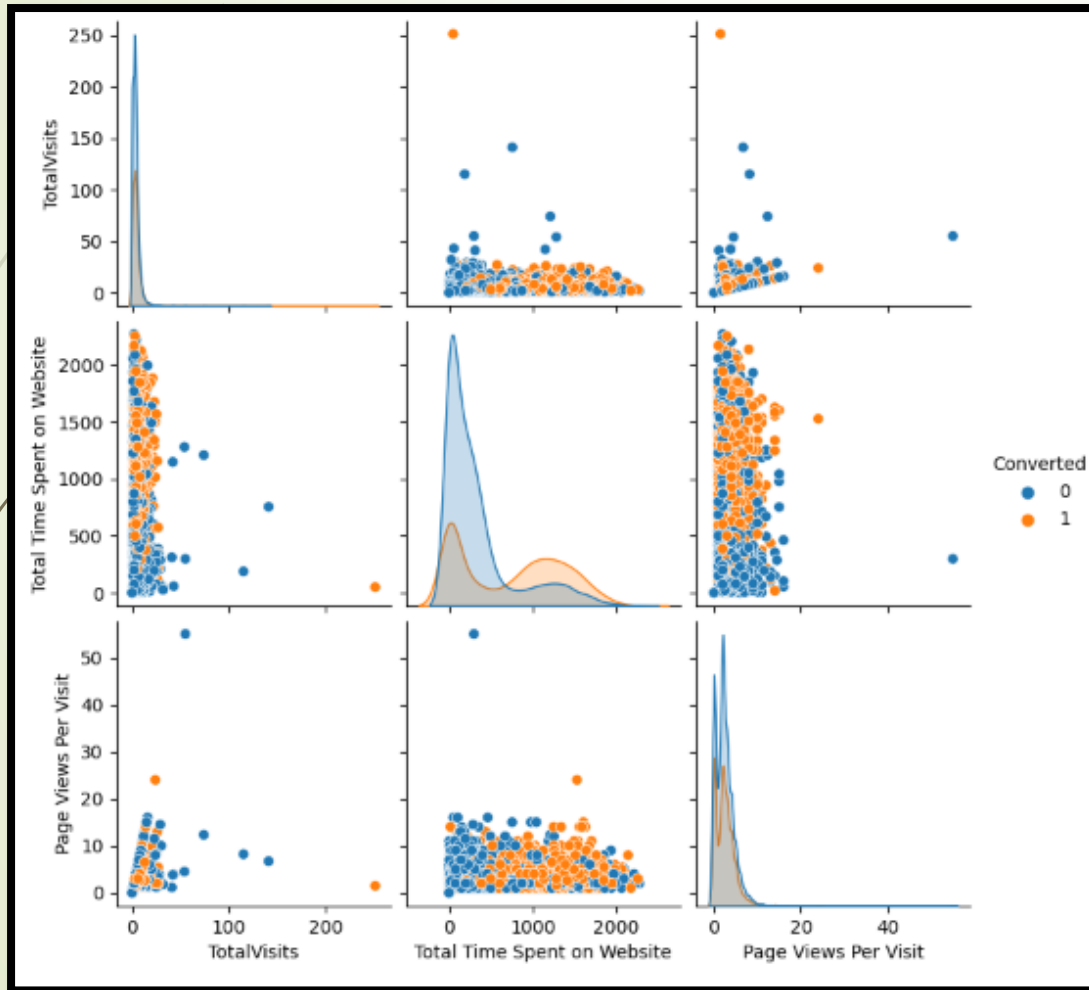
- Building a logistic regression model to assign a lead score between 0 and 100 of the each of the leads which is used by the company to find potential leads where higher score means people are more likely to convert and lower score means people are less likely to convert. There is some other problems presented by the company which model should be able to adjust to if the company requirement changes in the future so we will need to handle this as well.
- 



Source

- Import data for analysis
- Read and Understand the data
- Clean and Manipulate data for further analysis
- Finding and replaced missing values, Treated outliers
- Explanatory Data Analysis
- Data Preparation
- Model Building
- Train Test Split
- Feature Scaling
- VIF Checking
- Making Prediction on Train Data
- Calculation of Confusion Matrix, Accuracy, Sensitivity and Specificity
- Calculation of ROC Curve
- Calculation of Precision and Recall
- Making Prediction on Test Data
- Model Evaluation

Numerical Variables



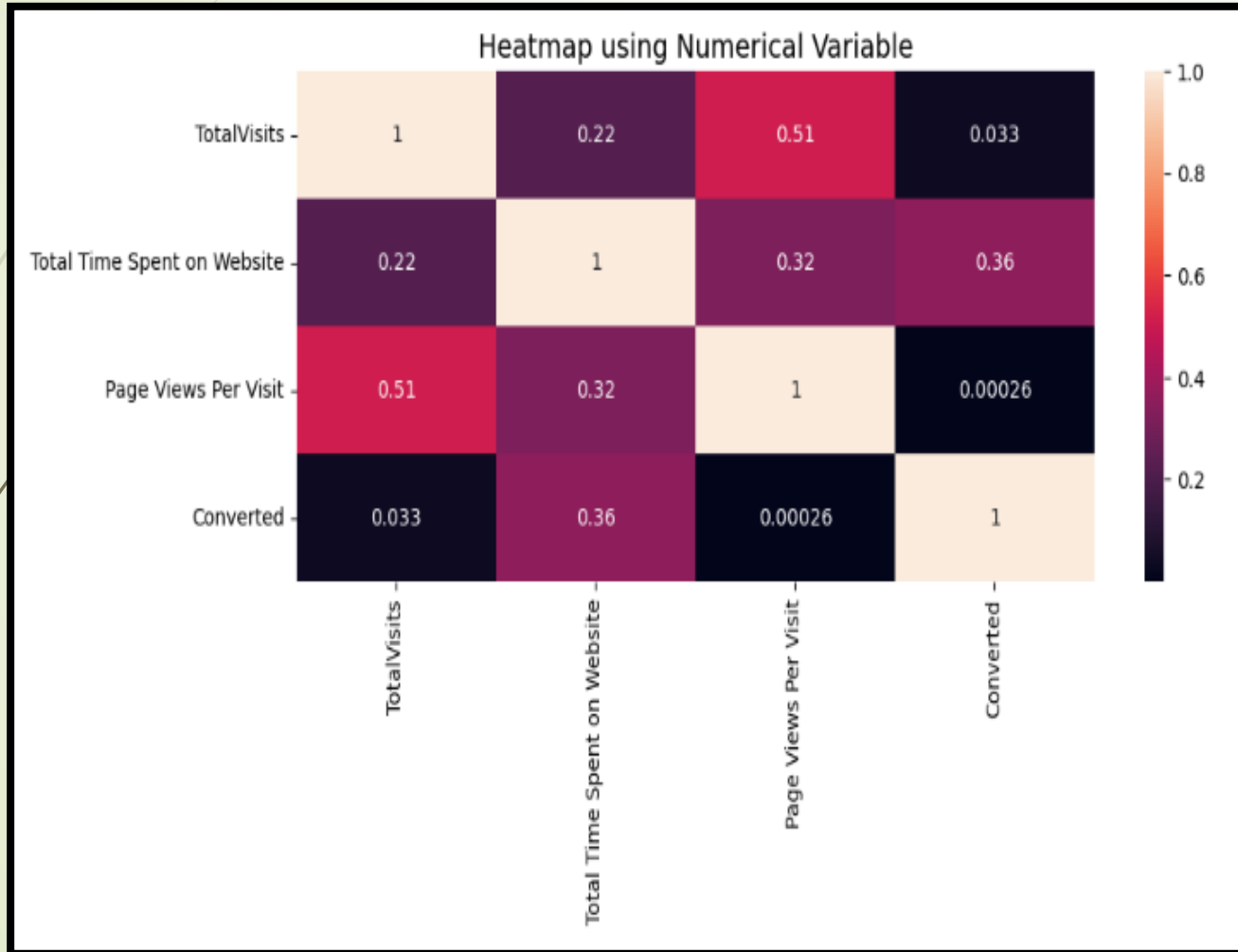
- TotalVisits- Number of visit is less by customer.
- Total Time Spent on Website – People are spending more time on website.
- Page Views Per Visit – Page visit per customer is low.



Data Cleaning, Manipulation and Preparation

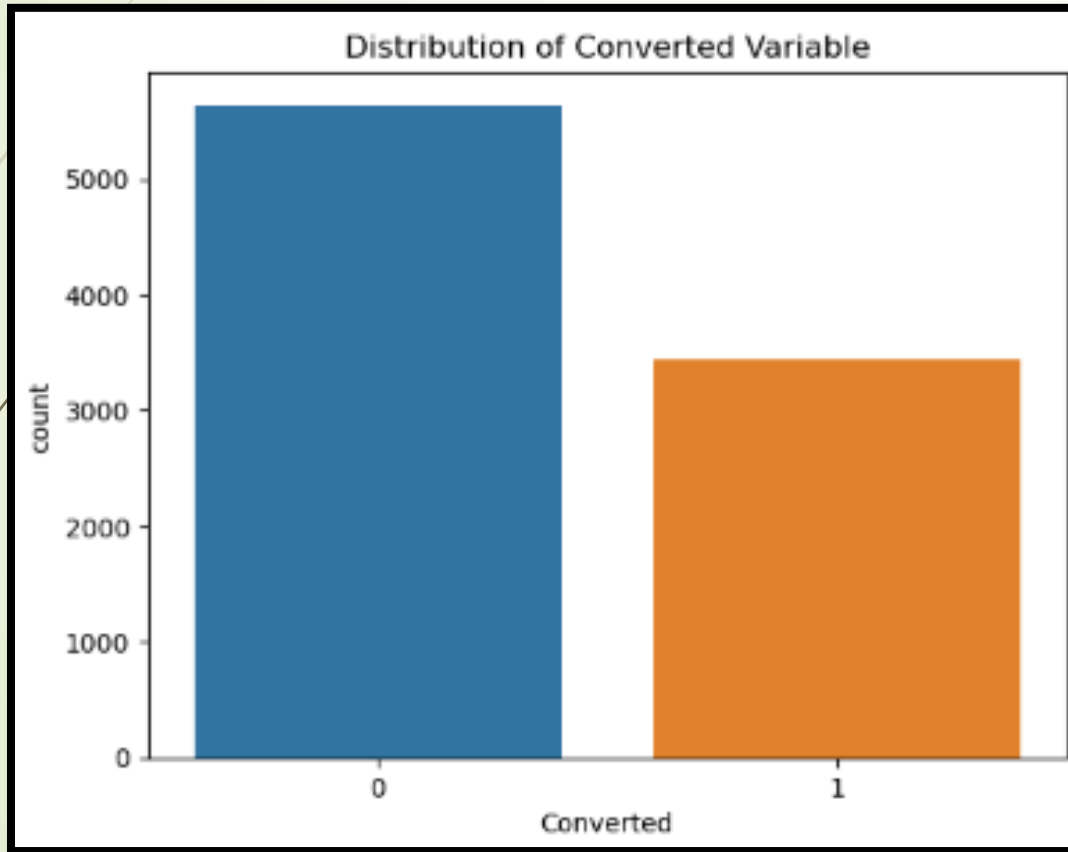
- Understand dimension of the dataset
- Statistical aspects of data
- Checked data type and percentage of missing values
- Imputation of missing values
- Removed columns with more than 40% missing value
- Replaced 'Select' values with null value and updated as 'Not Specified'
- Dropped all irrelevant columns with one variable
- Identified and worked on Outlier
- Converted categorical columns from Yes/No to binary value 1/0
- Created dummy variable deleted actual columns
- Re verified all the data for null values.
- Data Analysis
- Feature Standardization

Heatmap using Numerical Variables



- Form this graph, 'Total Visits' and 'Page Views Per Visit' has high correlation.
- Form this graph, 'Total Visits' and 'Total Time Spent on Website' has low correlation.

