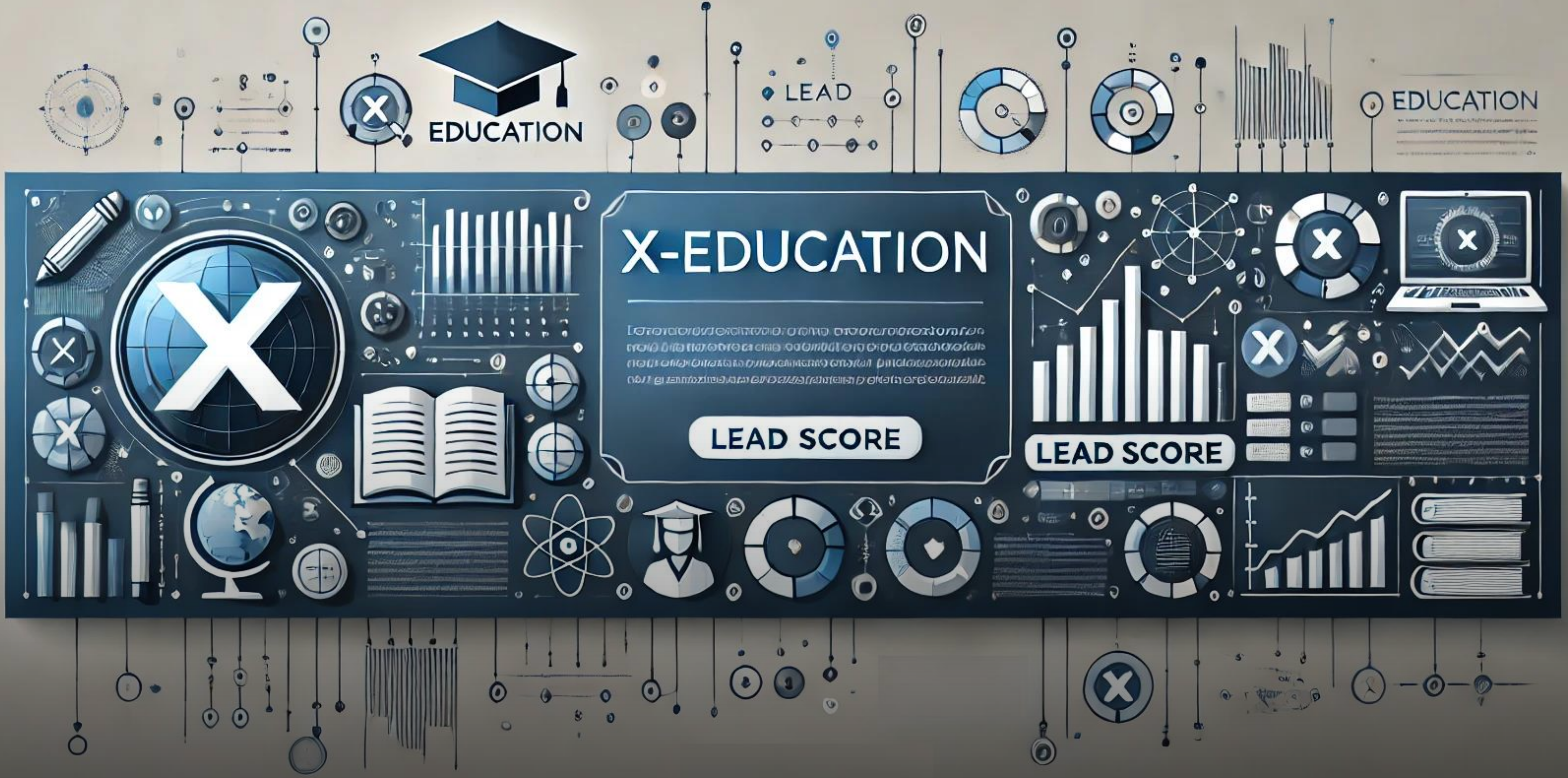
Lead Scoring Case Study



Rohit Vashishth

Rakesh Kumar Sahoo

Problem & Goal Statements

Problem Statement

An education company named X Education sells online courses to industry professionals on their website. 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. 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. The typical lead conversion rate at X education is around 30%, which is very poor.

To make this process more efficient, 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 as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goal Statement

The company requires us to build a Logistic Regression model wherein we need to assign a lead score (between 0 to 100) 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%.

Approach & Methodologies

Approach

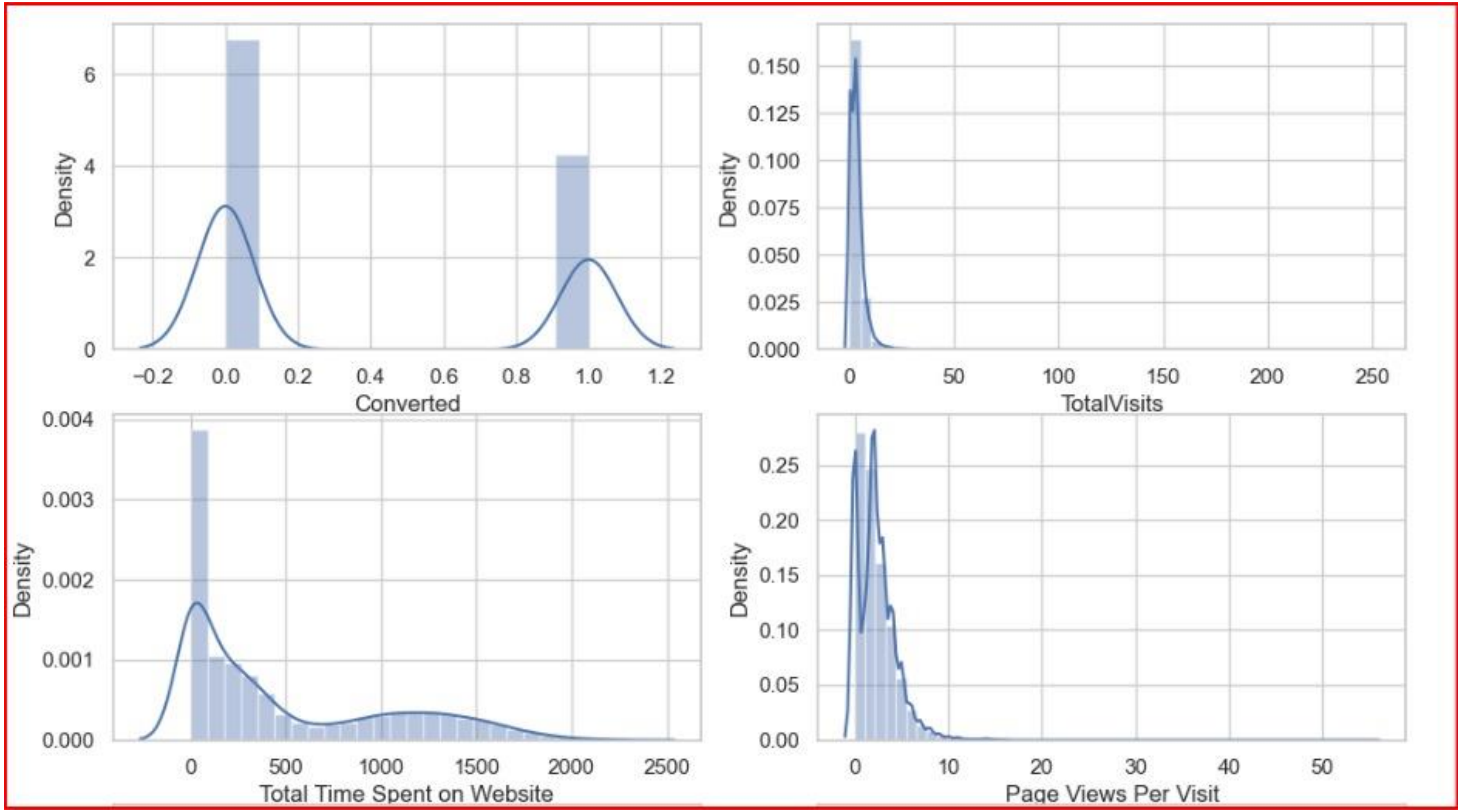
- Understand the domain variables
- Understand the data structure / metadata
- Handle missing, inconsistent, outlier values
- Exploratory Data Analysis (EDA)
- Data pre-processing for Model building
- Model Building
- Model Evaluation
- Generate Lead scores (Test data)

