



# Lead scoring Case study

Anshu Joshi  
Hammad Moin Siddiqui  
IIIT, B PGDSc Batch 026  
17<sup>th</sup> May, 2021

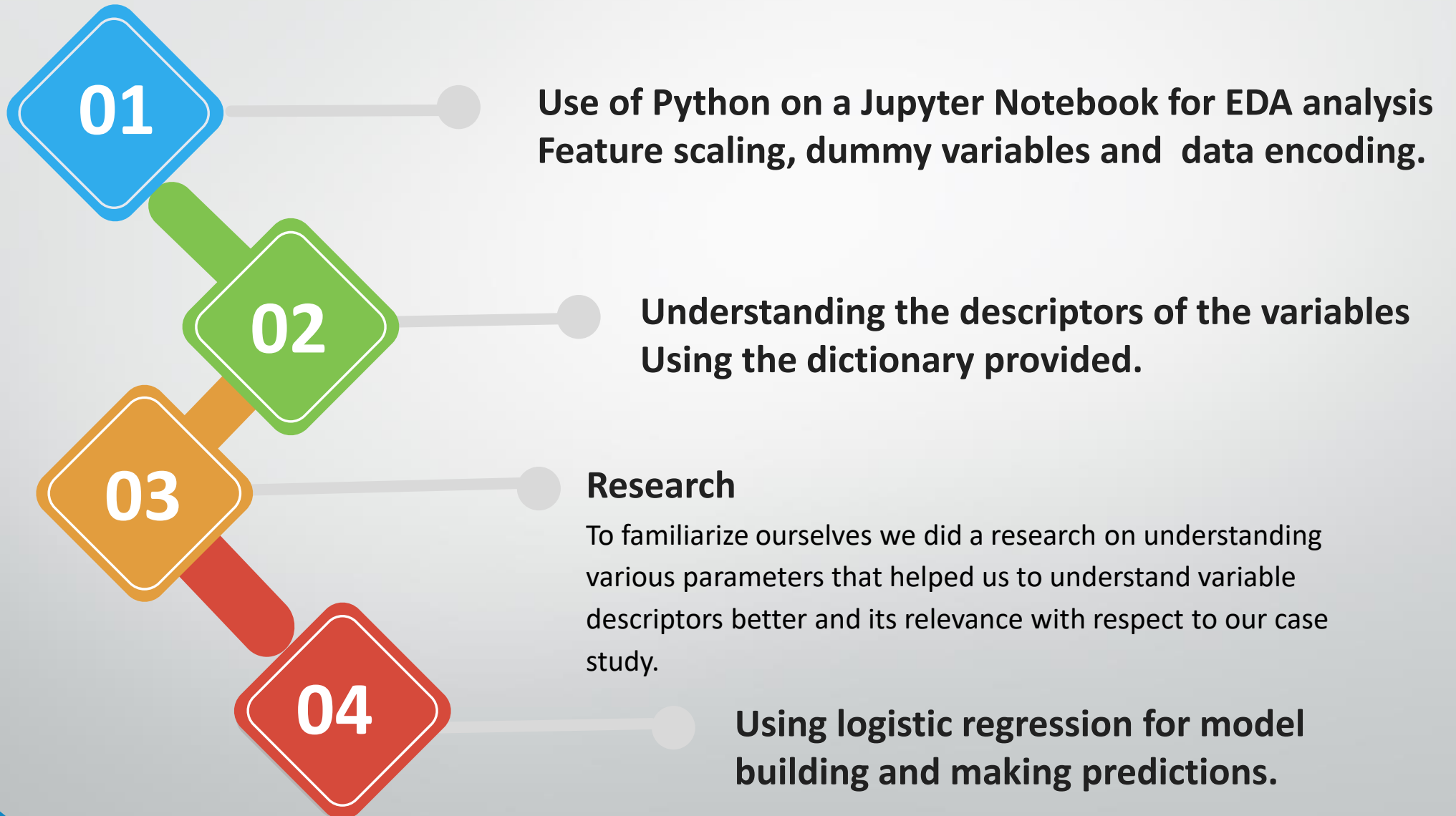
# Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites, the company also gets leads through past referrals.
- Leads are acquired through this process, 30% of the leads get converted while most do not.
- 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 .
- In the middle stage, you need to nurture the potential leads well in order to get a higher lead conversion.

# Objective

- X Education has appointed you to help them , The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.
- 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

# How did you go about studying the Lead scoring Case Study



# Data Cleaning

- Find the percentage of missing values of all the columns
- Remove columns with high missing percentage
- replacing the unique data with relevant category
- Checking and removing outliers
- Values that are not mentioned are replaced with other relevant values.

## Steps taken:

- We converted select as NaN as they are as good as null value.
- Dropped 'Asymmetrique Activity Index','Asymmetrique Profile Index', 'Asymmetrique Activity Score','Asymmetrique Profile Score','Leads Quaity','Lead Profile', 'How did you hear about X Education' as they have almost 50% or more missing values.
- Replacing the unique data with relevant category.
- Removed outliers.
- Removed null values as well as the city Mumbai.

# EDA

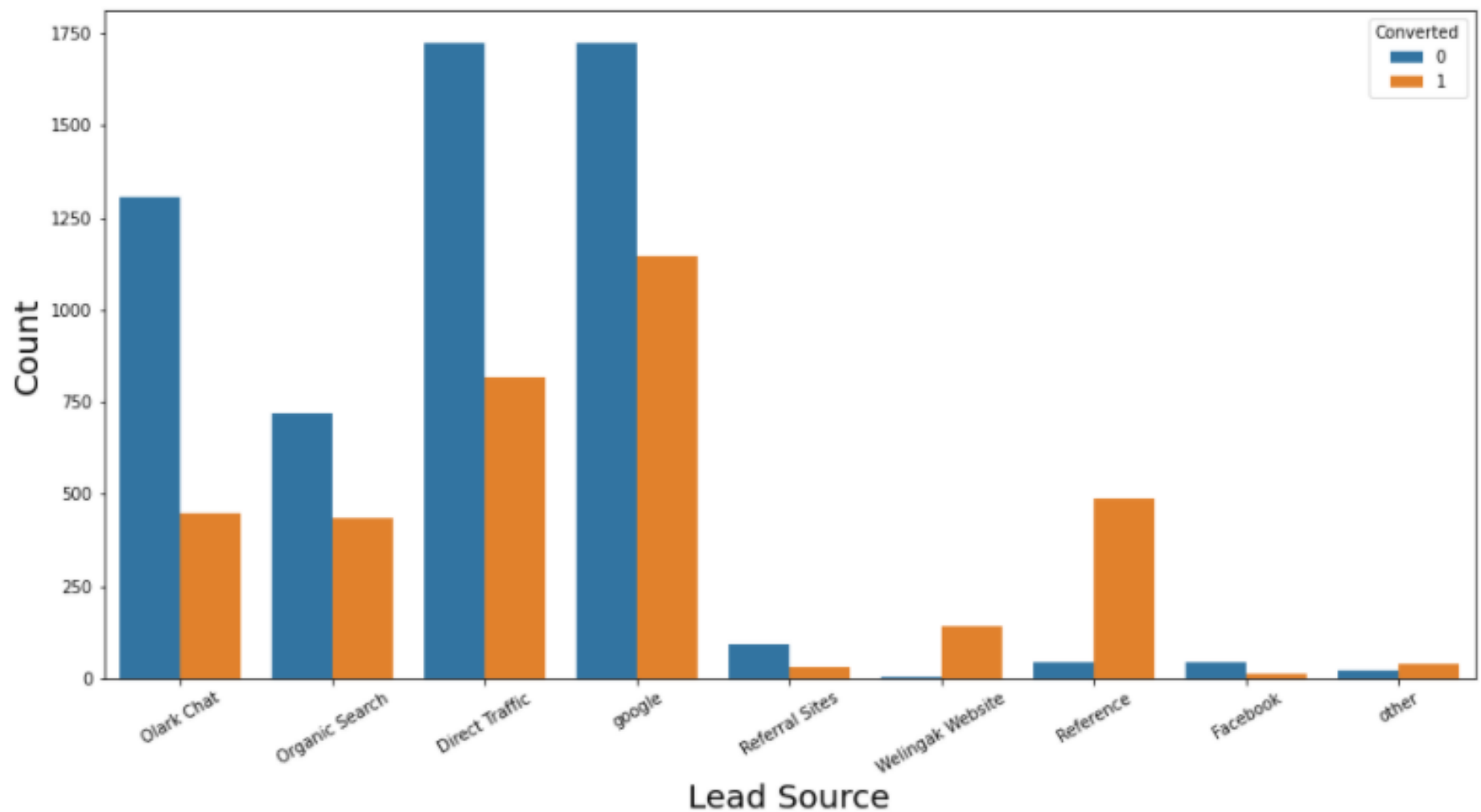
## Techniques

- Univariate Analysis
- Bivariate Analysis
- Categorical variable relation.



