Lead Scoring - Case Study of X Education Company

Sayali Sanjay Deshpande

GCP DS - C2

Background

- X Education specializes in offering online courses tailored for industry professionals.
- Numerous professionals seeking to enhance their skills frequently visit their website in search of a diverse range of upskilling courses.
- The company promotes its courses across multiple online platforms and search engines.
- Upon arriving at the website, visitors have the option to explore the available courses, submit course inquiry forms, or engage with instructional videos.

Background

- When individuals complete a form with their contact information, they are designated as "leads."
- Subsequently, upon acquiring these leads, members of the sales team initiate contact through phone calls, emails, and other methods.
- During this outreach process, some leads transition into customers, while the majority do not.
- The average lead conversion rate at X Education stands at approximately 30%.

Problem Statement

- X Education generates numerous leads, yet its conversion rate remains notably low.

- In a bid to enhance efficiency, the company aims to pinpoint the most promising leads, termed as 'Hot Leads'.

- By successfully discerning this subset of leads, the conversion rate is expected to surge as the sales team can then prioritize engaging with these prospects rather than reaching out to every lead indiscriminately.

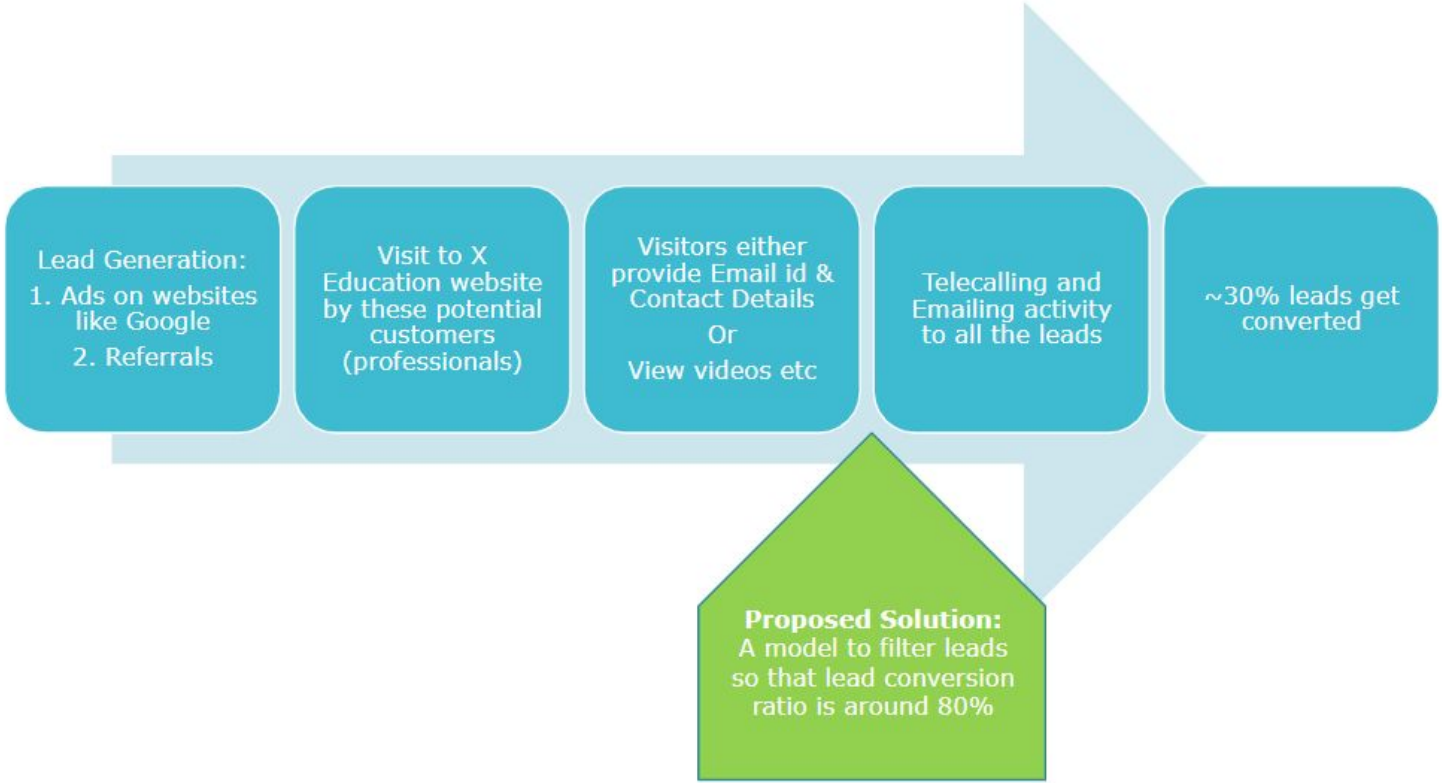
Problem Statement

- We will assist them in identifying the most viable leads, specifically those with the highest likelihood of becoming paying customers.

- Our task involves developing a model to assign a lead score to each prospect, ensuring that those with higher scores are more likely to convert.

- The CEO has indicated a rough target lead conversion rate of 80%.

A Quick Look at Lead Conversion Process



Proposed Solution:

- **Leads Clustering**

We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

- **Focus on Communication**

Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

- **Increase conversion**

Since we focussed on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.

Implementation

- **Data Gathering**

Loading & Observing the past data provided by the Company

- **Data Cleaning**

Duplicate removal, null value treatment, unnecessary column elimination, etc.

- **Performing EDA**

Univariate, Bivariate, and Heatmap for numerical and categorical columns

Implementation

- **Data Preparation**

Outlier Treatment, Feature-Standardization

- **Model Building**

Performing pre-requisites for RFE and Logistic Regression

- **Model Improvement**

Reduction of columns and Model re-building

Implementation

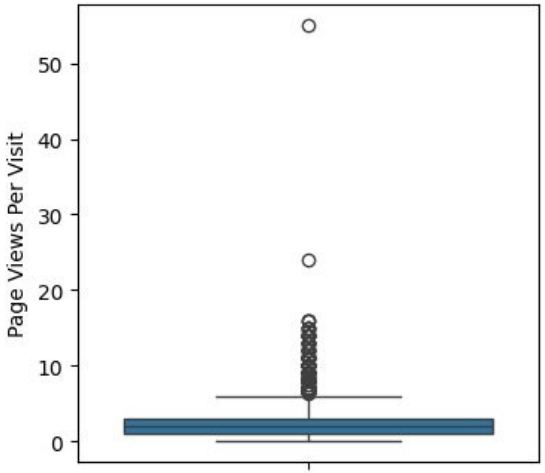
- **Final Model**

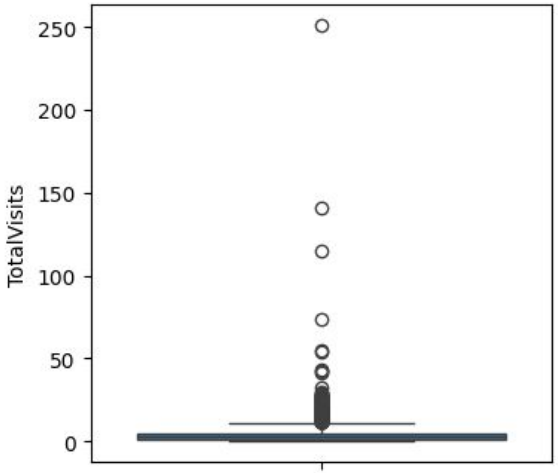
Final Model Analysis and performance on Test Data

