EDA Assignment Summary

by

Susmit Chakraborty

Debrathi Das

&

Samarjit Sinha

Problem Statement

You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

Goals of the Case Study

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

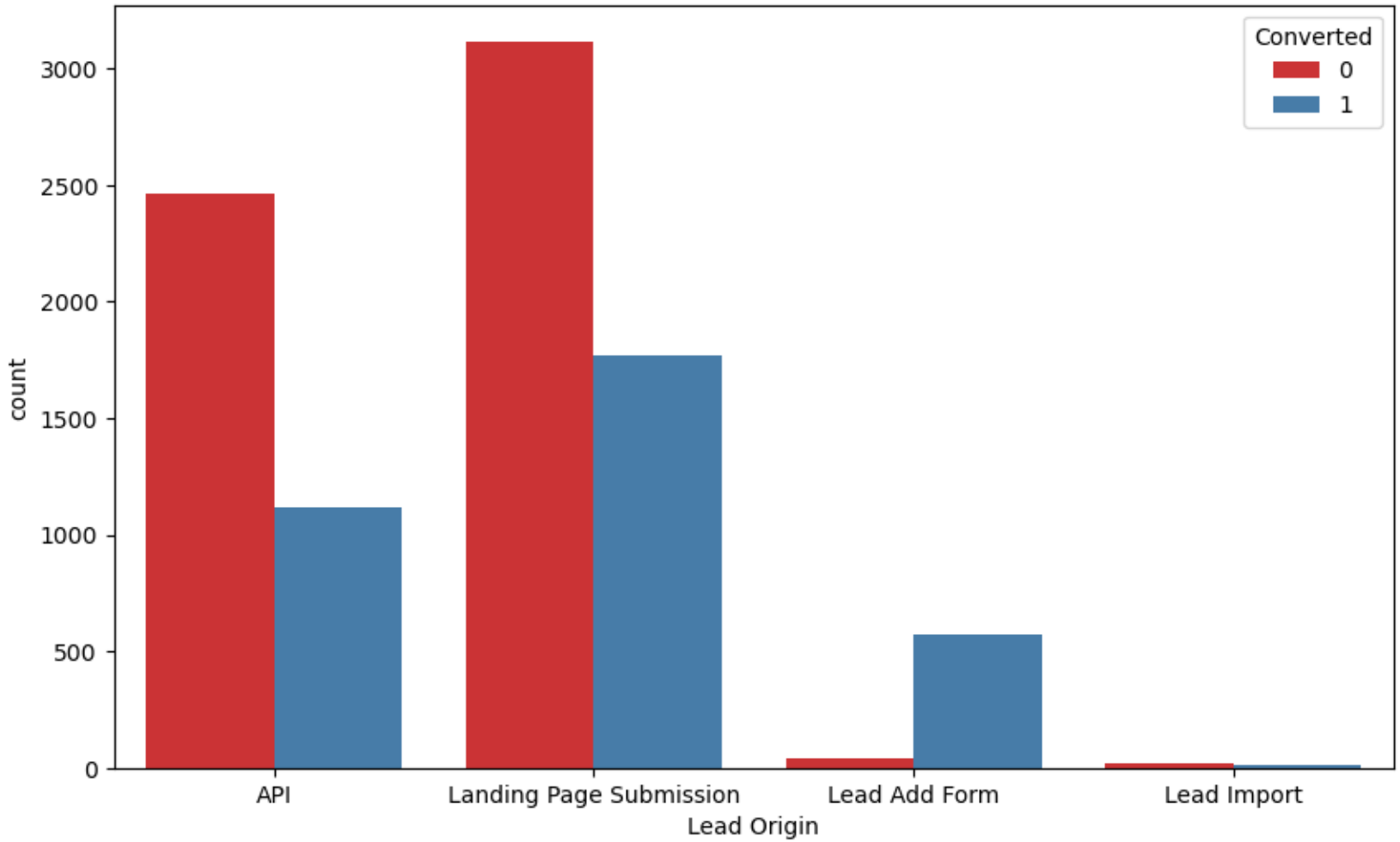
Approach & Methodology

1. Data importing and understanding
2. Checking Sanity for different columns
3. Removing the unimportant columns
4. Checking and handling the missing values with either mode/median
5. Univariate analysis and Bivariate analysis.
6. Data Preparation using 1/0 mapping and Dummy Variables incorporation
7. Model Building and Feature Elimination
8. Prediction
9. Model Evaluation using The ROC curve, Accuracy ,Sensitivity ,Specificity ,True Postive Rate ,False Postive Rate, Precision ,Recall
10. Predictions on the test dataset
11. The conclusion

