LEAD SCORING CASE STUDY

- GOURAV BAGGA
- ATISH BANERJEE
- NAVYA REDDY

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- 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.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVES

1.Build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead scores have a higher conversion chance, and the customers with lower lead scores have a lower conversion chance.

2.The CEO has given a ballpark of the target lead conversion rate to be around 80%.

3.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.

4.Some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.

Data understanding

1.The dataset consists of two files Leads and Leads Data Dictionary
2.Leads dataset consists of 9240 rows and 37 columns.
3.This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
4.The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
5.The data dictionary has been provided to learn more about the data set.
6.Another thing that we also need to check out for is the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value

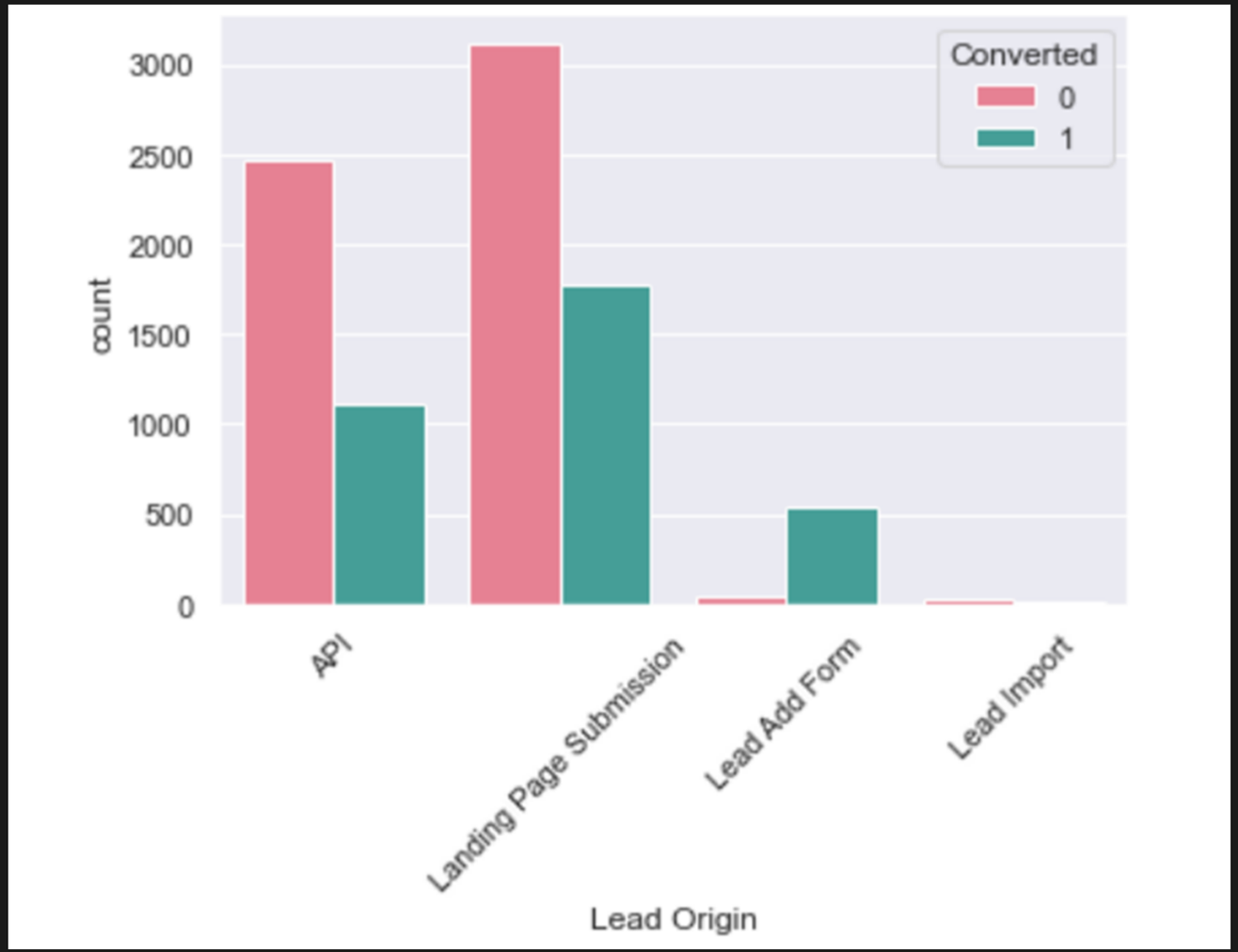
Data Cleaning

1. Handling the 'Select' level that is present in many of the categorical variables: Upon examination, it appears that numerous columns contain the value 'Select'. This occurrence likely signifies that the customer opted not to choose any option from the provided list, resulting in the display of 'Select'. Functionally, these 'Select' values can be considered equivalent to NULL. Hence, it is advisable to convert these values to NULL in order to address this issue.
2. After analyzing the dataset, we noticed that certain columns exhibit a substantial percentage of missing values. Considering this observation, a suitable approach would be to remove columns from the dataset if their missing value count exceeds 40%.
3. Dropped the columns with missing values greater than or equal to 40%.
4. The specialization column had 37% missing values, so a new category called others is created. This category will encompass all the leads who fall into any of these scenarios, allowing us to account for and include these individuals in our analysis.
5. The tags column has the most values as 'Will revert after reading the email', so we imputed missing values in this column with this value.
6. Due to the high skewness in the column, “What matters most to you in choosing a course”, it was advisable to remove it from the dataset,
7. In the columns “What is your current occupation” and “Country” the majority of values in the column are 'Unemployed‘ and ‘India’ respectively, so we can fill the missing values in this column with the same value.
8. After performing data cleaning, it has been determined that approximately 98% of the rows have been preserved in the dataset.