Method

- Remove variables that have >= 30% missing values
- Data imputation using Mean and Mode values
- Create a deep copy of primary dataframe for visualization
- Variable encoding – Yes/No to 1/0 and One hot encoding
- Normalization of Test and Train datasets
- Usage of RFE for automatic variable reduction
- Statsmodels and VIF used for model training as well as manual variable reduction
- Predict probabilities and evaluate optimal cut off for probability value to find Hot leads
- Model Evaluation through Recall (80% target provided by CEO)

Graphs & Insights

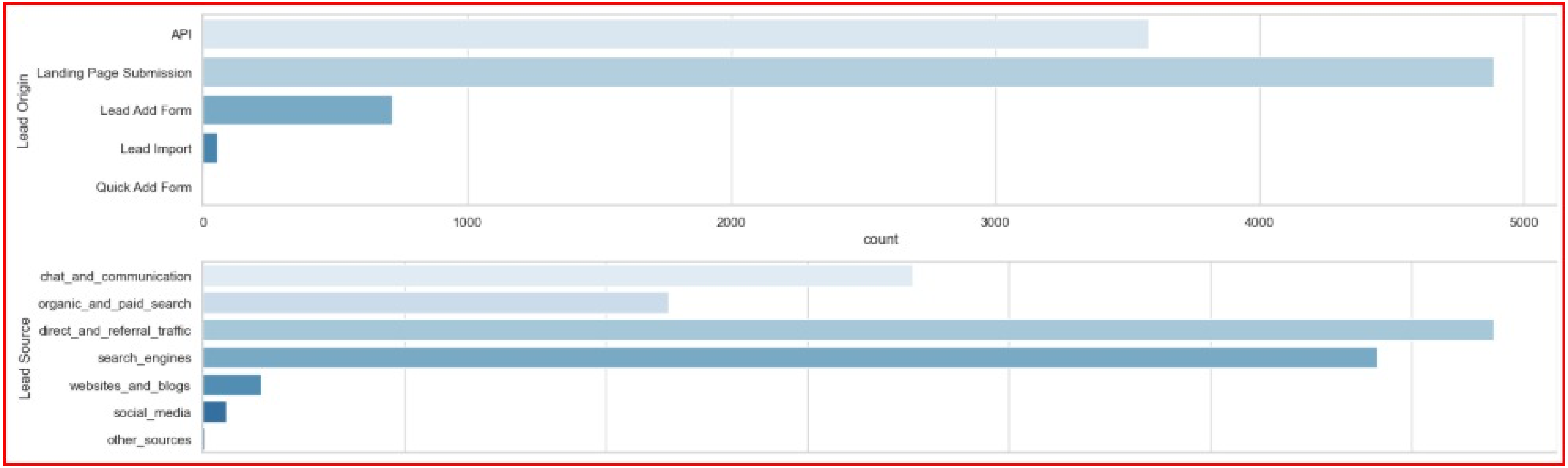
Univariate Analysis (Numerical columns)



- 'Converted' variable shows somewhat imbalance, however not significant
- 'Total Time Spent on Website' has positively skewed distribution. Most of the people aren't spending any time on website as 0 has the highest frequency

- Most of the customers view <= 5 pages per visit

Graphs & Insights
Univariate Analysis (Categorical columns)



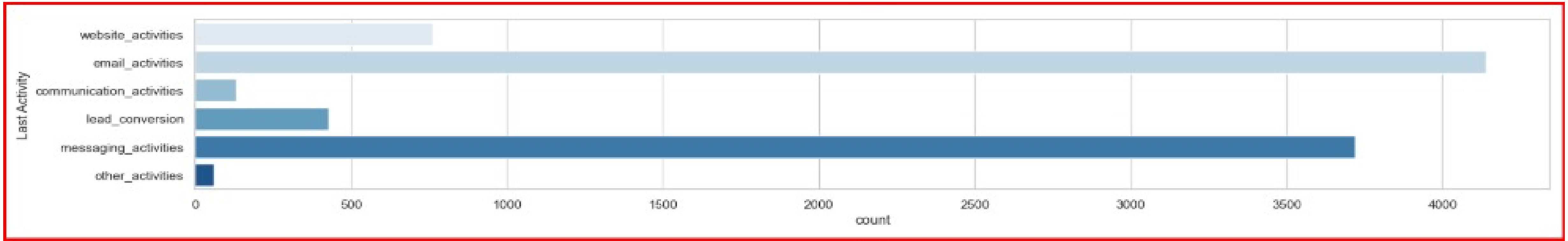
- ➢ For 'Lead Origin' variable, 'API' and 'Landing Page Submission' are the top 2 categories
- ➢ For 'Lead Source' variable, 'direct_and_paid_research' and 'search_engines' are the top 2 categories

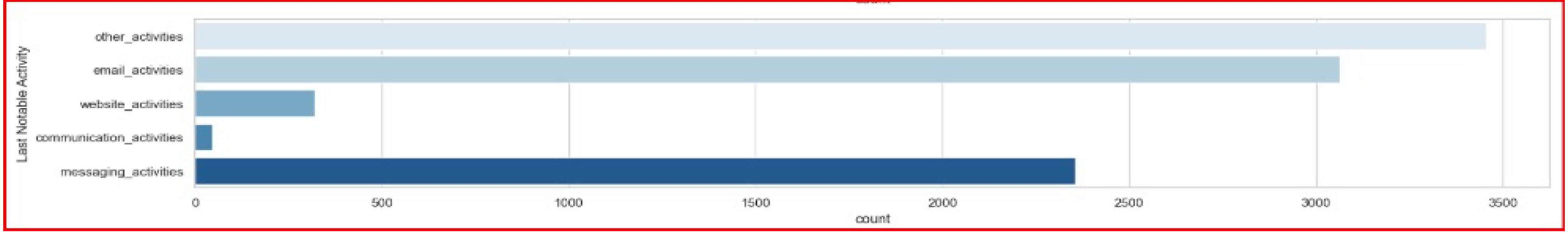
Note: Below groups were created to make visualization better:

- search engines: google, Google, bing
- direct and referral traffic: Direct Traffic, Reference, Referral Sites
- chat and communication: Olark Chat, Live Chat, Click2call
- social media: Facebook, Social Media, youtubechannel
- organic and paid search: Organic Search, Pay per Click Ads
- websites and blogs: Welingak Website, blog, WeLearn, welearnblog_Home
- other sources: Press Release, NC EDM, testone

Graphs & Insights

Univariate Analysis (Categorical columns)



