



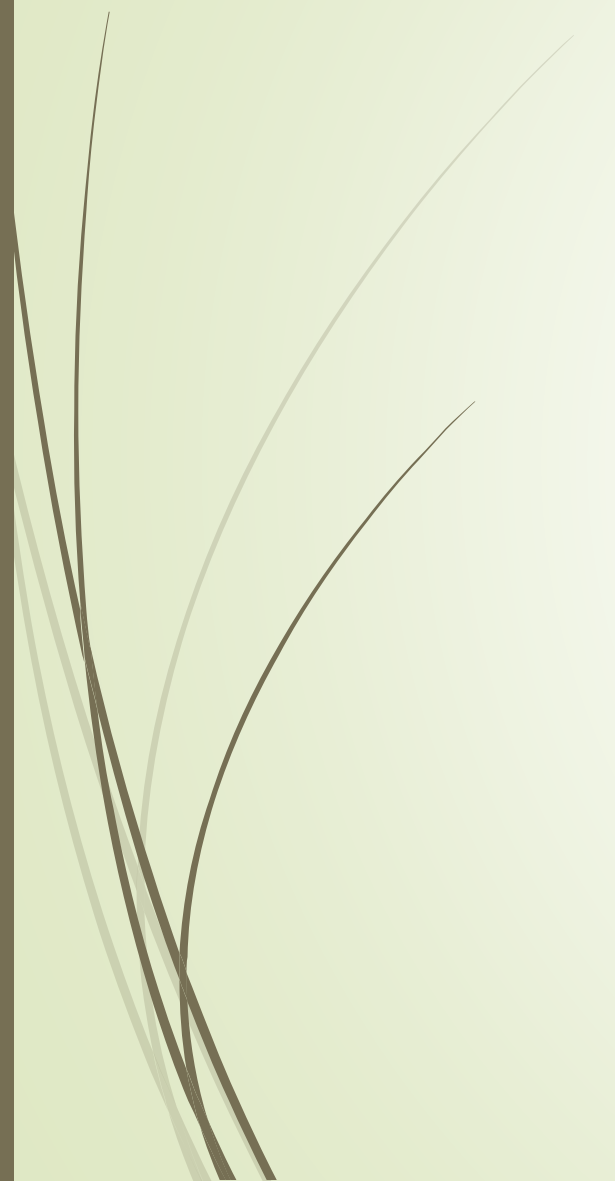
LEAD SCORING CASE STUDY

- Arun Tyagi & Hitik Negi



TABLE CONTENTS

Problem Statement	3
Objectives	4
Approach	5
Data Insights	7
Factors responsible for driving leads	17
Model Metrics	24
Conclusion	26



Problem Statement

- ❑ An education company named X Education sells online courses to industry professionals.
- ❑ Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- ❑ The company wants to increase the conversion rate to 80% as a target given by the CEO.



Objectives

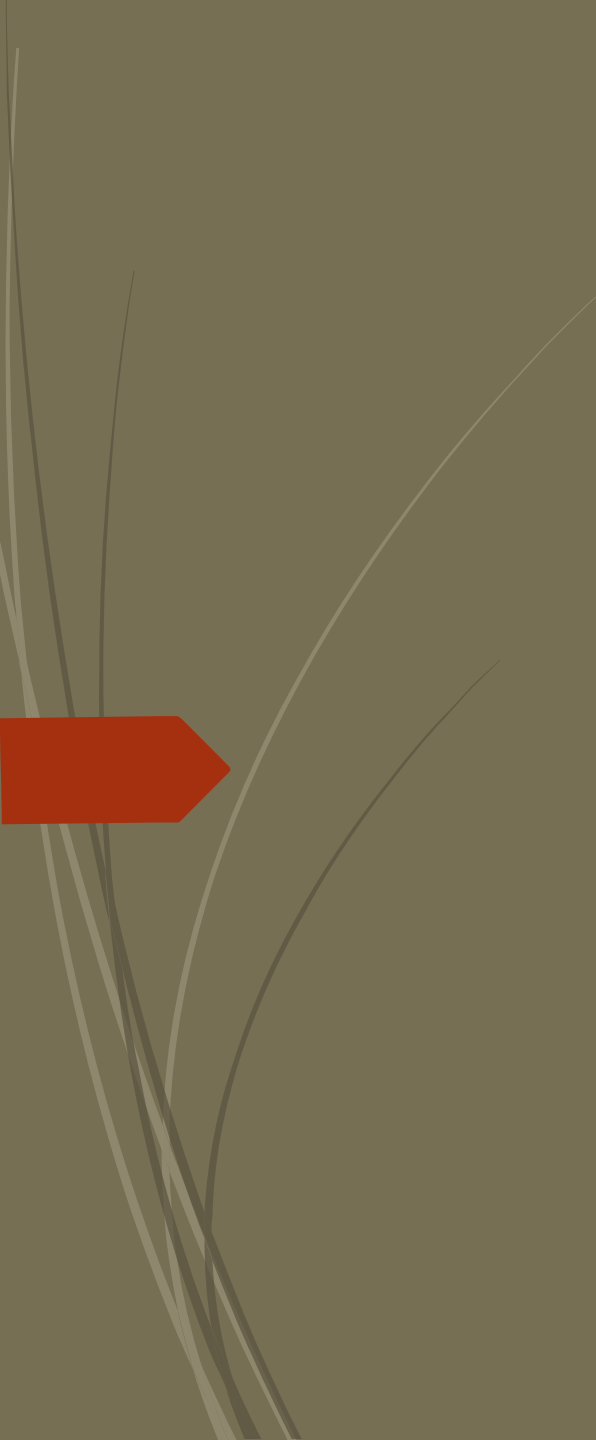
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

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.

