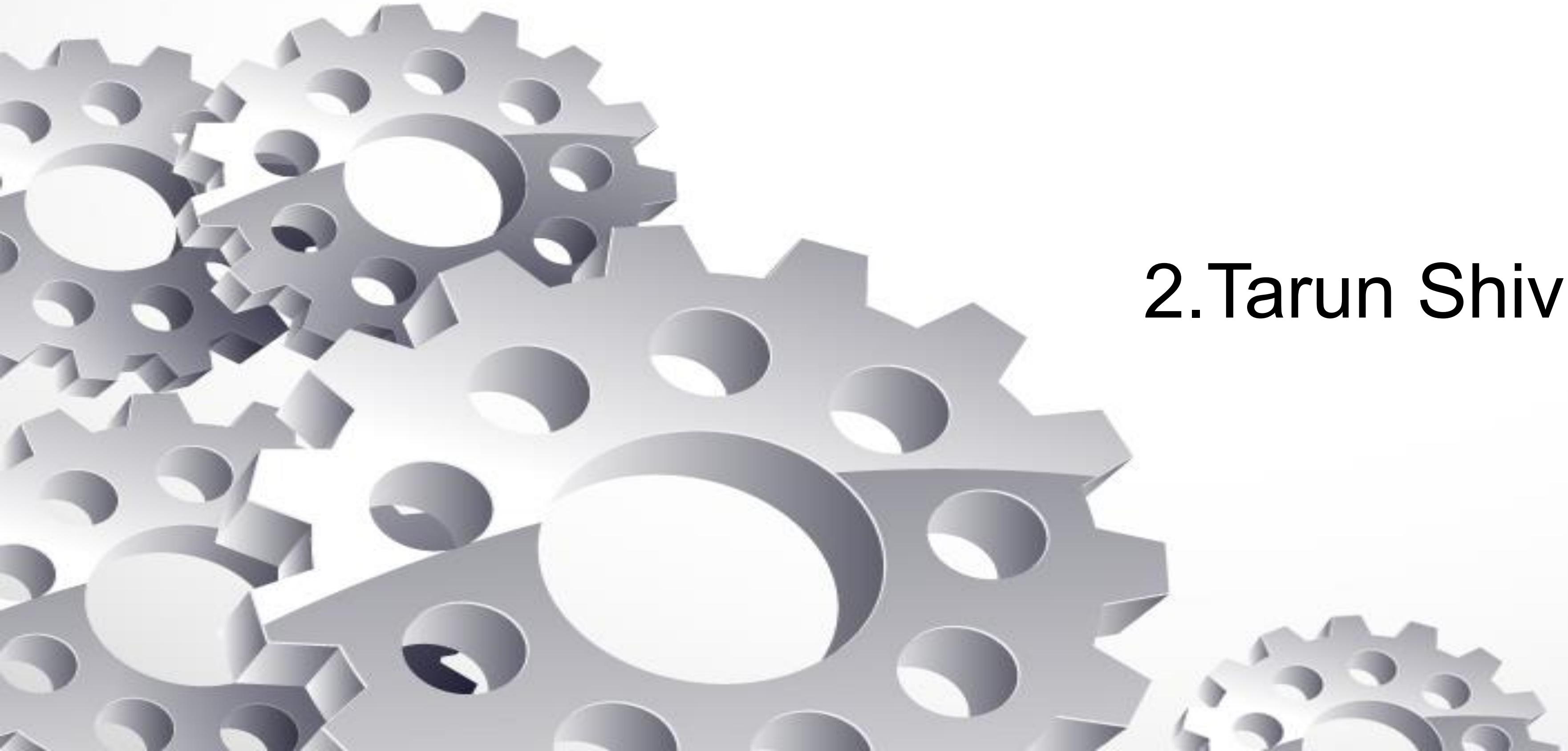
Lead Scoring Case Study Using Logistic Regression



Submitted by:

1.Soma Prasad

2.Tarun Shivkumar Ramanathan

3.Satweek Bandi

Contents:-

□Statement of the Problem

□Objective of the Business

□Solution Methodology

□EDA

□Building of the Model

□Evaluation of Model

□Conclusion

Statement Of Problem:-

- X Education is selling the online courses to the industry professionals.
- XEducation is having ample amount of leads.those leads conversion rates are lower.If we consider an example,they had 100 leads,only 30 of them joins the organisation.
- This process can be made optimistic is surfing for the most potential leads,called “Hot Leads”
- If everything works great then those set of leads are more into the conversion rate,through which the sales leads will follow up the “Hot Leads”.

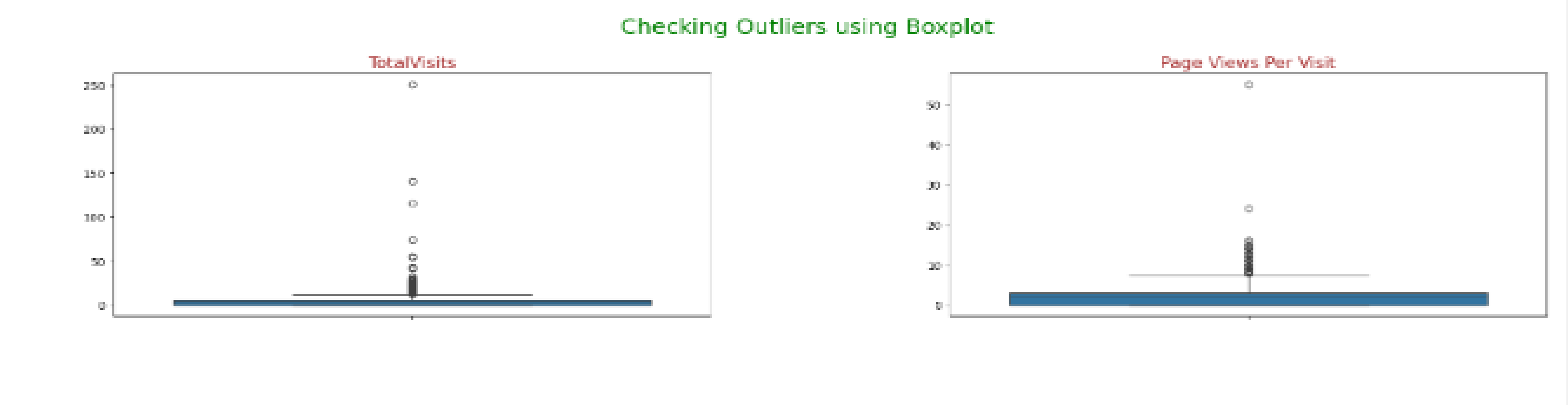
Objective of the Business :-

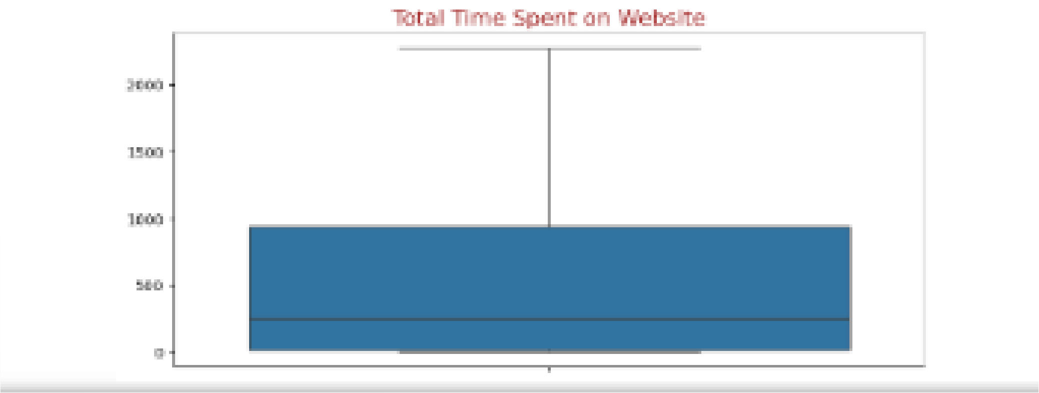
- The Lead X wants us to build a model that analyzes each lead and assigns them a score between 0 and 100. The higher the score, the "hotter" the lead, meaning they're more interested and ready to buy.
- CEO wish us to achieve a lead conversion rate of 80%.
- They wanted a Model to adapt to peak times, optimize manpower, and outline post-target strategies for effective implementation and sustained success.