Exploratory Data Analysis

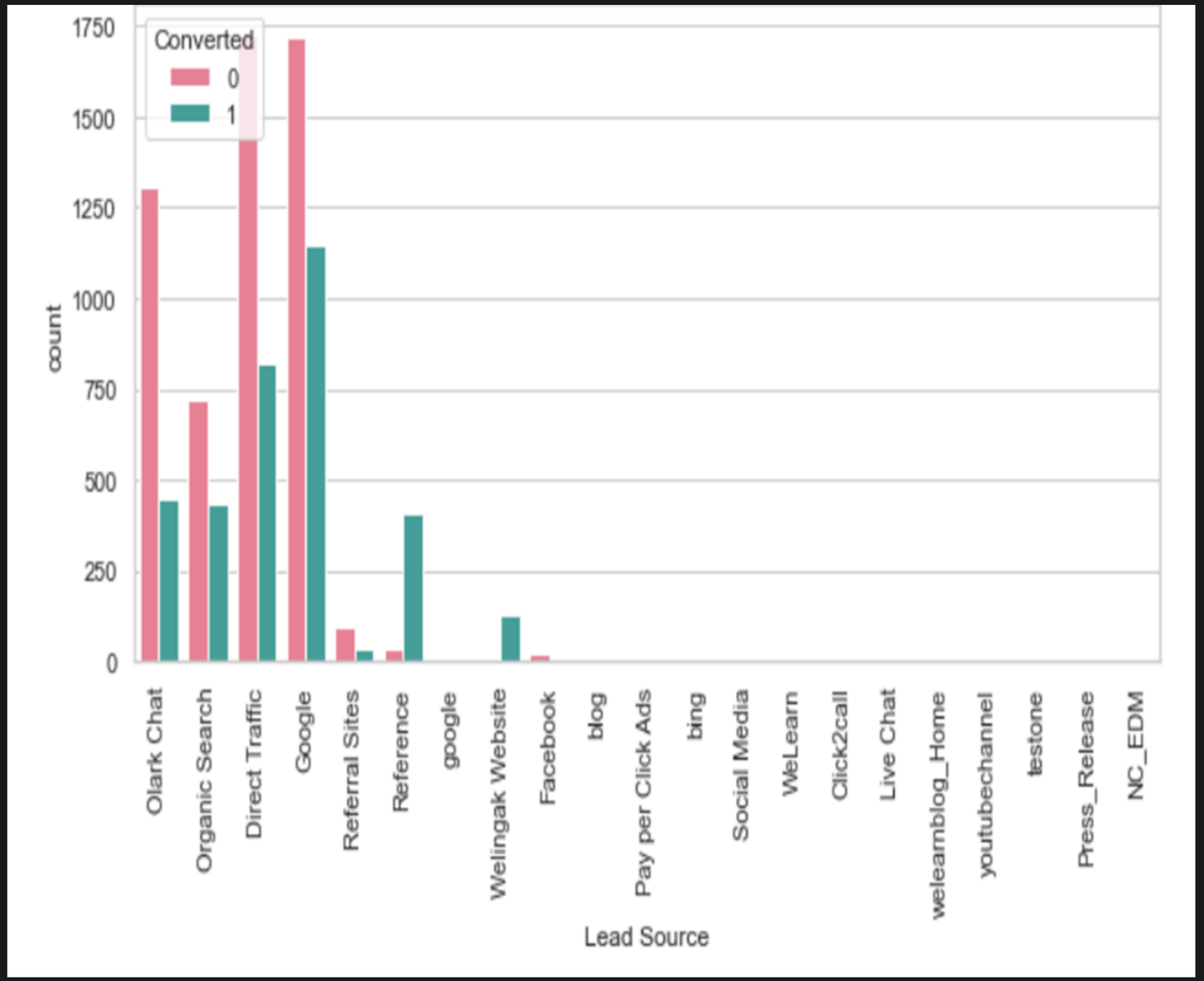
LEAD CONVERSION RATE OF LEAD ORIGIN

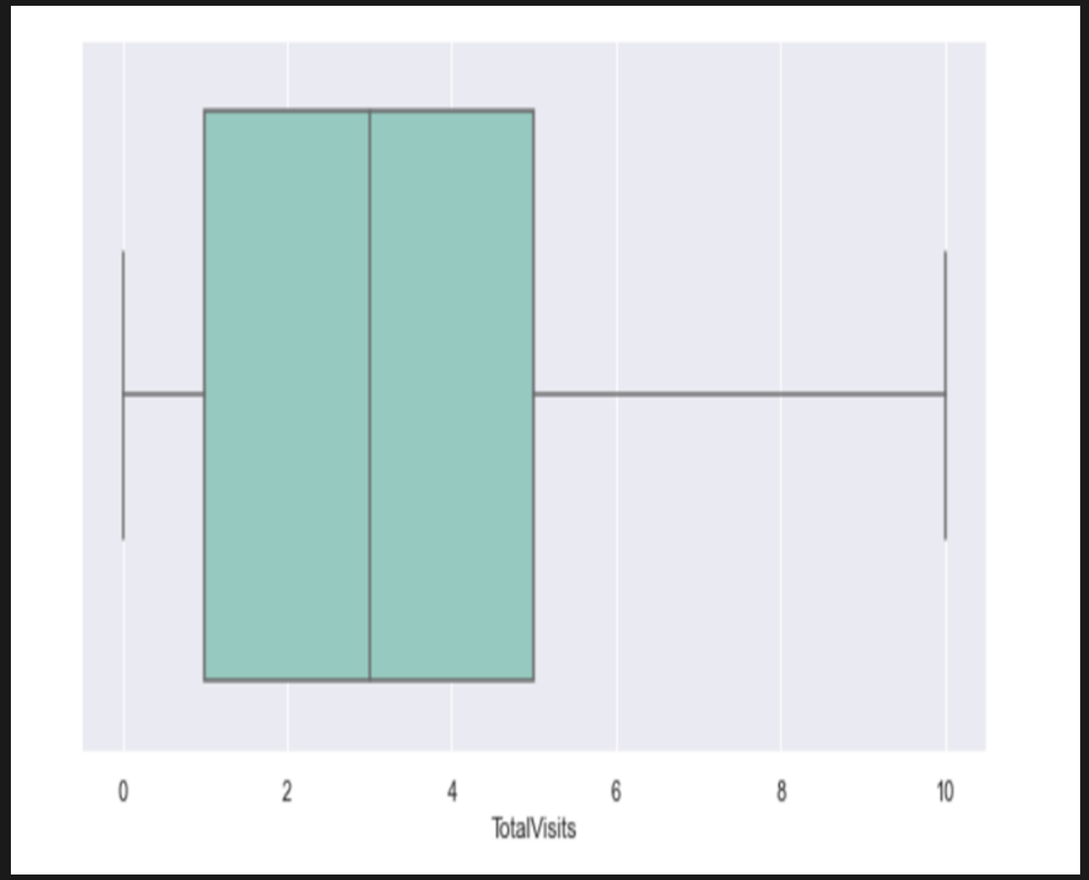
1. API AND LANDING PAGE SUBMISSION HAVE 30-35% CONVERSION RATE BUT THE COUNT OF LEAD ORIGINATING FROM THEM ARE CONSIDERABLE.
2. THE LEAD ADD FORM HAS AN EXCEPTIONAL CONVERSION RATE OF OVER 90%, WHICH MEANS THAT THE MAJORITY OF PEOPLE WHO SUBMIT THEIR INFORMATION THROUGH THIS FORM BECOME CUSTOMERS. HOWEVER, IT'S IMPORTANT TO NOTE THAT THE OVERALL NUMBER OF LEADS GENERATED FROM THE LEAD ADD FORM IS NOT VERY LARGE.
3. THE NUMBER OF LEADS IMPORTED IS QUITE LOW.