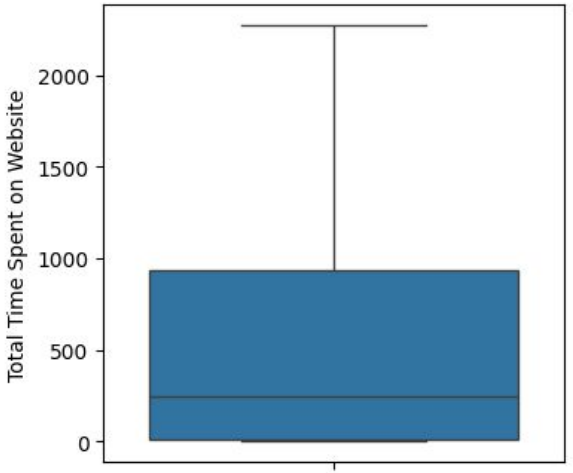
- **Verifying with PCA**

Verifying our Final Model Accuracy etc. with model built with PCA

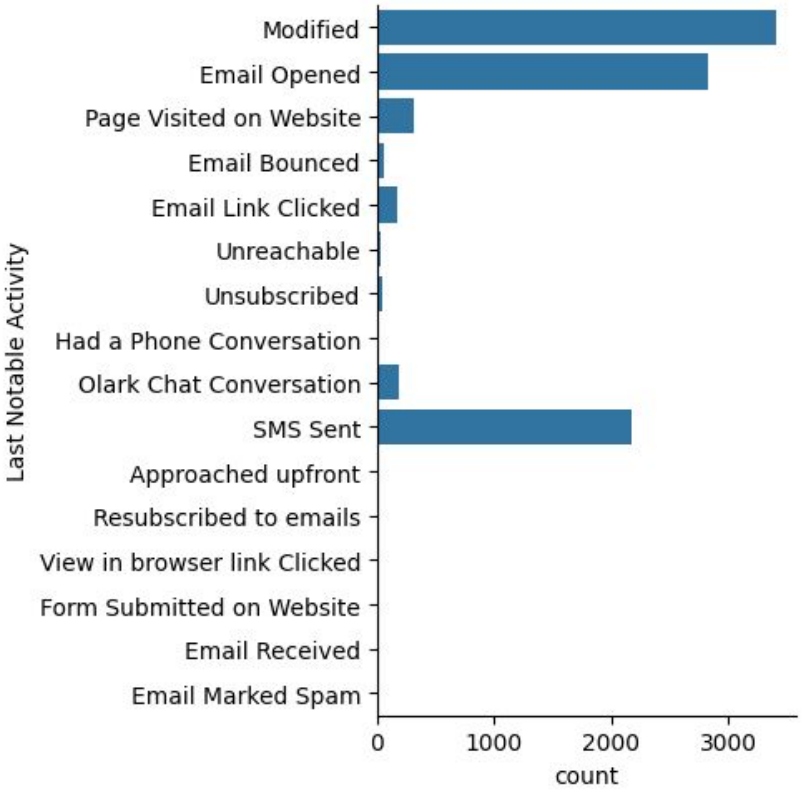
Univariate Analysis of Numerical Columns