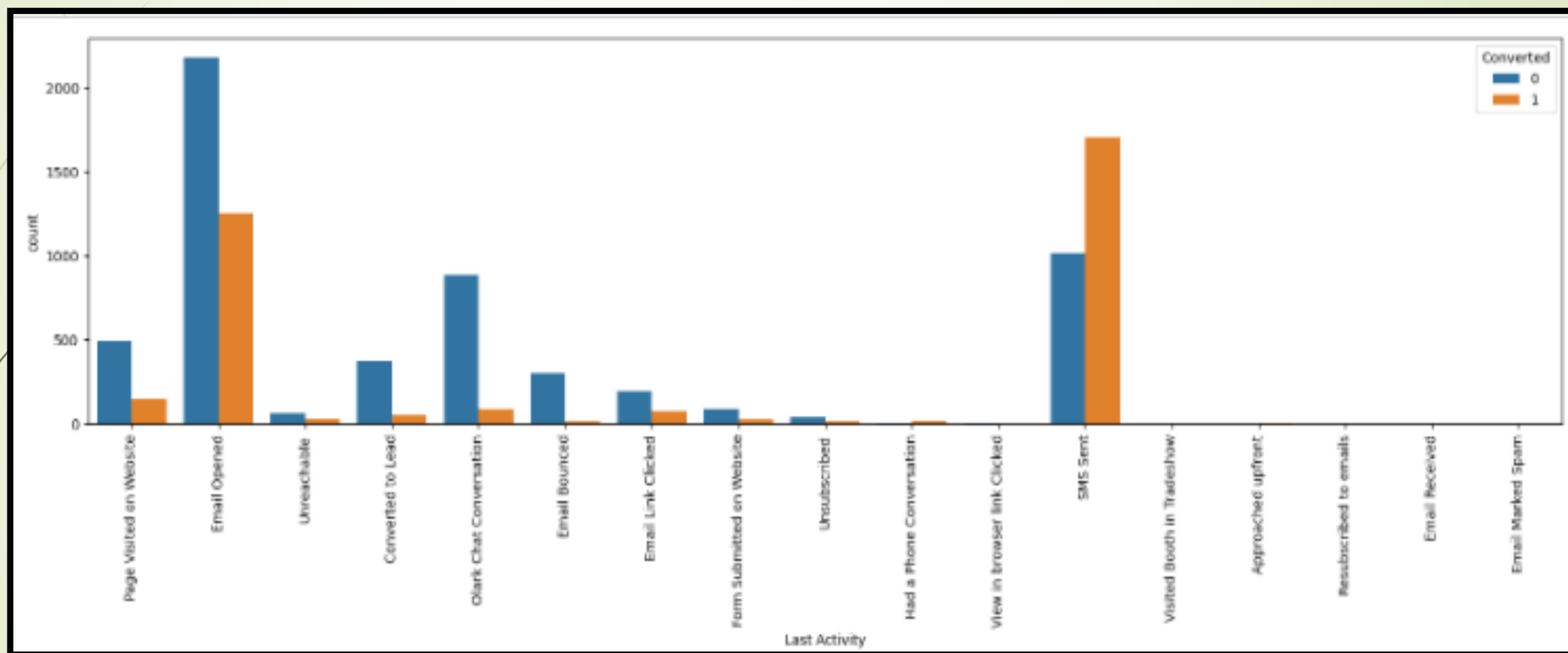
Target Variable - Converted



1 – Converted or Lead people
0 – Non-converted or Not Lead

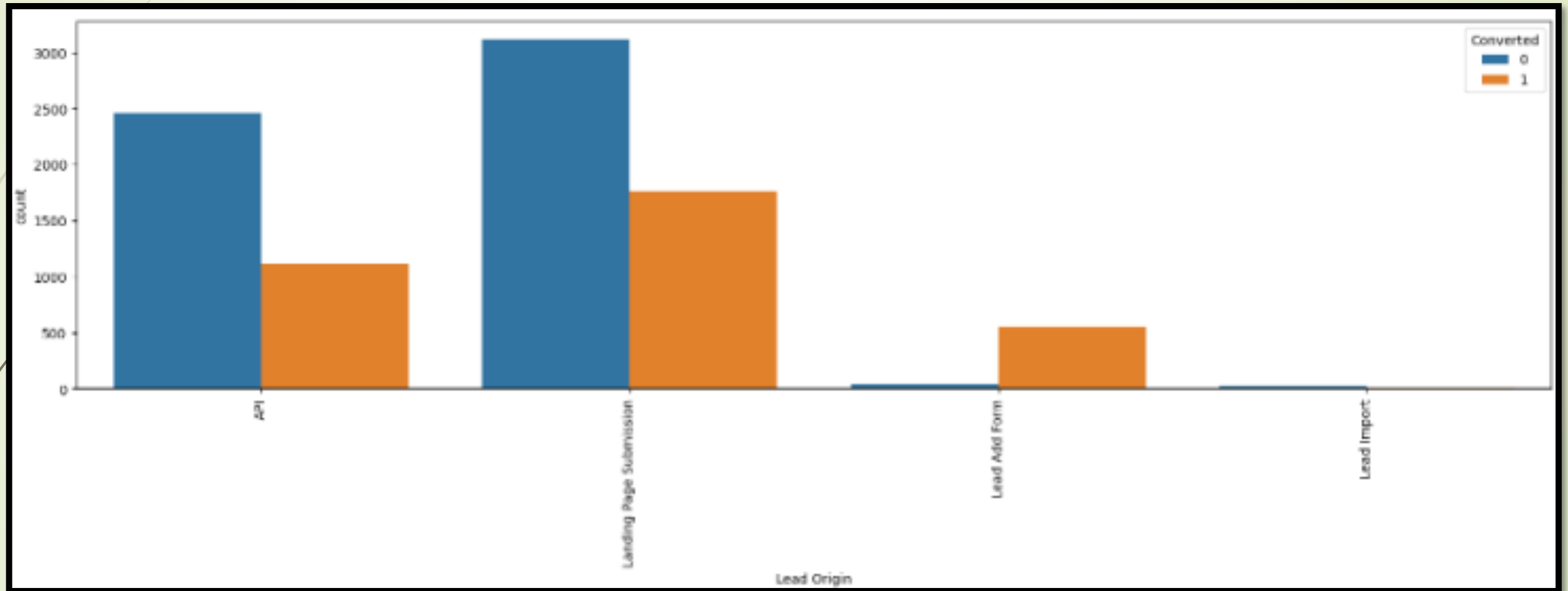
- People are less interested in becoming leads.

Last Activity vs Converted



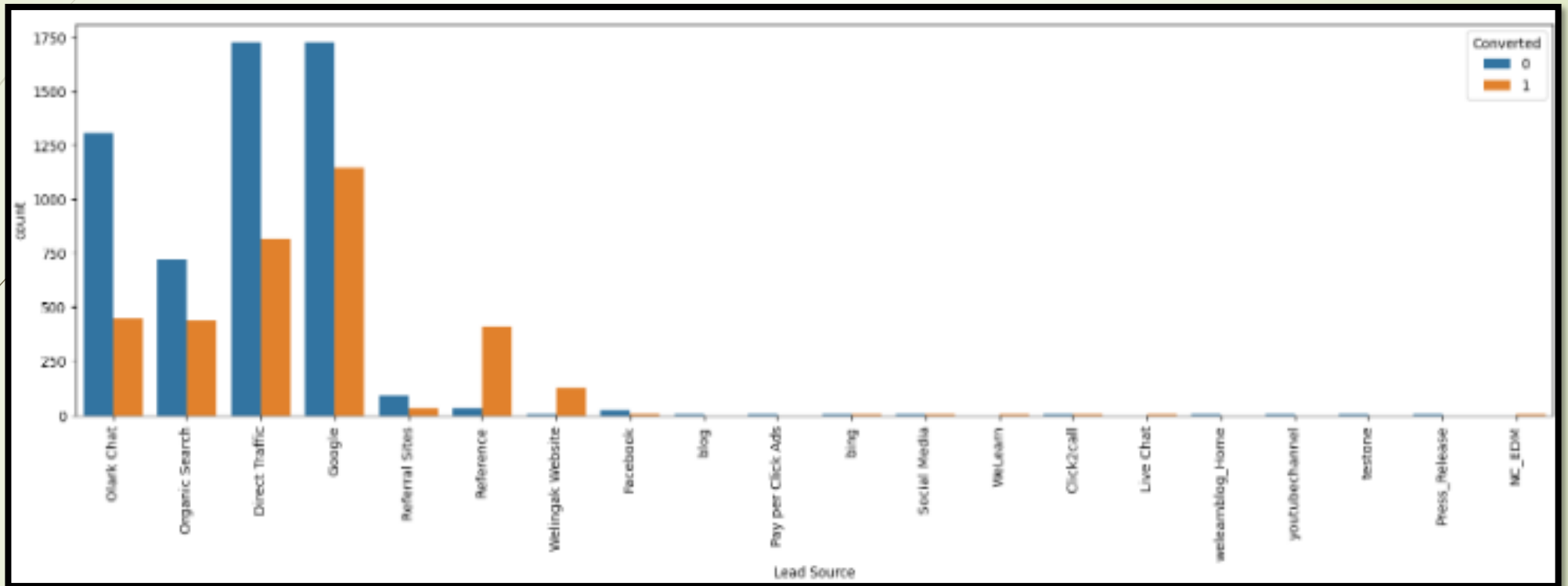
- SMS sent and Email opened are mostly providing highest number of leads.

Lead Origin vs Converted



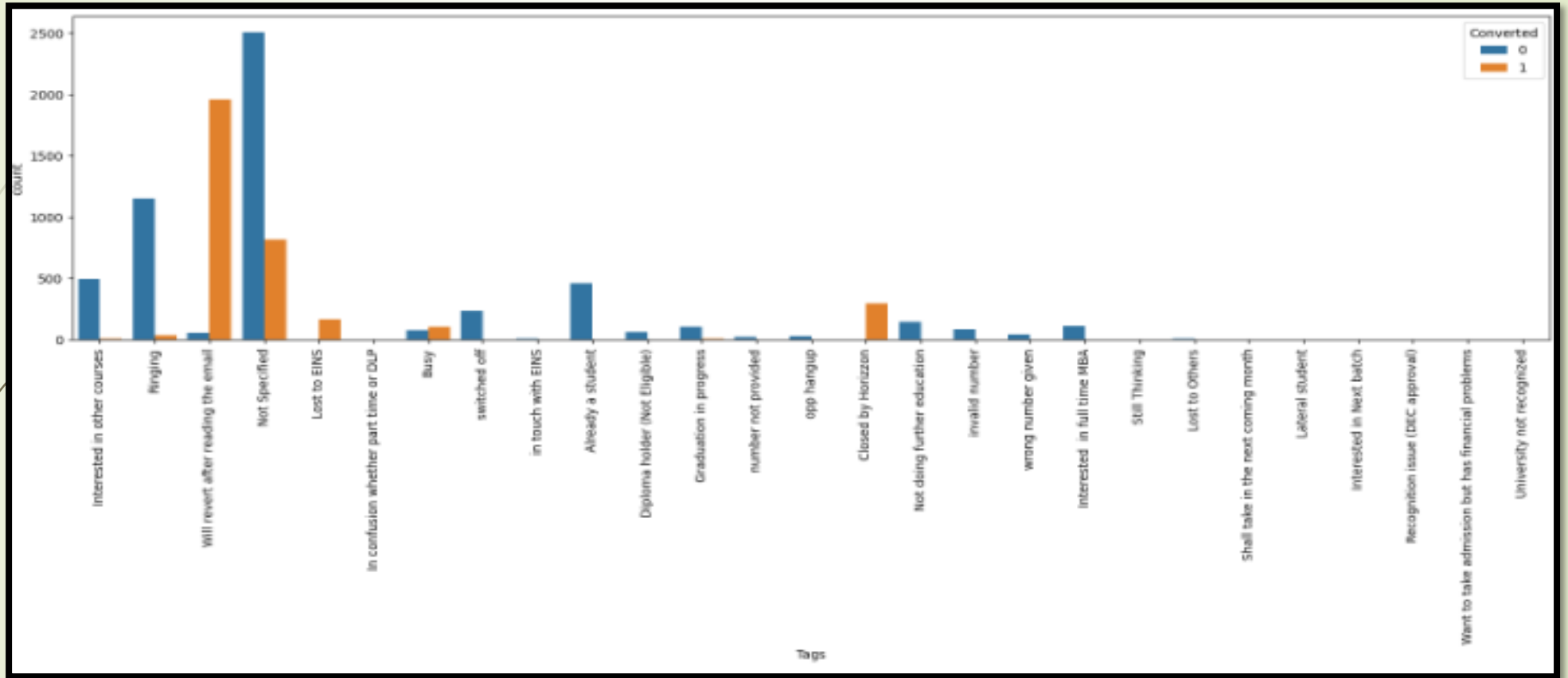
- Landing Page submission provides most no of leads with API and Lead add form.