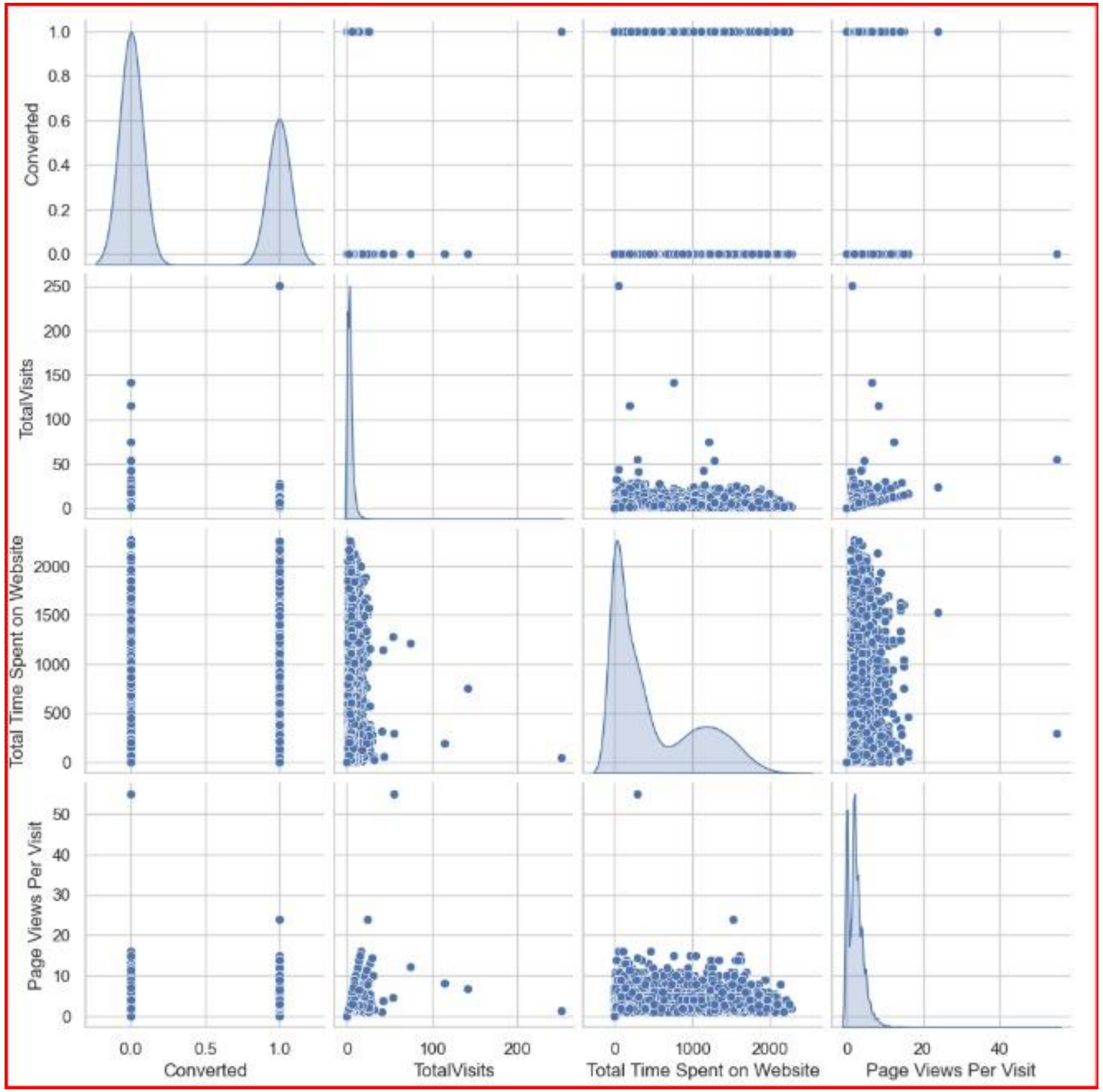


- For 'Last Activity' variable, 'email_activities' and 'messaging_activities' are the top 2 categories
- For 'Last Notable Activity' variable, 'other_activities' and 'messaging activites' are the top 2 categories

Note: Below groups were created for **'Last Activity'** & **'Last Notable Activity'** to make visualization better:

- **email activities:** Email Opened, Email Bounced, Email Link Clicked, Email Received, Email Marked Spam, Resubscribed to emails
- **messaging activities:** SMS Sent, Olark Chat Conversation
- **website activities:** Page Visited on Website, Form Submitted on Website, View in browser link Clicked
- **lead conversion:** Converted to Lead
- **communication activities:** Had a Phone Conversation, Approached upfront, Unreachable
- **other activities:** Unsubscribed, Visited Booth in Tradeshow

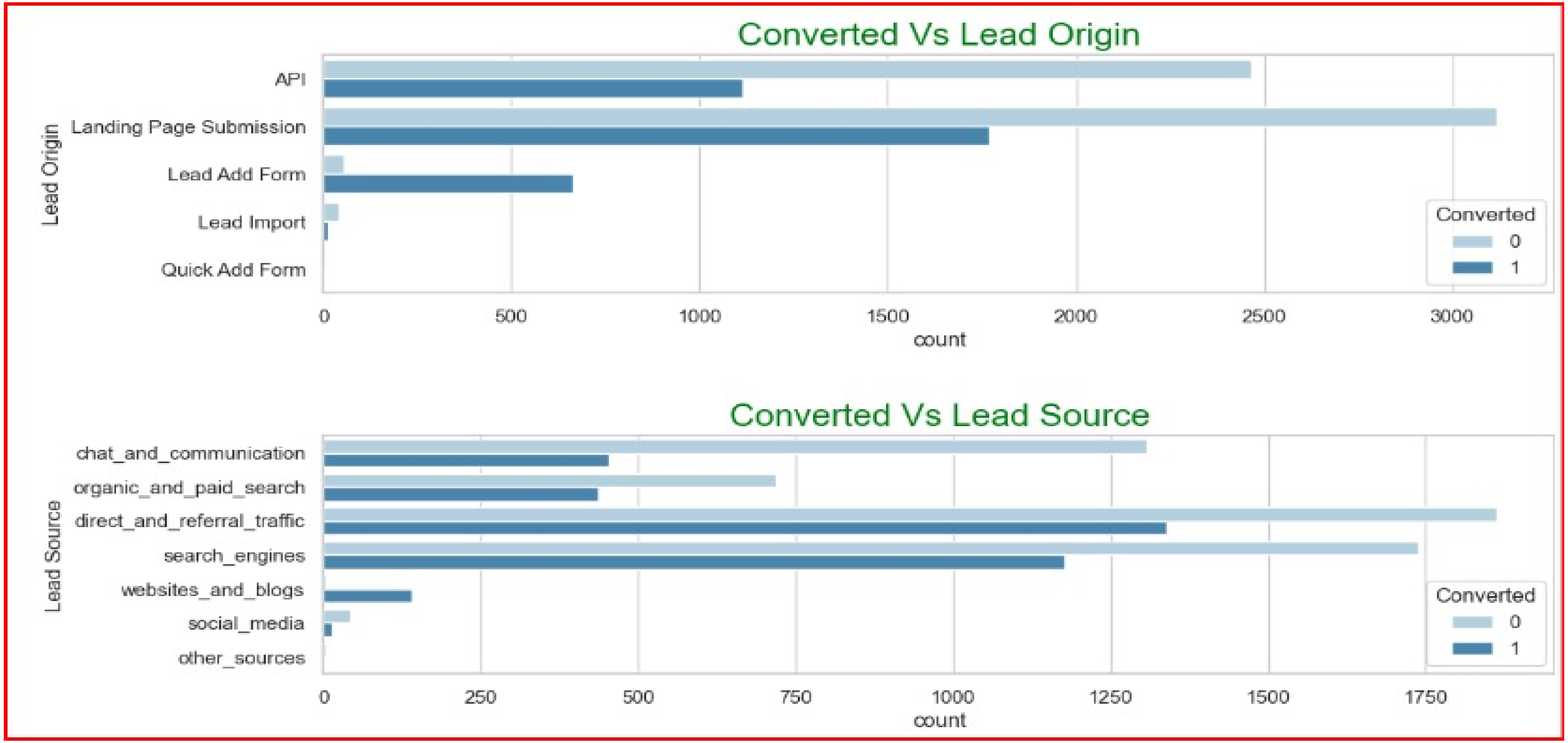
Graphs & Insights
Bivariate Analysis (Numerical columns)



