

Lead Score Case Study

By:

Sudeshna Mahapatra

Dip Ghosh



Problem Statement



- X Education sells online courses to industry professionals.
- company markets its courses on several websites and search engines like Google.
- Once people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- people fill up a form providing their email address or phone number, they are classified to be a lead.
- company also gets leads through past referrals.
- Once leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. ⚡ lead conversion rate is very poor.
- To make process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'

Lead Conversion Process - Demonstrated as a funnel

- lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom.
- middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- ballpark of the target lead conversion rate to be around 80%.





Data

- leads dataset from the past with around 9000 data points
- dataset consists of various attributes such as **Lead Source**, **Total Time Spent on Website**, **Total Visits**, **Last Activity**, etc.
- target variable is the column **'Converted'** which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
- Many of the categorical variables have a level called **'Select'** which needs to be handled because it is as good as a null value (think why?).





Goals of the Case Study:

- 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.
- 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.





Results Expected:

- 1. A well-commented Jupyter notebook with at least the logistic regression model, the conversion predictions and evaluation metrics.
- 2. The word document filled with solutions to all the problems.
- 3. The overall approach of the analysis in a presentation
 - Mention the problem statement and the analysis approach briefly
 - Explain the results in business terms
 - Include visualisations and summarise the most important results in the presentation
- 4. A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.





Approach used for analysis:



Step-1: Reading and Understanding Data

A *Importing Necessary library modules*

B *Understanding the structure of the data*

1 *Reading the dataset/Importing Data*

2 *Sanity checks for the loaded Data*

Inferences:

```
----> Leads_data_df contains 36 features, 1 target variable  
----> 4 features are of float64 datatype  
----> 3 features are of int64 datatype  
----> 30 features are of object datatype
```




Step-2: Data Pre-processing

1 Checking for categorical variables

✦ It can be observed that 30 columns are of categorical type.

2 Checking for necessary information

✦ This is the step where data manipulation is done based on the inspection of the given data.

In this step both missing values and outliers are identified.

Following are the assumptions taken during the manipulation phase.

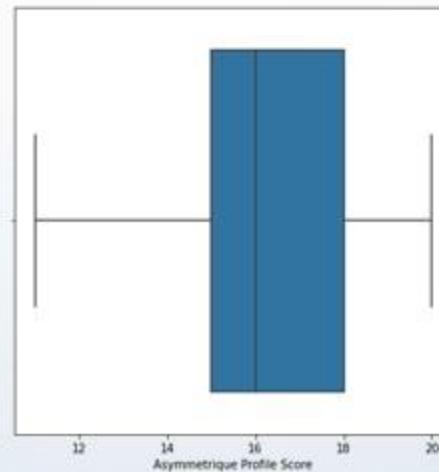
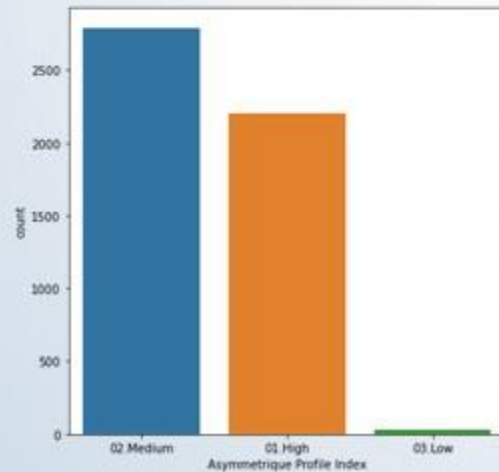
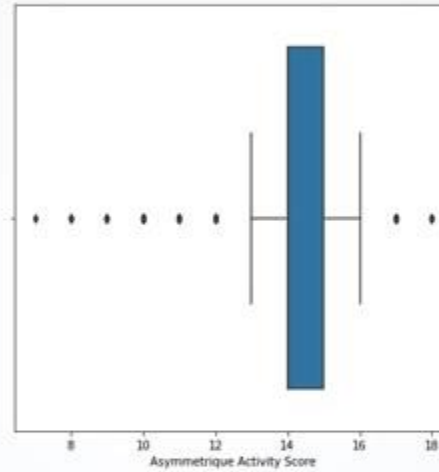
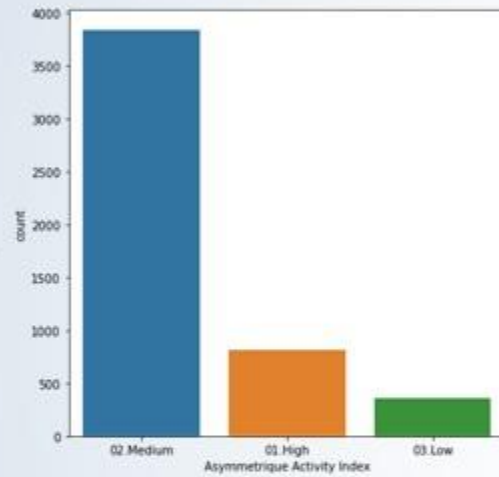


DATA MANIPULATION

Data Manipulation based on assumptions

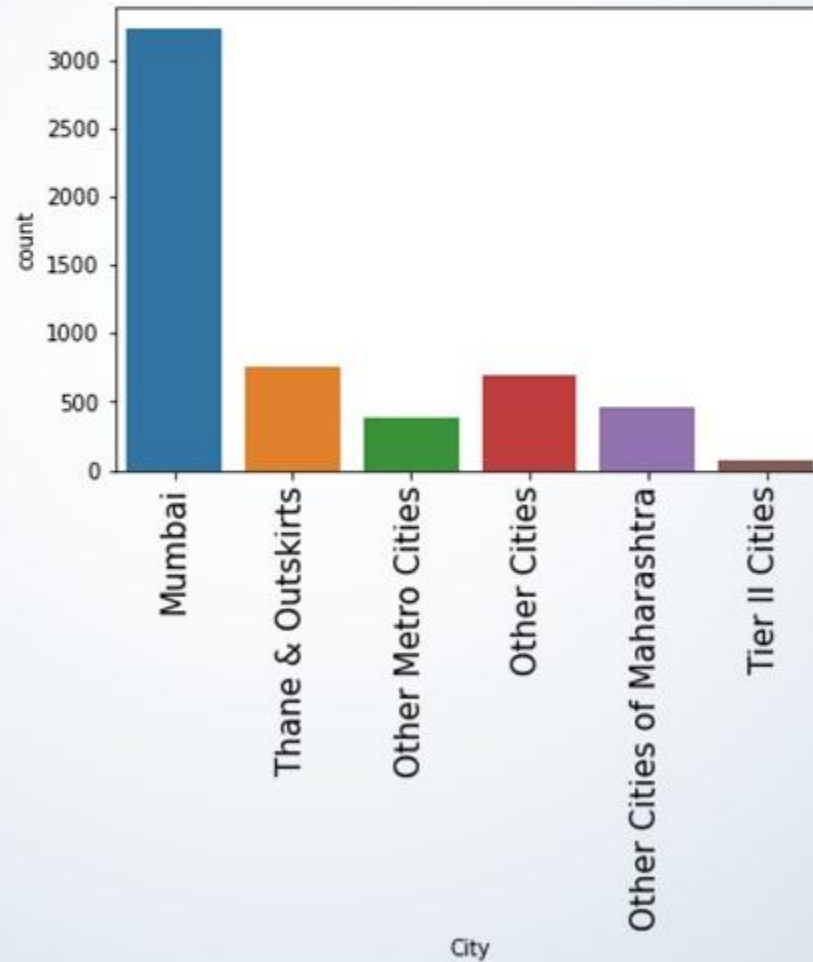
- ❁ **First Assumption:** *As previously observed the columns there are some values where just select is present. This is because the students have not filled that particular and so it can be considered as null as well.*
- ❁ **Second Assumption:** *Regardless of empty values mails have been sent to every student visiting the website in the Tags feature so maybe they might be interested after viewing the mail and will revert back after reading the mail.*
- ❁ **Third Assumption:** *In Lead Quality feature all the empty values are imputed with the Not Sure since both are to be considered for the same scenario.*

✿ **Fourth Assumption:** Asymmetric parameters have wide variations in data so values are dropped above the threshold of 45%.

