LEAD SCORING CASE STUDY

- *By Rajat Kawalkar*

PROBLEM STATEMENT

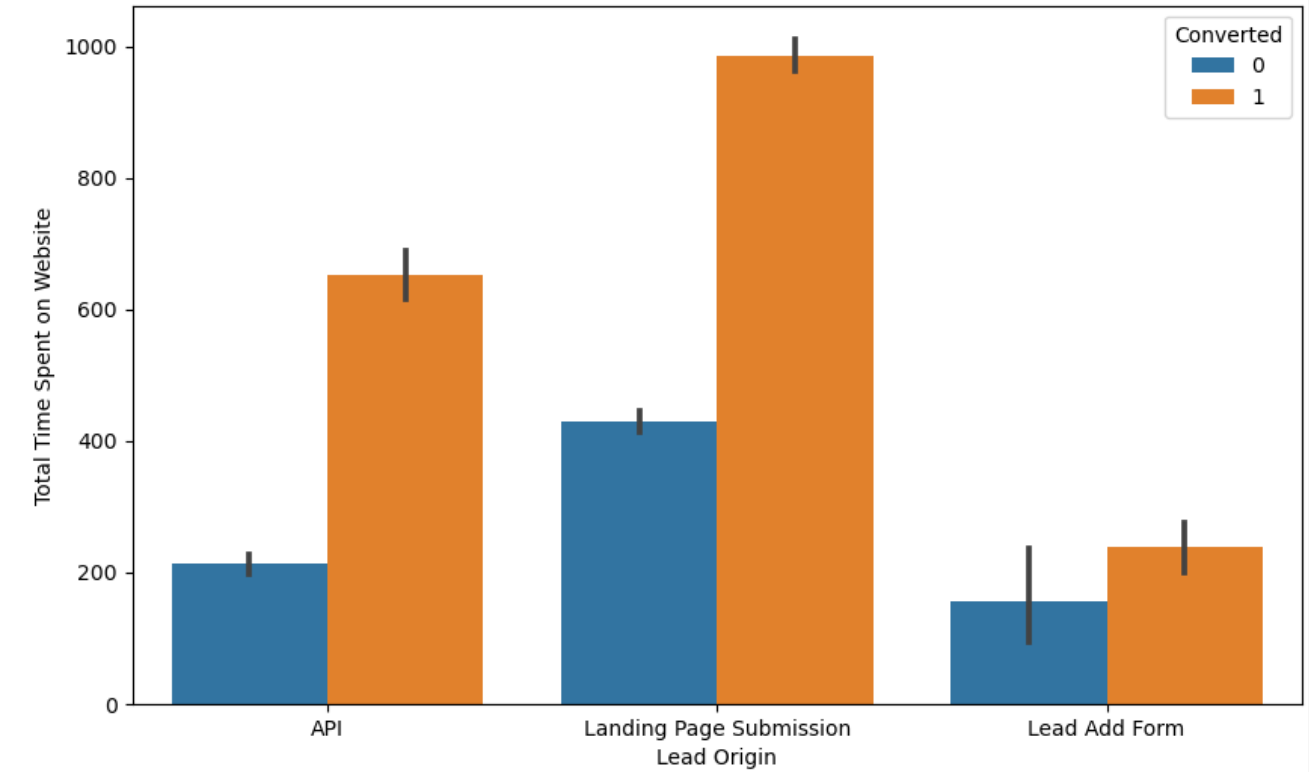
- X Education, an online education provider catering to industry professionals, typically achieves a lead conversion rate of approximately 30%.
- Despite generating a significant number of leads, the company struggles with a low conversion rate. For instance, out of 100 leads acquired in a day, only around 30 result in conversions.
- The pipeline exhibits a disparity where numerous leads enter the initial stage but only a fraction proceed to become paying customers. Effective nurturing of potential leads during the intermediate stage—such as providing product education and maintaining regular communication—is crucial for improving lead conversion rates.
- To enhance efficiency in this process, the company aims to pinpoint the most promising leads, commonly referred to as 'Hot Leads'.

BUISNESS SOLUTION

- The company needs you to develop a model that assigns a lead score to each lead.
- This score should reflect the likelihood of conversion, with higher scores indicating a greater chance of conversion and lower scores indicating a lower chance.
- To achieve this, we'll construct a logistic regression model that assigns a lead score ranging from 0 to 100.
- This score will help the company effectively target potential leads.

