# Lead Scoring Case Study

By
Himaja Sri K
Sangita Deb
Aishwarya Jagatap

# Problem Statement:

- An education company named X Education sells online courses to industry professionals. The X Education company is looking for the leads who are most likely to get converted into paying customers.

- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

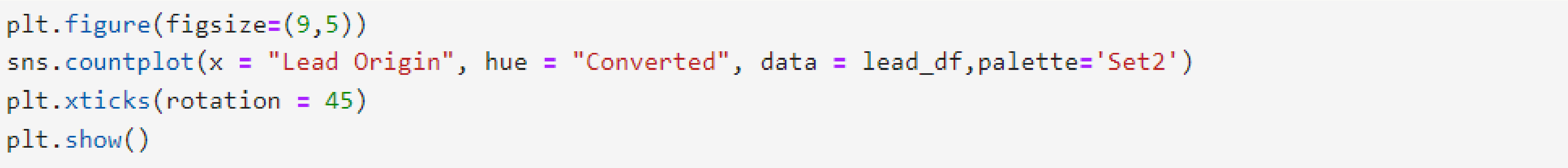
# Processing Steps

- Reading and Understanding the Data
- Data Cleaning
- Data Analysis
- Data Preparation
- ROC Curve
- Optimal Cutoff Value
- Model Evaluation
- Making Prediction’s on Test dataset

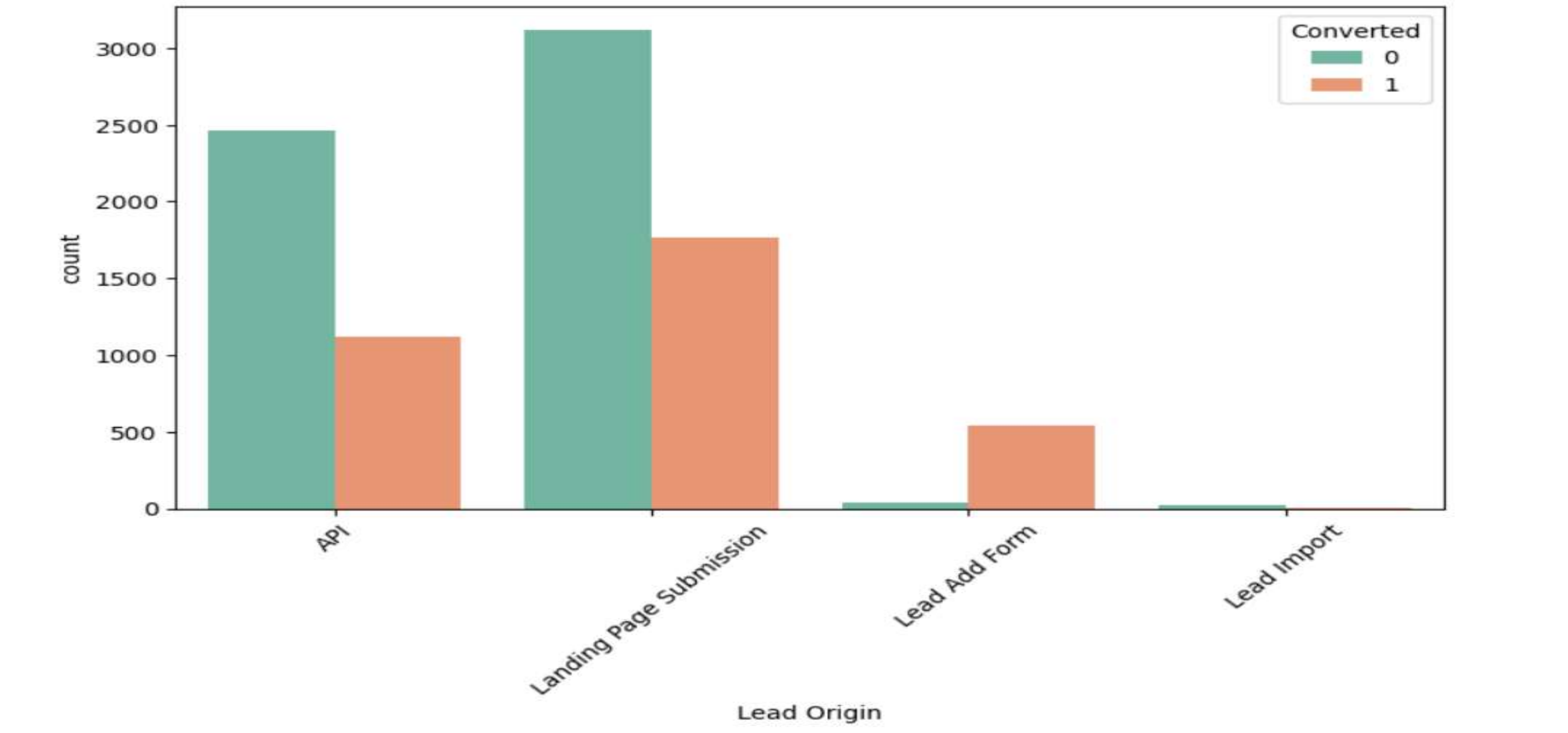
# Reading and Understanding the Data & Data Cleaning

- Read the data using CSV file and Understand the dataset
- Remove all unwanted data from the dataset.
- Identify null values in the data set using isnull() method.
- Calculate the null values percentage using isnull().sum() and len() functions.
- Drop the columns with null values percentage greater than 40%
- We should create dummy variables for categorical varibales.

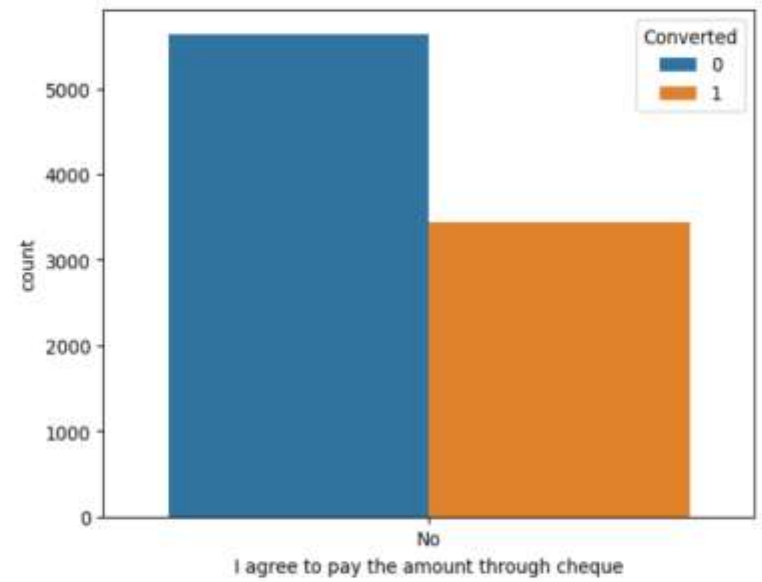
# Data Analysis - Univariate and Bivariate