Approach

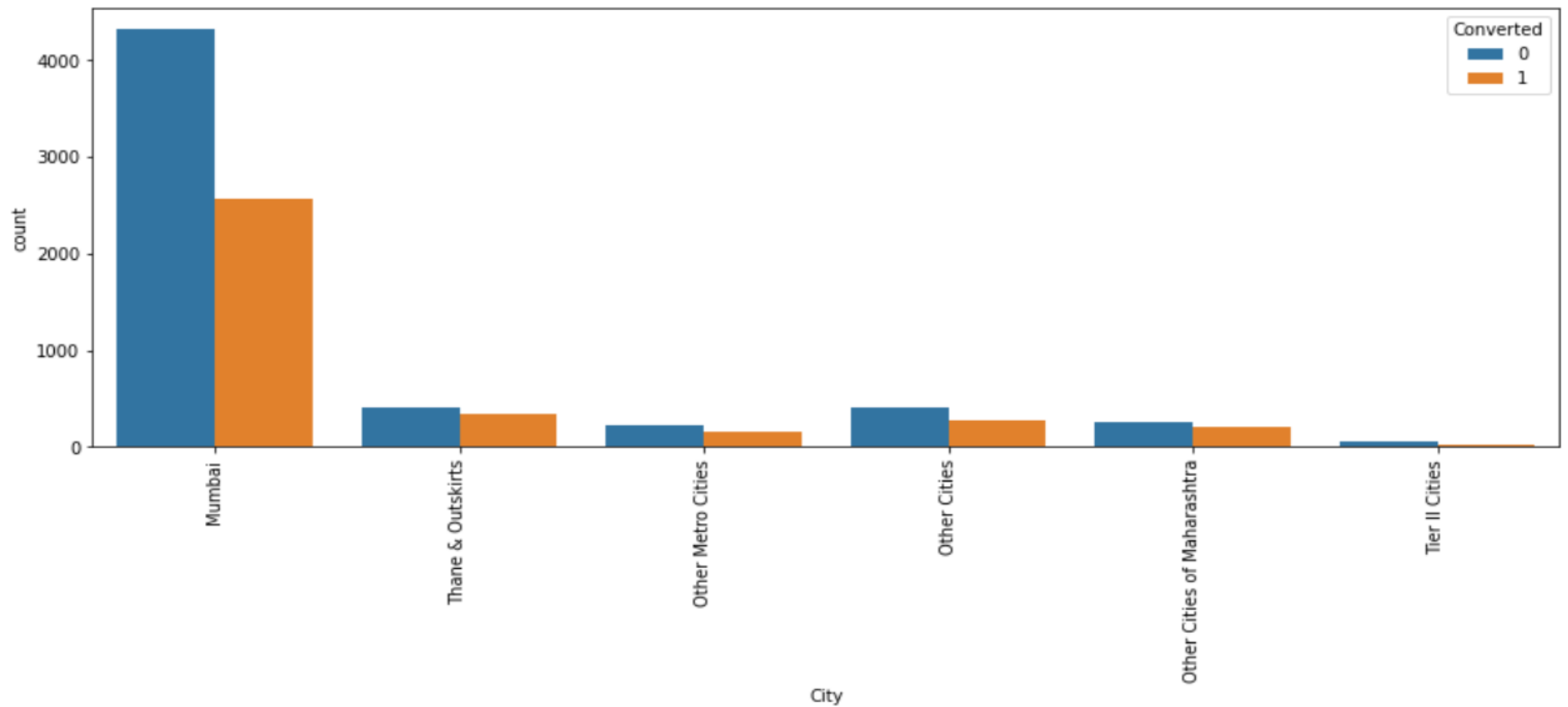
- ❑ **Analyzing Patterns:** Using Exploratory Data Analysis, we have analyzed the patterns present in the Dataset which will provide us intuition that the which features will help in driving the lead conversion.
- ❑ **Driving Factors:** Looking at the below data we get an intuition that how the variables are distributed.

