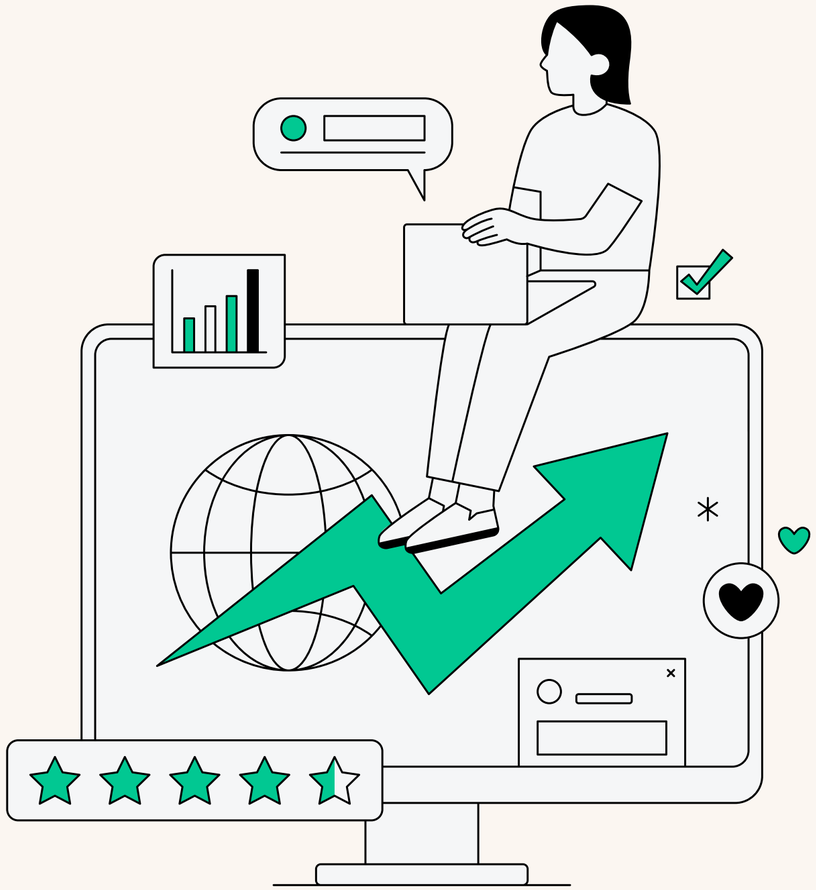
Presented by

Mark
Divya
Eshwar

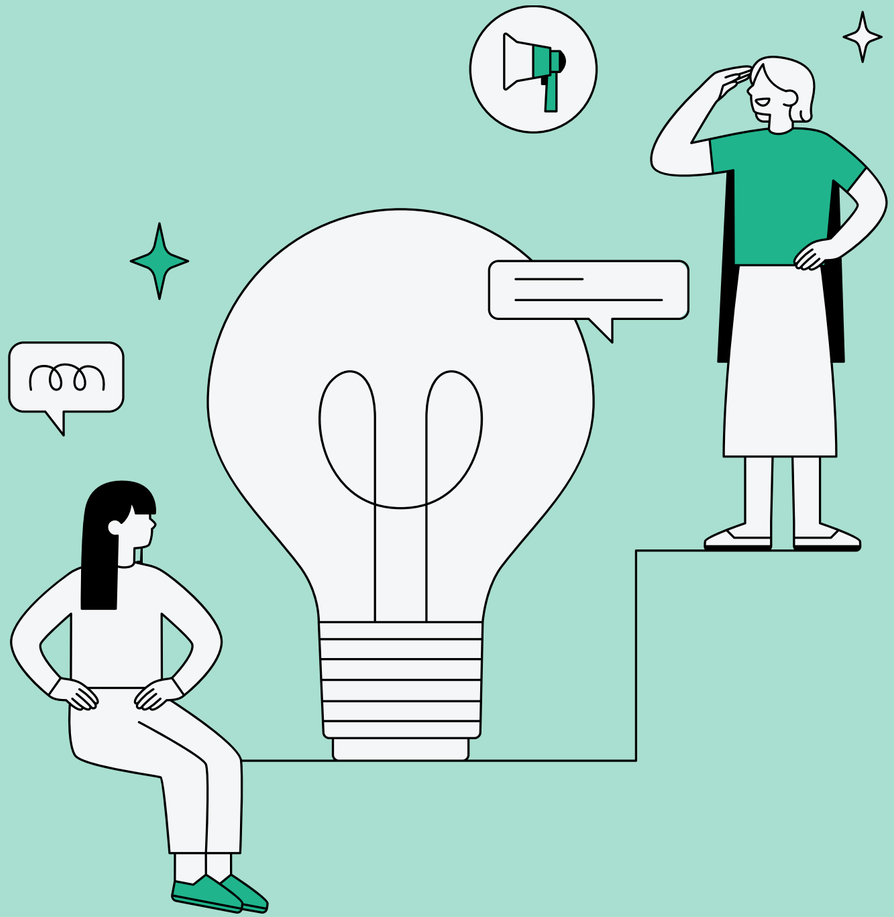
Lead Score Case Study

Building a Logistic Regression Model for X Education to Maximize Conversion Rates.



Problem Statement

X Education, an online course provider for industry professionals, faces a low lead conversion rate of around 30% despite acquiring a large number of leads daily through marketing efforts. To improve efficiency, the company wants to identify high-potential leads ("Hot Leads"), allowing the sales team to focus on the most promising prospects instead of reaching out to every lead. The objective is to develop a predictive model that assigns a lead score to each lead, ensuring that leads with a higher score have a higher chance of conversion. The ultimate goal is to optimize the sales process and achieve a target conversion rate of 80%.



Methodology used in the analysis

1. DATA CLEANING AND IMPUTING MISSING VALUES
2. EXPLORATORY DATA ANALYSIS
3. LOGISTIC REGRESSION MODEL BUILDING
4. MODEL EVALUATION
5. CONCLUSION AND RECOMMENDATION

Exploratory Data Analysis (EDA)

- Identified and replaced missing values with most dominant values
- Created graphs for visualizations
- Dropped columns where only one value was dominant to avoid bias
- Analyzed Numerical Values based on Correlation using heatmaps
- Used Boxplots to visualize and remove outliers

Model Building

- Built a Logistic Regression model, Data Split into - 70% train, 30% test
- RFE for Feature Selection and model building with selected features
- Dropped high P-Values
- VIF was used to find correlation between the variables
- Optimized model performance using selected features

Model Evaluation

Training Data:

- Accuracy: 92.11% | Sensitivity: 89.85% | Specificity: 93.50%
- ROC Curve AUC = 0.97
- Optimal cutoff at 0.3 for best accuracy

Testing Data:

- Accuracy: 92.51% | Sensitivity: 91.39% | Specificity: 93.20%

Conclusion

- Compared Model Performances
- Evaluated Business Impact
- Provided Recommendations (Identified hot leads)

