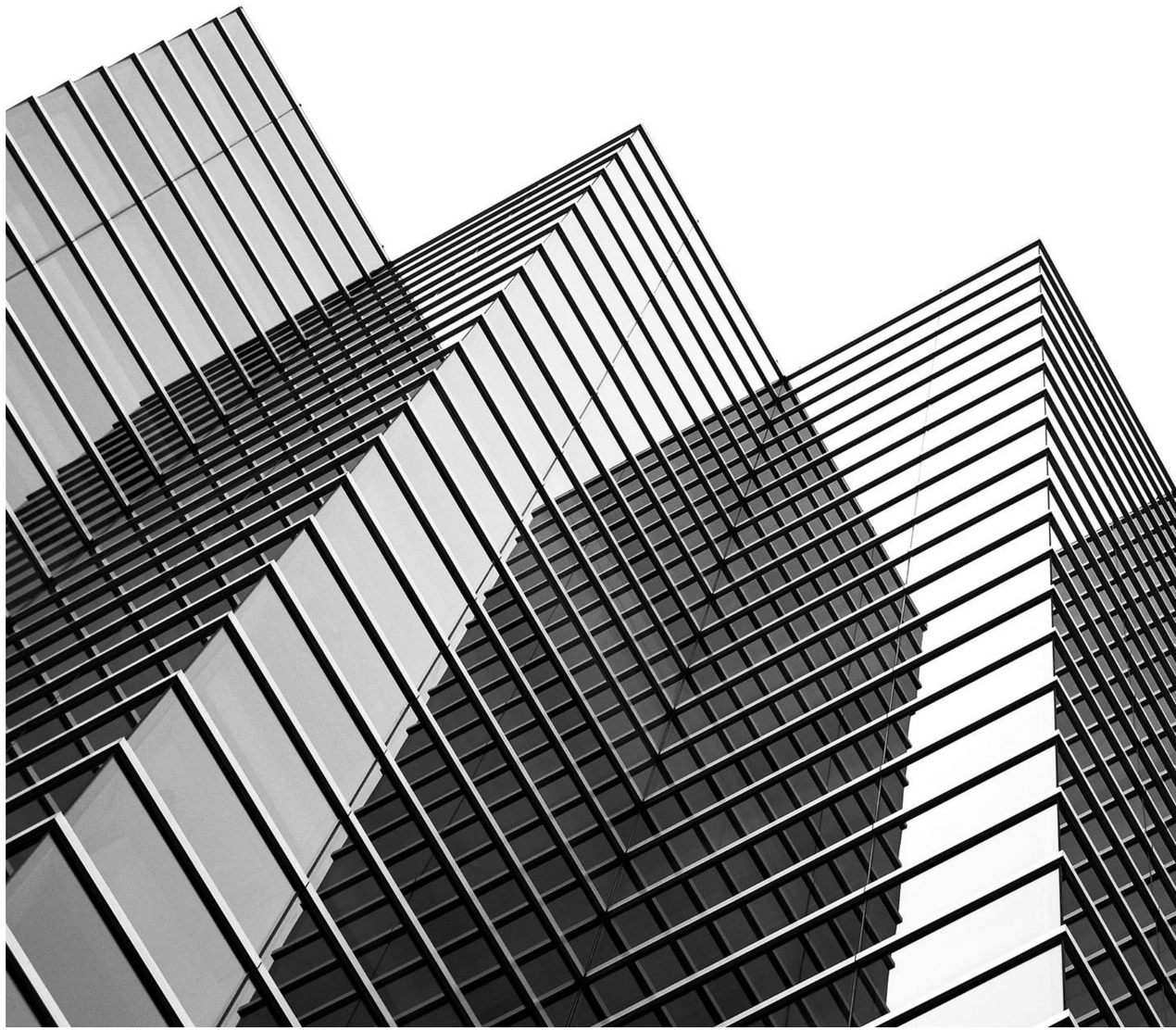
Lead scoring Case study

Data Science

Harsh Pandya
IIITB, Nov 2022 Batch DS40
20/May/2023

Problem Statement

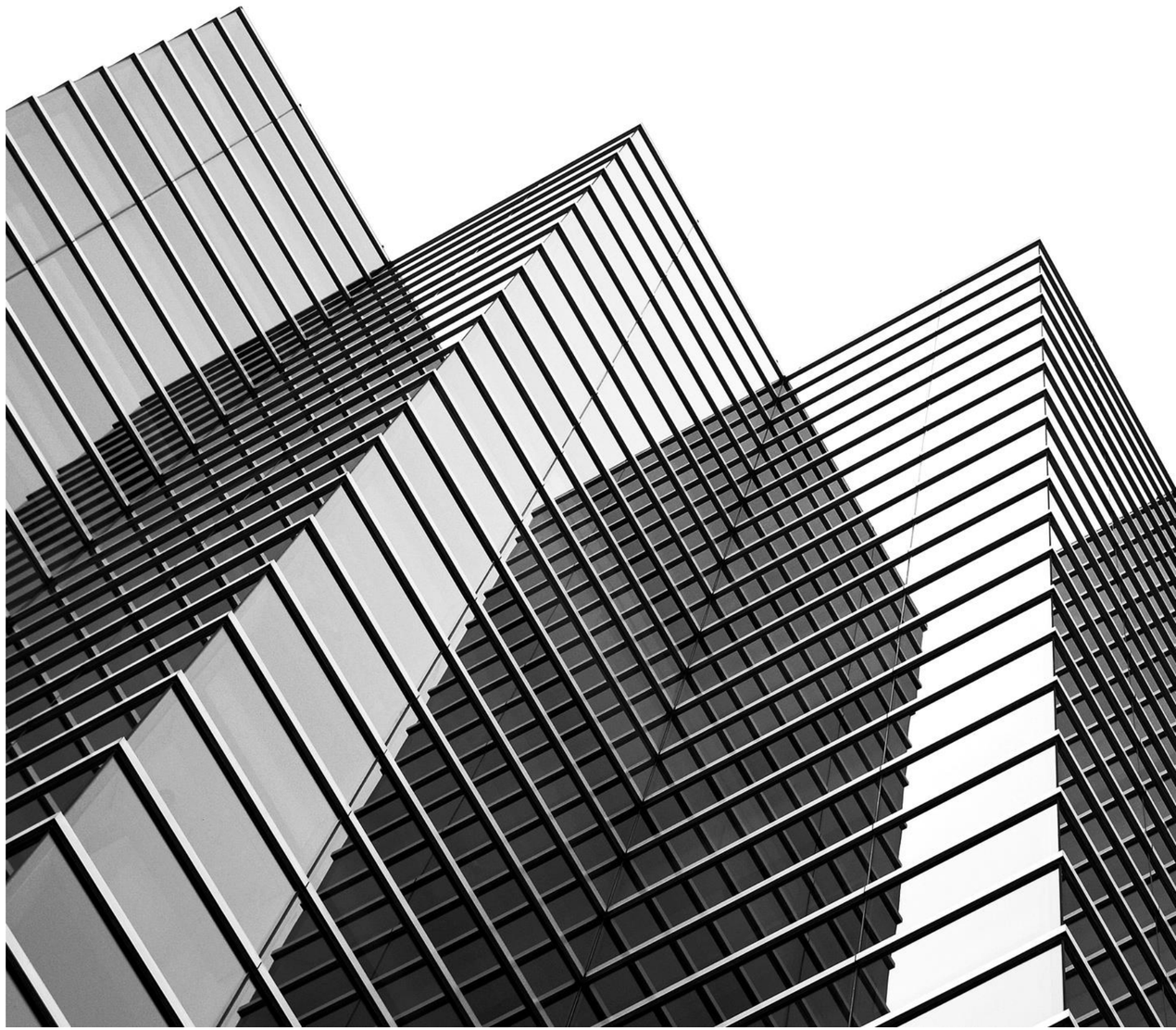
- X Education, an educational company, specializes in selling online courses to professionals in various industries. The company promotes its courses on multiple websites and also receives leads through referrals from previous customers. Through this process, the company acquires leads, but only 30% of them end up being converted into customers, with the majority not converting. In order to increase the conversion rate, the company aims to identify the most promising leads, referred to as "Hot Leads." By successfully identifying this set of leads, the company expects the conversion rate to improve. During the intermediate stage, it is crucial to effectively nurture the potential leads to achieve a higher conversion rate.