Lead Source vs Converted



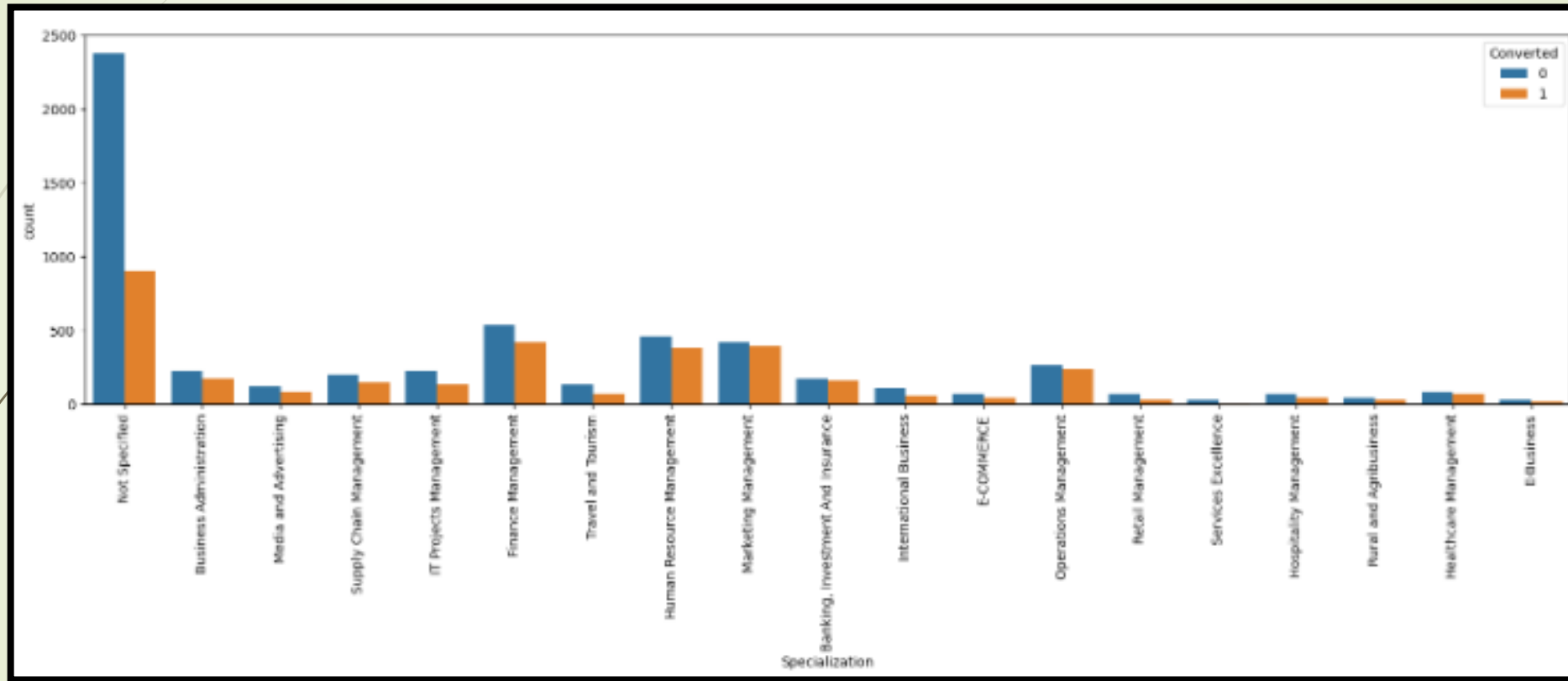
- Google provides highest number of leads apart from direct traffic, reference, organic search and Olark chat.

Tags vs Converted



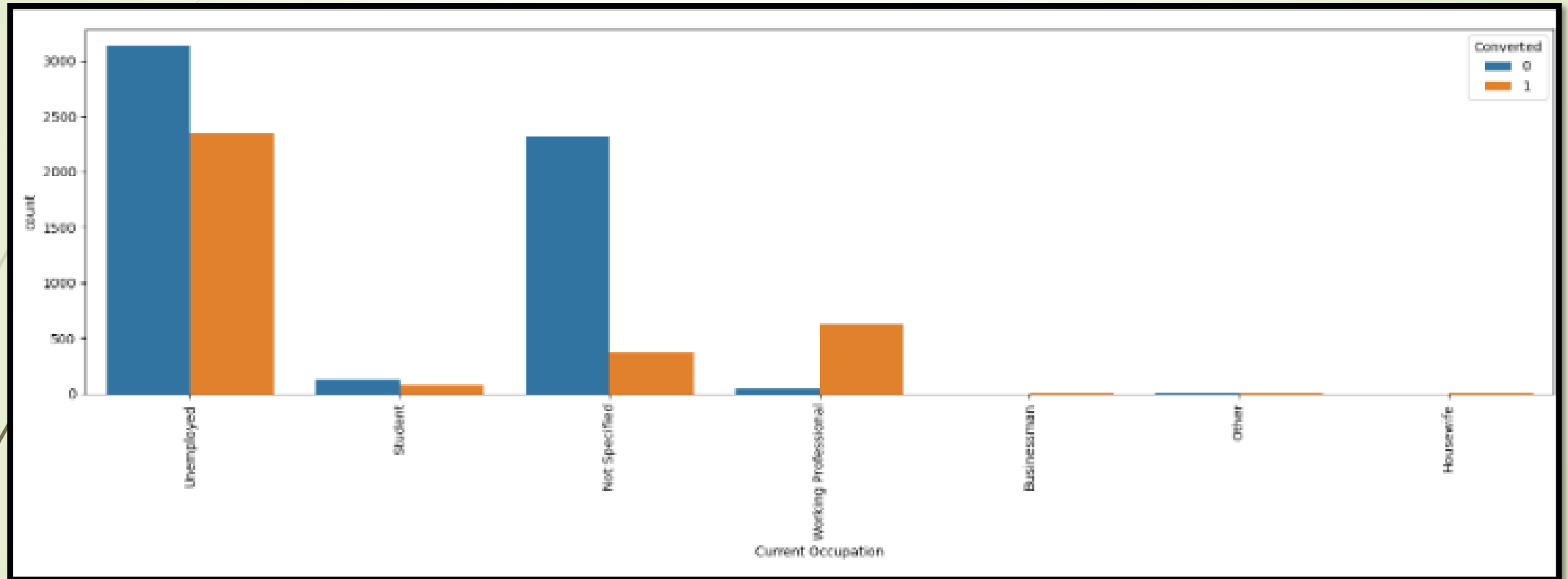
- People revert after reading the email has high no of converted leads.

Specialization vs Converted



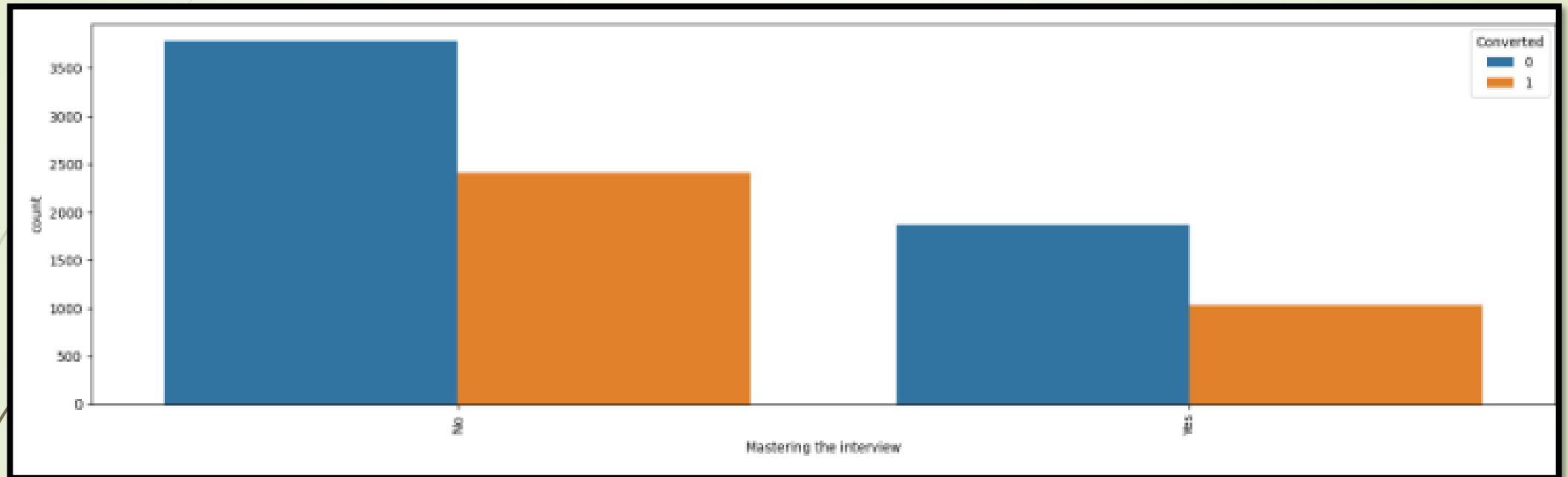
- Most of the customers have not mentioned about their specialization, Apart from that management peoples are mostly focusing on specialization and they are the leads.