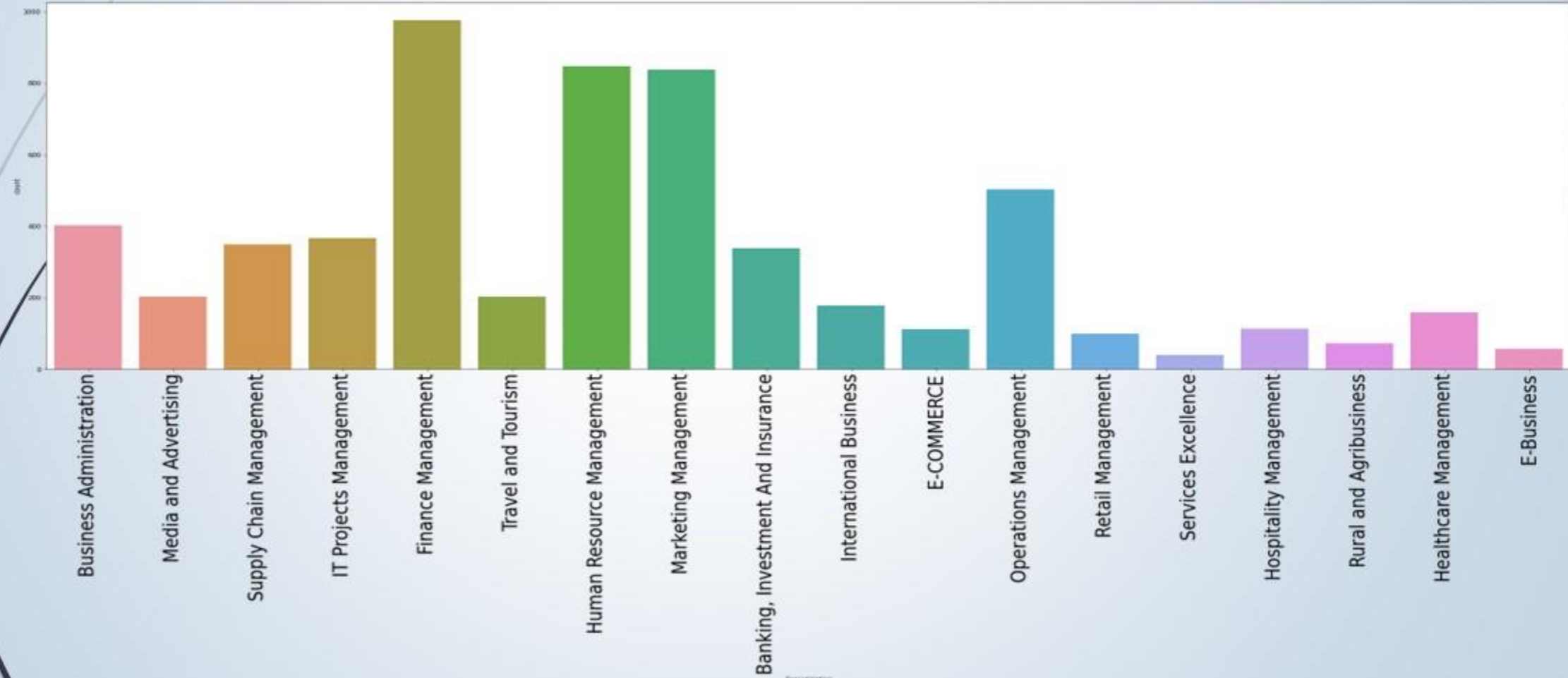


Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	

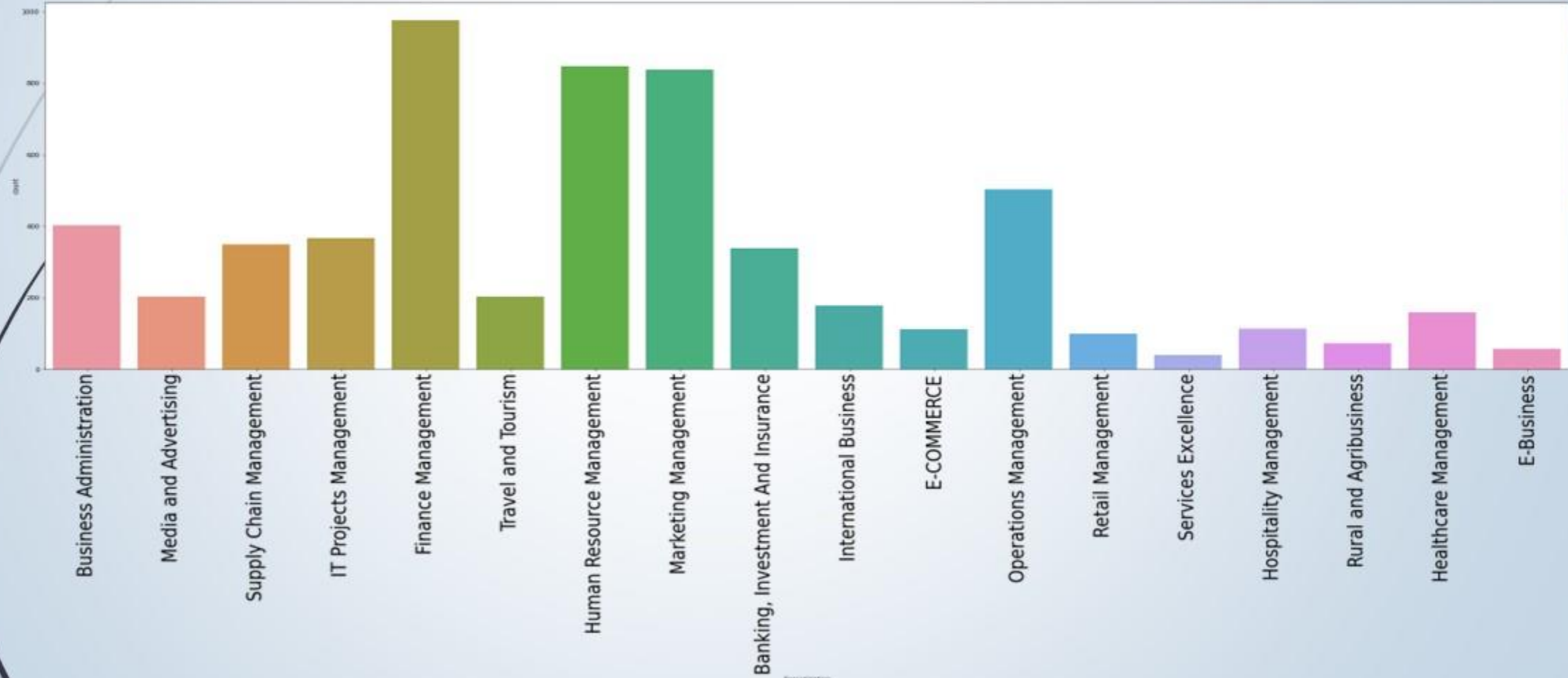
✿ **Fifth Assumption:** City feature has more than 60% values of the city 'Mumbai'. So wherever NaN values are there they can be imputed with it.



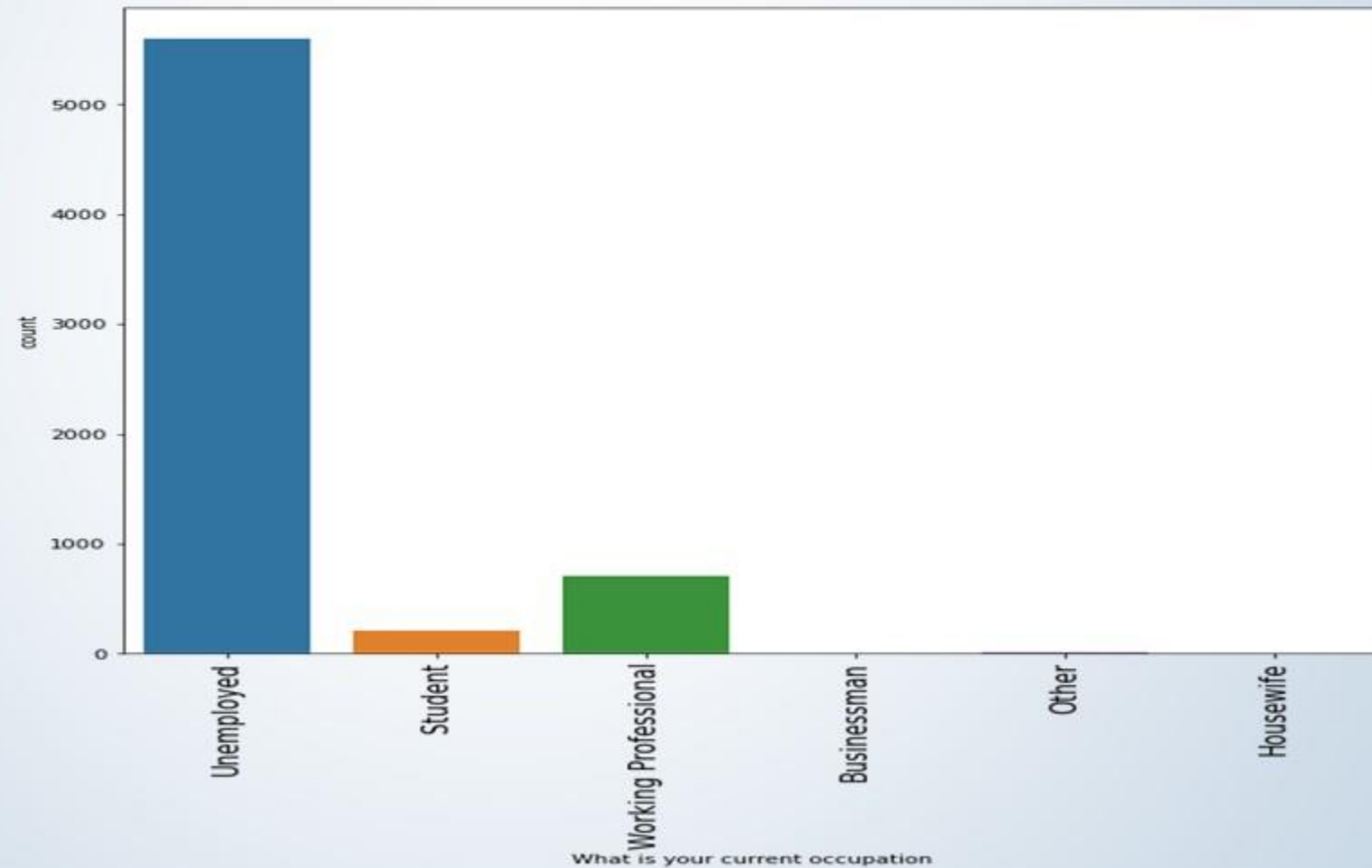
✿ **Sixth Assumption:** *In the Specialisation feature most of the students have mentioned their specialisation so it would be wise to impute the NaN values here into others or unknown category.*