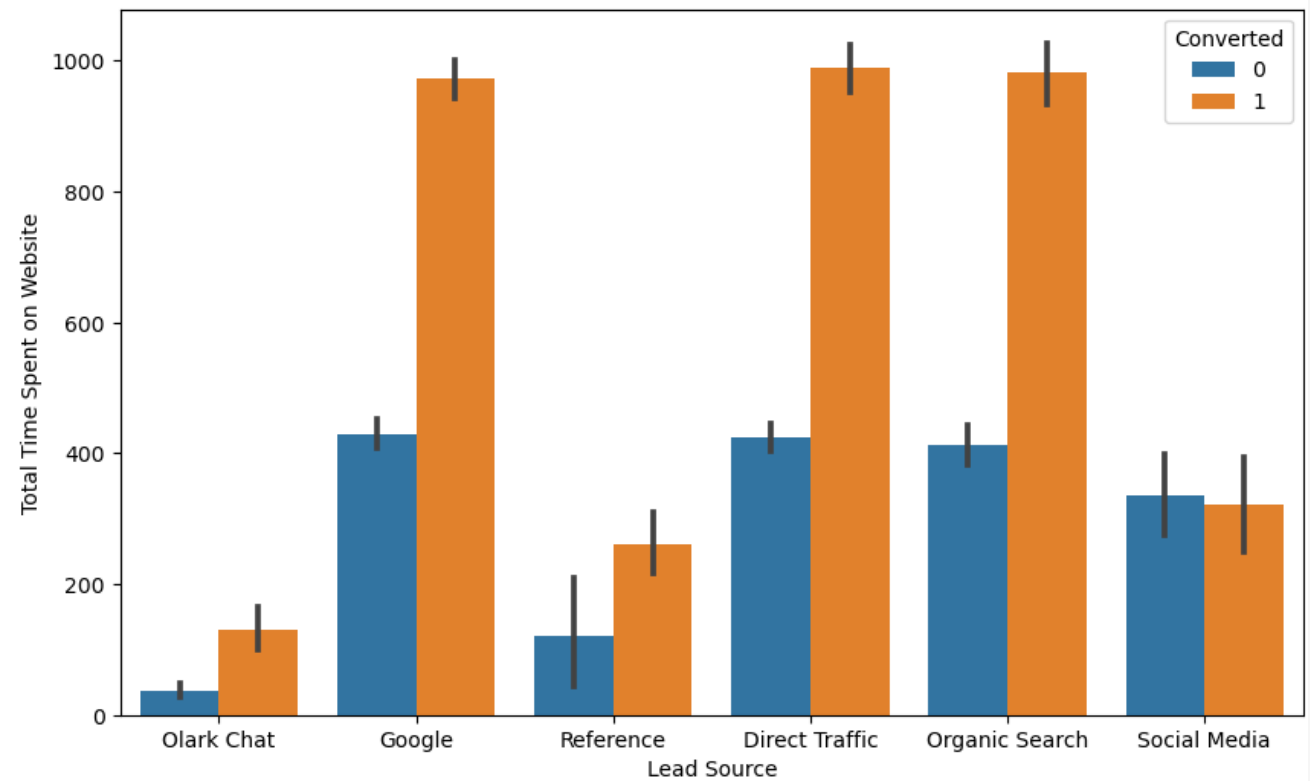
FLOWCHART TO ACHIEVE BUSINESS SOLUTION

- The initial steps involve comprehending the business problem at hand and thoroughly examining the datasets provided.
- This is followed by sourcing the data and conducting a detailed inspection.
- Subsequent stages include cleaning and manipulating the data, which entails identifying and addressing missing values through imputation.
- Exploratory Data Analysis (EDA) is then performed, comprising both univariate and bivariate analyses.
- Additionally, dummy variables are created for certain categorical variables.
- The dataset is further divided into training and testing sets, and feature scaling is applied to ensure uniformity.
- Model building ensues, involving feature selection using Recursive Feature Elimination (RFE), plotting Receiver Operating Characteristic (ROC) curves, determining the optimal cutoff point, and evaluating precision and recall metrics.
- Finally, predictions are made on the test set.