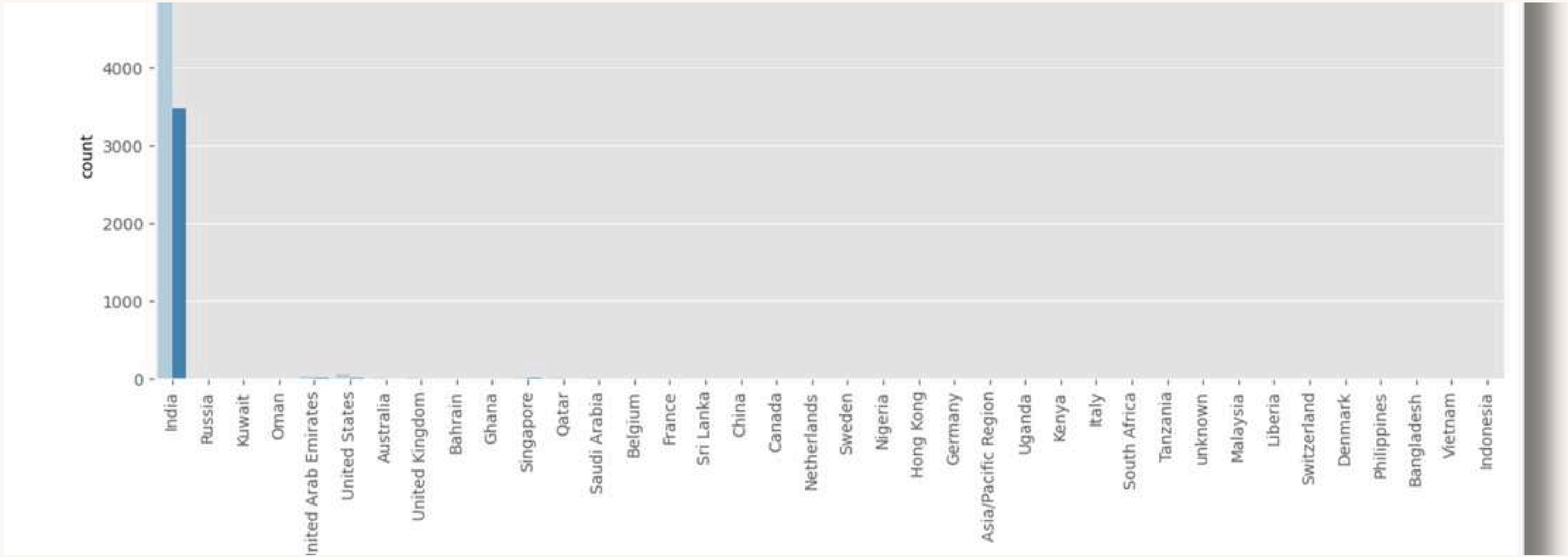
Data Cleaning and Preparation

- Read data and check for duplicates
- Dropping columns not useful for analysis
- Handled missing values (convert "Select" to NaN, dropping >40% null columns)

EXPLORATORY DATA ANALYSIS

We carried out exploratory data analysis on the categorical variables. Displaying most important factors leading to conversions

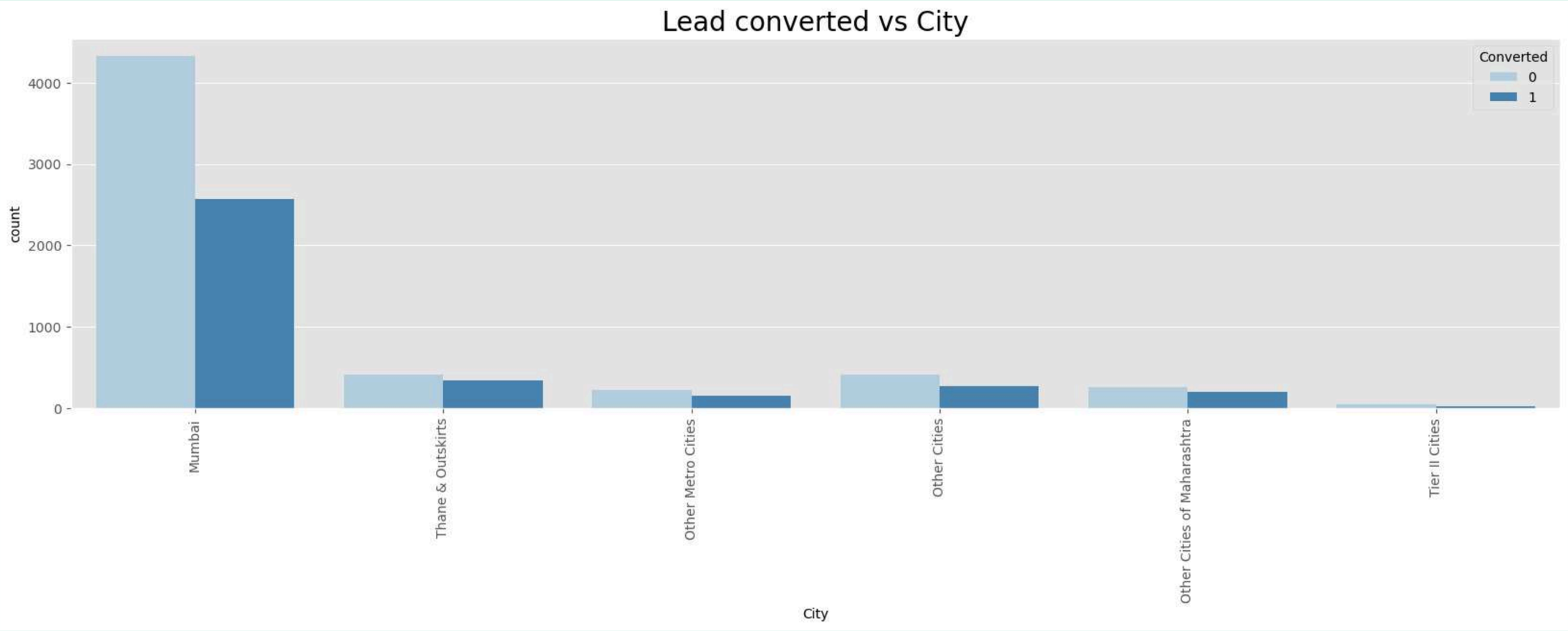
Number of Leads Vs Country



Inferences

- We observed that India was the country from where majority of the leads were generated
- We dropped the country column as it could lead to bias during analysis affecting efficacy of our solution.

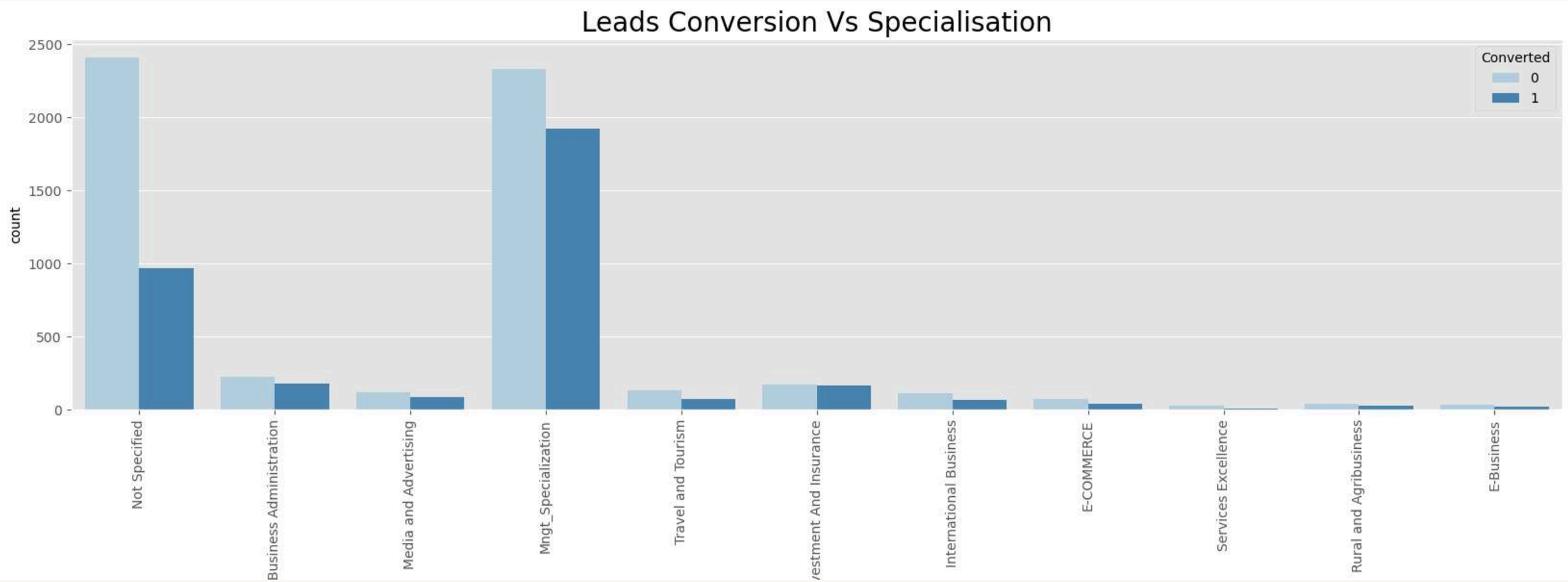
Lead Converted Vs City



Inferences

- Majority of the leads were generated from Mumbai
- It also featured the highest conversion rate