✿ **Seventh Assumption:** *In the 'What matters most to you in choosing a course' feature most of the students have chosen 'Better career Prospects'. So it would be better to impute NaN values with the same.*



✿ **Eight Assumption:** *In the Occupation feature 96% entries are of unemployed. So it would be better to impute NaN values with it.*



✿ **Nineth Assumption:** *In the Country feature 98% entries are of India.
So it would be better to impute NaN values with it.*





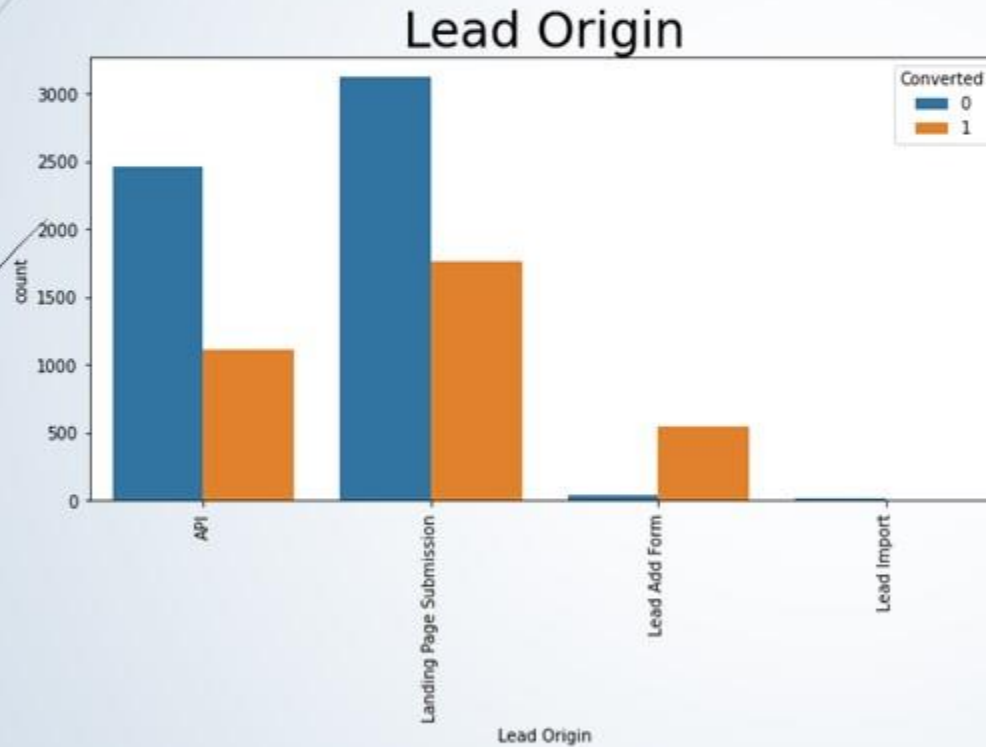
Step-3: Exploratory Data Analysis

Converted

The target variable. Indicates whether a lead has been successfully converted or not.

```
1 # Converted is the target variable, Indicates whether a Lead has been successfully converted (1) or not (0).
2 Converted = (sum(Leads_data_df['Converted'])/len(Leads_data_df['Converted'].index))*100
3 Converted
: 37.85541106458012
```


(🌸'☺'🌸) Univariate Analysis

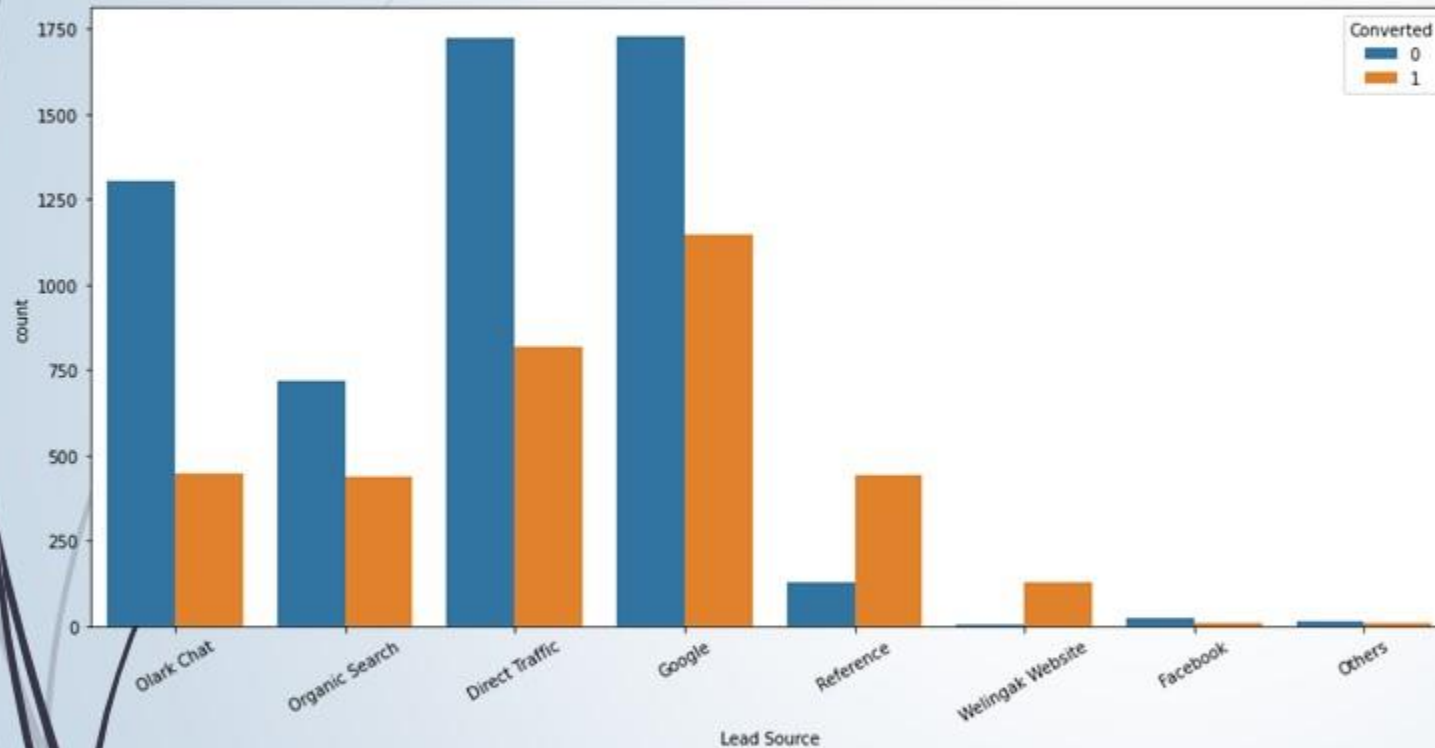


Insights:

- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead Import are very less in count.

Inference:

- Thus, we can conclude that to improve overall lead conversion rate, the focus should be on improving lead conversion of API and Landing Page Submission origin as they generate more leads from Lead Add Form and Lead Import.