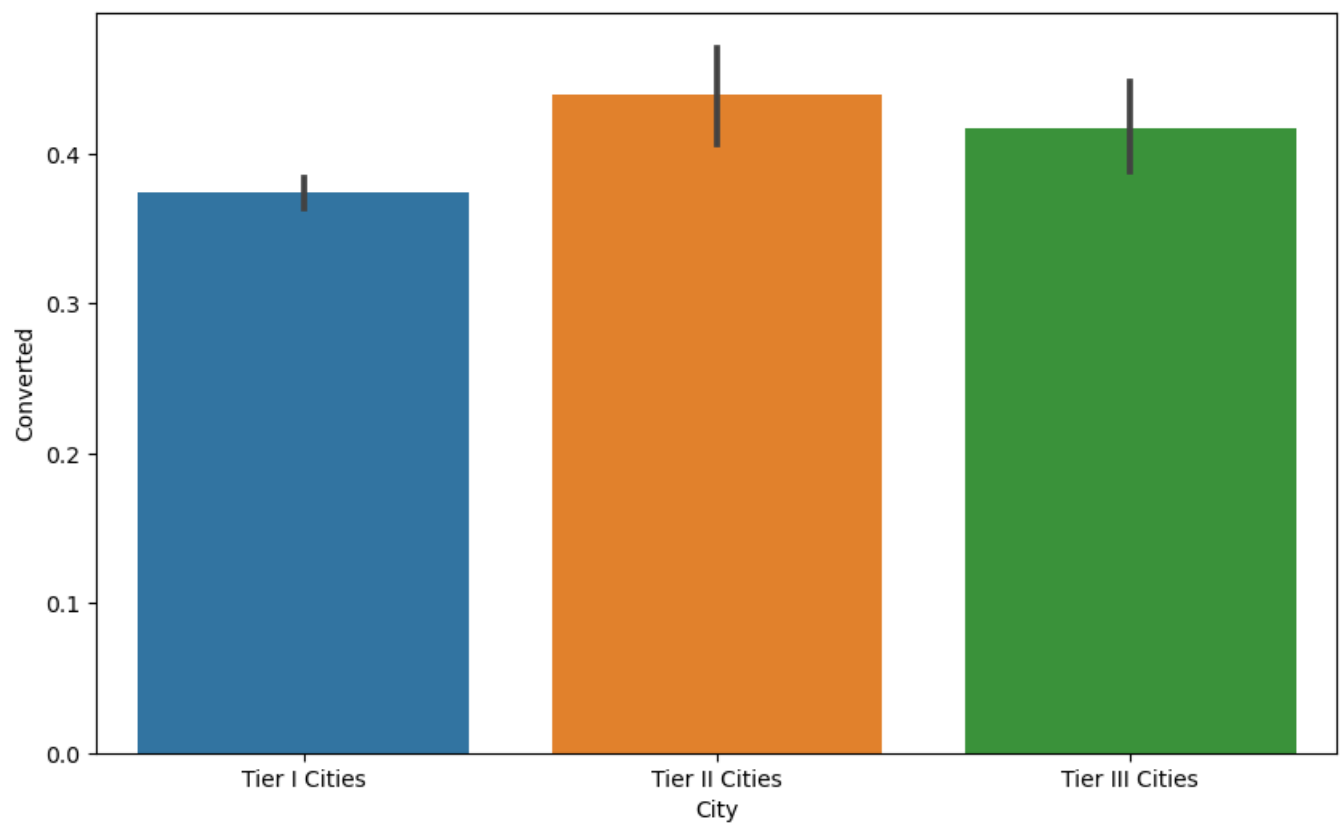
1. MULTIVARIATE ANALYSIS: ‘Lead Origin’ Column

- Through analyzing API and Landing Page submissions, it's evident that they generate the highest volume of leads as well as conversions.
- The Lead Add form, although generating fewer leads, exhibits a notably higher conversion rate.
- Based on this observation, the strategy entails generating more leads through the Lead Add Form and enhancing the conversion rate of leads originating from API and Landing Page Submissions.
- This approach aims to elevate the overall lead conversion rate.



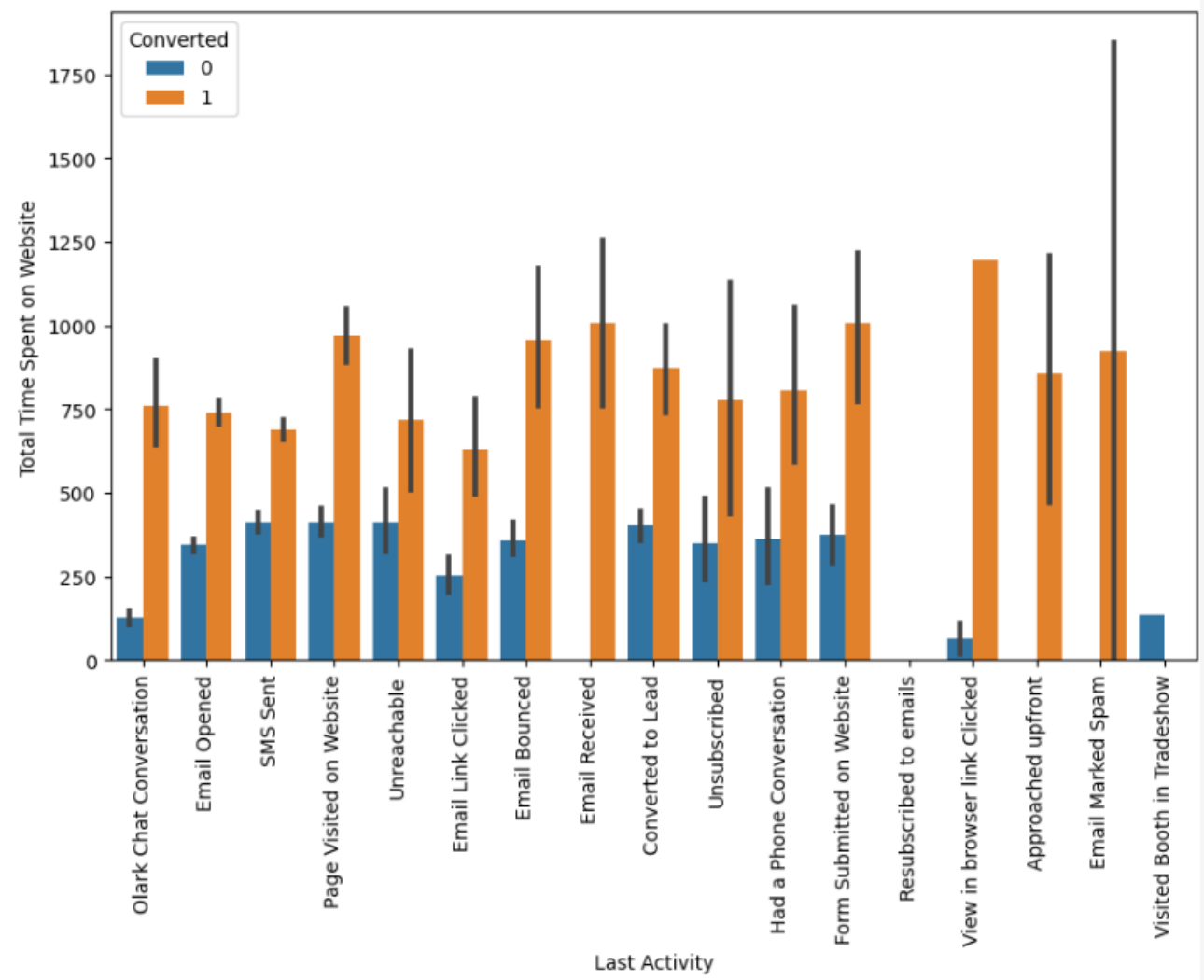
1. MULTIVARIATE ANALYSIS: ‘Lead Source’ Column