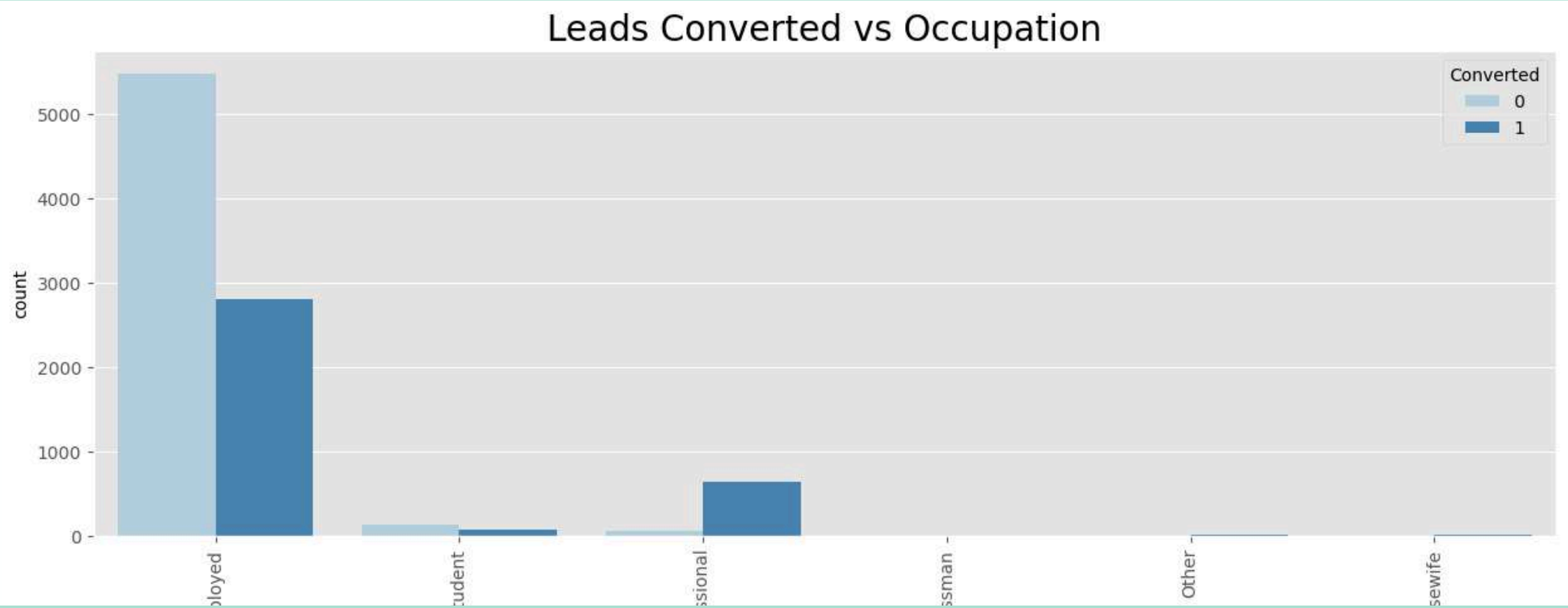
Lead Converted Vs Specialization



Inferences

- Some "Management" related specializations like Finance, HR, and Operations, generate more leads and have higher conversion rates
- As there were different types of management specializations, we combined them under heading Mngt_Specialization for simplicity.

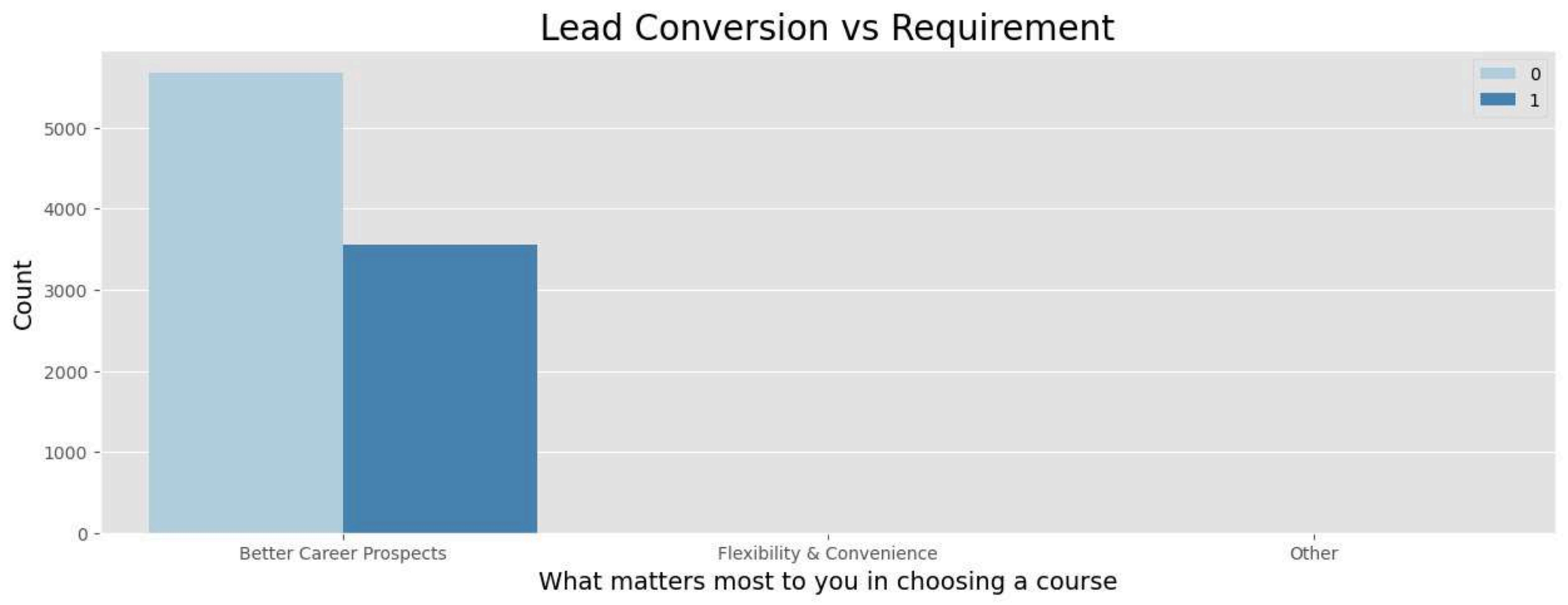
Lead Converted Vs Occupation



Inferences

- Majority of the leads come from people who are unemployed.
- Working professionals are more likely to become customers
- Students, Businessmen, Others and Housewives do not show much interest and are less likely to become customers.

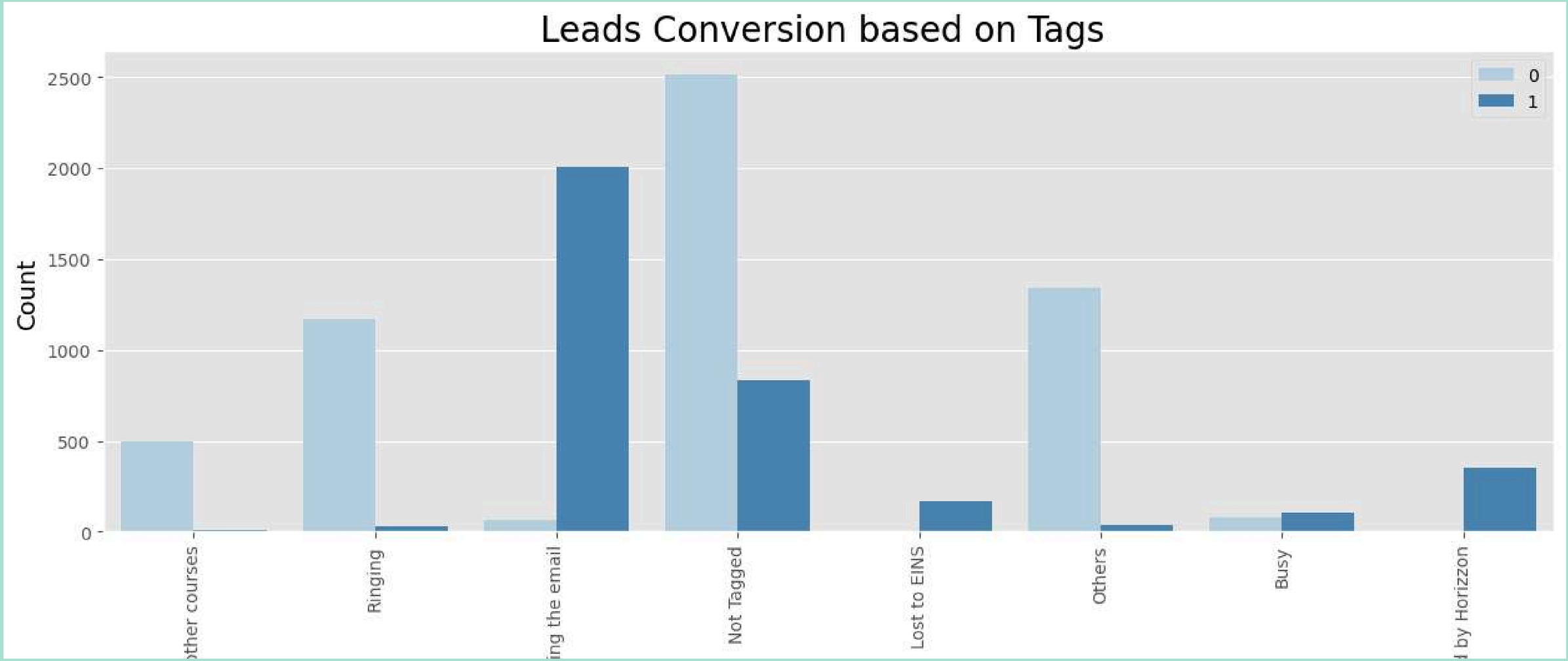
Lead Converted Vs Requirement



Inferences

- Most of the leads selected “Better Career Prospects”as a requirement of pursuing a course.
- This variable heavily influenced the column and hence we removed it to prevent biased analysis.