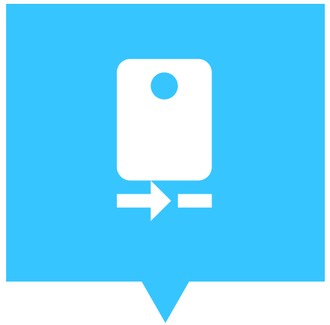
Objective

- X Education to assist the CEO in achieving a target lead conversion rate of approximately 80%. Your task is to develop a logistic regression model that assigns a lead score ranging from 0 to 100 to each lead. This scoring system will enable the company to effectively target potential leads.

- Additionally, the model should be flexible enough to adapt to future changes and address any other issues presented by the company. This means that your solution should be capable of accommodating modifications in the company's requirements as they arise.



How did you go about studying the Lead scoring Case Study



Use of Python on a Jupyter Notebook for EDA analysis
Feature scaling, dummy variables and data encoding.

Understanding the descriptors of the variables
Using the dictionary provided.

Research:

To familiarize ourselves we did a research on understanding various parameters that helped us to understand variable descriptors better and its relevance with respect to our case study.

Using logistic regression for model
building and making predictions.

Data Cleaning

1.

- Find the percentage of missing values of all the columns
- Remove columns with high missing percentage

2.

- replacing the unique data with relevant category
- Checking and removing outliers

3.

- Values that are not mentioned are replaced with other relevant values.

Steps taken

1.

- We converted select as NaN as they are as good as null value.

2.

- 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.

3.

- Removed outliers.
- Removed null values as well as the city Mumbai.