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

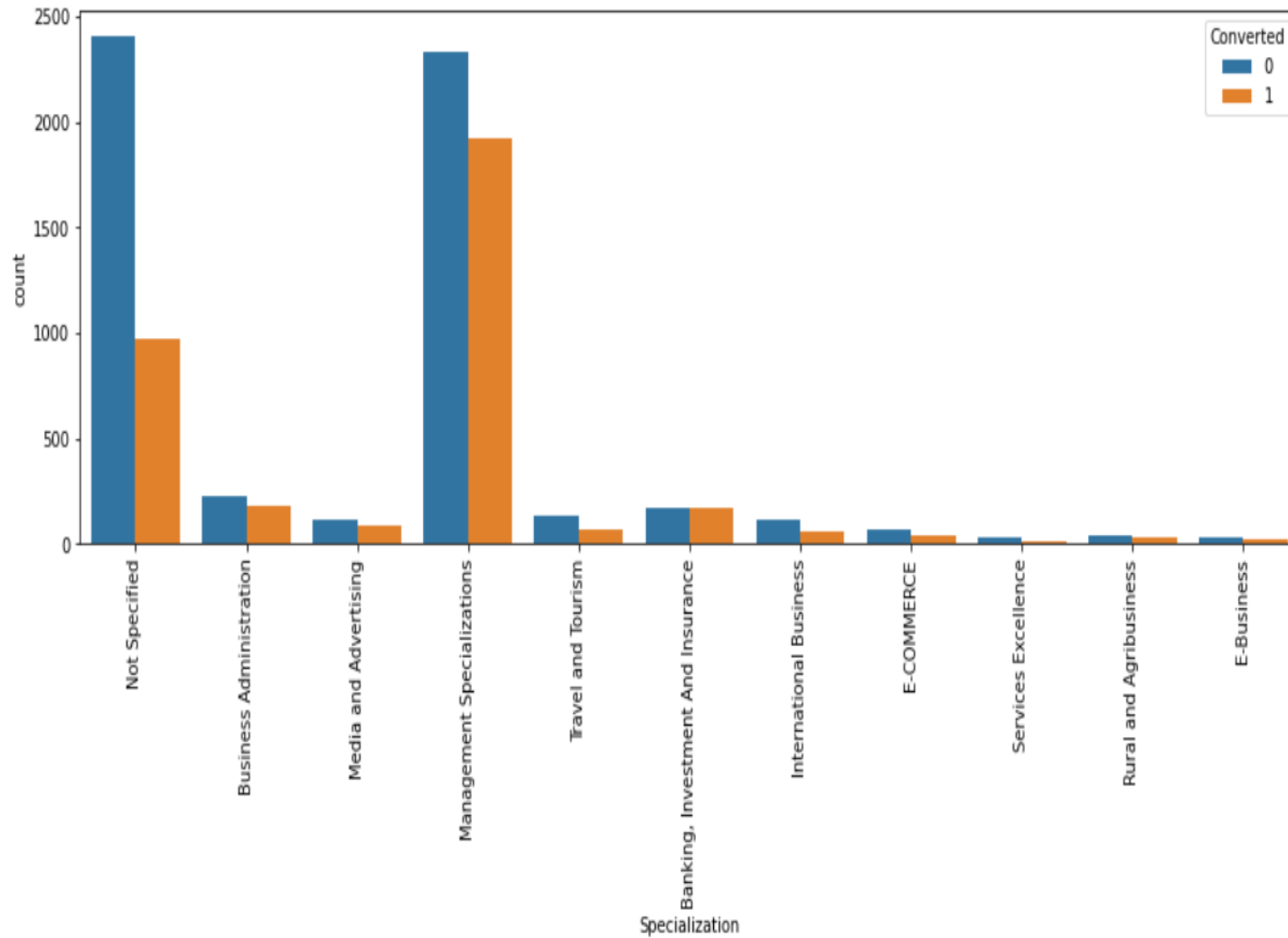
- 
- ❑ **Correlations:** Identifying correlations amongst variables to identify the variability in data and identify most important features that can help in driving the conversion of leads.
 - ❑ **Recommendations:** Focus on features that can expedite the conversion of leads.

Data Insights:

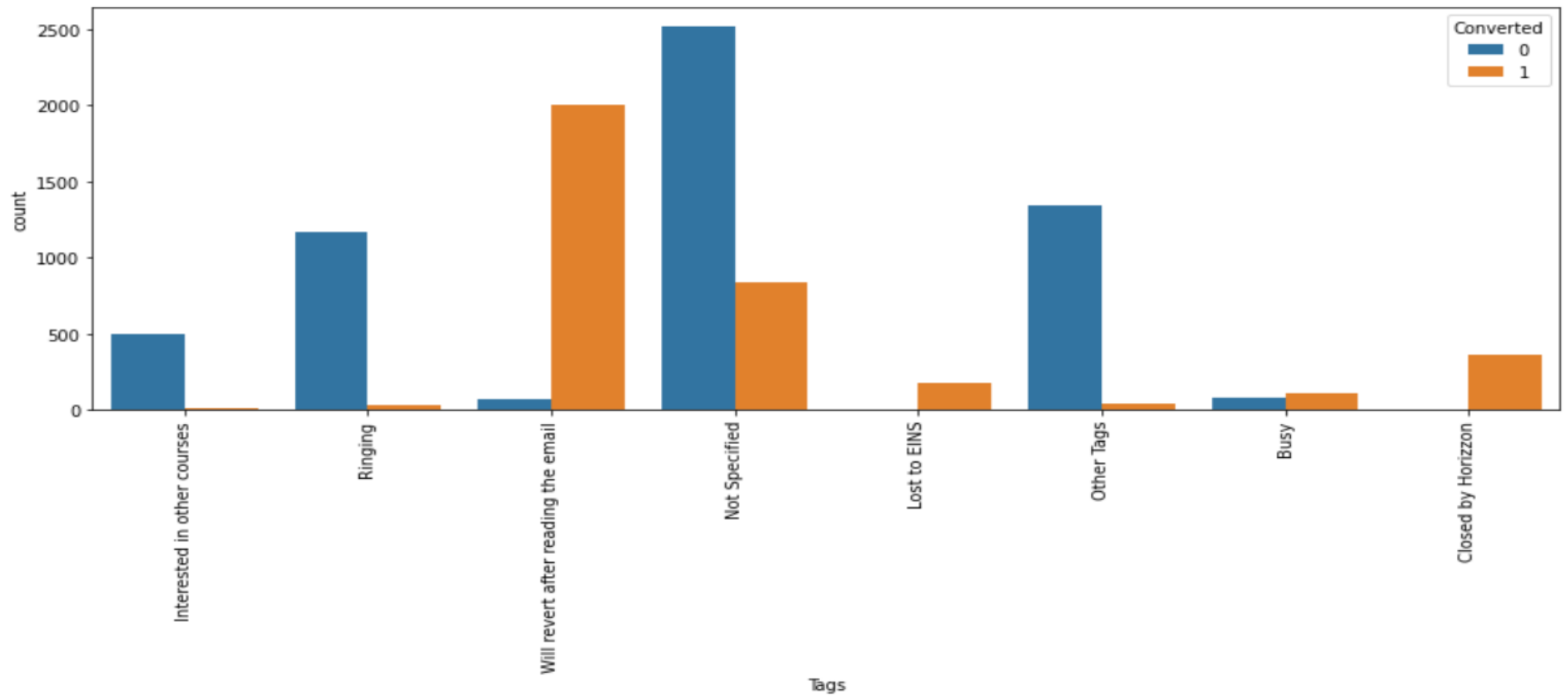
Categorical Columns w.r.t Converted
Columns