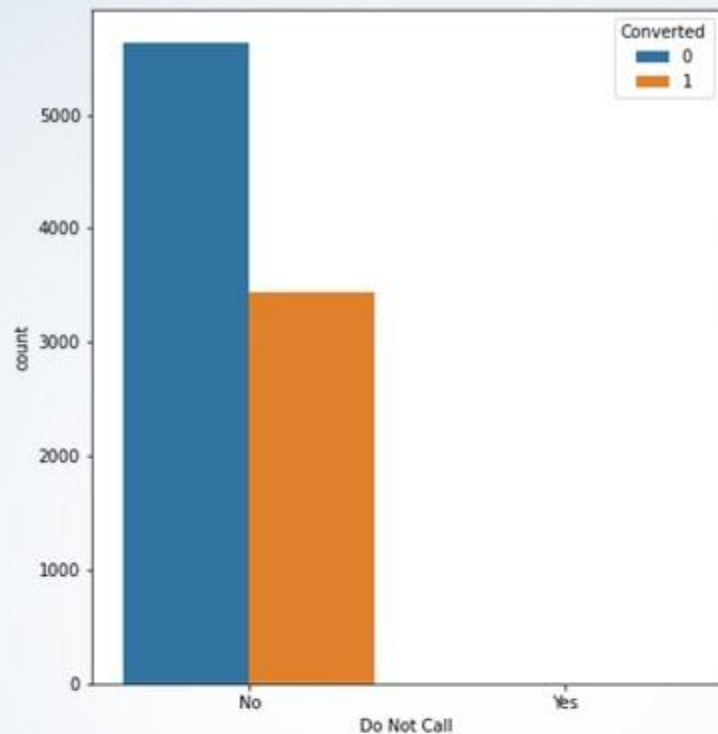
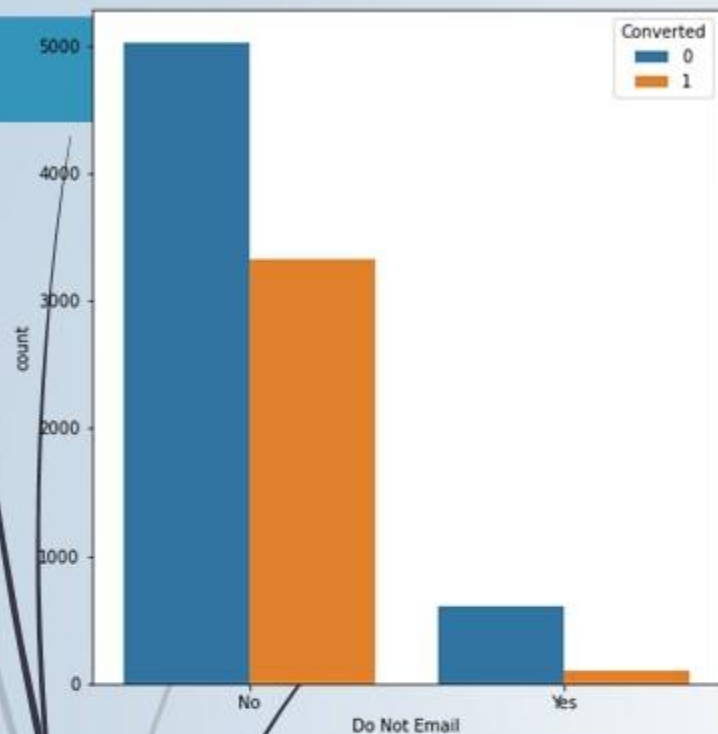


Insights:

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.

Inference:

- Thus, we can conclude that to improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads as they generate more leads from reference and welingak website.

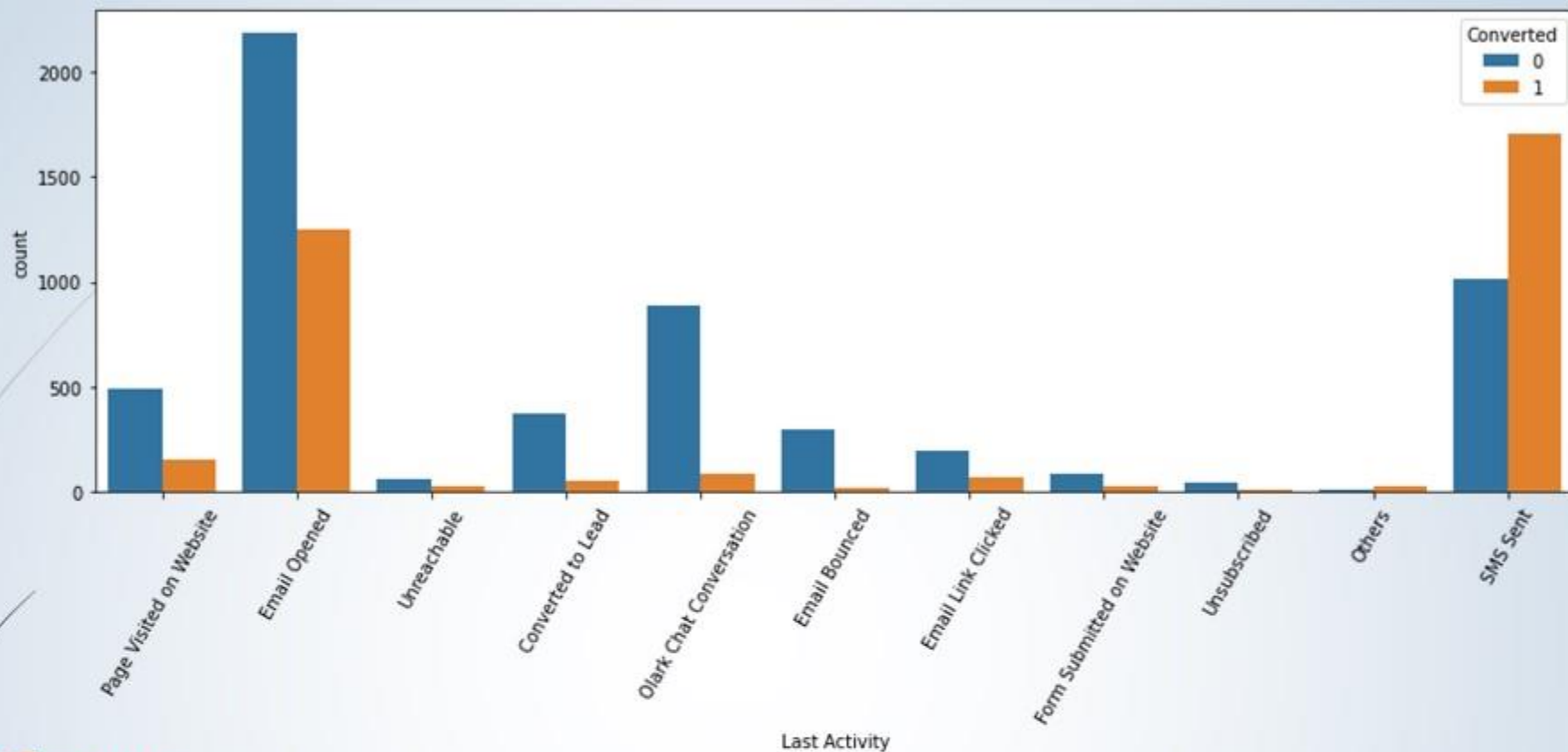


Insights:

- The conversion rate of leads for Do Not Email candidates is close to 63% when mailed but it is 20% when not mailed.
- Conversion Rate of Do Not Call Candidates is 56% approximately when called but zero when not called.

Inference:

- Thus we can conclude that to improve overall lead conversion rate, focus should be on making awareness or intimating the candidates either through call or email after a candidate has visited the website.