EDA

1

Univariate Analysis

2

Bivariate Analysis

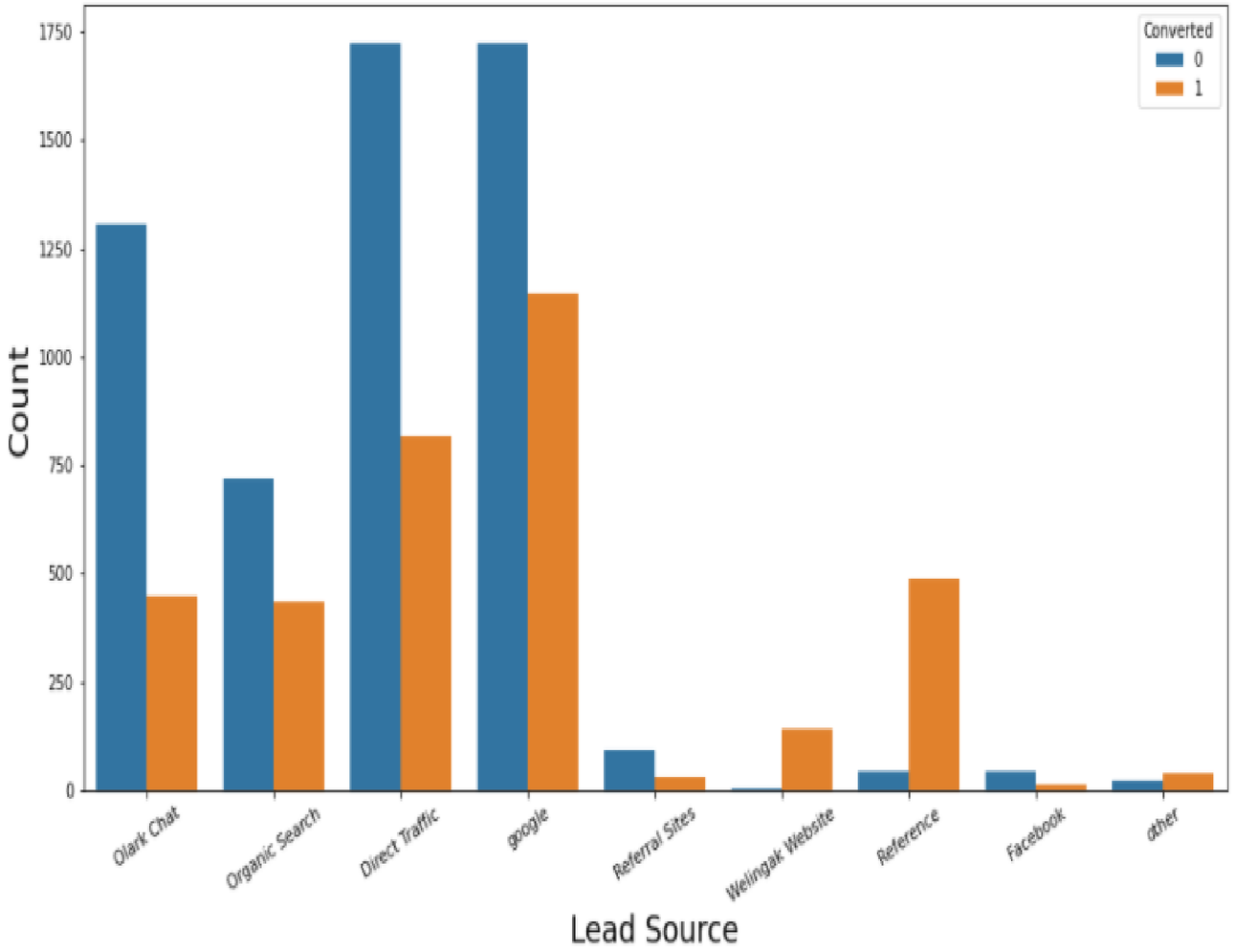
3

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

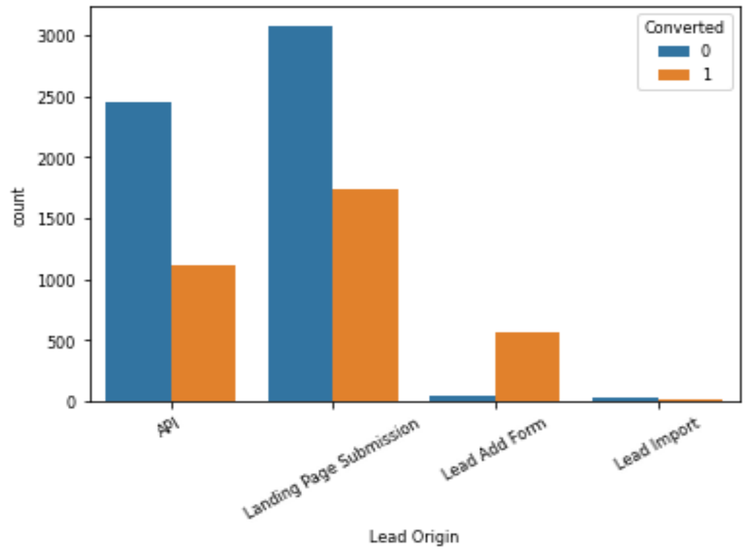
To improve the overall lead conversion rate, we should focus on the Organic Search,Olark Chat,Direct Traffic and google leads in the Lead Source and generates more leads.

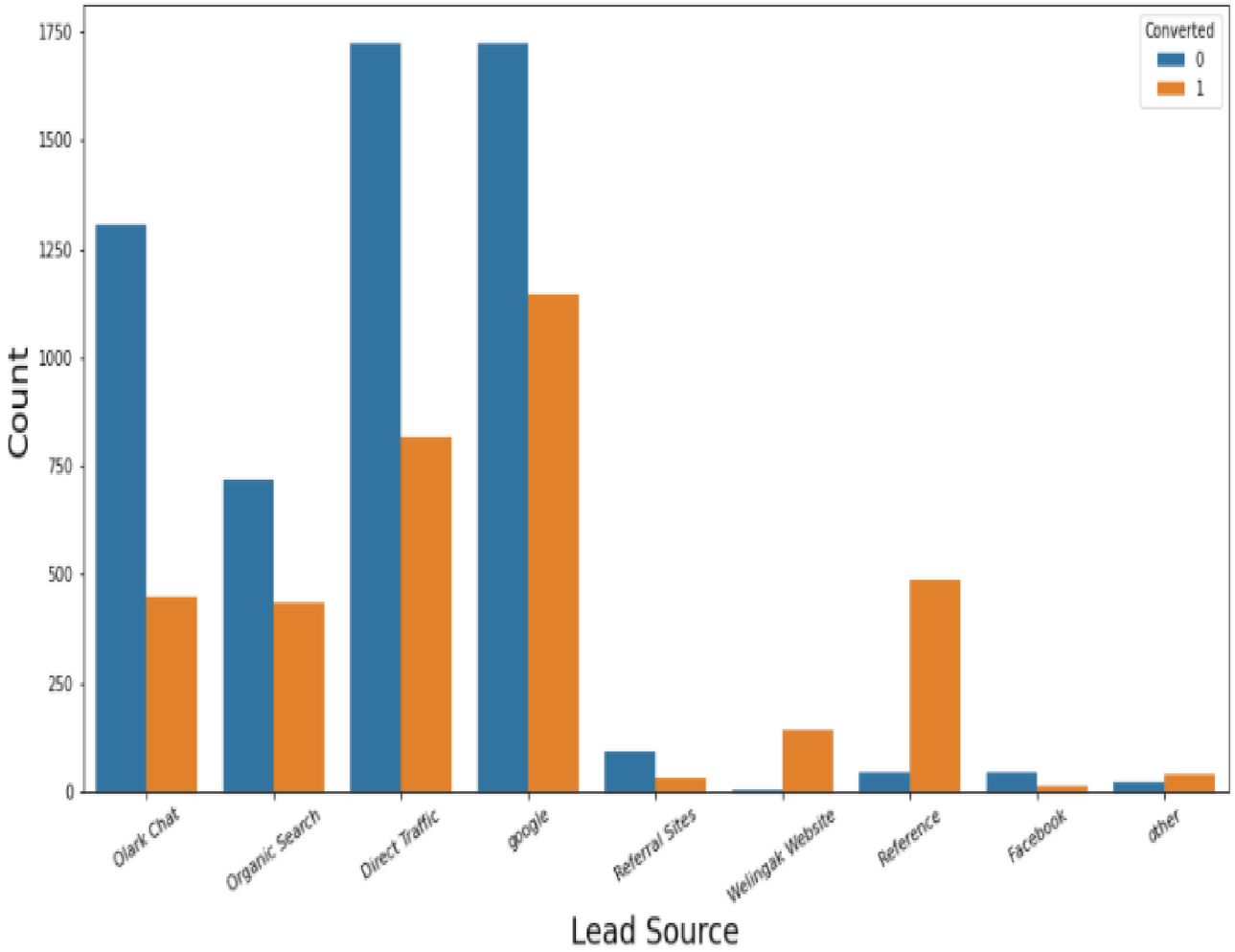


Visualization and inferences

- 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.