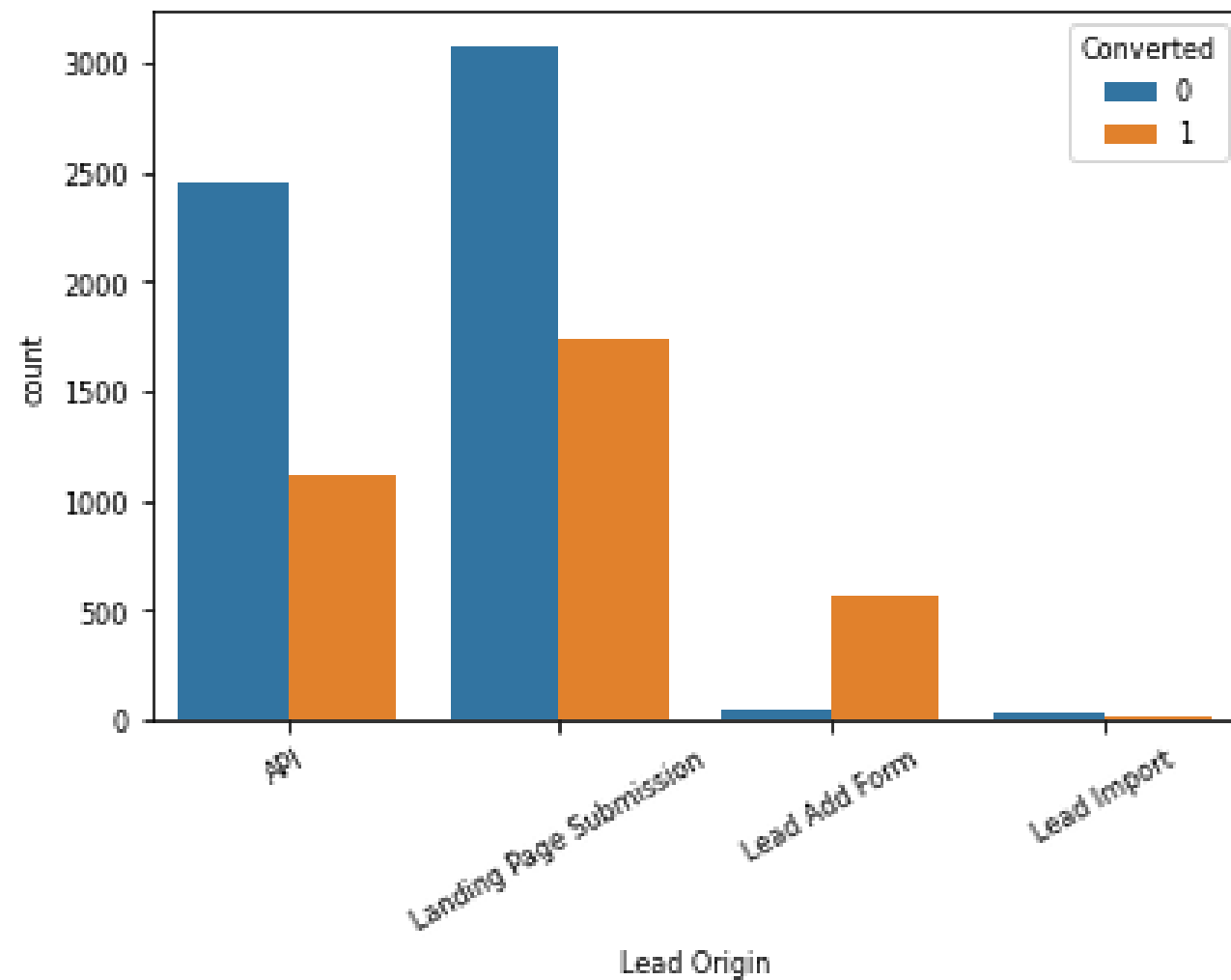
# Visualization and inferences



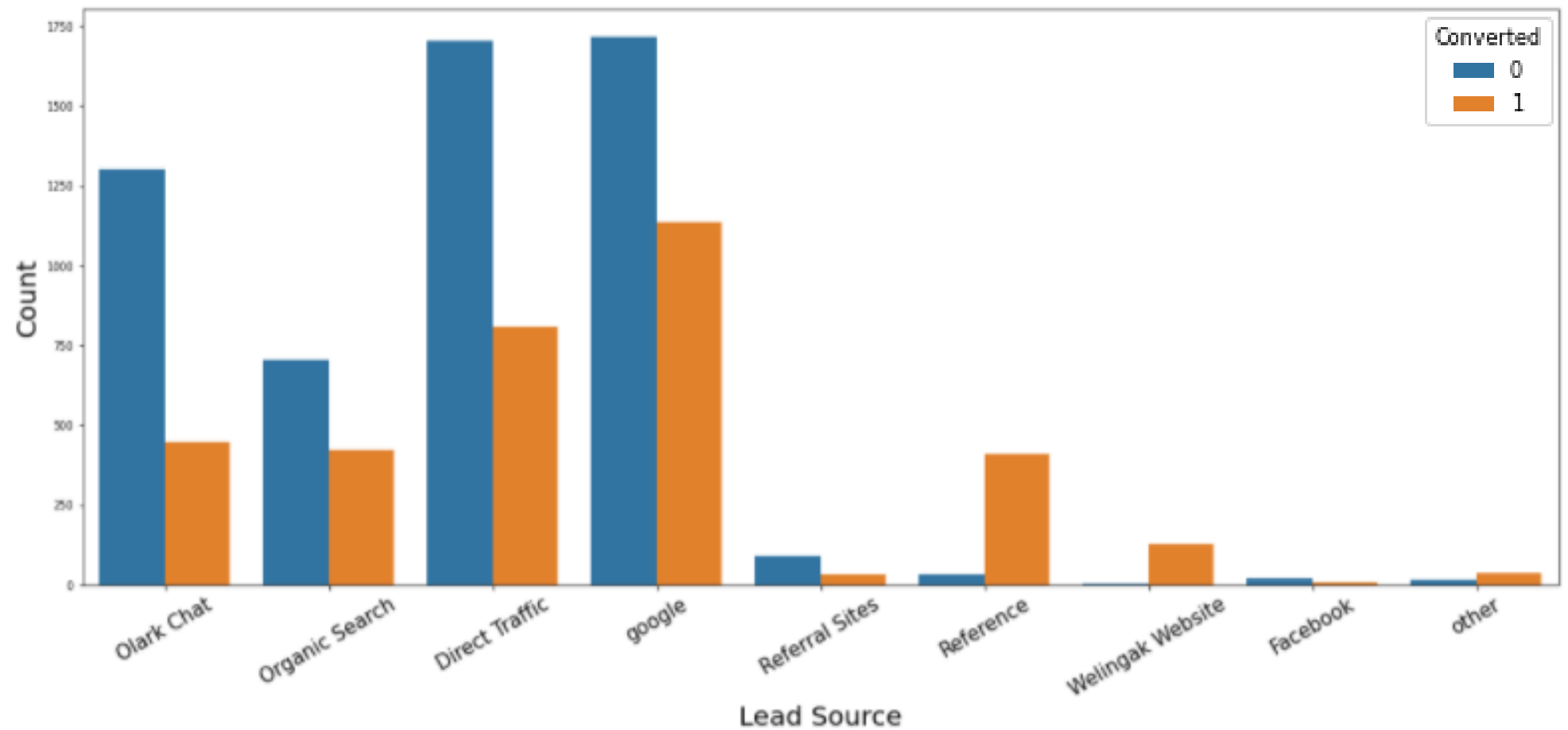


- Google & Direct traffic generates maximum number of the leads.
- Conversion rate of the welingak website and reference leads is high