➢ 'Total Visits' and 'Page Views Per Visit' show some level of linearity

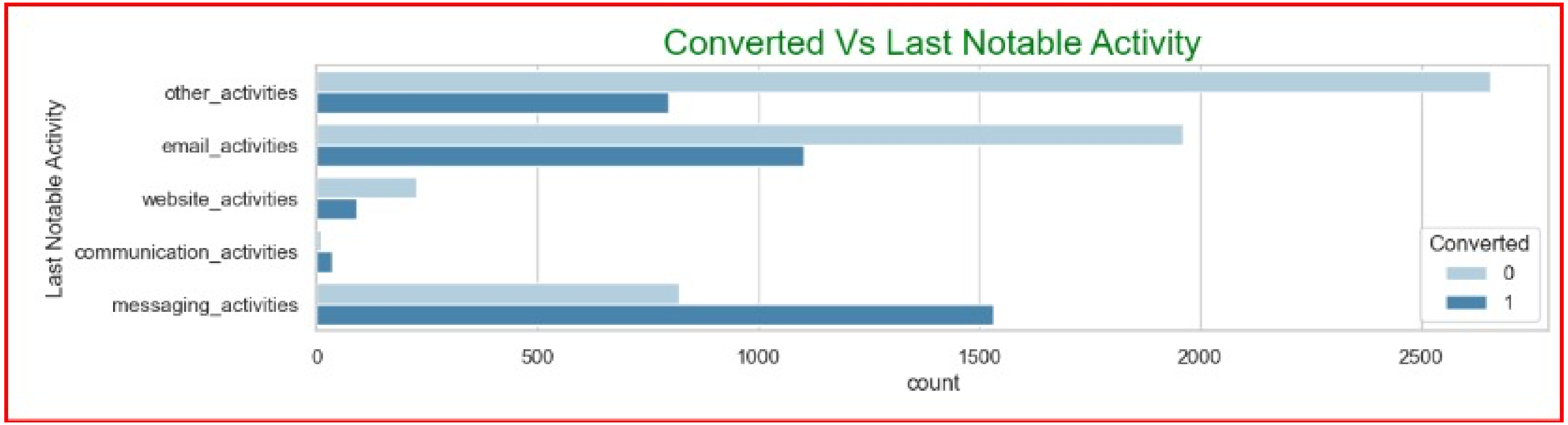
Graphs & Insights

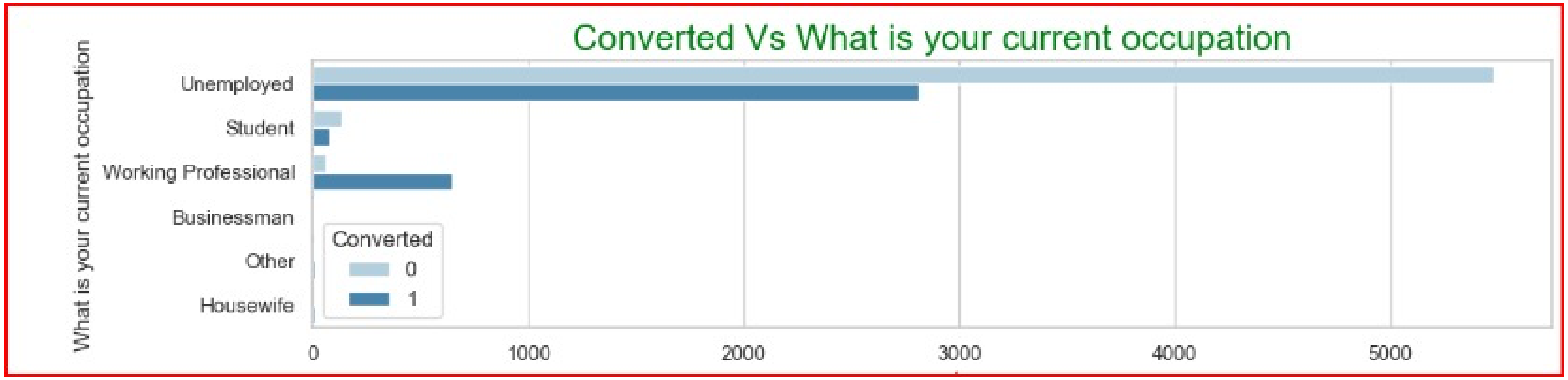
Bivariate Analysis (Categorical columns)



- For 'Lead Origin' variable, 'API' and 'Landing Page Submission' are the top 2 categories for 'Converted' (Target variable)
- For 'Lead Source' variable, 'direct_and_paid_research' and 'search_engines' are the top 2 categories for 'Converted' (Target variable)

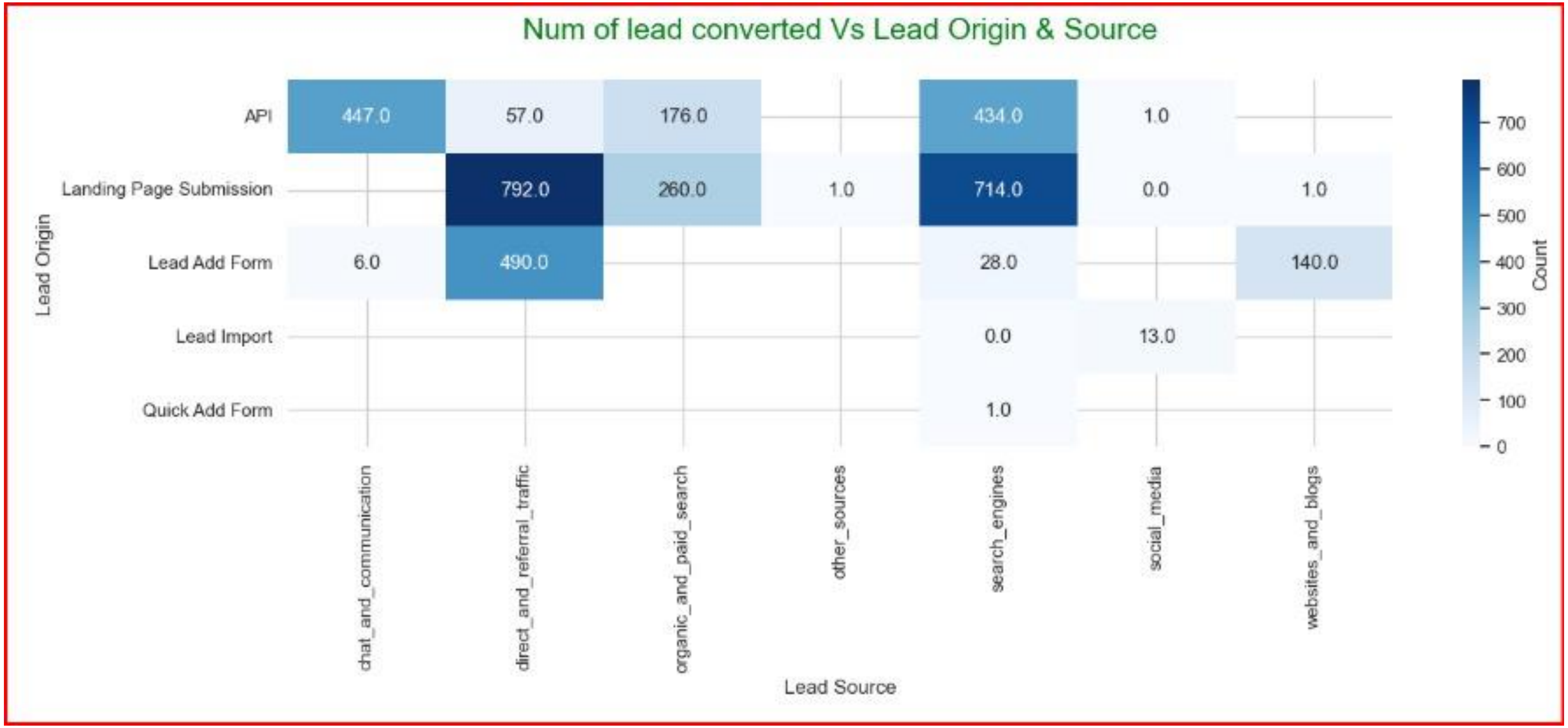
Graphs & Insights
Bivariate Analysis (Categorical columns)