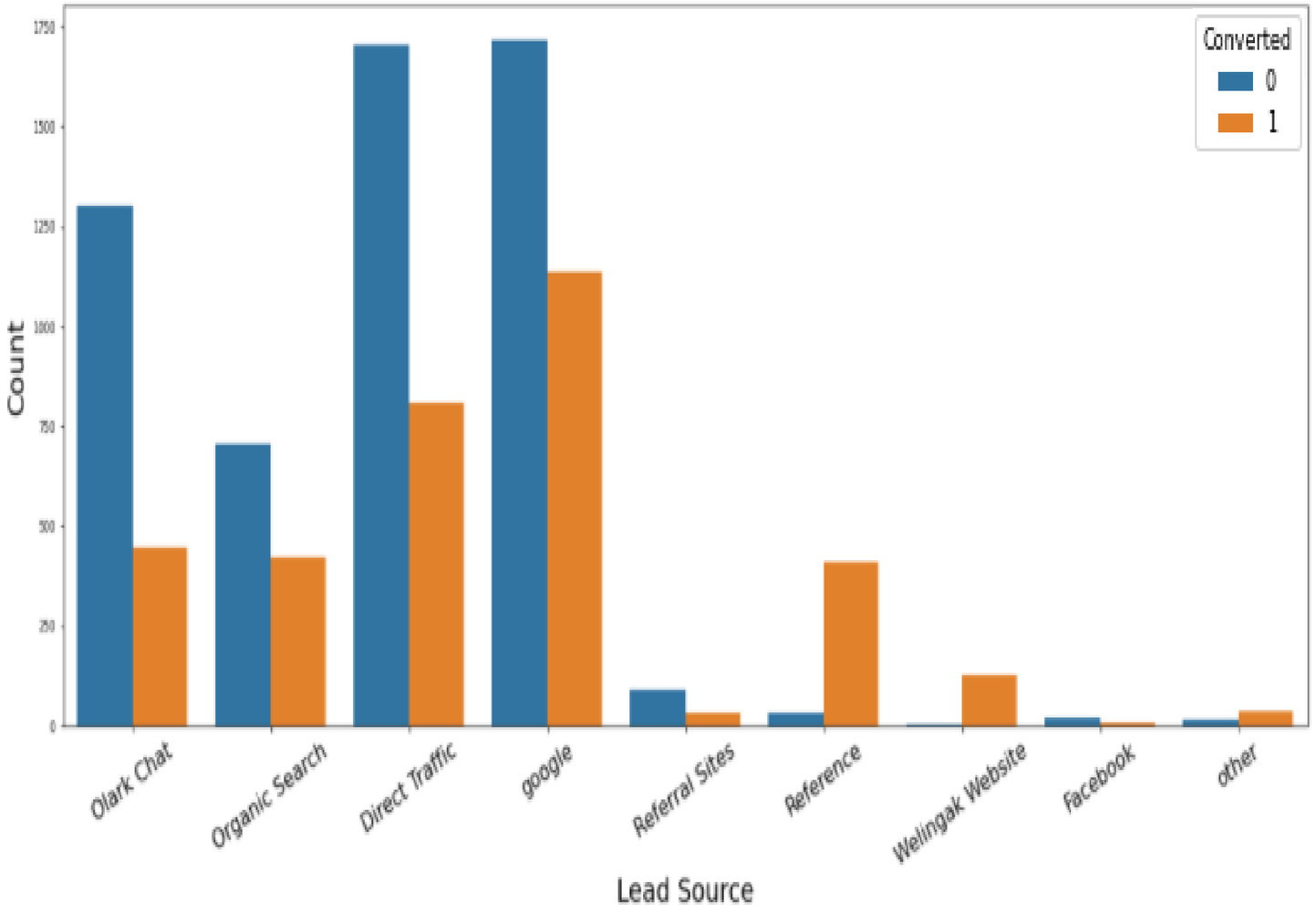




- API and Landing page show a good conversion rate and are good enough in number.
- While Lead Add form has high conversion rate but do not have enough counts.

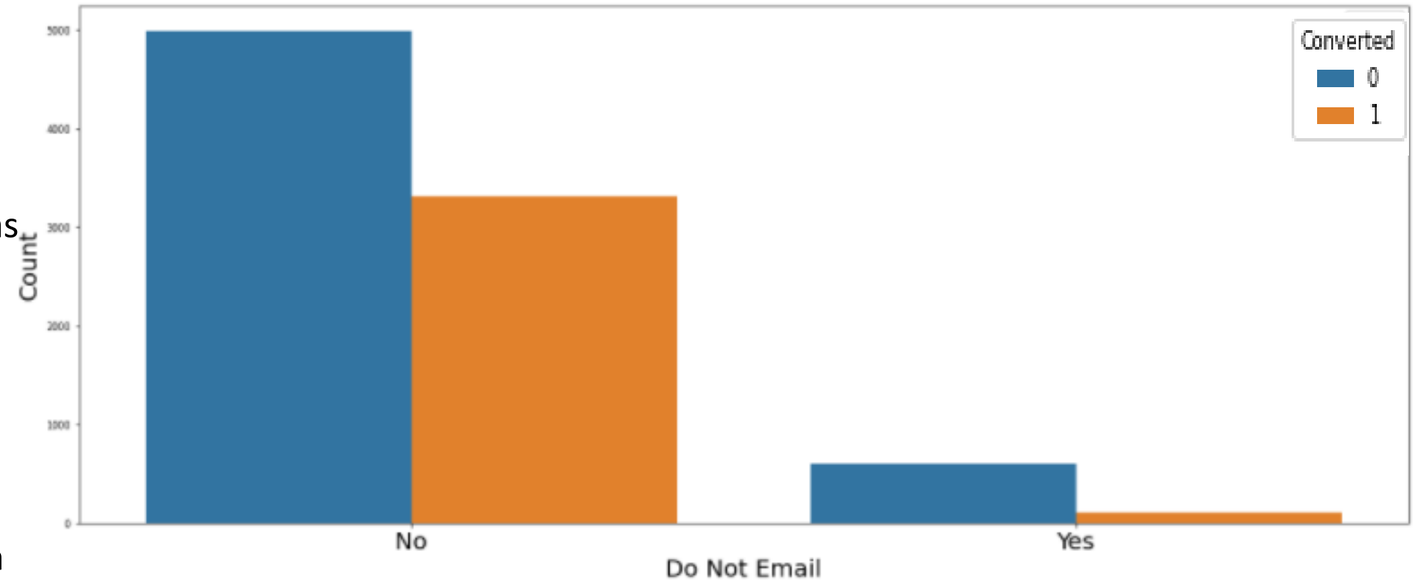
Visualization and inferences

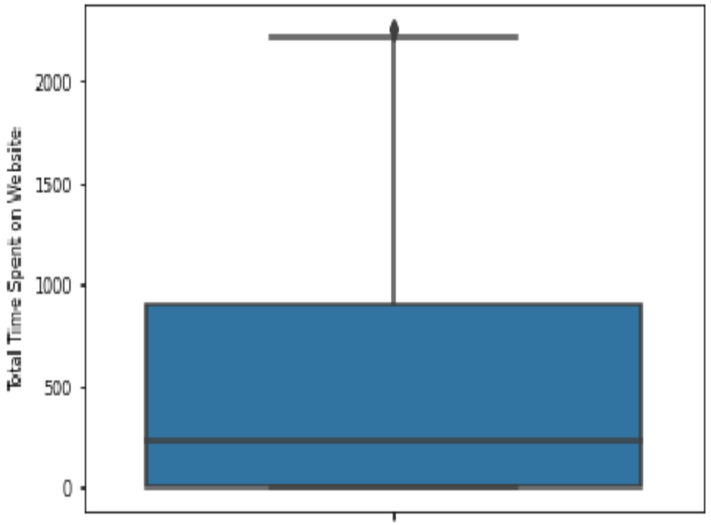
- Most of the leads are generated by Google and Direct Traffic which also has good conversion rate.
- Leads by reference is mostly converted a same goes with Welingak Website.