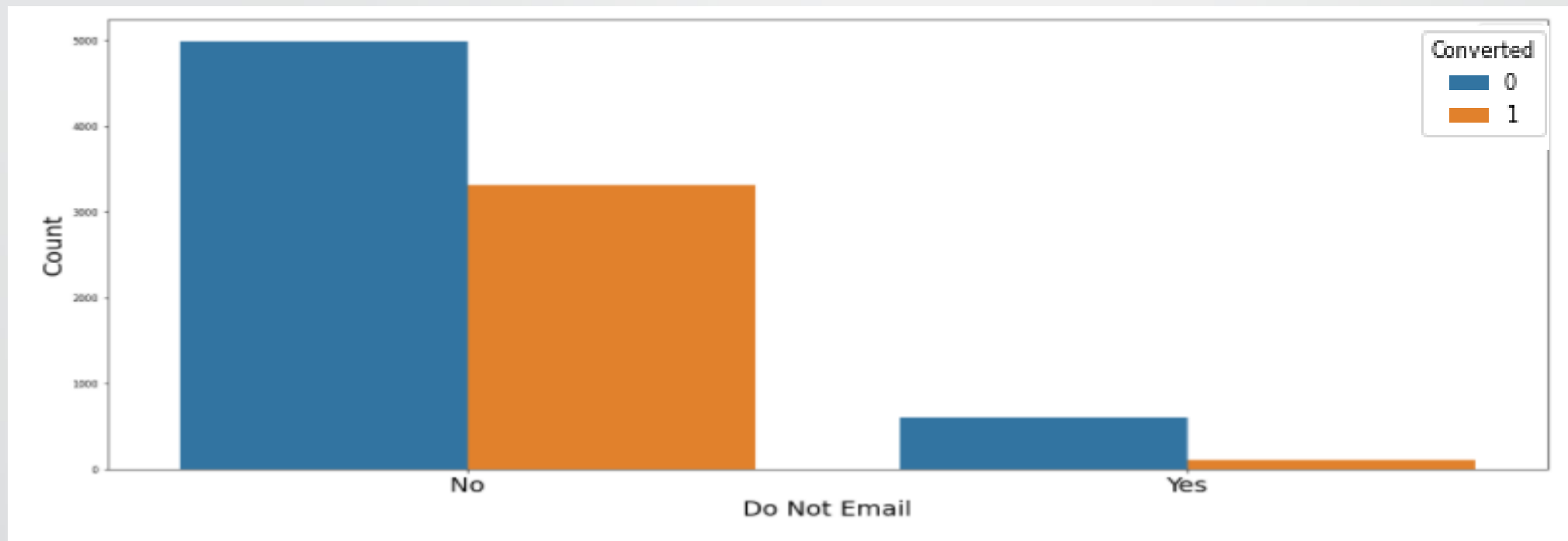
To improve the overall lead conversion rate, we should focus on the Organic Search, Olark Chat, Direct Traffic and google leads in the Lead Source and generates more leads.



- API and Landing page show a good conversion rate and are good enough in number.
- While Lead Add form has high conversion rate but do not have enough counts.

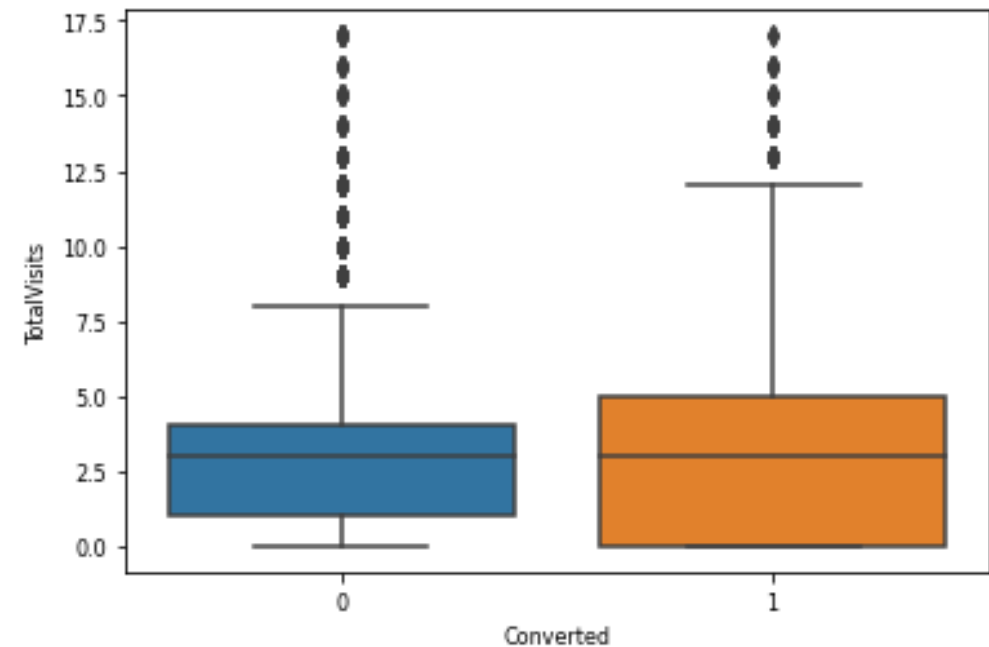


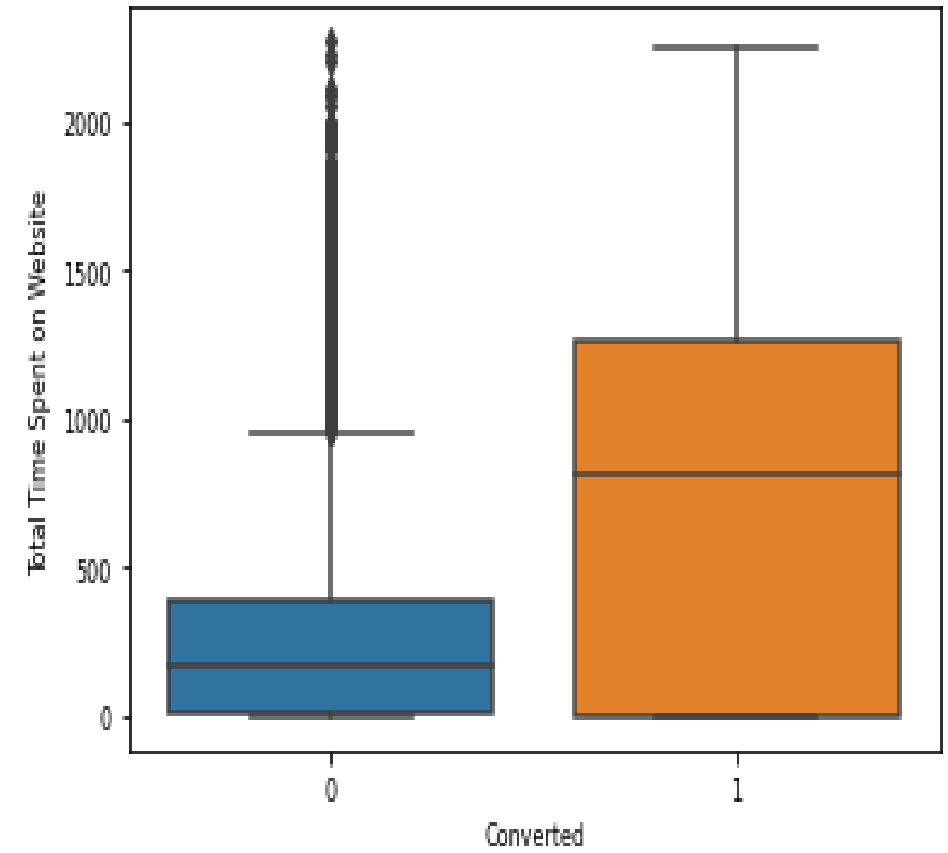
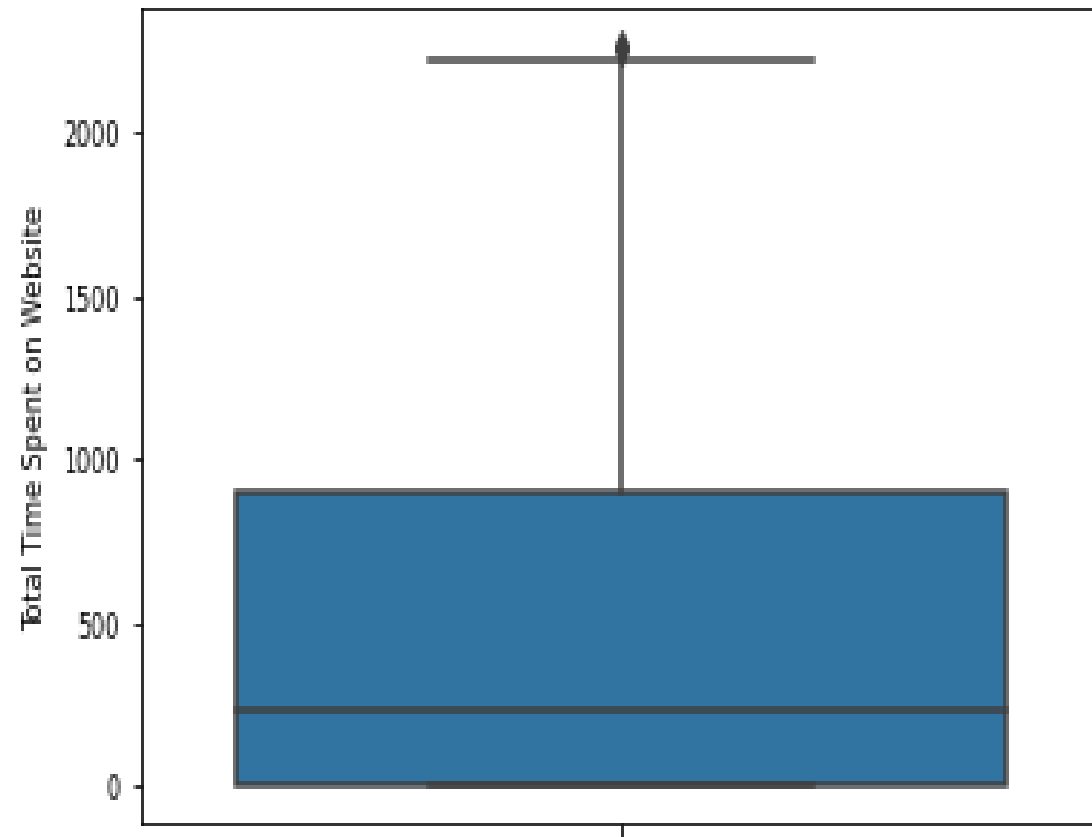
- Most of the leads are generated by Google and Direct Traffic which also has good conversion rate.
- Leads by reference is mostly converted and same goes with Welingak Website.