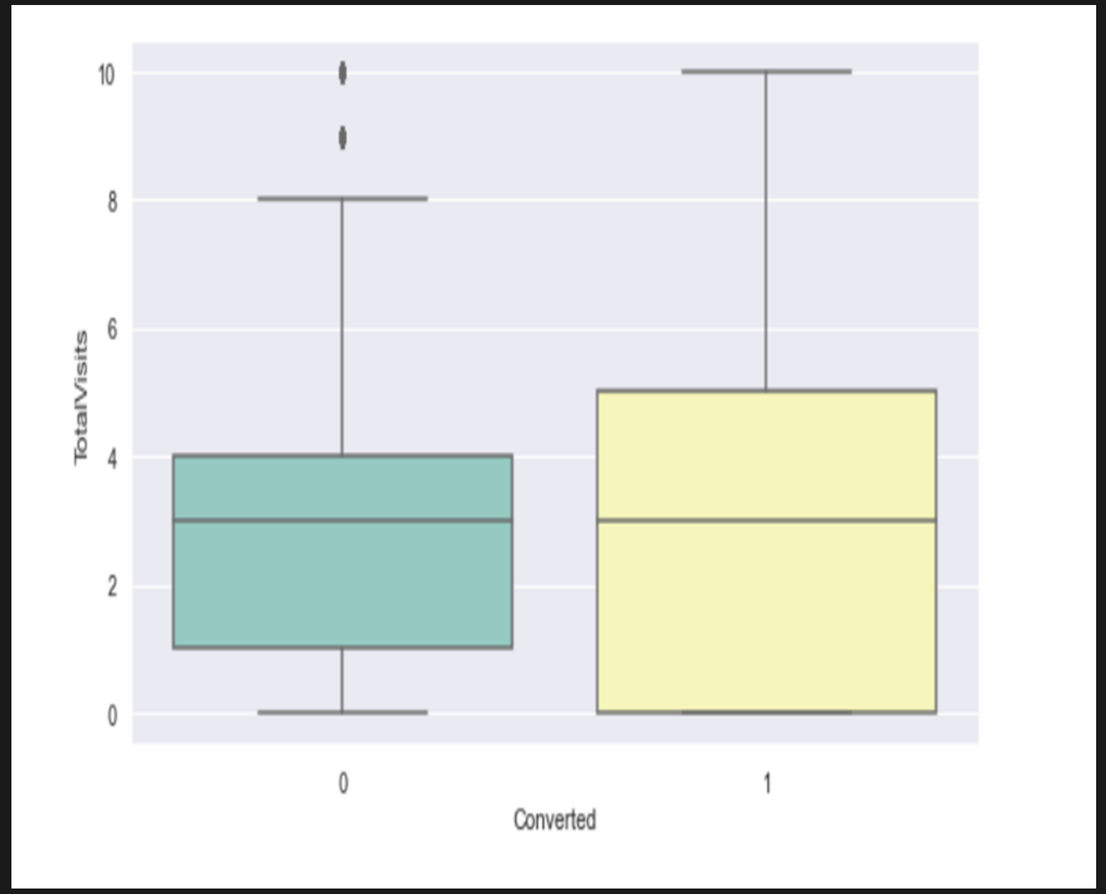


LEAD CONVERSION RATE OF LEAD SOURCE

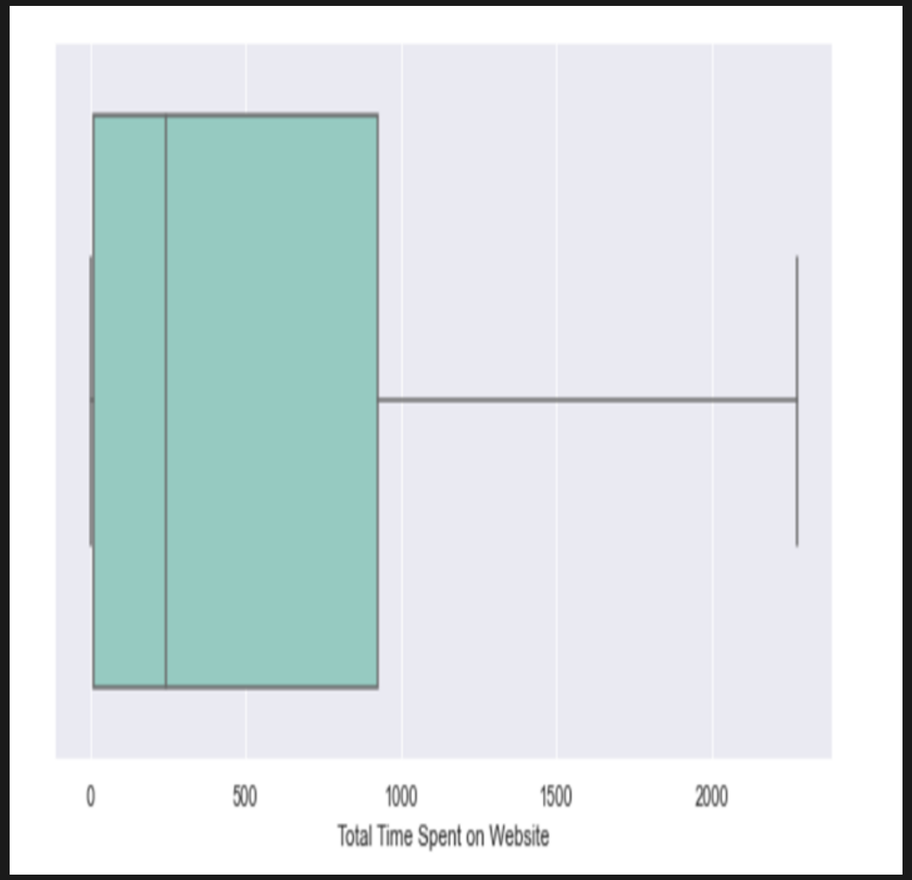
1. THE MAJORITY OF LEADS ARE GENERATED THROUGH GOOGLE AND DIRECT TRAFFIC, INDICATING THAT THESE SOURCES CONTRIBUTE THE HIGHEST NUMBER OF LEADS.
2. THE CONVERSION RATE FOR LEADS ORIGINATING FROM REFERENCES AND THROUGH THE WELINGAK WEBSITE IS NOTABLY HIGH.
3. TO ENHANCE THE OVERALL RATE AT WHICH LEADS ARE CONVERTED INTO CUSTOMERS, IT IS CRUCIAL TO PRIORITIZE IMPROVING LEAD CONVERSION RATES FOR OLARK CHAT, ORGANIC SEARCH, DIRECT TRAFFIC, AND GOOGLE LEADS. ADDITIONALLY, EFFORTS SHOULD BE DEDICATED TO GENERATING A HIGHER NUMBER OF LEADS FROM REFERENCE SOURCES AND THE WELINGAK WEBSITE.

