WELCOME

INTRODUCTION

Problem Statement:

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these 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.

Business Objectives:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology

- Data cleaning and data manipulation.

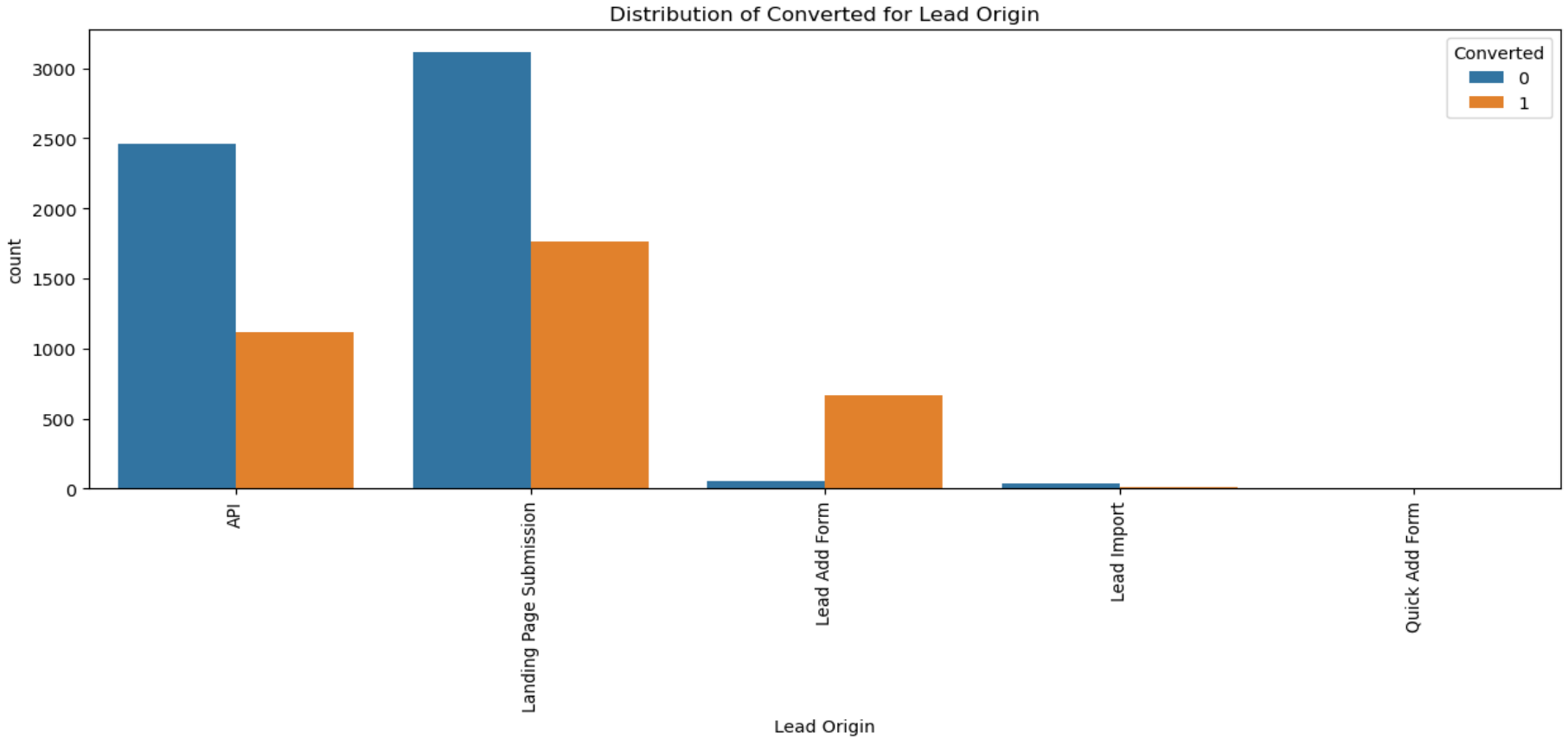
1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns.
4. Imputation of the values.
5. Check and handle outliers in data.

