Lead Score Case Study

Batch : DS C56

Meghna, Moksh & Nagaraj T K S

Problem statement

- X Education is an online education company catering to professionals in various industries. Their website attracts numerous visitors daily, thanks to marketing efforts on platforms like Google. These visitors explore the **course offerings**, **complete inquiry forms**, and engage with **educational content**.

- Leads are generated when individuals provide their contact information by filling out forms or through referrals. After acquiring these leads, the sales team initiates outreach, including calls and emails, with the goal of converting them into paying customers. However, the current lead conversion rate stands at a modest **30%**.

- To enhance the efficiency of their lead conversion process, X Education aims to pinpoint the most promising leads, often referred to as '**Hot Leads.'** By identifying this select group, the company can focus its sales efforts more effectively, potentially leading to a higher conversion rate. This strategic shift will enable the sales team to concentrate their efforts on engaging with individuals who are more likely to become valued customers, as opposed to making contact with everyone indiscriminately.

Business Objective

- To assist the company in identifying the most promising leads, often referred to as **'Hot Leads,'** with an impressive lead conversion rate of approximately **80%**:

- **Develop a model** that assigns a **lead score** to each **potential customer**, effectively distinguishing those with a higher **likelihood** of **conversion** from those with a lower chance.

- Empower the sales team to redirect their efforts towards **engaging** with **potential leads**, thus preventing them from making fruitless phone calls.

Solution Strategy

- Data Sourcing
- Data **Cleaning** and **Preparation**
- Conducting **Exploratory Data Analysis**
- Scaling Features
- Data Division into **Test** and **Train** Sets
- Constructing a Logistic Regression Model and Computing Lead Scores
- Assessing Model Performance through Various Metrics, such as **Specificity** and **Sensitivity**, or **Precision** and **Recall**
- Applying the Optimal Model to the Test Data, Informed by Sensitivity and Specificity Metrics

Data Sourcing, Cleaning & Preparation:

- Retrieve Data from the Source
- Transform the data into a clean, analytically suitable format
- Eliminate duplicate entries
- Address any outliers in the data
- Conduct Exploratory Data Analysis
- Standardize Features for consistency and comparability

Feature Scaling and Splitting data set:

- Standardize Numeric Data Features
- Divide the data into training and testing sets.

Model building:

- Employ R**ecursive Feature Elimination (RFE)** for Feature Selection
- Establish the Optimal Model Utilizing Logistic Regression
- Compute and Evaluate Model Performance by Assessing Multiple Metrics such as **accuracy**, **sensitivity**, **specificity**, **precision**, and **recall**.

Result:

- Calculate the lead score and verify if it results in an 80% conversion rate in the final predictions.
- Assess the final prediction's performance on the test set by applying a cutoff threshold derived from sensitivity and specificity metrics.

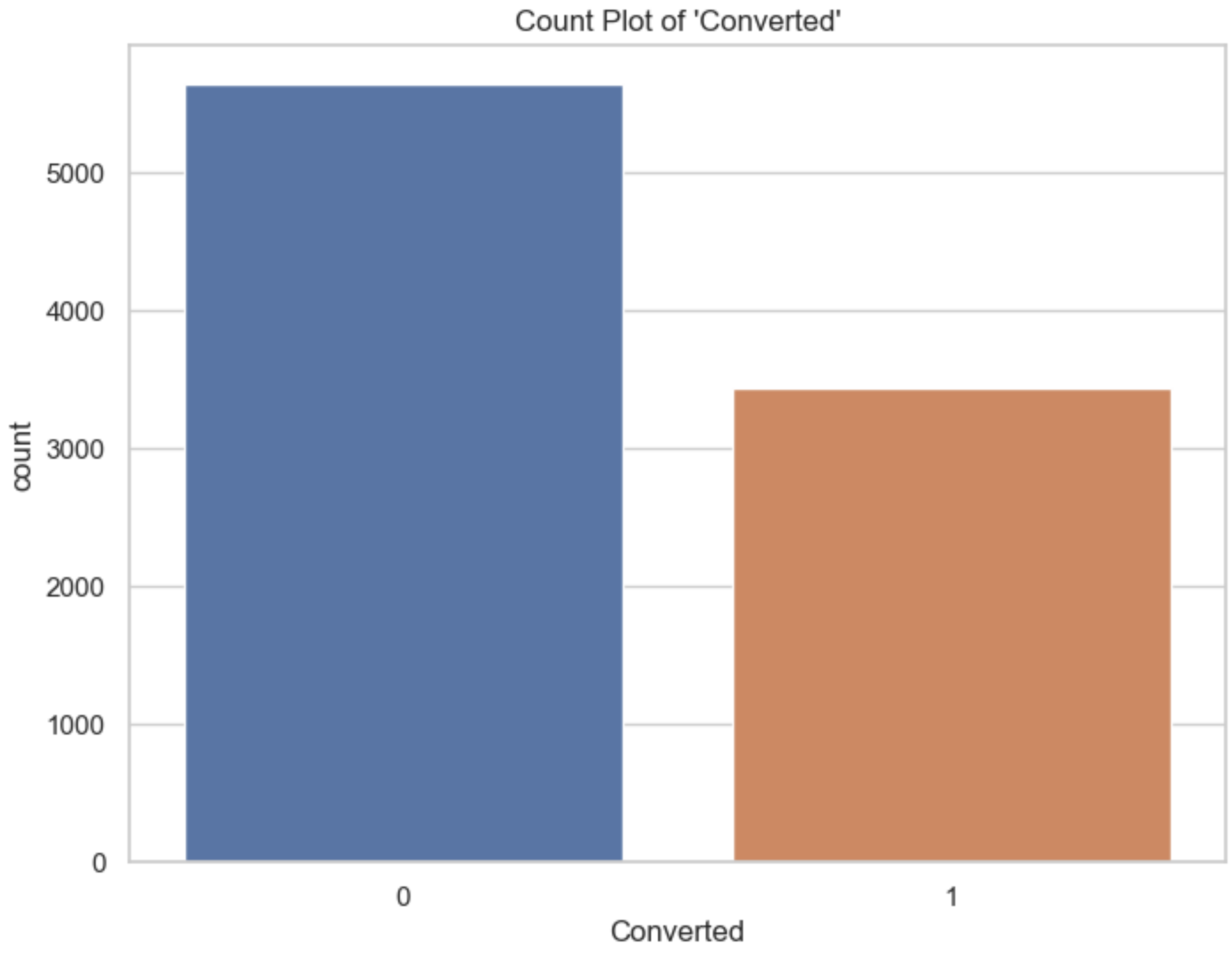
Exploratory Data Analysis

Key observations derived from the data:

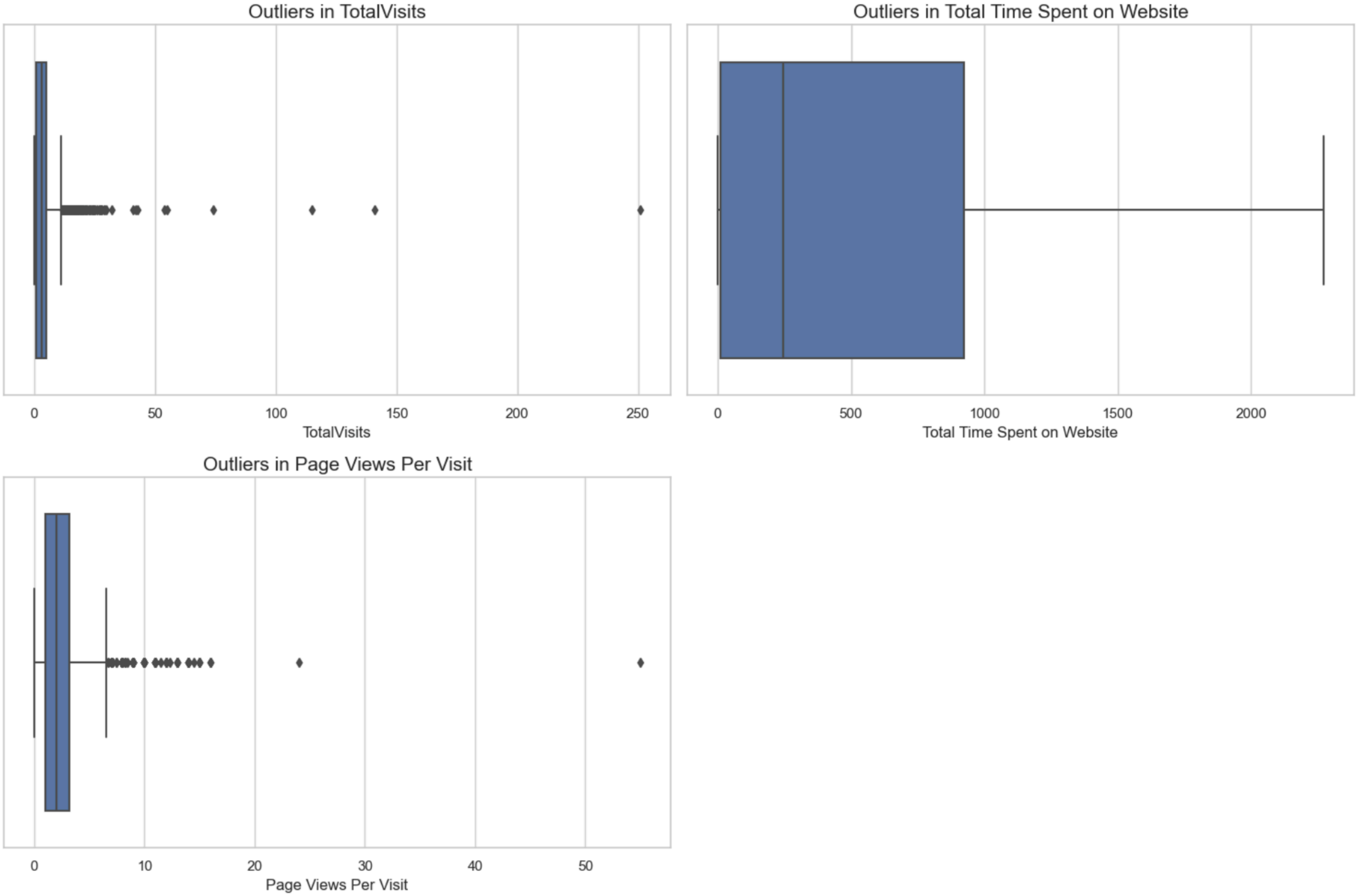
1. Within our dataset, there are a total of **9,240** distinct customer entries, and our objective is to pinpoint those with the greatest likelihood of conversion.

We can categorize potential leads based on their **Lead Score**, which represents their probability of conversion.

Among the 9,240 entries, approximately **37%** of leads have successfully converted, while the remaining **73%** have not.