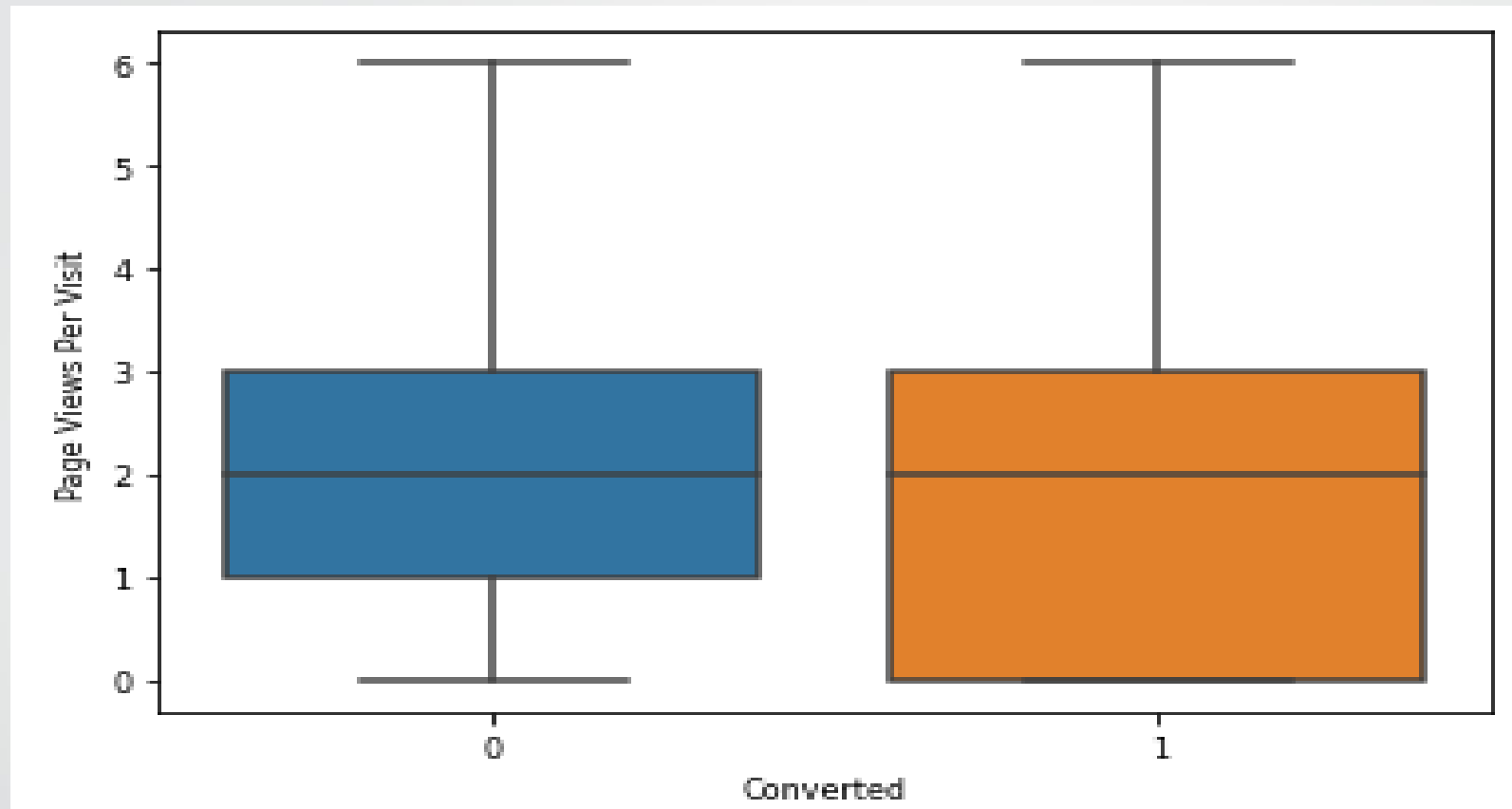
Most customers don't want to get emailed about the course but still get converted more as compared to people who choose to get emailed.

Both converted and not converted has almost same median from total visits.