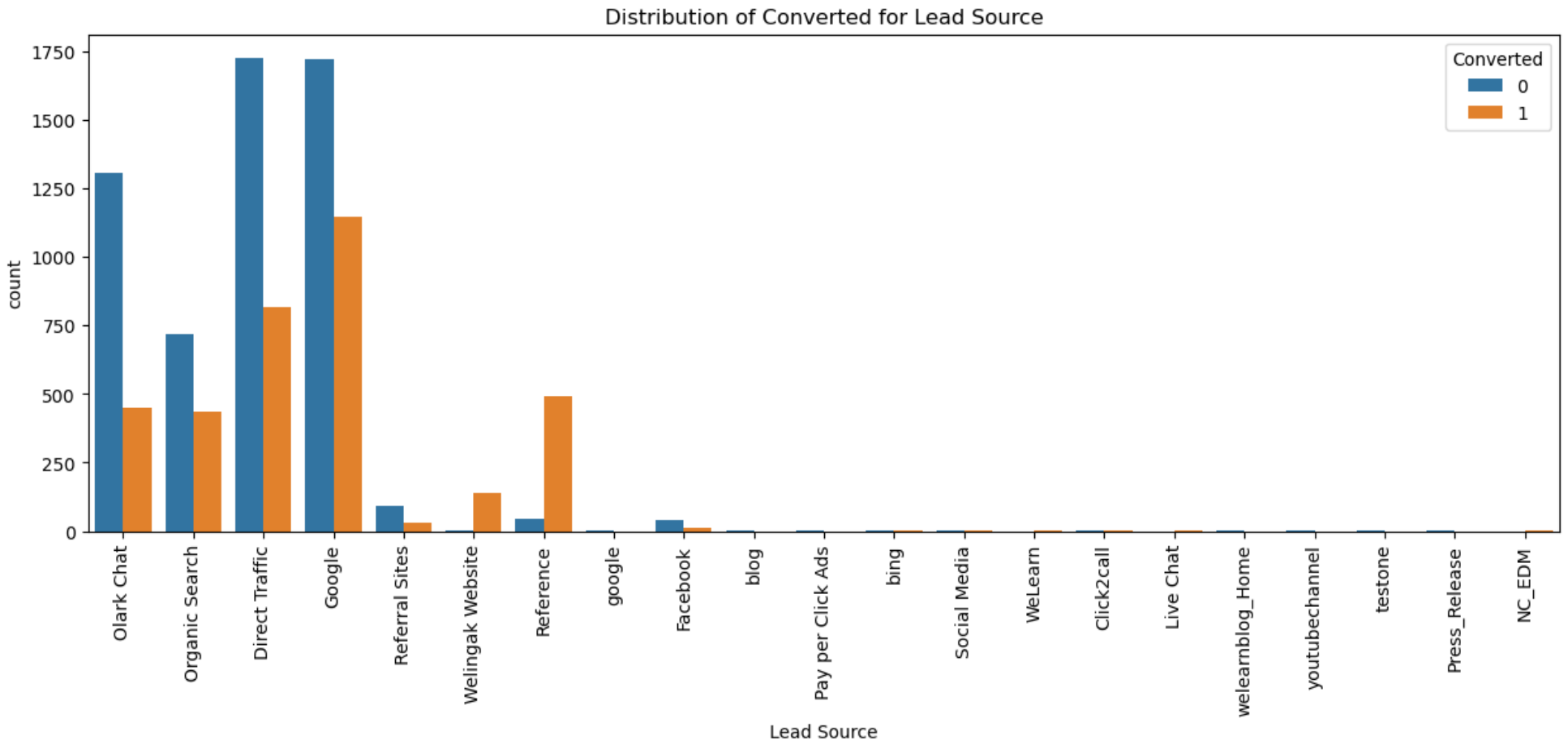
- EDA
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.