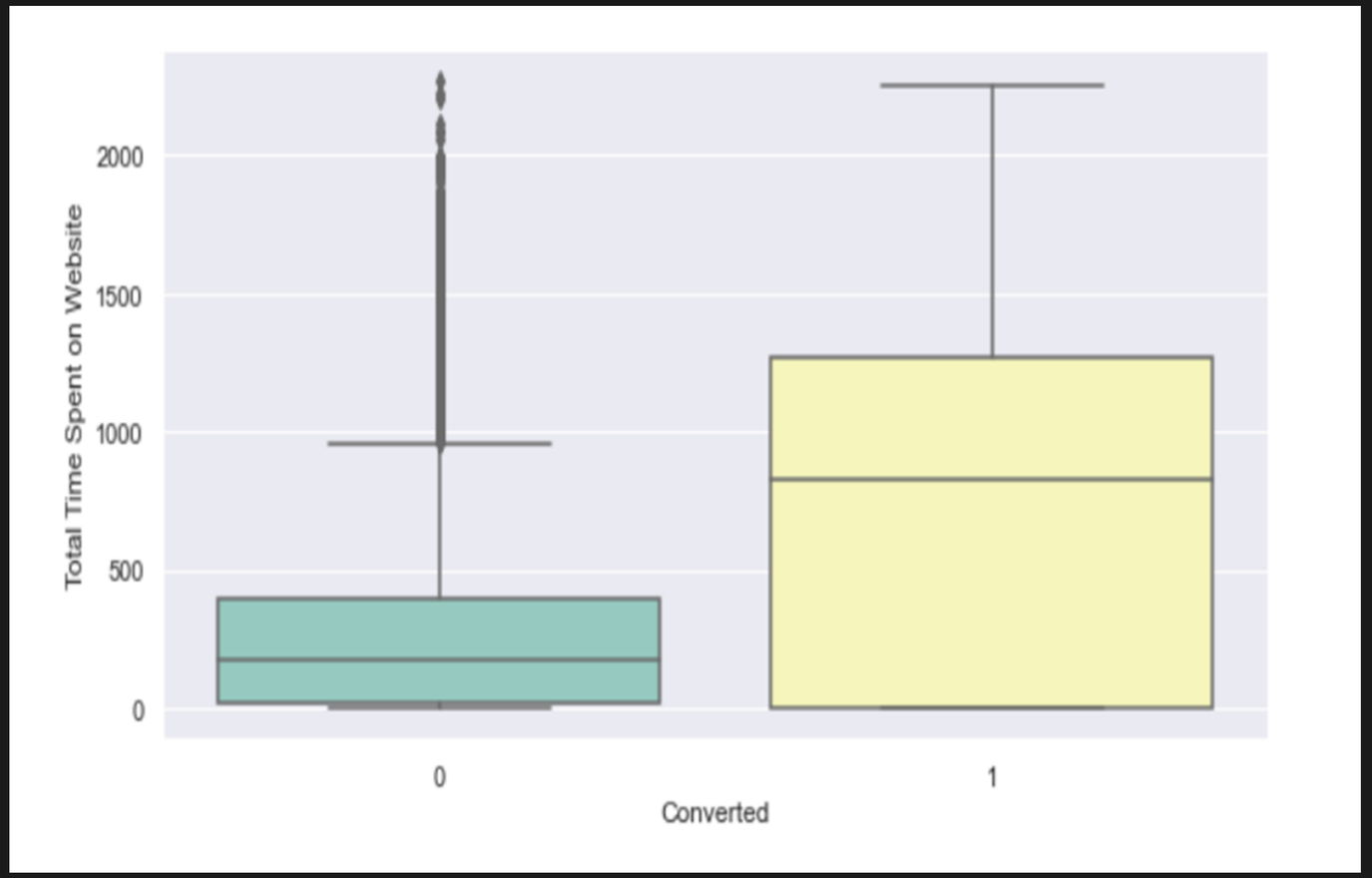




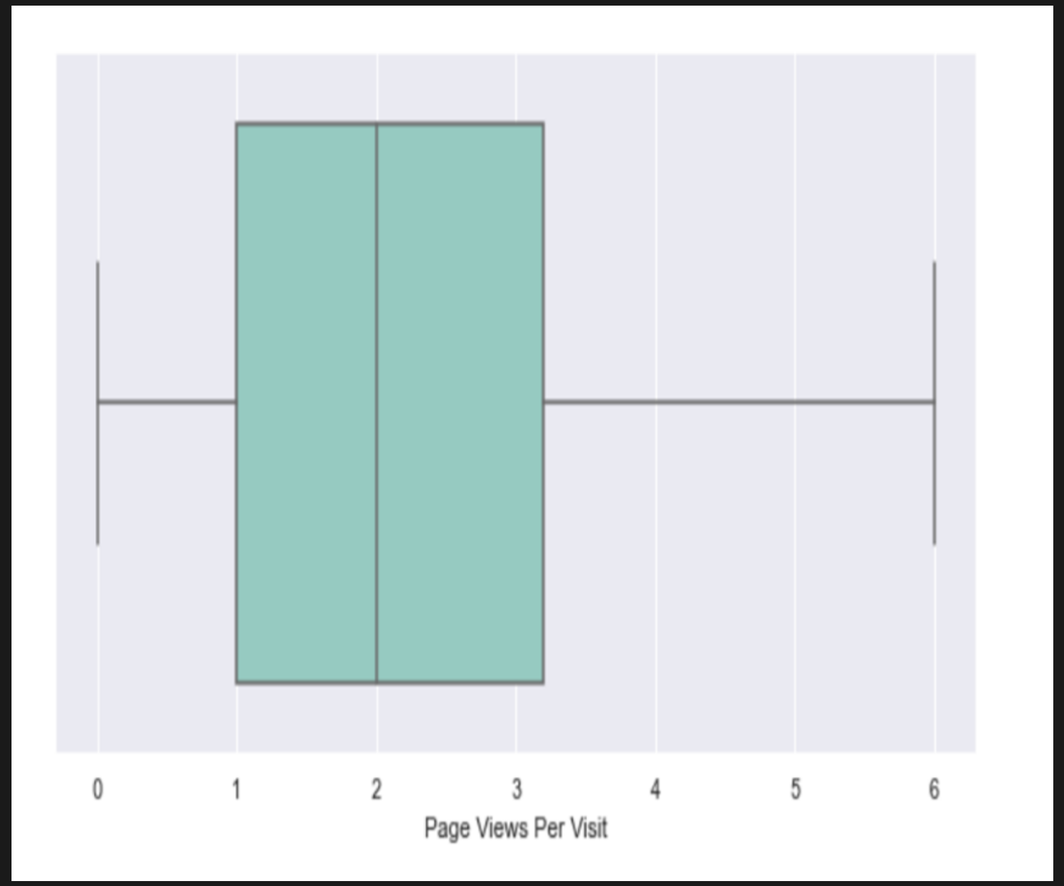


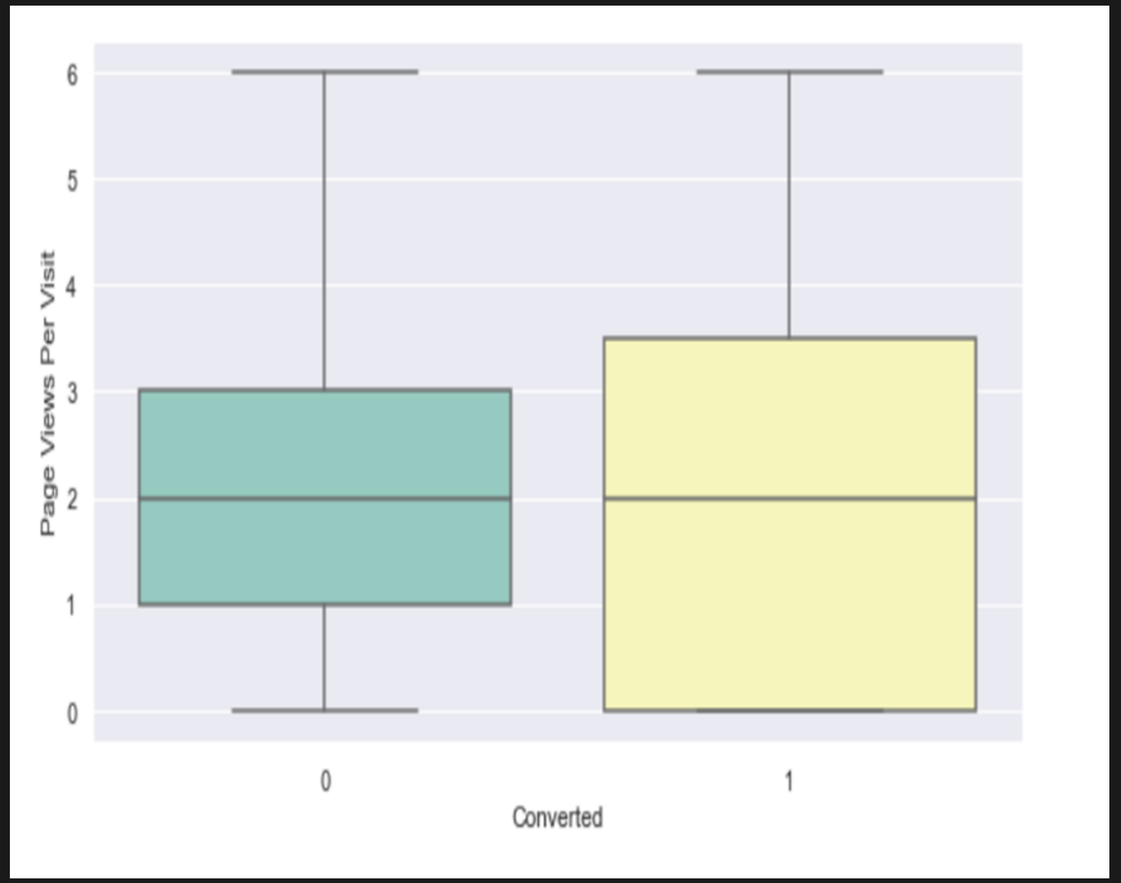
- THE MEDIAN VALUE REMAINS CONSISTENT ACROSS BOTH CONVERTED AND NON-CONVERTED LEADS.
- DRAWING ANY MEANINGFUL CONCLUSIONS BASED ON THE 'TOTAL VISITS' PARAMETER IS NOT POSSIBLE DUE TO ITS LIMITED INFORMATIVE VALUE.