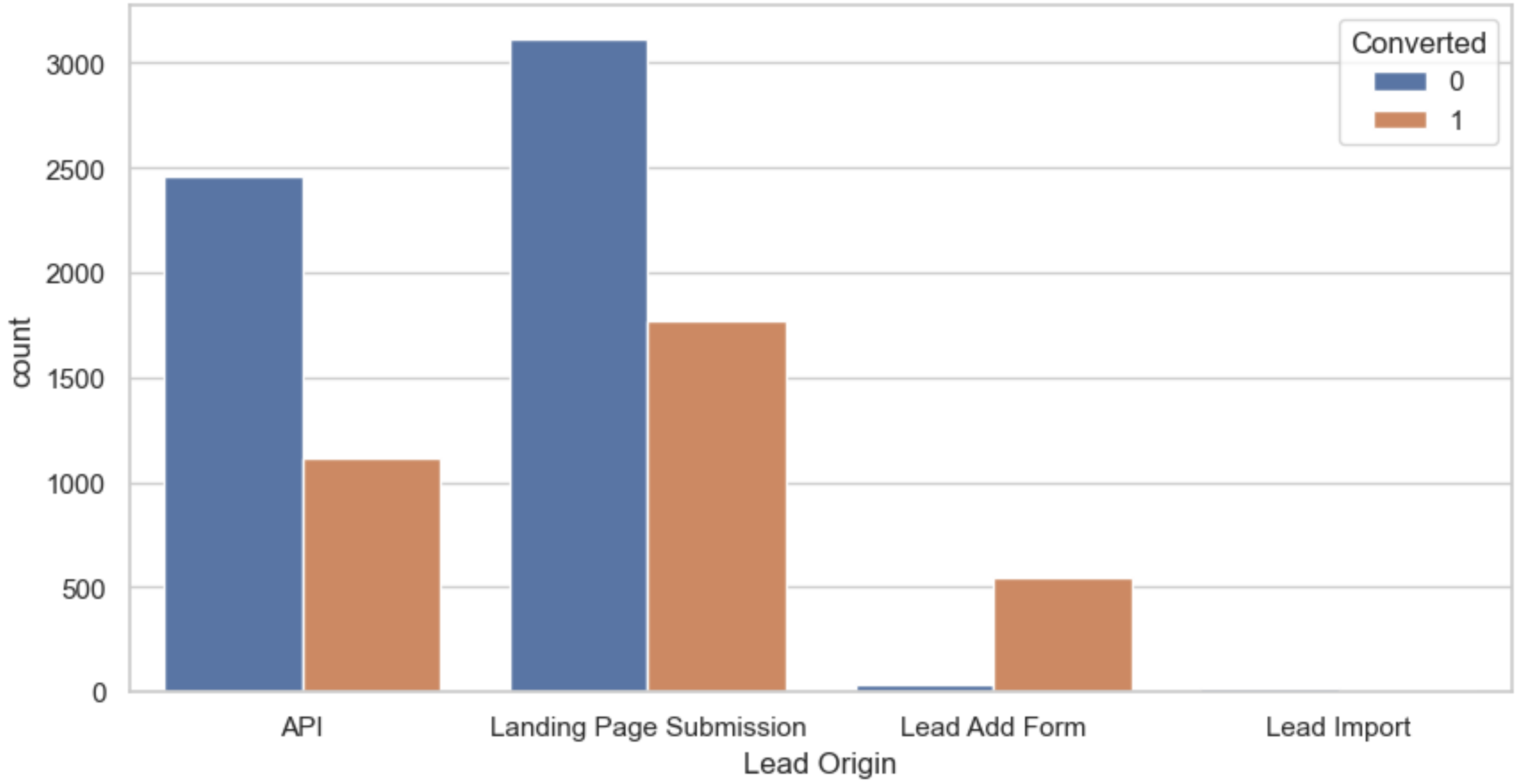


2. For numerical columns **‘TotalVisits’**, **‘’Time Spent on Website, ‘Views Per Visit’**, we observe some outliers present as below



3. For categorical columns,

A) **Lead Origin:**

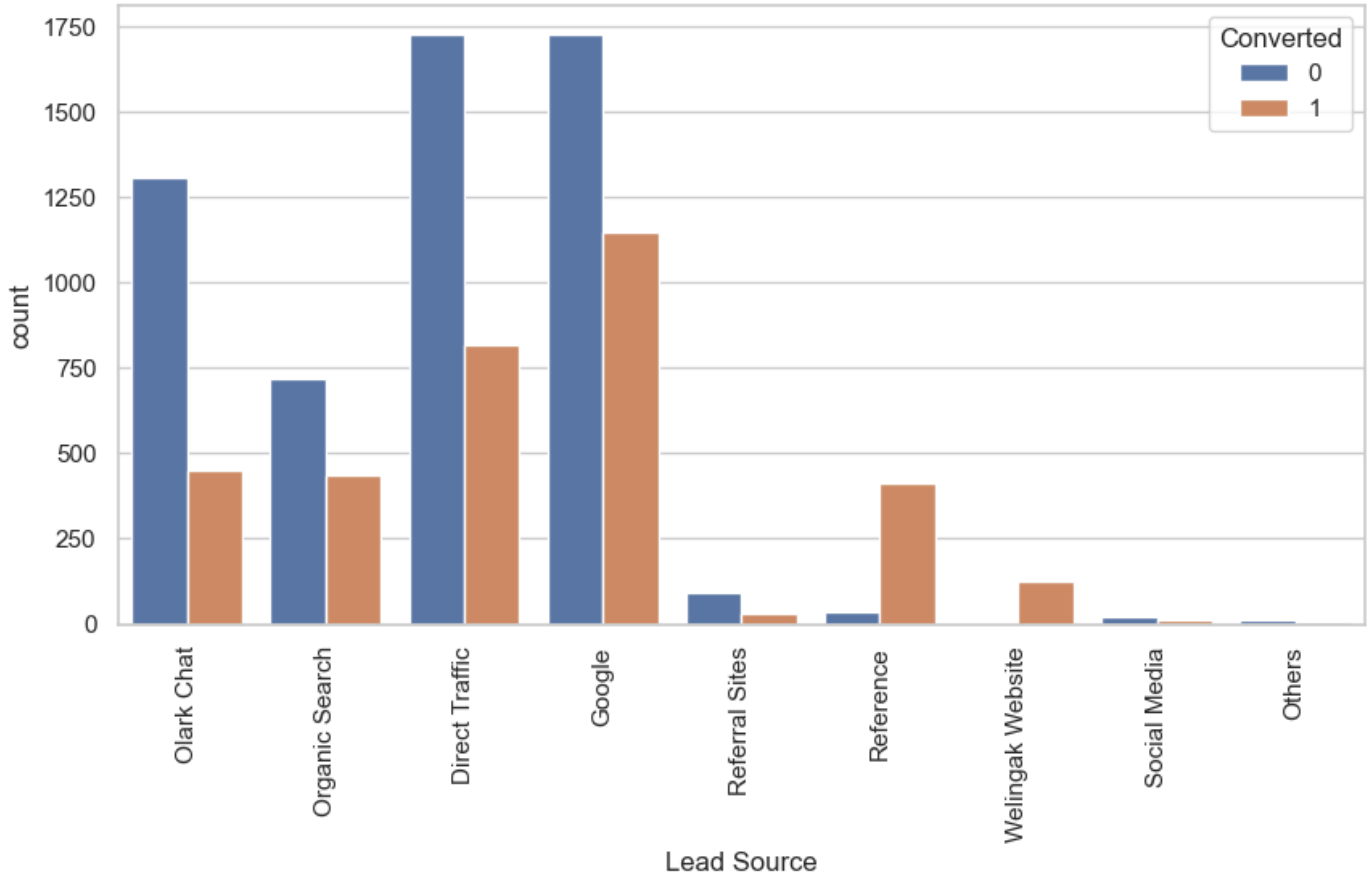


1. The majority of leads are generated from customers who were initially identified as leads through Landing Page submissions.