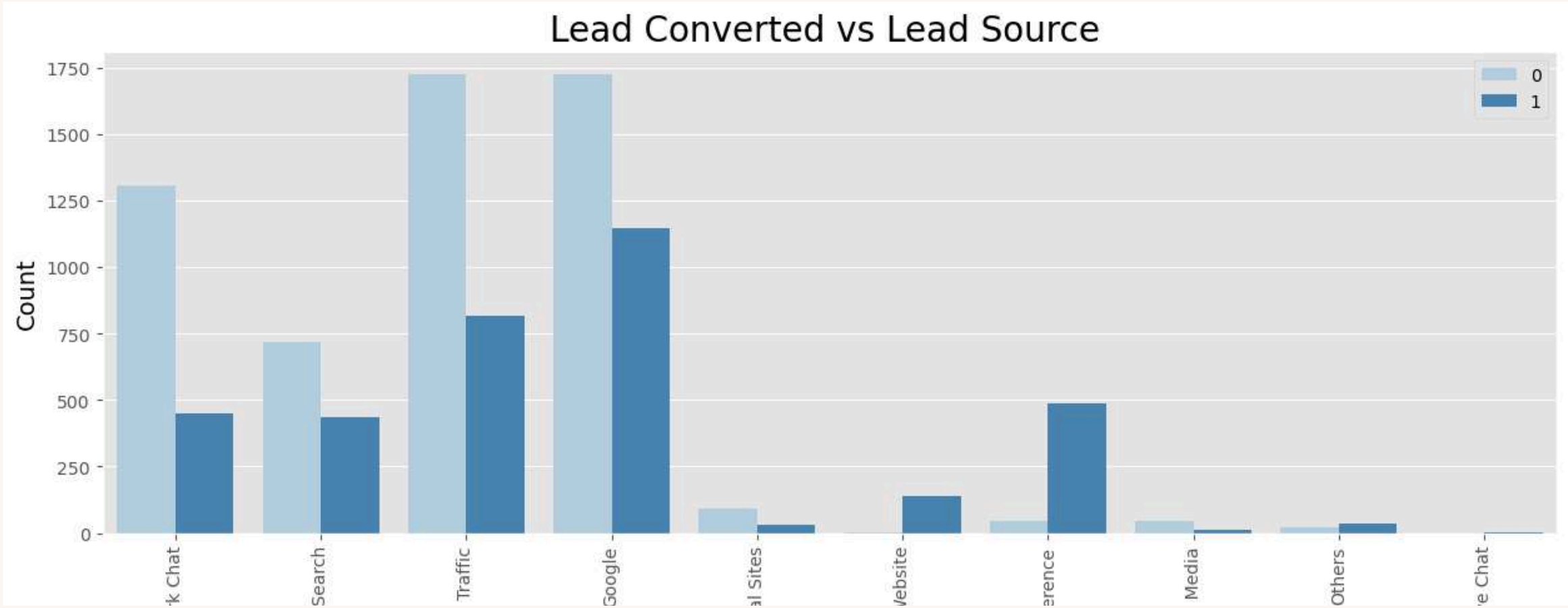
Lead Converted Vs Tags



Inferences

- Initially we found that many of the leads were not tagged. We replaced “NaN” present earlier in the data shared with 'Not tagged ’. We later clubbed them with “others” as their counts were low in comparison to other lead sources
- Leads who read the email had the highest conversion rate which could be improved further by sharing more emails with leads
- Lead conversion was also good from Horizon and more leads could be generated from here.

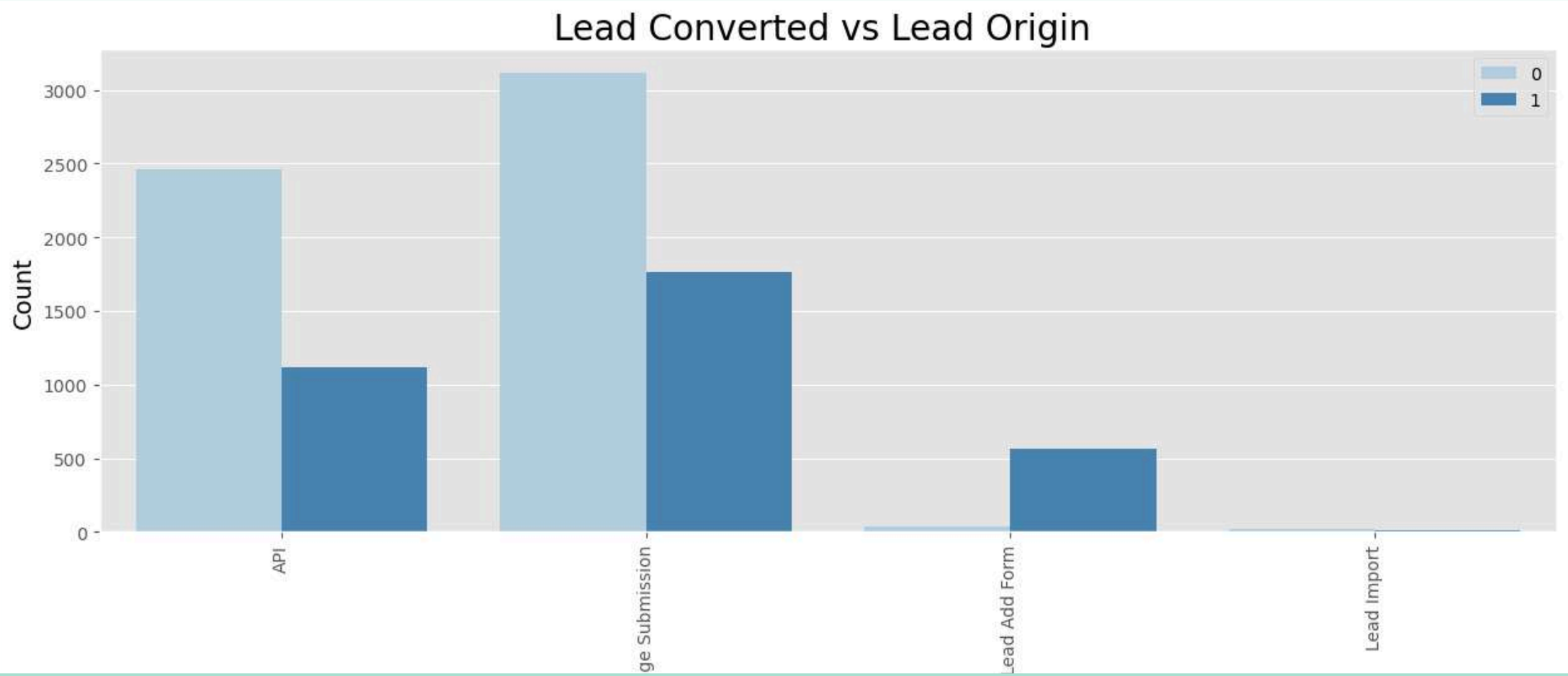
Lead Converted Vs Lead Source



Inferences

- Most of the lead conversion is taking place through google and direct traffic
- It would be beneficial to increase leads from references and the Welingak website.