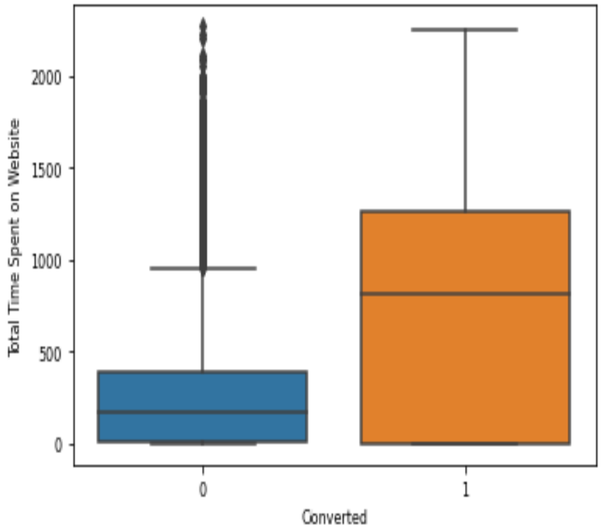


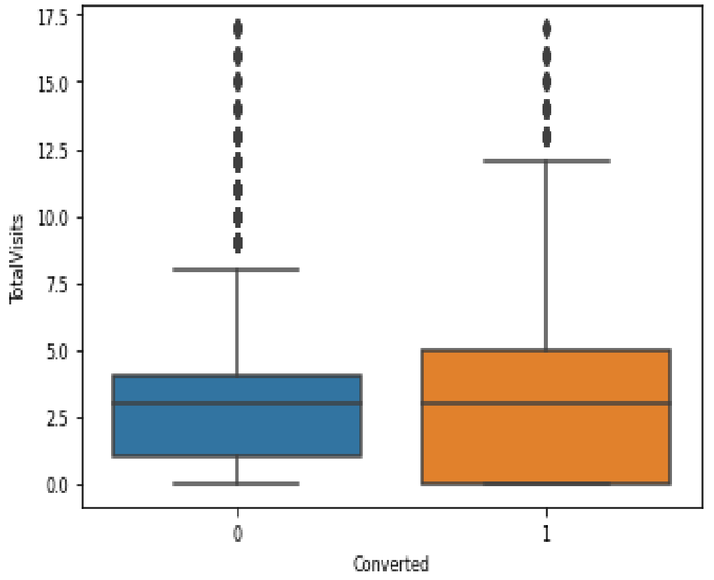
Visualization and inferences

- 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.