- Organic search, direct traffic, and Google exhibit a notable conversion rate.
- Reference and Olark Chat demonstrate the highest lead generation.
- Social media shows the lowest conversion rate in comparison to other channels.
- Based on these observations, there's a need to prioritize efforts on leads generated from 'Social Media', 'Organic Search', 'Direct Traffic', and 'Google' to enhance the overall lead conversion rate.



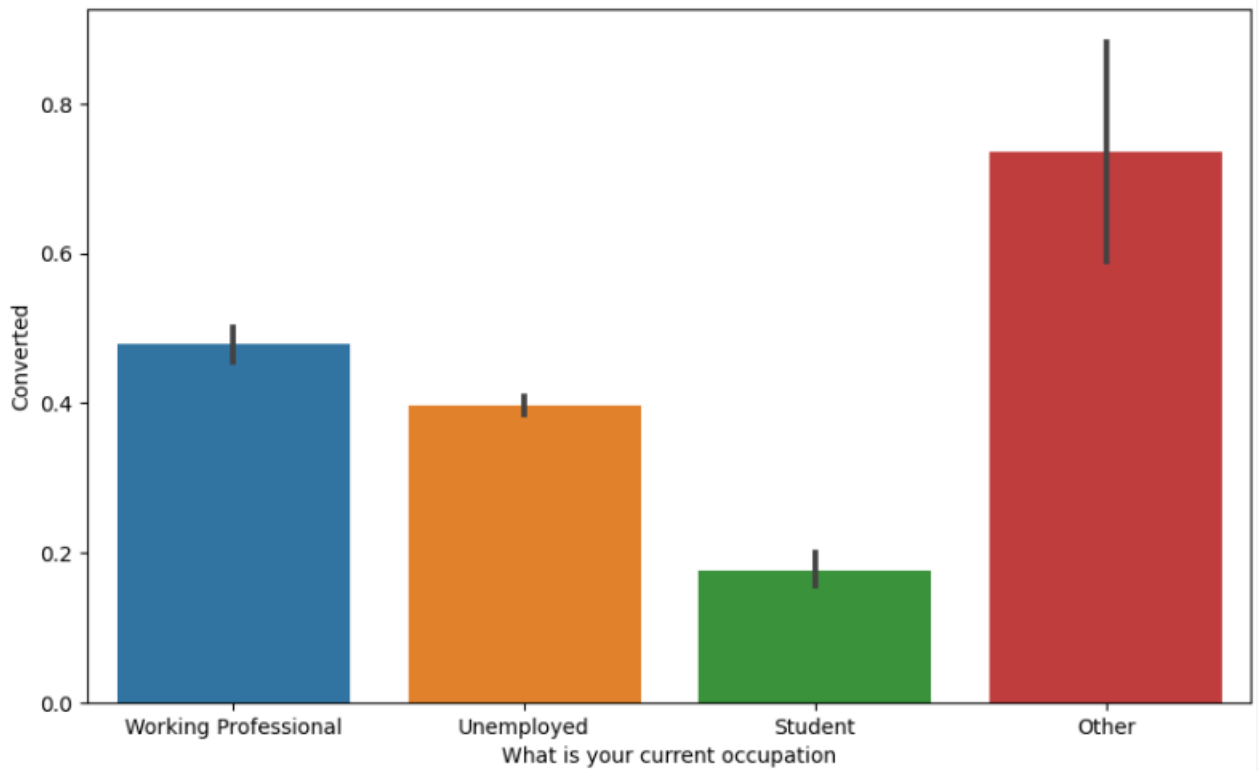
1. MULTIVARIATE ANALYSIS: ‘City’ Column

- All tiers of cities, including Tier I, Tier II, and Tier III, hold equal significance for analysis.