What is your current occupation vs Converted



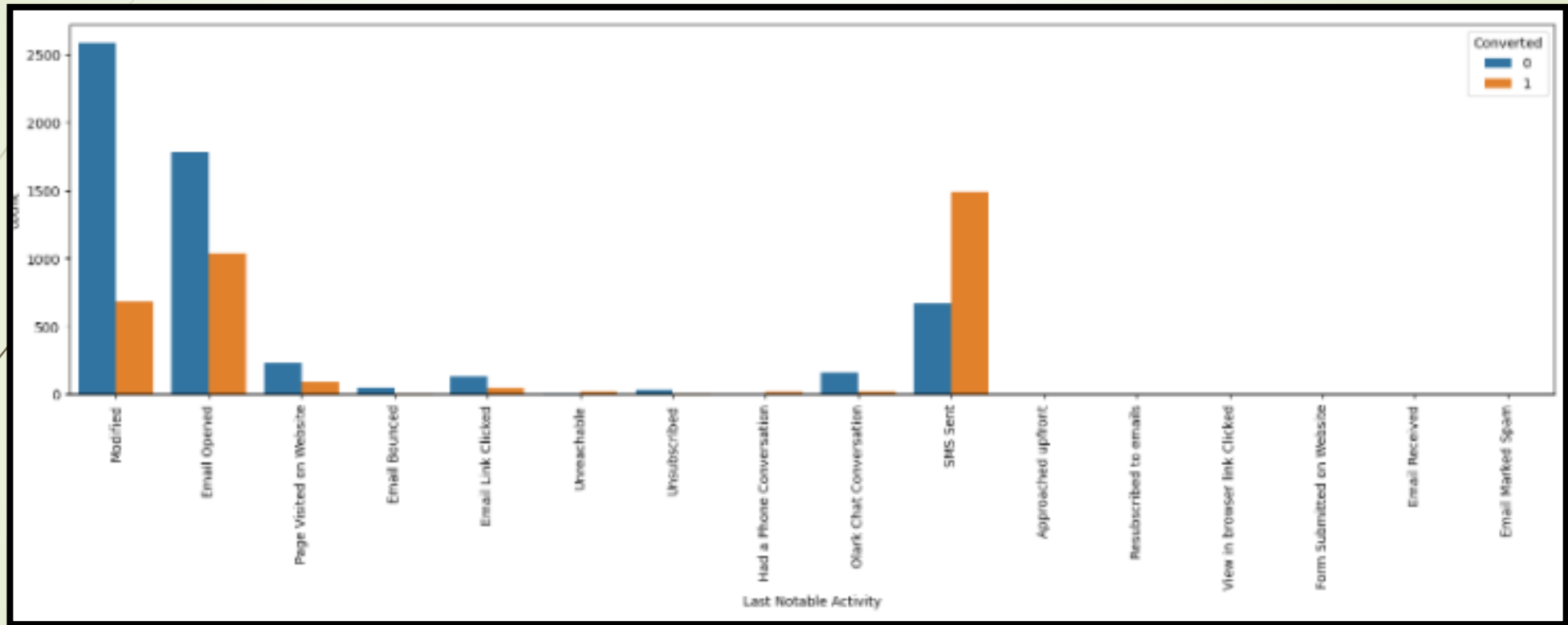
- Most of the people are unemployed who wants to become leads. Apart from that working professionals are mostly focusing to become leads.

A free of mastering the interview vs Converted



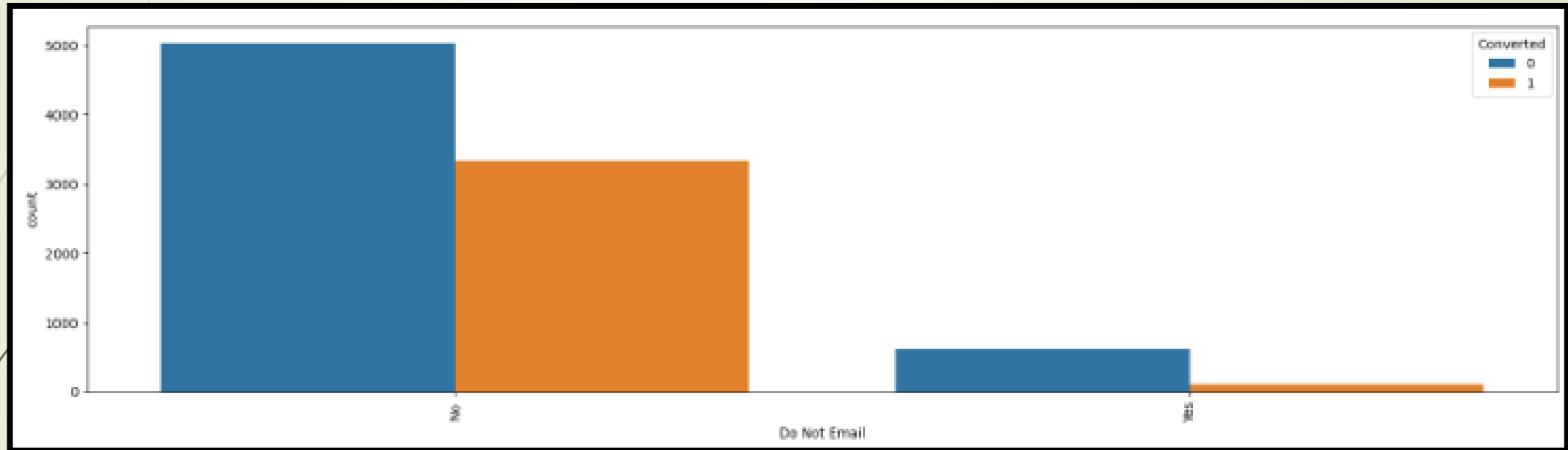
- Customers are not interested for mock interview.

Last Notable Activity vs Converted



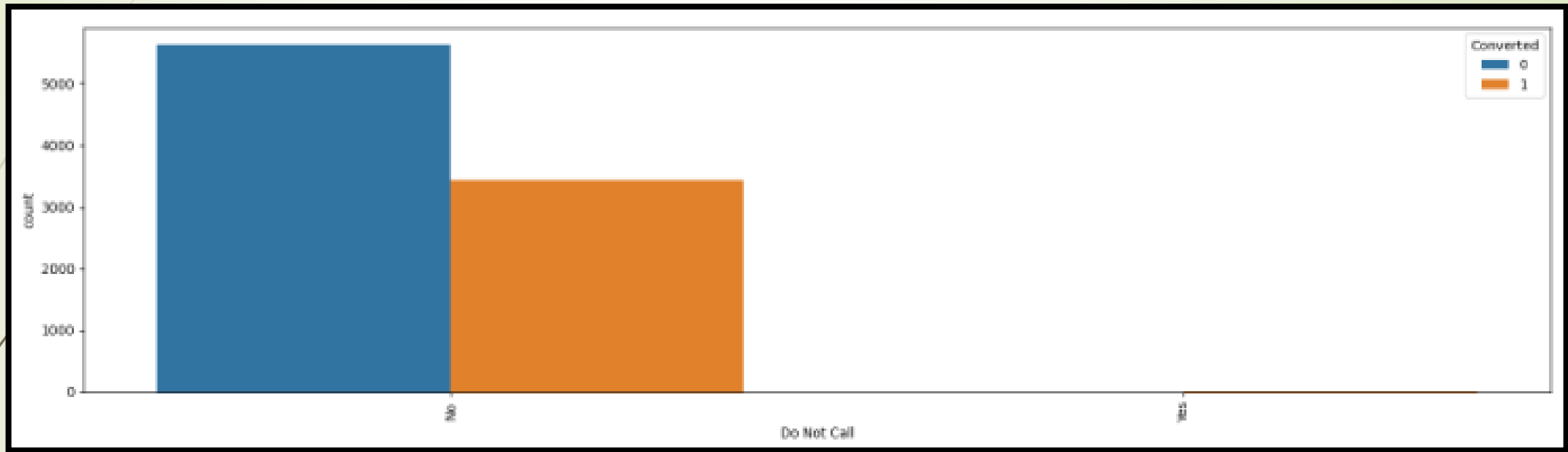
- As per last notable activity, customers want messages for converting leads.

Do not email vs Converted



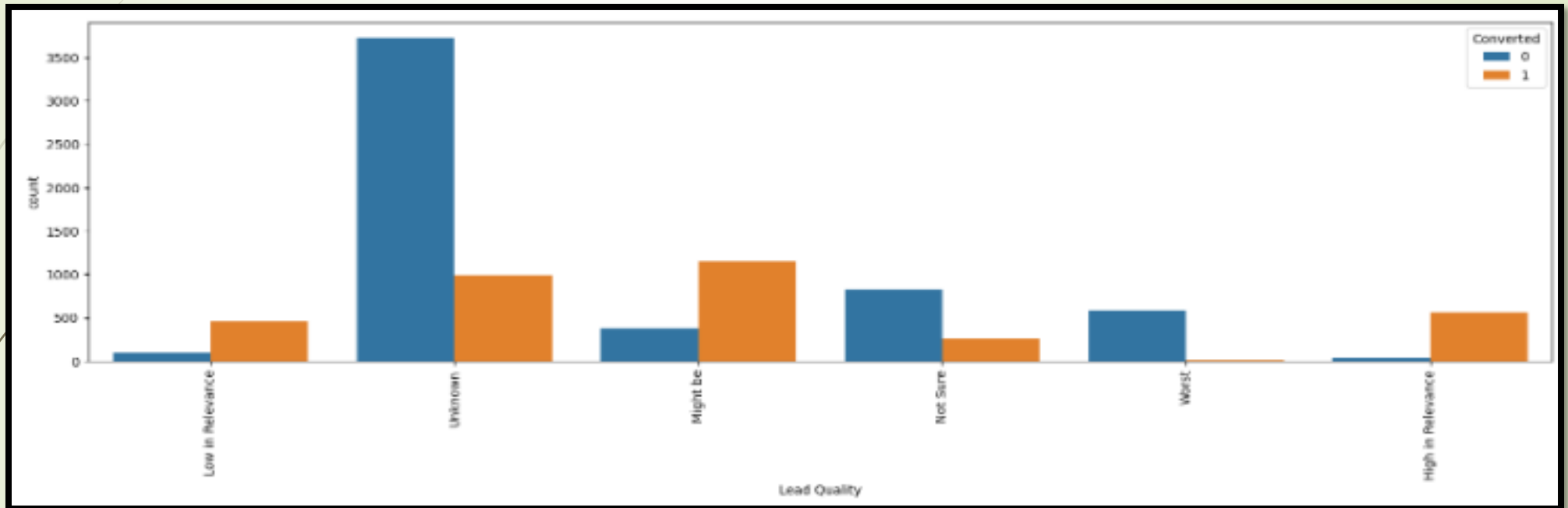
- Customers are less interested to receive emails for leads.

Do not call vs Converted



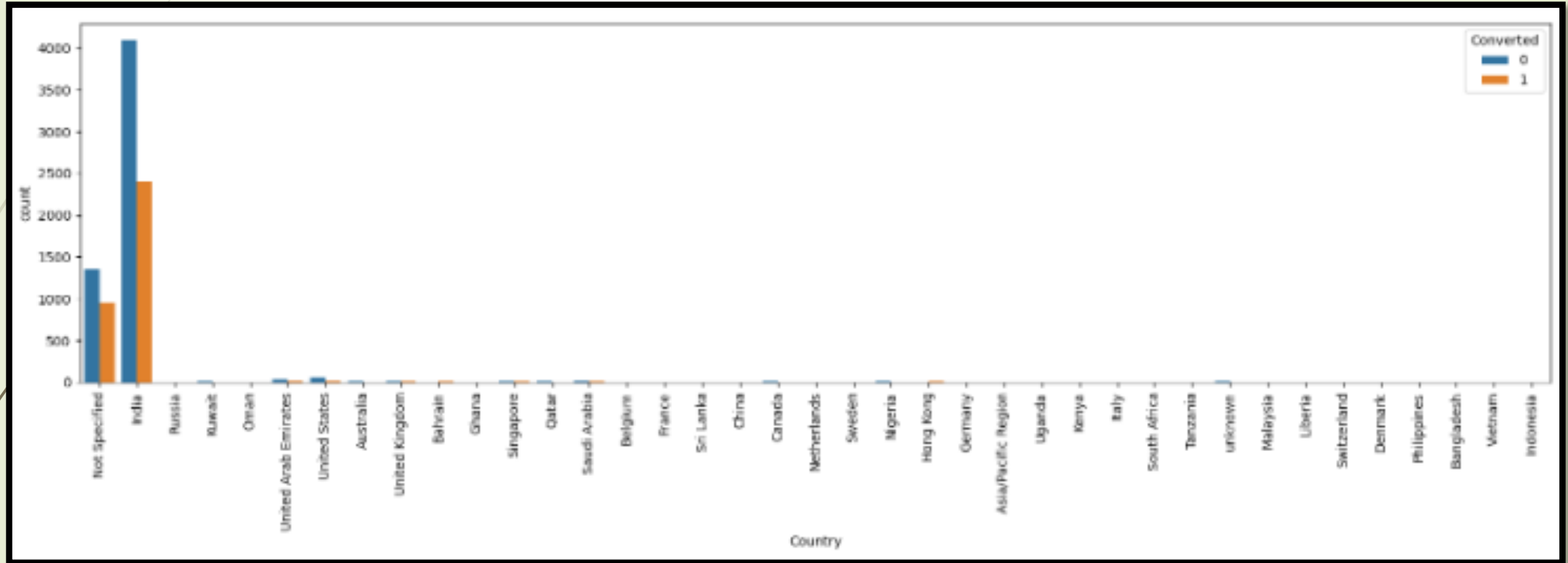
- Customers are less interested in getting calls for leads.