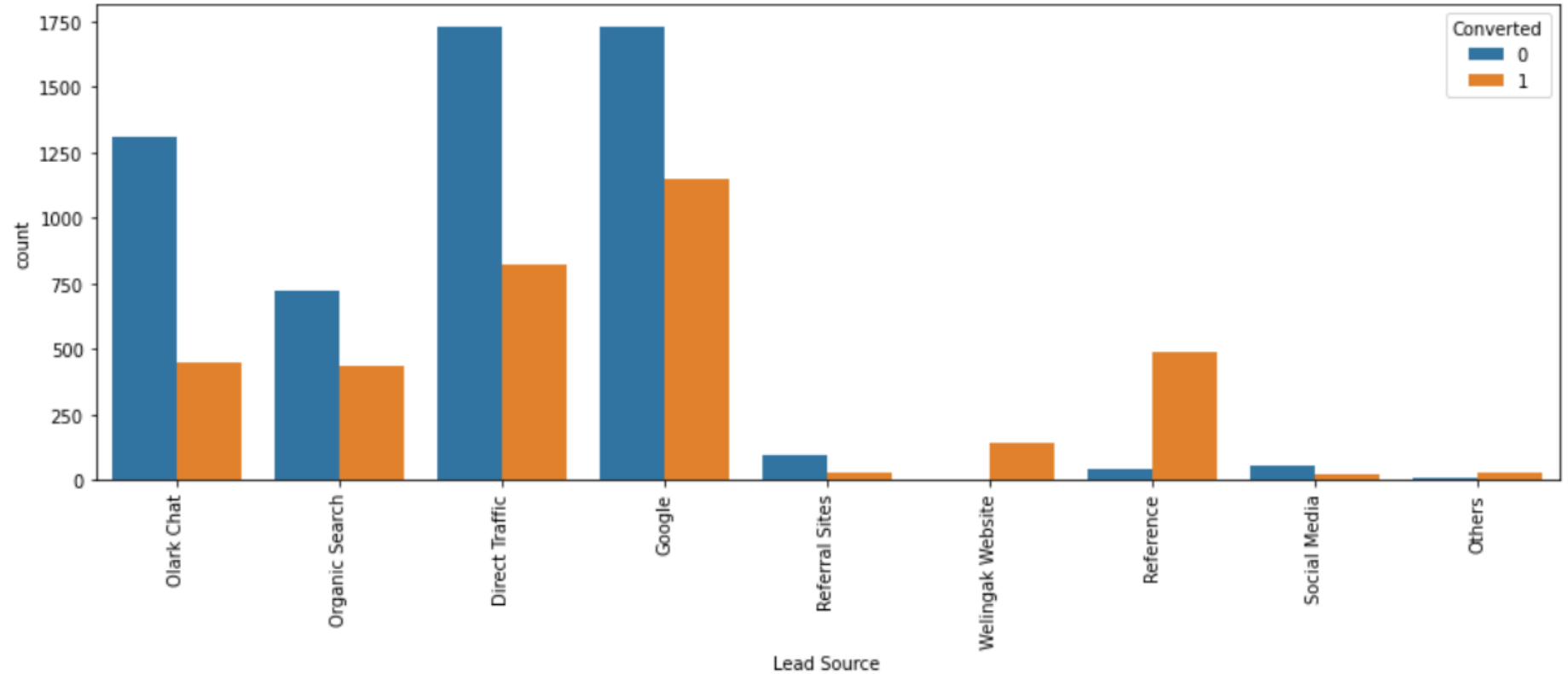
► We can see that majority of our leads are from Mumbai and least number of leads are from Tier 2 cities.



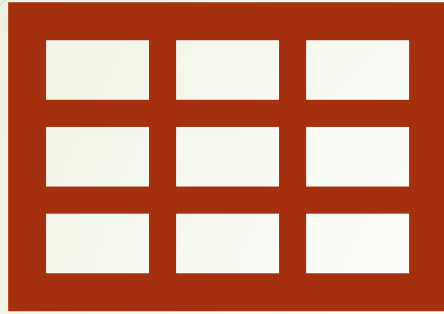
- Majority of leads have Not Specified their specialization and from the pool of leads who have Not Specified their specialization the conversion rate is low as well.
- High number of leads are from Management Specialization with high conversion rate.