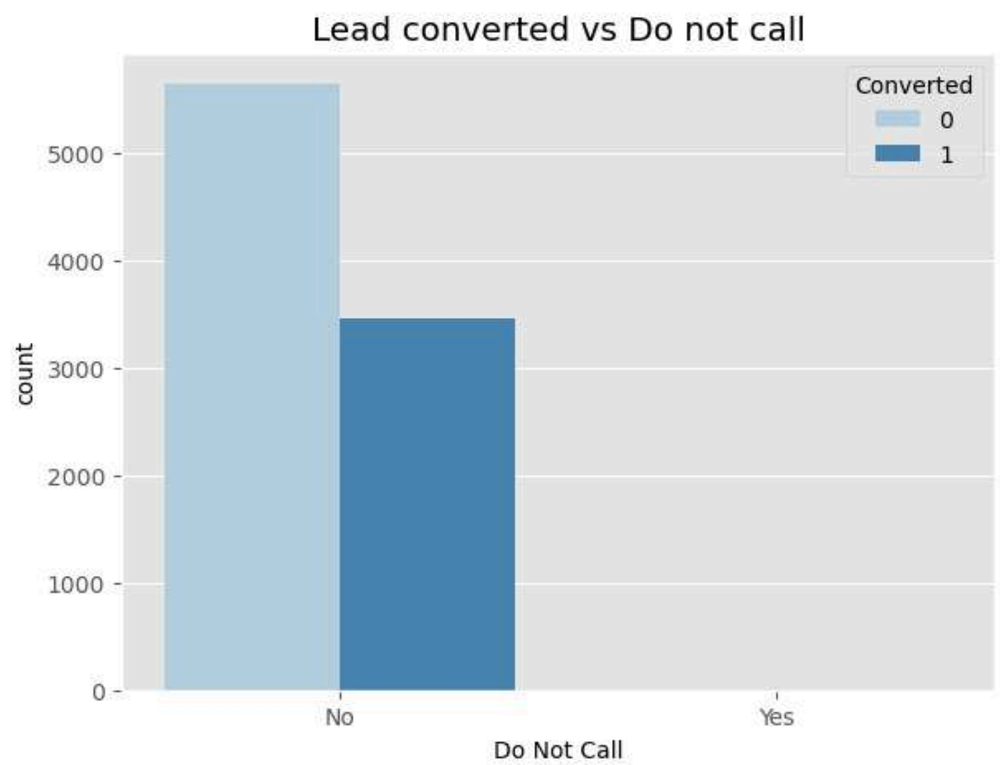
Lead Converted Vs Lead Origin

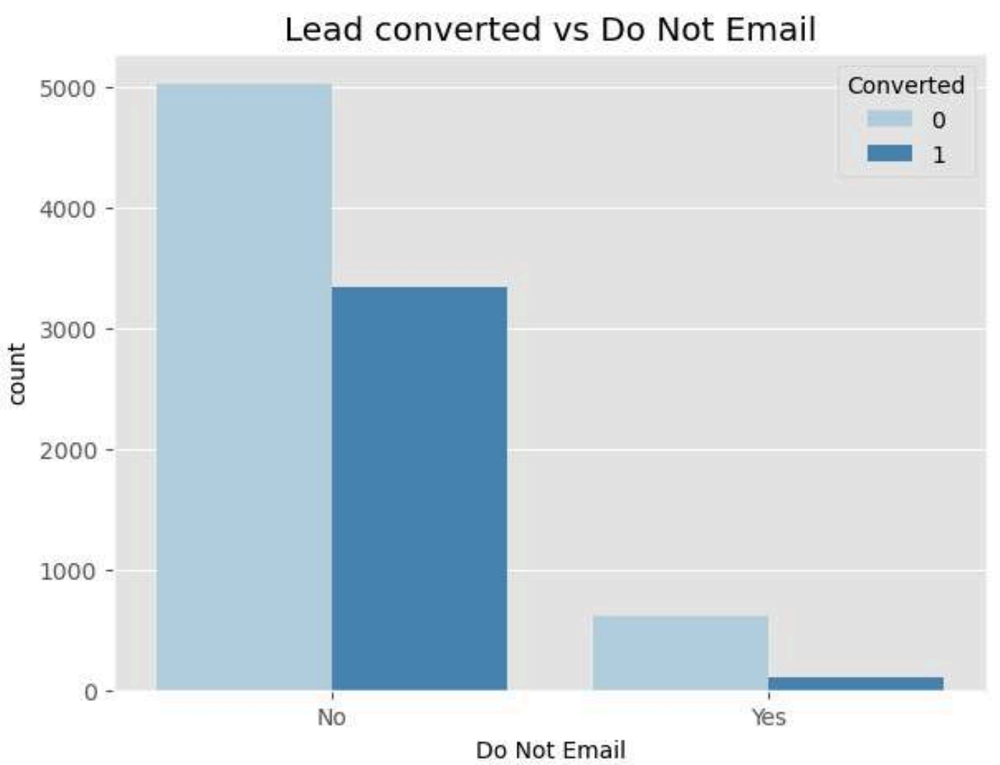


Inferences

- Lead conversion form ‘Lead Add Form’ has the highest conversion rate and lead generation should be increased from the same
- Large number of leads are coming from "Landing page submission" followed by "API", however the conversion rate could be improved.

Lead Converted Vs Do Not Call/Email

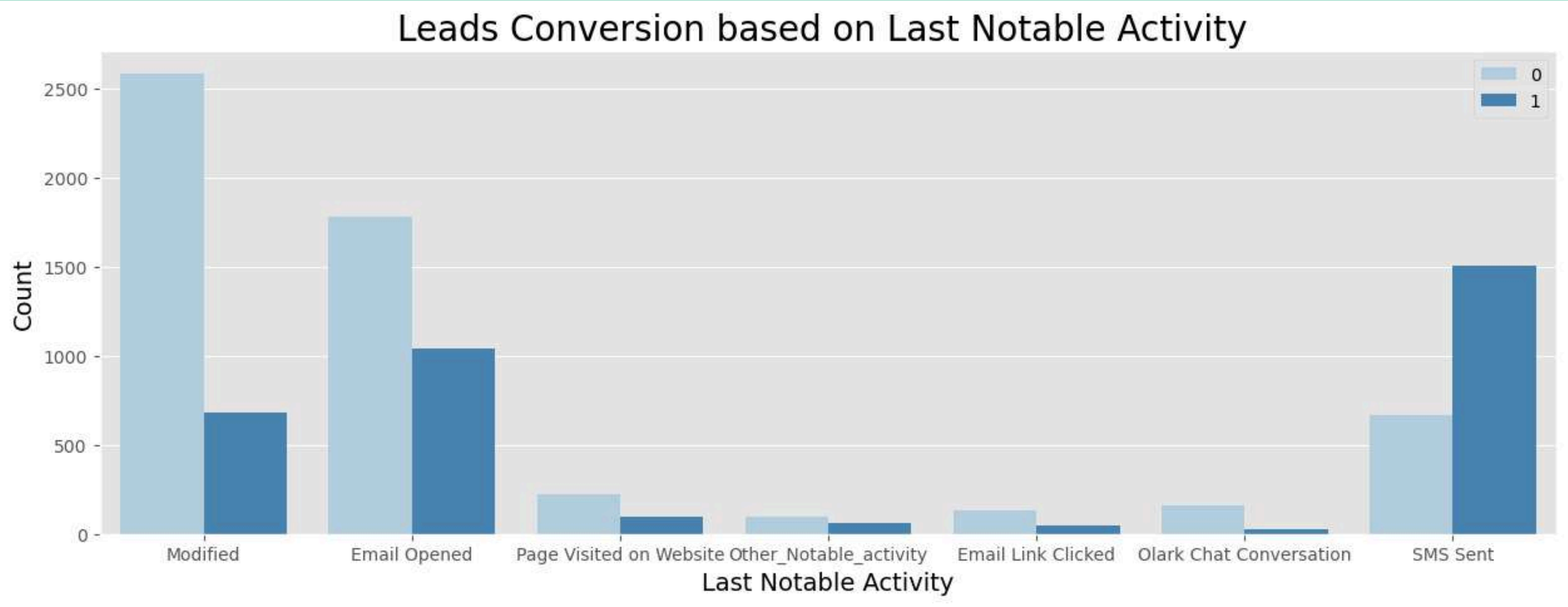




Inferences

- As 'Do not call’ had 'No' as most occurring category which could create bias, we dropped it from the data
- There were a good number of conversions from leads who opted for Do not call/Email