2. Customers who originate from Lead Add Forms exhibit a notably high likelihood of conversion, although their number is relatively low.

3. Lead origins such as API and Lead Import demonstrate the lowest conversion rates, with only a limited number of customers originating from Lead Import.

B) Lead Source :

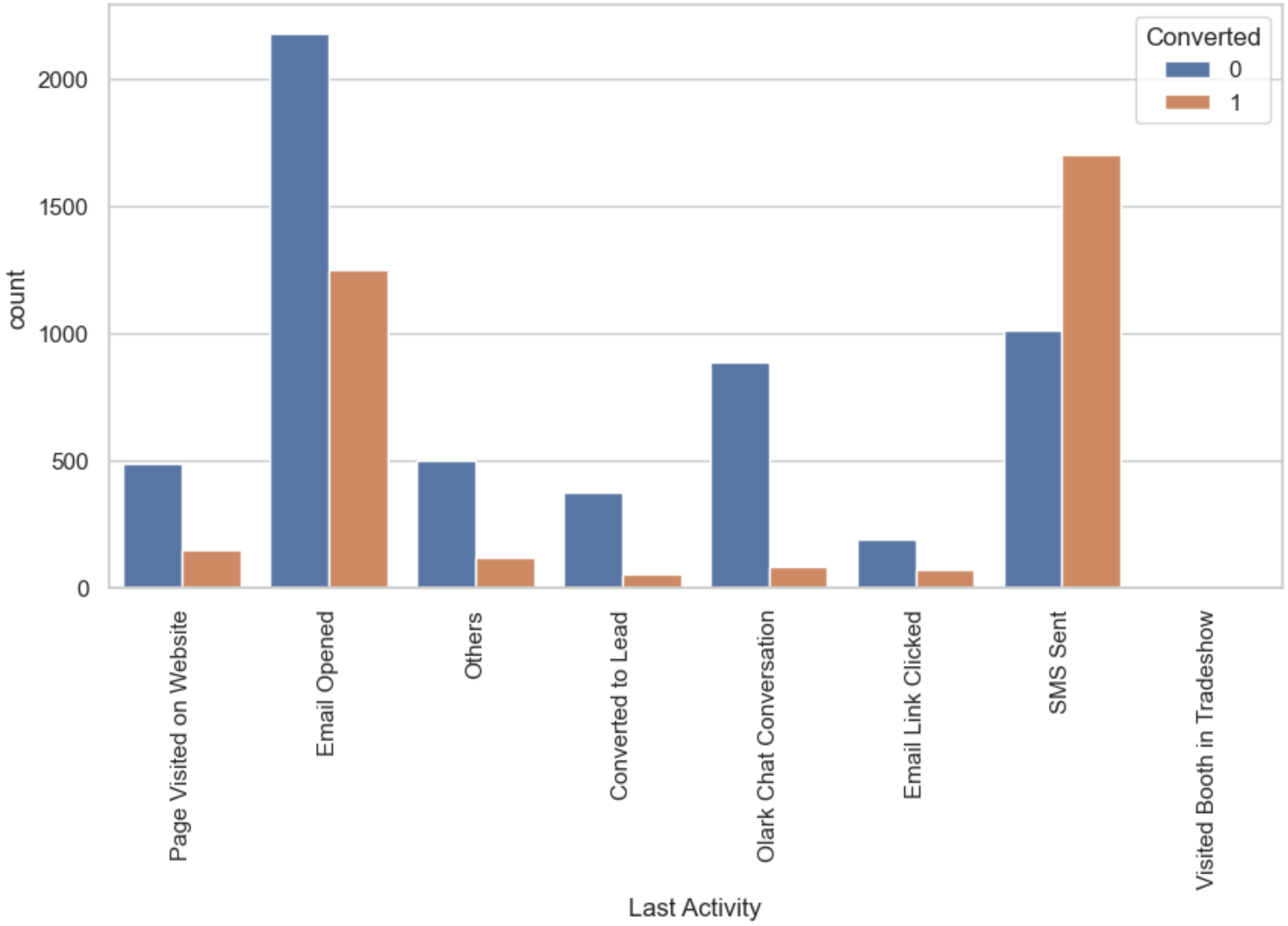


1. The primary sources of leads predominantly stem from Google and Direct Traffic.

2. Among these sources, leads originating from Google exhibit the highest probability of conversion.