Lead Quality vs Converted



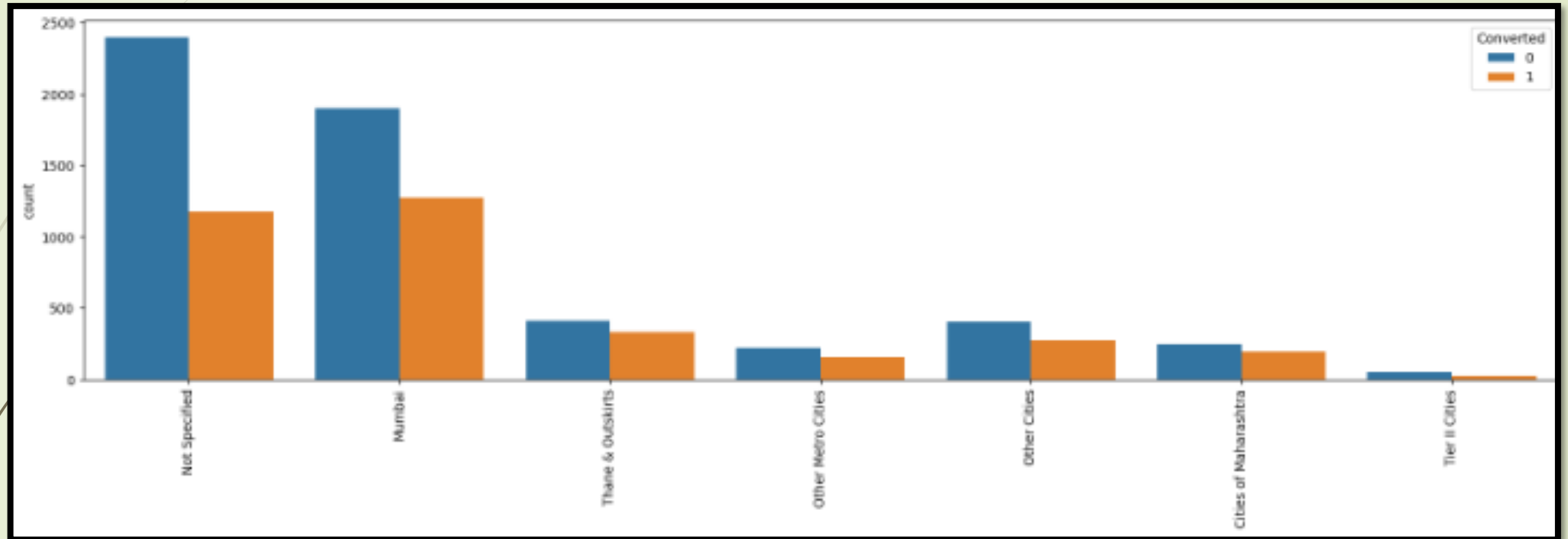
- Highest no of leads are from low and relevance might be and high and relevance category.

Country vs Converted



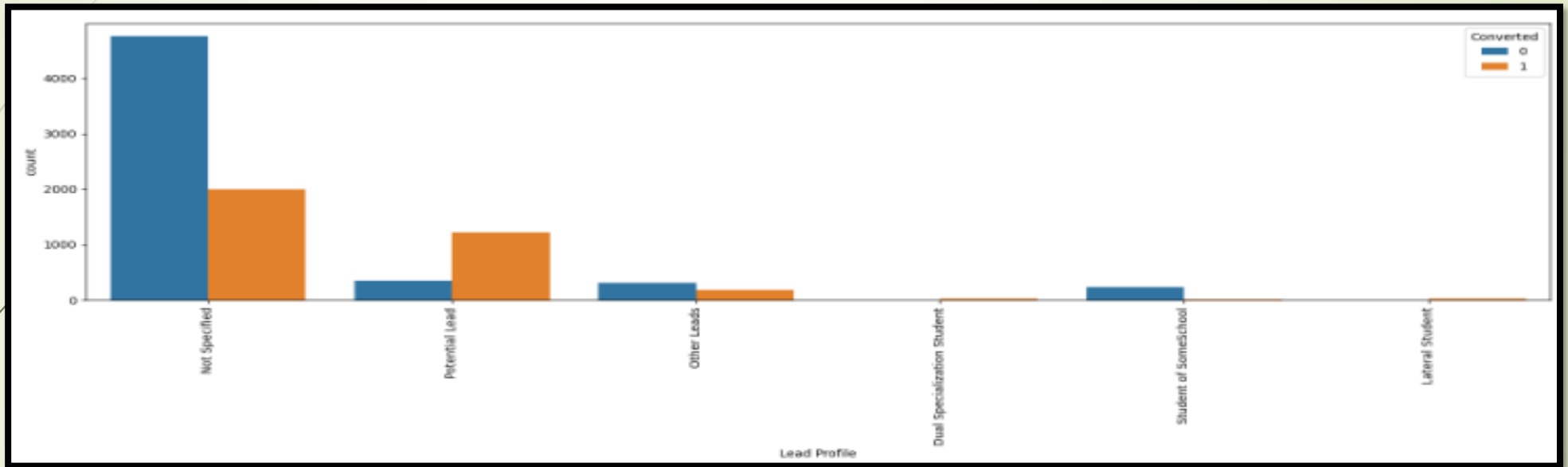
- Most of the customer are from country India but very few of them want to become leads.

City vs Converted



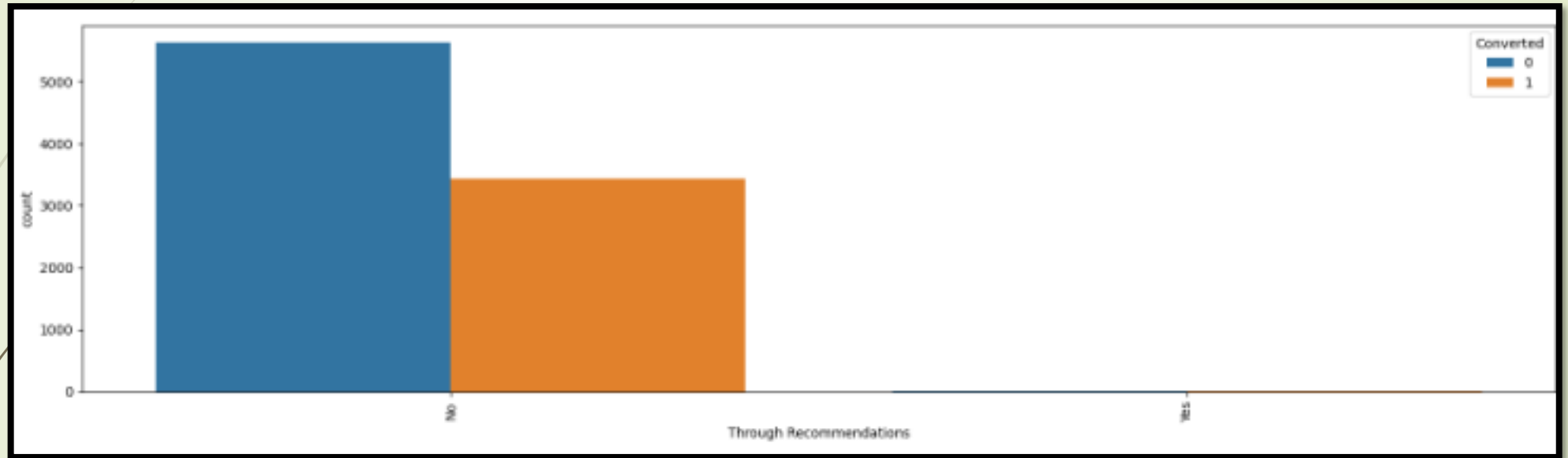
- Half of the people have not mentioned about city, those who mentioned mostly from Mumbai, became leads, others are from other cities

Lead Profile vs Converted



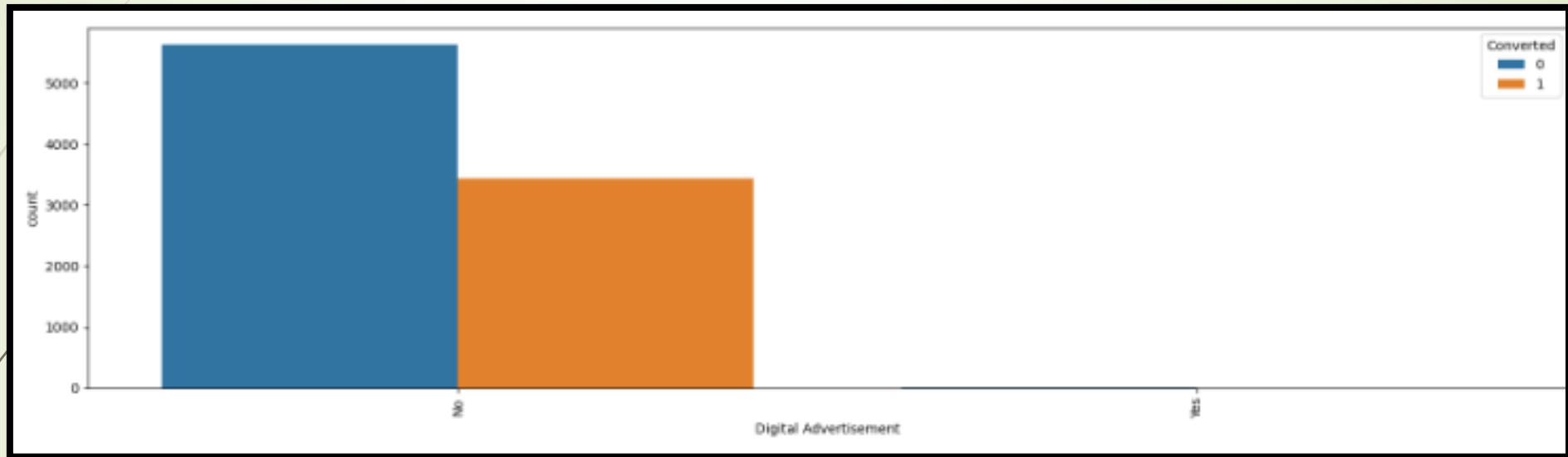
- As per lead profile, potential leads has very high conversion rate and large number of people are not interested to mention anything.