Solution Methodology:-

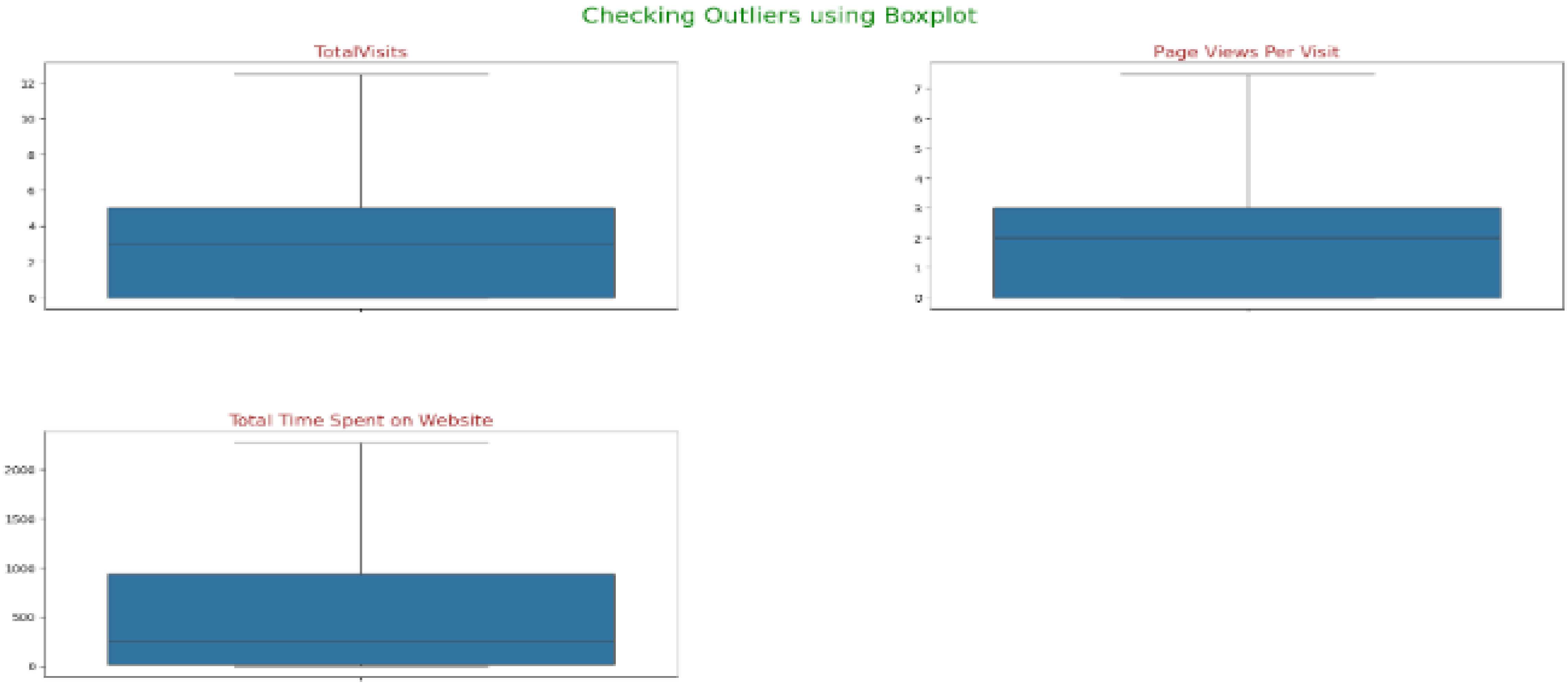
- o Data cleaning and data Inspection.
- o EDA
- o Dummy Variable Creation
- o Test-Train split
- o Feature scaling
- o Dropping highly correlated dummy variables
- o Model Building (RFE Rsquared VIF and pvalues)
- o Model Evaluation
- o Checking Accuracy
- o Finding Optimal Cutoff Point
- o Making predictions on test set

Outlier Analysis





Checking Outliers with Boxplot :-

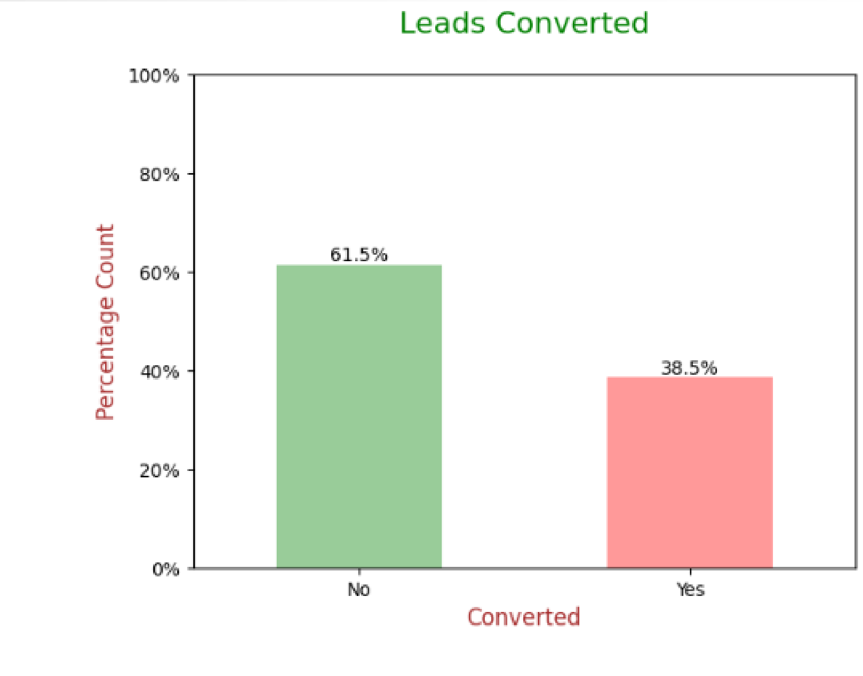


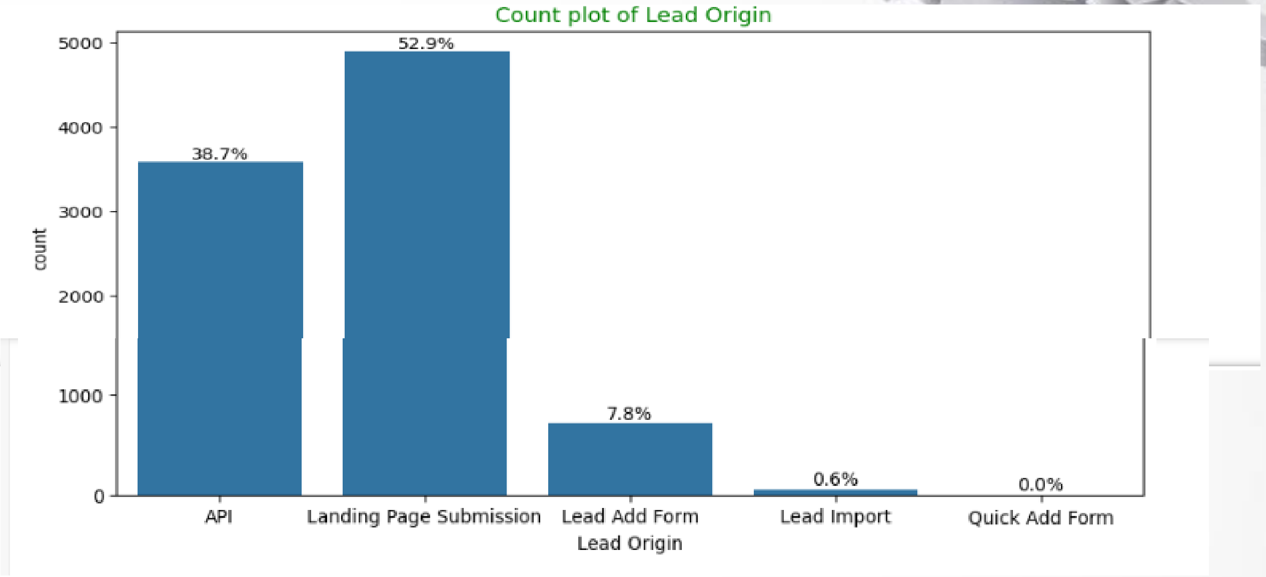
Exploratory Data Analysis

Univariate Analysis

- Most people find us through Google or directly visiting our site, making these the primary lead sources.
- The majority of site visitors are unemployed individuals, highlighting a potential demographic focus for outreach and support.
- Email opening marks the final step for most leads, emphasizing the importance of email communication in our engagement strategy.
- Most leads originate from submitting landing pages, indicating the effectiveness of our landing page design in capturing interest.