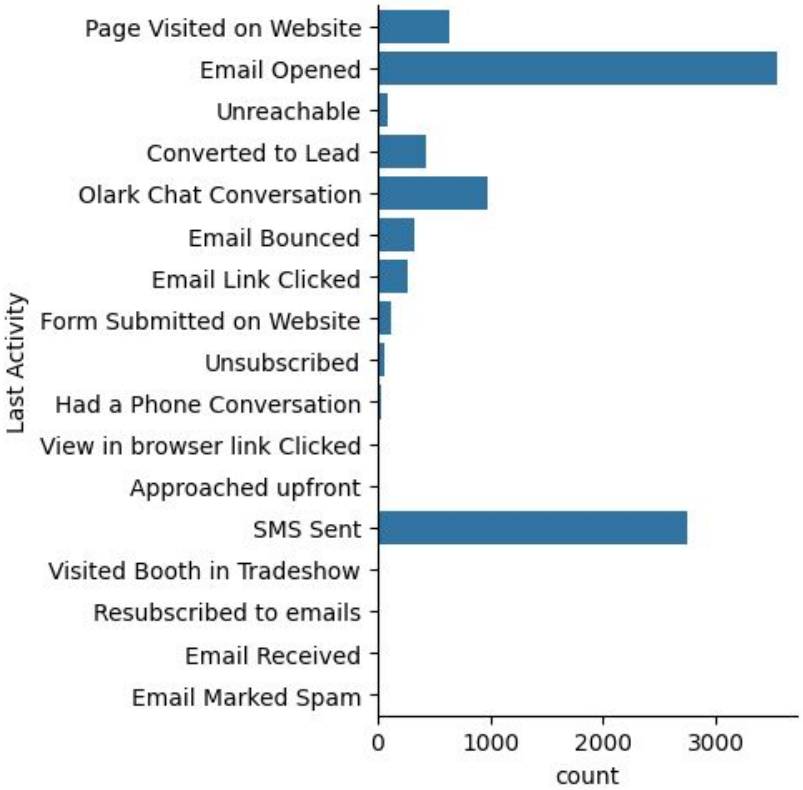




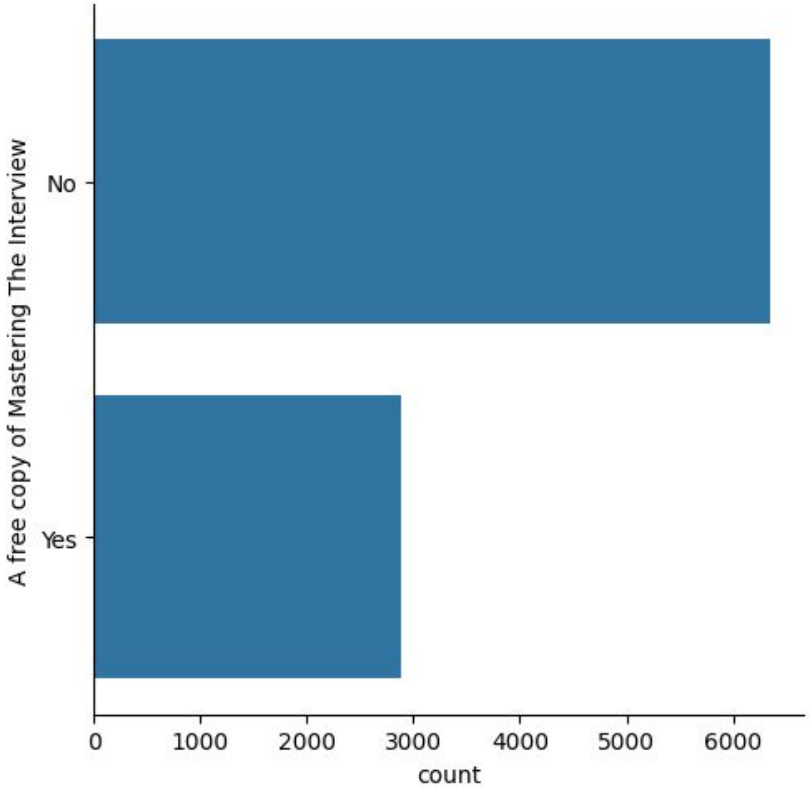


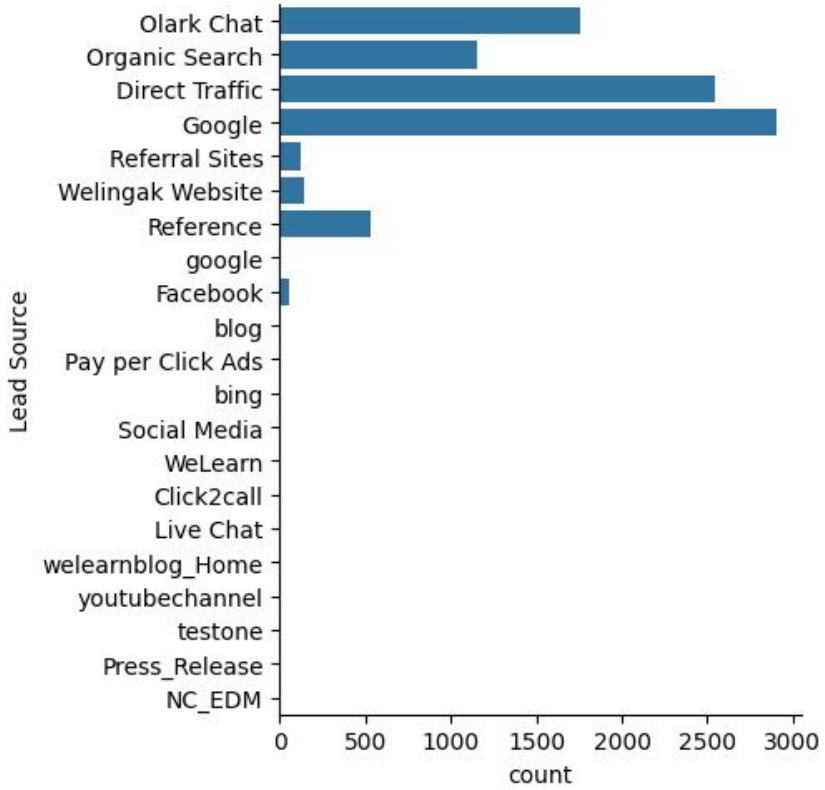
Univariate Analysis of Numerical Columns



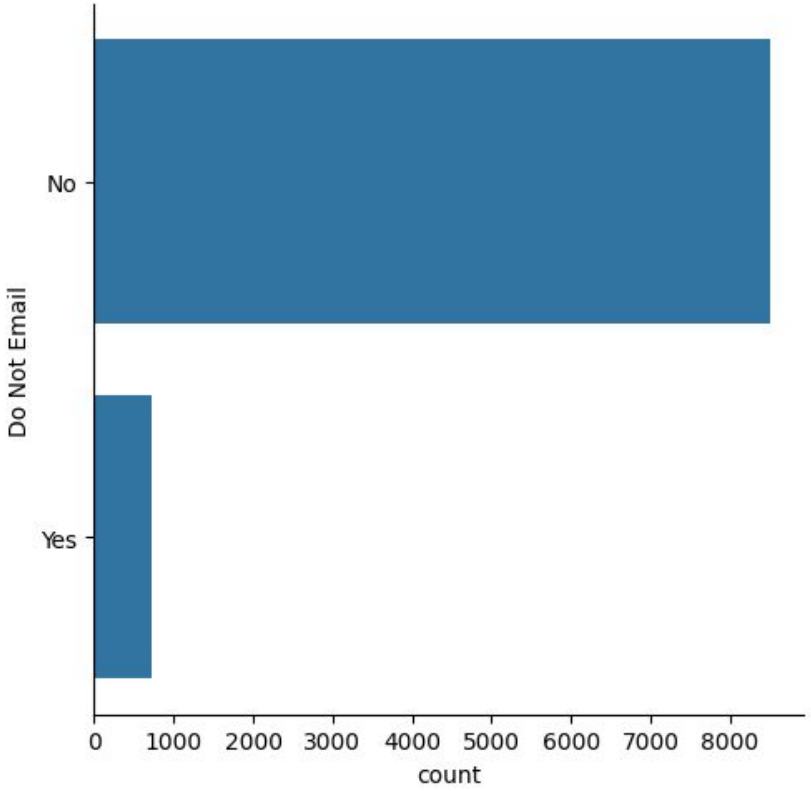


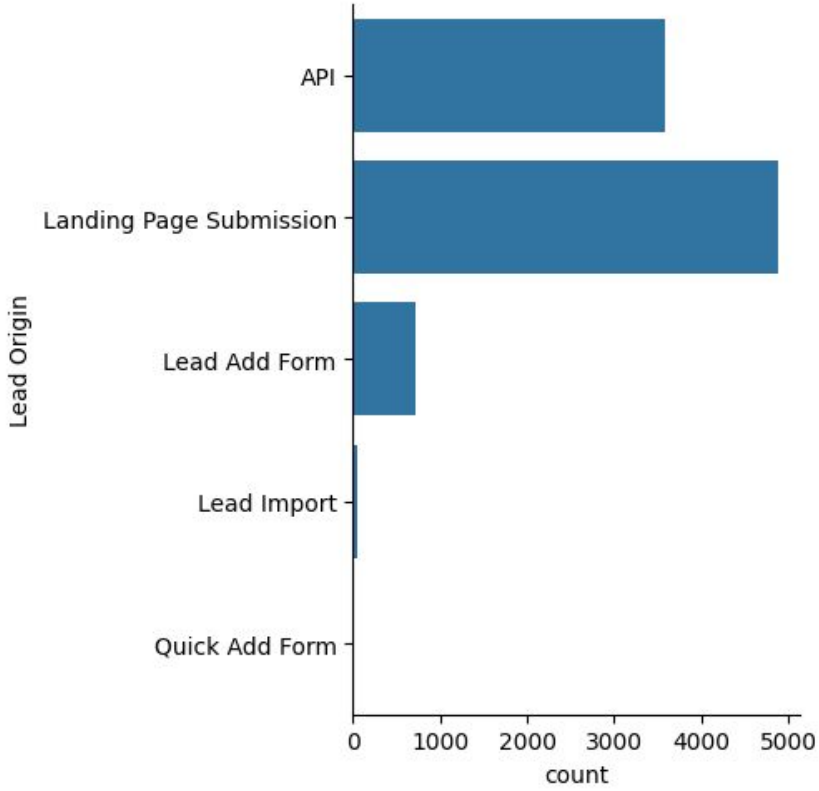
Univariate Analysis of Numerical Columns



