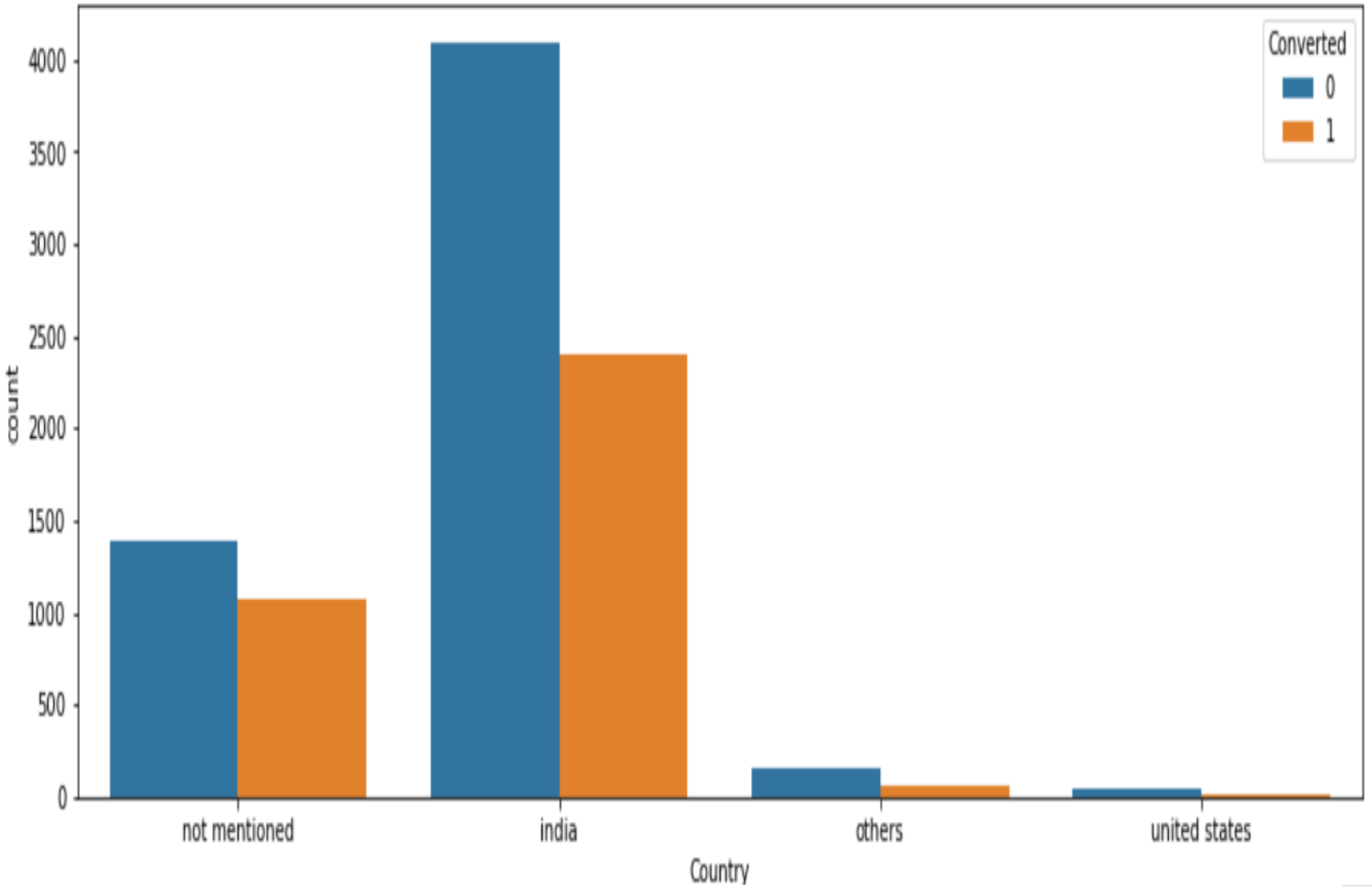
Lead Score Case Study

By Somnath Banerjee, Sri Harini & Harish B V

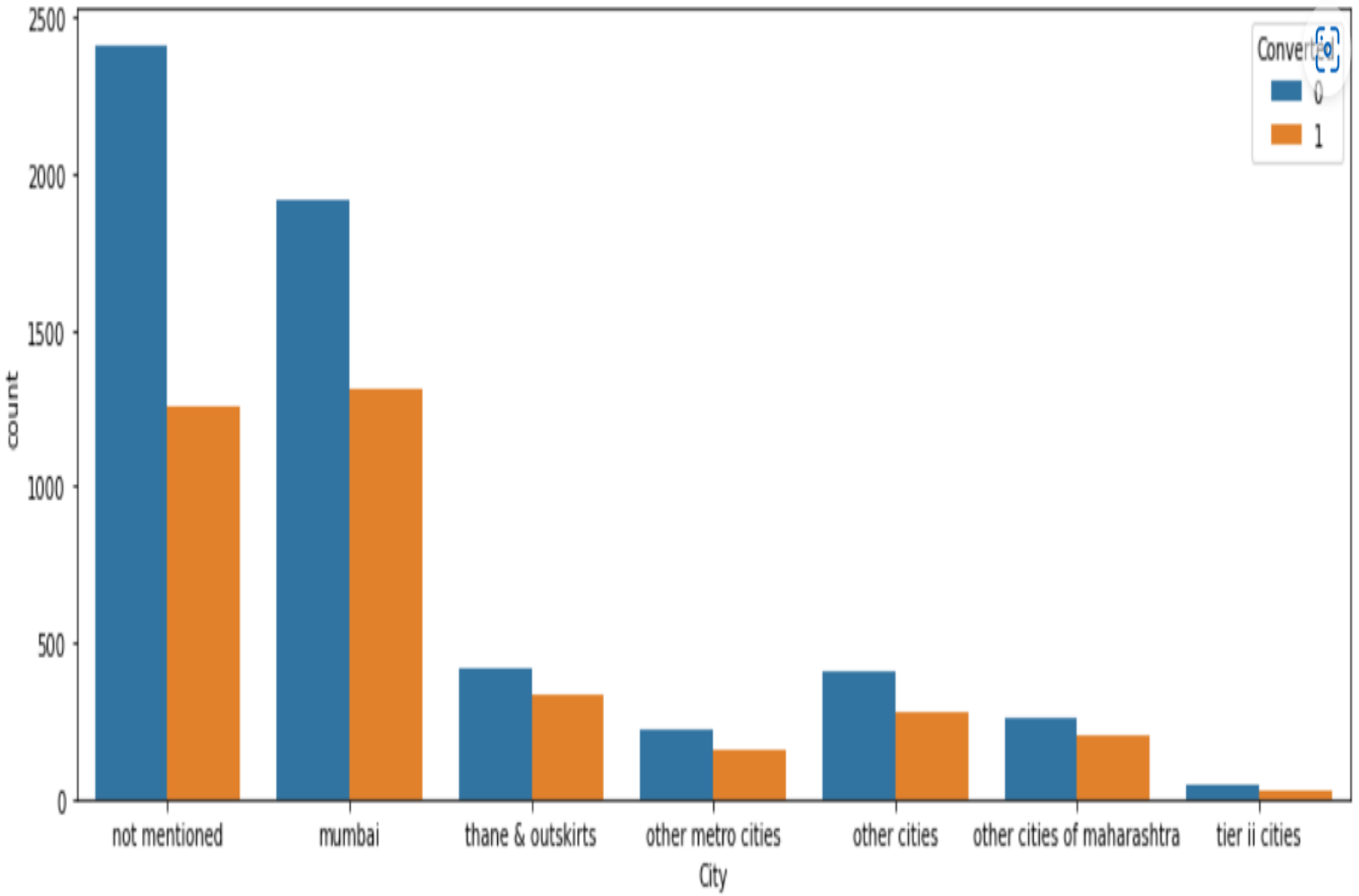
Univariate and Bivariate Analysis

Observation: As we can see converting