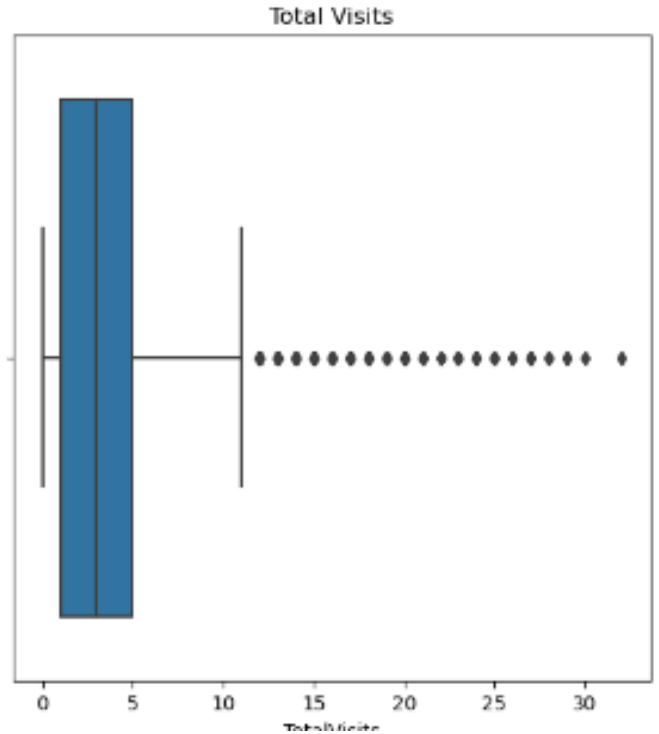
1. MULTIVARIATE ANALYSIS: ‘Last Activity’ Column

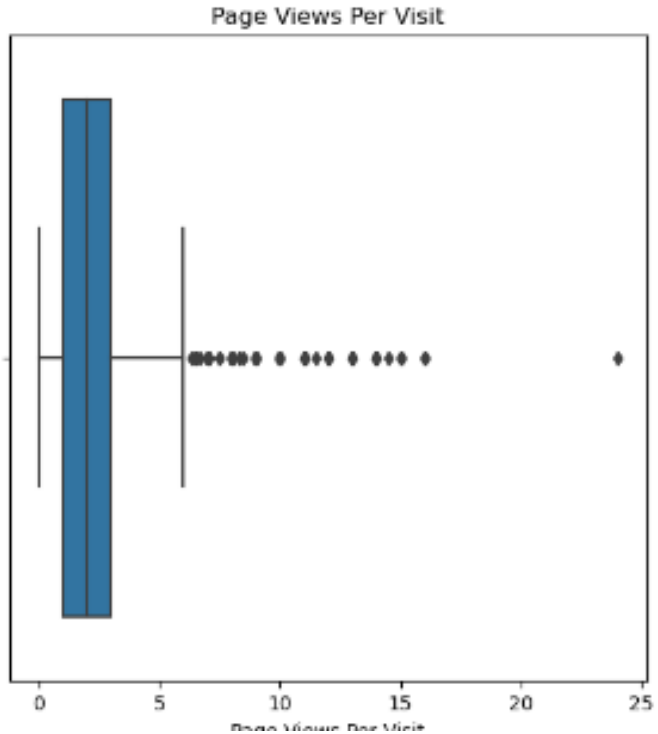
- The highest conversion rate is observed with the "View in Browser" link clicked, with Olark Chat Conversation following closely behind in terms of effectiveness.

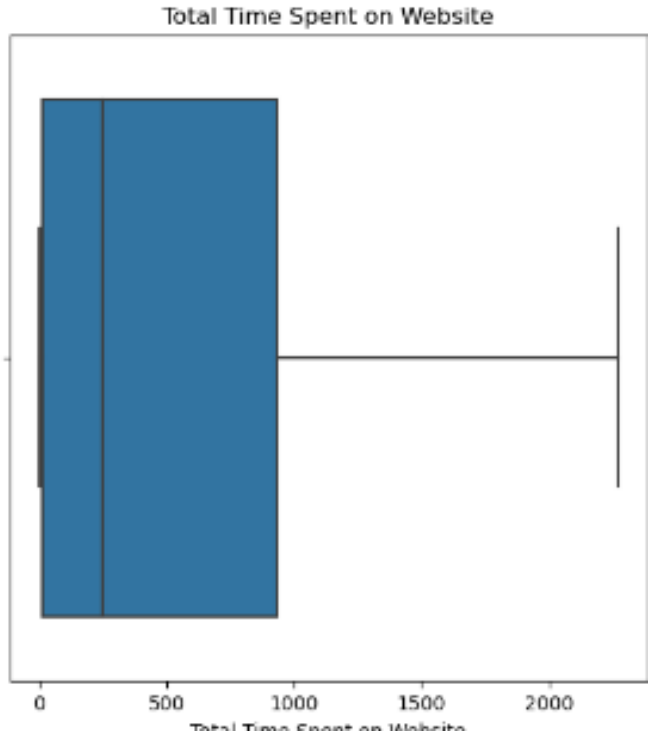


1. MULTIVARIATE ANALYSIS: ‘What is Your Current Occupation’ Column

- The largest category in terms of quantity is "Other," with "Working Professionals" following closely behind.