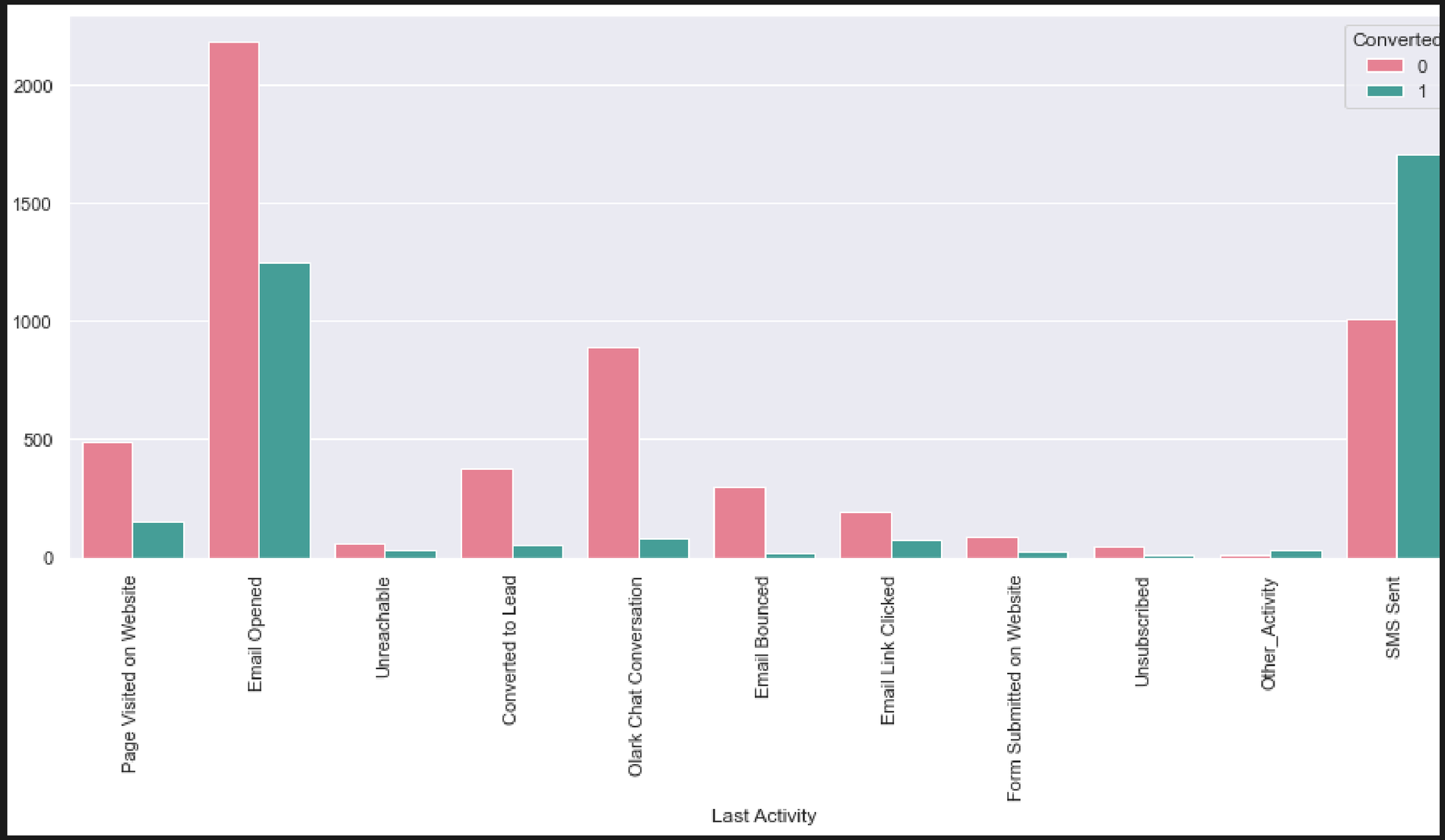


- THERE IS A HIGHER LIKELIHOOD OF LEAD CONVERSION WHEN THERE IS A CORRELATION BETWEEN INCREASED TIME SPENT ON THE WEBSITE AND THE CONVERSION RATE.
- IN ORDER TO ENHANCE USER ENGAGEMENT AND ENCOURAGE LEADS TO INVEST MORE TIME ON THE WEBSITE, IT IS IMPERATIVE TO MAKE THE WEBSITE MORE CAPTIVATING AND INTERACTIVE.