Through recommendation vs Converted



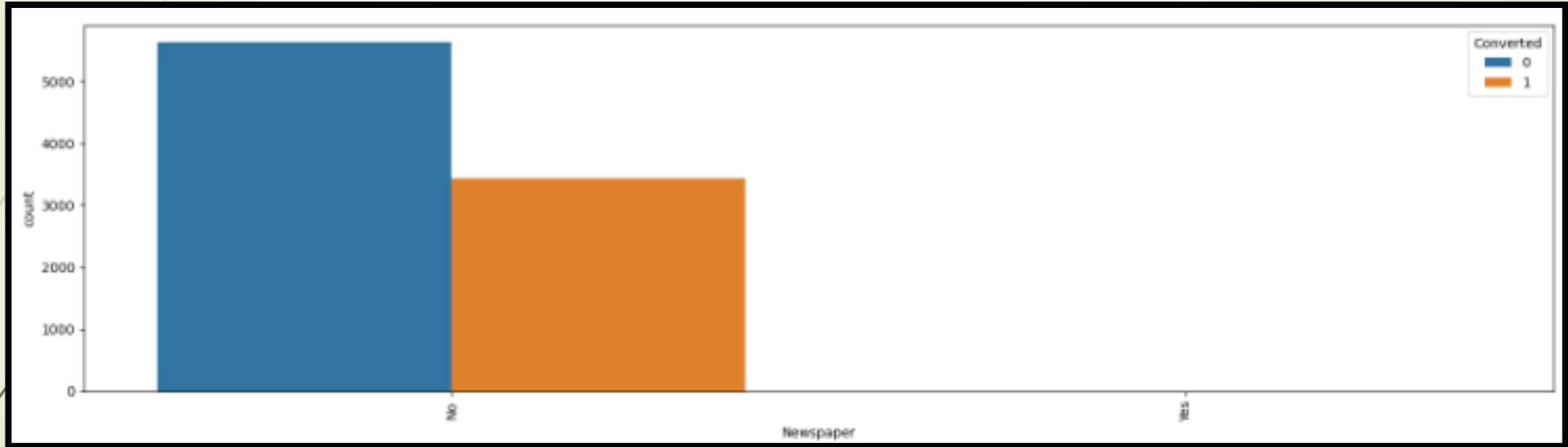
- Through recommendation is not a good option for promising leads.

Digital Advertisement vs Converted



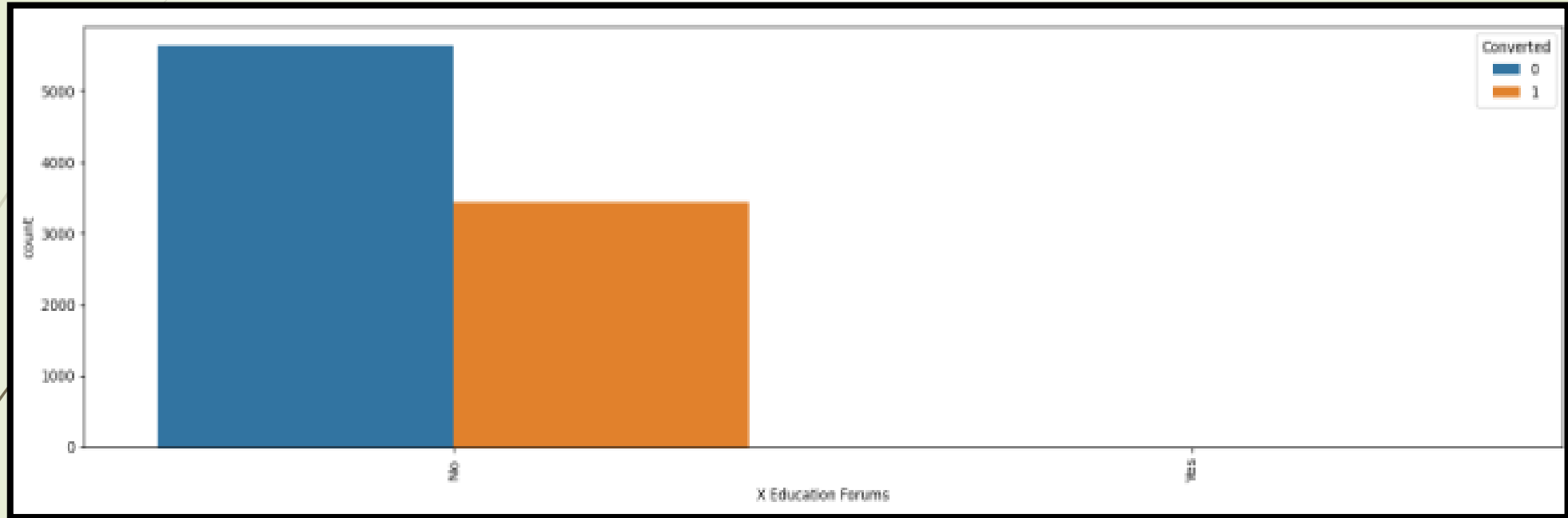
- Digital advertisement is not a good source for lead conversion

Newspaper vs Converted



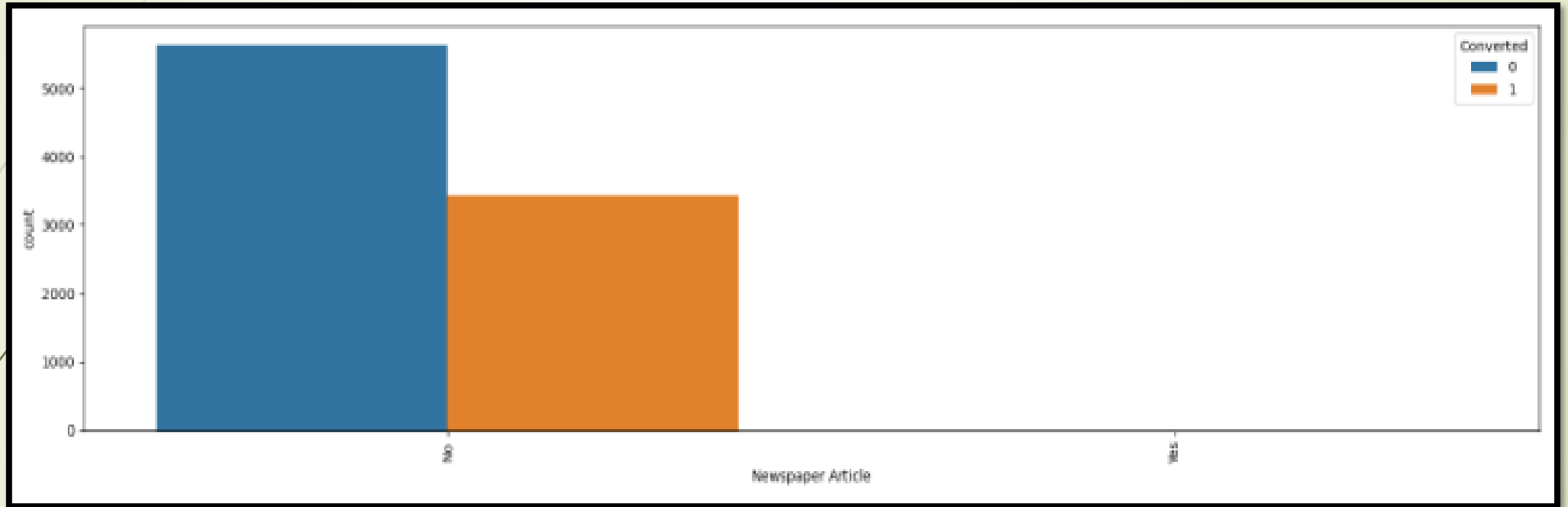
- Company got less leads from newspaper

X Education Forums vs Converted



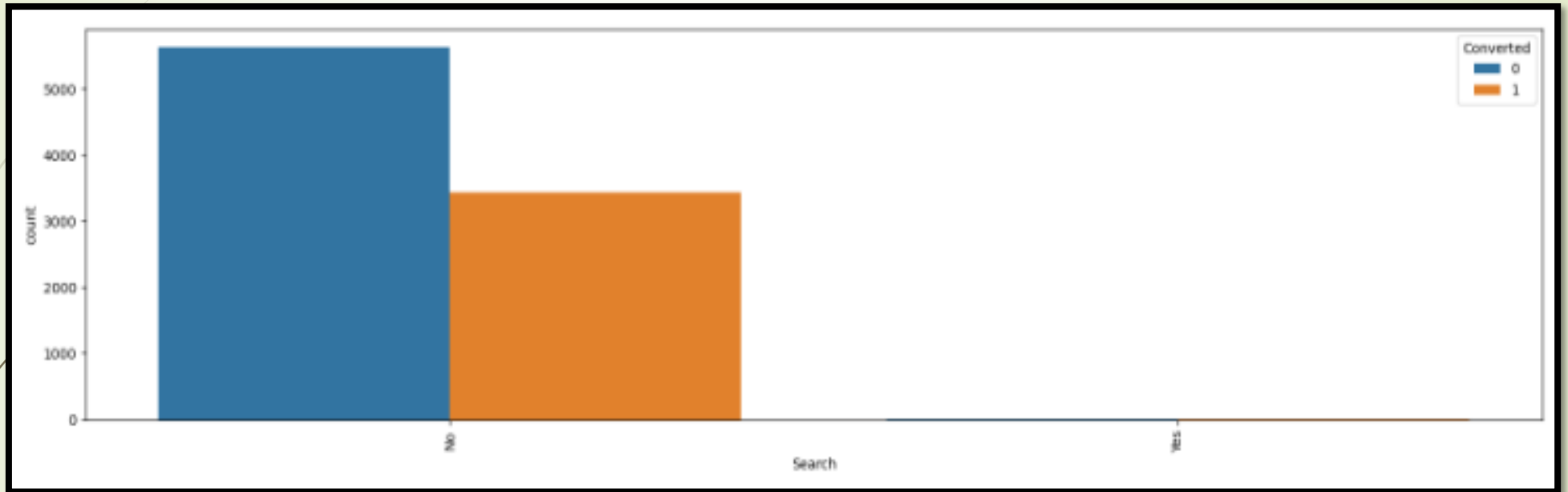
- X Education forums is also not very encouraging for increasing leads.