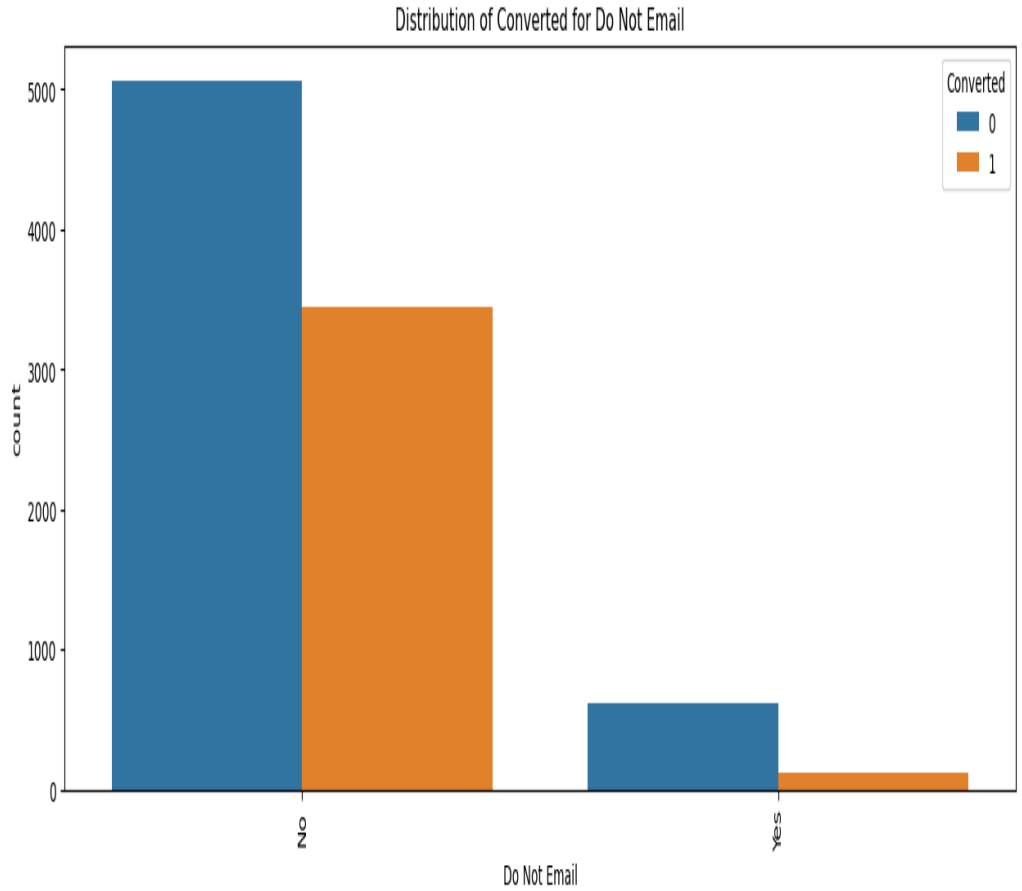
Data cleaning and data manipulation

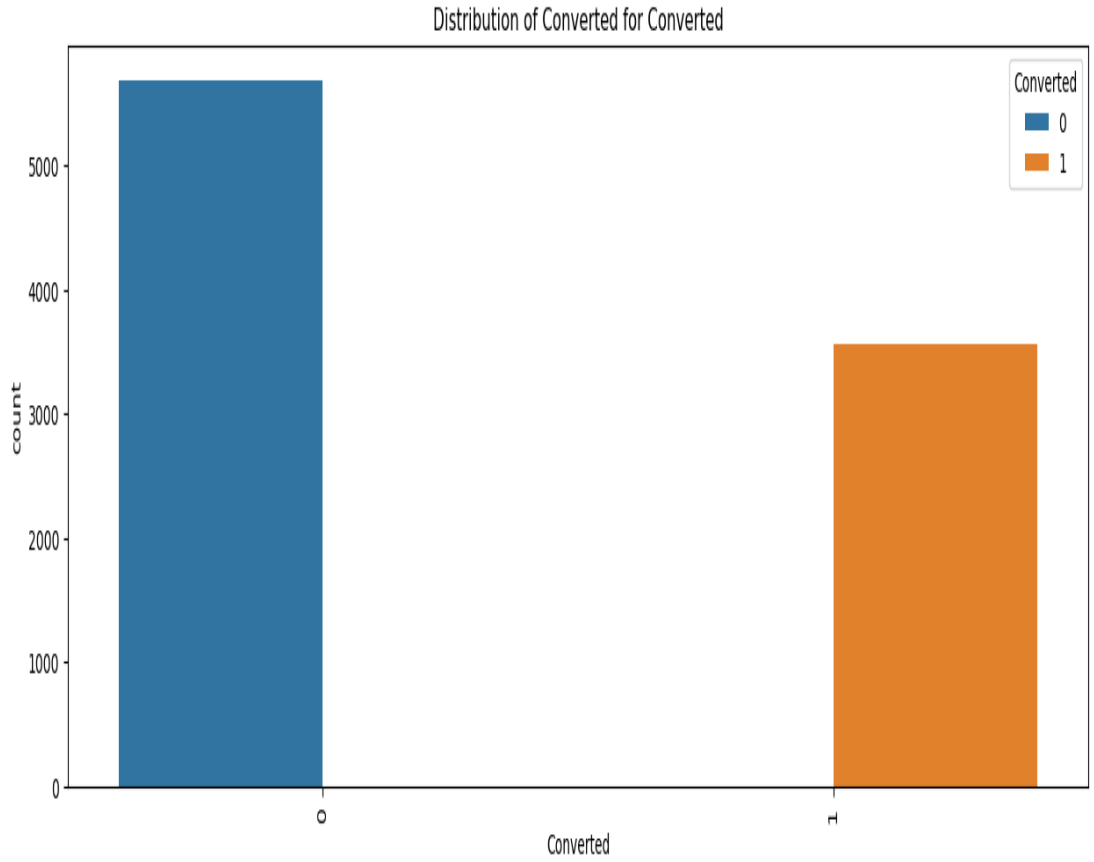
- Raw data is having 9240 rows and 37 columns.
- Raw data is having 39% conversion rate.
- Out of 37 columns 16 columns having missing values.
- There are lots of columns which are having ‘select’ as value probably due to columns are not marked as mandatory.
- There are no duplicates in the data.
- 'Prospect ID', 'Lead Number‘ are dropped as this are unique value and not required for our analysis.
- Columns(Lead_Quality,Asymmetrique_Activity_Index,Asymmetrique_Profile_Score,Asymmetrique_Activity_Score,Asymmetrique_Profile_Index) which are having more the 40% missing values are dropped.