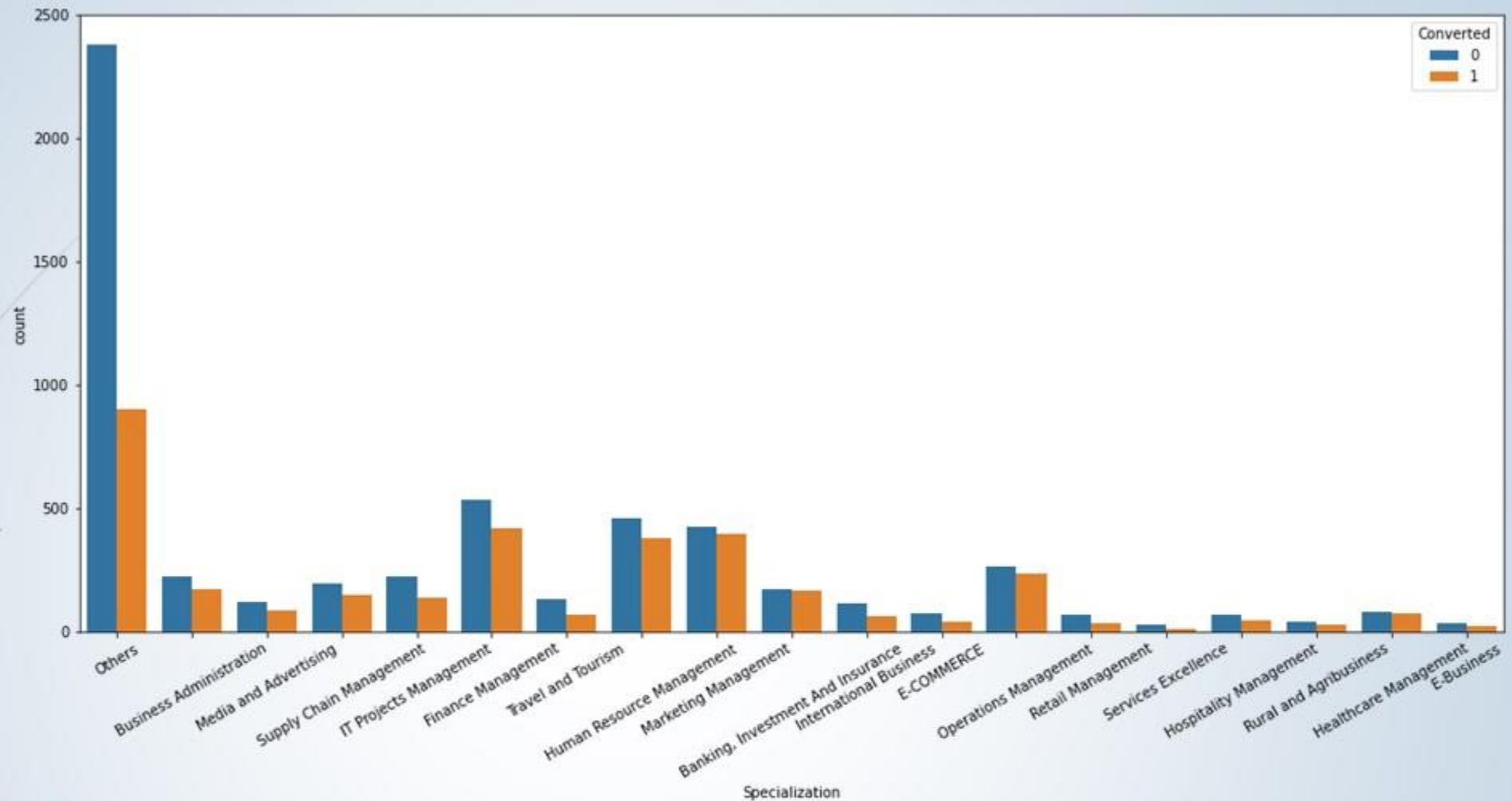


Insights:

- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost 60%.

Inference:

- Thus we can conclude that to improve overall lead conversion rate, focus should be on making awareness or intimating the candidates either through SMS or email after a candidate has visited the website.

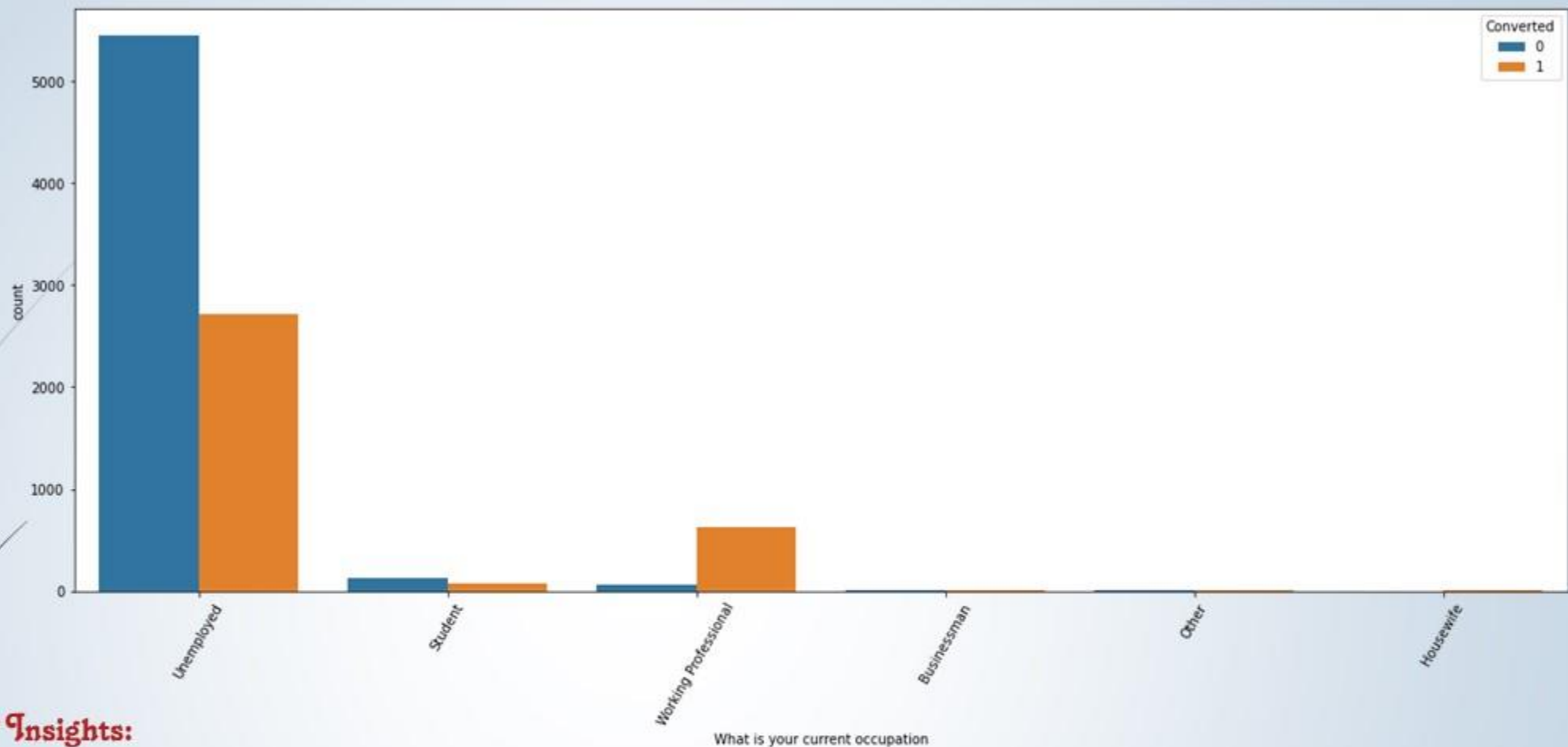


Insights:

- The conversion rate of leads for Business Administration, Supply chain management, etc are more.

Inference:

- Thus we can conclude that to improve overall lead conversion rate, focus should be on management professionals.



Insights:

- The conversion rate of leads working professionals are more.
- Conversion Rate of Unemployed candidates are close to 55%.

Inference:

- Thus we can conclude that to improve overall lead conversion rate, focus should be on working professionals.