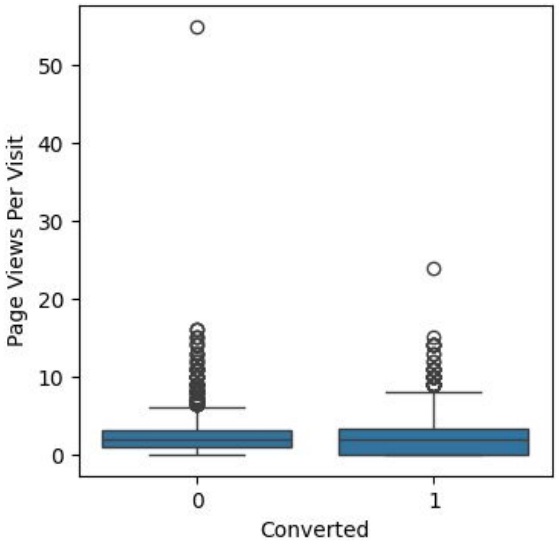


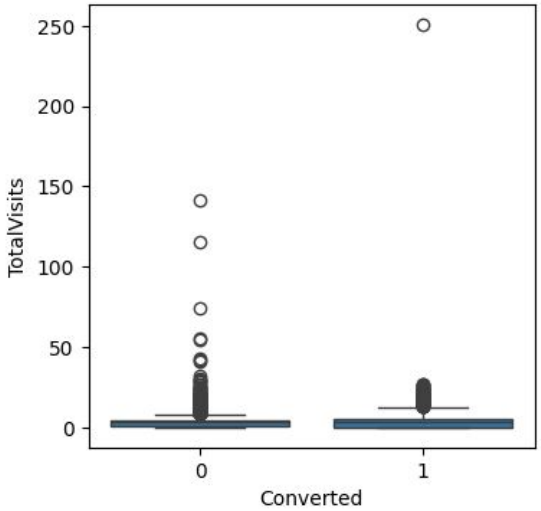
Univariate Analysis of Numerical Columns