- Majority of leads converted into final customers are from the tag "Will revert after reading the email"

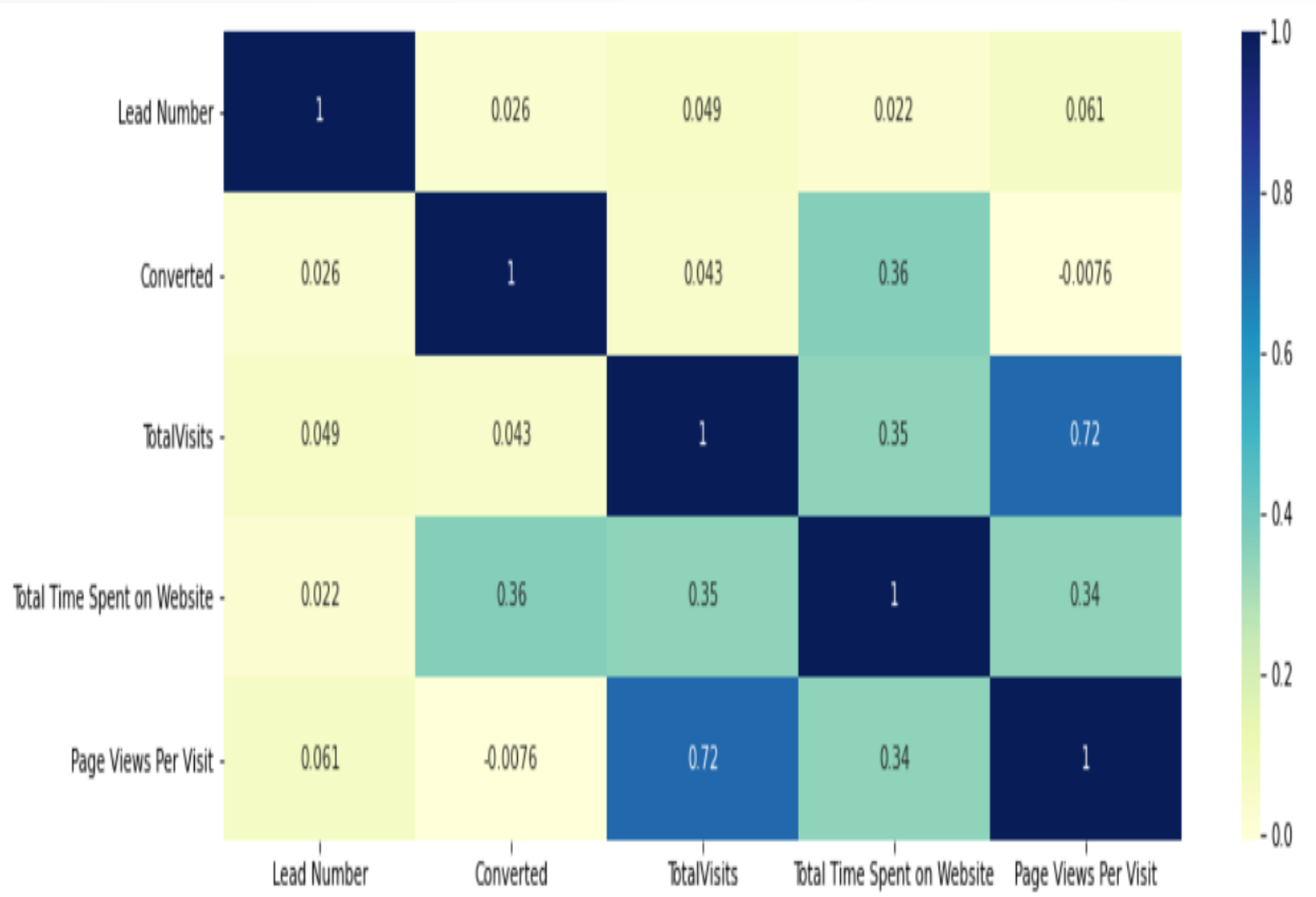


- As we can see majority of leads were generated through Direct Traffic and Google.
- Whereas maximum number of leads converted into final customer were from Google.