Leads Converted ,Univariate Analysis for Categorical---Variables:-

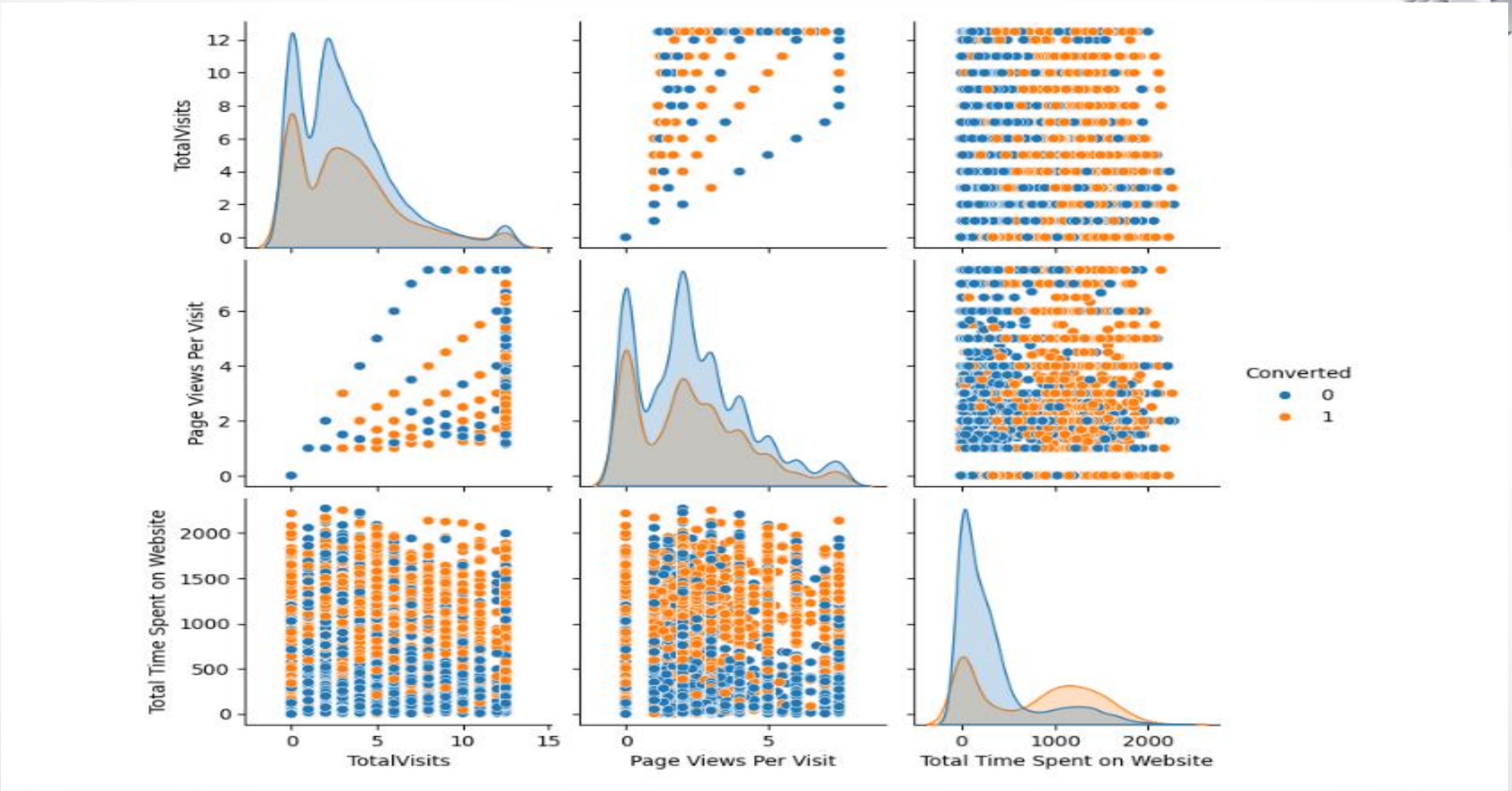




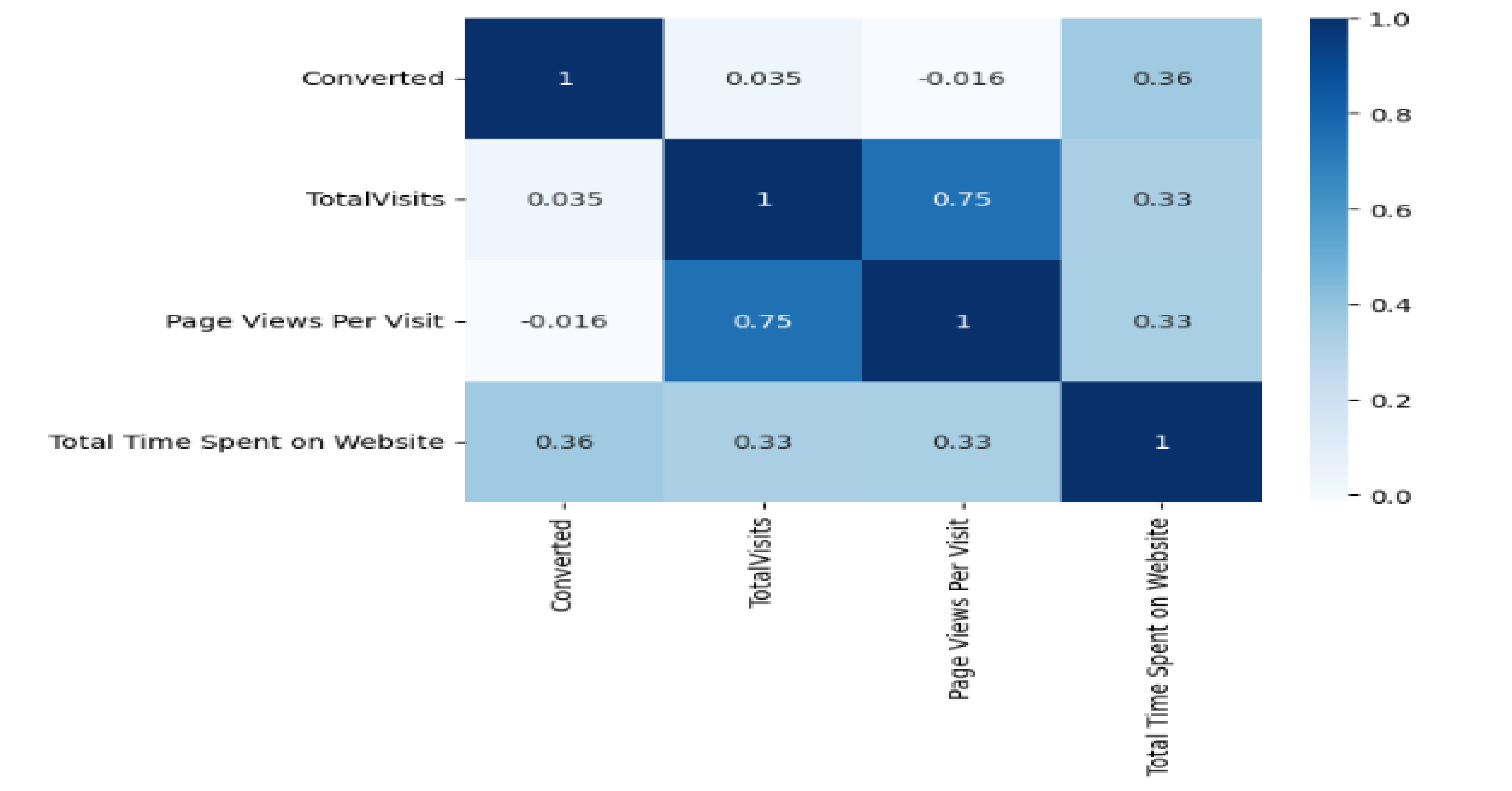
Bivariate Analysis

- "Lead forms & landing pages yield most conversions, prioritize Google/Direct Traffic sources."
- "Email/SMS activity drives highest conversions, prioritize outreach to these leads."
- "India leads in conversions, focus on finance, marketing, HR. Professionals prioritize over housewives."
- "Career prospects drive course opt-ins, prioritize clients with this motivation."
- "Tag 'revert after email' has highest conversion rate among others."
- "Target Mumbai for lead conversion first, then Tier II cities."
- "SMS leads in conversions, followed by email opens."