Newspaper Article vs Converted



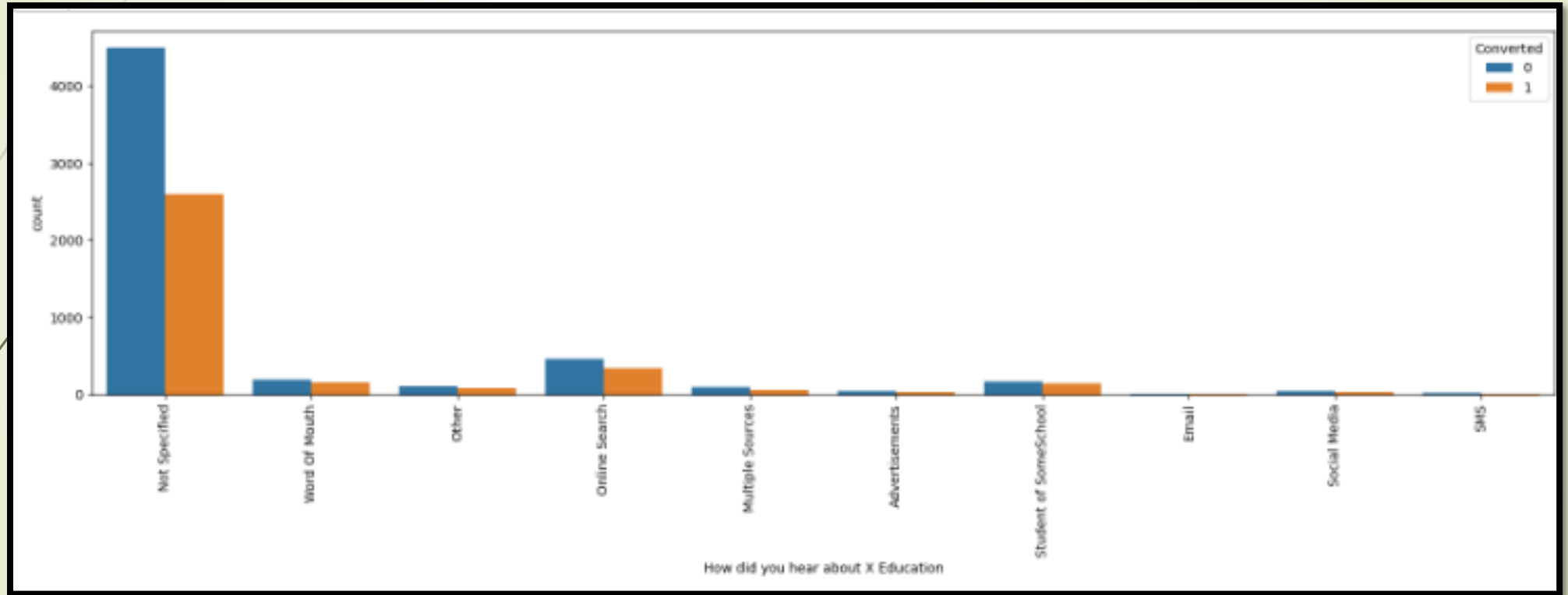
- Newspaper Article is not a good source for promising leads.

Search vs Converted



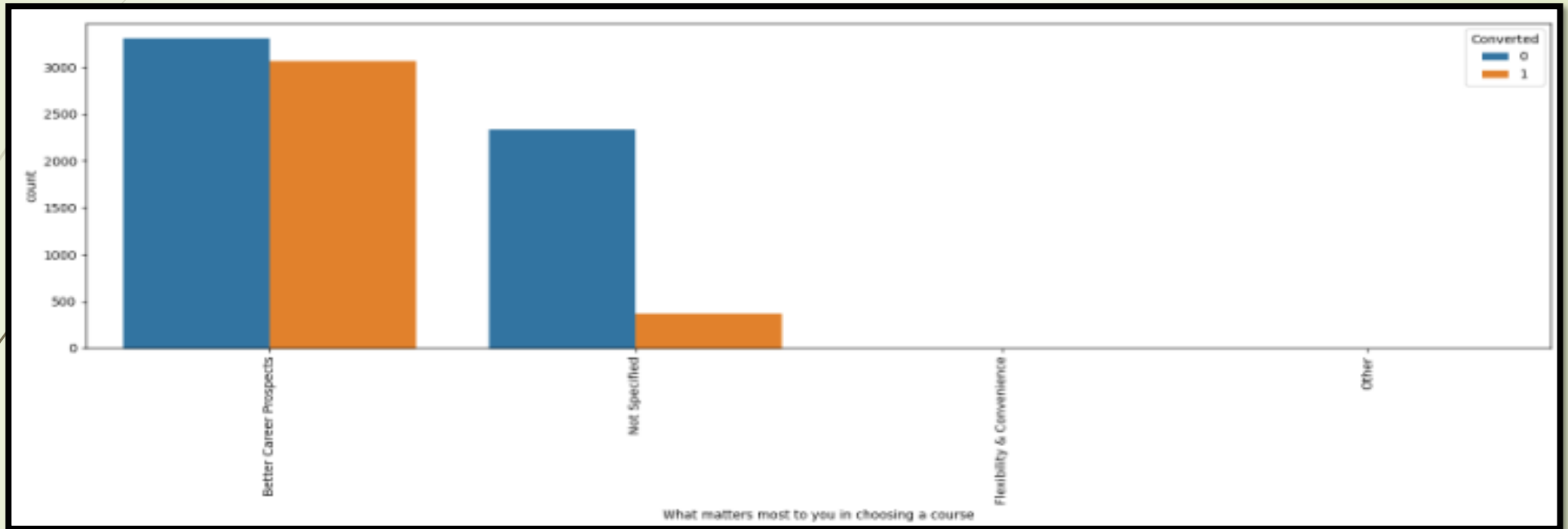
- Search is not a good option for getting leads.

How did you hear about X Education vs Converted



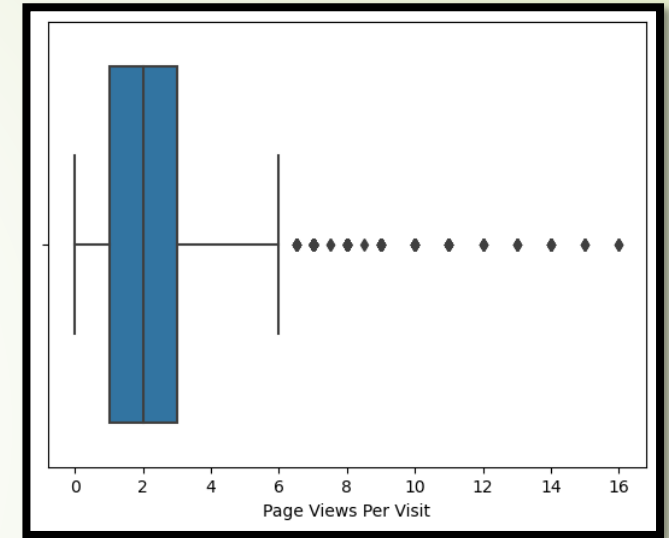
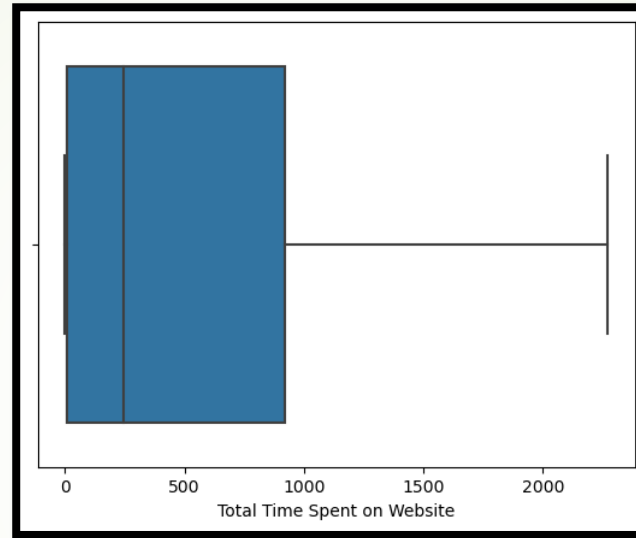
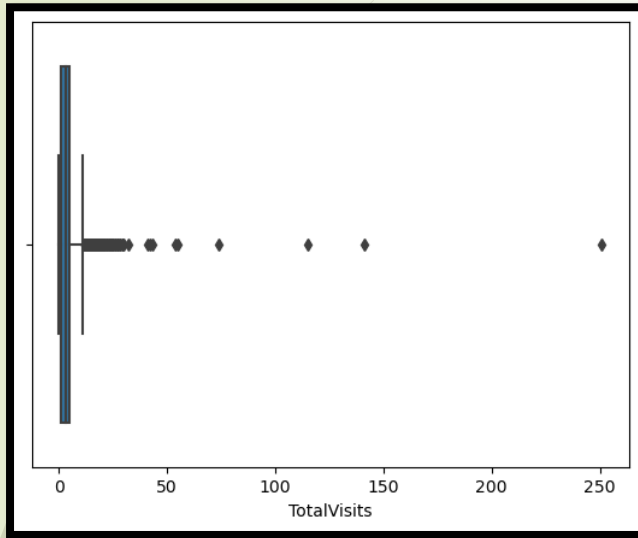
- Most of the customers are not interested to mention about it, few people hear about it from online search, words of mouth etc and became lead.

What matters most to you in choosing a course vs Converted



- Most of leads are looking for better career opportunities

Outlier Treatment



- 'TotalVisits', 'Page Views Per Visit' has outliers.
- 'Total Time Spent on Website' has no outlier

Logistic Regression Model Building

- Split the data into 70% train and 30% test data
- Used Feature Scaling by StandardScaler
- Checked Feature Scaling using RFE with 15 features
- Dropped column whose p value is > 0.05
- Checked VIF and dropped column who are outside the range of 5.
- Make prediction using train dataset
- Created data frame for actual converted and predicted converted and created column for converted probability
- Calculated and plotted Confusion Matrix, Accuracy, Sensitivity and Specificity
- Calculated and plotted ROC curve and find optimal cut off point
- Checked Precision and Recall and find optimal tradeoff between them
- Make predictions using test dataset
- Calculated predicted value and its probability
- Checked Confusion Matrix, Accuracy, Sensitivity and Specificity

Final Model Visualization with RFE and VIF

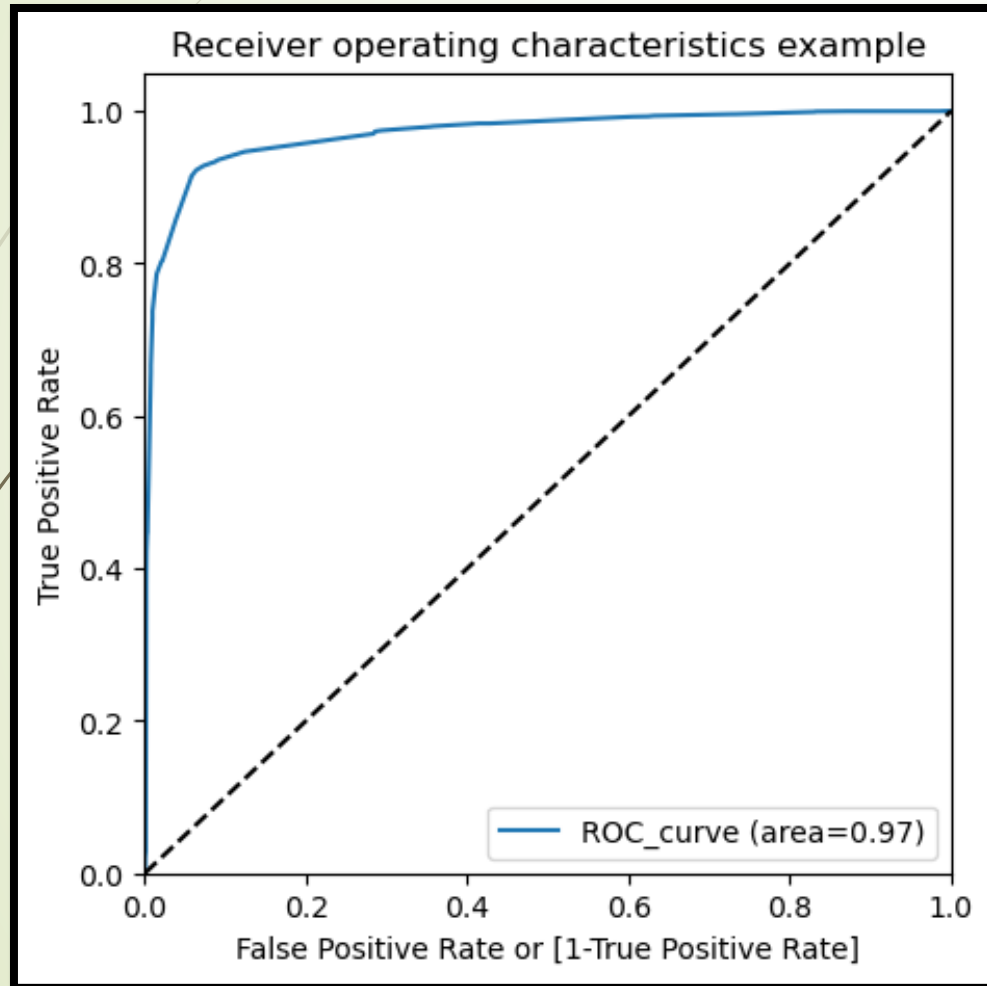
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6246
Model:	GLM	Df Residuals:	6232
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1238.9
Date:	Mon, 22 Jul 2024	Deviance:	2477.8
Time:	23:08:58	Pearson chi2:	1.09e+04
No. Iterations:	8	Pseudo R-squ. (C S):	0.6061
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-3.0601	0.203	-15.046	0.000	-3.459	-2.662
Lead_Source_Welingak Website	2.2764	0.749	3.039	0.002	0.808	3.744
Last Activity_SMS Sent	2.1377	0.120	17.820	0.000	1.903	2.373
What matters most to you in choosing a course_Not Specified	-2.5206	0.145	-17.420	0.000	-2.804	-2.237
Tags_Busy	2.5259	0.286	8.835	0.000	1.966	3.086
Tags_Closed by Horizzon	8.6033	0.744	11.564	0.000	7.145	10.061
Tags_Lost to EINS	8.8985	0.582	15.297	0.000	7.758	10.039
Tags_Not Specified	3.5478	0.229	15.494	0.000	3.099	3.997
Tags_Ringing	-1.7019	0.300	-5.666	0.000	-2.291	-1.113
Tags_Will revert after reading the email	6.3737	0.257	24.780	0.000	5.870	6.878
Tags_switched off	-2.5039	0.744	-3.363	0.001	-3.963	-1.045
Lead Quality_Worst	-2.0031	0.651	-3.079	0.002	-3.278	-0.728
Last Notable Activity_Modified	-1.5277	0.126	-12.156	0.000	-1.774	-1.281
Last Notable Activity_Olark Chat Conversation	-1.5835	0.416	-3.809	0.000	-2.398	-0.769