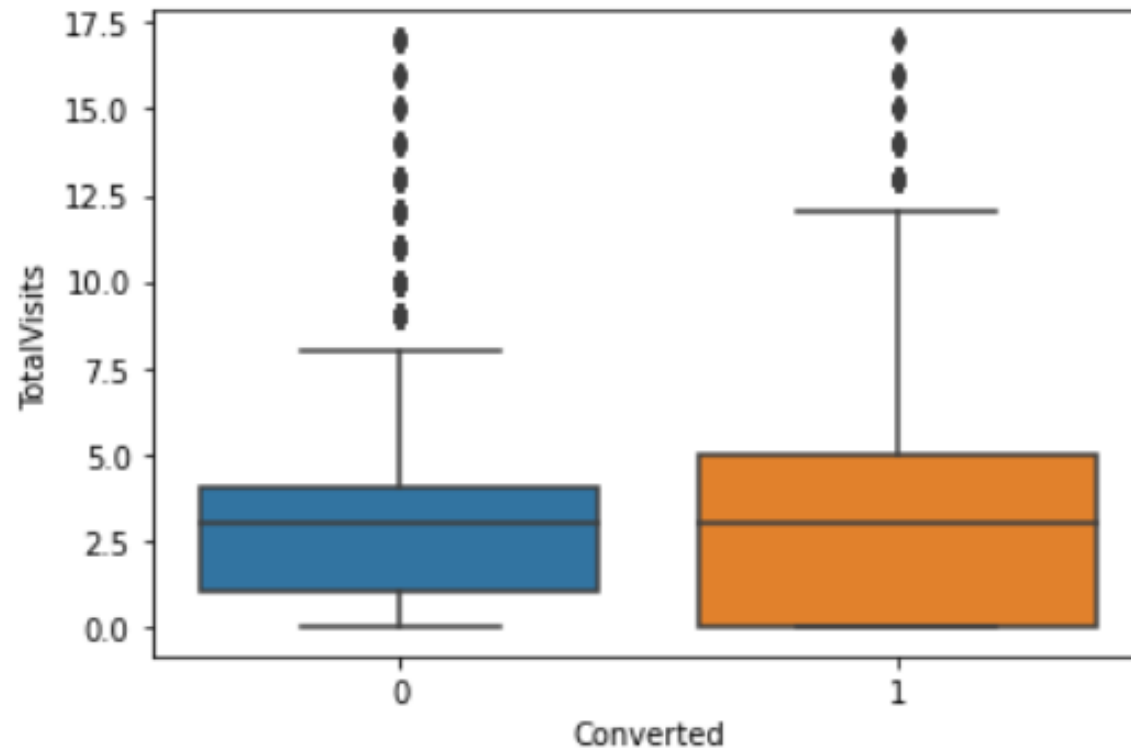


Data Insights:

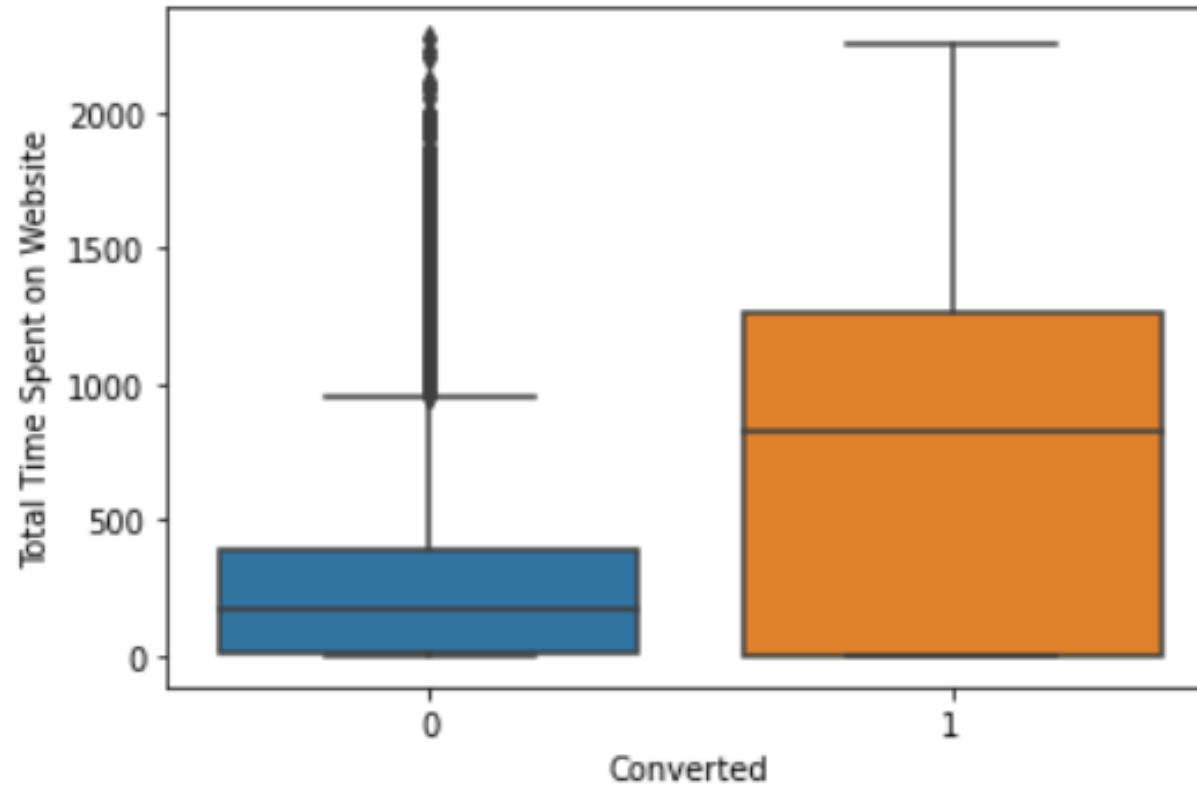
Numerical Columns w.r.t Converted Columns