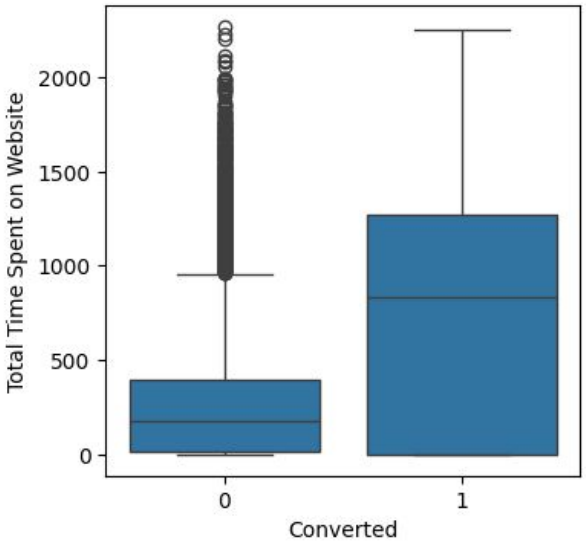




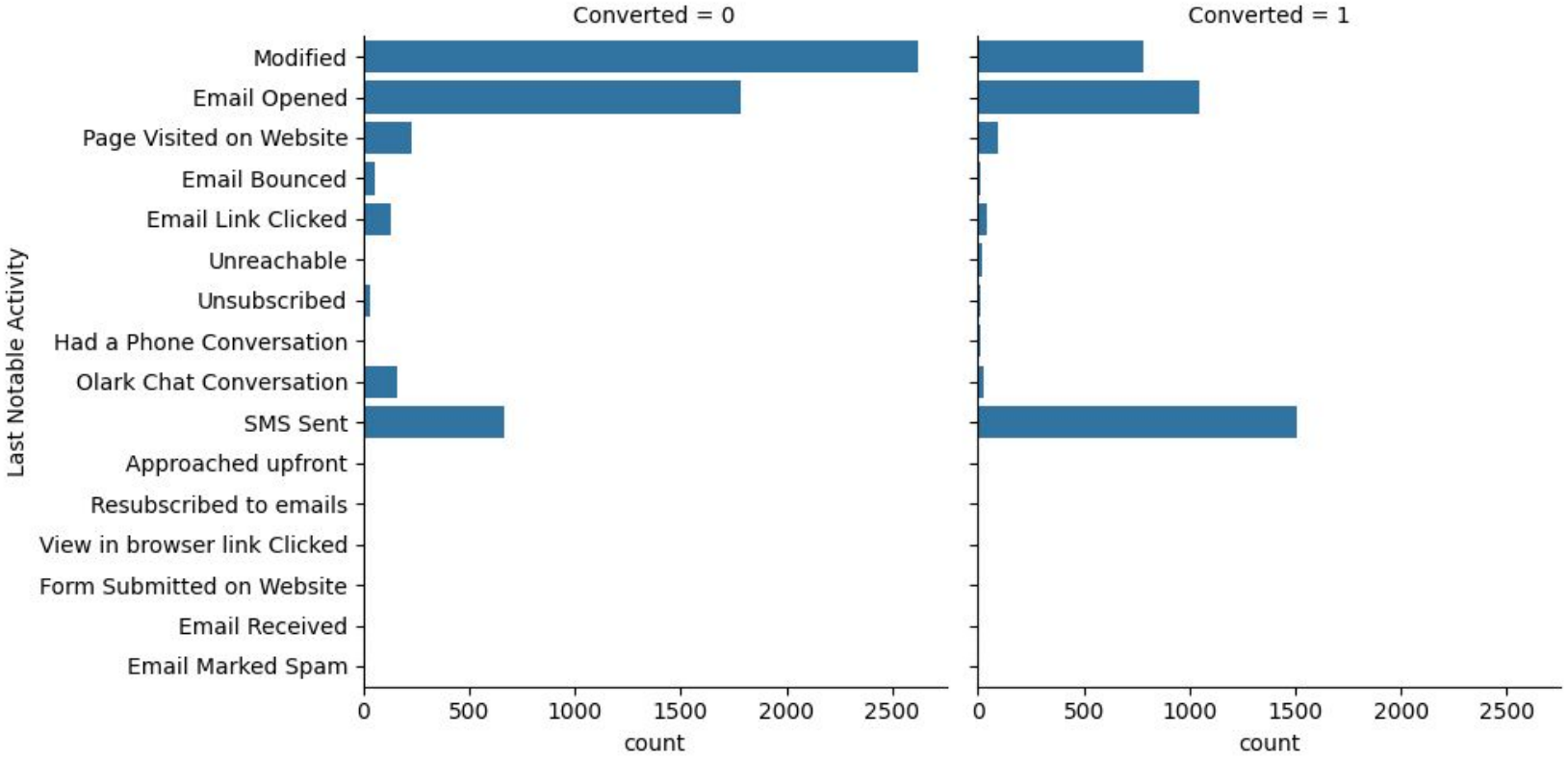
Bivariate Analysis of Numerical Columns



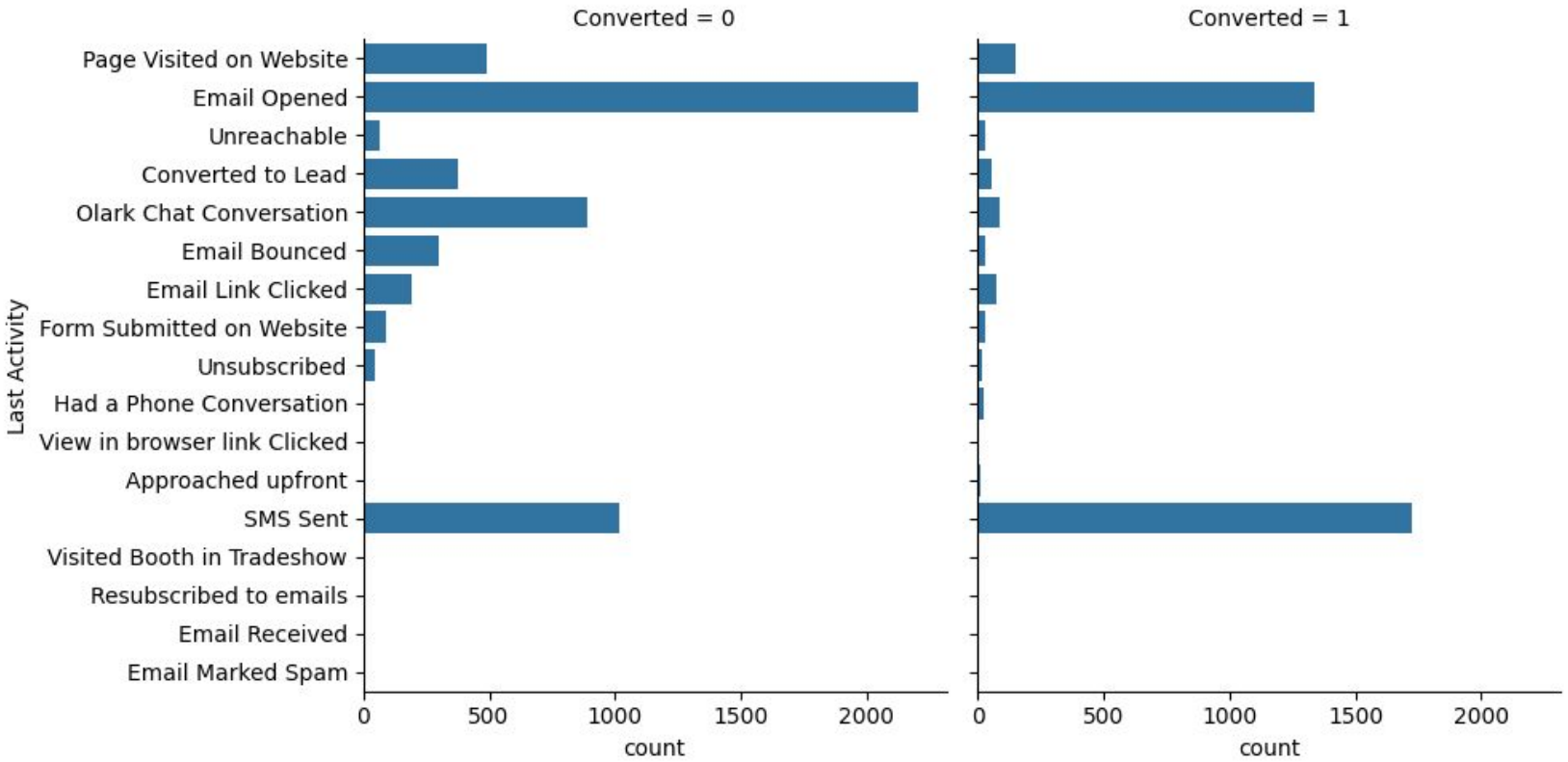




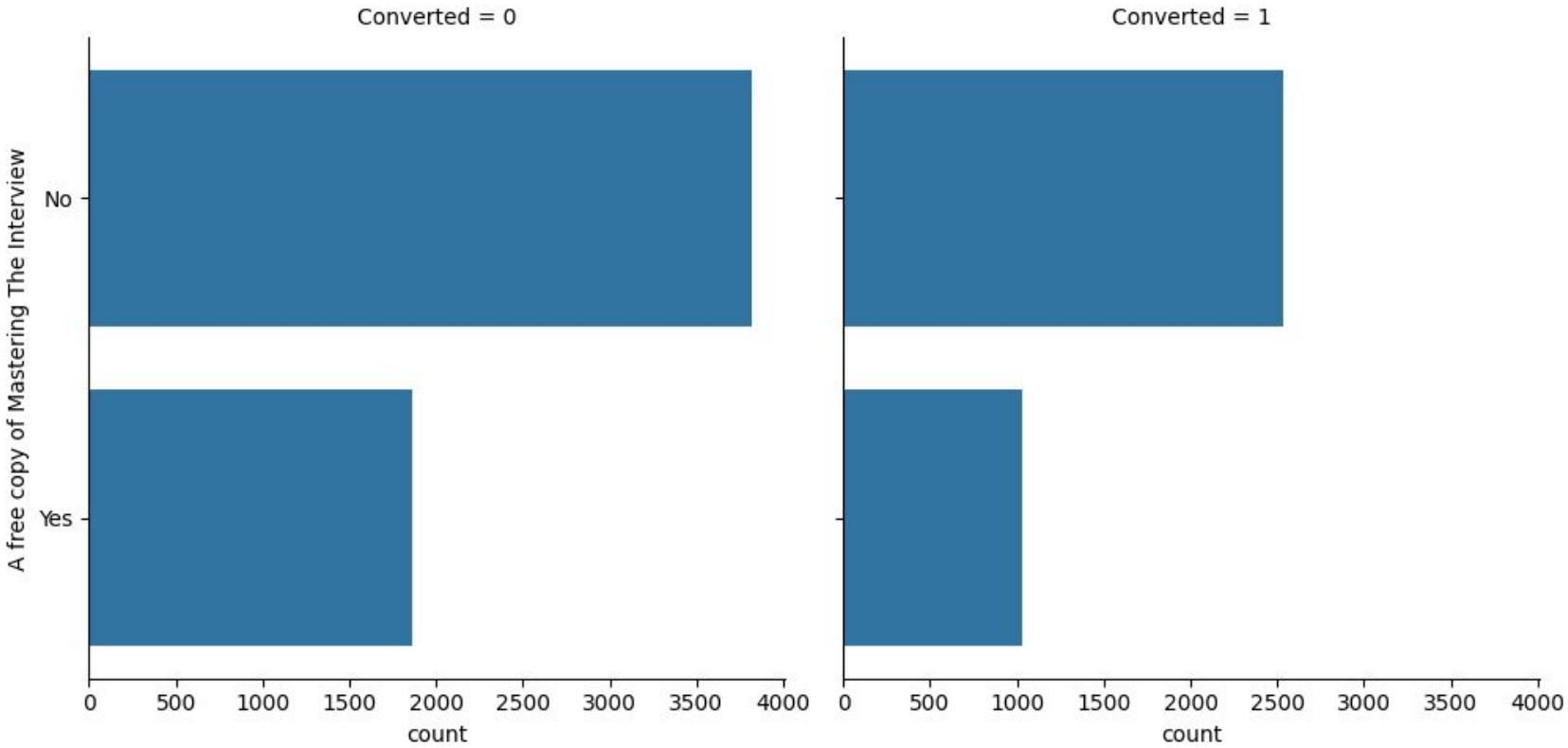
Bivariate Analysis of Categorical Columns