ratio is higher than the all other countries



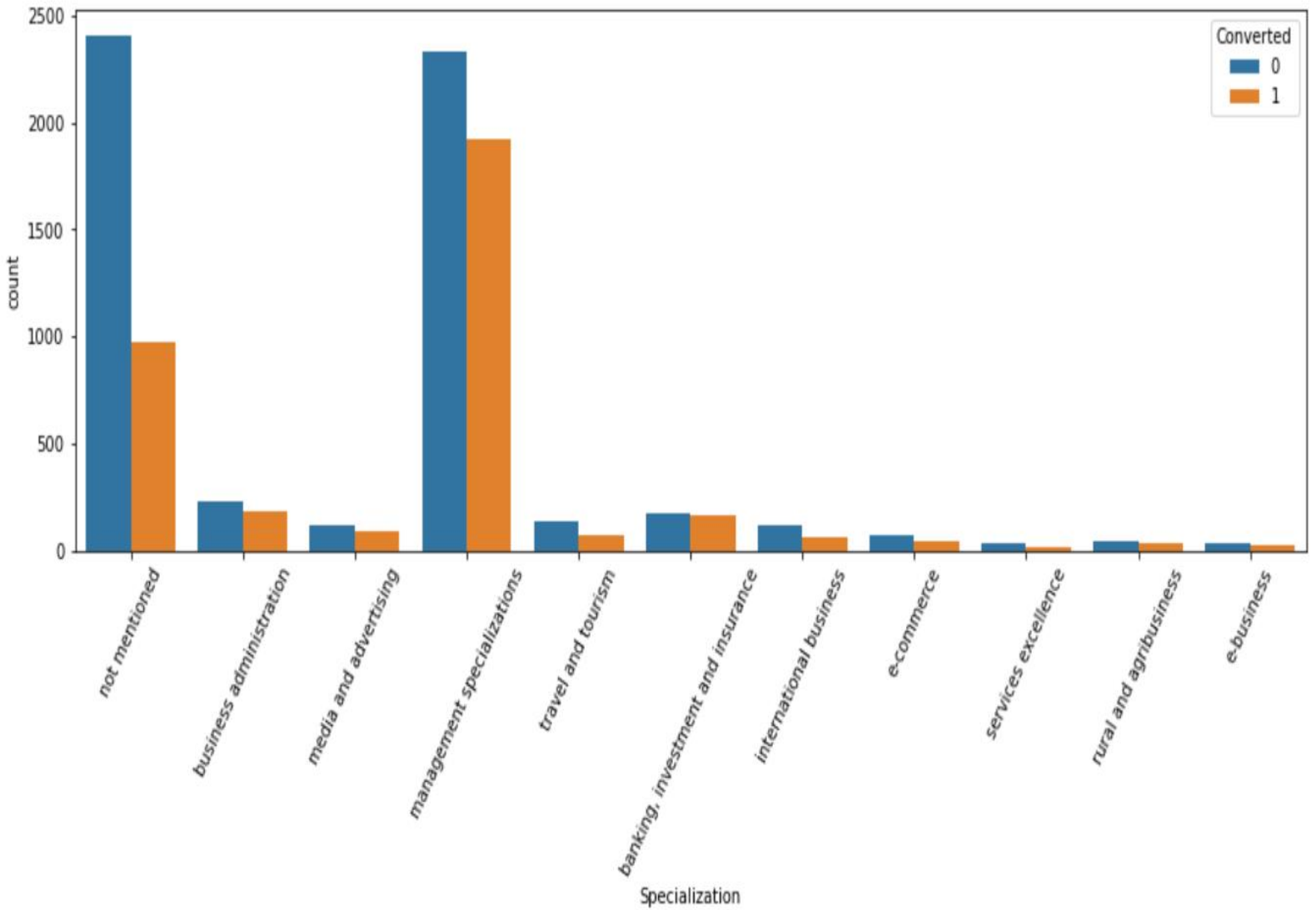
plotting spread of City column

Observation: City Bombay and Not mentioned category are having higher conversion rate in India