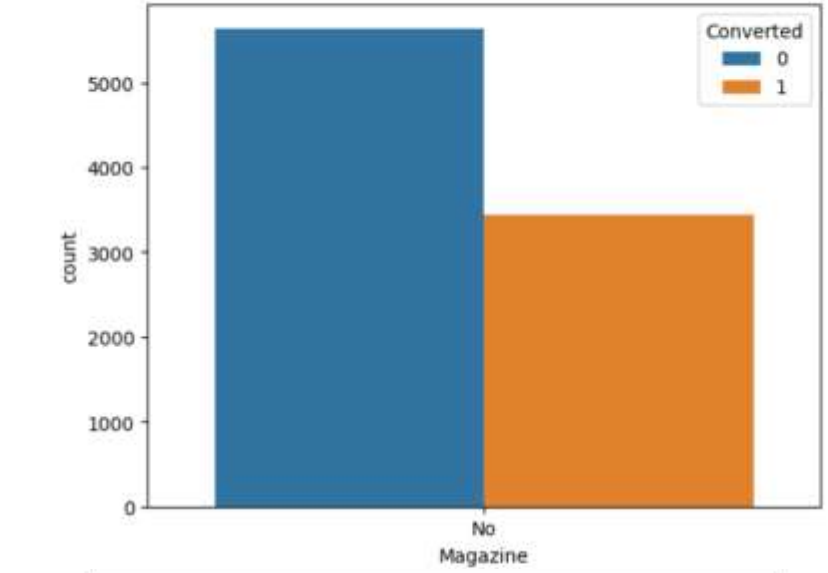


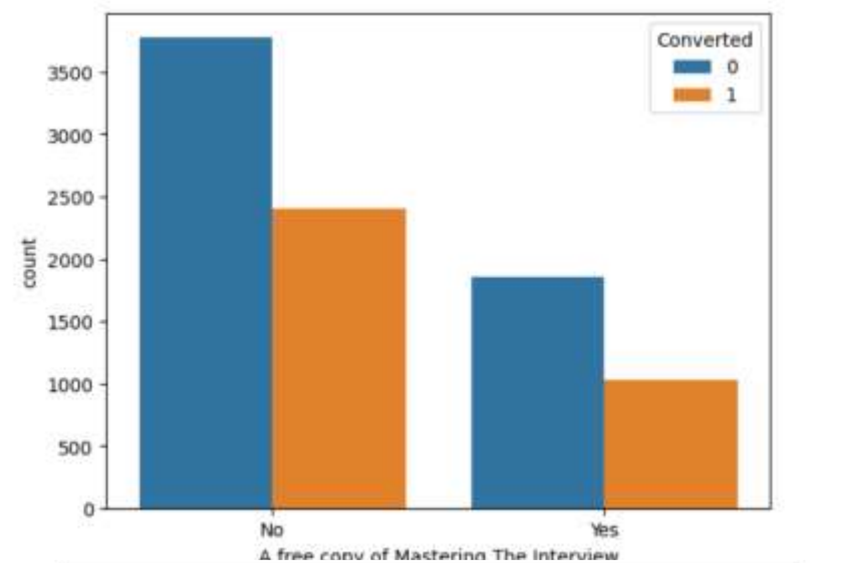
```
plt.figure(figsize=(9,5))
sns.countplot(x = "Lead Origin", hue = "Converted", data = lead_df,palette='Set2')
plt.xticks(rotation = 45)
plt.show()
```

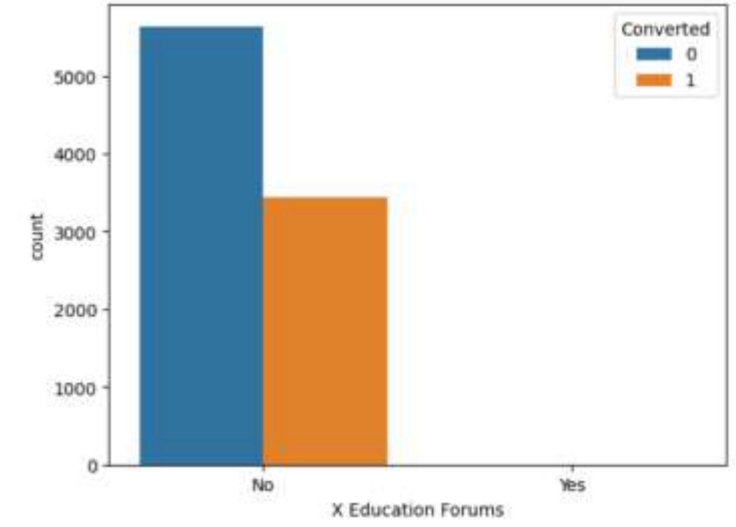


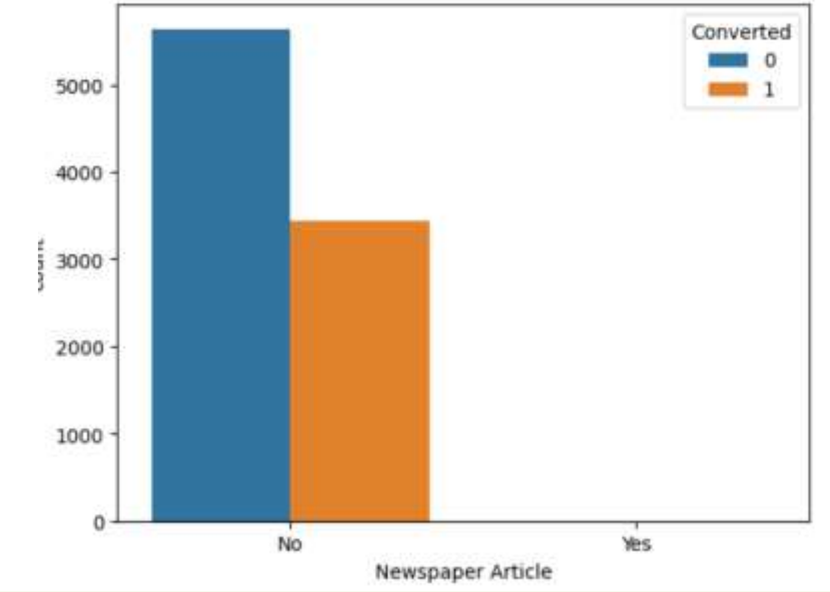
# There is No much impact on conversion rates for these fields.