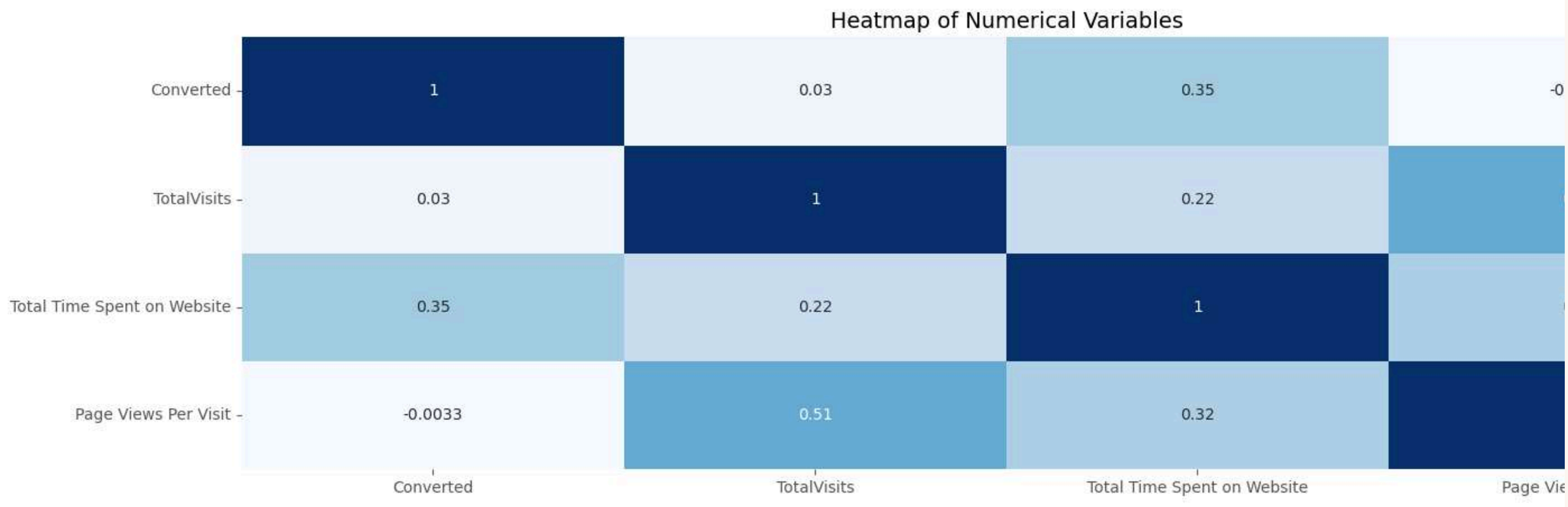
Lead Converted Vs Last Notable Activity



Inferences

- Lead conversion via SMS sent has the highest conversion rate and lead generation should be increased from the same
- Modified has the lowest conversion rate even though leads generated from here are the highest
- Leads who opened the email also have a good conversion rate which could be improved further.

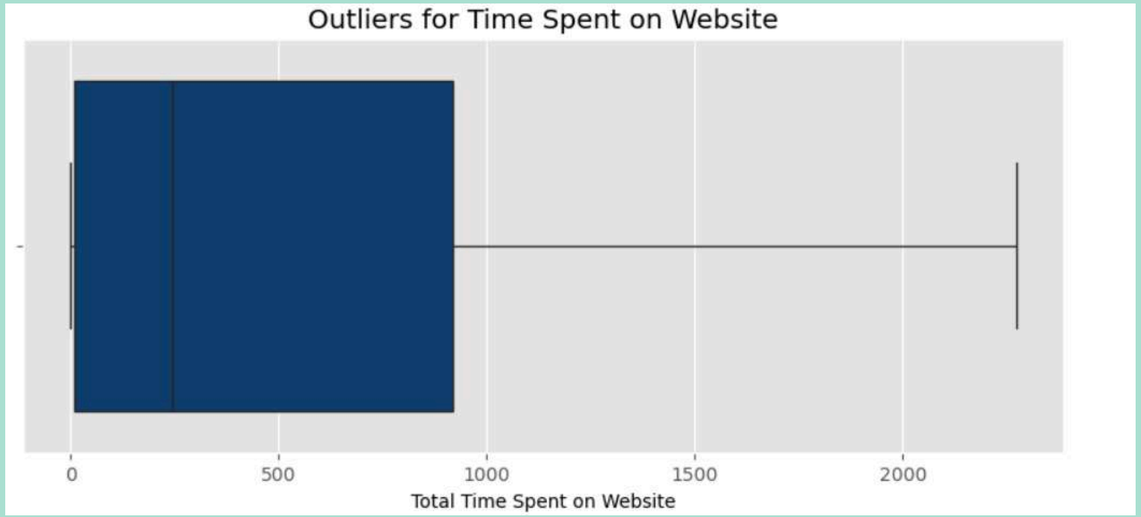
Analysis of Numerical Values based on Correlation

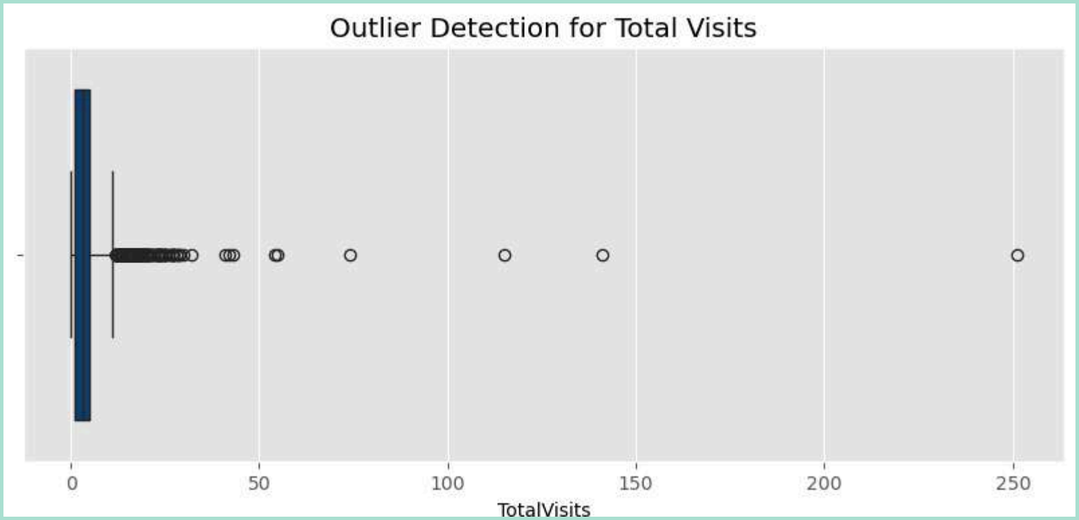


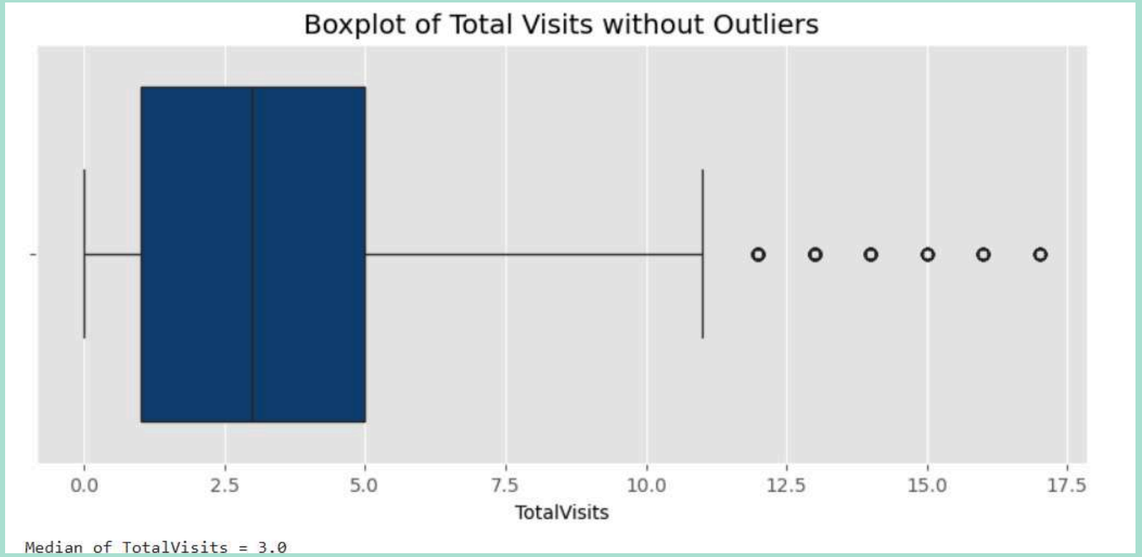
Inferences

- There is a higher correlation between total time spent on website and converted leads
- We further analyzed the total time spent on website and Total Visits to check for ‘Outliers’