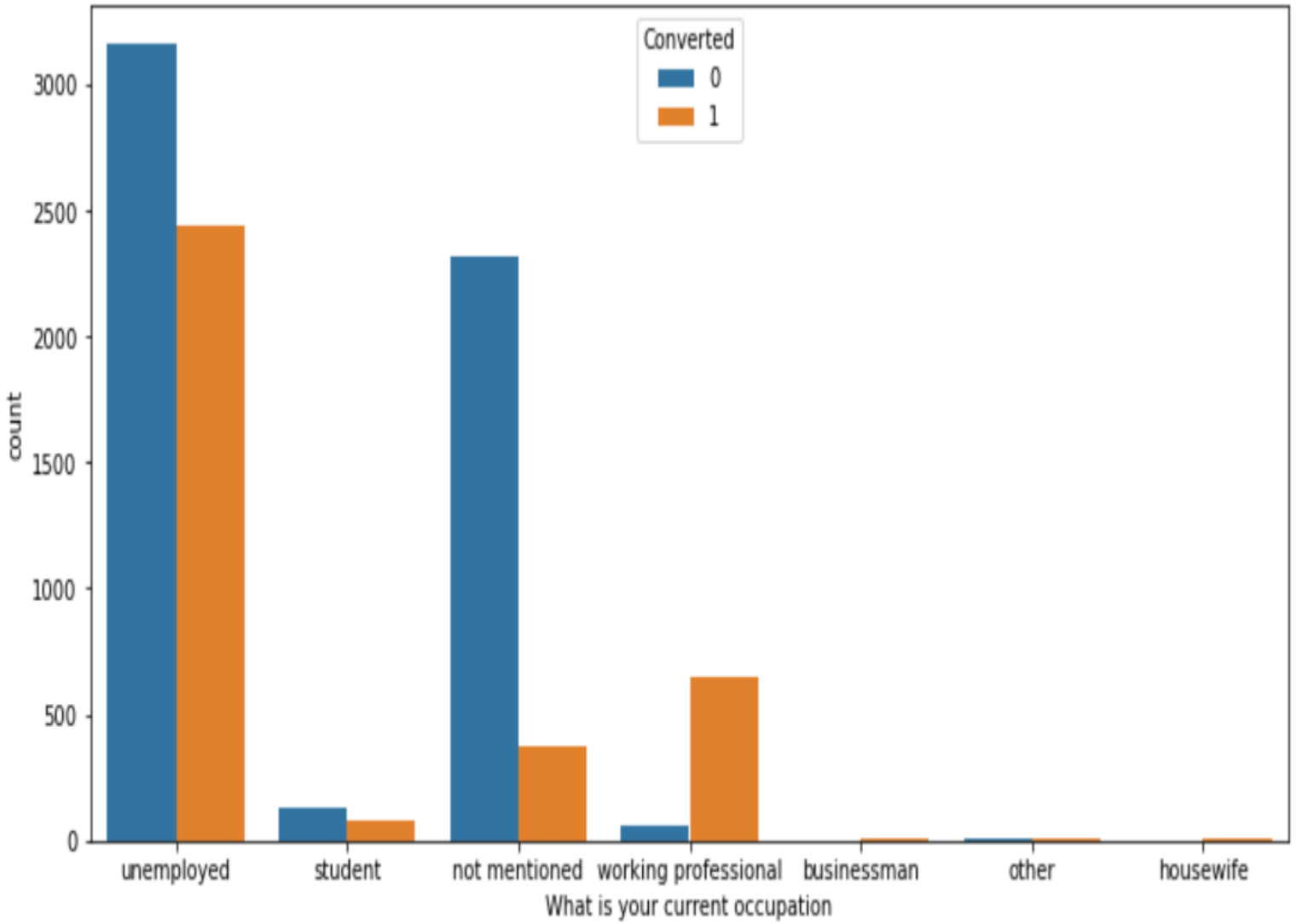
plotting spread of Specialization column

Observation: Management Specialization is having high conversion rate



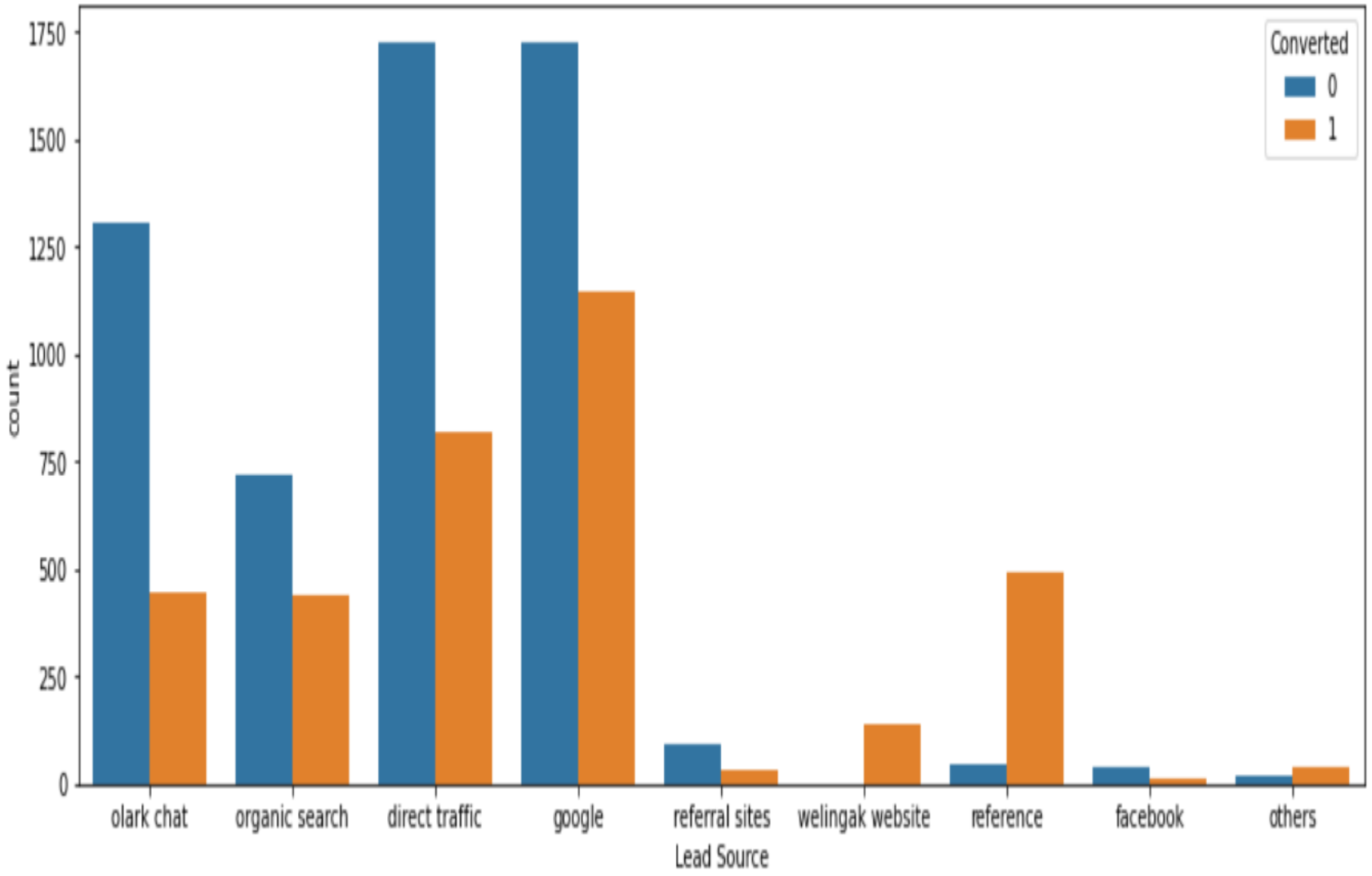
Visualizing count of Variable based on Converted value:

Unemployed candidates are having higher conversion rate among all the categories

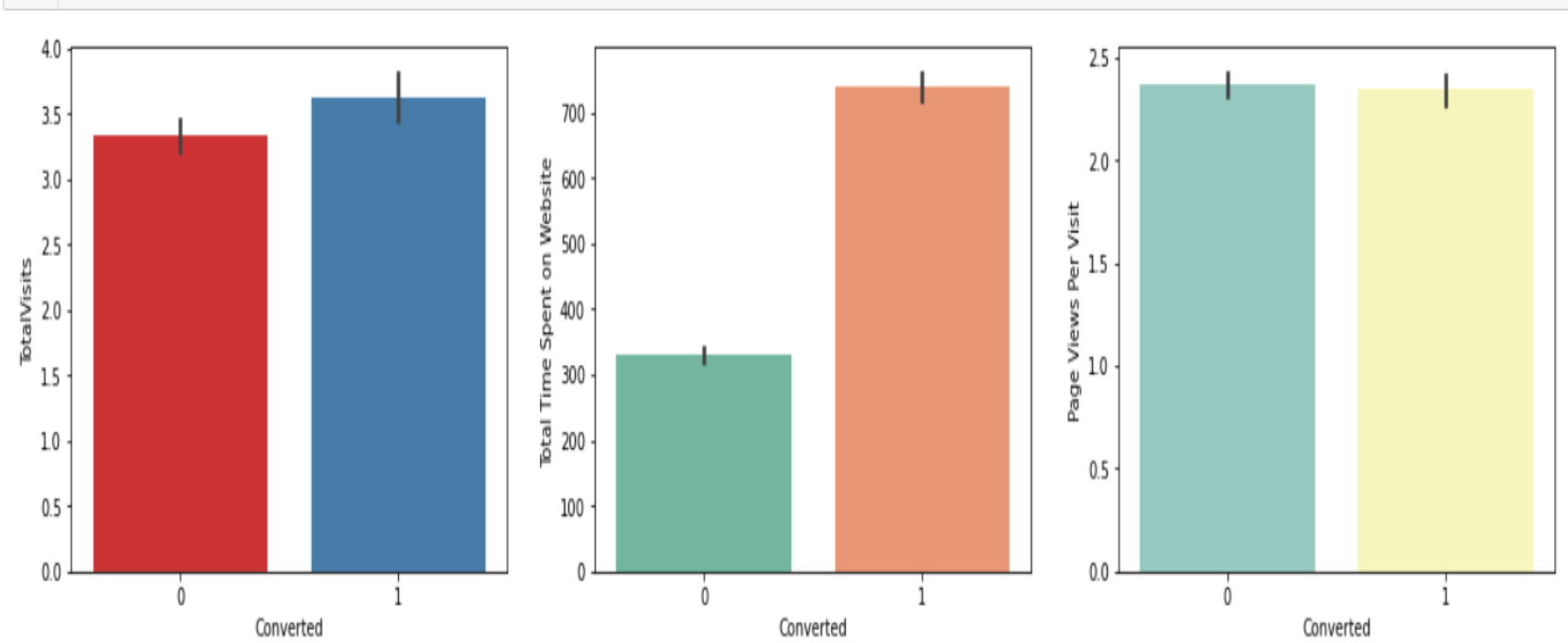


Visualizing count of Variable based on Converted value

Conversion Rate of reference leads and leads through welingak website is high.
To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