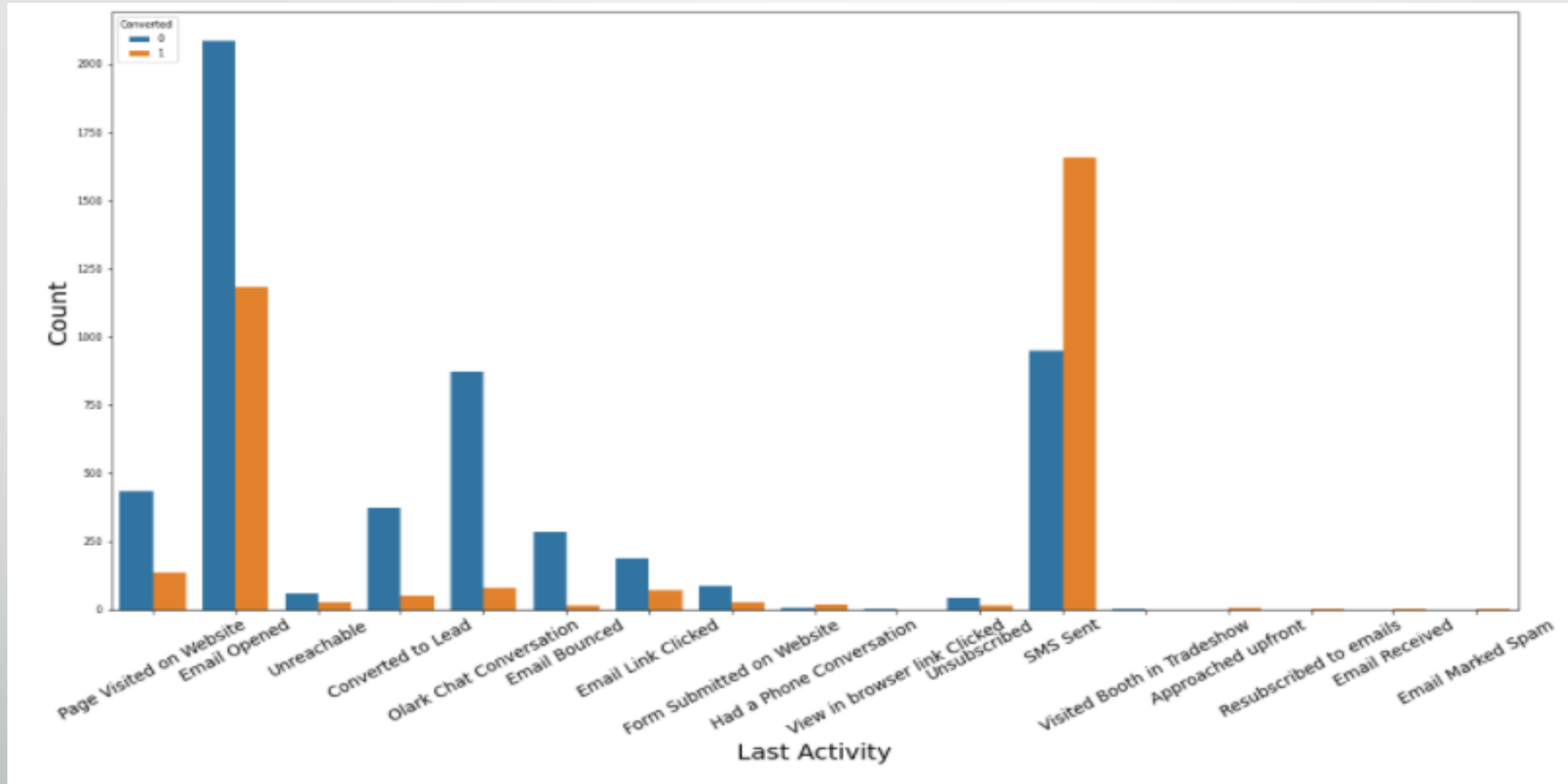


- The average time spent on website is around 250.
- We can see that when the total time spent by a person on website is around 1000 , the lead shows a positive conversion sign.

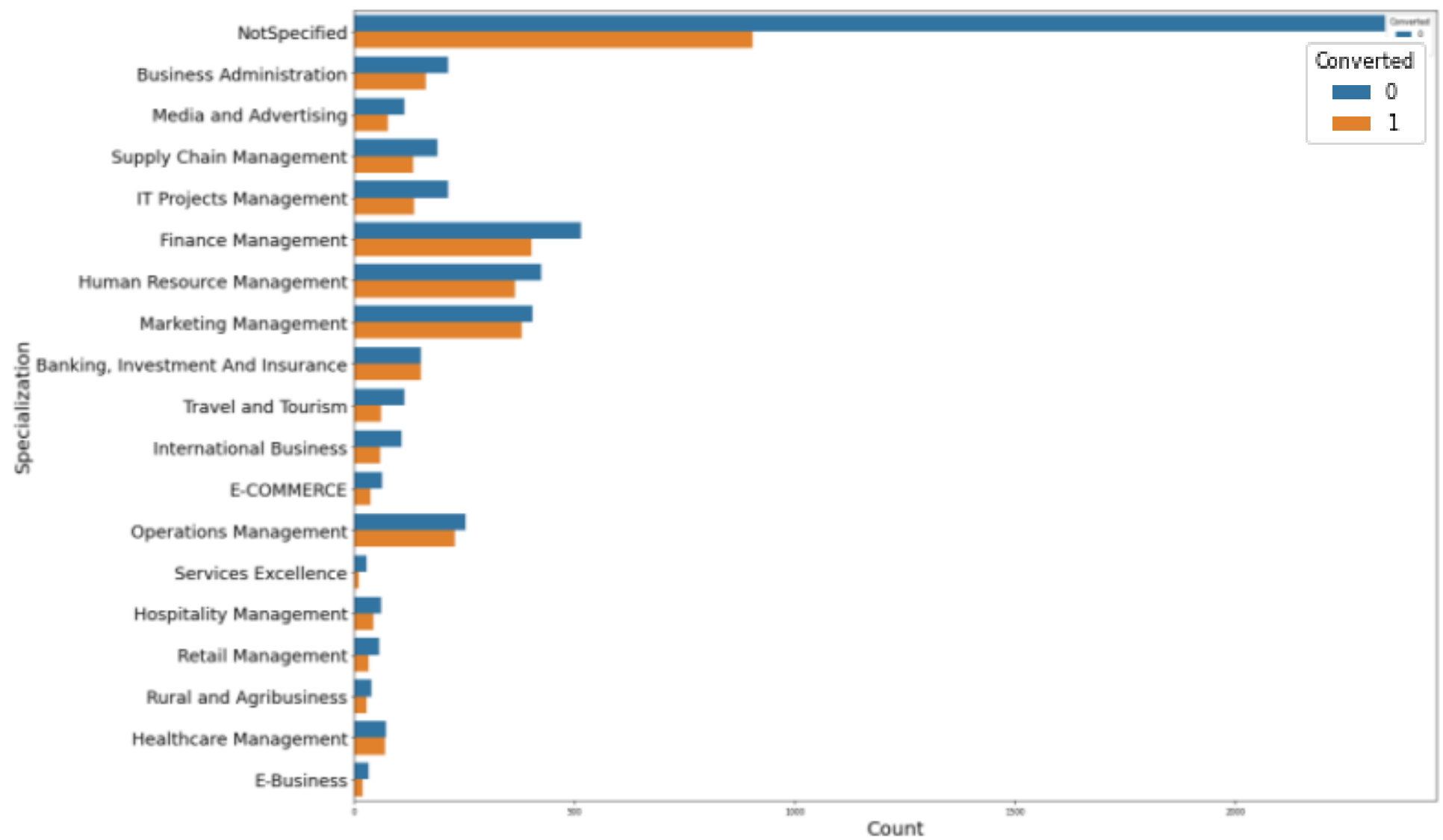


There is no difference in median of the page views for converted and not converted leads.