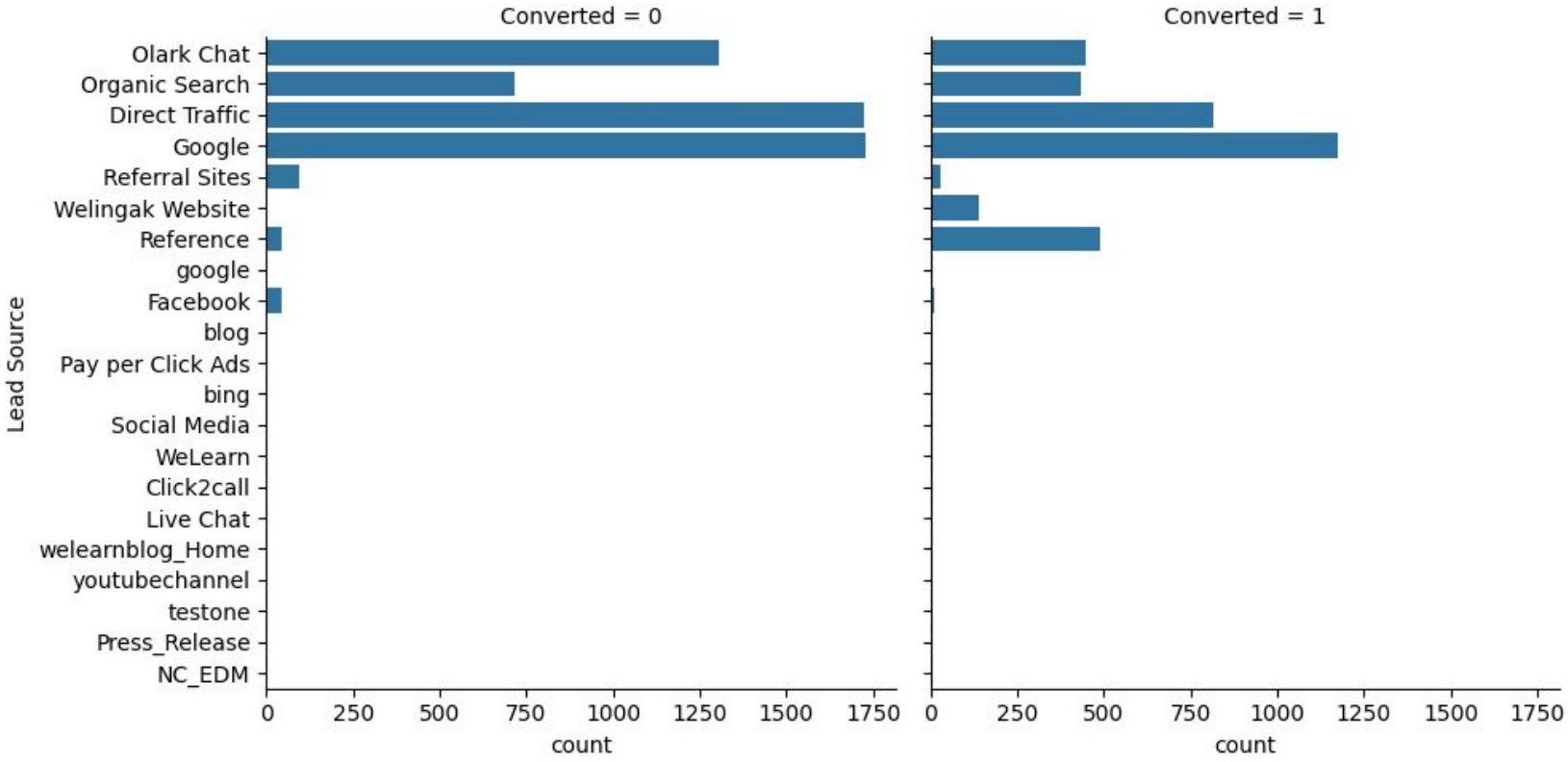
Bivariate Analysis of Categorical Columns



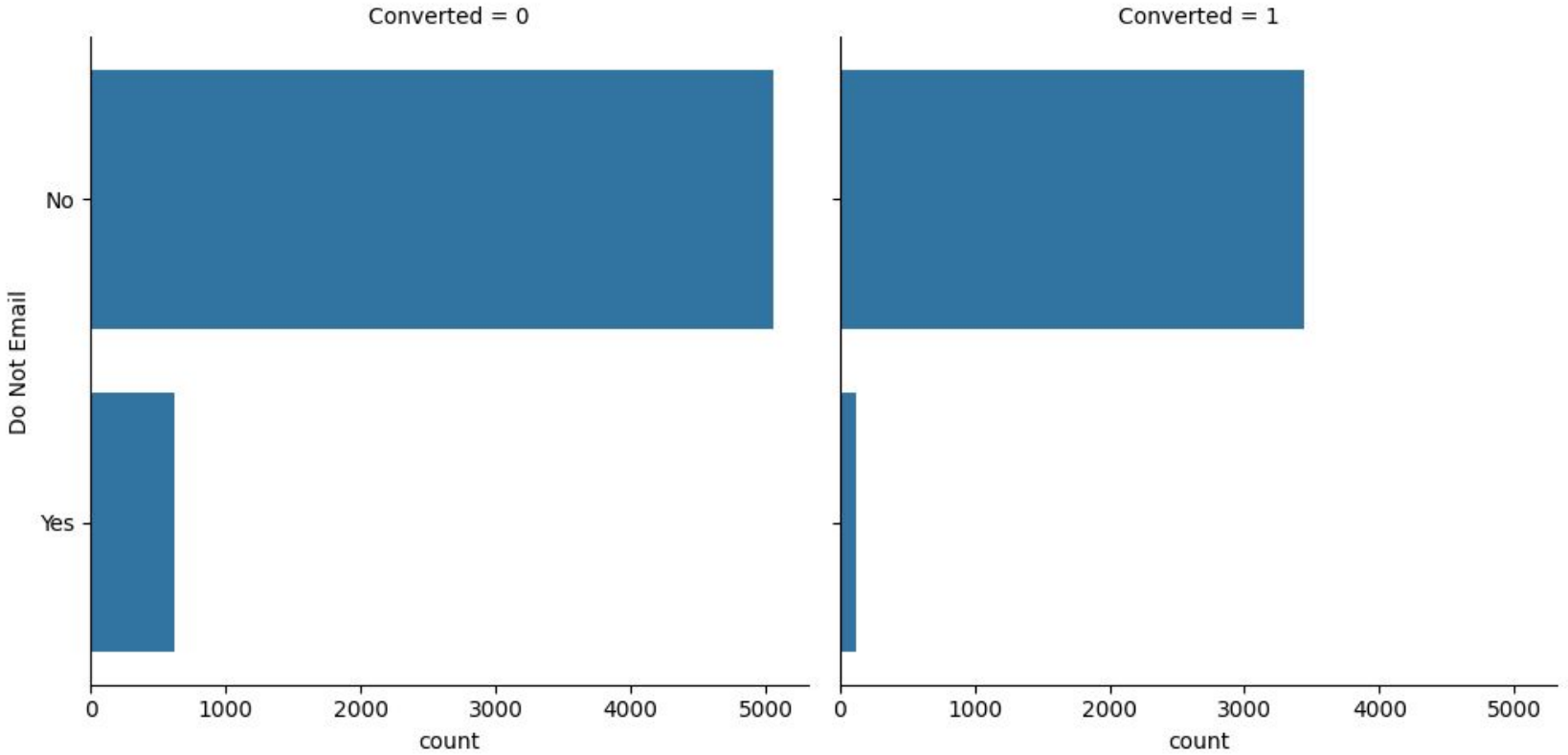
Bivariate Analysis of Categorical Columns