Graphs and Insights

INSIIGH 1 :

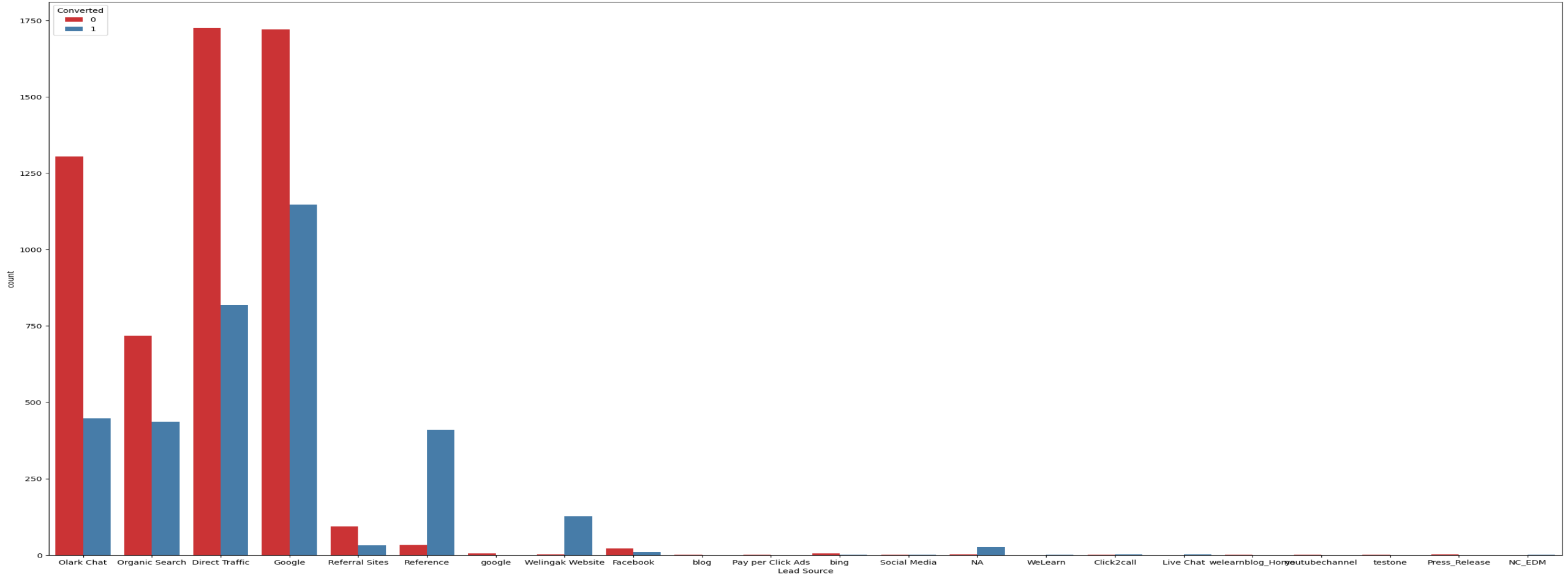
Landing Page Submission got maximum conversion rate



Graphs and Insights

INSIIGH 2:

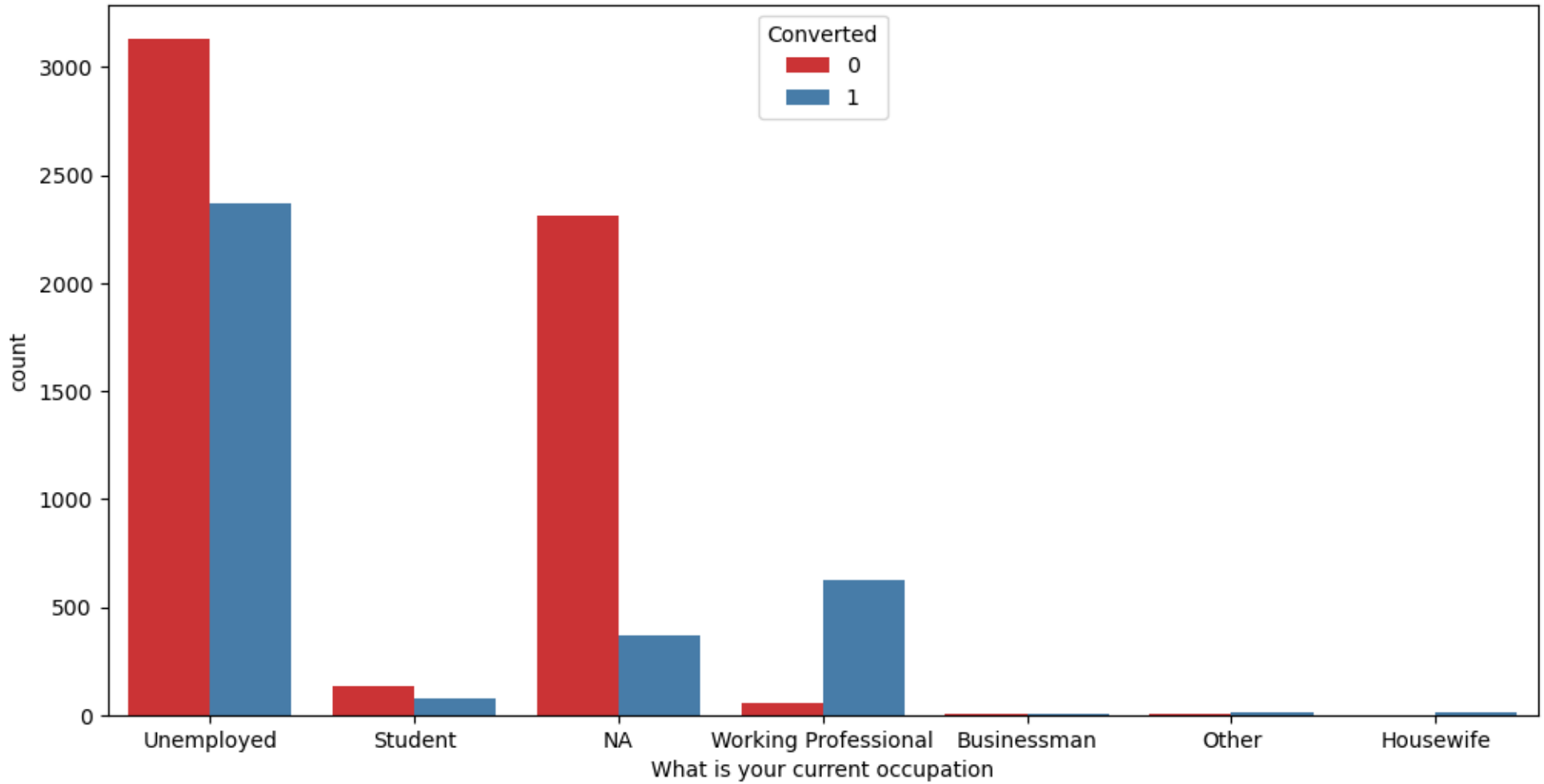
Google is the highest lead source that converts maximum.



Graphs and Insights

INSIIGH 3:

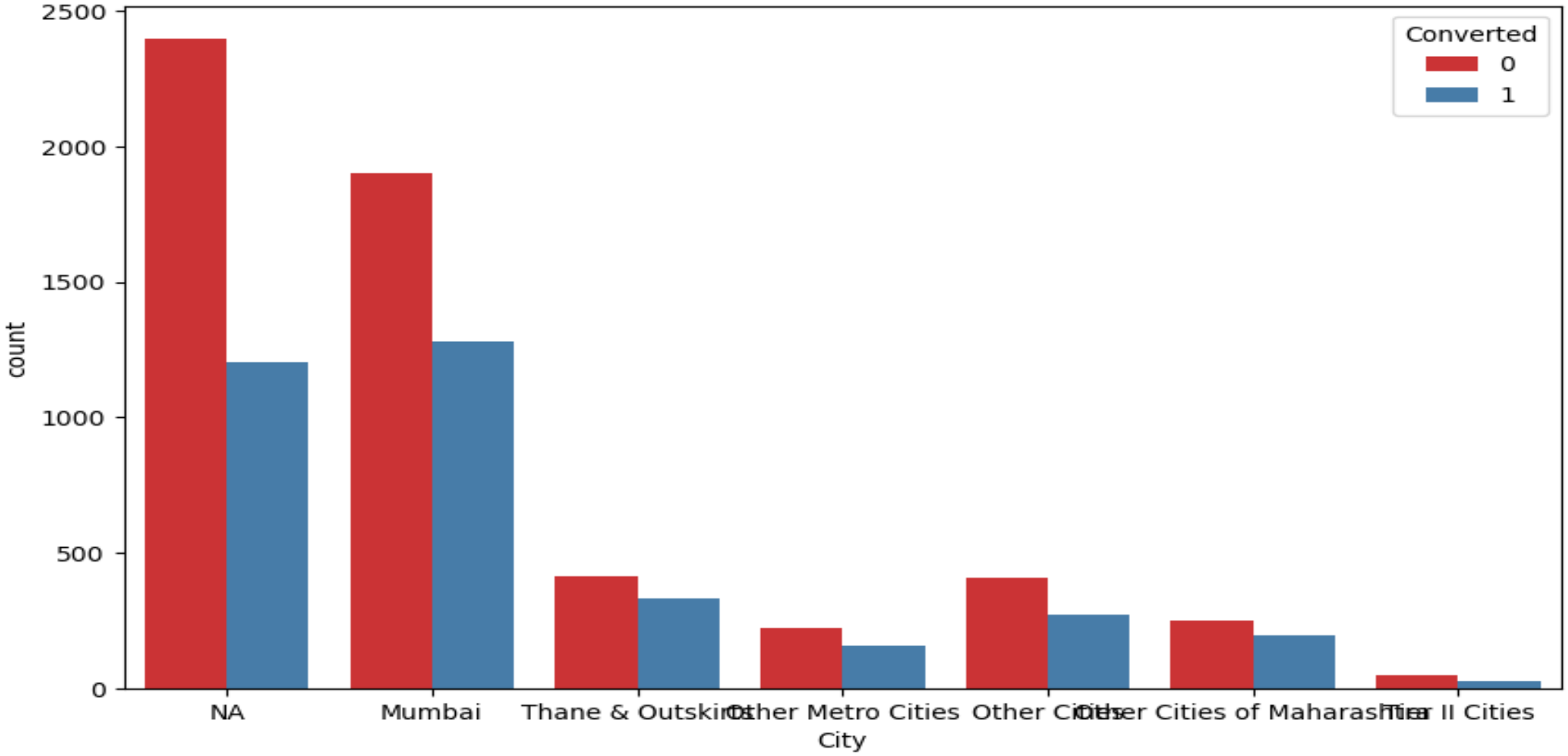
Unemployment got the maximum conversion rate



Graphs and Insights

INSIIGH 4:

Mumbai should be the targeted city



Graphs and Insights

INSIIGH 5 :