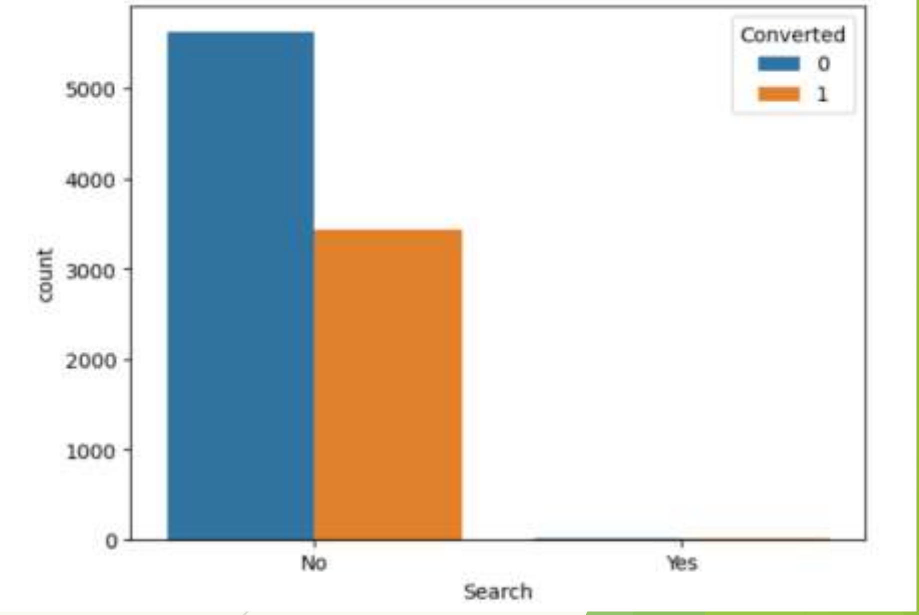




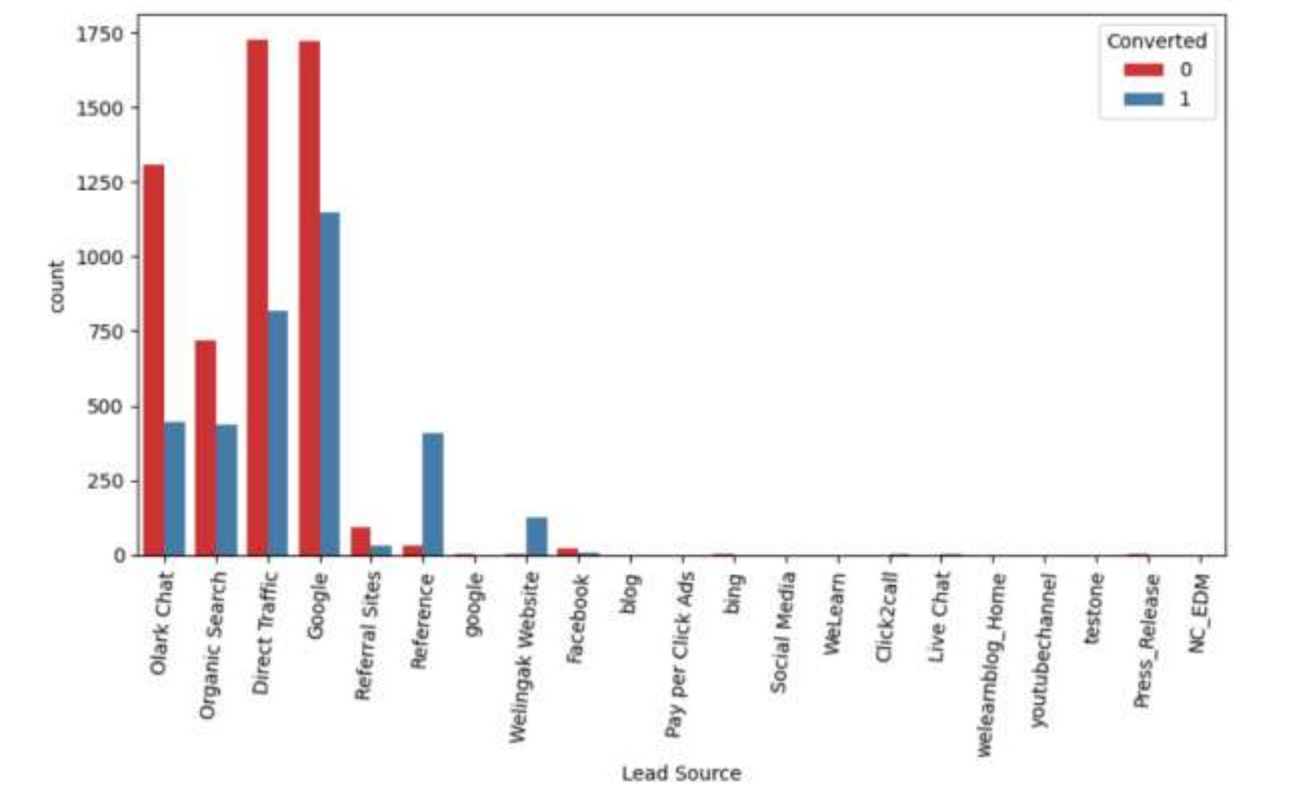


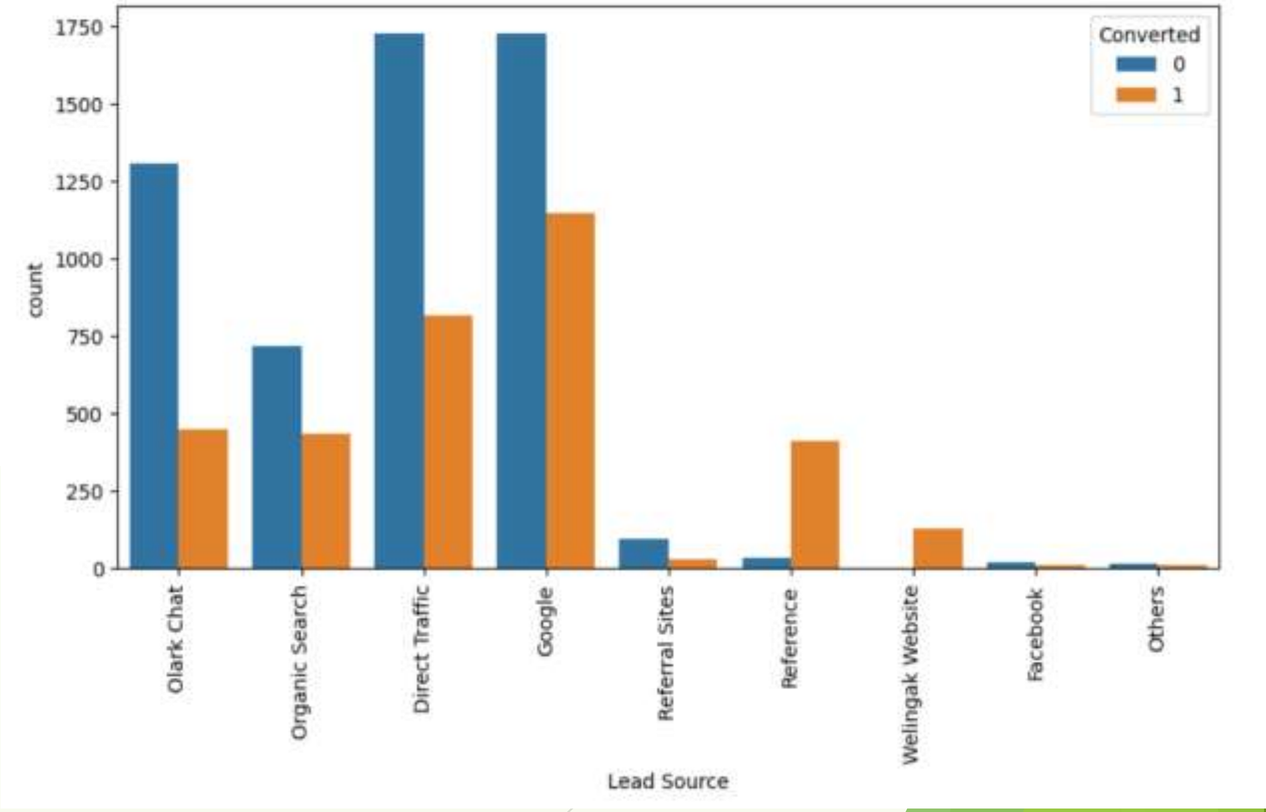




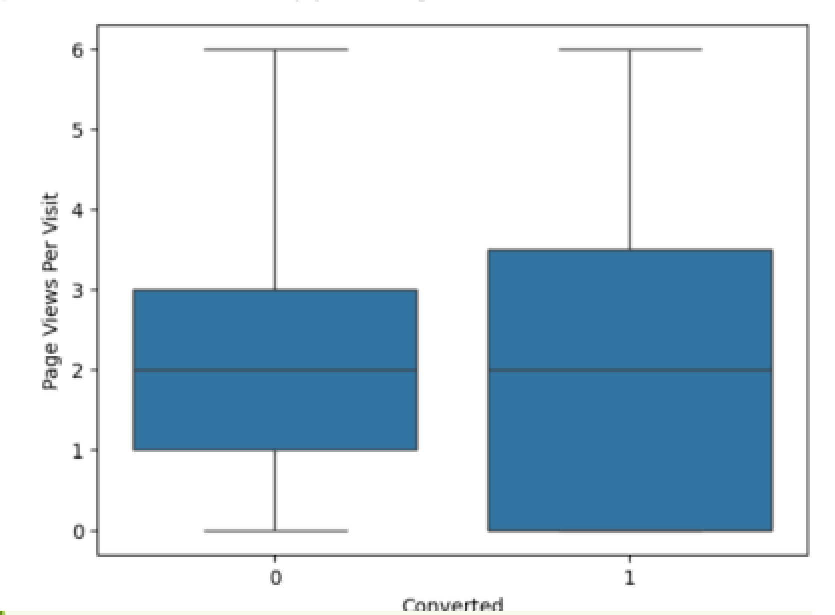