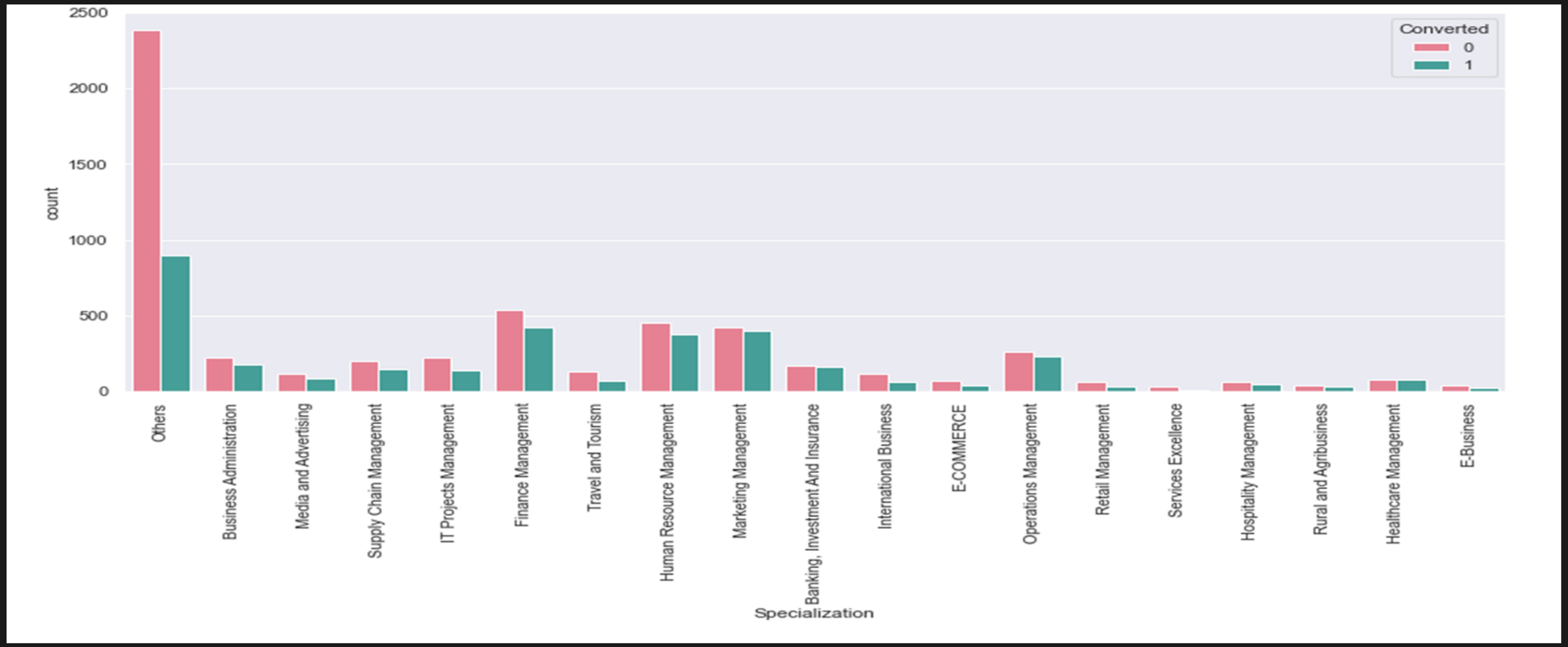




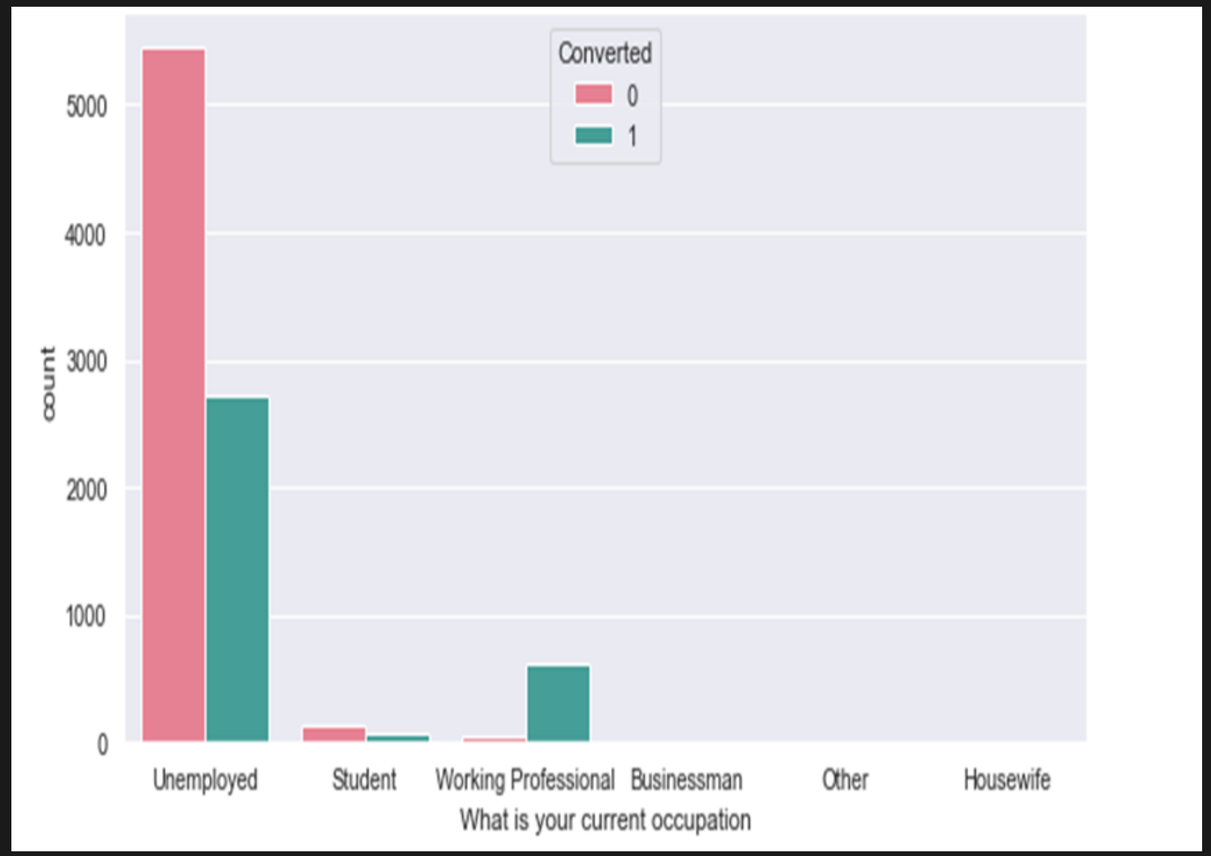
- THE MEDIAN VALUE REMAINS CONSISTENT ACROSS BOTH CONVERTED AND NON-CONVERTED LEADS.
- REGARDING LEAD CONVERSION, NO DEFINITIVE CONCLUSIONS CAN BE DRAWN SPECIFICALLY BASED ON THE PAGE VIEWS PER VISIT PARAMETER.



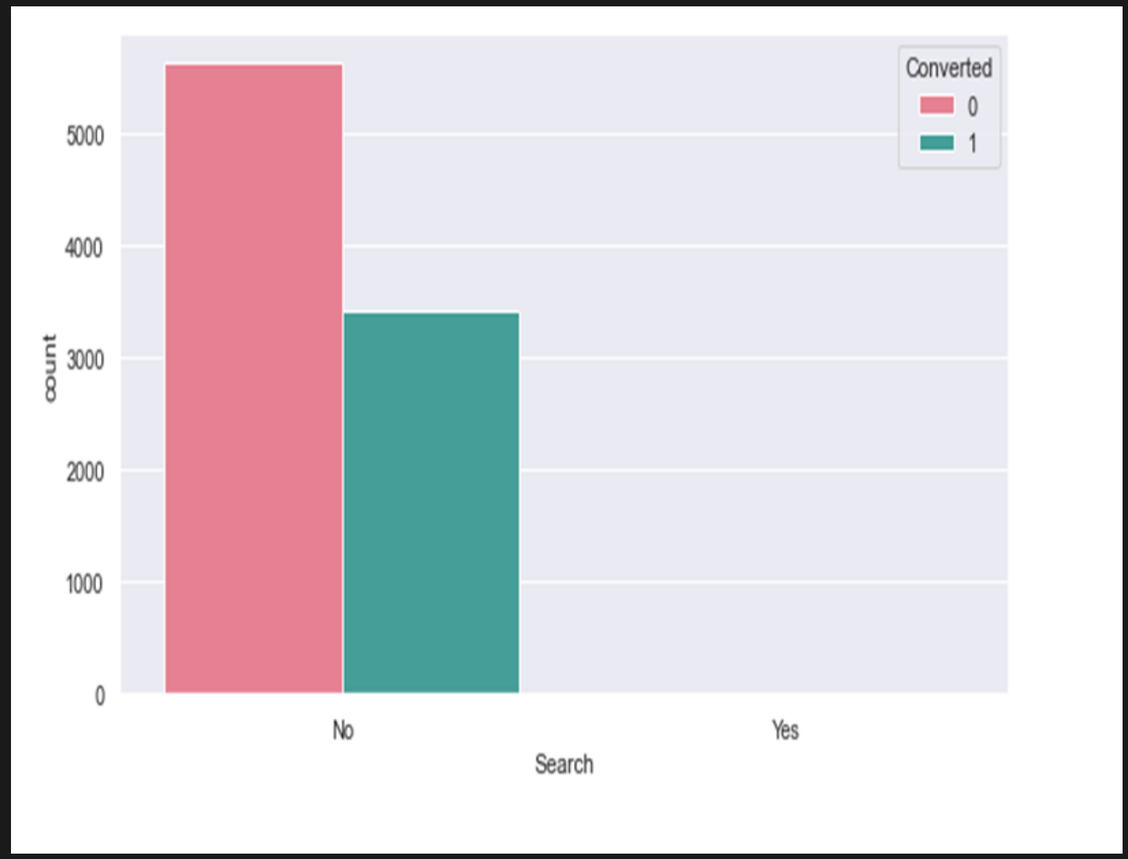
1. THE MAJORITY OF LEADS INDICATE THEIR LAST ACTIVITY AS HAVING OPENED THEIR EMAIL.
2. CONVERSION RATE FOR LEADS WITH LAST ACTIVITY AS SMS SENT IS ALMOST 60%.



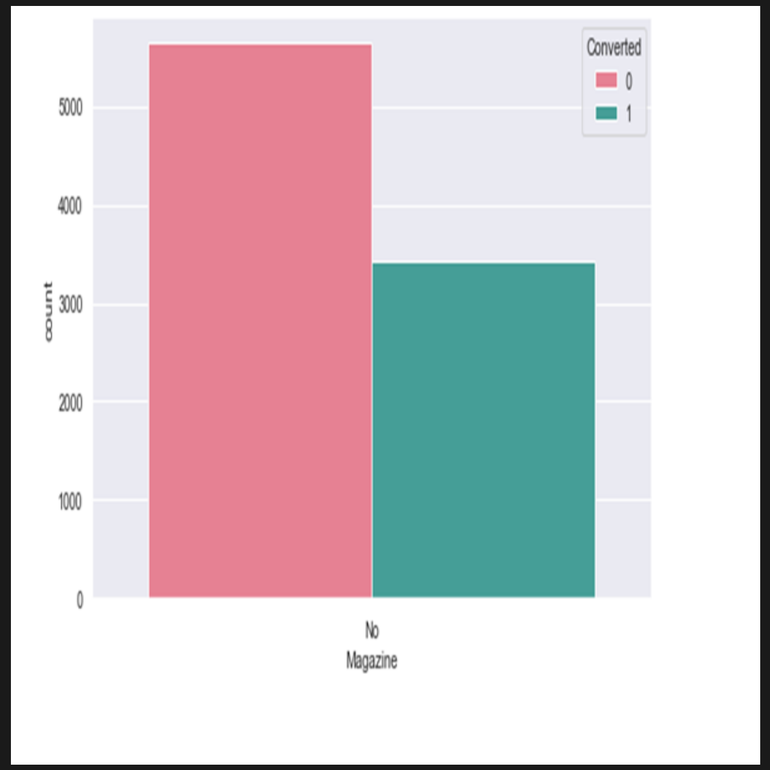
THE EMPHASIS SHOULD BE DIRECTED TOWARDS THE SPECIALIZATION CATEGORIES THAT EXHIBIT A HIGH CONVERSION RATE.