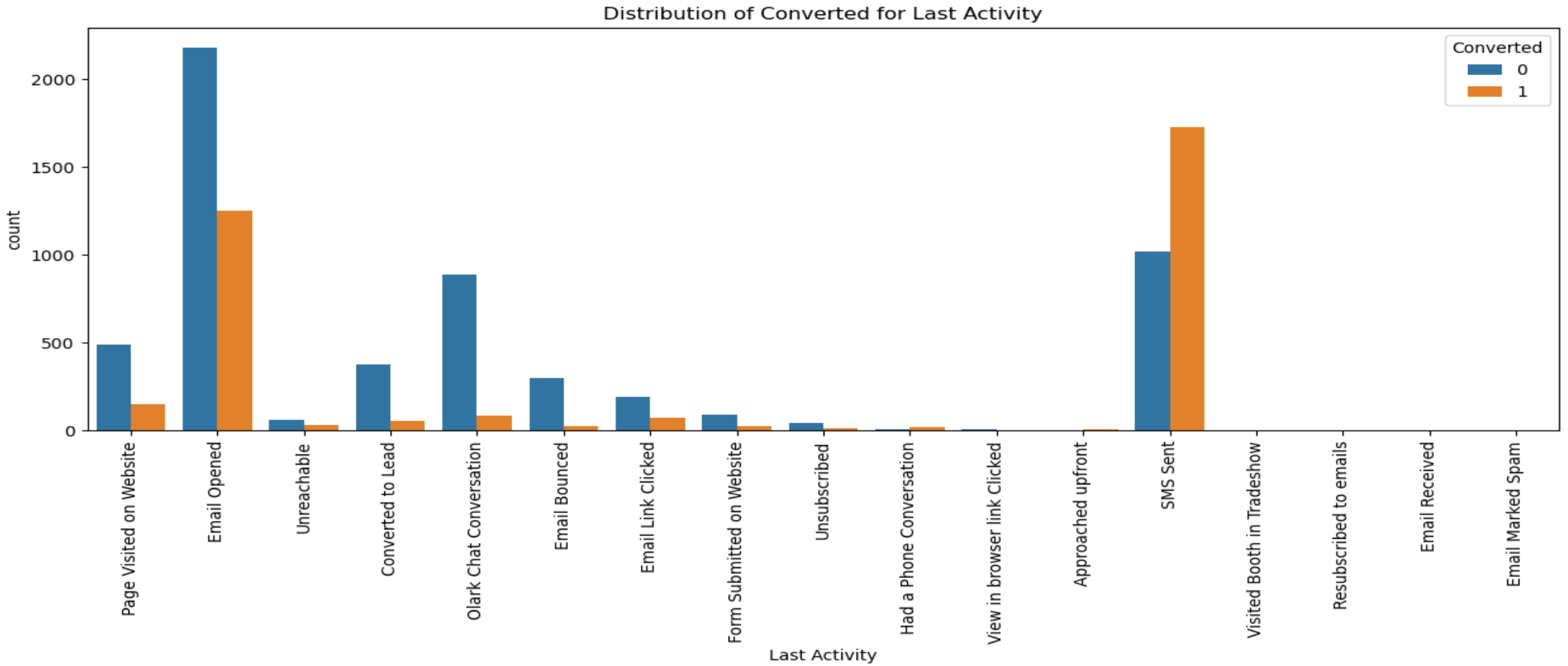
EDA

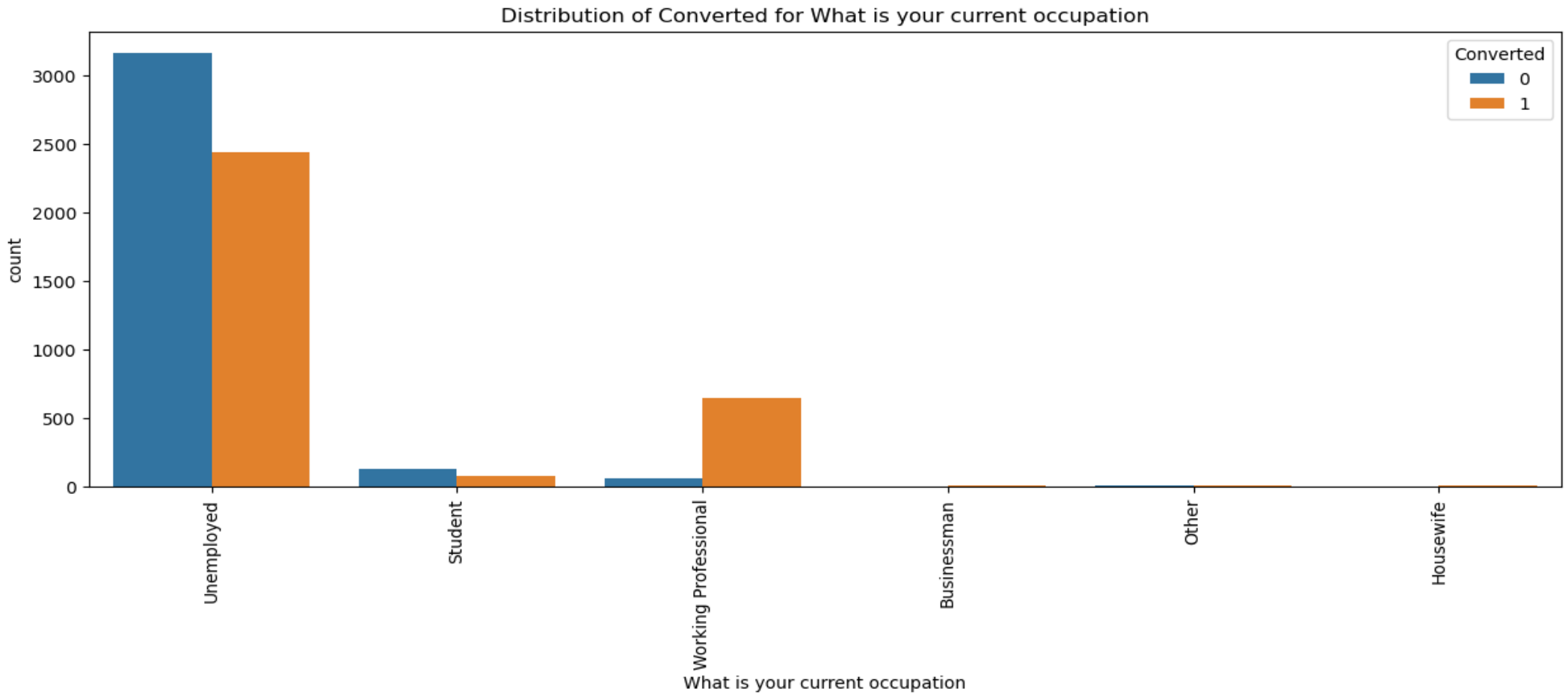

Distribution of Converted for Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