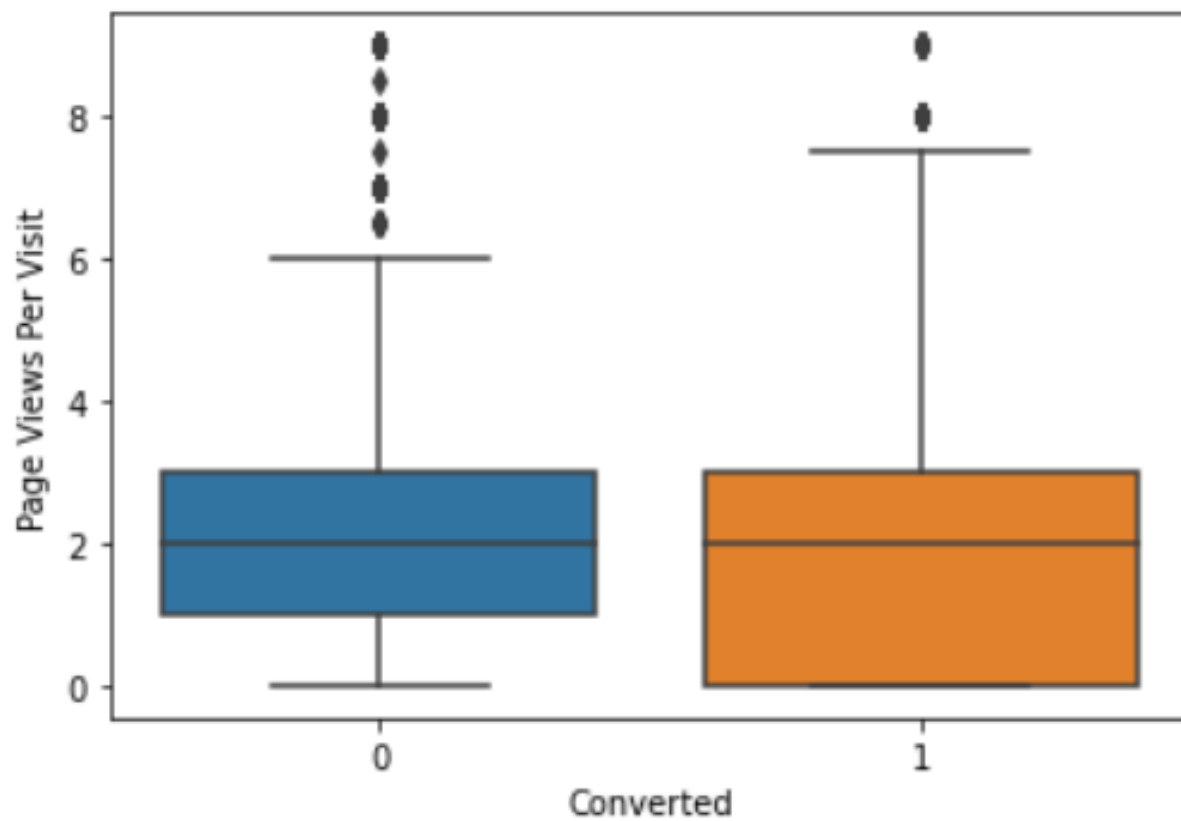
➡ There's some correlation between the variable "Total Visits" and "Page Views Per Visit".



➤ From the above boxplot we can see that median for both converted and not converted leads is almost the same.



➤ From the above boxplot we can clearly infer that converted leads has spend more time on website than non-converted leads.



► Median for converted leads and non-converted leads is almost the same.

Factors Responsible in Driving Leads

The background features several thin, light green lines. One line is a long, shallow curve starting from the top center and extending towards the right edge. Another line is a steeper curve starting from the top right and extending towards the bottom right. A third line is a straight vertical line located to the right of the main text.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6363
Model:	GLM	Df Residuals:	6347
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3000.2
Date:	Tue, 10 Aug 2021	Deviance:	6000.5
Time:	13:23:13	Pearson chi2:	7.59e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-0.4776	0.078	-6.123	0.000	-0.630	-0.325
Total Time Spent on Website	1.0858	0.037	28.995	0.000	1.012	1.159
Lead Origin_Lead Add Form	2.7911	0.197	14.172	0.000	2.405	3.177
Lead Origin_Lead Import	-1.1648	0.526	-2.213	0.027	-2.196	-0.133
What is your current occupation_Working Professional	2.7139	0.179	15.170	0.000	2.363	3.064
Specialization_Banking, Investment And Insurance	0.8176	0.174	4.686	0.000	0.476	1.160
Specialization_Business Administration	0.3784	0.165	2.287	0.022	0.054	0.703
Specialization_E-COMMERCE	0.9070	0.289	3.135	0.002	0.340	1.474
Specialization_International Business	0.5717	0.231	2.478	0.013	0.120	1.024
Specialization_Management Specializations	0.5712	0.081	7.010	0.000	0.411	0.731
Specialization_Media and Advertising	0.6650	0.219	3.039	0.002	0.236	1.094
Lead Source_Direct Traffic	-1.3846	0.120	-11.518	0.000	-1.620	-1.149
Lead Source_Google	-0.8759	0.108	-8.088	0.000	-1.088	-0.664
Lead Source_Organic Search	-1.0004	0.128	-7.844	0.000	-1.250	-0.750
Lead Source_Referral Sites	-1.0744	0.319	-3.370	0.001	-1.699	-0.449
Lead Source_Welingak Website	3.0870	1.022	3.020	0.003	1.083	5.091

Important
factors which
influence the
conversion of
leads -

Total Time Spent on Website
Lead Origin_Lead Add Form
Lead Origin_Lead Import
What is your current occupation_Student
What is your current occupation_Working Professional
Specialization_Banking, Investment And Insurance
Specialization_Business Administration
Specialization_E-COMMERCE
Specialization_International Business
Specialization_Management Specializations
Specialization_Media and Advertising
Lead Source_Direct Traffic
Lead Source_Google
Lead Source_Organic Search
Lead Source_Referral Sites
Lead Source_Welingak Website