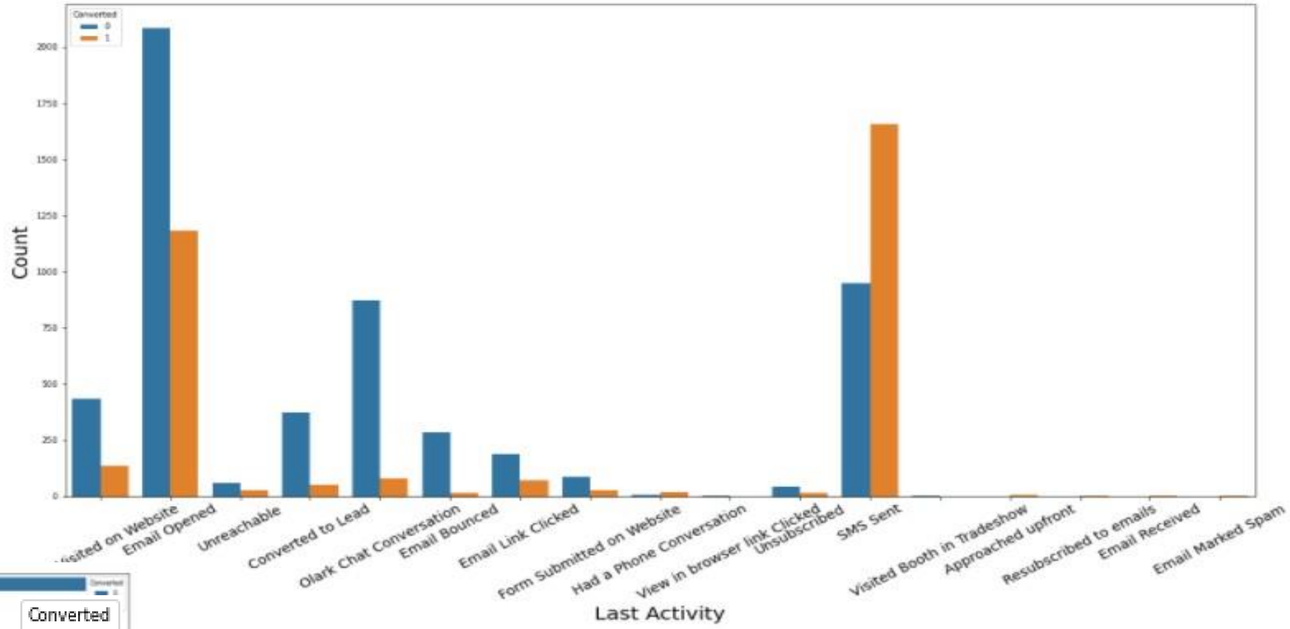


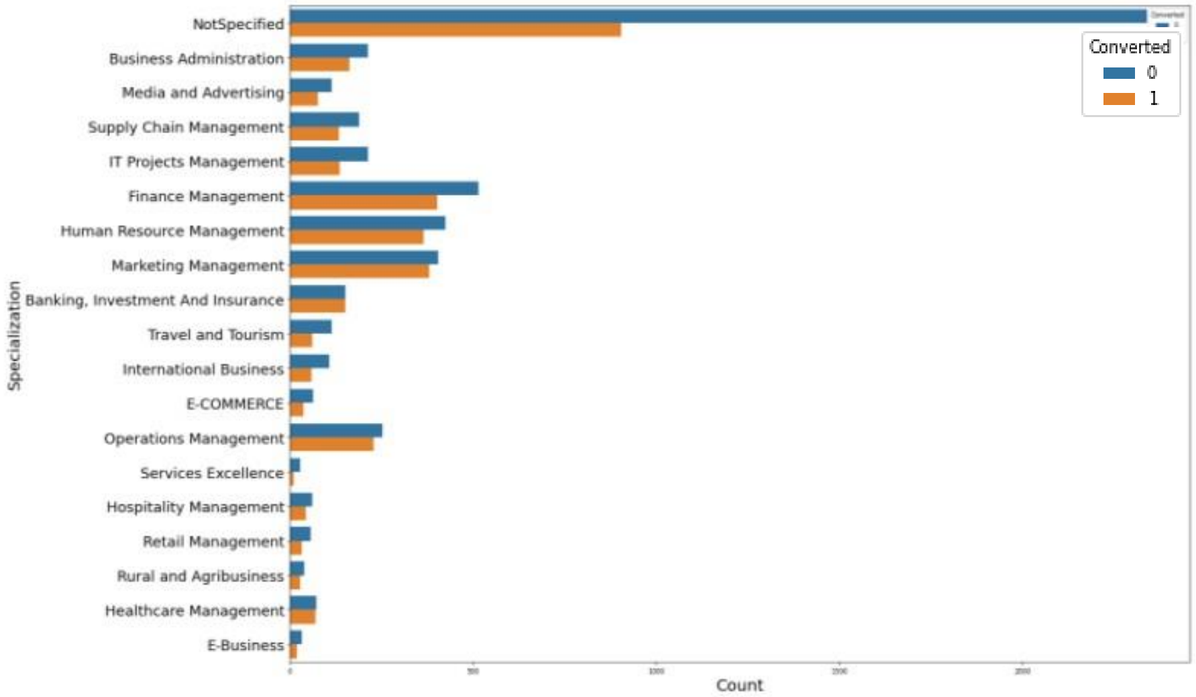




Categorical Variable relation

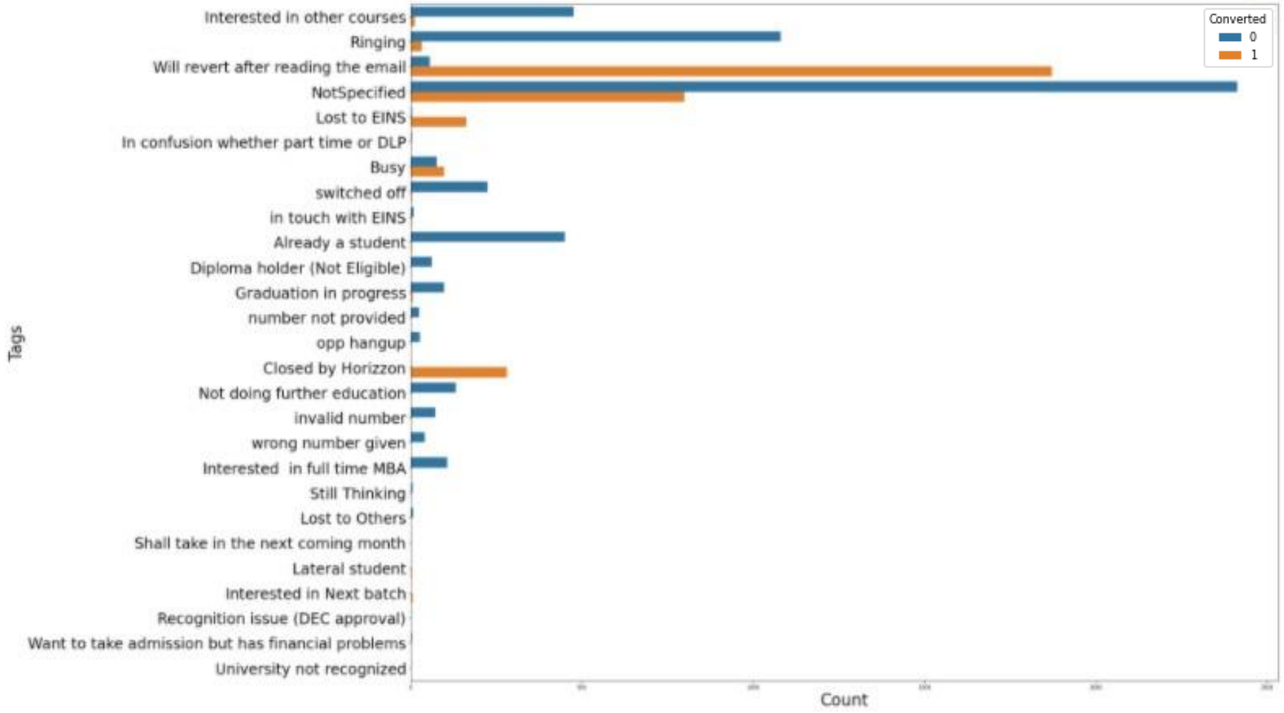
- Most of the leads are from email opened.
- SmS sent has the highest conversion rate among all of them.



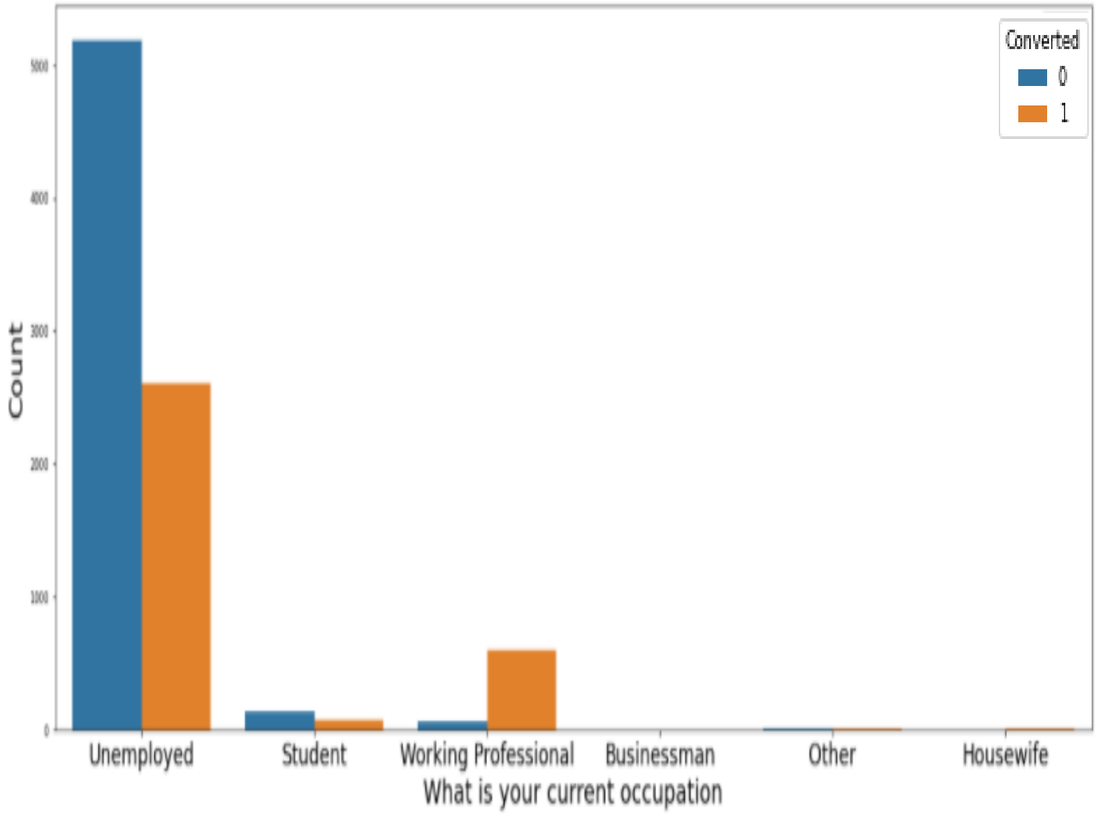


As we can see that those who have not specified their industry domain in which they worked before have the highest conversion rate followed by Finance management ,HR management and marketing management.