- For 'Last Notable Activity' variable, 'email_activities' and 'messaging_activities' are the top 2 categories for 'Converted' (Target variable)
- For 'What is your current occupation' variable, 'Unemployed' and 'Working professional' are the top 2 categories for 'Converted' (Target variable)

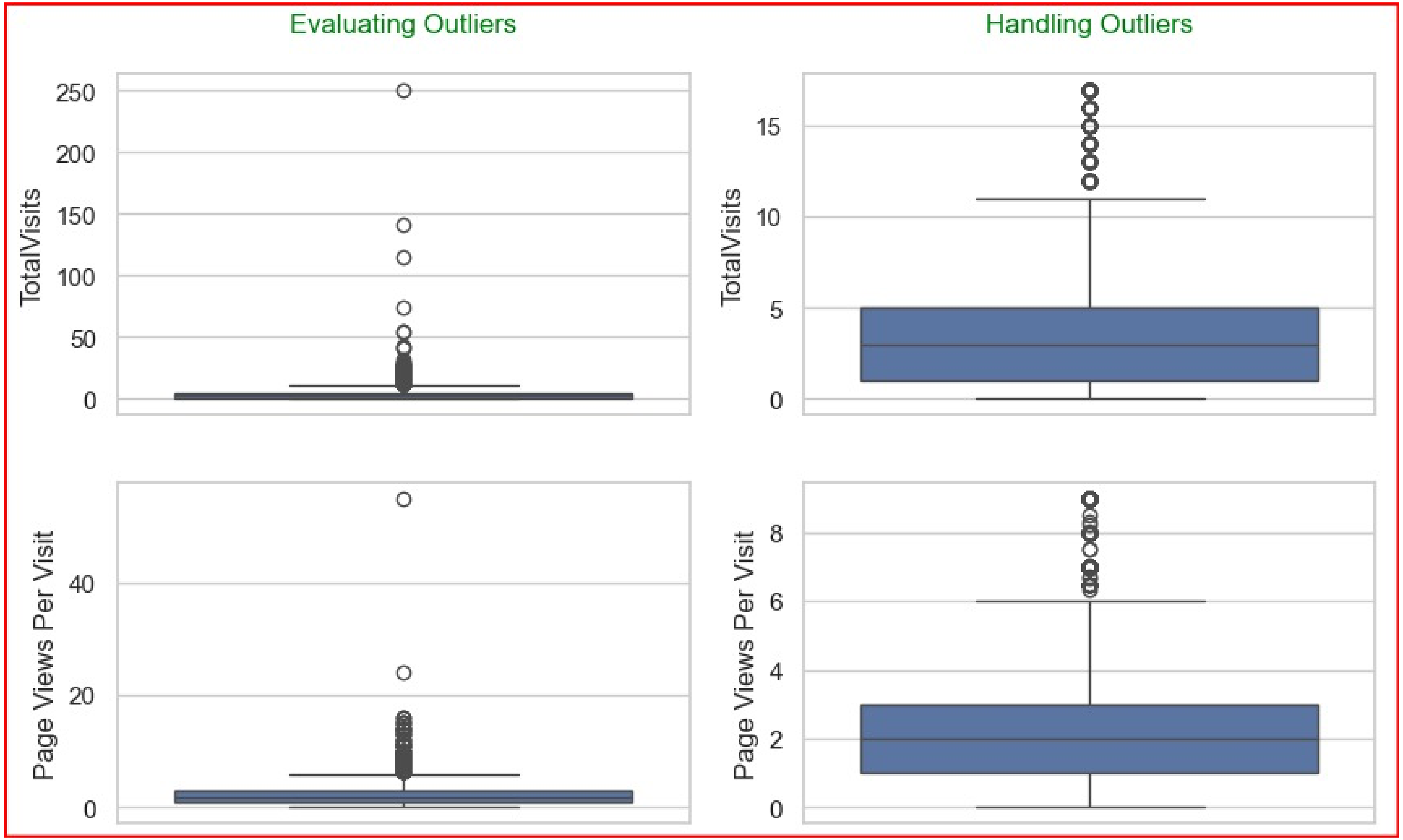
Graphs & Insights
Multivariate Analysis (Categorical columns)



- The below combination of categories from 'Lead Origin' and 'Lead Source' drive better conversions
  - 'Landing page submission' and 'direct_and_referral_traffic'
  - 'Landing page submission' and 'search_engines'
  - 'Lead Add Form' and 'direct_and_referral_traffic'

Graphs & Insights

Outlier Analysis (Numerical columns)