3. Leads sourced from Reference demonstrate the maximum likelihood of conversion.

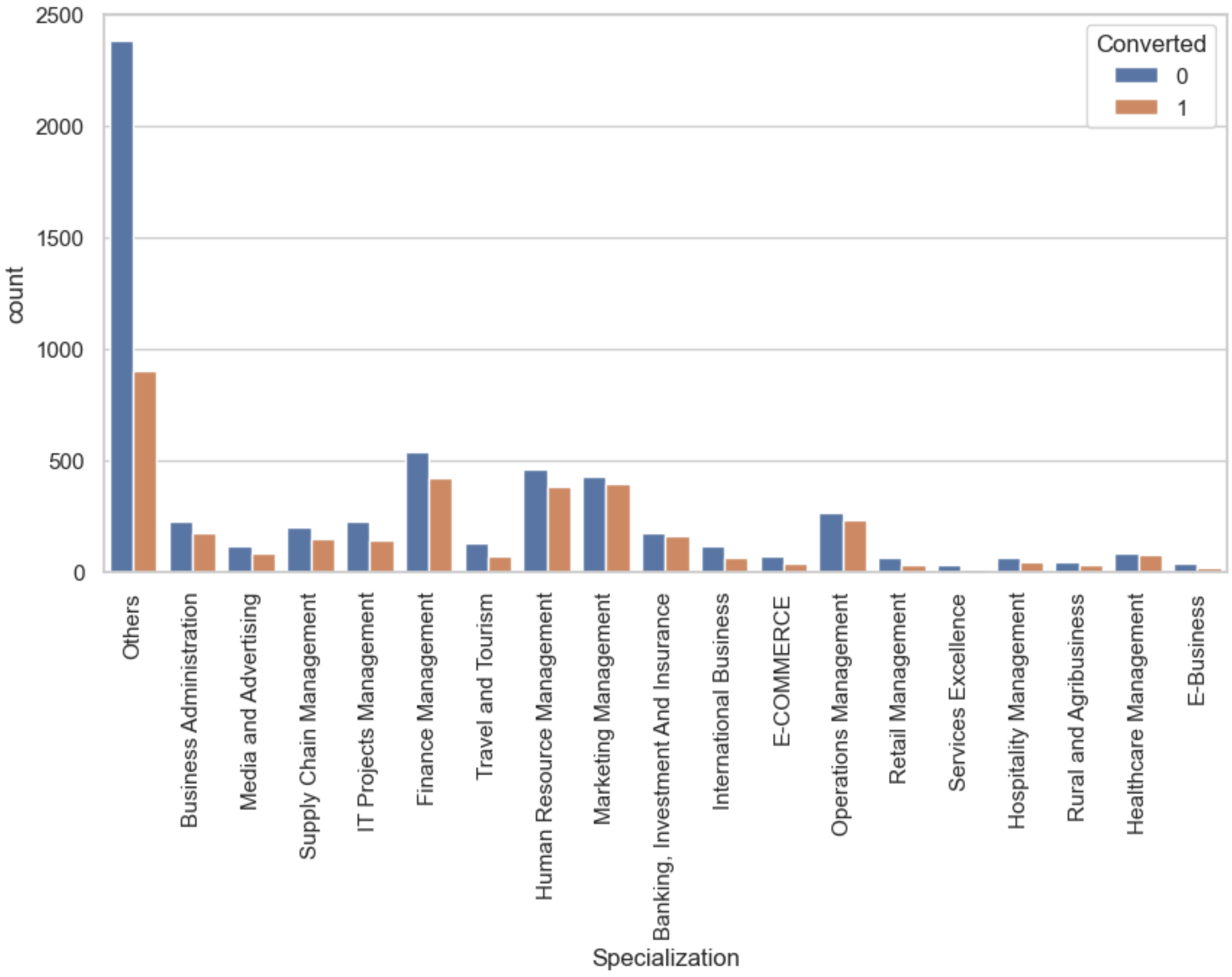
C) Last Activity:



1. Customers whose last recorded activity was sending SMS messages show a notably higher conversion rate, reaching approximately 63%.

2. The majority of customers have their last activity recorded as Email Opened, and this group maintains a conversion rate of around 36%.

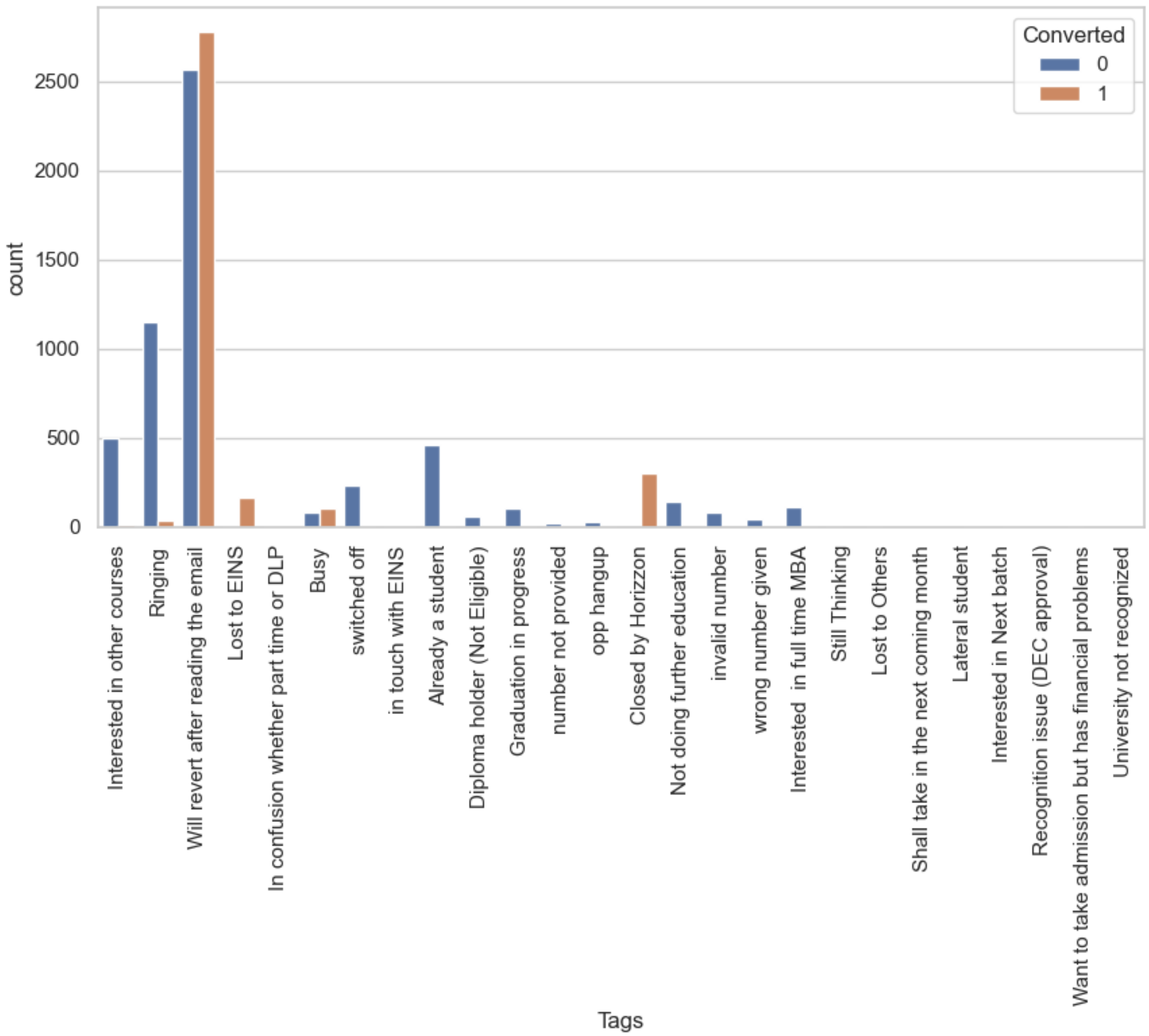
D) Specialization:



1. The majority of leads have indicated their specialization as Management or fall into the **"Others"** category.

2. Leads with a specialization in Rural & Agribusiness demonstrate the lowest probability of conversion.

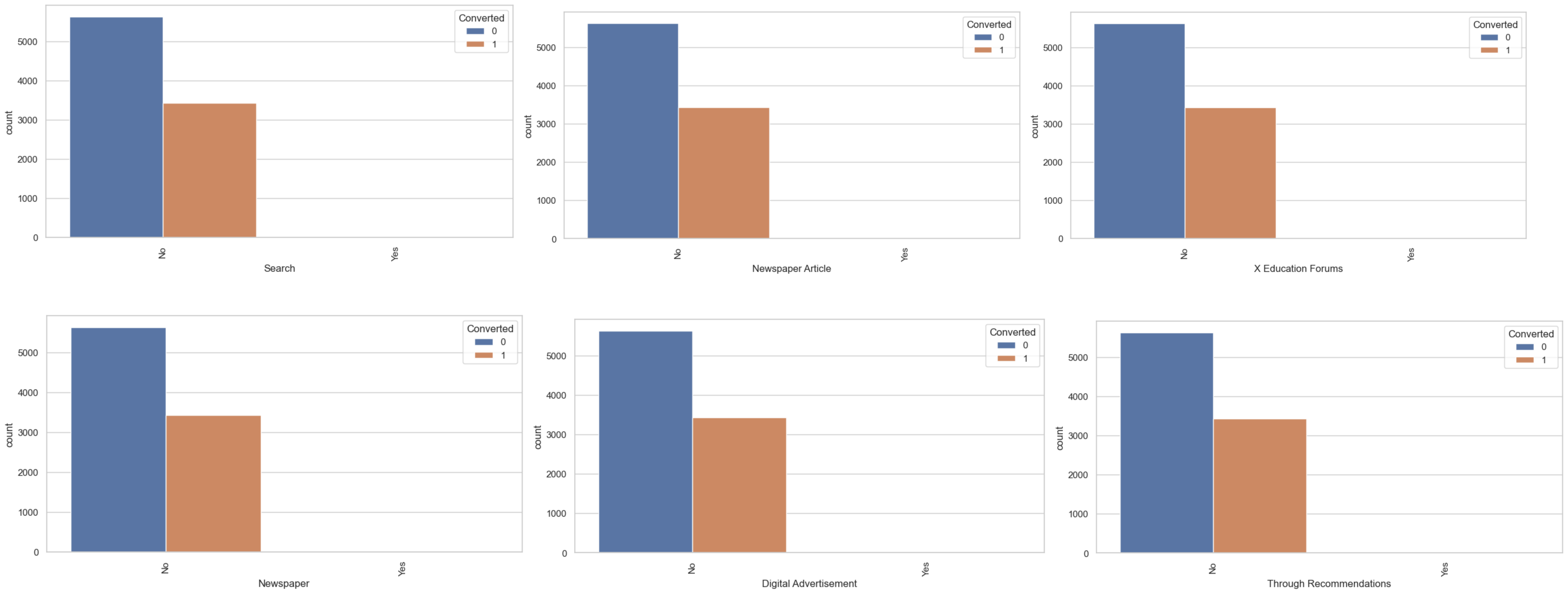
E) Tags:



Greater attention should be directed towards leads that exhibit a tendency to respond after reading emails, given their potential as promising leads with higher conversion rates.

F) Other Columns:

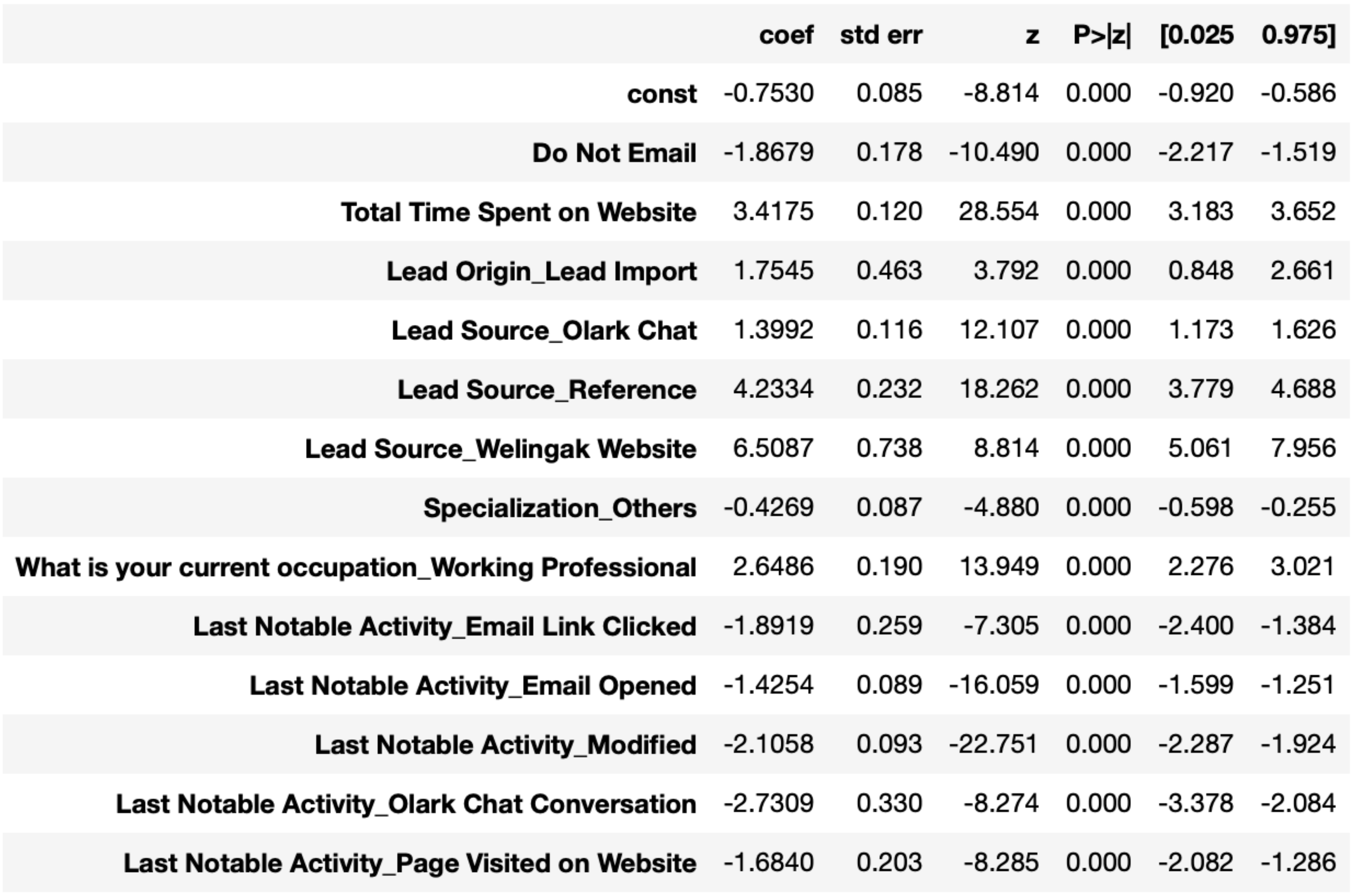
There is relatively little influence on conversion rates observed from **Search, Newspaper Article, X Education Forums, Digital advertisements, and recommendations.**



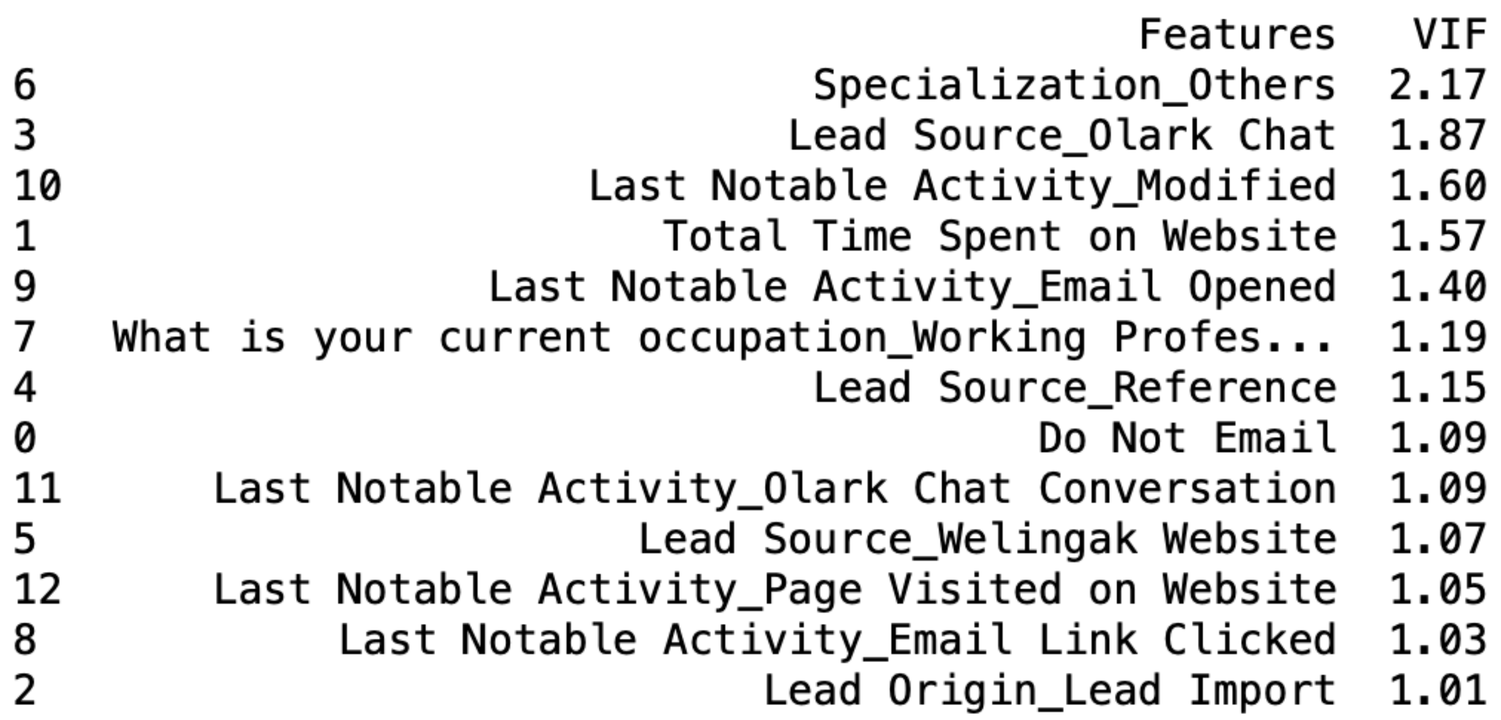
Key Factors Influencing Lead Generation

The below features influence great in lead conversion,

- Lead Source_Welingak Website
- Lead Source_Reference
- Total Time Spent on Website
- What is your current occupation_Working Professional
- Lead Origin_Lead Import
- Lead Source_Olark Chat
- Specialization_Others
- Last Notable Activity_Email Opened
- Last Notable Activity_Page Visited on Website
- Do Not Email
- Last Notable Activity_Email Link Clicked
- Last Notable Activity_Modified
- Last Notable Activity_Olark Chat Conversation