Analysis of Numerical Values based on Correlation



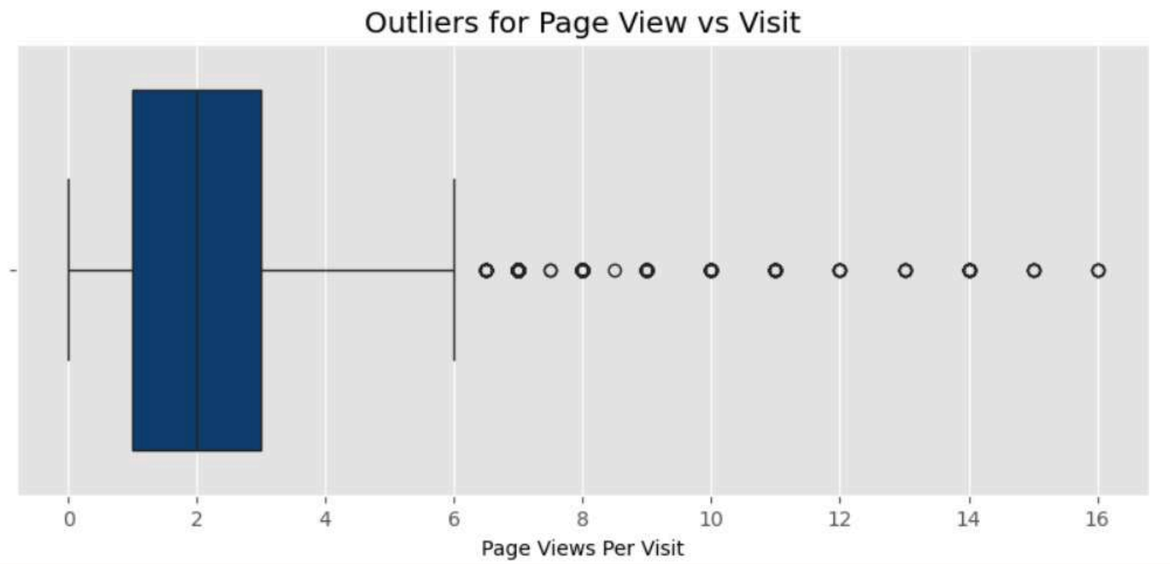


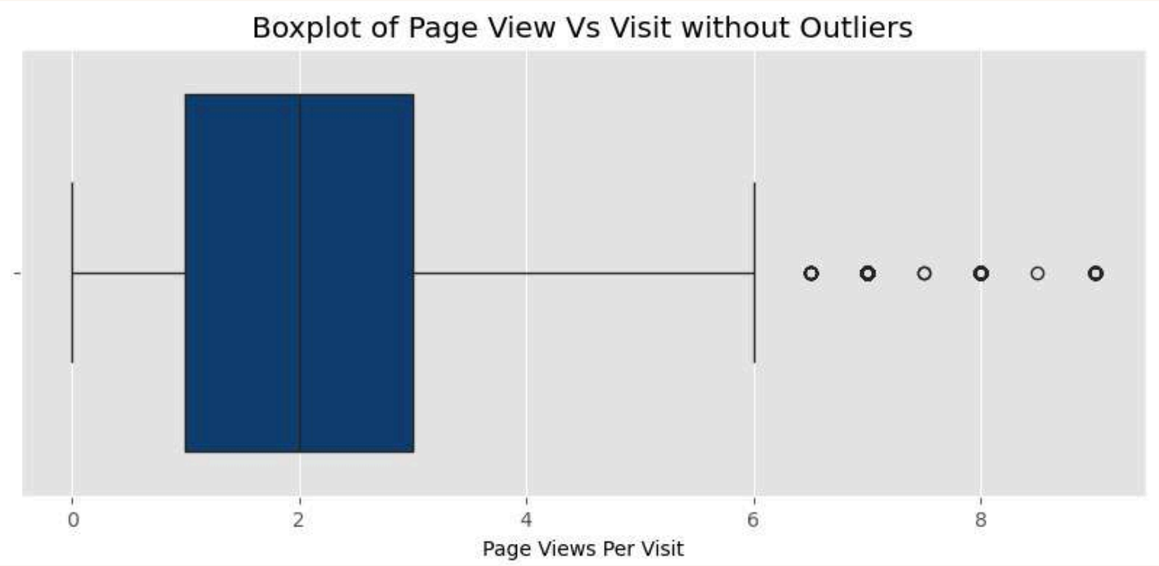


Inferences

- There are no outliers for 'Total time spent on website’
- We removed outliers from ‘Total Visits’ and found the median at 3.0.

Analysis of Numerical Values based on Correlation





Inferences

- There were outliers in 'page view per visit' hence we limited the capping to the top 1% and removed them.
- Post this step we found that there were no 'null' values in the data and we could proceed with Model Building.

Model Building

Generalized Linear Model Regression Results

:	Converted	No. Observations:	6267
	GLM	Df Residuals:	6253
	Binomial	Df Model:	13
:	Logit	Scale:	1.0000
	IRLS	Log-Likelihood:	-1319.2
	Tue, 18 Feb 2025	Deviance:	2638.3
	14:36:10	Pearson chi2:	1.02e+04
s:	8	Pseudo R-squ. (CS):	0.5965
pe:	nonrobust		

	coef	std err	z	P>|z|	[0.0
	-1.4581	0.085	-17.246	0.000	-1.6
ent on Website	1.2797	0.093	13.745	0.000	1.0
	1.1742	0.455	2.581	0.010	0.2
irect Traffic	-0.6710	0.124	-5.431	0.000	-0.9
elingak Website	4.0102	1.119	3.585	0.000	1.8
_SMS Sent	2.0471	0.110	18.559	0.000	1.8
Activity_Modified	-1.6963	0.122	-13.899	0.000	-1.9
Activity_Olark Chat Conversation	-1.8724	0.481	-3.895	0.000	-2.8
y Horizzon	7.2690	1.018	7.143	0.000	5.2
ed in other courses	-2.0074	0.398	-5.045	0.000	-2.7
EINS	5.9177	0.607	9.746	0.000	4.7
	-2.3531	0.205	-11.490	0.000	-2.7
	-3.4529	0.237	-14.574	0.000	-3.9
ert after reading the email	4.6300	0.186	24.946	0.000	4.2

Training Data Stats before optimal cut off

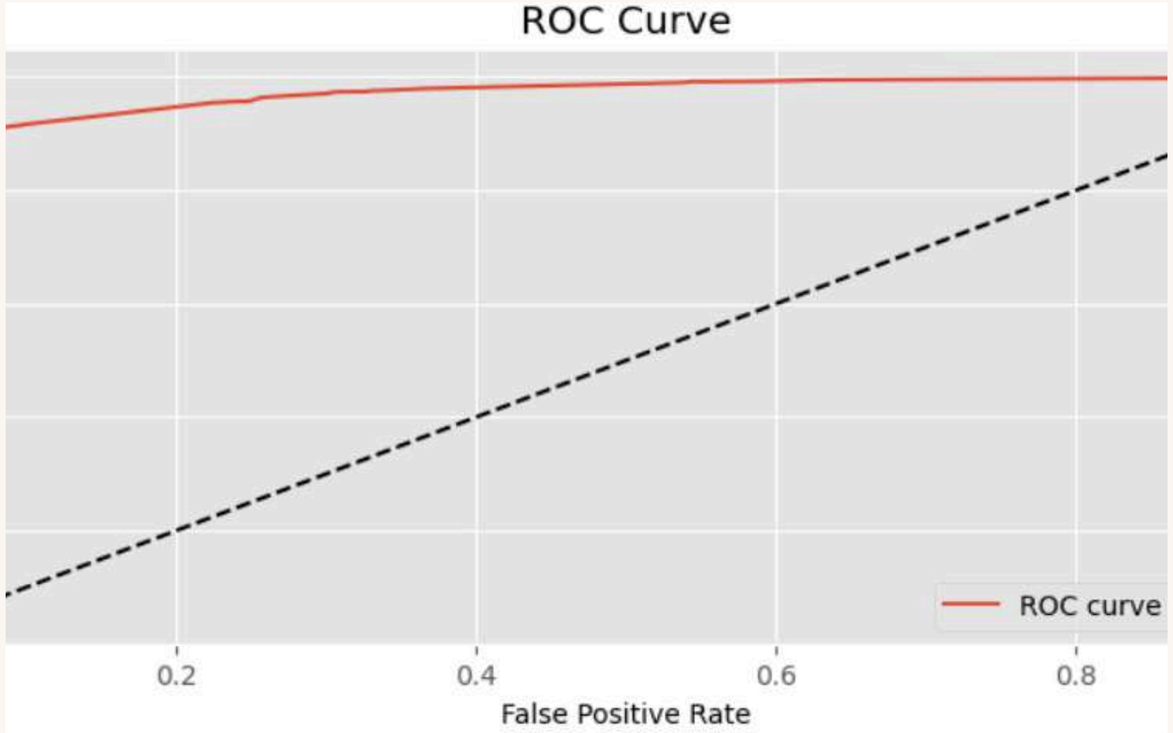
- Accuracy = 92.30%
- Sensitivity = 86.12%
- Specificity = 96.11%