Categorical Variable relation

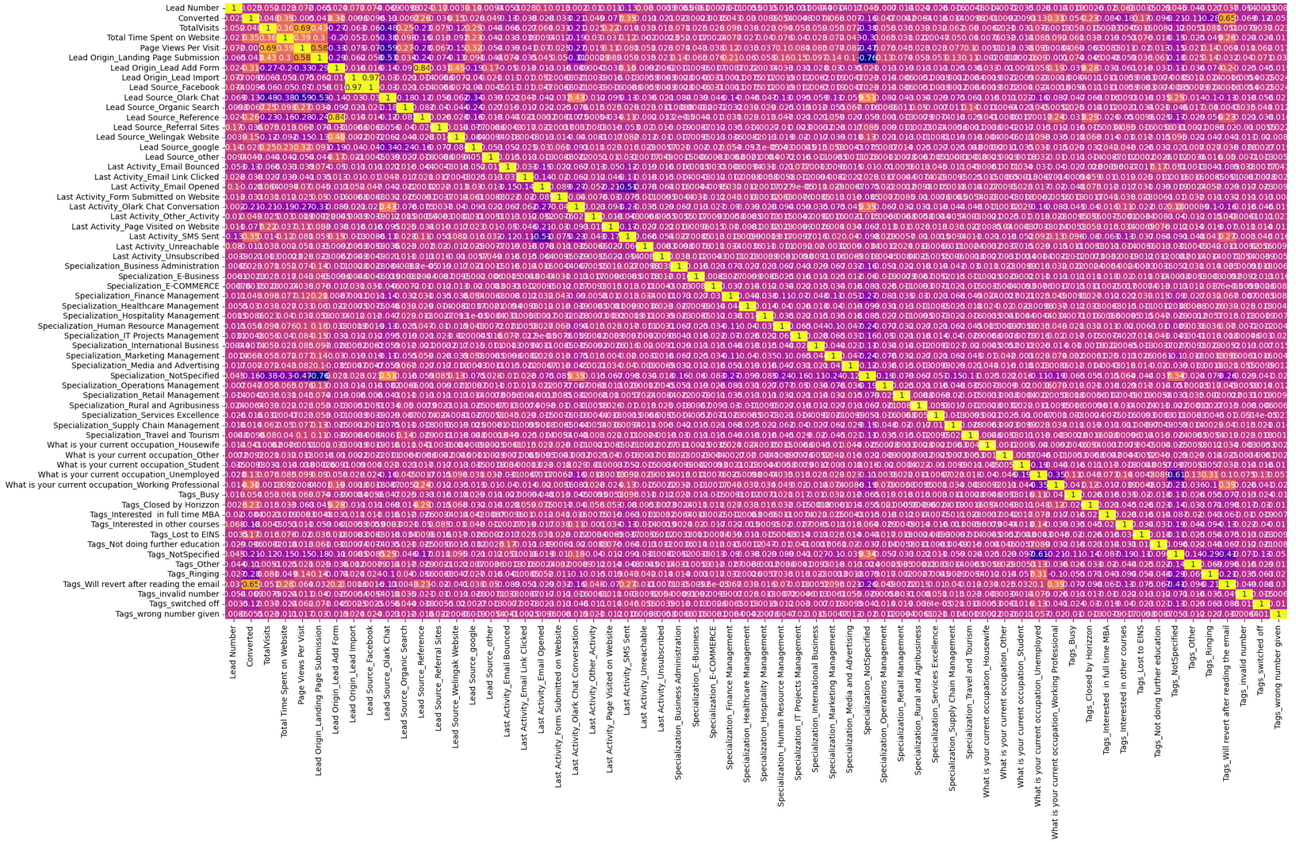


Customers who said they will revert after reading the email have maximum conversions , that means the content in emails have a positive impact on customers.



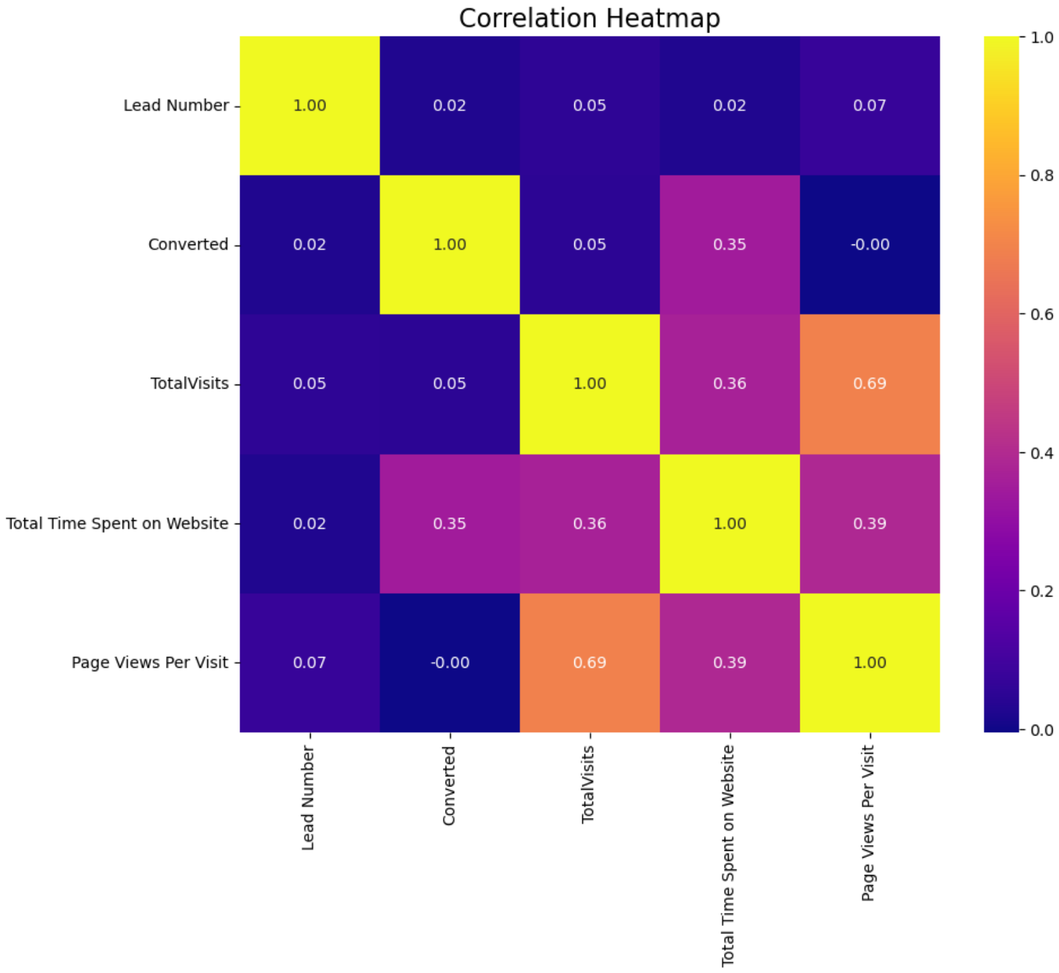
- Working Professionals have high conversion rate as compared to others.
- Most of the leads are generated by Unemployed.

Heatmap after dropping the unnecessaries.



Heatmap after Encoding ,Scaling and creating dummy variables.