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7530	0.085	-8.814	0.000	-0.920	-0.586
Do Not Email	-1.8679	0.178	-10.490	0.000	-2.217	-1.519
Total Time Spent on Website	3.4175	0.120	28.554	0.000	3.183	3.652
Lead Origin_Lead Import	1.7545	0.463	3.792	0.000	0.848	2.661
Lead Source_Olark Chat	1.3992	0.116	12.107	0.000	1.173	1.626
Lead Source_Reference	4.2334	0.232	18.262	0.000	3.779	4.688
Lead Source_Welingak Website	6.5087	0.738	8.814	0.000	5.061	7.956
Specialization_Others	-0.4269	0.087	-4.880	0.000	-0.598	-0.255
What is your current occupation_Working Professional	2.6486	0.190	13.949	0.000	2.276	3.021
Last Notable Activity_Email Link Clicked	-1.8919	0.259	-7.305	0.000	-2.400	-1.384
Last Notable Activity_Email Opened	-1.4254	0.089	-16.059	0.000	-1.599	-1.251
Last Notable Activity_Modified	-2.1058	0.093	-22.751	0.000	-2.287	-1.924
Last Notable Activity_Olark Chat Conversation	-2.7309	0.330	-8.274	0.000	-3.378	-2.084
Last Notable Activity_Page Visited on Website	-1.6840	0.203	-8.285	0.000	-2.082	-1.286



```
                                             Features   VIF
6                               Specialization_Others  2.17
3                              Lead Source_Olark Chat  1.87
10                     Last Notable Activity_Modified  1.60
1                         Total Time Spent on Website  1.57
9                  Last Notable Activity_Email Opened  1.40
7   What is your current occupation_Working Profes...  1.19
4                               Lead Source_Reference  1.15
0                                        Do Not Email  1.09
11     Last Notable Activity_Olark Chat Conversation  1.09
5                        Lead Source_Welingak Website  1.07
12     Last Notable Activity_Page Visited on Website  1.05
8           Last Notable Activity_Email Link Clicked  1.03
2                             Lead Origin_Lead Import  1.01
```

Terminologies Required

Before moving forward, it's essential to grasp several key concepts:

- Transforming **categorical columns** into **numerical format** is necessary because our algorithm exclusively operates on numerical data.

- **Feature Scaling** is performed to standardize the data, ensuring it is on the same scale.

- **Data Splitting** involves dividing the dataset into an 80% portion designated as the training data and a 20% portion designated as the test data. The model is trained on the training data and validated on the test data.

- The concept of a Confusion Matrix is also fundamental.

Confusion Matrix:

	Predicted NO	**Predicted YES**
Actual NO	True Negative	False Negative
Actual YES	False Positive	True Positive

In the realm of model evaluation, we rely on a set of fundamental metrics known as Confusion Metrics, which encompass the following key components:

1. **True Positive (TP):** These are the correct positive predictions.
2. **False Positive (FP):** Representing incorrect positive predictions.
3. **True Negative (TN):** These denote the accurate negative predictions.
4. **False Negative (FN):** Indicating erroneous negative predictions.

These metrics serve as the foundation for assessing model performance, and from them, we derive the following essential metrics:

1. **Accuracy:** Calculated as **(True Negative + True Positive) divided** by the **total**, where the total comprises TP, FN, FP, and FN. Accuracy provides an overall measure of model correctness.

2. **Sensitivity (SN):** Also referred to as recall **(REC) or true positive rate (TPR)**, it is computed as **True Positive divided** by t**he sum of True Positive and False Positive**. Sensitivity quantifies the model's ability to correctly identify positive instances, with the best sensitivity score being 1.0, and the worst being **0.0.**

3. **Specificity (SP):** Specificity, also known as the **true negative rate (TNR)**, is determined by the **division** of **True Negative** by the s**um of True Negative and False Negative.** It gauges the model's proficiency in correctly recognizing negative instances, with **1.0** representing the highest level of specificity and 0.0 being the lowest.

4. **Precision:** Precision signifies the number of **true positives divided by the total of true positives and false positives**. It characterizes the model's precision in identifying positive cases accurately.

5. **Recall:** Defined as **the number of true positives divided by the sum of true positives and false negatives.** It quantifies the model's ability to recapture positive instances correctly. True positives refer to data points correctly classified as positive by the model, while false negatives denote data points that the model incorrectly identifies as negative despite being positive.

Final Model Metrics

The following are the metrics that we got from the final model,

1. Train Data:

Confusion Matrix:

	Not Converted	Converted Leads
Not Converted	3141	764
Converted Leads	468	1978

- Accuracy: **0.81** or **81%**
- Sensitivity (True Positive Rate): **0.81** or **81%**
- Specificity (True Negative Rate): **0.80** or **80%**
- Precision: **0.79** or **79%**
- Recall : **0.70** or **70%**

2. Test Data:

Confusion Matrix:

	Not Converted	**Converted Leads**
Not Converted	1392	342
Converted Leads	206	783

- Accuracy: **0.80** or **80%**
- Sensitivity (True Positive Rate): **0.80** or **80%**
- Specificity (True Negative Rate): **0.80** or **80%**
- Precision: **0.70** or **70%**
- Recall : **0.79** or **79%**

The model demonstrates strong predictive capabilities for the Conversion Rate, providing valuable insights for the education company. This enables the identification of the most promising leads, often referred to as **"Hot Leads."**

Conclusion

Focus Areas:

- The company should give precedence to leads originating from **"Welingak Websites"** and **"Reference"** sources, as these are more likely to culminate in conversions.
- Leads who have invested a significant amount of time on the company's websites warrant special consideration, as their prolonged engagement signals a heightened potential for conversion.
- Targeting working professionals among the leads is a prudent strategy, given their elevated probability of conversion.
- Particular emphasis should be placed on leads acquired through **"Lead Import,"** as they exhibit a greater likelihood of conversion.
- It is advisable to channel efforts towards leads sourced through **"Olark Chat,"** as these leads boast a higher propensity for conversion.