- **2. UNIVARIATE ANALYSIS: ‘Total Visits’, ‘Page Views Per Visit’ and ‘Total Time Spent on Website’ Column**

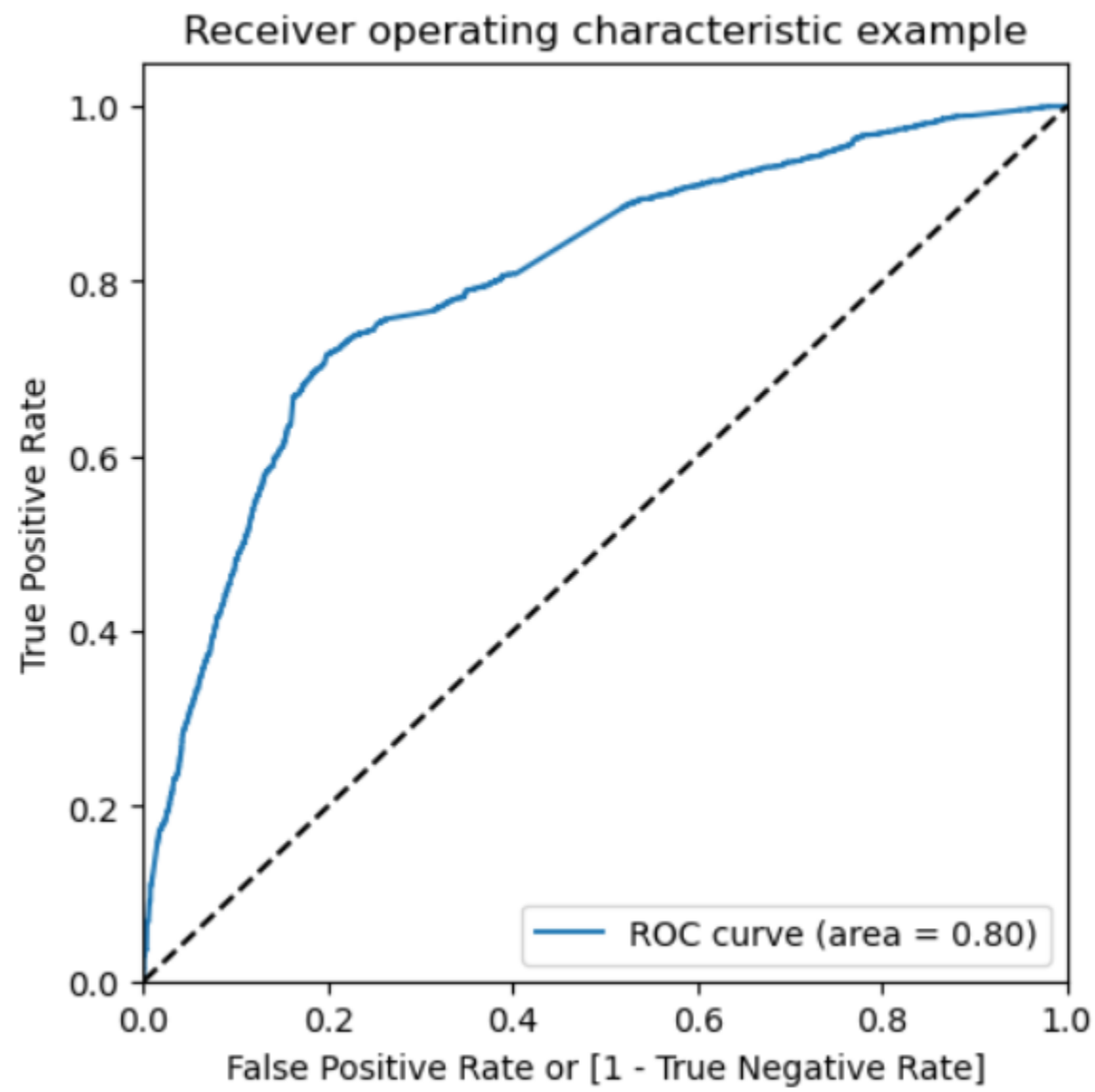
3. CORRELATION MAP: Before (top) and After (bottom) Removing Multi-collinearity