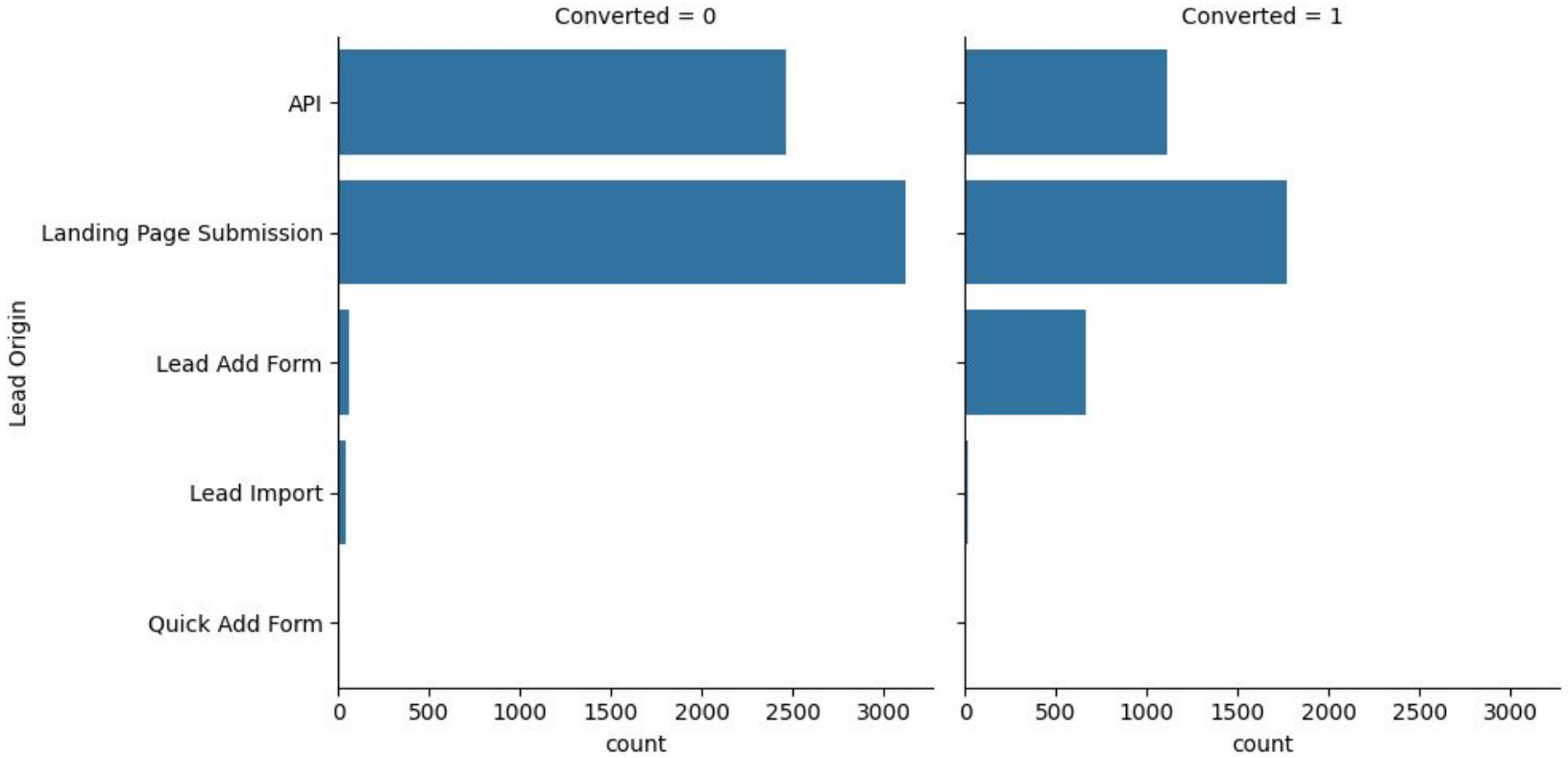
Bivariate Analysis of Categorical Columns



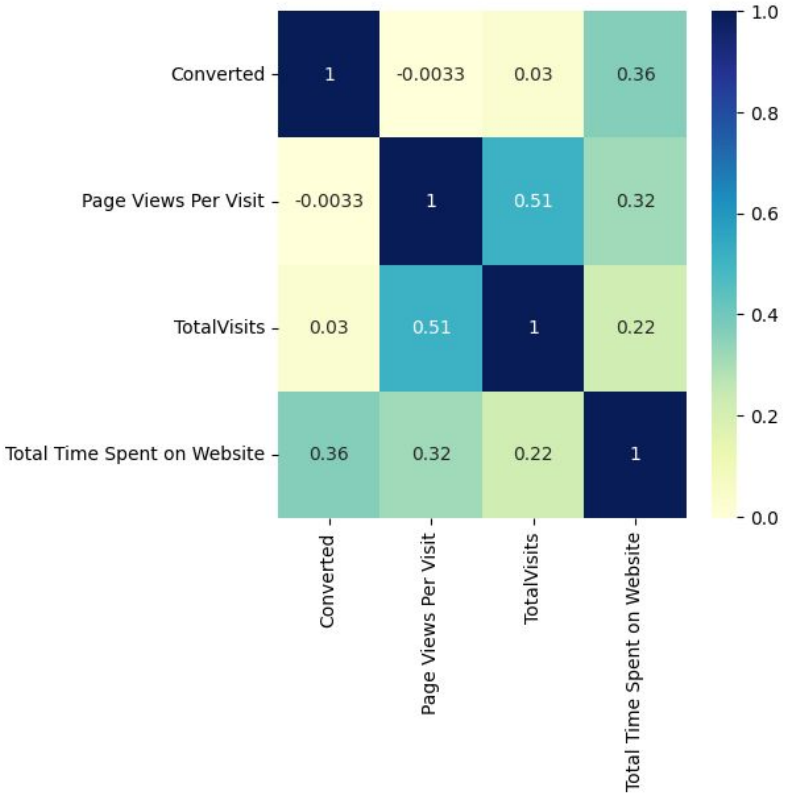
Bivariate Analysis of Categorical Columns



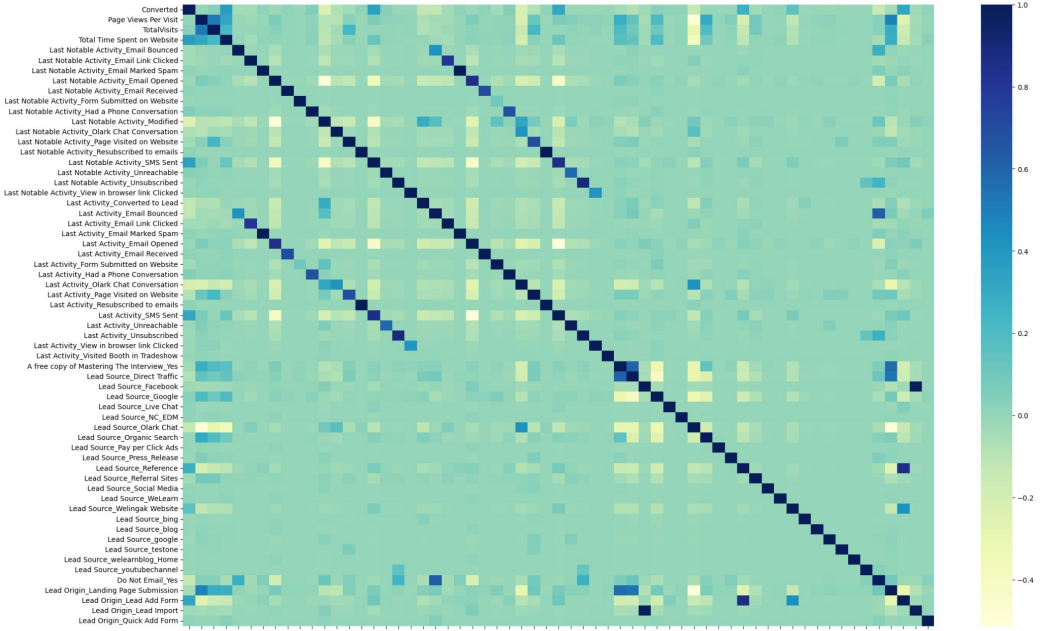
Bivariate Analysis of Categorical Columns