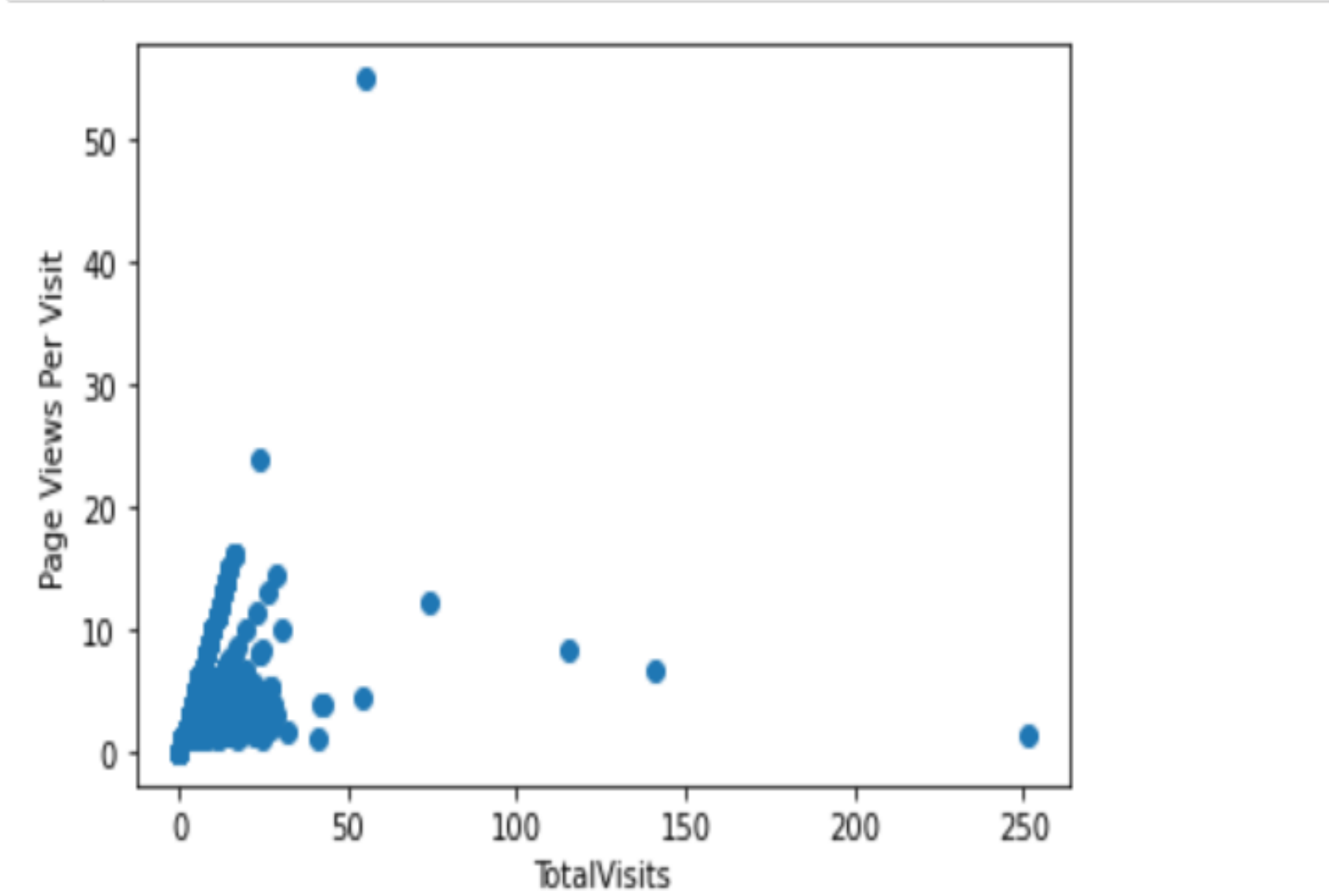


Check the conversions for all numeric values

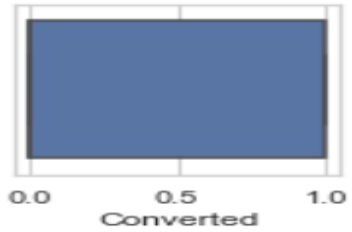


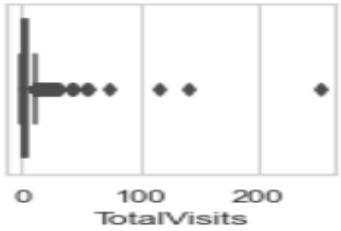
Outlier:

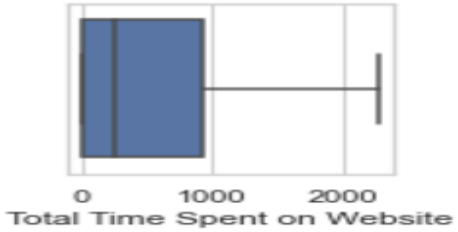
Scatter plot to see the distribution between Total no of visits and Page views per visits. We have seen the outliers as well as trend.