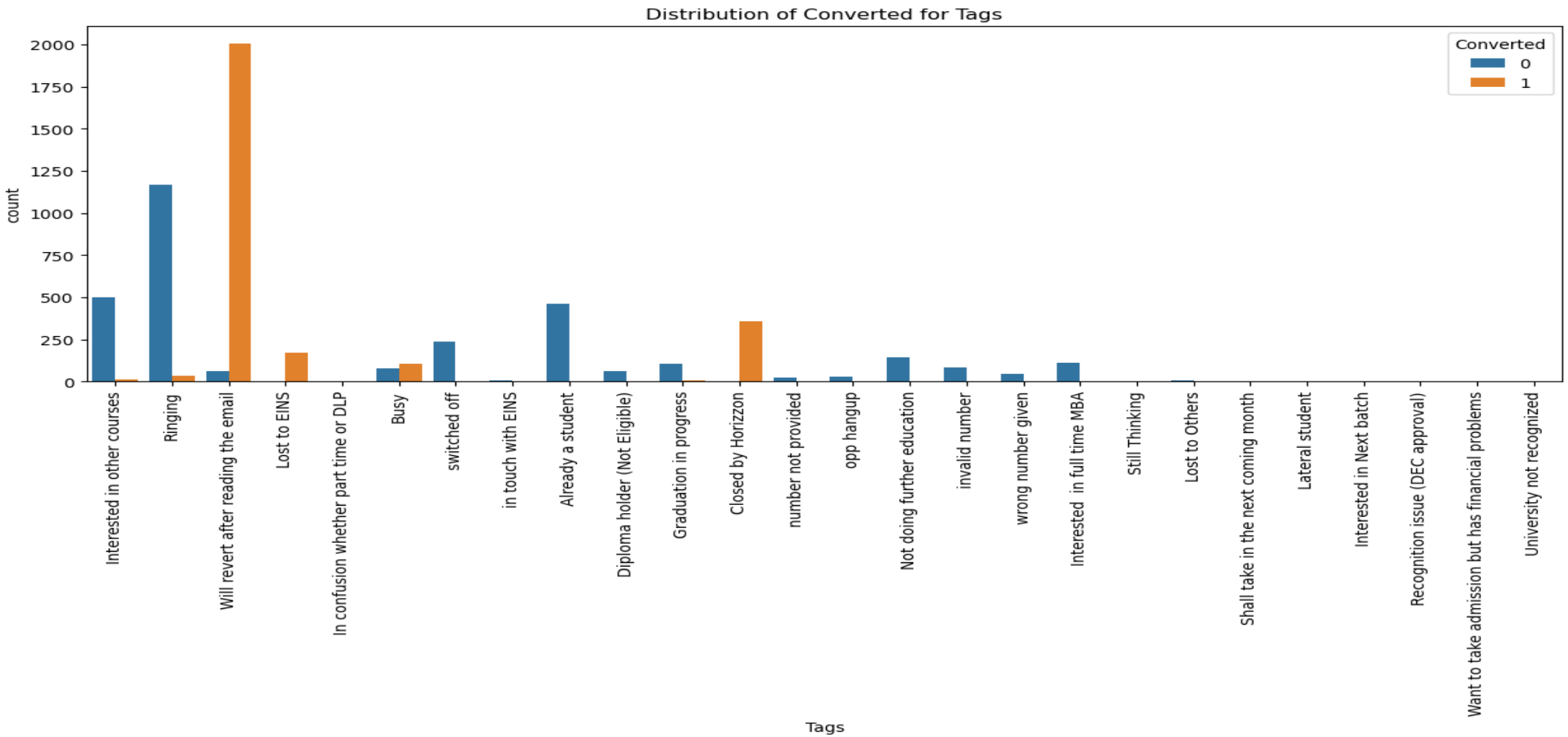

Distribution of Converted for Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


Distribution of Converted for Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