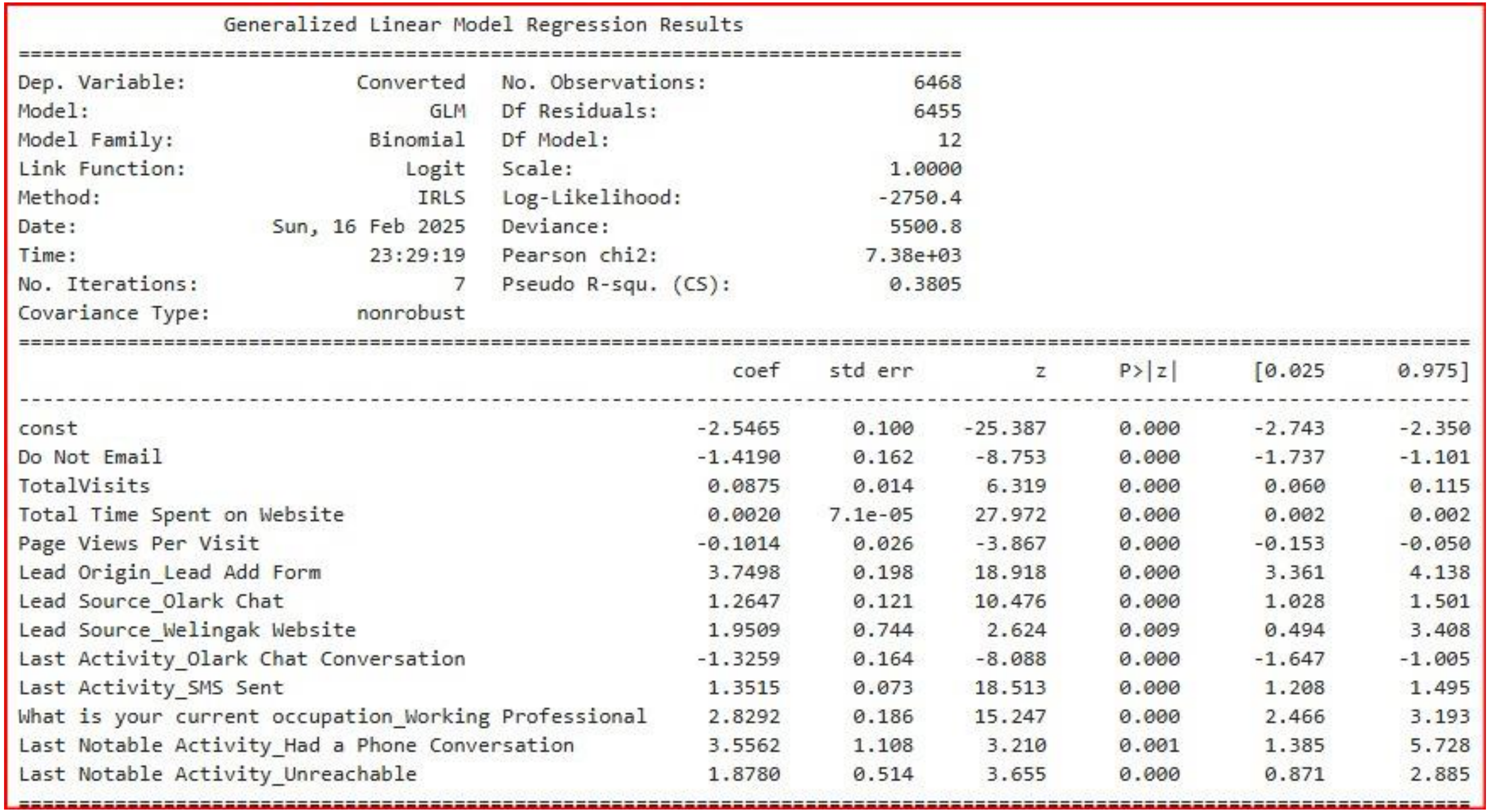
Evaluating Outliers

Handling Outliers

- ➢ Both these variables have outliers

- ➢ Values for both variables were capped to 99th Percentile to treat outliers

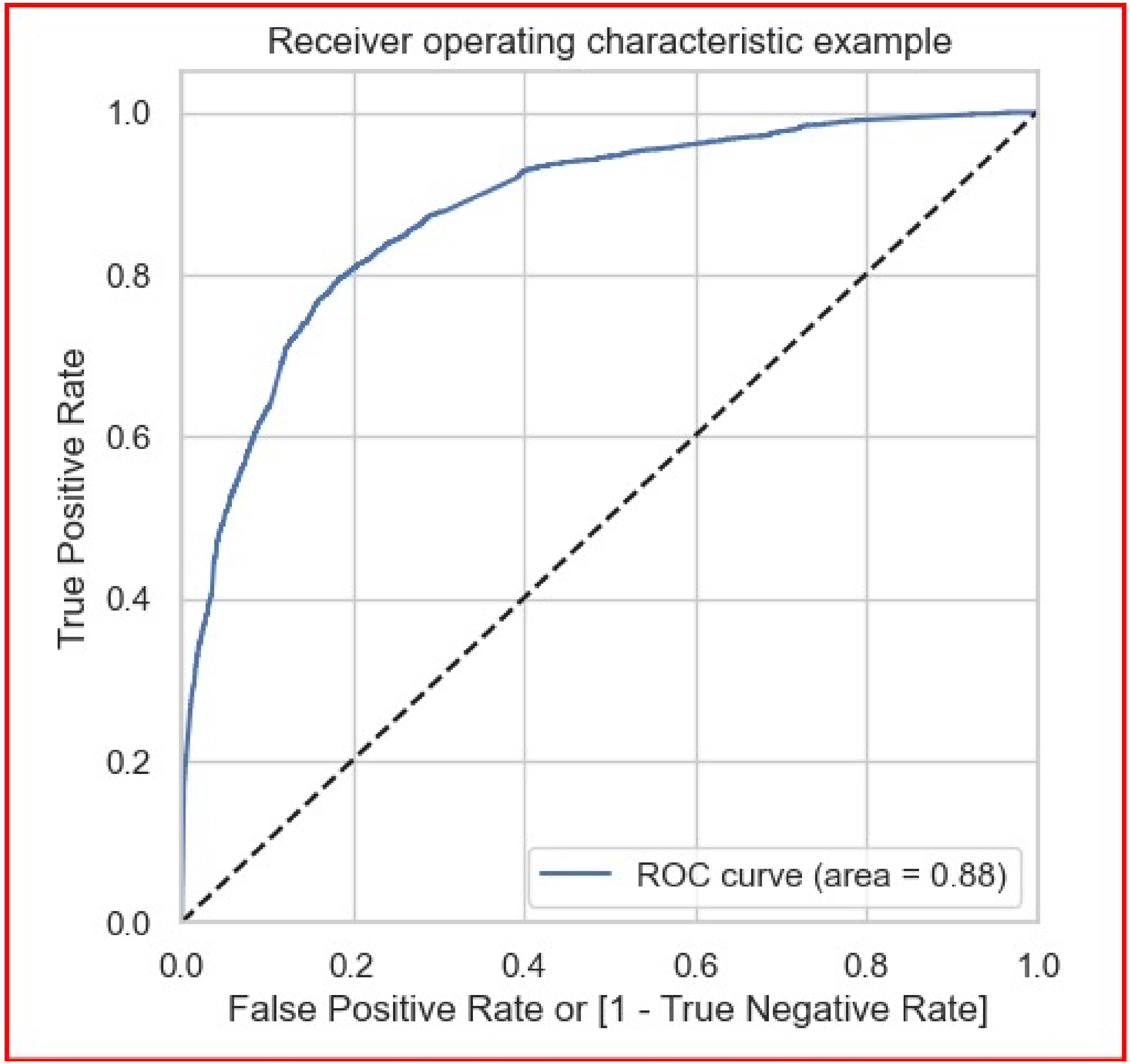
Final Logistic Regression Model



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6455
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2750.4
Date:                Sun, 16 Feb 2025   Deviance:                       5500.8
Time:                        23:29:19   Pearson chi2:                 7.38e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3805
Covariance Type:            nonrobust
=========================================================================================================================
                                                            coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------------
const                                                    -2.5465      0.100    -25.387      0.000      -2.743      -2.350
Do Not Email                                             -1.4190      0.162     -8.753      0.000      -1.737      -1.101
TotalVisits                                               0.0875      0.014      6.319      0.000       0.060       0.115
Total Time Spent on Website                               0.0020    7.1e-05     27.972      0.000       0.002       0.002
Page Views Per Visit                                     -0.1014      0.026     -3.867      0.000      -0.153      -0.050
Lead Origin_Lead Add Form                                 3.7498      0.198     18.918      0.000       3.361       4.138
Lead Source_Olark Chat                                    1.2647      0.121     10.476      0.000       1.028       1.501