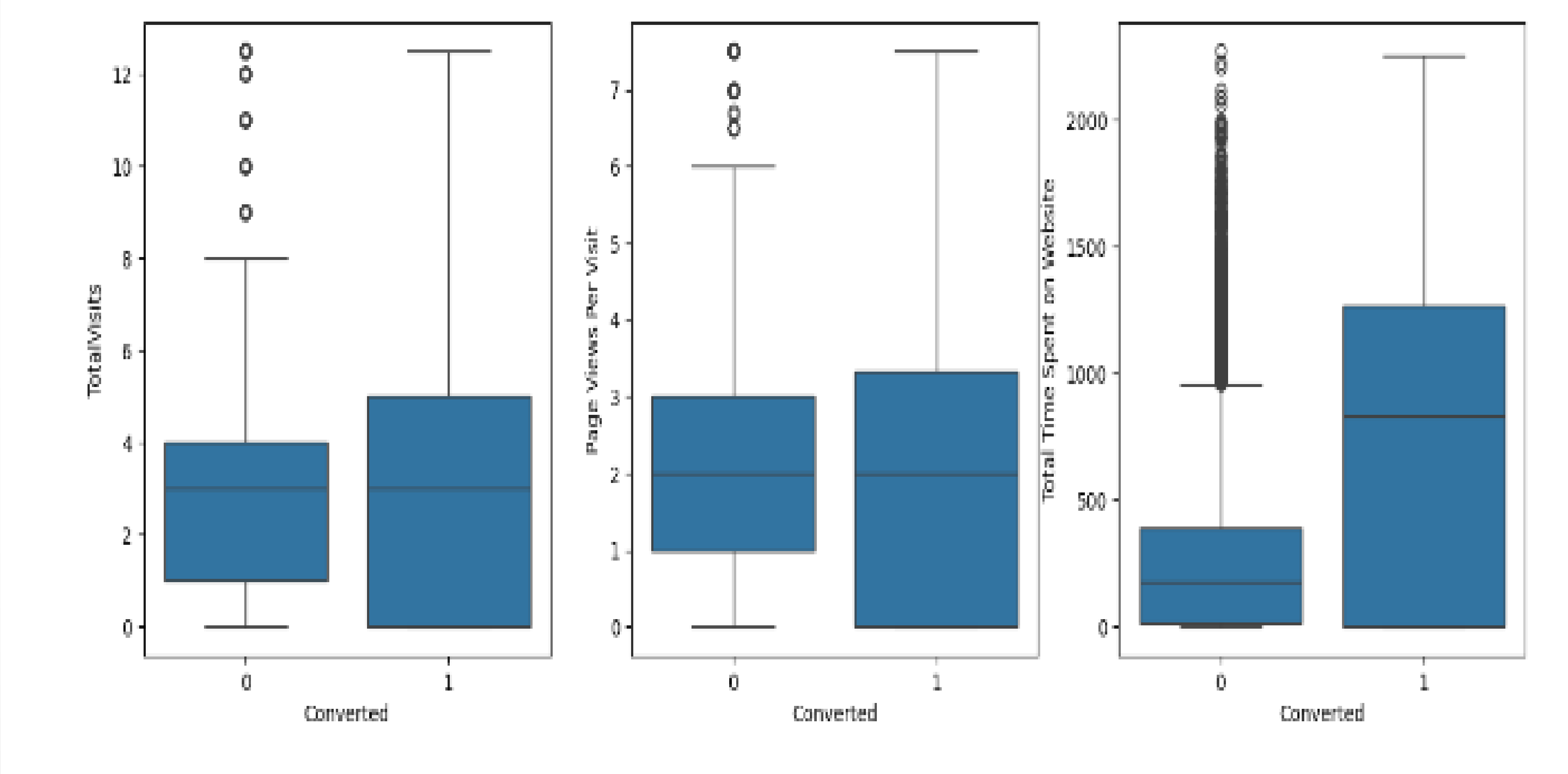
EDA BIVARIATE ANALYSIS:-

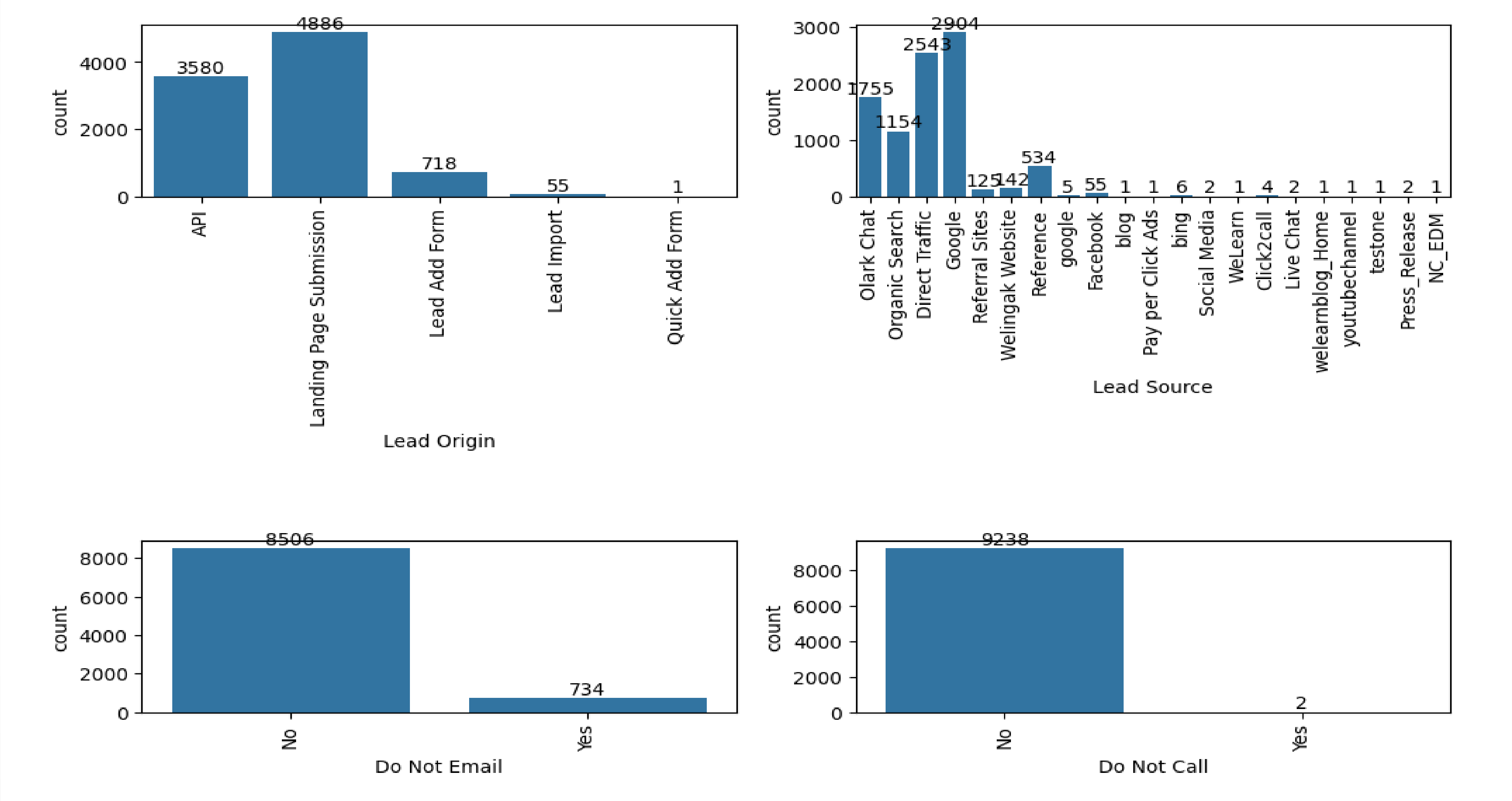


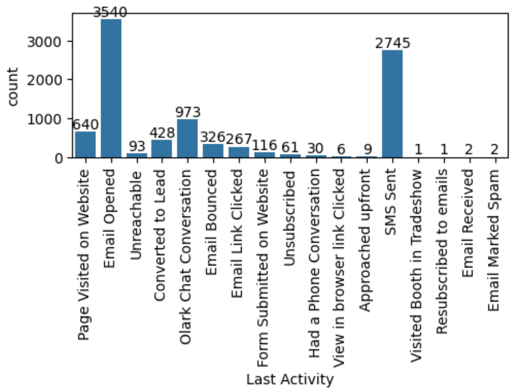
HEATMAP :-

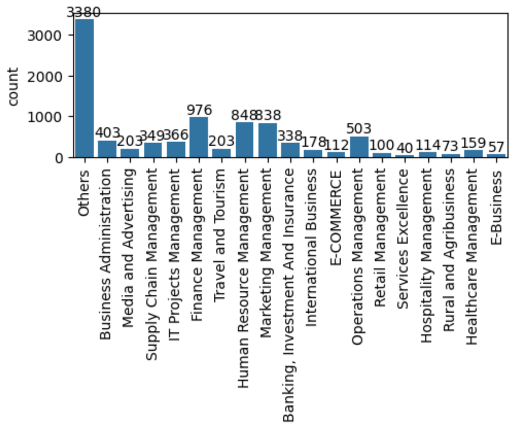

Converted
TotalVisits
Page Views Per Visit
Total Time Spent on Website
1
0.035
-0.016
0.36
0.035
1
0.75
0.33
-0.016
0.75
1
0.33
0.36
0.33
0.33
1
Converted
TotalVisits
Page Views Per Visit
Total Time Spent on Website
1.0
0.8
0.6
0.4
0.2
0.0

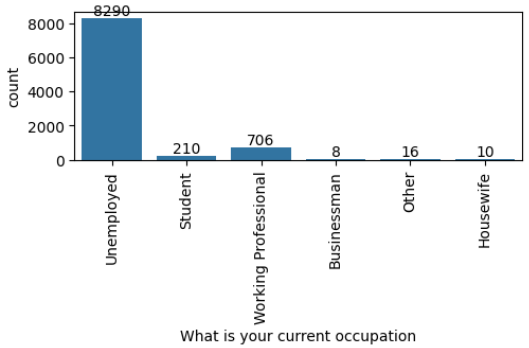
Boxplot with Converted as hue

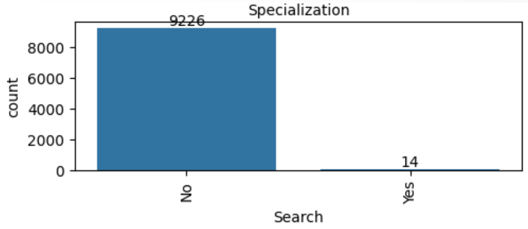


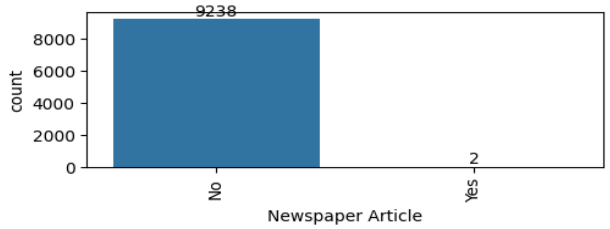

count
4000
2000
0
3580
4886
718
55
1
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin
count
3000
2000
1000
0
1755
1154
2543
2904
125
142
534
5
55
1
1
6
2
1
4
2
1
1
1
2
1
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source
count
8000
6000
4000
2000
0
8506
734
No
Yes
Do Not Email
count
8000
6000
4000
2000
0
9238
2
No
Yes
Do Not Call

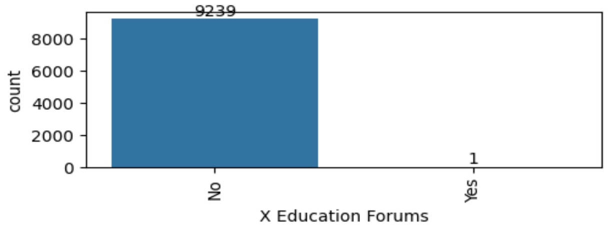

count
3000
2000
1000
0
3540
640
93
428
973
326
267
116
61
30
6
9
2745
1
1
2
2
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity

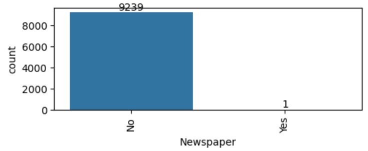

count
3000
2000
1000
0
3380
403
203
349
366
976
203
848
838
338
178
112
503
100
40
114
73
159
57
Others
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business

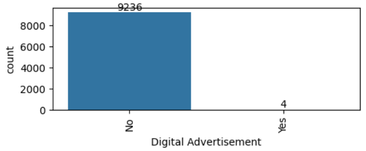

count
8000
6000
4000
2000
0
8290
210
706
8
16
10
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