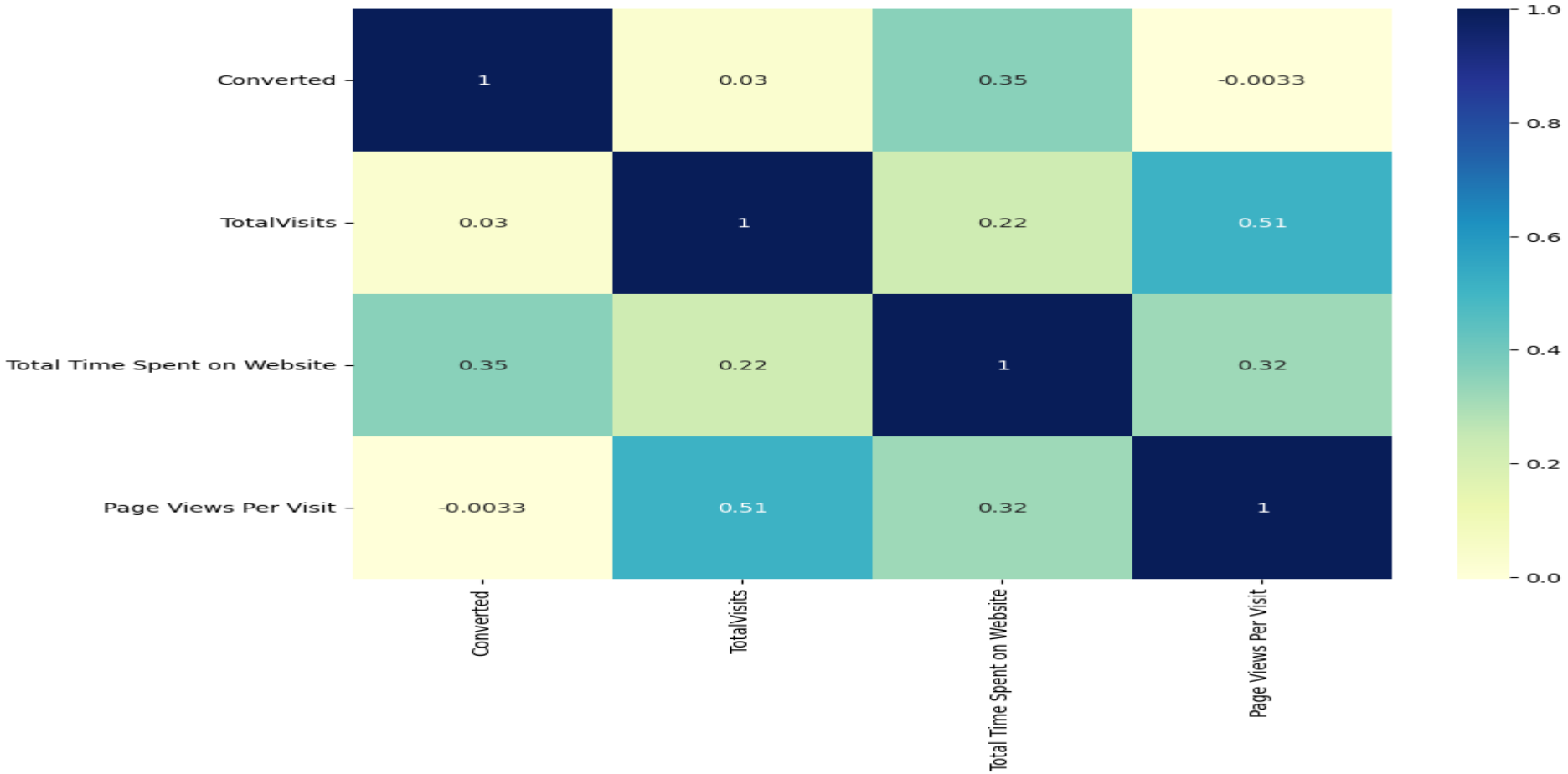

Distribution of Converted for Converted
Converted
0
1
count
5000
4000
3000
2000
1000
0
0
1
Converted


Distribution of Converted for Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

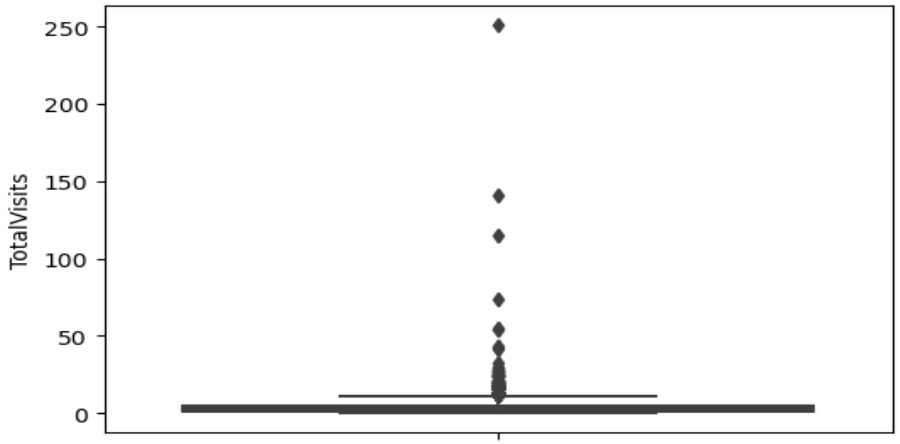

Distribution of Converted for What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