Lead Source_Welingak Website                              1.9509      0.744      2.624      0.009       0.494       3.408
Last Activity_Olark Chat Conversation                    -1.3259      0.164     -8.088      0.000      -1.647      -1.005
Last Activity_SMS Sent                                    1.3515      0.073     18.513      0.000       1.208       1.495
What is your current occupation_Working Professional      2.8292      0.186     15.247      0.000       2.466       3.193
Last Notable Activity_Had a Phone Conversation            3.5562      1.108      3.210      0.001       1.385       5.728
Last Notable Activity_Unreachable                         1.8780      0.514      3.655      0.000       0.871       2.885
=========================================================================================================================
```

- This is the final model with 12 statistically significant features, with VIF values less than 2.5 which ensures that there is no significant multi-collinearity in the model

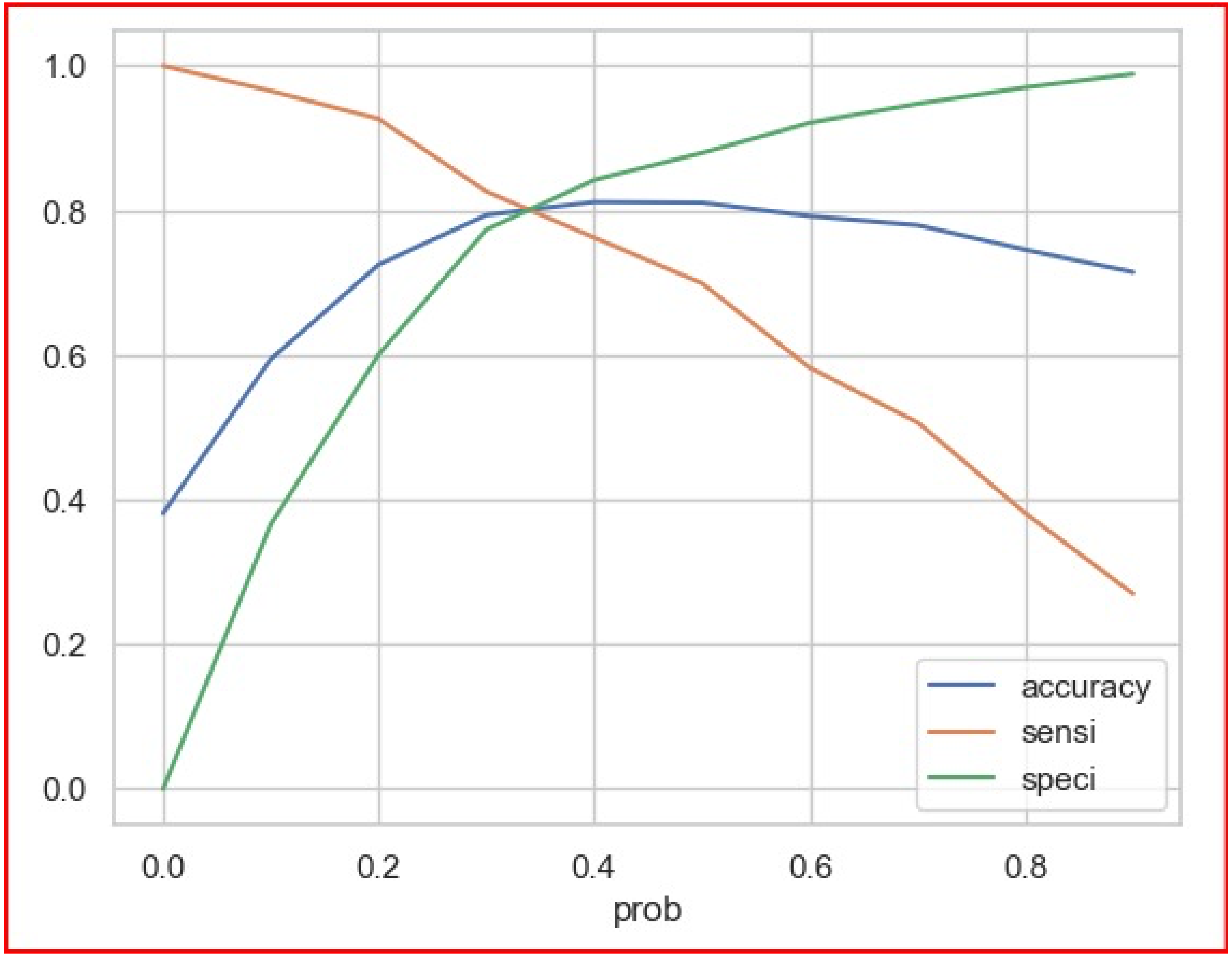
Receiver Operating Characteristic (ROC) Curve



- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. Our curve is also following the same trajectory

- Area Under the Curve (AUC) for this **ROC is 0.88** which means that the model is **very good** and is able to differentiate positive and negative classes effectively

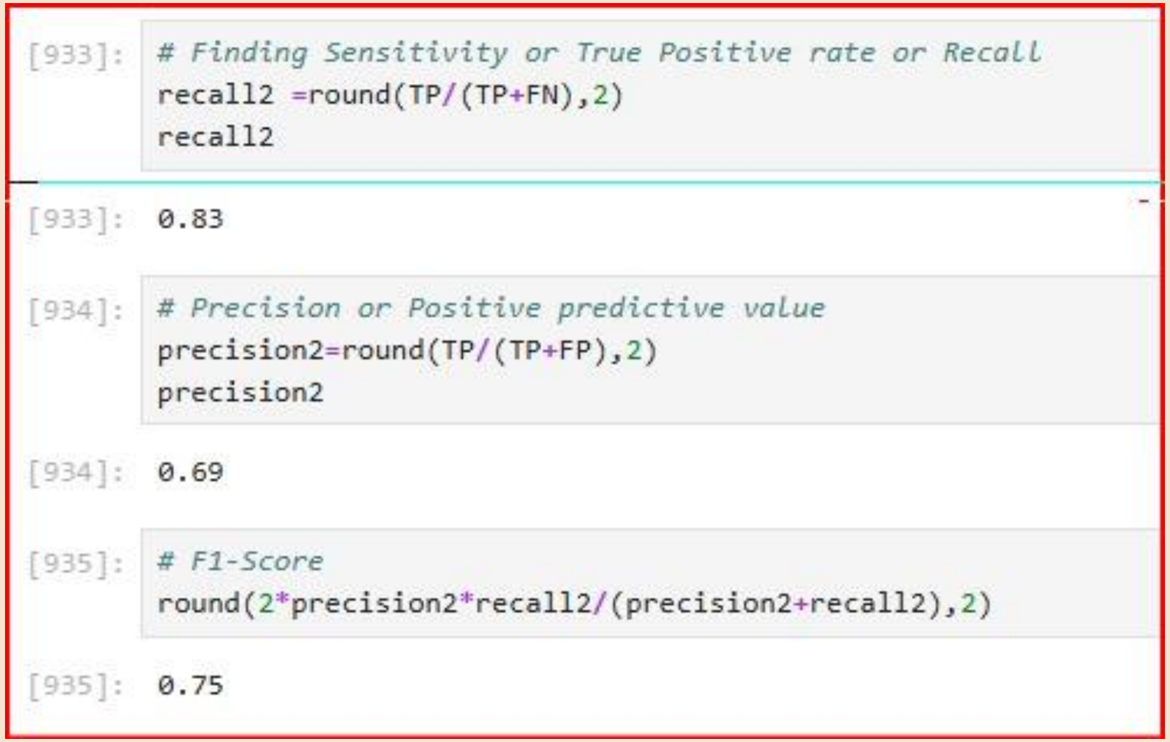
Finding Optimal Cut Off value



- Accuracy, Sensitivity and Specificity are intercepting at 0.30, which we'll now consider as the Optimal Cut off value
- This informs that any lead with a lead score of >= 30 should be considered as 'Hot Lead'

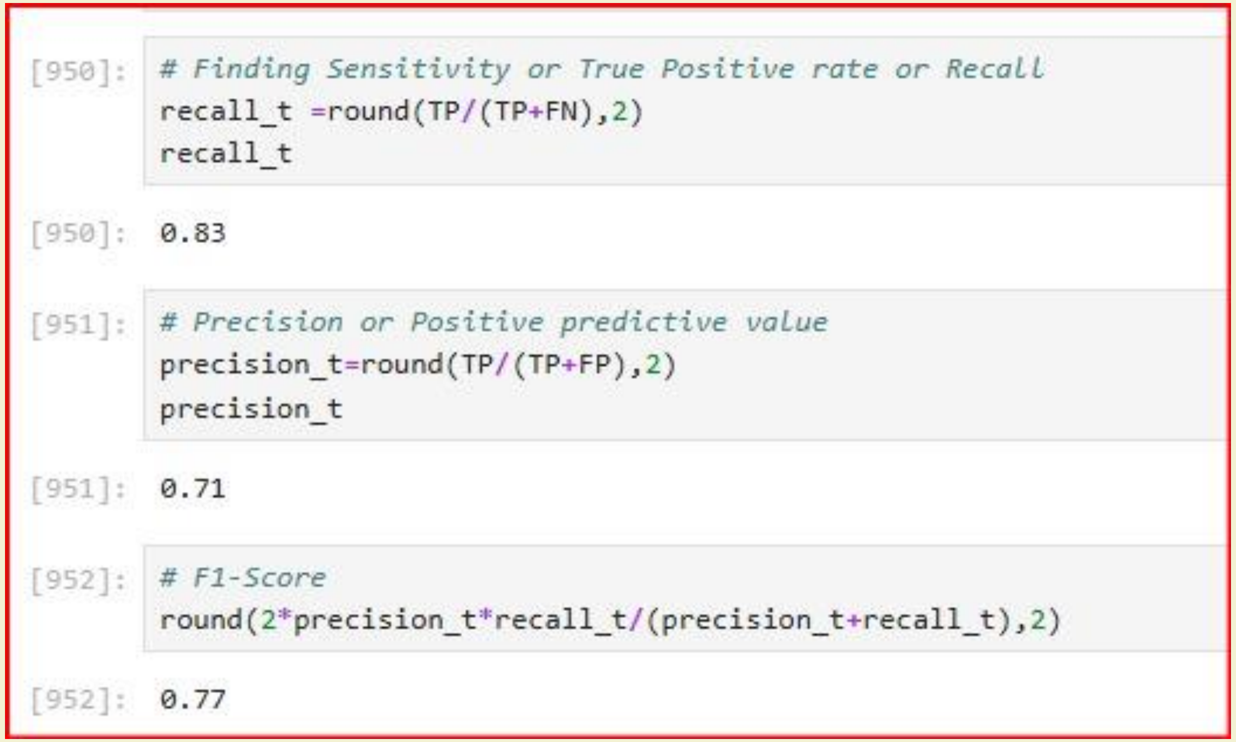
Evaluation Metrics – Train and Test datasets

Train dataset evaluation scores