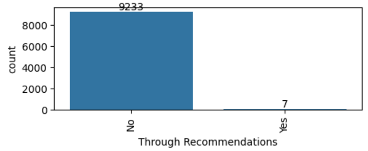

Specialization
count
8000
6000
4000
2000
0
9226
14
No
Yes
Search

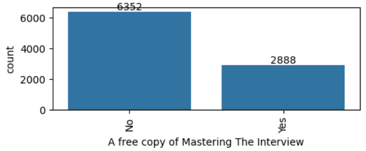

count
8000
6000
4000
2000
0
9238
2
No
Yes
Newspaper Article


count
8000
6000
4000
2000
0
9239
1
No
Yes
X Education Forums


count
8000
6000
4000
2000
0
9239
1
No
Yes
Newspaper

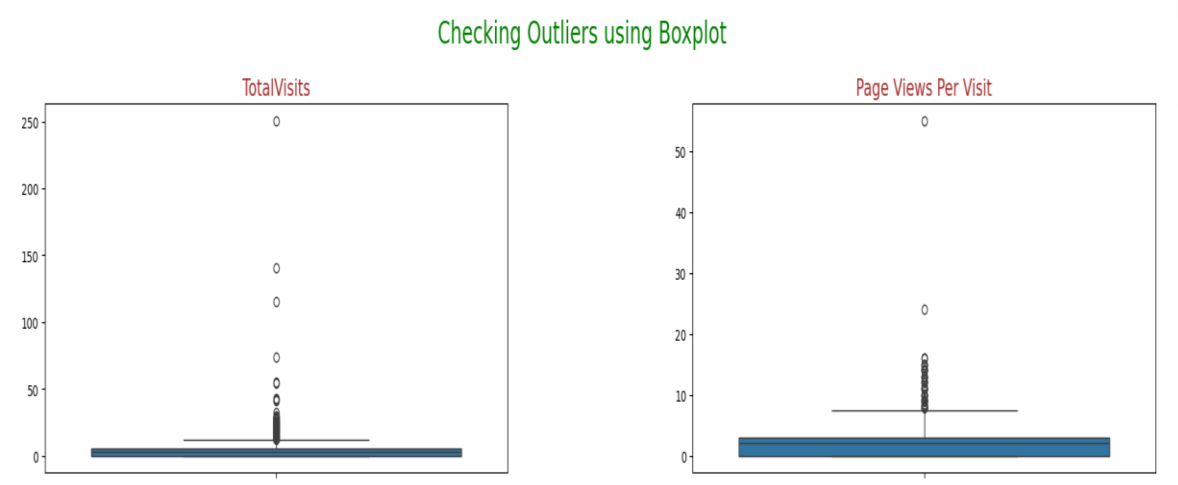

count
8000
6000
4000
2000
0
9236
4
No
Yes
Digital Advertisement

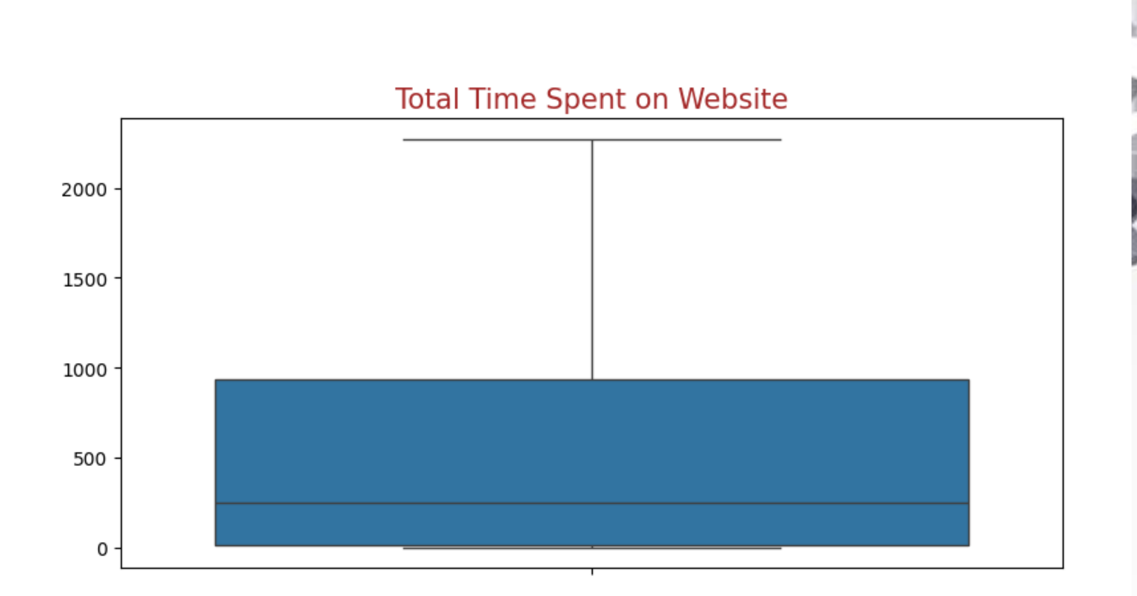

count
8000
6000
4000
2000
0
9233
7
No
Yes
Through Recommendations

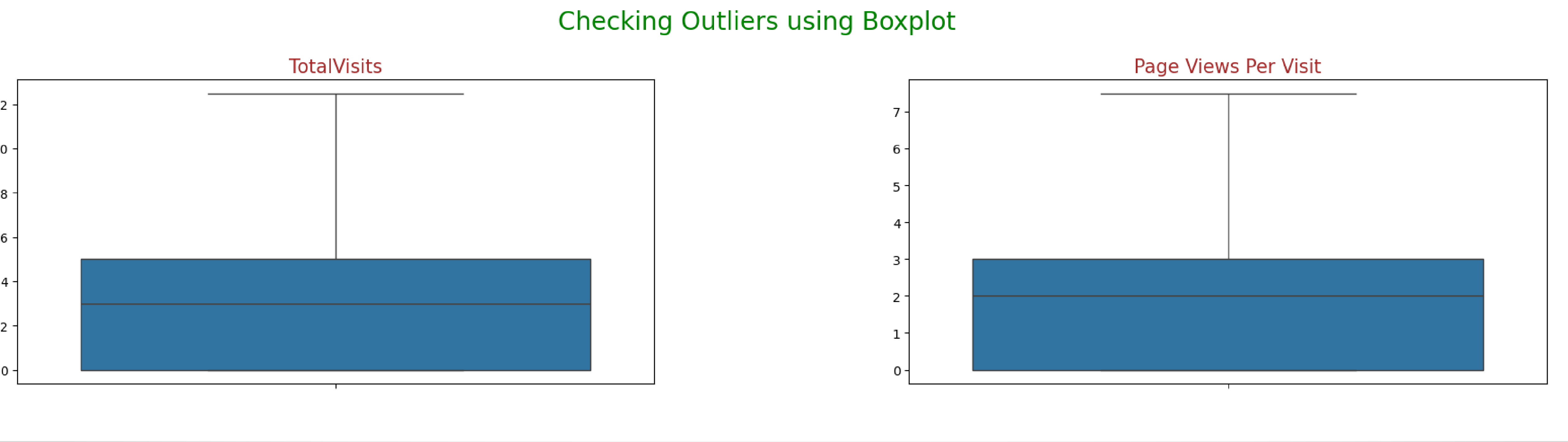

count
6000
4000
2000
0
6352
2888
No
Yes
A free copy of Mastering The Interview

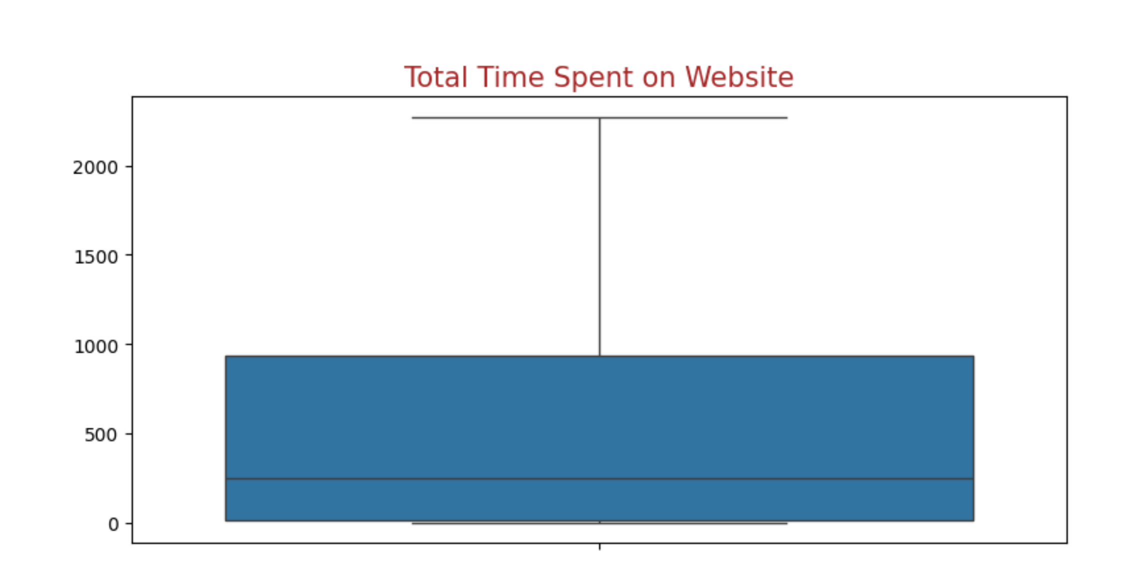
MODEL BUILDING :-

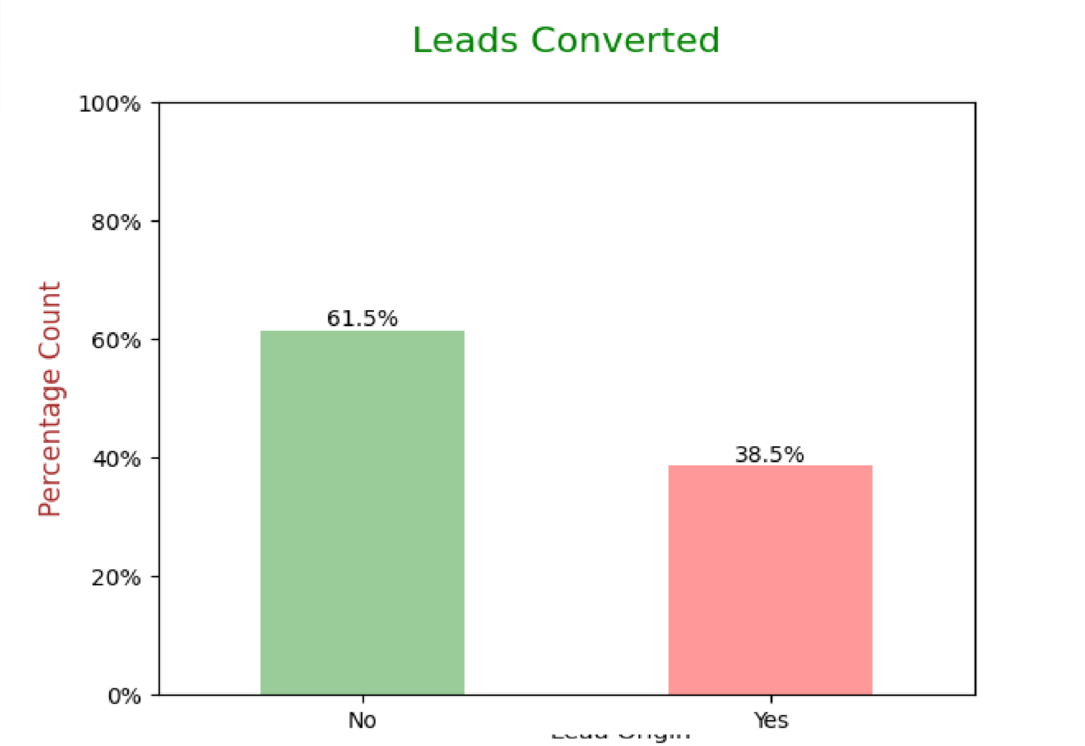
- "Split data into training and testing sets."
- "Scale variables in the training set for consistency."
- "Construct initial model."
- "Use Recursive Feature Elimination (RFE) to remove less important variables."
- "Refine model by eliminating variables with high p-values."
- "Assess multicollinearity with Variance Inflation Factor (VIF)."
- "Make predictions using the training set."
- "Evaluate model accuracy and other metrics."
- "Apply model to predict outcomes using the test set."
- "Analyze precision and recall of test predicting future insights."