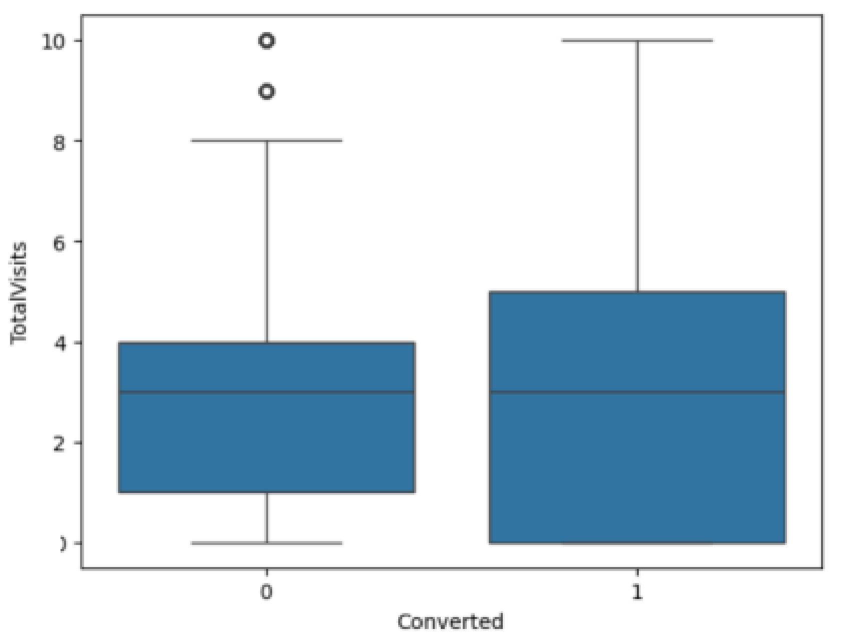
# There are fields with no counts those are removed and names as Others for Lead Source and major conversion in the lead