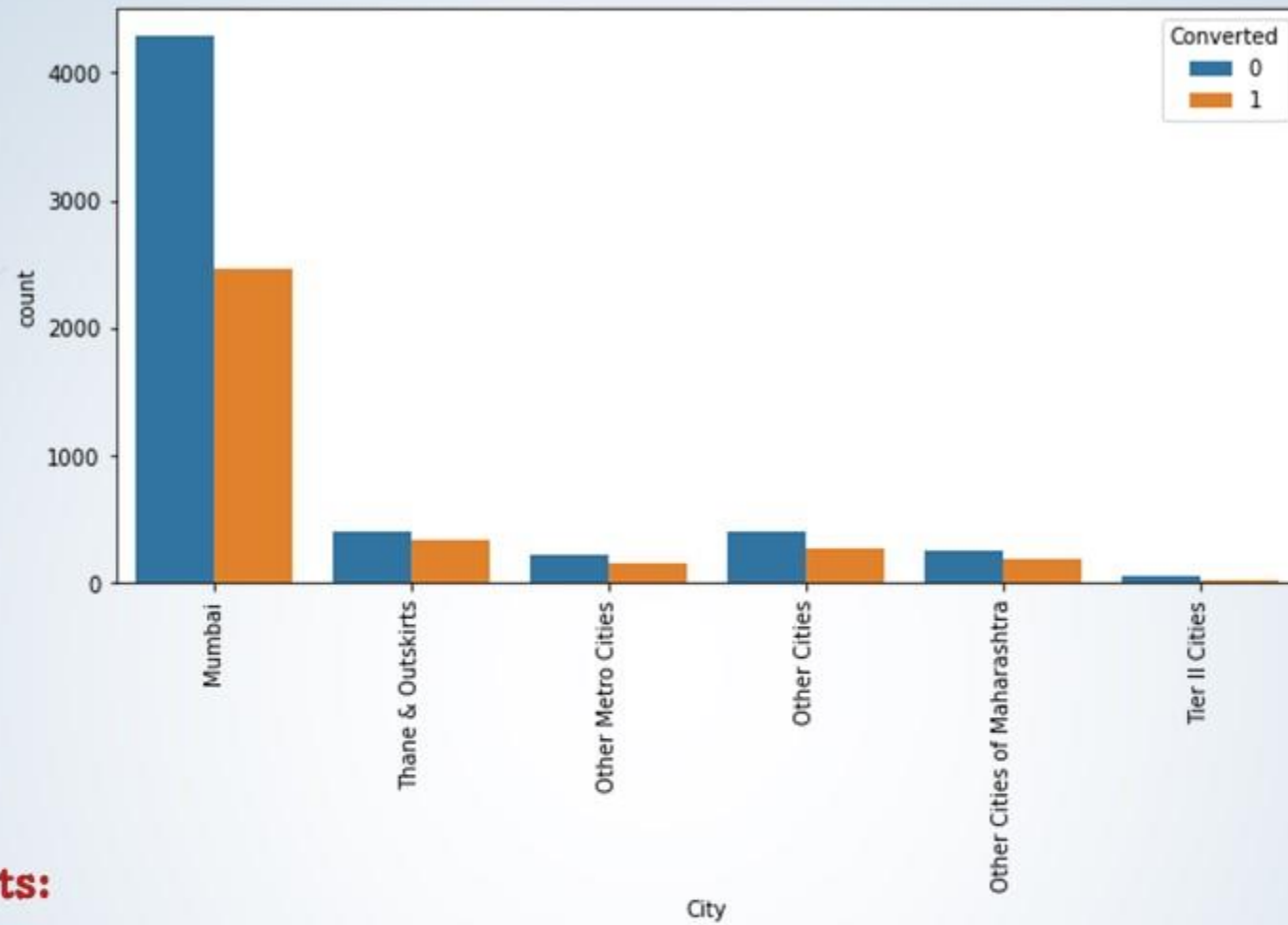


Insights:

- The conversion rate of leads with High relevance in the field is the highest followed by low relevance in field.
- Conversion Rate of not sure and worst candidates are the lowest.

Inference:

Thus we can conclude that to improve overall lead conversion rate, focus should be on making awareness the type of content being offered by the company to the candidates for the course to be appealing.

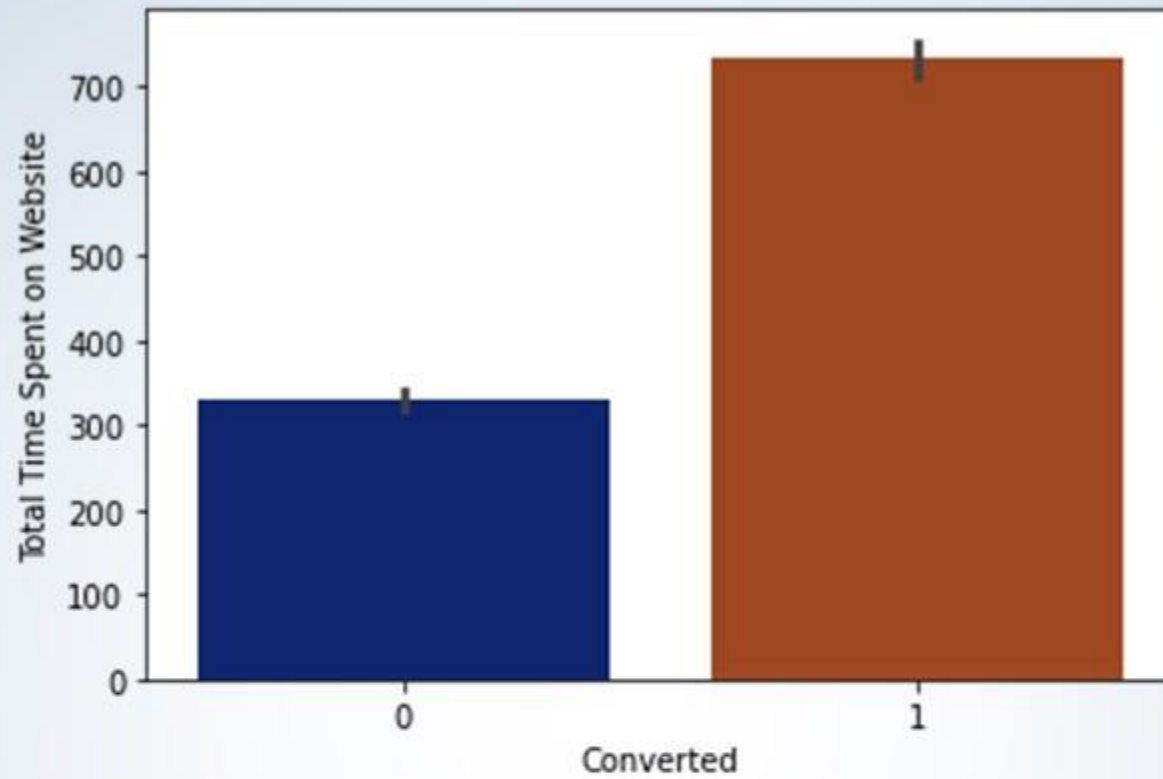


Insights:

- Here we can observe that most of the entries are from 'Mumbai'.

Inference:

- Here it can be concluded that around 37% conversion of leads happens from 'Mumbai' city.



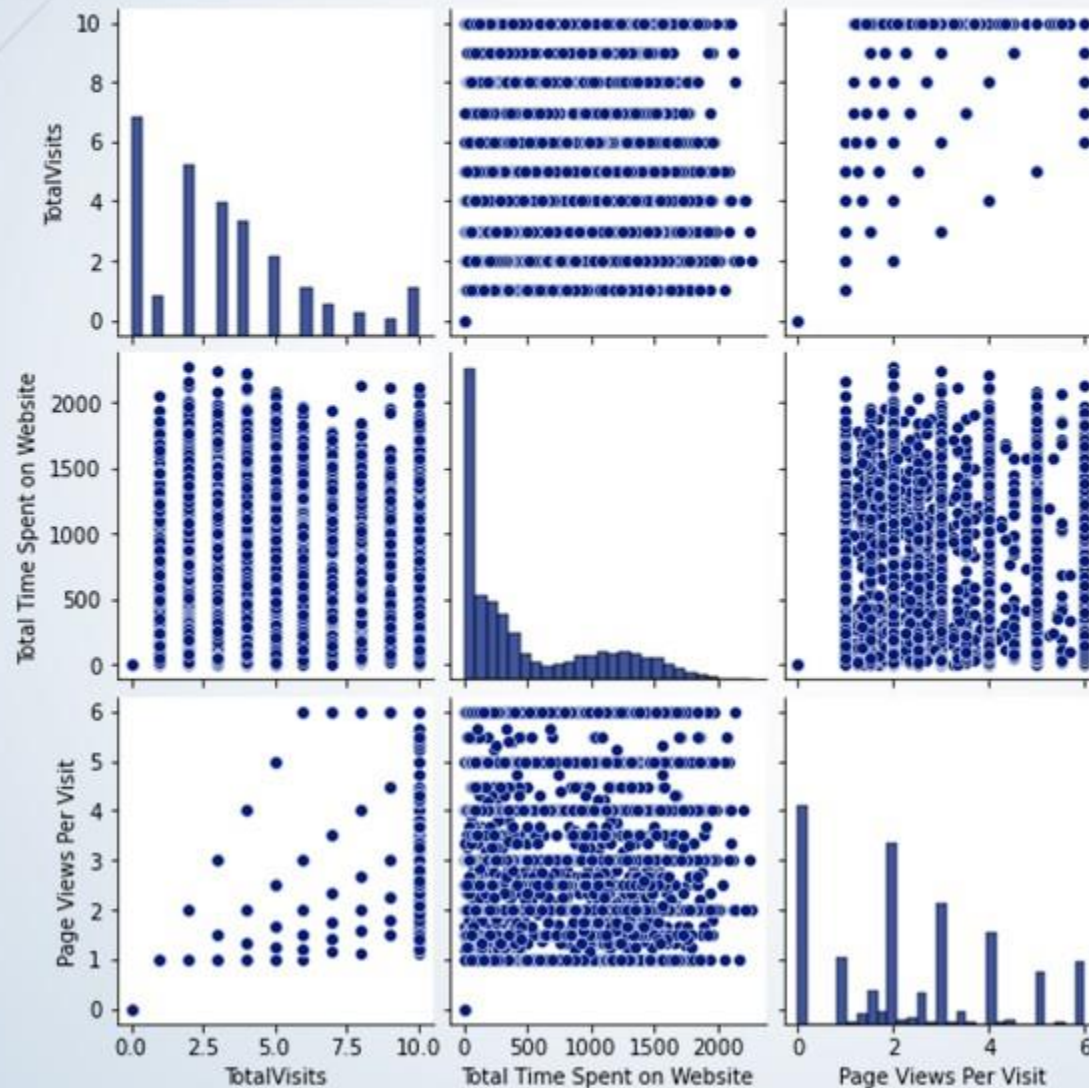
Insights:

- Leads spending more time on the website are more likely to be converted.