# Categorical Variable relation

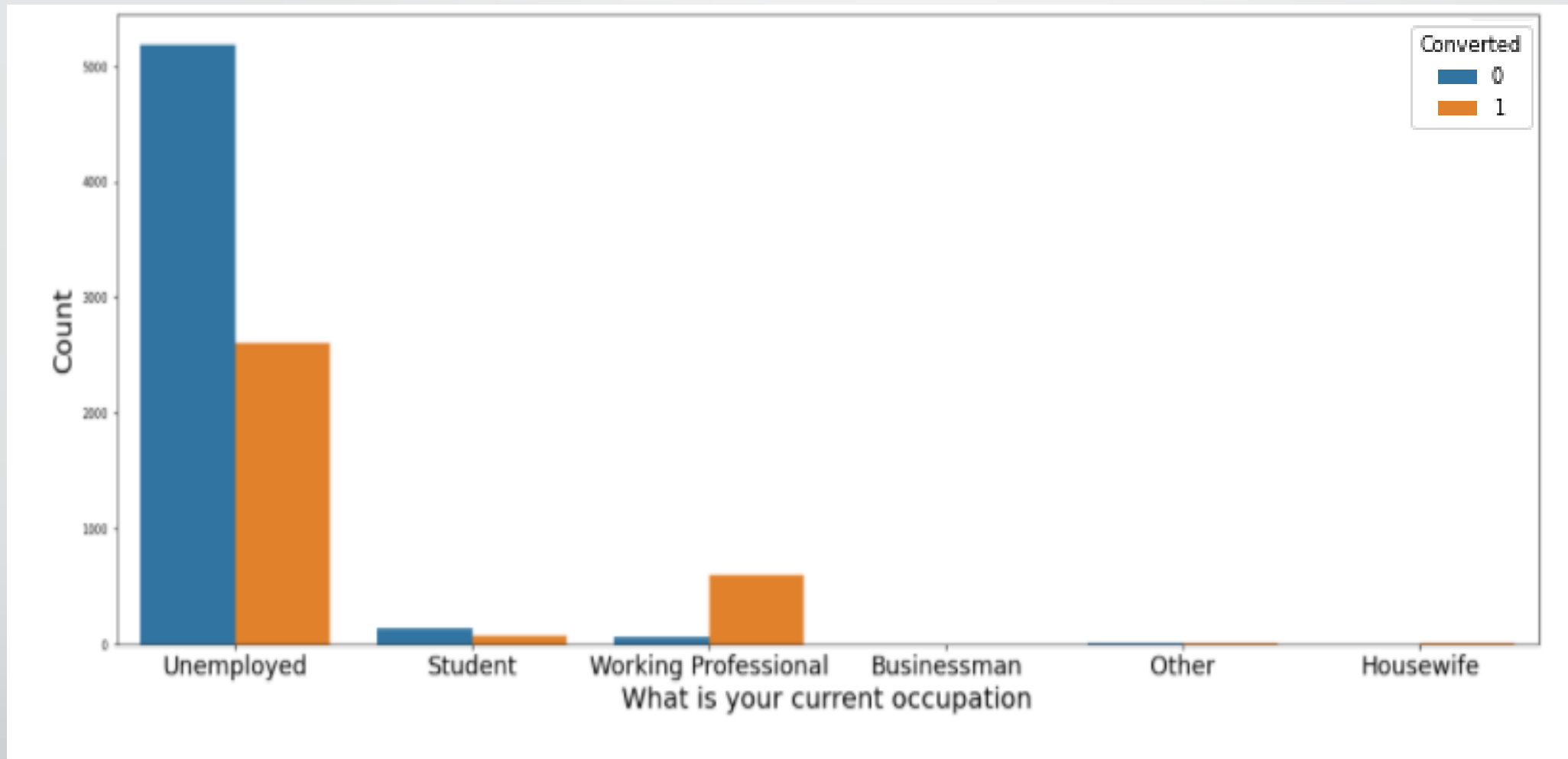


- Most of the leads are from email opened.
- SmS sent has the highest conversion rate among all of them.

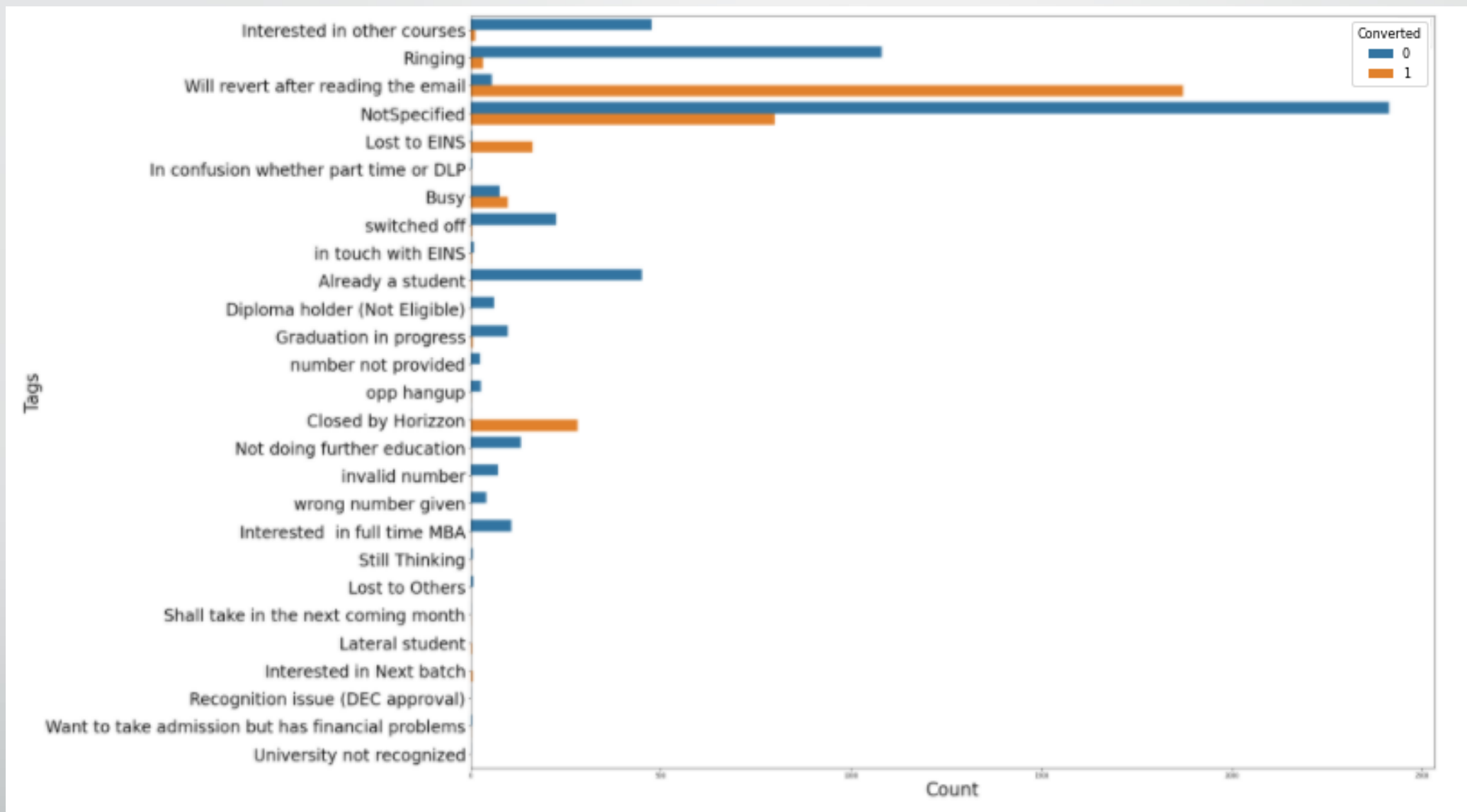


As we can see that those who have not specified their industry domain in which they worked before have the highest conversion rate followed by Finance management ,HR management and marketing management.