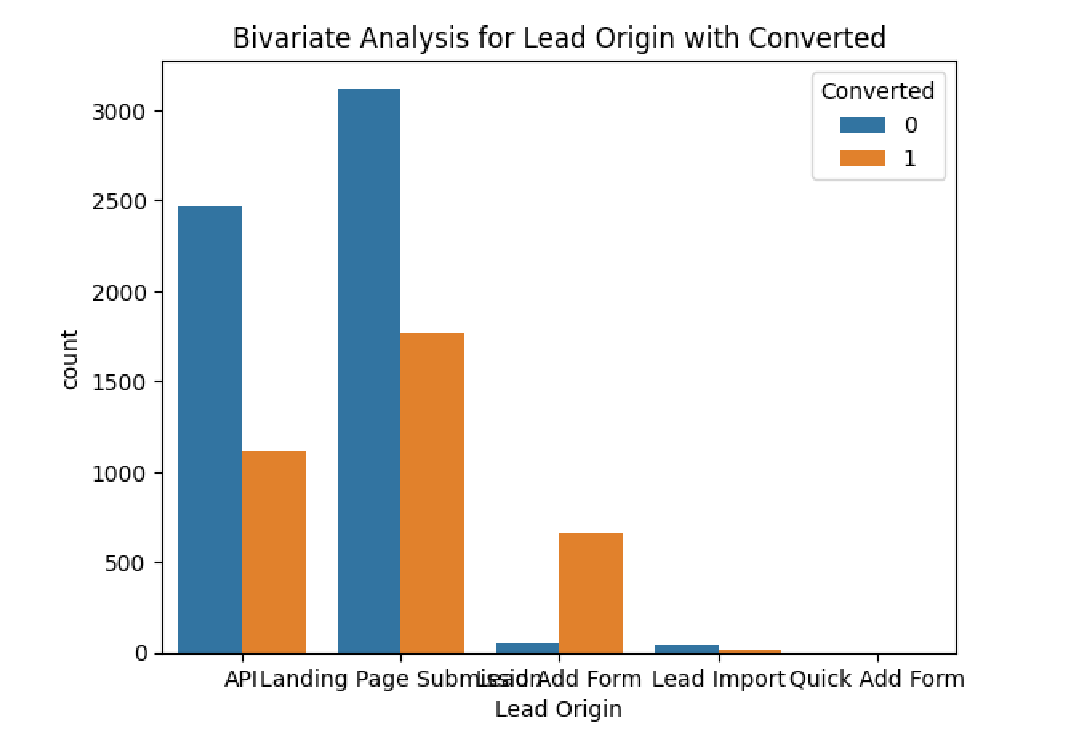

Checking Outliers using Boxplot
TotalVisits
250
200
150
100
50
0
Page Views Per Visit
50
40
30
20
10
0

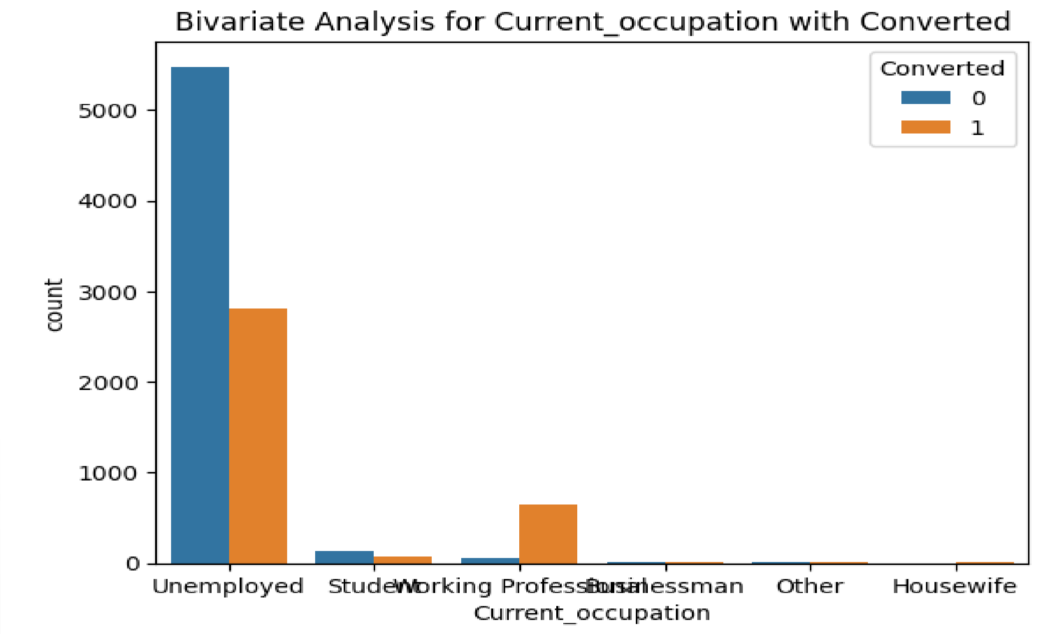

Total Time Spent on Website
2000
1500
1000
500
0

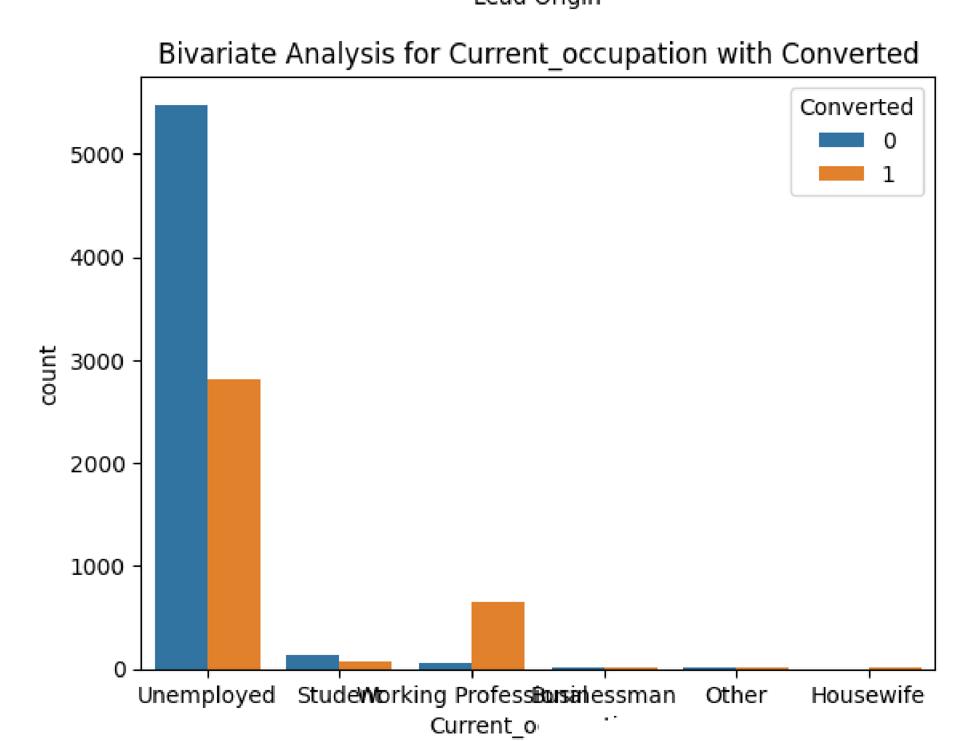

Checking Outliers using Boxplot
TotalVisits
2
0
8
6
4
2
0
Page Views Per Visit
7
6
5
4
3
2
1
0