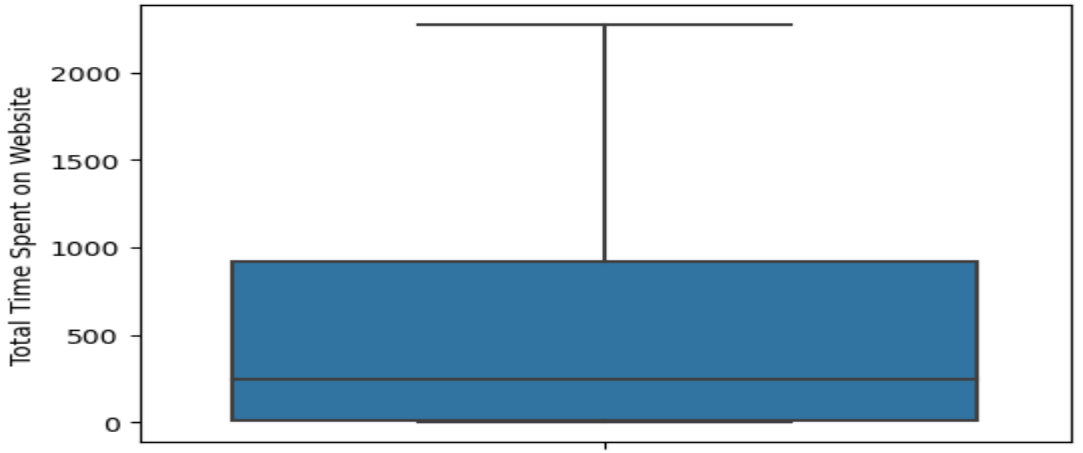

Distribution of Converted for Tags
Converted
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
Closed by Horizzon
number not provided
opp hangup
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags

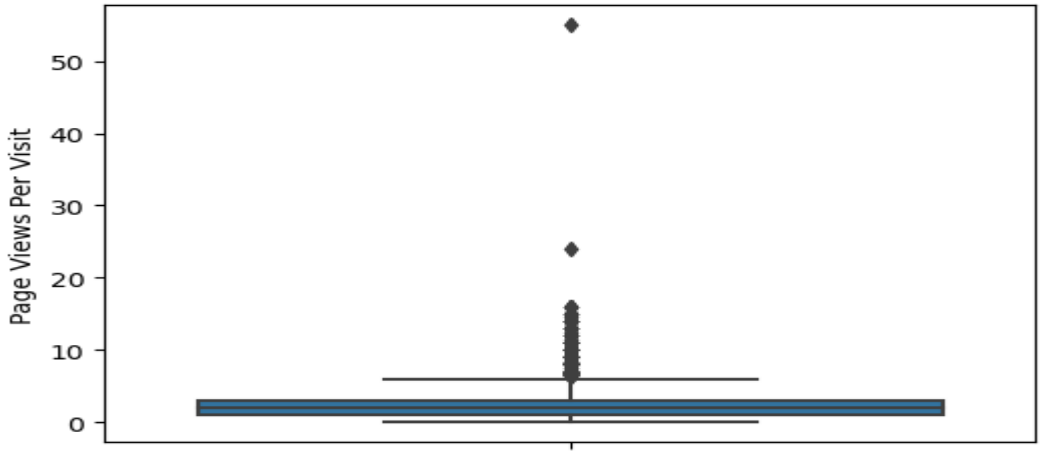
Correlation Matrix For Numerical Variables



Numerical Outlier Checking







Feature Scaling & Dummy Variables and encoding of the data.

- All numerical outlier variables are normalized.
- Dummy Variables for binary columns and categorical columns are created.
- Total no of columns 50 and total no of rows 8953 are created for Logistic regression model building and analysis

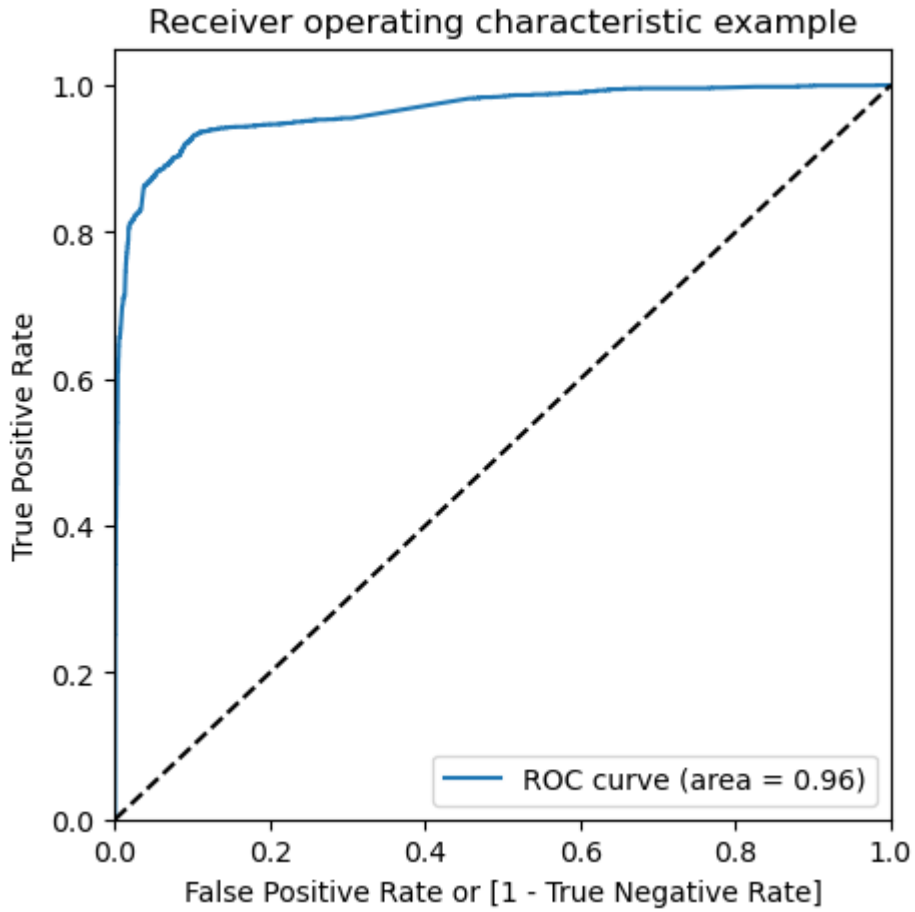
Classification technique: logistic regression used for the model making and prediction