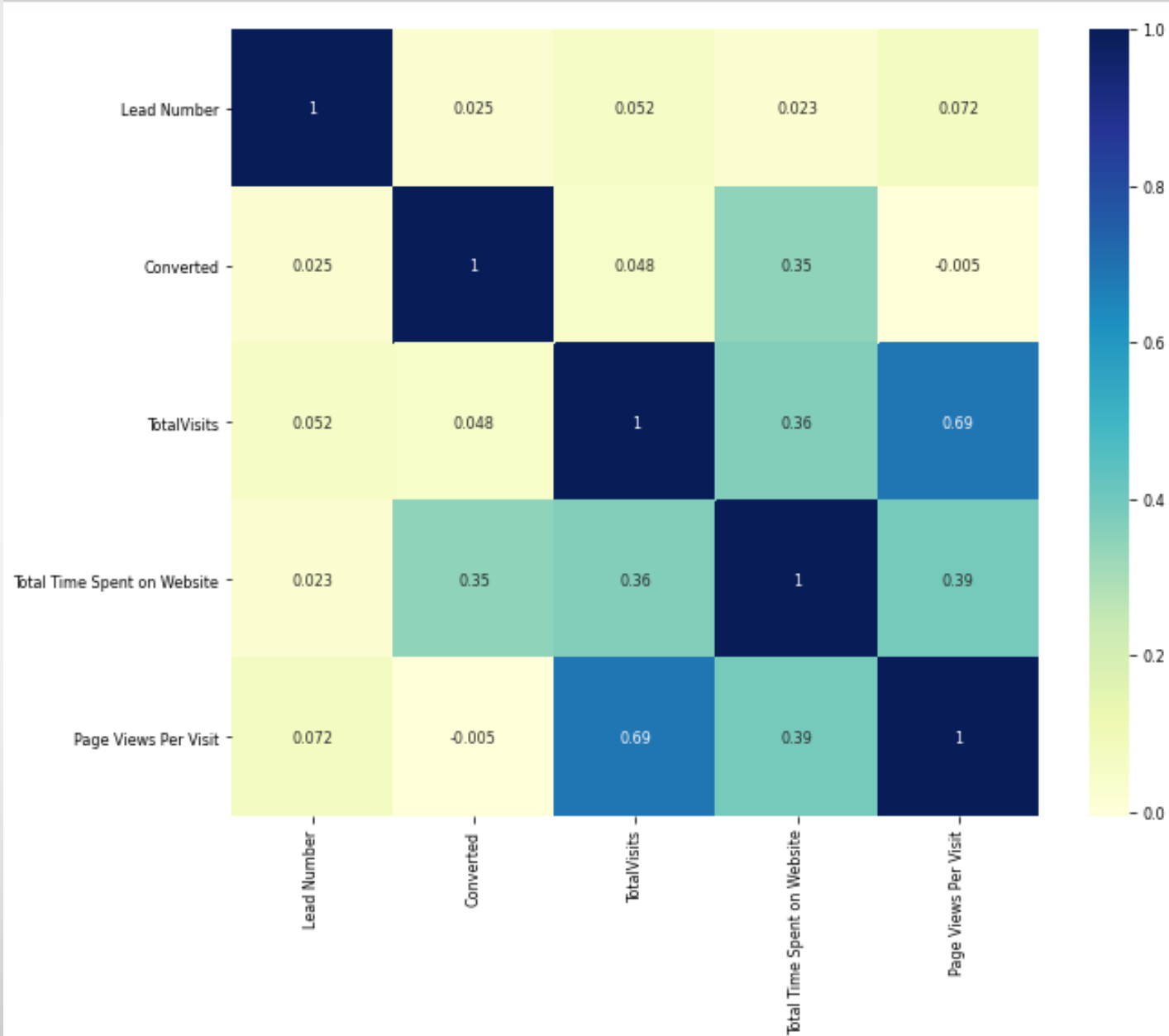
- Working Professionals have high conversion rate as compared to others.
- Most of the leads are generated by Unemployed.

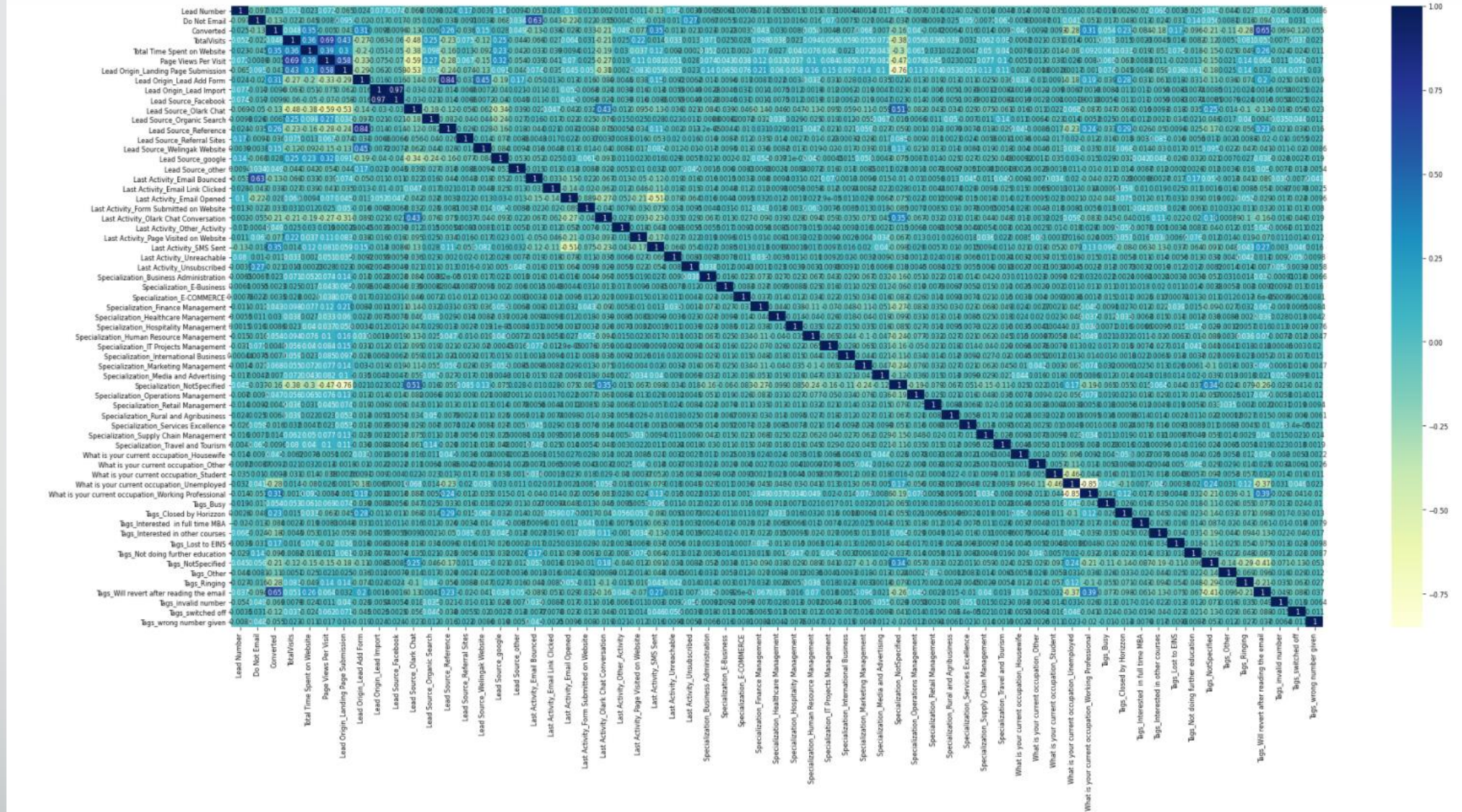


Customers who said they will revert after reading the email have maximum conversions , that means the content in emails have a positive impact on customers.

## Heatmap after dropping the unnecessaries.

Here we can clearly see the high correlation between Total visits and Page Views Per Visit.





# Heatmap after Encoding ,Scaling and creating dummy variables.

Since this heatmap is very crowded and the correlations are not perfectly visible so we revert to RFE.