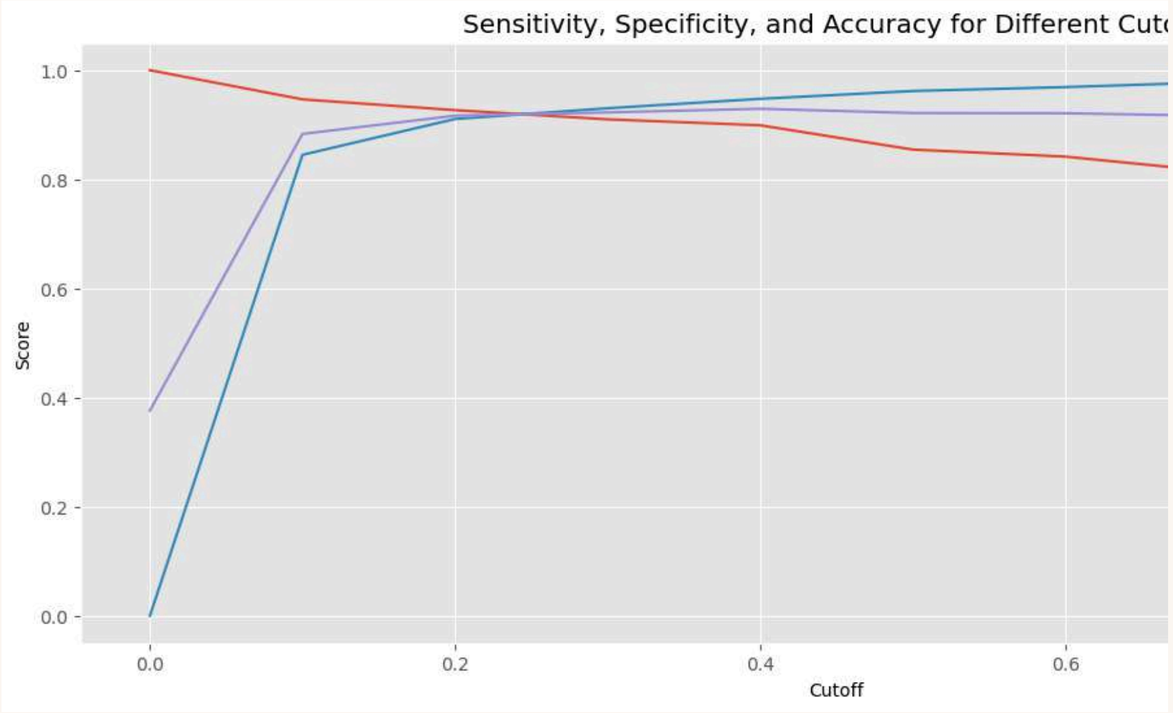
	Features	VIF
1	Lead Add Form	1.72
12	Tags_Will revert after reading the email	1.60
4	Last Activity_SMS Sent	1.48
2	Lead Source_Direct Traffic	1.38
5	Last Notable Activity_Modified	1.37
0	Total Time Spent on Website	1.34
3	Lead Source_Welingak Website	1.34
10	Tags_Others	1.24
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.07
6	Last Notable Activity_Olark Chat Conversation	1.01

Steps taken for Model Building

- Binary Mapping - replacing Yes/No responses to 0 and 1.
- Adding dummy variables
- Logistic Regression Model building, by first splitting the data into 70% Train and 30% Test sets
- Numeric values were standardized using StandardScaler
- RFE for Feature Selection and built a model with selected features
- Dropping high P-Values
- VIF Variance Inflation Factor to find correlation between the variables
- Dropping variables with high correlation to maintain VIF values below 5
- Interpreting Probabilities, Lead Score and Predictions on Train Data.

Model Building





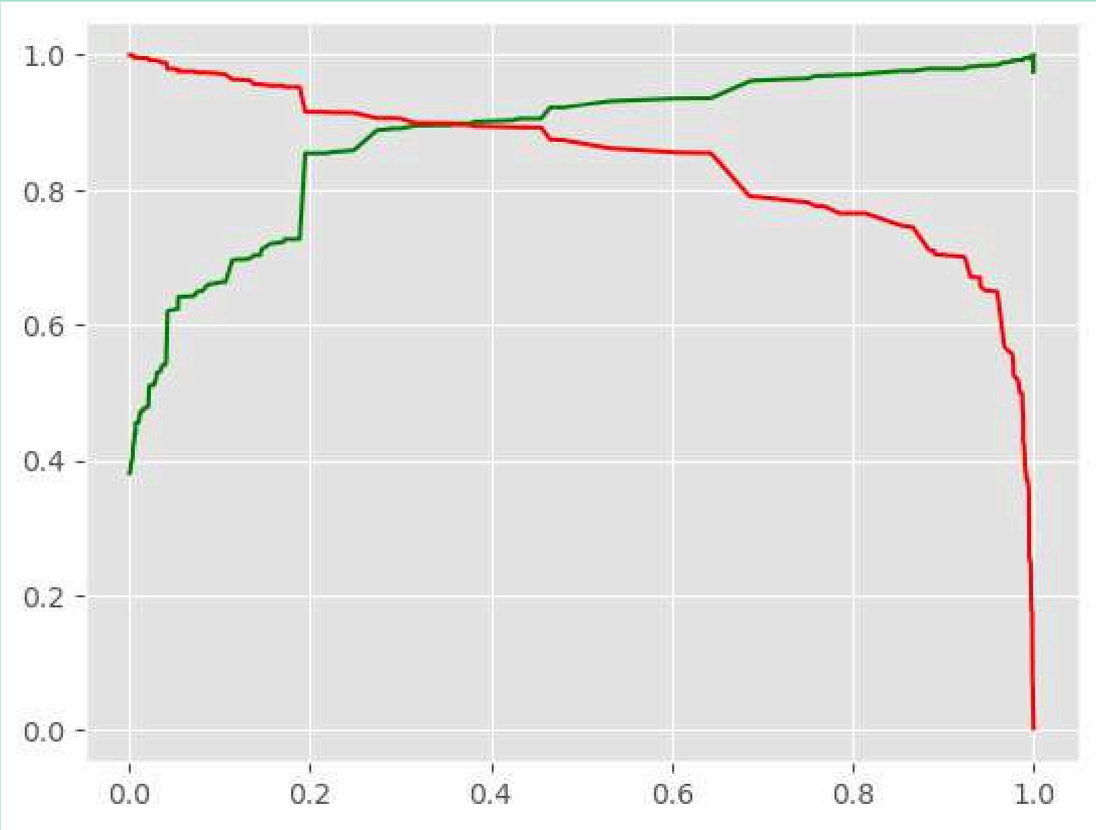
Inferences

- ROC curve area of 0.97 also indicating a good predictive model
- Optimal cut off point at 0.3 after plotting a graph leading to accuracy of 92.11%

Training Data Stats after optimal cut off

- Accuracy = 92.11%
- Sensitivity = 89.85%
- Specificity = 93.50%

Model Evaluation



Precision Recall Curve

Precision = 89.47% | Recall = 89.85%

Precision Recall Curve

- Precision Recall curve shows Sensitivity (True Positive Rate) and Specificity (True Negative Rate) as a function of probability threshold
- Precision is concerned with the quality of positive predictions. A high precision indicates that the model has a low rate of false positives.
- Recall is concerned with the quantity of the relevant instances captured by the model. A high recall indicates that the model has a low rate of false negatives.
- As X Education asked for a ballpark of the target lead conversion rate to be around 80%, this model does justice to their requirement.

Comparison of Testing and Training Data

Training Data

- Accuracy = 92.11%
- Sensitivity = 89.85%
- Specificity = 93.50%

Testing Data

- Accuracy = 92.51%
- Sensitivity = 91.39%
- Specificity = 93.20%

Recommendations

By comparing the training and testing data we can conclude that our model is predicting the conversion rate with high accuracy. X Education can use above model to successfully scale up their lead conversion rate.

Key observations for improvement

- It would be beneficial to increase leads from references and the Welingak website
- Working professionals are more likely to become customers
- Most of the lead conversion is taking place through google and direct traffic
- Lead conversion via SMS sent has the highest conversion rate and lead generation should be increased from the same
- Leads who opened the email also have a good conversion rate which could be improved further.

Thank you very much!