	Features	VIF
6	Tags_Not Specified	4.98
2	What matters most to you in choosing a course_...	4.73
1	Last Activity_SMS Sent	1.65
11	Last Notable Activity_Modified	1.54
8	Tags_Will revert after reading the email	1.38
7	Tags_Ringing	1.12
0	Lead_Source_Welingak Website	1.11
10	Lead Quality_Worst	1.10
4	Tags_Closed by Horizzon	1.08
5	Tags_Lost to EINS	1.06
12	Last Notable Activity_Olark Chat Conversation	1.05
3	Tags_Busy	1.04
9	Tags_switched off	1.03

Plot using ROC Curve

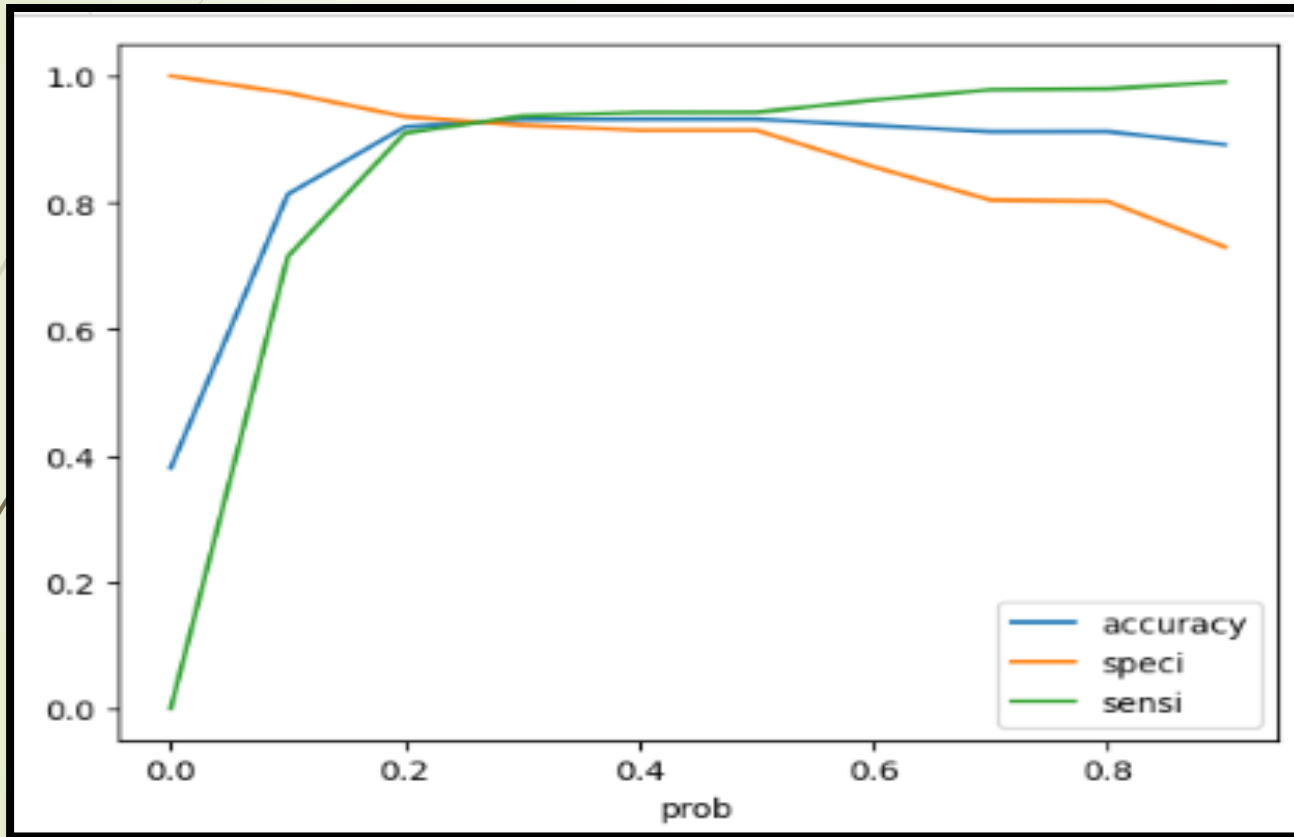


Area under the curve = 0.97

Created this graph for model stability with the area under the curve. Area is leaned towards left side of the border which indicates good accuracy

Model Evaluation – Train Data

Accuracy, Sensitivity and Specificity



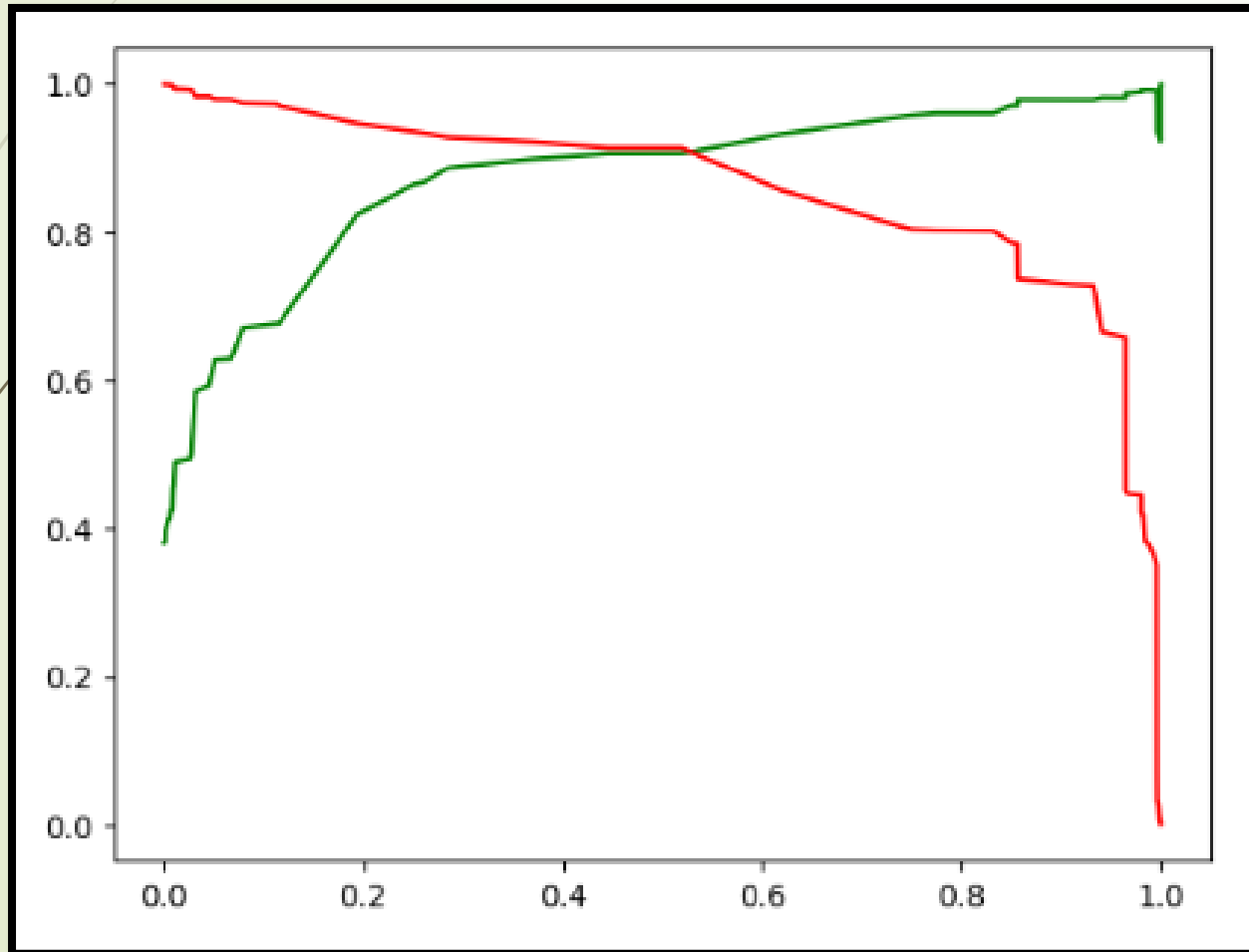
- **Accuracy** – 93.09%
- **Sensitivity** – 92.21%
- **Specificity** – 93.64%

Creating this model for finding optimal cut off probability among accuracy, sensitivity and specificity.

- From the above graph 0.3 is the optimal cut off point

Model Evaluation – Train Data

Precision and Recall



- **Precision** – 89.90%

- **Recall** – 92.21%

We have created this graph to showcase the trade off between Precision and Recall.



Model Evaluation – Test Data

Accuracy, Sensitivity and Specificity

- **Accuracy** – 93.39%
- **Sensitivity** – 92.65%
- **Specificity** – 93.82%




Conclusion based on EDA

- People who are investing more time on their website and engaging themselves in courses they are becoming leads.
- Mostly from Mumbai, India people are more career centric and enrolling courses for better career opportunities.
- Unemployed and working professionals like management professional people have very high conversion rate.
- People who revert after reading the email they become promising leads.
- Company getting most number of leads via organic search, direct traffic, google, reference and Welingak website.
- API, Landing Page Submission and Lead add form has very high conversion rate.
- People who received SMS they are good source for promising leads.



Conclusion based on Logistic Regression Model

- Built a logistic regression model
 - Calculated ROC, confusion matrix, accuracy, sensitivity and specificity based on train and test data. Also checked Precision and Recall and find optimal point and got Recall value which is higher than Precision value, which indicated model is accurate.
 - Accuracy, Sensitivity and Specificity of train data are very close to test data.
 - This model specifies number of people who can become leads and who are not interested for conversion.
 - This model has the ability to change as per future company requirement.
- 



THANK YOU