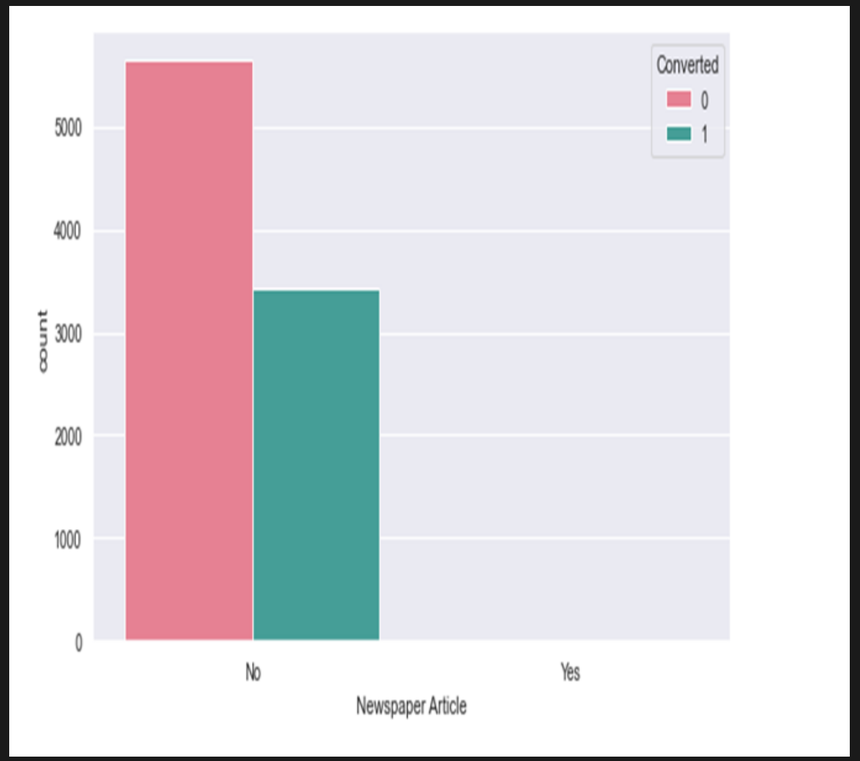


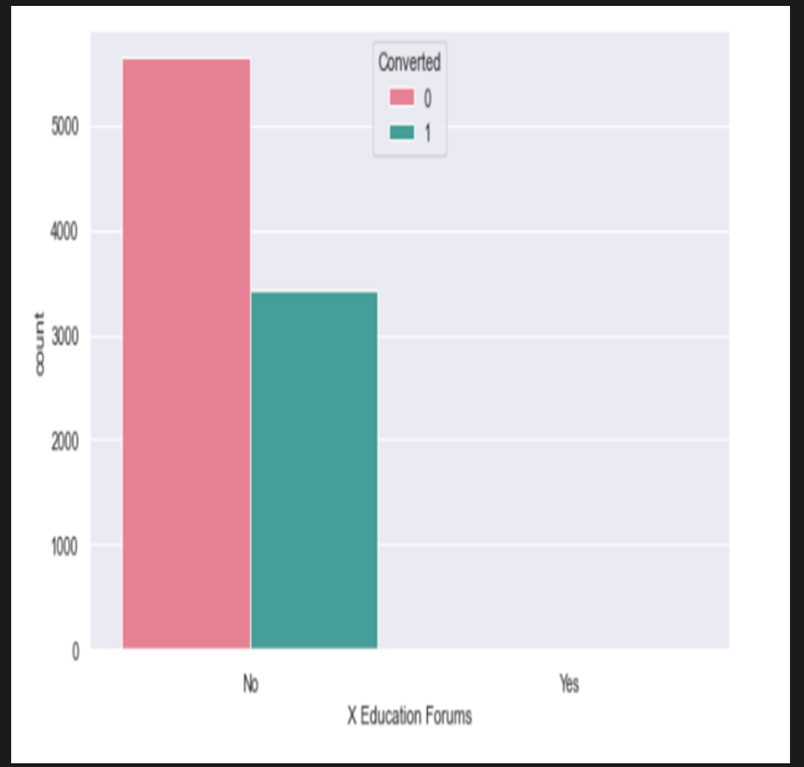
1. THERE IS A STRONG LIKELIHOOD OF WORKING PROFESSIONALS ENROLLING IN THE COURSE AND SUBSEQUENTLY JOINING IT.
2. ALTHOUGH THE MAJORITY OF LEADS ARE CLASSIFIED AS UNEMPLOYED, IT IS NOTEWORTHY THAT THIS CATEGORY EXHIBITS A CONVERSION RATE OF APPROXIMATELY 30-35%.



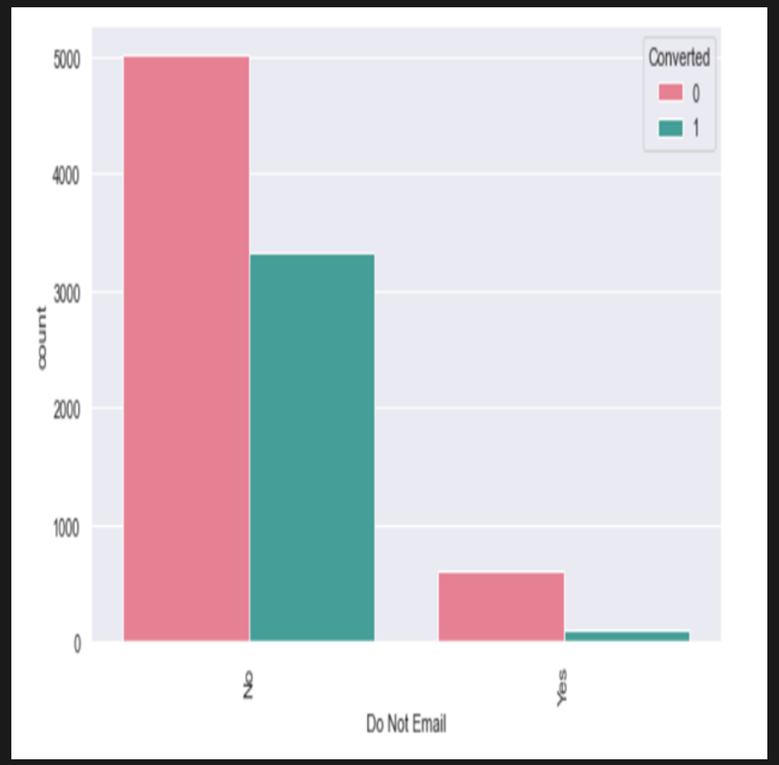
1. THERE IS A STRONG LIKELIHOOD OF WORKING PROFESSIONALS ENROLLING IN THE COURSE AND SUBSEQUENTLY JOINING IT.
2. ALTHOUGH THE MAJORITY OF LEADS ARE CLASSIFIED AS UNEMPLOYED, IT IS NOTEWORTHY THAT THIS CATEGORY EXHIBITS A CONVERSION RATE OF APPROXIMATELY 30-35%.