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



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

Building our model

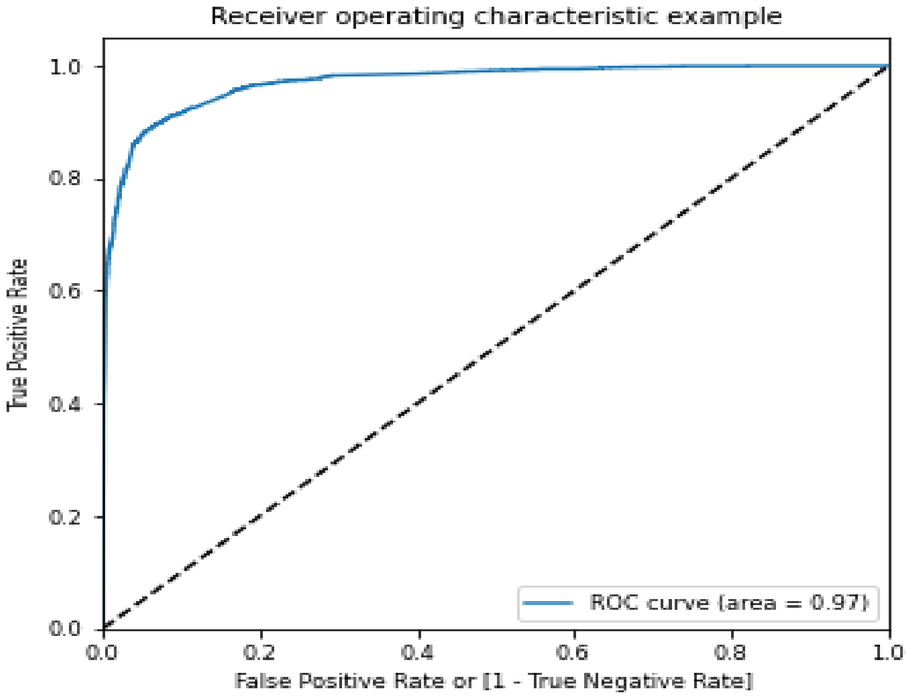
Steps performed:

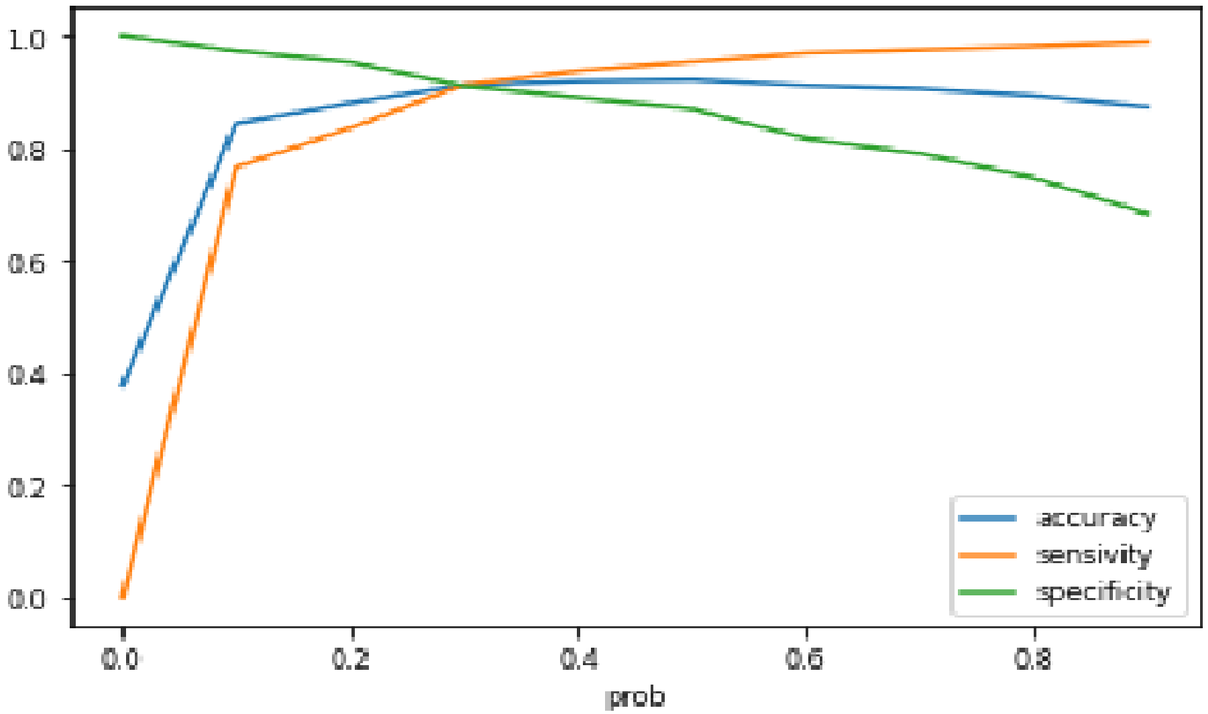
- Splitting the data into training and test sets.
- While performing test –train split , we choose 70 : 30 ratio.
- Using RFE for feature selection.
- Running RFE with 15 variables as output.

.

- 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

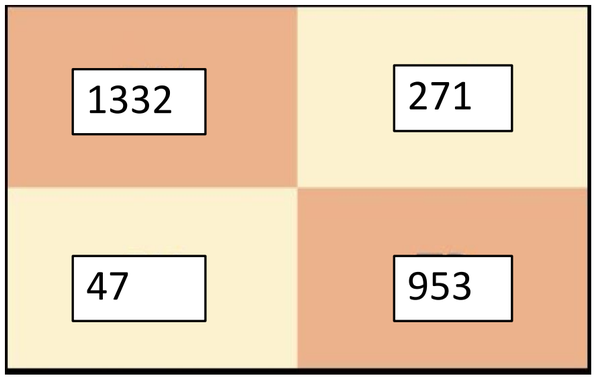
Confusion Matrix

- “**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.

Matrix

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Matrix



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 Not Specified
- Lead Source Olark Chat

- Lead Origin Lead Add Form
- Last Activity SMS Sent
- Tags Ringing

- Total Time Spent on Website
- Last Activity Email Opened
- Lead Source We lingak Website

- Tags Closed by Horizzon
- Tags Busy
- Tags Lost to EINS

Conclusion

Based on our analysis, the following insights were obtained:

- Lead Generation Sources: Google and Direct Traffic generate the highest number of leads, with good conversion rates. Focusing on these sources can be beneficial for lead acquisition and conversion.
- Effective Channels: Referrals and the Welingak website have shown potential for generating quality leads with good conversion rates. Allocating resources towards these channels can yield positive results.
- Email Marketing: Sending targeted emails to customers has proven to be highly effective in lead conversion. This approach should be emphasized to improve the conversion rate.
- SMS Campaigns: Among different communication channels, SMS has demonstrated the highest conversion rate. Giving attention to SMS campaigns can lead to increased conversion rates.
- Targeting Undesignated Industry: Leads who do not specify their industry domain have shown the highest conversion rate. Additionally, focusing on Finance Management, HR Management, and Marketing Management domains can also yield favorable conversion rates.

Furthermore, the top features associated with a good conversion rate are as follows:

- Tags_Closed by Horizon
- Tags_Lost to EINS
- Tags_Will revert after reading the email

To increase the probability of lead conversion, the top three categorical/dummy variables in the model that should be prioritized are:

- Tags
- Lead Source
- Lead Origin

By focusing on these variables and leveraging the insights obtained from the analysis, you can enhance the effectiveness of lead conversion strategies and improve overall conversion rates.

Thanks

If you have any specific questions or need further elaboration on any of the points mentioned, please feel free to ask. I'm here to assist you!