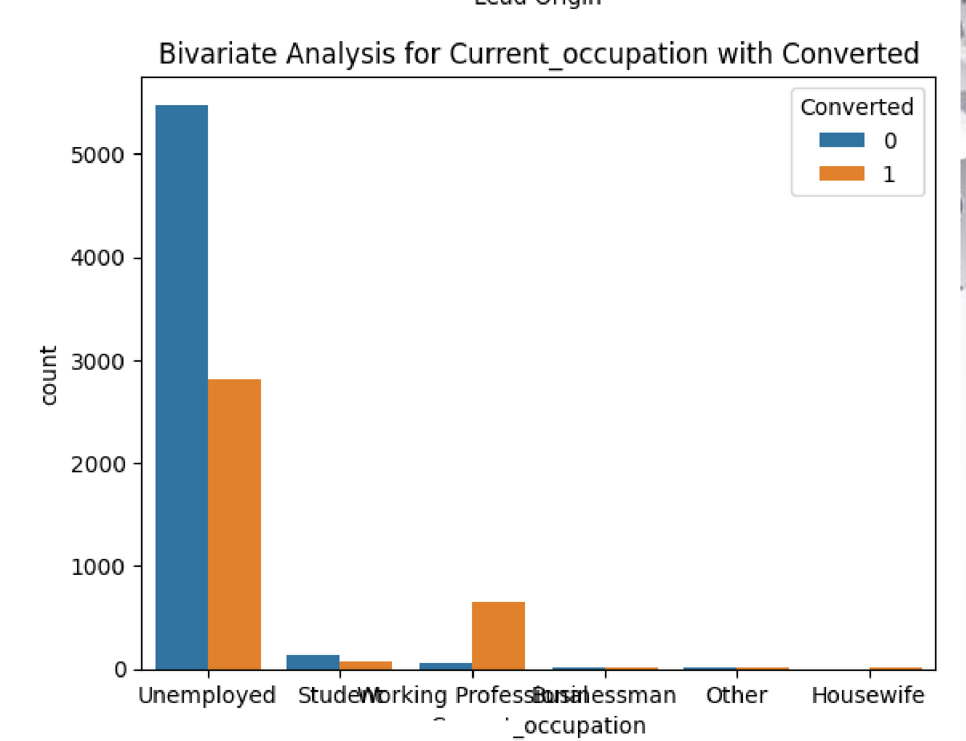

Total Time Spent on Website

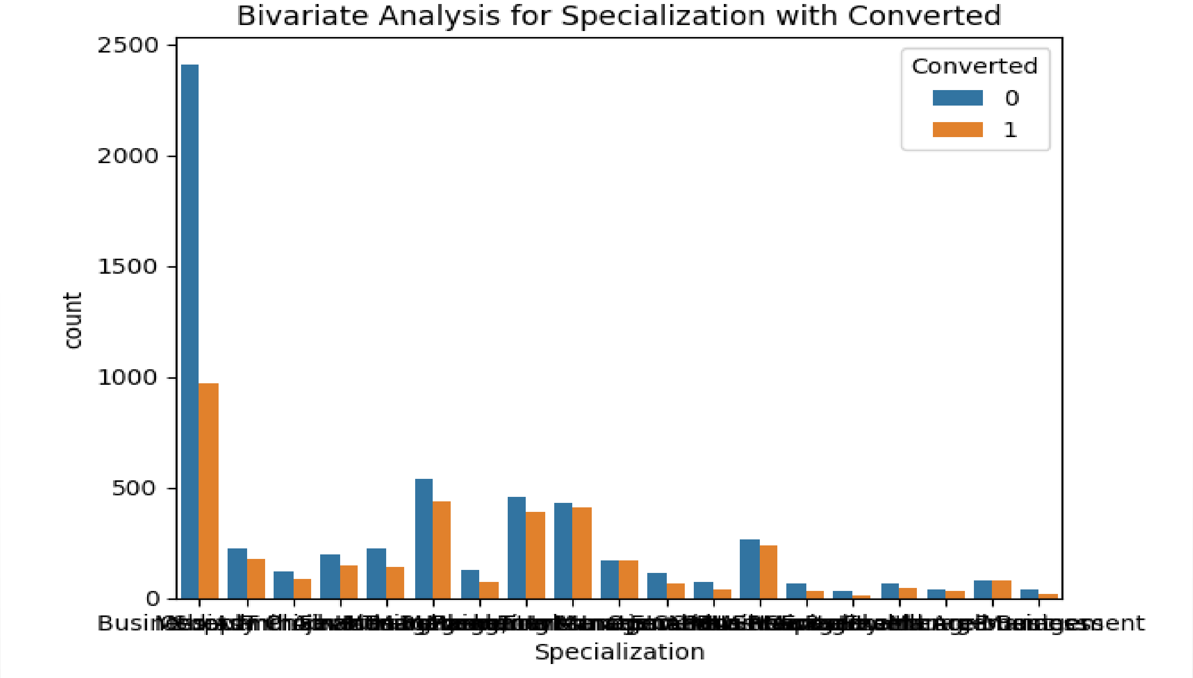

Leads Converted


Bivariate Analysis for Lead Origin with Converted

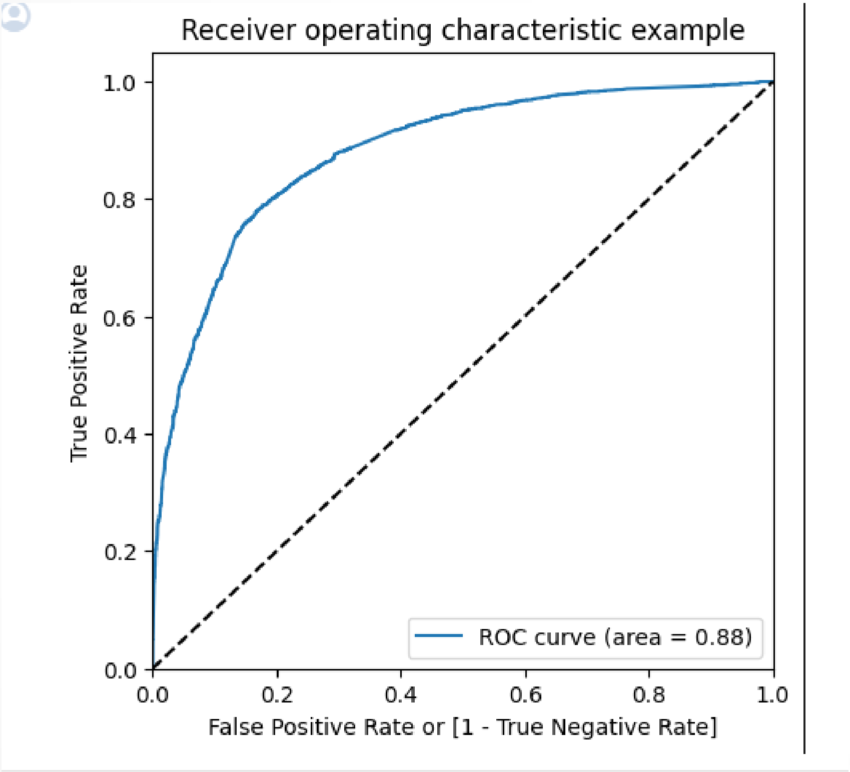

Bivariate Analysis for Current_occupation with Converted

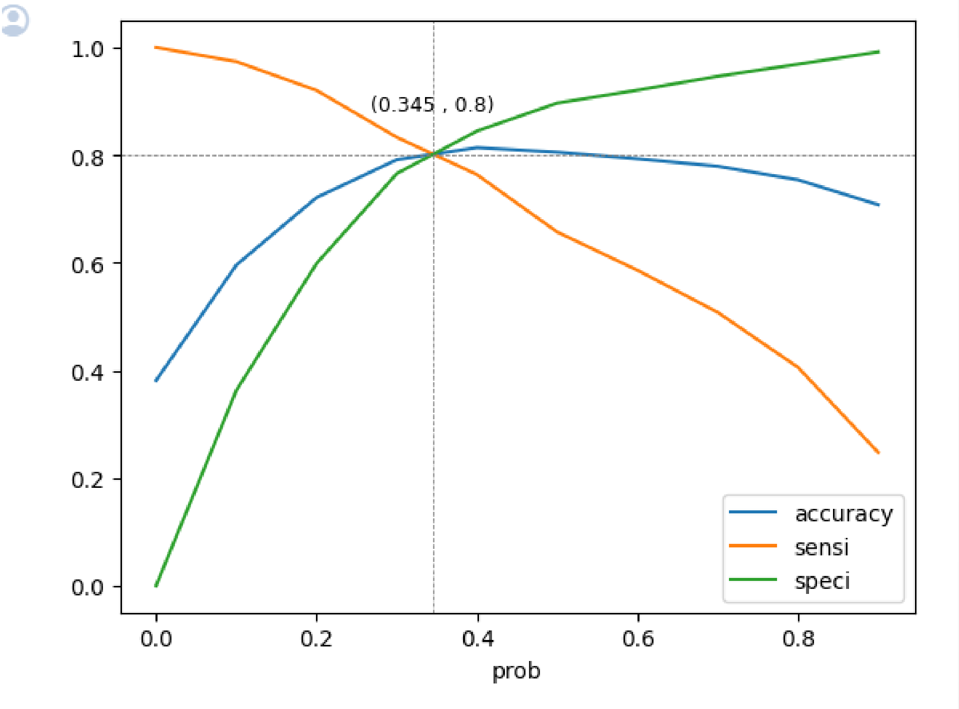

Lead Origin
Bivariate Analysis for Current_occupation with Converted
Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
Current_occupation

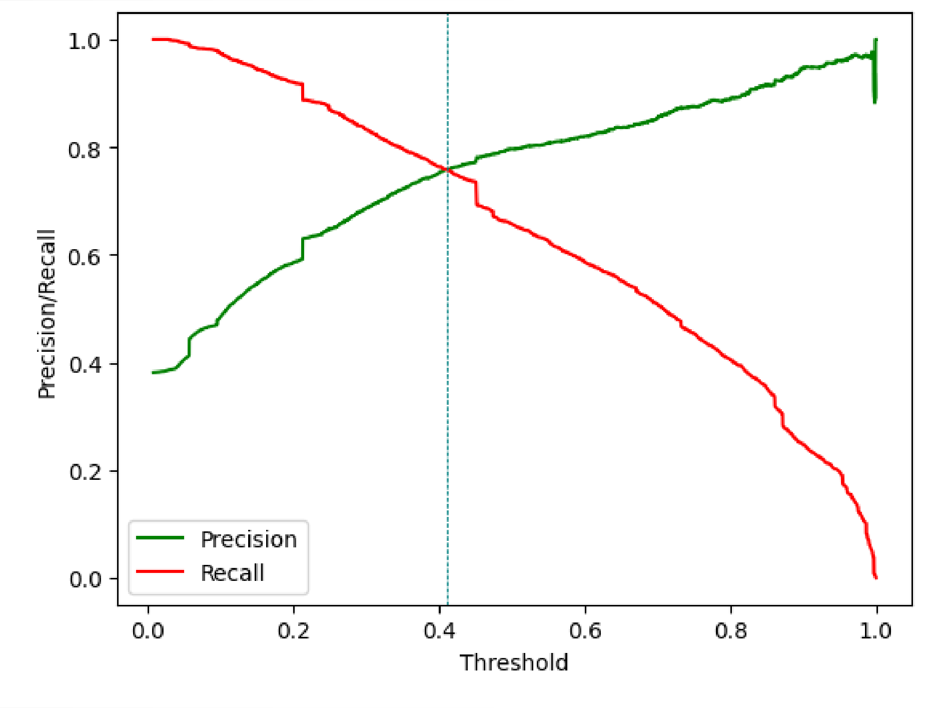

Lead Origin
Bivariate Analysis for Current_occupation with Converted
Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
Current_occupation