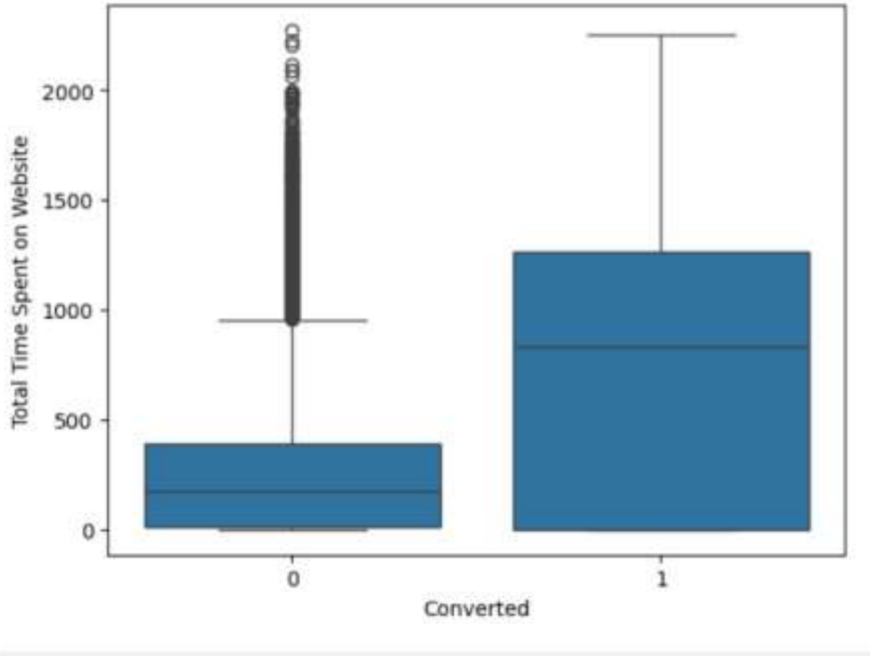




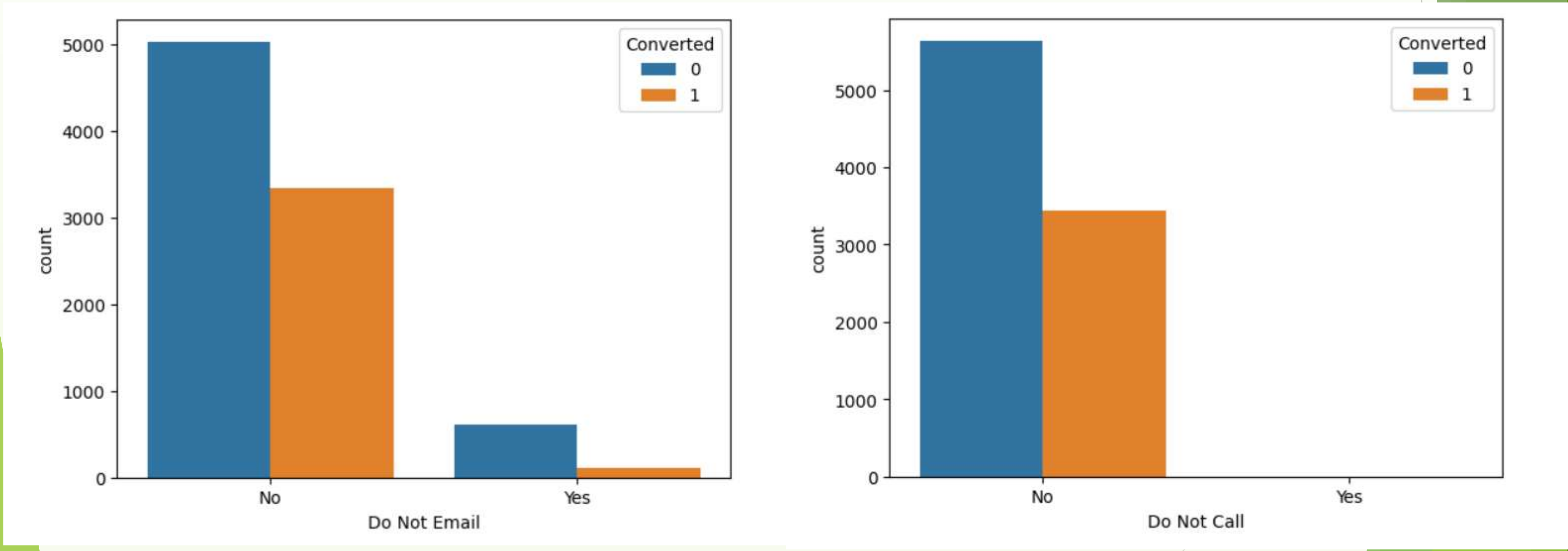
# The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit