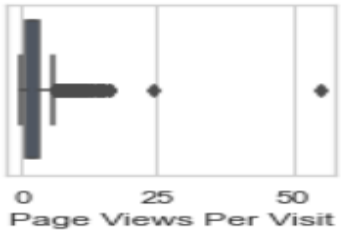


Outlier Treatment

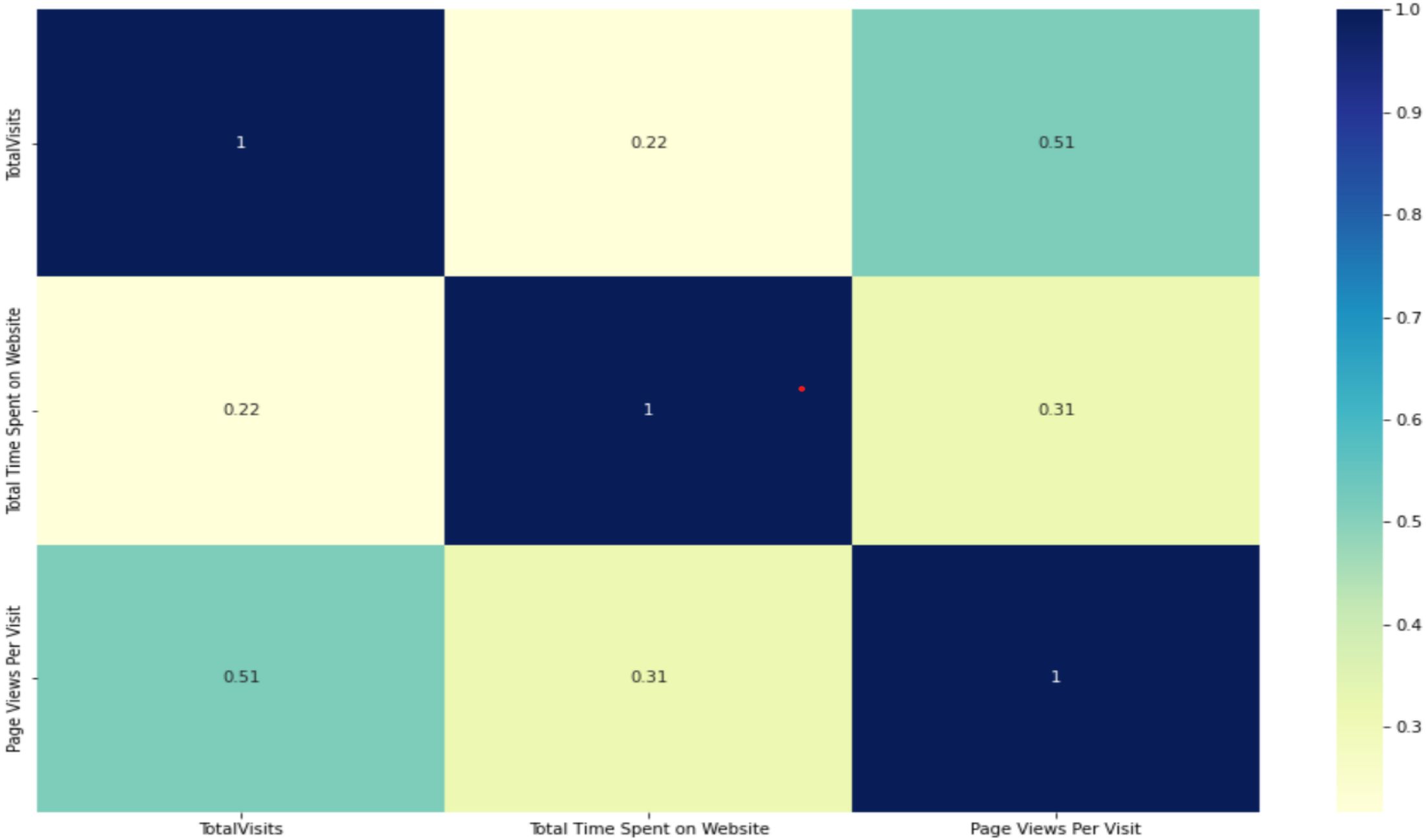








Correlation Matrix


TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.22
0.51
0.22
1
0.31
0.51
0.31
1
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3

Confusion Matrix

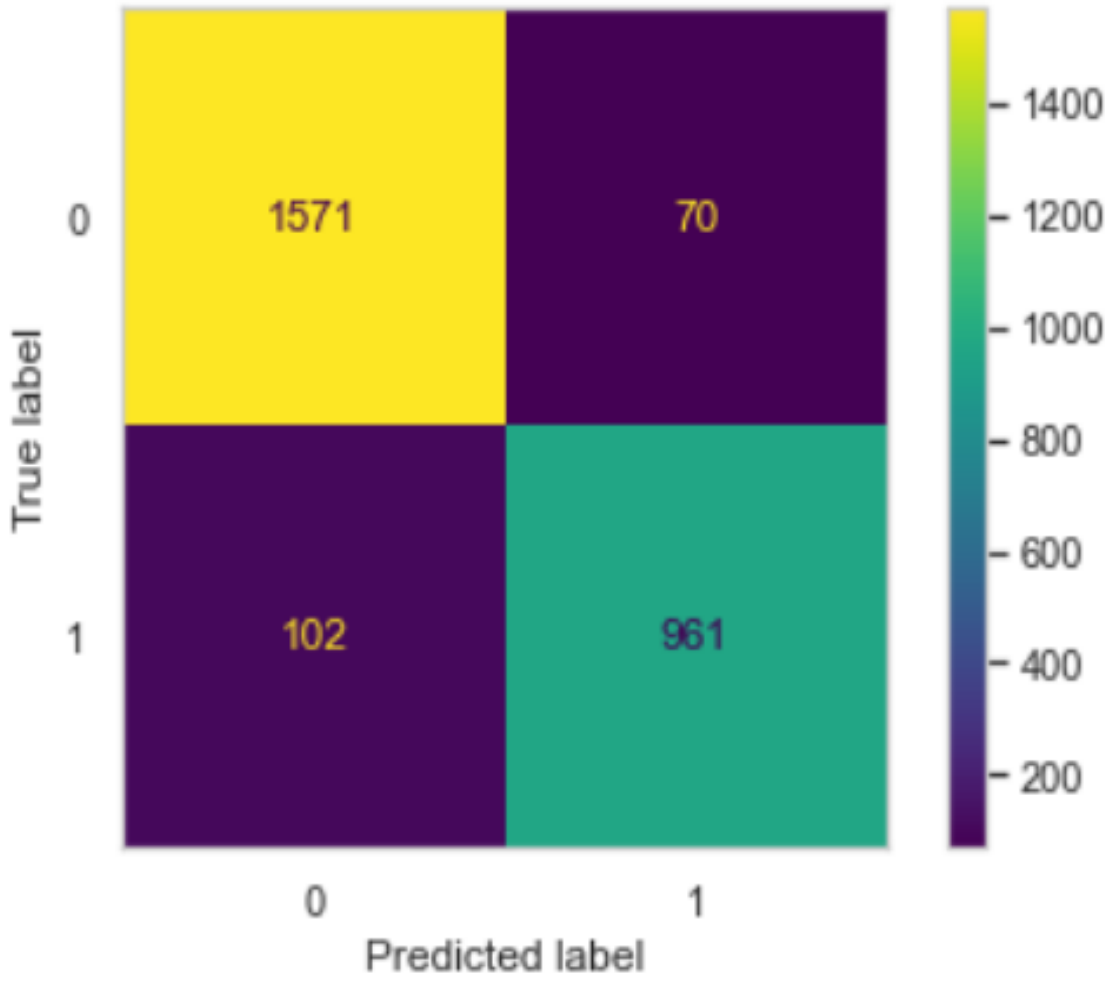
Insights:

ROC-AUC score test dataset: 0.98

precision score test dataset: 0.93

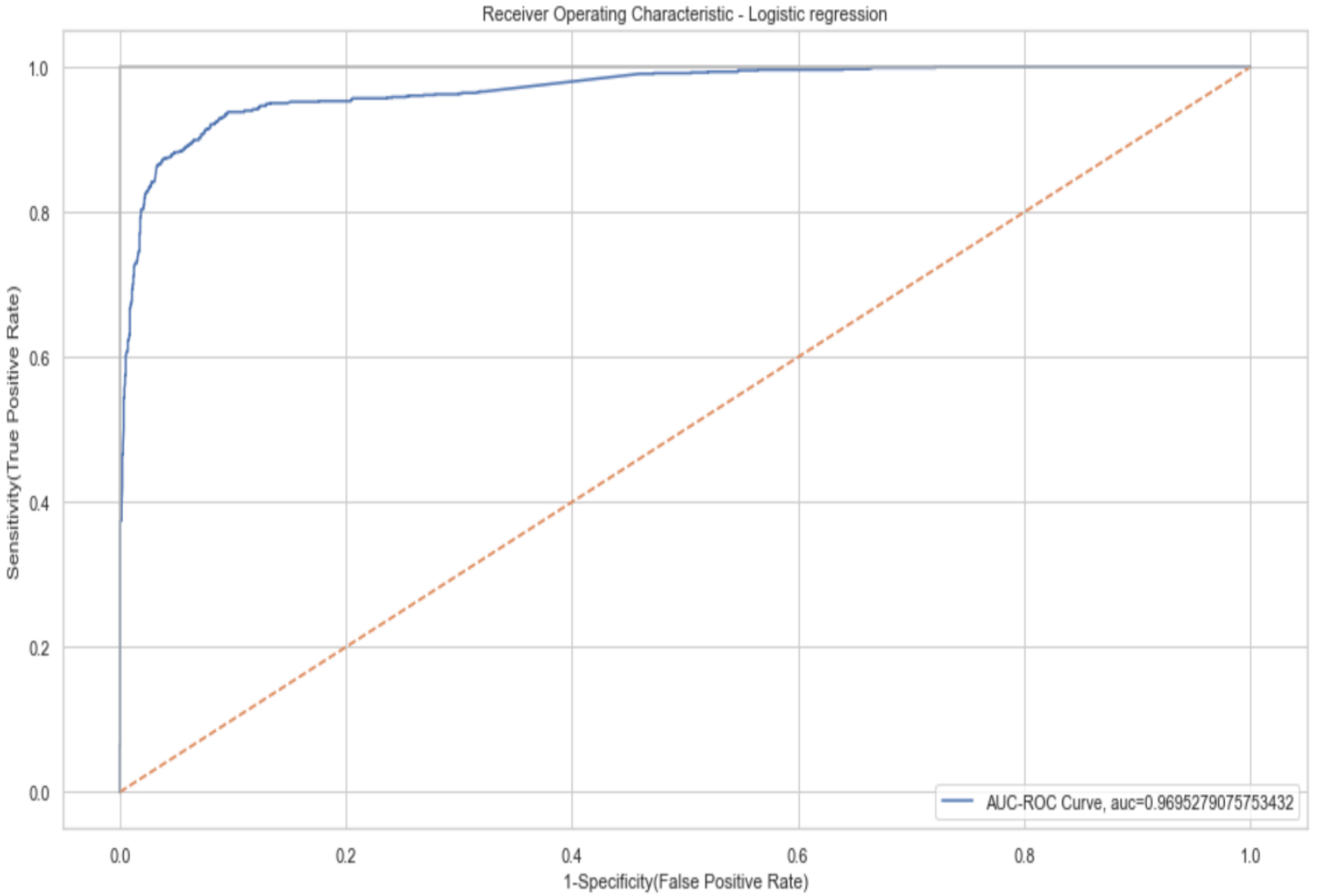
Recall score test dataset: 0.90

We have achieved a pretty good Recall score and ROC-AUC score



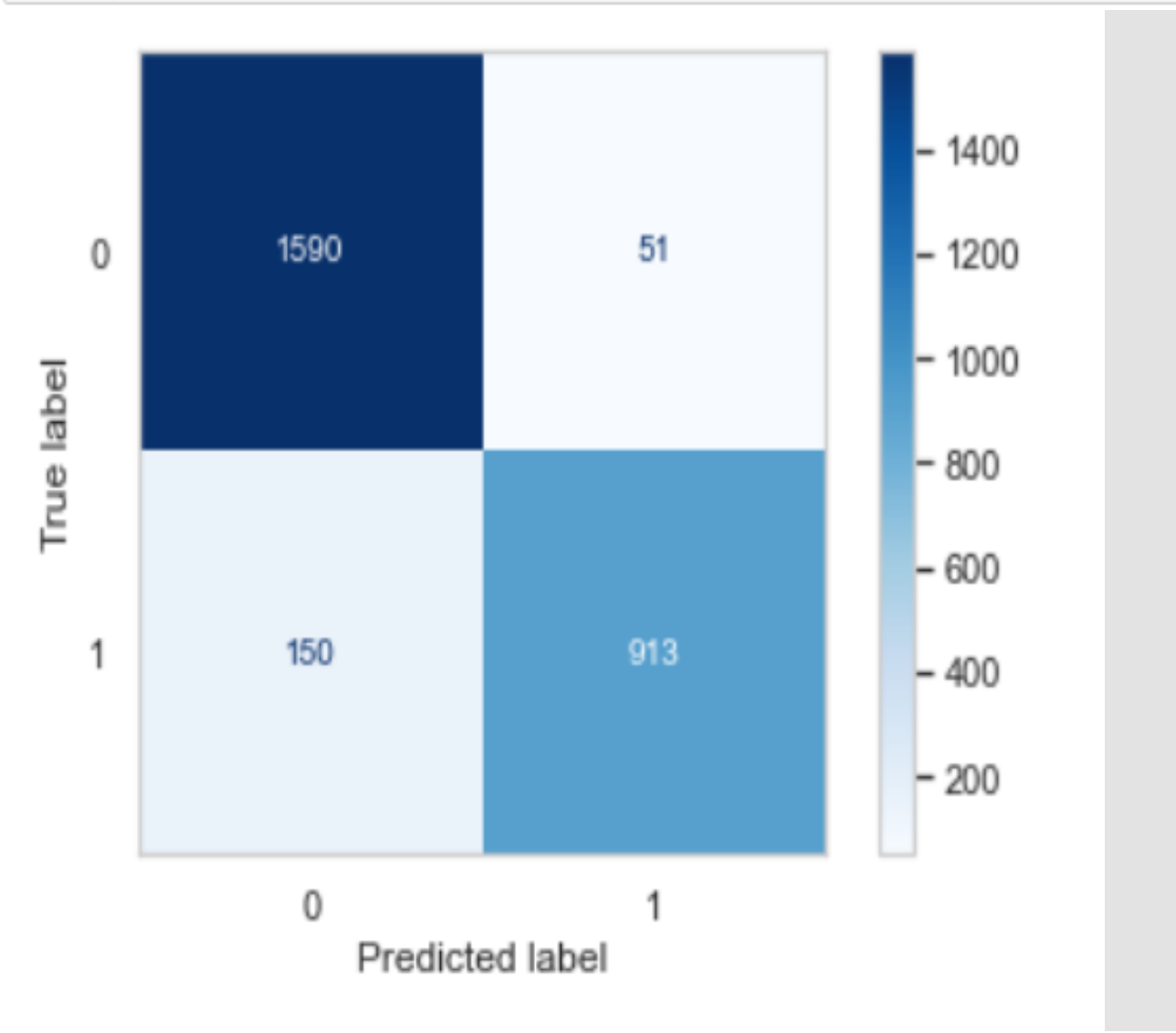
Specificity vs Sensitivity Trade-off

- Sensitivity is the proportion of True Positives or the proportion of cases correctly identified by the test as meeting a certain condition.

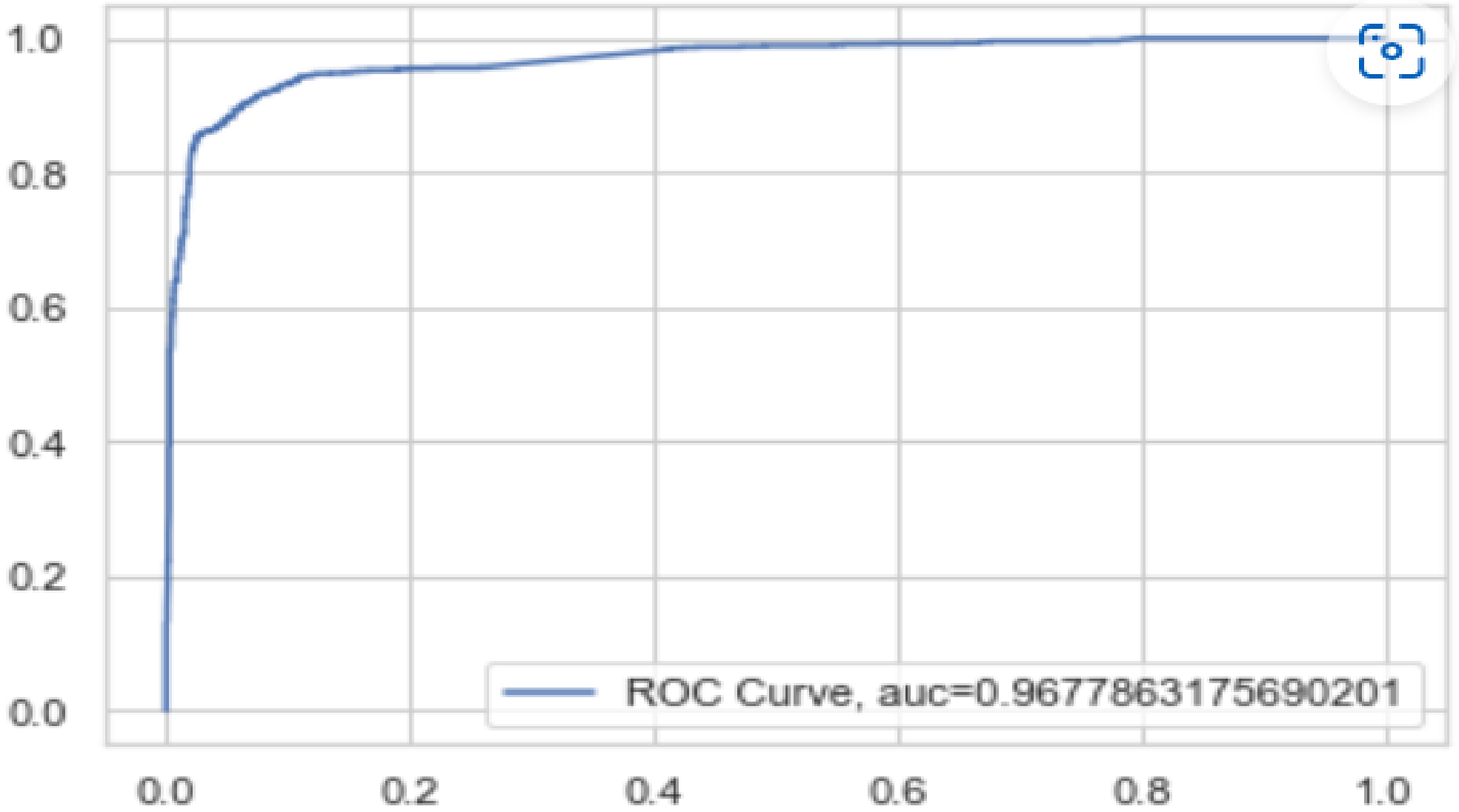


True Label vs Predict label

Precision: Precision is about being precise, i.e., how accurate your model is. In other words, you can say, when a model makes a prediction, how often it is correct.
Recall: If there are patients who have diabetes in the test set and your Logistic Regression model can identify it 83% of the time.



ROC Curve


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, auc=0.9677863175690201

Conclusion

We have balanced dataset in this case study and here, the ROC-AUC score on test dataset is 98.1%, which is mind blowing.

Features which contribute more towards the probability of a lead getting converted are:

1. Tags_will revert after reading the email
2. Total Time Spent on Website
3. Tags_ringin

Thank you