Dep. Variable:	Converted	No. Observations:	6461
Model:	GLM	Df Residuals:	6450
Model Family:	Binomial	Df Model:	10
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3368.5
Date:	Fri, 16 Feb 2024	Deviance:	6737.1
Time:	21:54:39	Pearson chi2:	6.62e+03
No. Iterations:	5	Pseudo R-squ. (CS):	0.2505
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4380	0.115	-12.490	0.000	-1.664	-1.212
Total Time Spent on Website	4.3336	0.147	29.458	0.000	4.045	4.622
Page Views Per Visit	-0.6692	0.430	-1.556	0.120	-1.512	0.174
Lead Source_Google	0.2962	0.071	4.169	0.000	0.157	0.436
Lead Source_Olark Chat	0.8220	0.110	7.495	0.000	0.607	1.037
Lead Source_Reference	3.8917	0.203	19.156	0.000	3.494	4.290
Lead Source_Social Media	1.6542	0.151	10.962	0.000	1.358	1.950
Specialization_Other	-0.2861	0.066	-4.353	0.000	-0.415	-0.157
What is your current occupation_Student	-1.3620	0.126	-10.824	0.000	-1.609	-1.115
What is your current occupation_Unemployed	-0.2483	0.078	-3.201	0.001	-0.400	-0.096
City_Tier III Cities	0.1375	0.093	1.472	0.141	-0.046	0.321

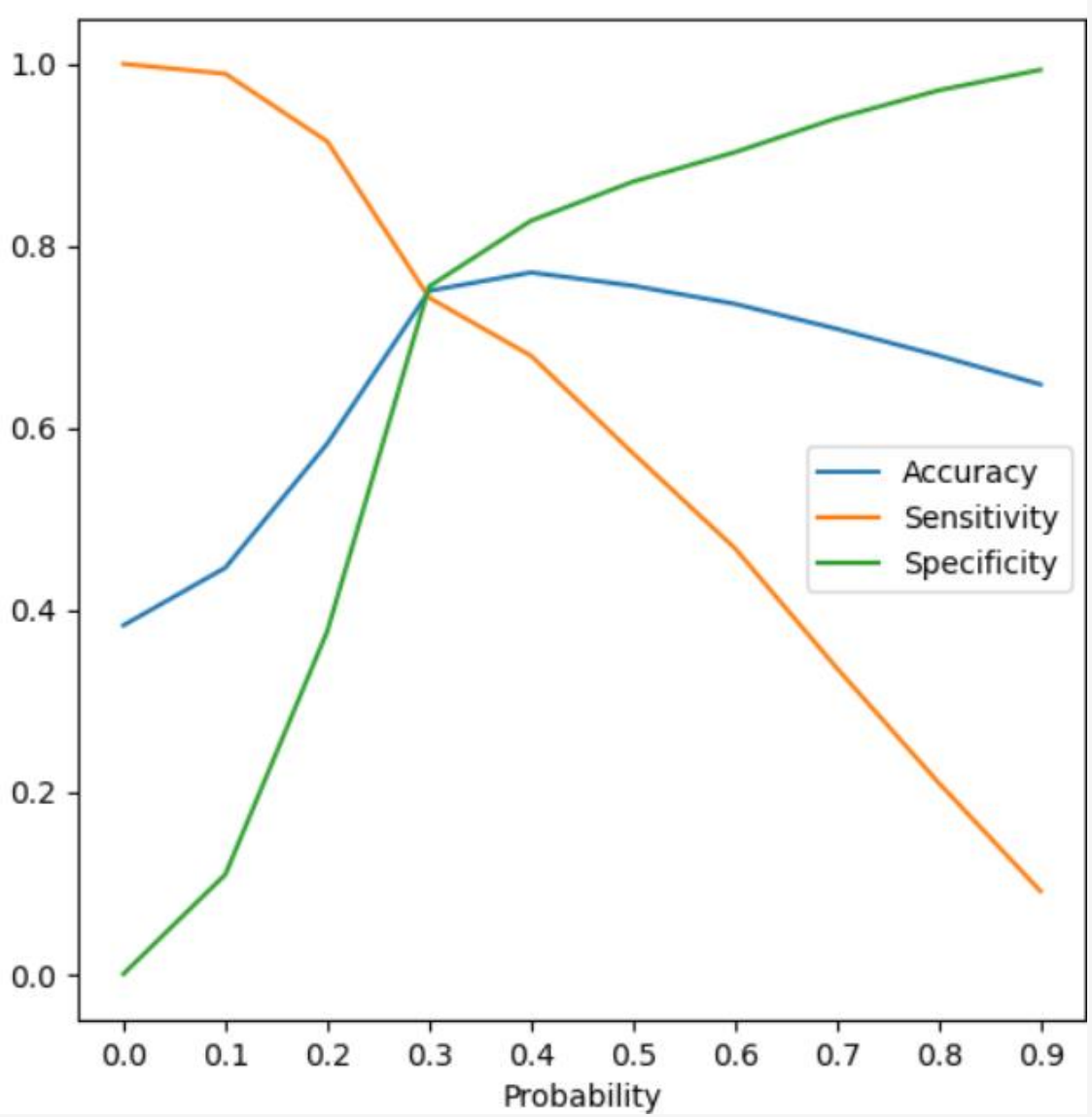
	Features	VIF
0	What is your current occupation_Unemployed	3.41
1	Specialization_Other	2.77
2	Page Views Per Visit	2.74
3	Total Time Spent on Website	2.04
4	Lead Source_Olark Chat	1.77
5	Lead Source_Google	1.65
6	What is your current occupation_Student	1.44
7	City_Tier III Cities	1.16
8	Lead Source_Social Media	1.11
9	Lead Source_Reference	1.10