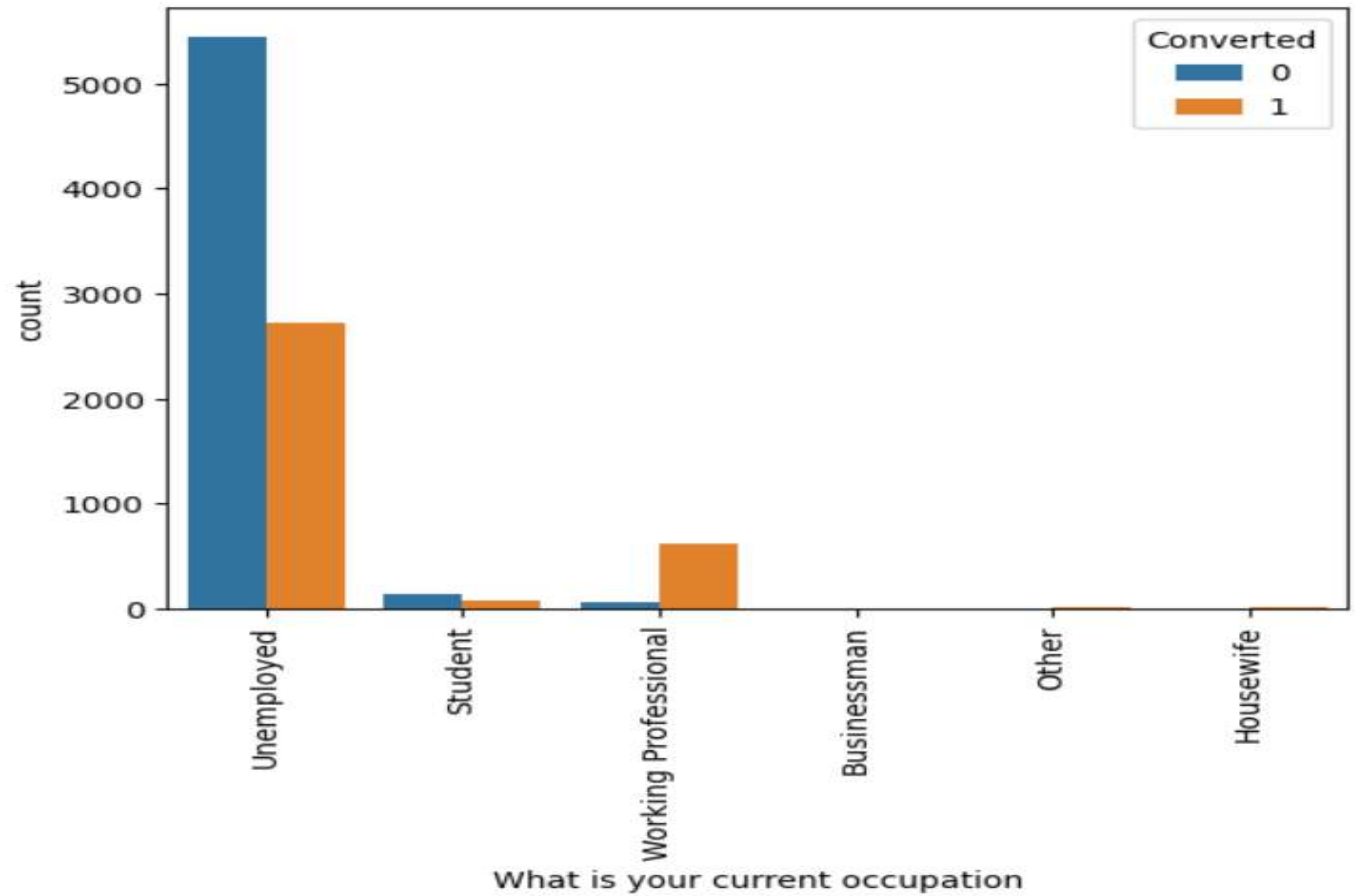




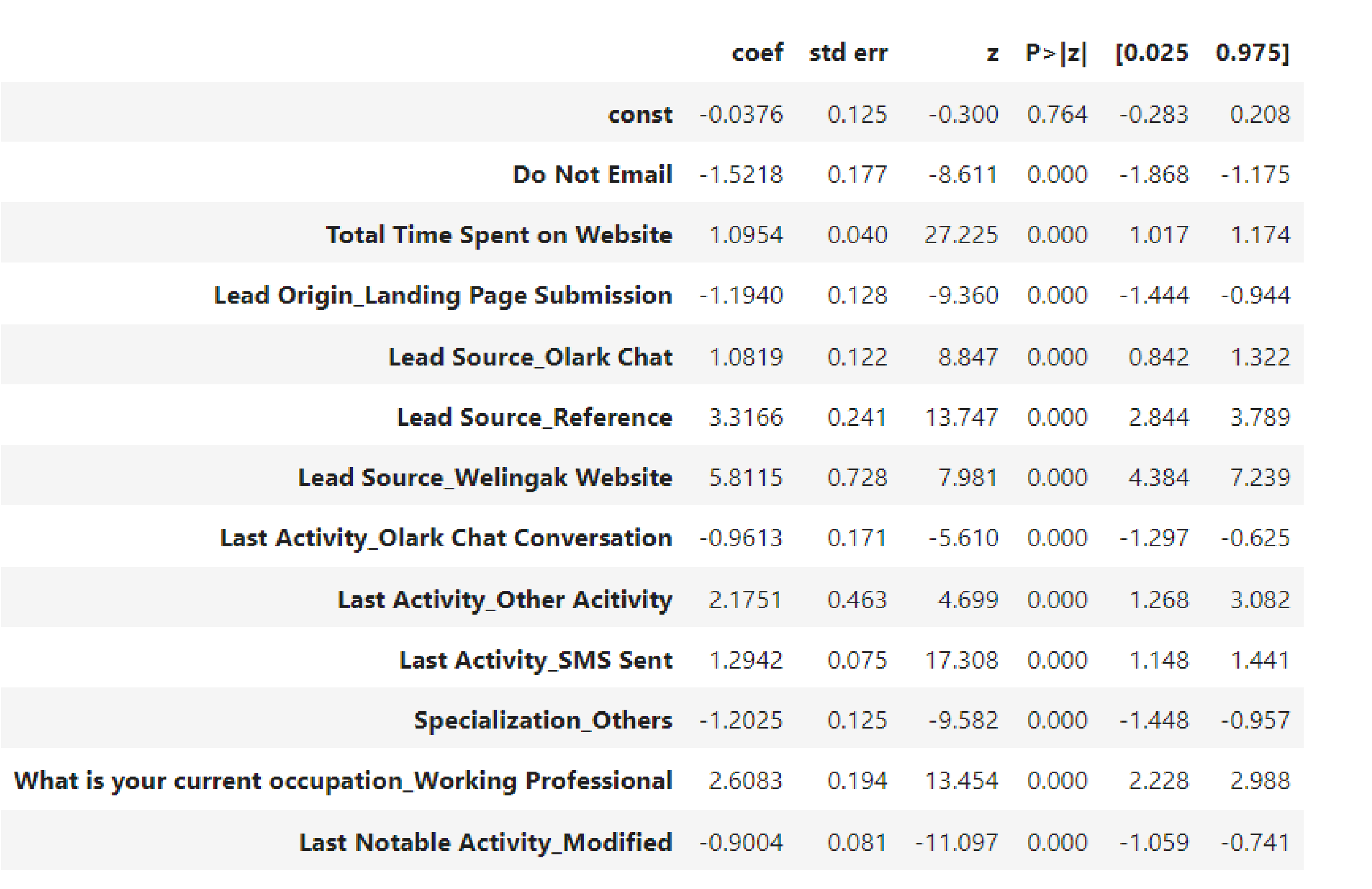
# Major conversion has happened from Emails sent and Calls made.



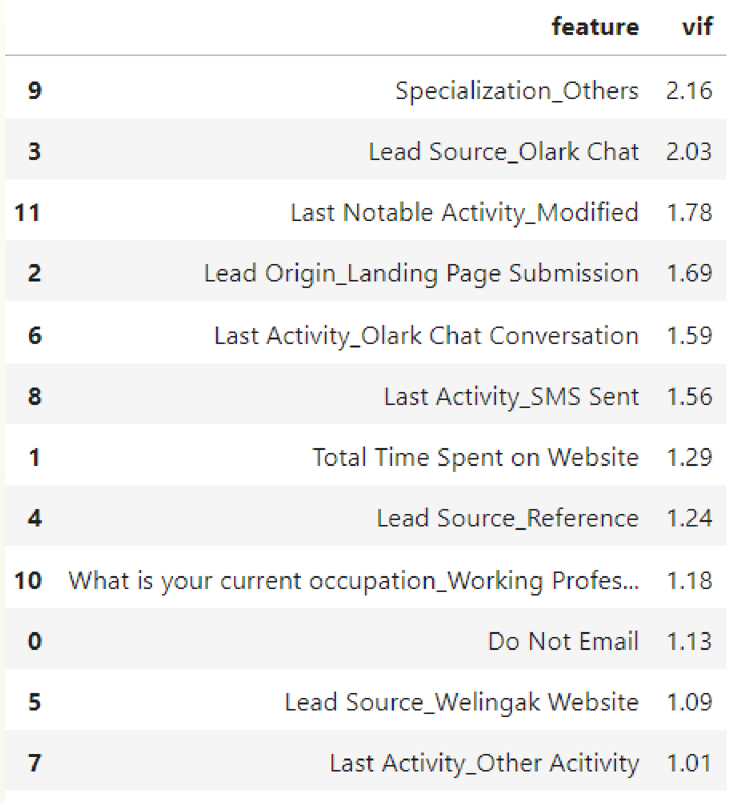
# More conversion happened with people who are unemployed, followed by Working Professionals.