```
[933]: # Finding Sensitivity or True Positive rate or Recall
       recall2 =round(TP/(TP+FN),2)
       recall2

[933]: 0.83

[934]: # Precision or Positive predictive value
       precision2=round(TP/(TP+FP),2)
       precision2

[934]: 0.69

[935]: # F1-Score
       round(2*precision2*recall2/(precision2+recall2),2)

[935]: 0.75
```

Test dataset evaluation scores

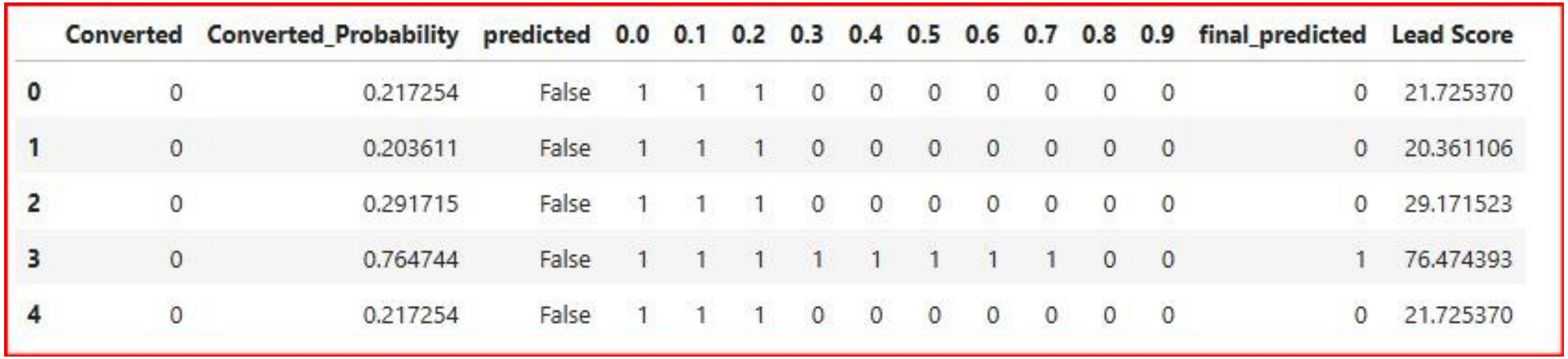


```
[950]: # Finding Sensitivity or True Positive rate or Recall
       recall_t =round(TP/(TP+FN),2)
       recall_t

[950]: 0.83

[951]: # Precision or Positive predictive value
       precision_t=round(TP/(TP+FP),2)
       precision_t

[951]: 0.71

[952]: # F1-Score
       round(2*precision_t*recall_t/(precision_t+recall_t),2)

[952]: 0.77
```

Lead Scores

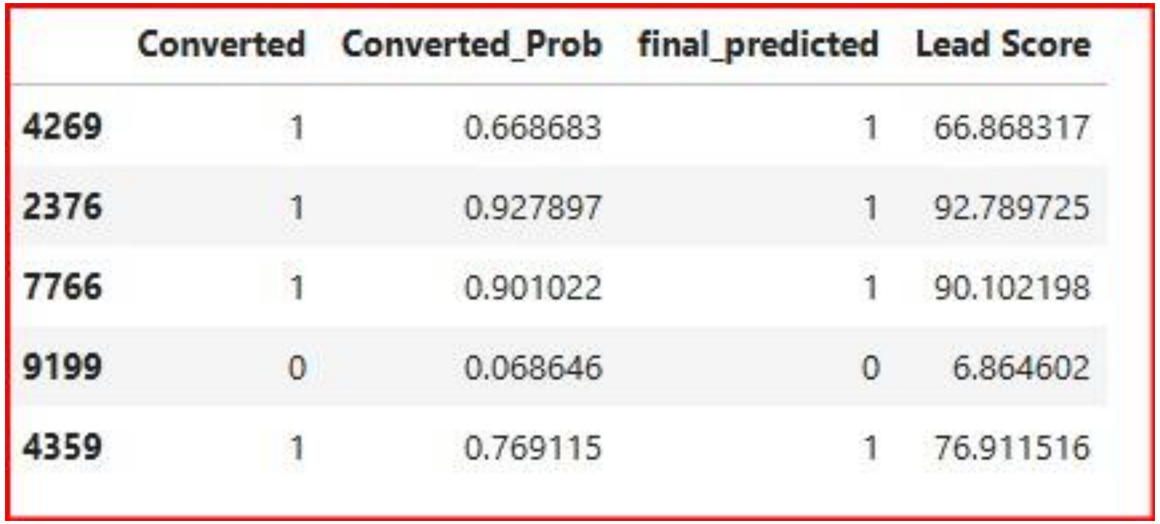
For both Train and Test datasets

Train Dataset



	Converted	Converted_Probability	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead Score
0	0	0.217254	False	1	1	1	0	0	0	0	0	0	0	0	21.725370
1	0	0.203611	False	1	1	1	0	0	0	0	0	0	0	0	20.361106
2	0	0.291715	False	1	1	1	0	0	0	0	0	0	0	0	29.171523
3	0	0.764744	False	1	1	1	1	1	1	1	1	0	0	1	76.474393
4	0	0.217254	False	1	1	1	0	0	0	0	0	0	0	0	21.725370

Test Dataset



	Converted	Converted_Prob	final_predicted	Lead Score
4269	1	0.668683	1	66.868317
2376	1	0.927897	1	92.789725
7766	1	0.901022	1	90.102198
9199	0	0.068646	0	6.864602
4359	1	0.769115	1	76.911516

Recommendations

- Model advises that Sales team should engage Working professionals via Calls, Emails and SMS
  - Coefficients are: - Working professional (2.8292), Emails (-1.6389), Calls (3.5562), SMS (1.3515)

- Business should focus on optimizing the following Lead Sources: - Lead Add Form, Olark Chat, Welingak Website since they are positively correlated and below are their coefficients
  - Lead Add Form (3.7498), Welingak Website (1.9509), Olark Chat (1.2647)

- It is recommended that leads (Hot Leads, i.e. lead score of >= 30) be distributed evenly among sales representatives while maintaining a balanced mix of high, medium, and low lead scores. Since lead scores range from 0 to 100, with higher scores indicating a greater likelihood of conversion, this approach will ensure that each salesperson has an equal opportunity to work with leads of varying potential.