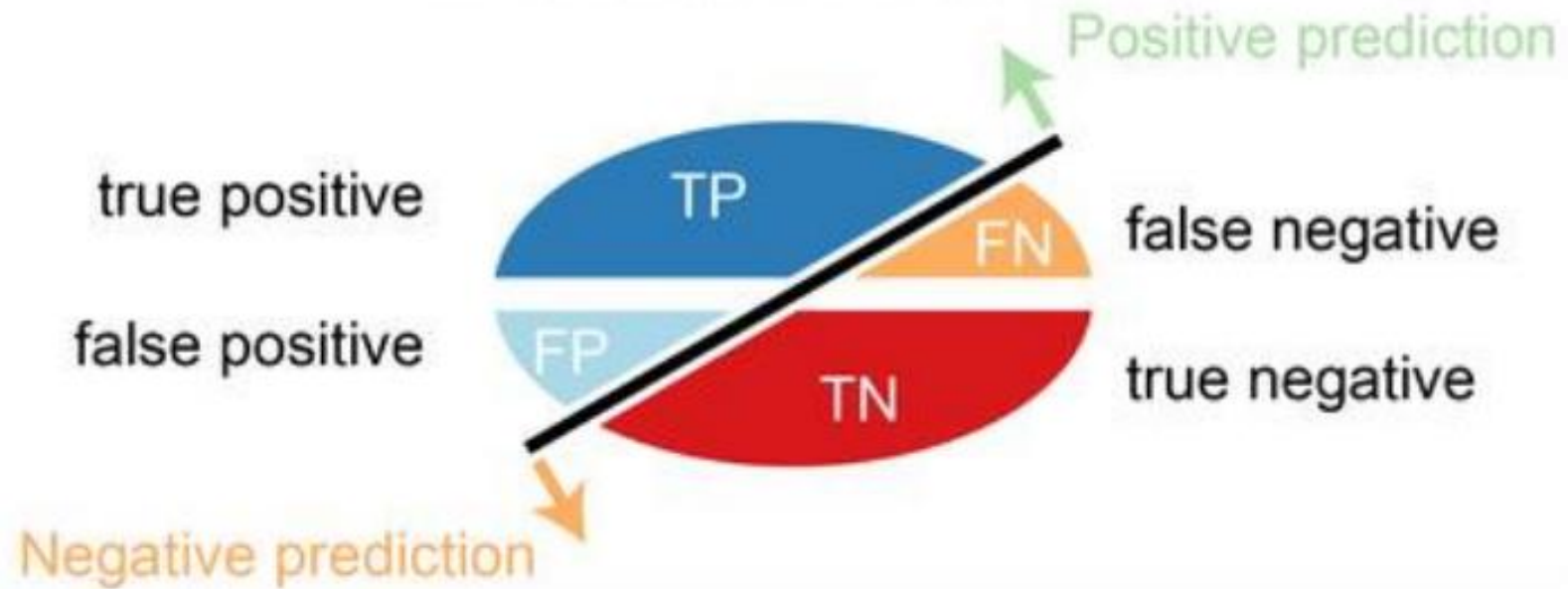
Terminologies Required

Before proceeding ahead, we need to understand few terminologies:

- **Conversion of categorical columns to numerical:** This step is done as our algorithm runs only on numerical data.
- **Feature Scaling:** This is done to bring our data into same scale.
- **Data Splitting:** We have split the data into 70:30 and named it as train data and test data. We run model on train data and validate our model on test data.
- **Confusion Matrix:**

	Predicted No	Predicted Yes
Actual No	True Negative	False Negative
Actual Yes	False Positive	True Positive

Four outcomes of a classifier





Whereas,

True positive (TP): correct positive prediction

False positive (FP): incorrect positive prediction

True negative (TN): correct negative prediction

False negative (FN): incorrect negative prediction


Above Metrics is Known as Confusion Metrics, using above metrics we derived following things:

1. Accuracy = (True Negative + True Positive)/Total

This metrics provides the accuracy of the model, where total is TP + FN + FP +FN.

2. Sensitivity = True Positive / (True Positive + False Positive)

Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0.



3. **Specificity** = True Negative/ (True Negative + False Negative)

Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0.

4. **Precision** = True Positive/ (True Positives +False Positives)

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

5. **Recall** = True Positives/(True Positives +False Negatives)

The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the model that are positive (meaning they are correct), and false negatives are data points the model identifies as negative that are positive (incorrect).

Model Metrics

Running model on features selected we get following metrics:

1. Train Data:

- Confusion Metrics

	Not Converted Leads	Converted Leads
Not Converted Leads	2993	942
Converted Leads	473	1955

- Accuracy: 77.76%
- Sensitivity: 80.51%
- Specificity: 76.06%
- Precision: 67%
- Recall: 80%

2. Test Data:

- Confusion Metrics

	Not Converted Leads	Converted Leads
Not Converted Leads	1270	390
Converted Leads	219	848

- Accuracy: 77.66%
- Sensitivity: 79.47%
- Specificity: 76.50%
- Precision: 68.49%
- Recall: 79.47%

The Model seems to predict the Conversion Rate very well. We should be able to help the education company select the most promising Leads or the Hot Leads.

Conclusion

- Company should focus on following features to increase the leads-
- **Lead Source_Welingak Website**
- **Lead Origin_Lead Add Form**
- **hat is your current occupation_Working Professional**
- Company should also focus on Lead Score (which are the probabilities obtained via algorithm) which are greater than 80% to expedite the conversion rate.