# Final model - P-values & VIF

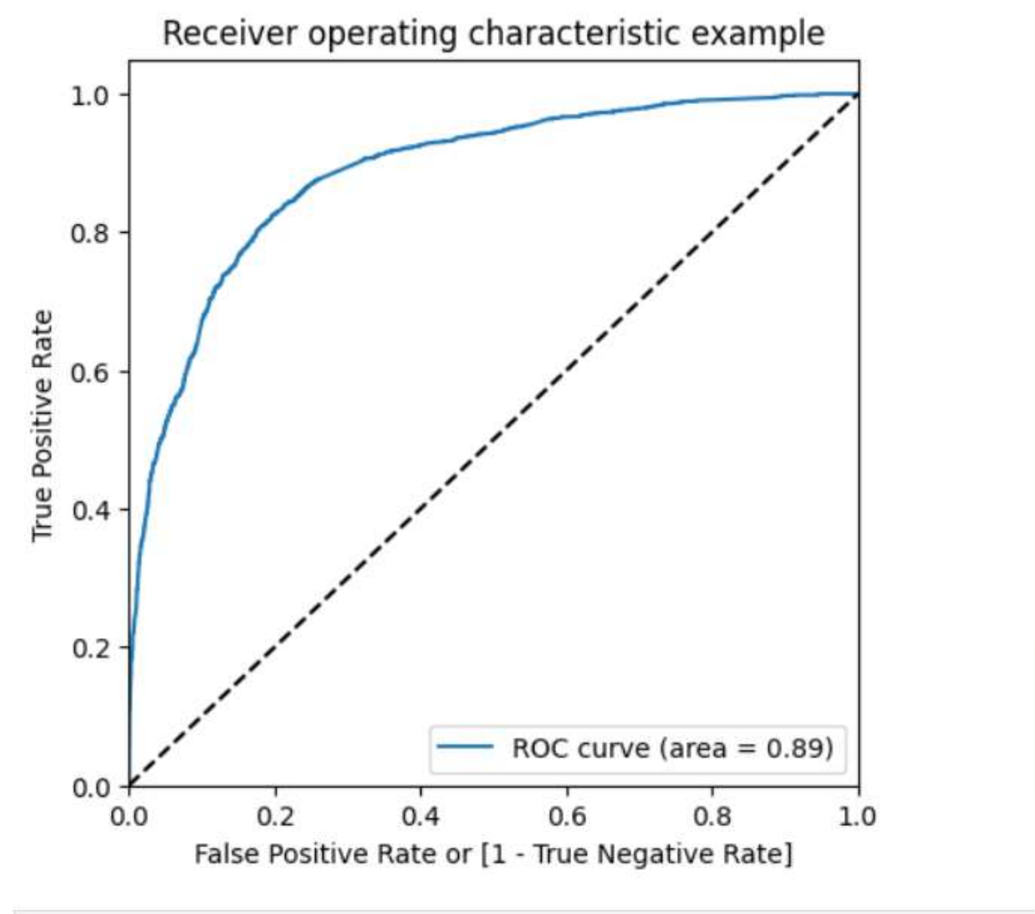


| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -0.0376 | 0.125 | -0.300 | 0.764 | -0.283 | 0.208 |
| **Do Not Email** | -1.5218 | 0.177 | -8.611 | 0.000 | -1.868 | -1.175 |
| **Total Time Spent on Website** | 1.0954 | 0.040 | 27.225 | 0.000 | 1.017 | 1.174 |
| **Lead Origin_Landing Page Submission** | -1.1940 | 0.128 | -9.360 | 0.000 | -1.444 | -0.944 |
| **Lead Source_Olark Chat** | 1.0819 | 0.122 | 8.847 | 0.000 | 0.842 | 1.322 |
| **Lead Source_Reference** | 3.3166 | 0.241 | 13.747 | 0.000 | 2.844 | 3.789 |
| **Lead Source_Welingak Website** | 5.8115 | 0.728 | 7.981 | 0.000 | 4.384 | 7.239 |
| **Last Activity_Olark Chat Conversation** | -0.9613 | 0.171 | -5.610 | 0.000 | -1.297 | -0.625 |
| **Last Activity_Other Acitivity** | 2.1751 | 0.463 | 4.699 | 0.000 | 1.268 | 3.082 |
| **Last Activity_SMS Sent** | 1.2942 | 0.075 | 17.308 | 0.000 | 1.148 | 1.441 |
| **Specialization_Others** | -1.2025 | 0.125 | -9.582 | 0.000 | -1.448 | -0.957 |
| **What is your current occupation_Working Professional** | 2.6083 | 0.194 | 13.454 | 0.000 | 2.228 | 2.988 |
| **Last Notable Activity_Modified** | -0.9004 | 0.081 | -11.097 | 0.000 | -1.059 | -0.741 |