4. FINAL MODEL VISUALIZATION STATISTICS



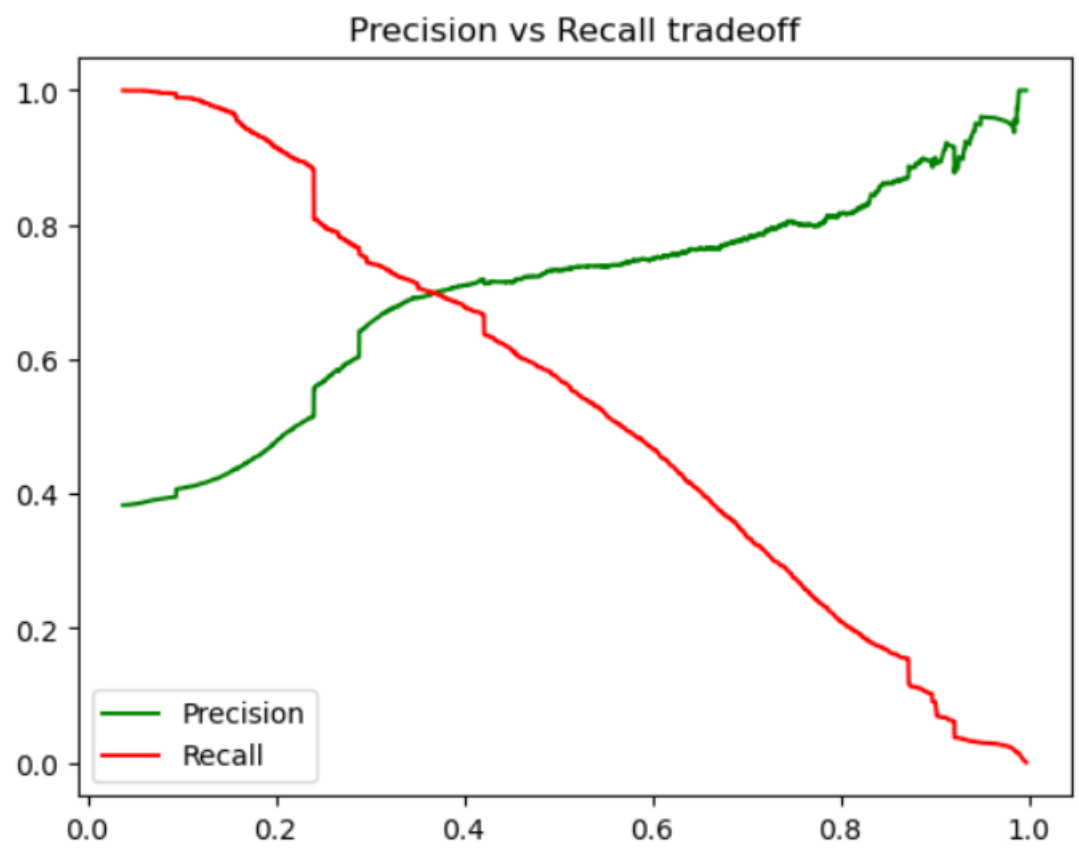
5. ROC CURVE INTERPRETATION

- The curve leans towards the left side of the boundary more than the right, indicating high accuracy of our model.
- The curve encompasses 80% of the total area beneath it.



6. OPTIMAL CUTOFF POINT

- From the plot depicted above, it's evident that the probability threshold is approximately 0.4.
- We're contemplating using 0.4 as the threshold to balance sensitivity and accuracy.



7. PRECISION AND RECALL TRADEOFF

- We can discern from the preceding plot that there exists a trade-off between Precision and recall, with the optimal point appearing to be around 0.36.

8. INFERENCES

- Observations regarding the comparison of values between the training and testing datasets are as follows:

- - In the training data:
- 1. Accuracy stands at 76.85%.
- 2. Sensitivity is recorded at 68.07%.
- 3. Specificity shows a value of 82.35%.

- - In the testing data:
- 1. Accuracy increases to 77.58%.
- 2. Sensitivity decreases slightly to 68.88%.
- 3. Specificity notably improves to 83.06%.

9. CONCLUSIONS

- Crucial features identified during the model training, contributing to a favorable conversion rate, include:

1. Total time spent on the website.
2. Lead source specified as 'Reference.'
3. Lead source originating from social media platforms.

- We observed a Recall value greater than the Precision value, which is deemed satisfactory from a business standpoint.
- Upon comparing the test dataset with the training dataset, the Sensitivity, Specificity, and Accuracy metrics fall within an acceptable range.
- The model demonstrates adaptability to align with company requirements, yielding positive outcomes.