Inference:

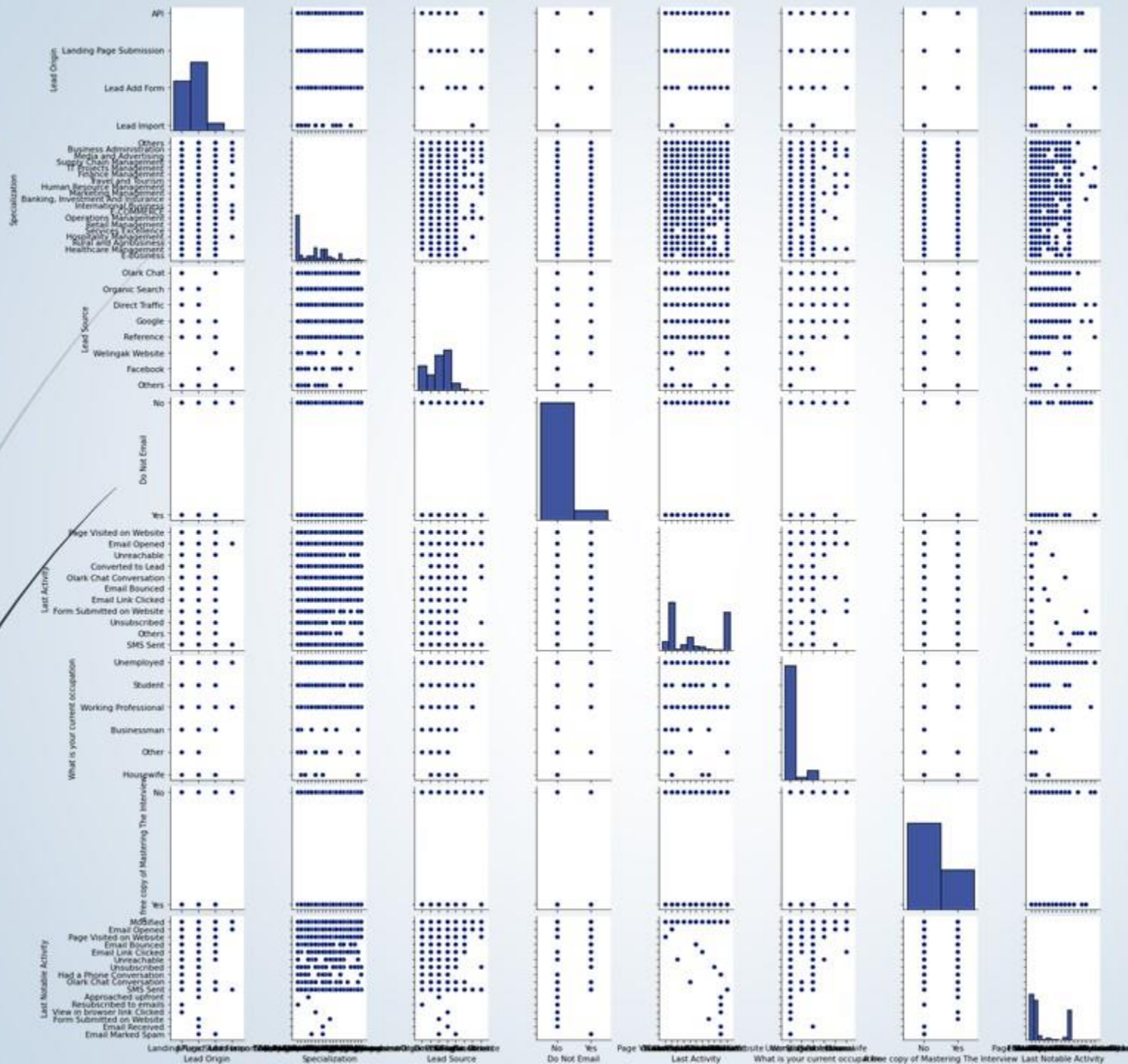
- Thus we can conclude that to improve overall lead conversion rate, focus should be on improving website infographics and interface.

(🌸'☺'🌸) Bivariate Analysis



Insights:

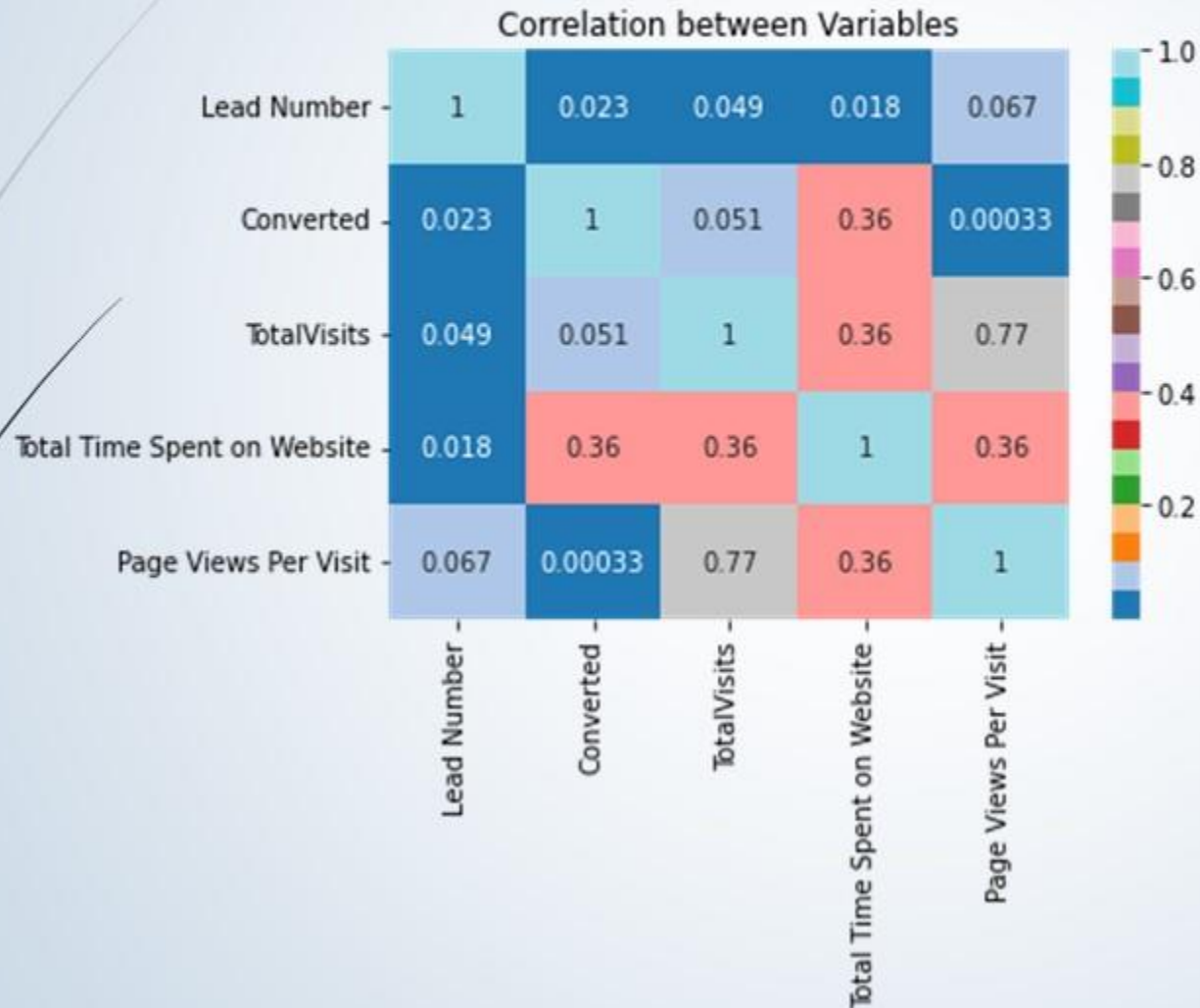
- As it can be observed that there is low variation in Page Views Per Visit and Total Visits but higher variation in Total Time Spent on Website



Insights:

- It can be observed that there is some correlation between Conversion and some categorical columns like Lead Origin and Lead Source.

(🌸😊🌸) Multivariate Analysis



Insights:

- It can be observed that there is positive correlation between Total Time Spent on Website and Conversion.
- It can also be observed that there is almost no correlation in Page Views Per Visit and Total Visits with Conversion



Step-4: Dealing With Categorical Variables/ Creation of Dummy Variables

- In this step dummy variables are created and a new data frame with the cleaned data and dummy variables is created.

✿ *Data Retained:*

- 98.2% data has been retained after data cleaning.