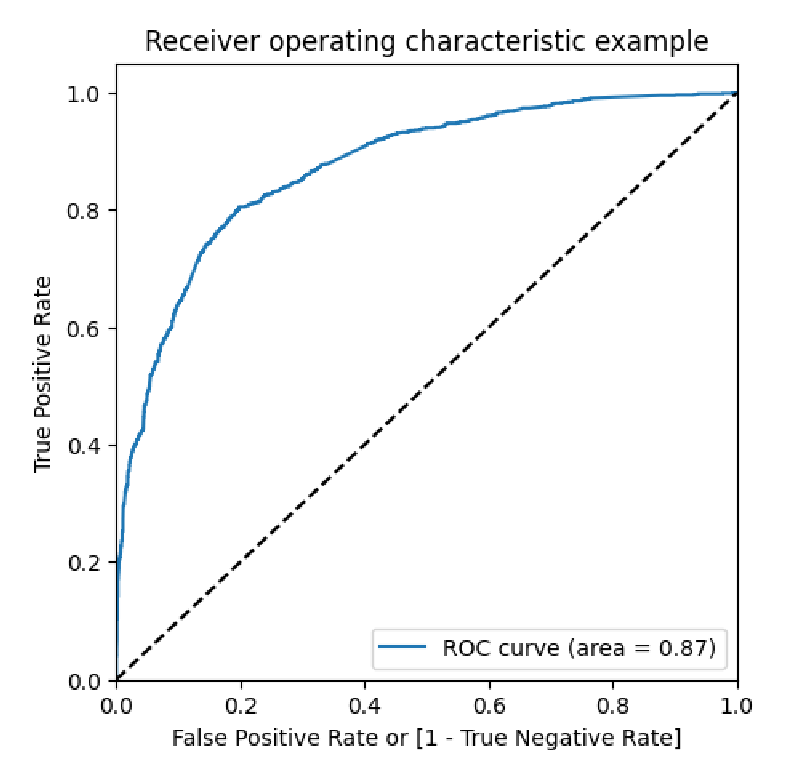

Bivariate Analysis for Specialization with Converted
Converted
0
1
count
2500
2000
1500
1000
500
0
Specialization

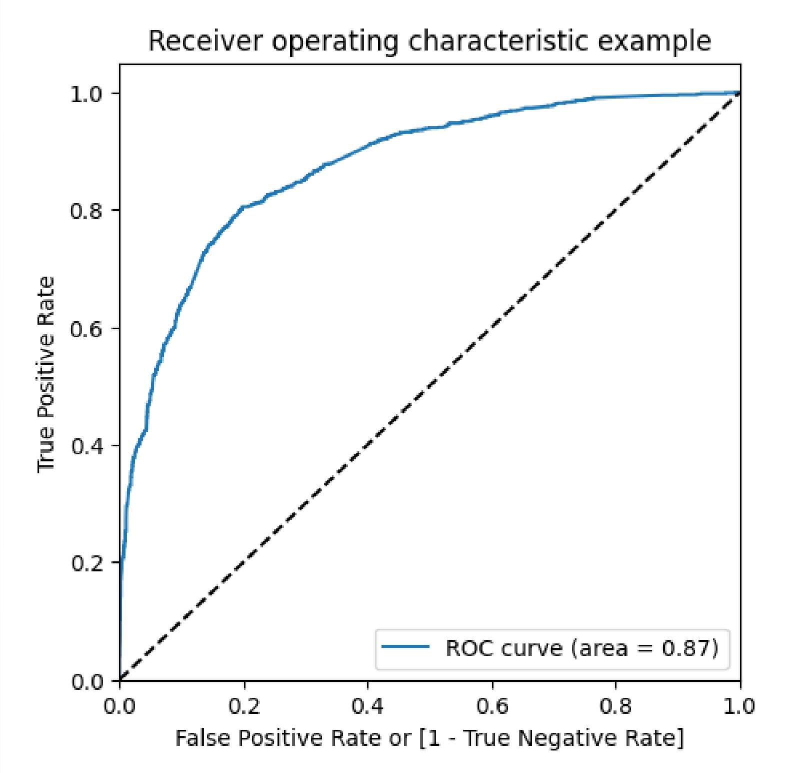
MODEL EVALUATION

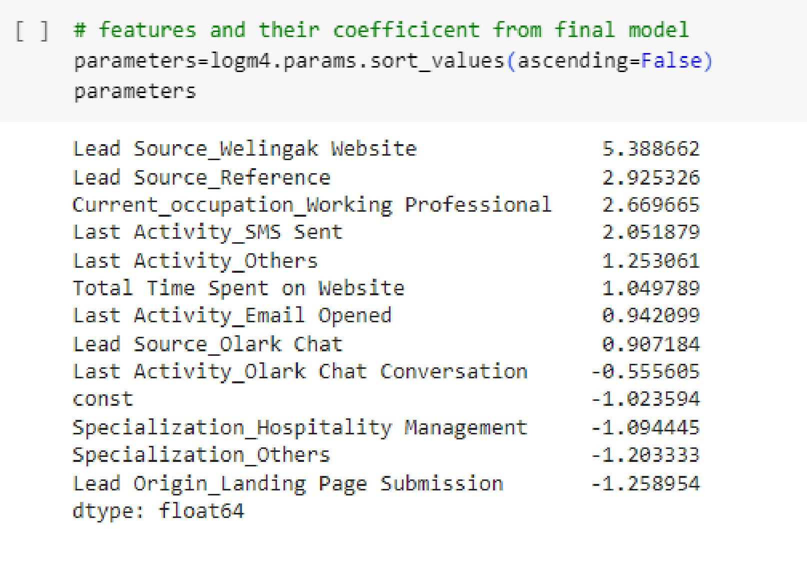

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)


(0.345 , 0.8)
prob
0.0
0.2
0.4
0.6
0.8
1.0
accuracy
sensi
speci


Precision/Recall
Threshold
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Precision
Recall

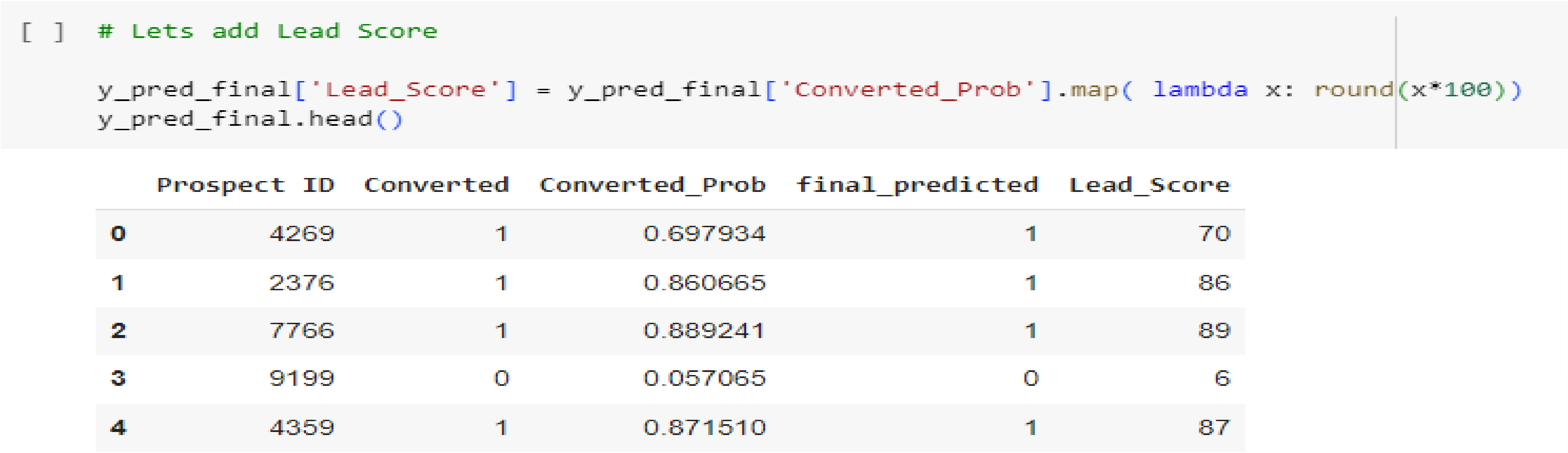

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.87)


Receiver operating characteristic example
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.87)



```
[ ]  # features and their coefficicent from final model
     parameters=logm4.params.sort_values(ascending=False)
     parameters
```

```
Lead Source_Welingak Website              5.388662
Lead Source_Reference                     2.925326
Current_occupation_Working Professional   2.669665
Last Activity_SMS Sent                    2.051879
Last Activity_Others                      1.253061
Total Time Spent on Website               1.049789
Last Activity_Email Opened                0.942099
Lead Source_Olark Chat                    0.907184
Last Activity_Olark Chat Conversation    -0.555605
const                                    -1.023594
Specialization_Hospitality Management    -1.094445
Specialization_Others                    -1.203333
Lead Origin_Landing Page Submission      -1.258954
dtype: float64
```



```
[ ]  # Lets add Lead Score

     y_pred_final['Lead_Score'] = y_pred_final['Converted_Prob'].map( lambda x: round(x*100))
     y_pred_final.head()
```

	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
0	4269	1	0.697934	1	70
1	2376	1	0.860665	1	86
2	7766	1	0.889241	1	89
3	9199	0	0.057065	0	6
4	4359	1	0.871510	1	87

Lead Score: Lead Score is assigned to the customers

- The customers with a higher lead score have a higher conversion chance
- The customers with a lower lead score have a lower conversion chance.

FINAL CONCLUSION

Evaluation Metrics :-

- Train set :
- o Accuracy -> 81.7%
- o Sensitivity-> 79.9%
- o Specificity-> 82.7%
- For Test set :
- o Accuracy : 79.8%
- o Sensitivity : 75.99%
- o Specificity : 82.15%
- Evaluation metrics in both test and train dataset are consistent. Therefore

Evaluation Metrics(Contd.) :-

- final model is performing good.
- Top 3 features contributing to predicting hot leads are:
- o Lead Origin_Lead Add Form
- o Current_occupation_Working Professional
- o Last Activity_SMS Sent

Recommendations:-

- Lead Origin, Current Occupation, and Last Activity are top contributors to lead conversion probability.
- Focus on Lead Add Form origin, Working Professional occupation, and Customer SMS activity for conversion.
- Improve Specialization-Others, Olark Chat Last Activity, and address issues with bounced emails.