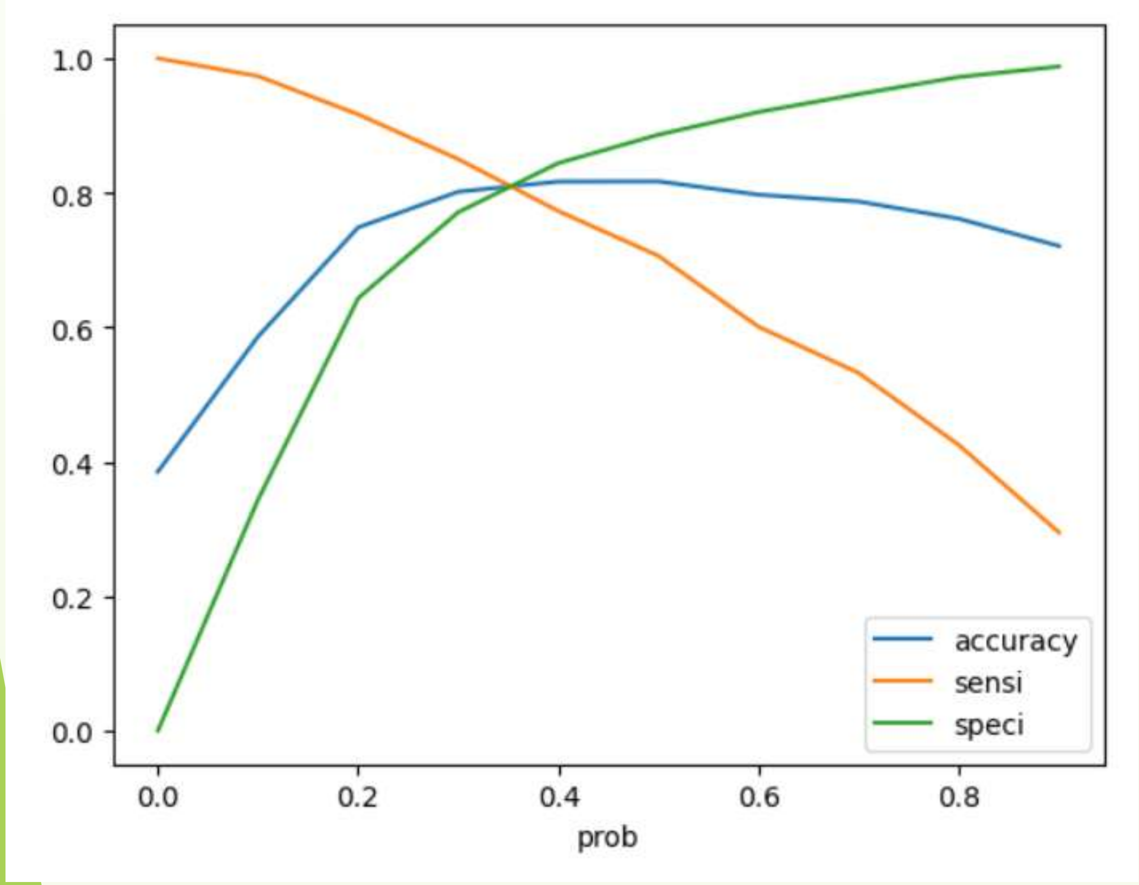
| | feature | vif |
|---|---|---|
| **9** | Specialization_Others | 2.16 |
| **3** | Lead Source_Olark Chat | 2.03 |
| **11** | Last Notable Activity_Modified | 1.78 |
| **2** | Lead Origin_Landing Page Submission | 1.69 |
| **6** | Last Activity_Olark Chat Conversation | 1.59 |
| **8** | Last Activity_SMS Sent | 1.56 |
| **1** | Total Time Spent on Website | 1.29 |
| **4** | Lead Source_Reference | 1.24 |
| **10** | What is your current occupation_Working Profes... | 1.18 |
| **0** | Do Not Email | 1.13 |
| **5** | Lead Source_Welingak Website | 1.09 |
| **7** | Last Activity_Other Acitivity | 1.01 |

**ROC Curve:** Then we plotted the ROC curve for the features and the curve came out be decent with an area coverage of 89%(0.89)



# Model Evaluation - Sensitivity and Specificity on Train Data Set.

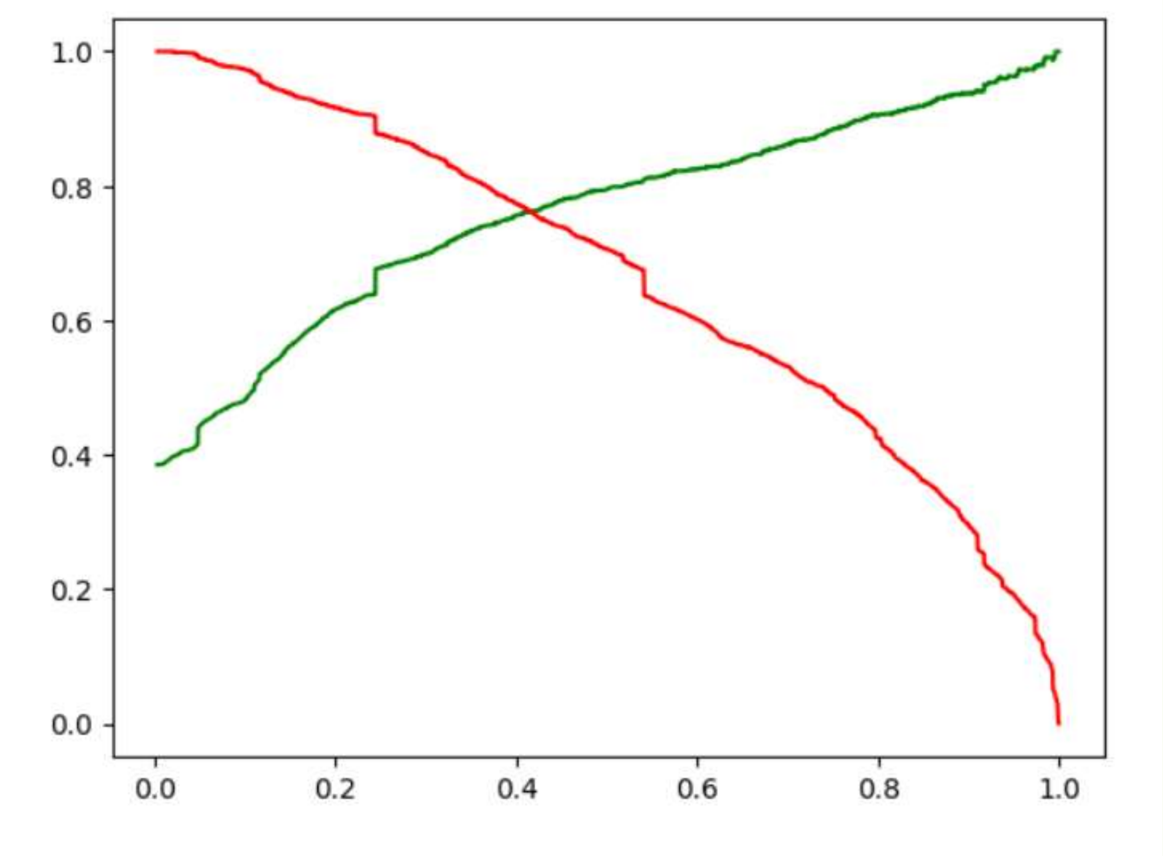
- The graph depicts an optimal cut off of 0.36 based on Accuracy,
- Sensitivity and Specificity



We Observed:

- accuracy=81%,
- sensitivity=80.4% (0.804),
- specificity=82.1% (0.821)

# The graph depicts an optimal cut off of 0.42 based on Precision and Recall



- Precision = 73.8%(0.738)
- Recall = 80.4%(0.804)

# Making Prediction’s on Test dataset:

- Sensitivity and Specificity for calculating the final prediction. –
- Then we started the learning's to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.0%(0.809), Sensitivity=79.4%(0.794) and Specificity= 81.8%(0.808).
- □ The top 3 variables that contribute for lead getting converted in the model are
- □ lead sources "Welingak Websites" & "Reference" & "Olark Chat"
- □ "Working professionals”
- □ “More time on the websites”
- □ Hence overall this model seems to be good.

# THANK YOU