Building Model (train:test = 70%:30%)

-  *Step-5: Splitting the Data into Training and Testing Sets*
-  *Step-6: Rescaling the features*
 -  *Running First Training Model*
 -  *Feature Selection Using RFE*
 -  *Assessing the model with StatsModels*

➤ 📄 *Step-7: Finalising with the training model*

➤ 📄 *Step-8: Predicting Values based on final model*

➤ 🌸 *Predicted Values on Training data*

➤ 🍉 *Creating new column 'predicted' for predictive values*

➤ 📄 *Step-9: Metrics Evaluation*

➤ 📄 *Step-10: VIF Evaluation*

➤ 📄 *Step-11: Metrics beyond simply accuracy*

➤ ✂ *Sensitivity of the final model is 87.1%*

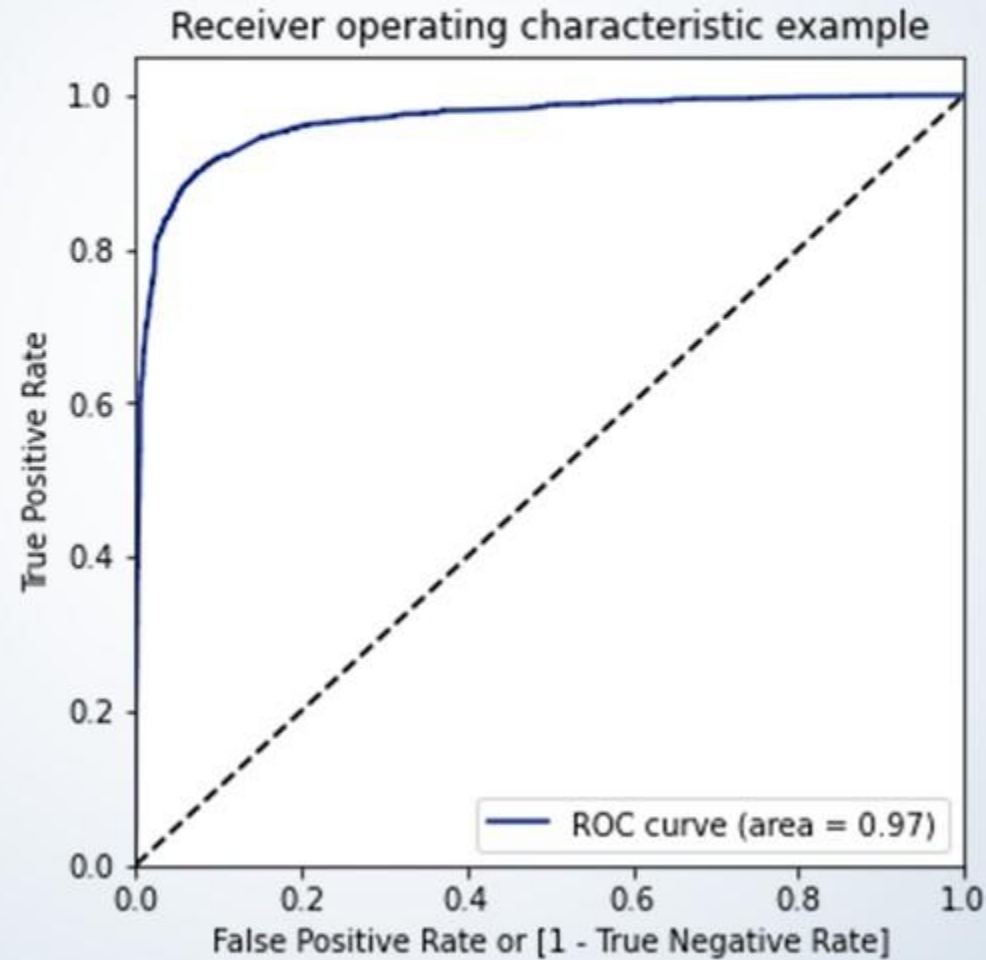
➤ ✂ *Specificity of the final model 94.6%*

➤ ✂ *Predictive churn found is 5.3%*

➤ ✂ *Positive churn is 91.1%*

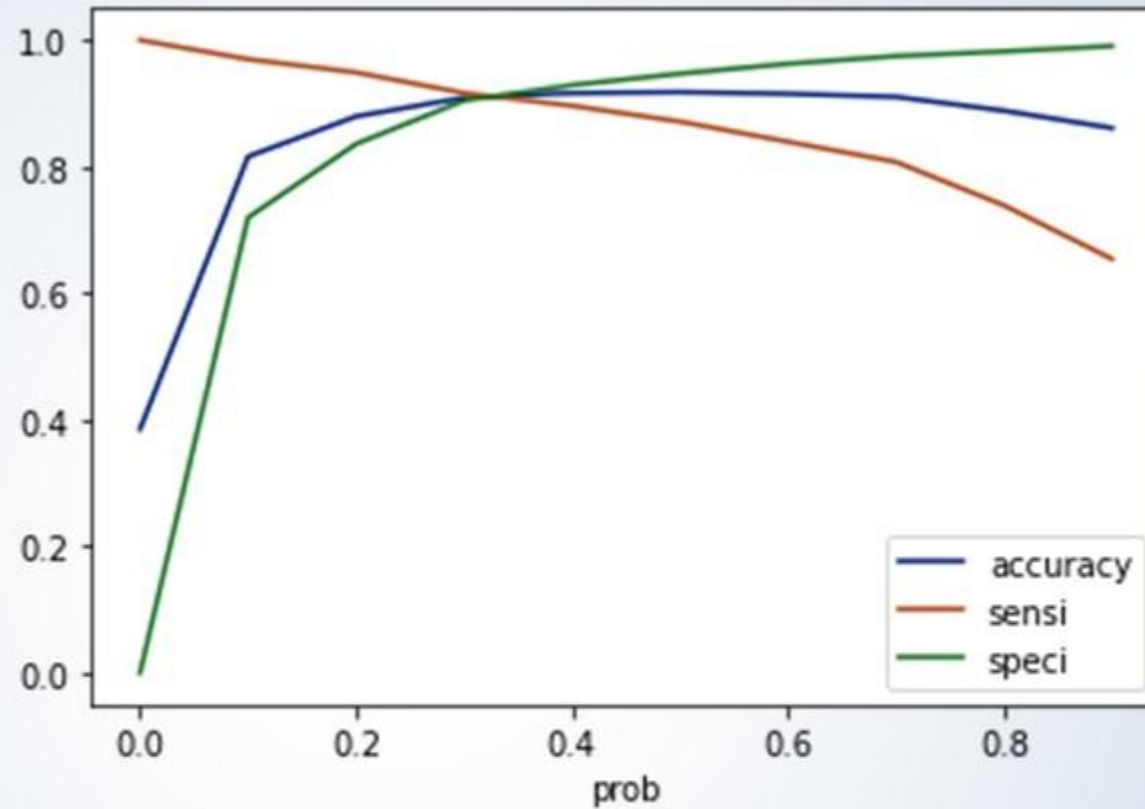
➤ ✂ *Negative churn is 92.1%*

Step-12: Plotting the ROC Curve





Step-13: Finding Optimal Cutoff Point



➡ **Observation:**

➡ From the curve above, 0.25 is the optimum point to take it as a cutoff probability.




Step-14: Making predictions on the test set

Observations:

- After running the model on the **Train Dataset** these are the figures we obtain:
 - Accuracy : 91.7%
 - Sensitivity : 87.1%
 - Specificity : 94.6%
- After running the model on the **Test Dataset** these are the figures we obtain:
 - Accuracy : 89.75%
 - Sensitivity : 91.1%
 - Specificity : 88.9%

LEARNING OUTCOMES OF THE ASSIGNMENT:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 89.75, 91.1 and 88.9% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is above 80%.
- Hence overall this model seems to be good.



❁ ❁ Important features responsible for good conversion rate or the ones which contributes more towards the probability of a lead getting converted are :

- 1. **Lead Origin_Lead Add Form**
- 2. **What is your current occupation_Working Professional**
- 3. **Total Time Spent on Website**

Predictors for HOT LEADS:

Positive Predictors for HOT LEADS:

- A customer with these TAGS assigned is a potential Lead: "Closed by Horizzon", "Lost to EINS", "Will revert after reading the email"
- A customer Lead sourced by "Welingak Website" is a Hot Lead.
- A customer who is currently "Working Professional" or "Unemployed" is a Hot Lead.

Negative Predictors for HOT LEADS

- A customer with these TAGS assigned is NOT a potential Lead: "Already a Student", "switched off", "Not doing further education", "Diploma holder (Not Eligible)", "Ringing", "Interested in other courses", "Interested in full time MBA"
- A customer whose Lead Quality is deemed as "Worst" is also NOT a Hot Lead.

RECOMMENDATIONS:

- It's good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.
- Along with phone calls, it's good to mail the leads also to keep them reminding as email is as powerful as cold calling.
- Reducing the number of call attempts to 2-4 and increasing the frequency of usage of other media like advertisements in Google, or via emails to keep in touch with the lead will save a lot of time.
- Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.