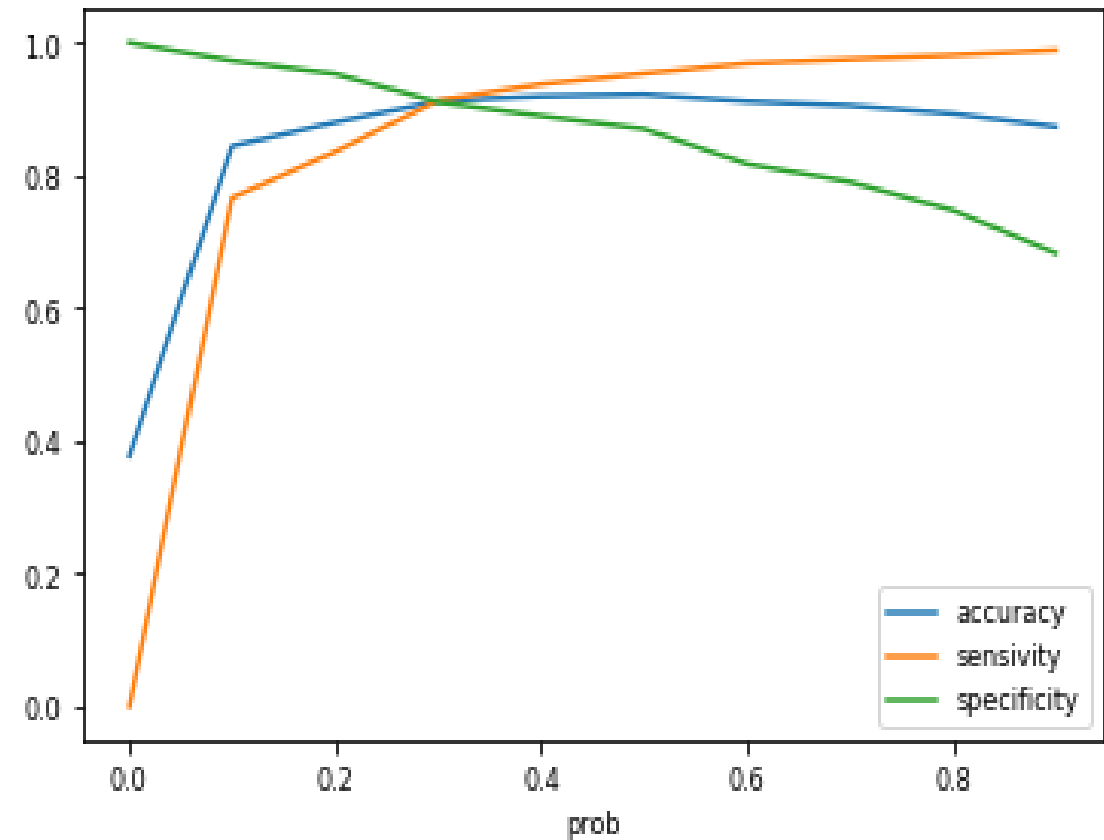
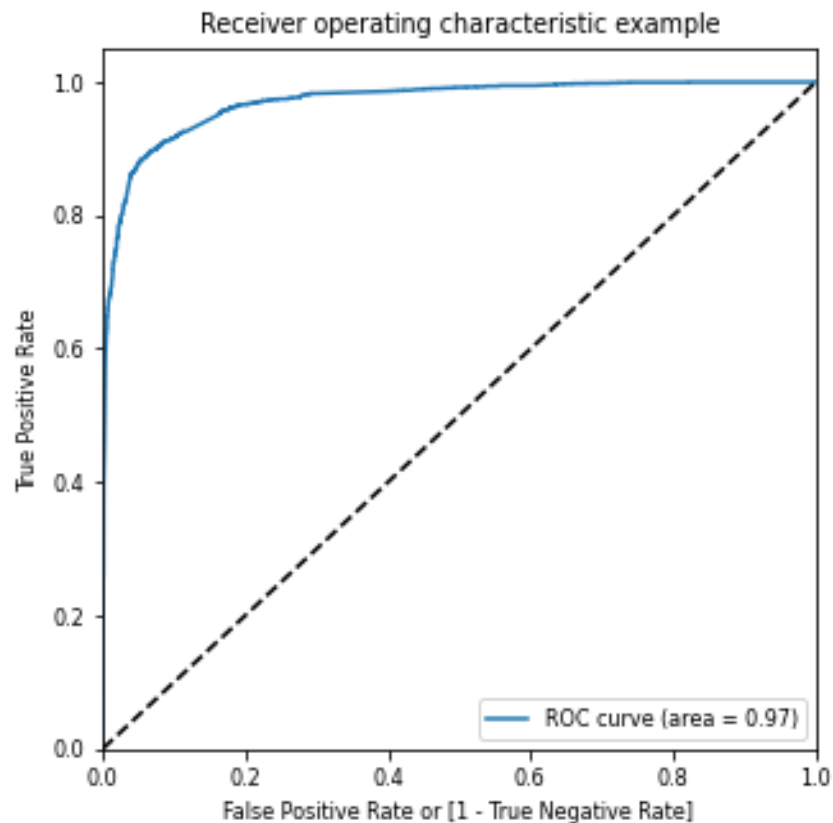


# Building our model

## Steps performed:

- Splitting the data into training and test sets.
- While performing test – train split , we choose 70 : 30 ratio.
- Using RFE for feature selection.
- Running RFE with 15 variables as output.
- Building model by removing the variables whose p-value is greater than 0.05 and VIF value is greater than 5.
- Predictions on test dataset.
- Creation of confusion matrix.
- Calculation of accuracy, sensitivity , specificity, precision and Recall.

# ROC Curve



- The area under ROC curve is 0.97 , which is very good.
- The optimal cutoff probability from second graph is 0.3

# Confusion matrix

	Predicted <b>0</b>	Predicted <b>1</b>
Actual <b>0</b>	TN	FP
Actual <b>1</b>	FN	TP

- **“true positive”** for correctly predicted event values.
- **“false positive”** for incorrectly predicted event values.
- **“true negative”** for correctly predicted no-event values.
- **“false negative”** for incorrectly predicted no-event values.

1332	271
47	953

The Precision and Recall comes as 0.91 and 0.86 respectively.



# Model Evaluation

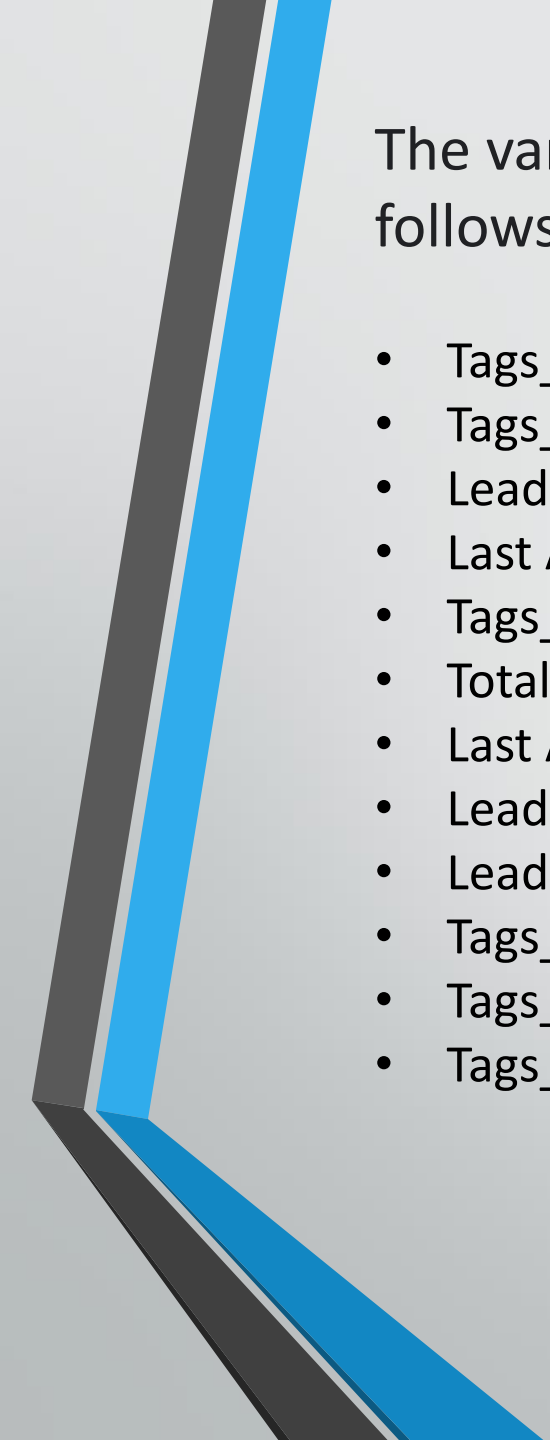
## Comparison of the values of test and train data.

### Train data:

- Accuracy : 87.99%
- Sensitivity : 95.28%
- Specificity : 83.57%

### Test data:

- Accuracy : 87.78%
- Sensitivity : 95.30%
- Specificity : 83.09%



The variables in our dataset which matters the most for good business are as follows:

- Tags\_Will revert after reading the email
- Tags\_NotSpecified
- Lead Origin\_Lead Add Form
- Last Activity\_SMS Sent
- Tags\_Ringing
- Total Time Spent on Website
- Last Activity\_Email Opened
- Lead Source\_Welingak Website
- Lead Source\_Olark Chat
- Tags\_Closed by Horizzon
- Tags\_Busy
- Tags\_Lost to EINS

# Conclusion

We found the following from our analysis:

- Google and Direct Traffic generate most leads and also has good conversion rate.
- Focusing on Reference and Welingak Website can get good leads and conversion.
- Focusing on emails to the customer shows a very positive lead conversion.
- SmS sent has the highest conversion rate among all of them
- Targeting people who don't specify their industry domain can get the highest conversion rate followed by Finance management ,HR management and marketing management.
- Top features for good conversion rate:
  1. Tags\_Closed by Horizon,
  2. Tags\_Lost to EINS,
  3. Tags\_Will revert after reading the email
- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion:
  - 1.Tags,
  2. Lead Source,
  3. Lead Origin



# Thank you

Happy to address should you need any clarification