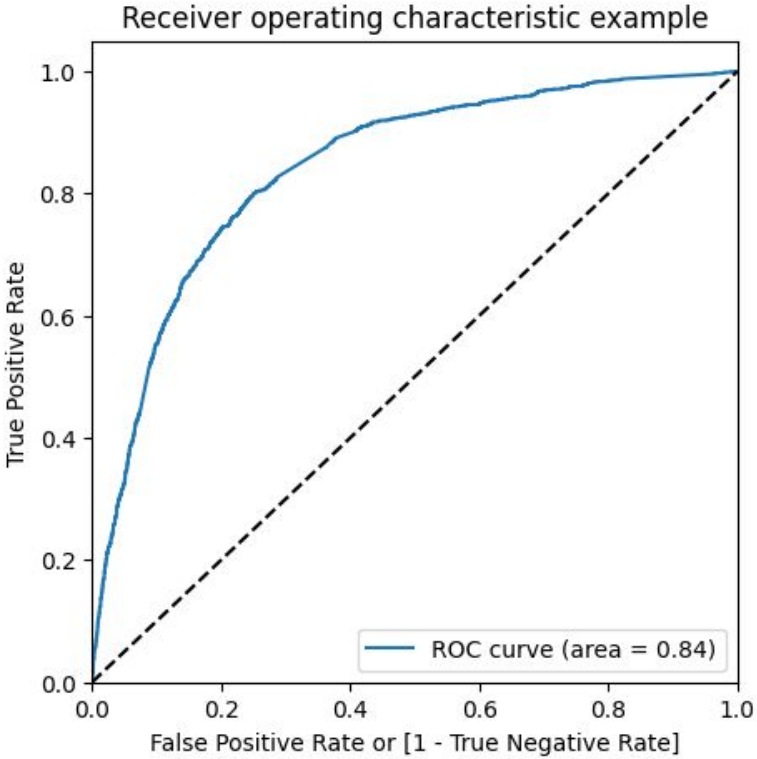
Correlation Analysis - Heat Map



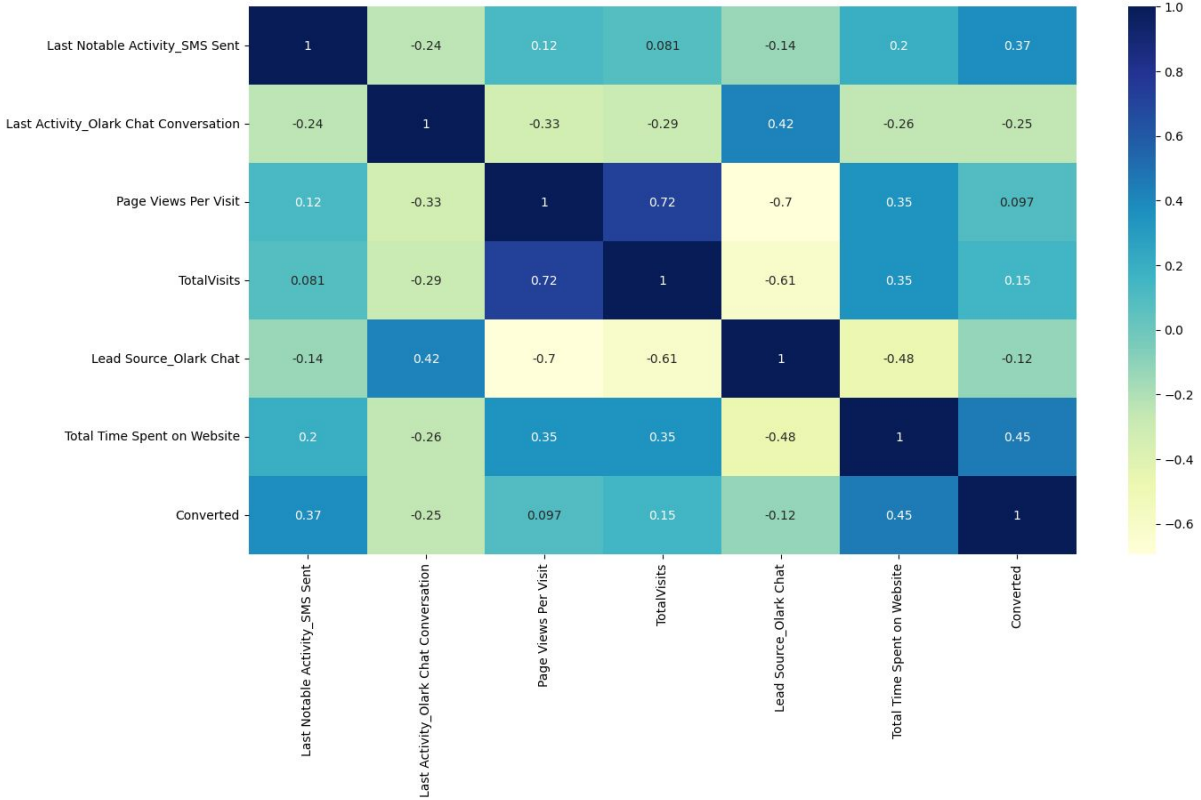
Correlation Analysis - Heat Map



Linear Regression – Final Model



Correlation Analysis – Heat Map – Linear Regression – Final Model



Inference

- Overall accuracy on Test set: 0.786
- Sensitivity of our logistic regression model: 0.733
- Specificity of our logistic regression model: 0.823

Business Insights

Top 3 variables in model, that contribute towards lead conversion are:

- Total Time Spent on Website
- Last Notable Activity_SMS Sent
- Total Visits

Business Insights

Top 3 variables in the model, that should be focused are:

- Last Activity_SMS Sent (positively impacting)
- Last Activity_Olark Chat Conversation (negatively impacting)
- Lead Source_Olark Chat (negatively impacting)

Conclusion - LR Model

- Our Logistic Regression Model performs satisfactorily and demonstrates a level of accuracy comparable to the model obtained through PCA, achieving an accuracy of 78.6% on the Test Set, along with 73.3% Sensitivity and 82.3% Specificity.

- By adjusting the cut-off value, we can manipulate these parameters to predict Hot leads based on factors such as the availability of additional resources, and conversely, anticipate other scenarios.

Recommendations

The X Education Company should prioritize the following key areas to enhance the overall conversion rate:

1. Enhance user engagement on their website, as it correlates with higher conversion rates.

2. Boost the frequency of sending SMS notifications, as this has shown to increase conversions.

3. Increase total visits through advertising initiatives, as this contributes to higher conversion rates.

4. Enhance the performance of the Olark Chat service, as its current state negatively impacts conversion rates.