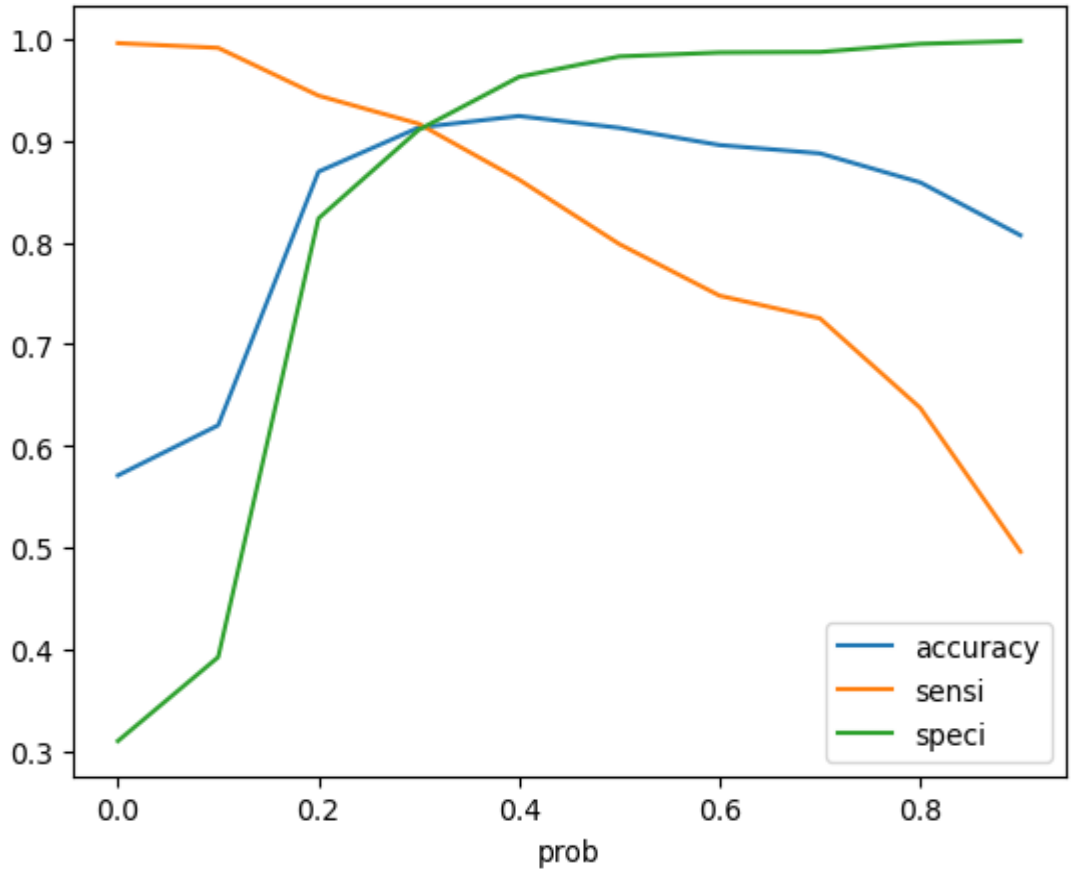
- Splitting the data in train and test for 70% and 30% ratio.
- Scaling and fitting the train data set using SKLEARN MinMaxScaler.
- Top 15 important feature selection using RFE.
- Automation for Stats model and VIF checking.
- Elimination of features which are having VIF >5 and P value >0.05.
- Making Prediction on train data set.
- Achieved accuracy of 91.28% on trained data.
- 91.65% Specificity : 91.06%
- Optimal cut off value found as 0.3

Validation of the model

- Achieved overall accuracy of 92.33%.
- Achieved Sensitivity : 91.49% Specificity : 92.84% on test data.
- Precision 88.50% achieved.
- Recall 91.48% achieved.
- Point where Accuracy, Sensitivity, Specificity intersects each other is the point for choosing optimal cutoff point.
- Area of ROC curve found 0.9 which is very nearly to 1.

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.96)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.2
0.4
0.6
0.8

Summary

- Our final model performance was attained using accuracy, sensitivity, specificity and ROC-AUC scores. The model demonstrated a good ability to differentiate between converting and non_converting leads, with sensitivity being particularly important given the business goal of identifying hot leads.
- We also observed that the Total Time Spent on Website, Lead Origin_Landing Page Submission and Last Activity_SMS Sent were the top predictors of conversion suggesting that engagement metrics are key drivers of customer interest.
- One key learning from this assignment was the importance of thorough data cleaning and preparation, especially handling missing values and redundant categories in categorical variables. Proper feature engineering significantly impacts model performance by reducing noise and focusing on the most relevant information.