India is the highest successful targeted country

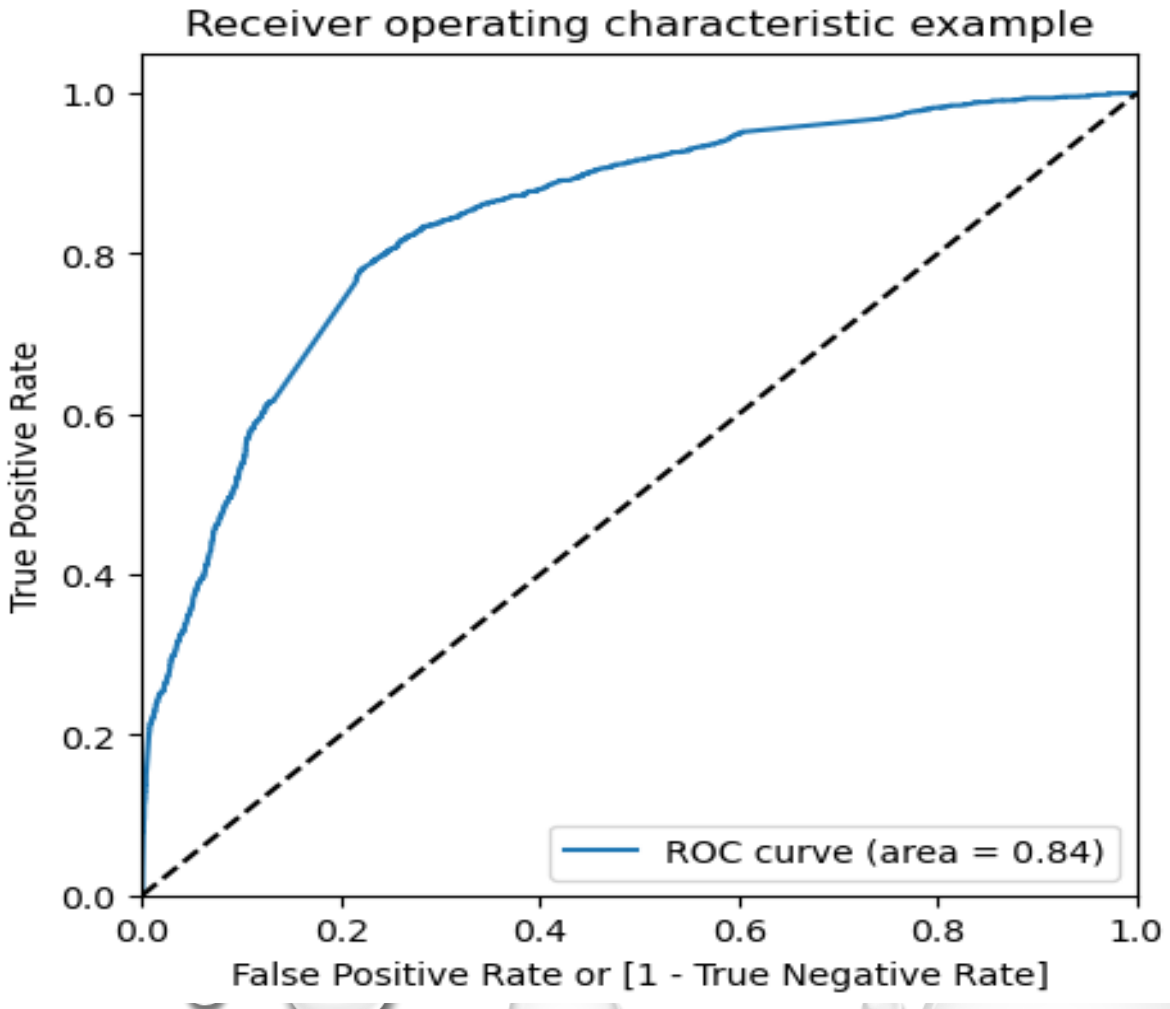
Final Model

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6360
Model Family:	Binomial	**Df Model:**	11
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3035.8
Date:	Sun, 13 Aug 2023	**Deviance:**	6071.6
Time:	21:49:52	**Pearson chi2:**	6.36e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3126
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.8306	0.087	-9.581	0.000	-1.001	-0.661
Total Time Spent on Website	1.1080	0.038	28.882	0.000	1.033	1.183
Lead Source_Direct Traffic	-1.6367	0.099	-16.496	0.000	-1.831	-1.442
Lead Source_Google	-1.2903	0.095	-13.597	0.000	-1.476	-1.104
Lead Source_NA	3.0460	1.035	2.944	0.003	1.018	5.074
Lead Source_Organic Search	-1.4195	0.117	-12.161	0.000	-1.648	-1.191
Lead Source_Referral Sites	-1.8580	0.335	-5.551	0.000	-2.514	-1.202
Lead Source_Welingak Website	4.1402	0.718	5.769	0.000	2.734	5.547
What is your current occupation_Other	1.4331	0.651	2.202	0.028	0.157	2.709
What is your current occupation_Student	1.3354	0.211	6.319	0.000	0.921	1.750
What is your current occupation_Unemployed	1.4837	0.081	18.378	0.000	1.325	1.642
What is your current occupation_Working Professional	4.1411	0.183	22.649	0.000	3.783	4.499

VIF & ROC

	Features	VIF
9	What is your current occupation_Unemployed	2.29
2	Lead Source_Google	1.63
1	Lead Source_Direct Traffic	1.57
4	Lead Source_Organic Search	1.27
10	What is your current occupation_Working Profes...	1.12
0	Total Time Spent on Website	1.10
6	Lead Source_Welingak Website	1.06
8	What is your current occupation_Student	1.04
5	Lead Source_Referral Sites	1.02
3	Lead Source_NA	1.01
7	What is your current occupation_Other	1.01



Results

Predictions on the train dataset

- The ROC curve has a value of 0.84
- Accuracy : 77.12%
- Sensitivity :82.51%
- Specificity : 72.57%
- True Postive Rate : 82.51%
- False Postive Rate : 27.42%
- Precision : 64.80%
- Recall : 82.51%

Predictions on the test dataset

- Accuracy : 76.82%
- Sensitivity :82.24%
- Specificity : 73.47%
- Precision : 65.67%
- Recall : 82.24%
- True Postive Rate : 82.24%
- False Postive Rate : 26.52%

Conclusion

After Model building and evaluation, it can be concluded that there are lots of insights in the database, but only important are prescribed the previous slides.

Besides the previous insights, here all are given below:

1. Based on the coefficient values from the Final model, the following variables contribute most:
- Lead Source
- What is your current occupation
- Total Time Spent on Website

2. Based on the coefficient values from the Final model, the following dummy variables contribute most:
- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- What is your current occupation_Unemployed

THANK YOU