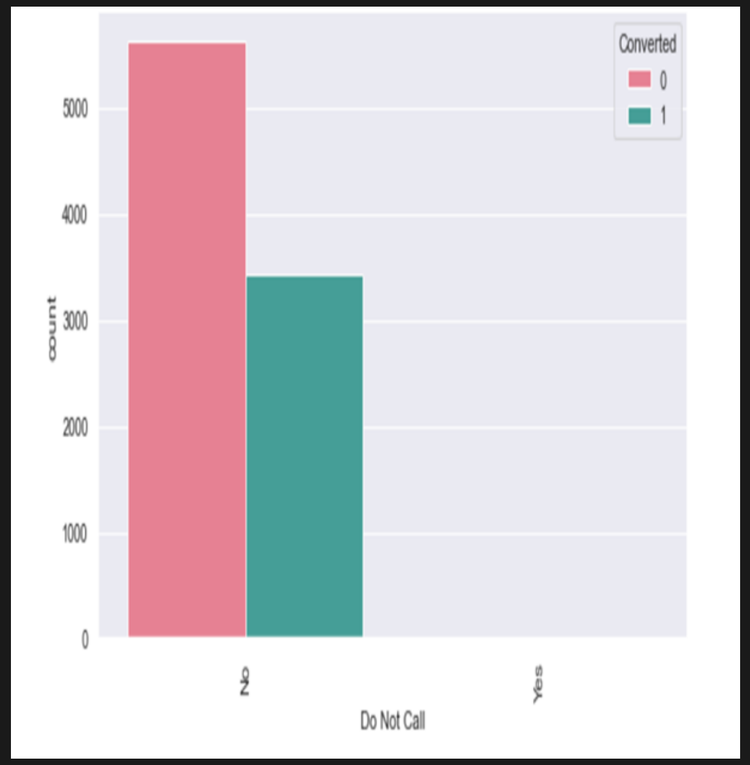


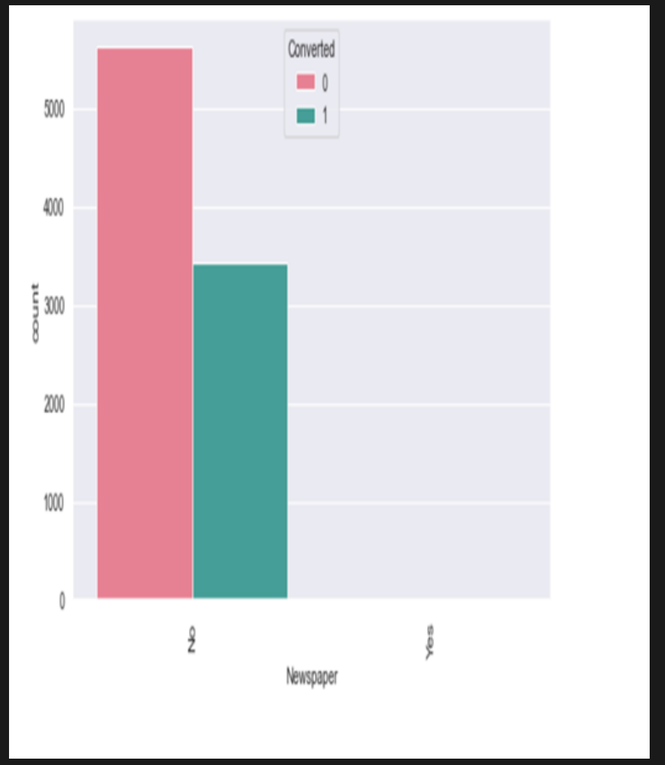




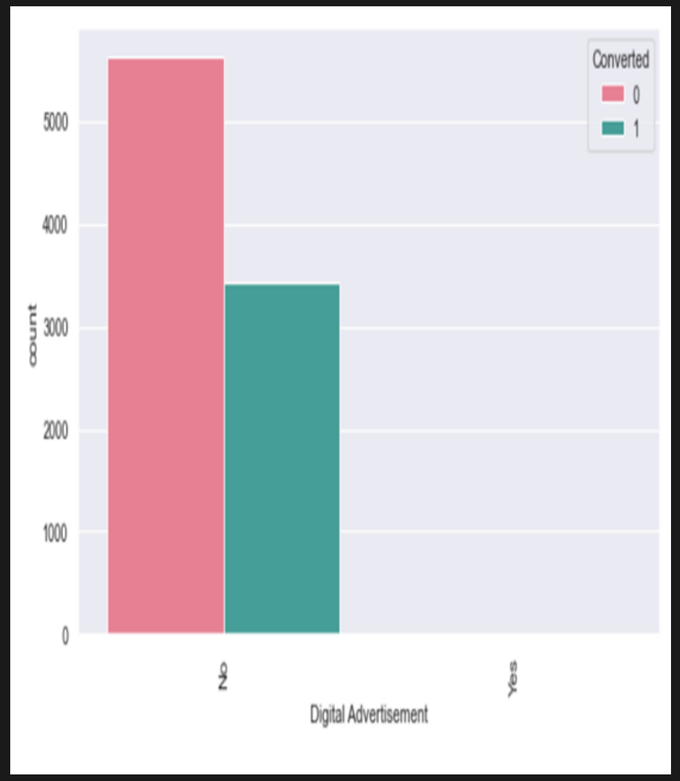
DUE TO THE OVERWHELMING PREVALENCE OF 'NO' ENTRIES IN THIS PARAMETERS, IT IS INCONCLUSIVE AND OFFERS NO MEANINGFUL INFERENCE.

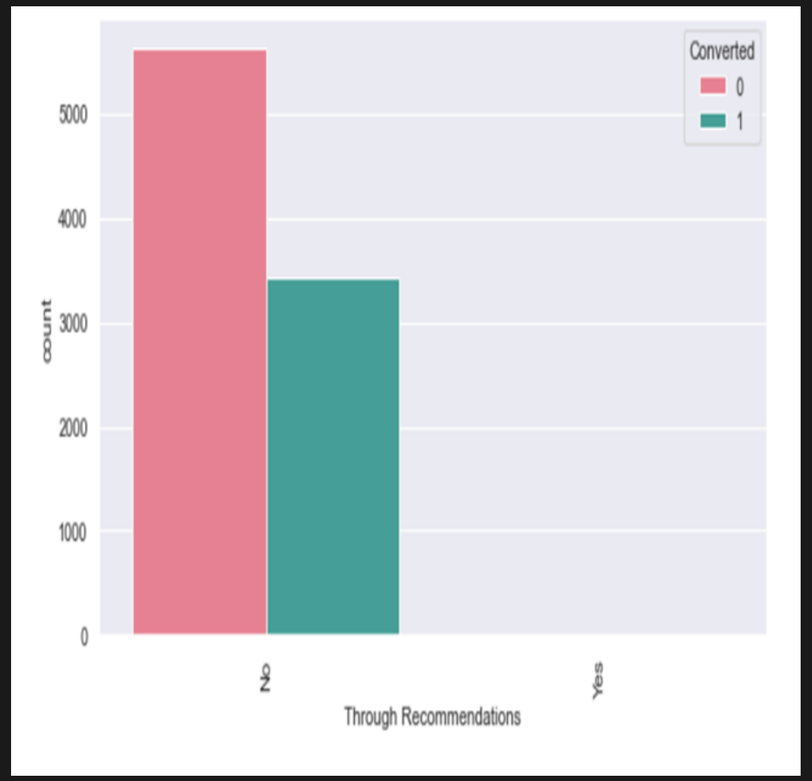


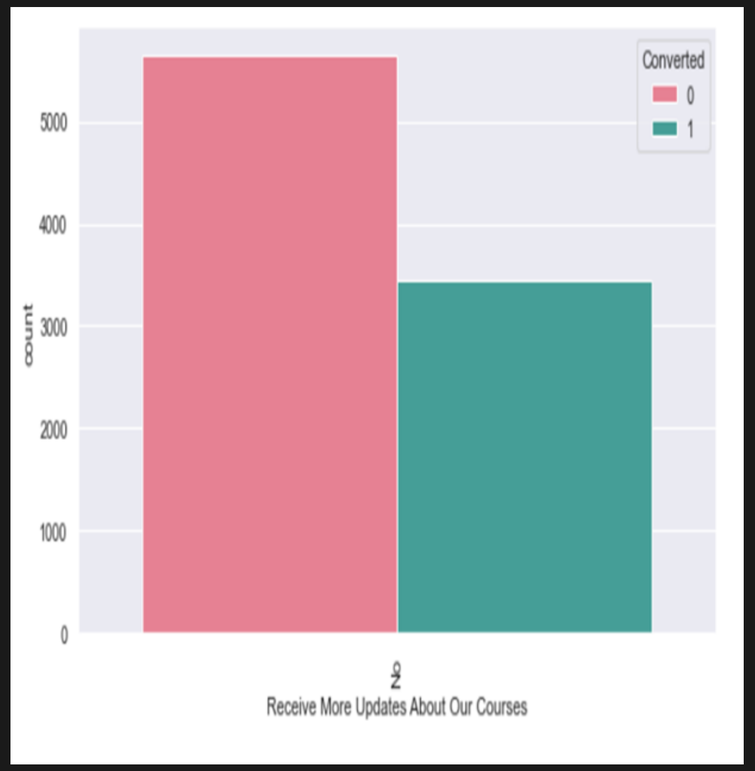




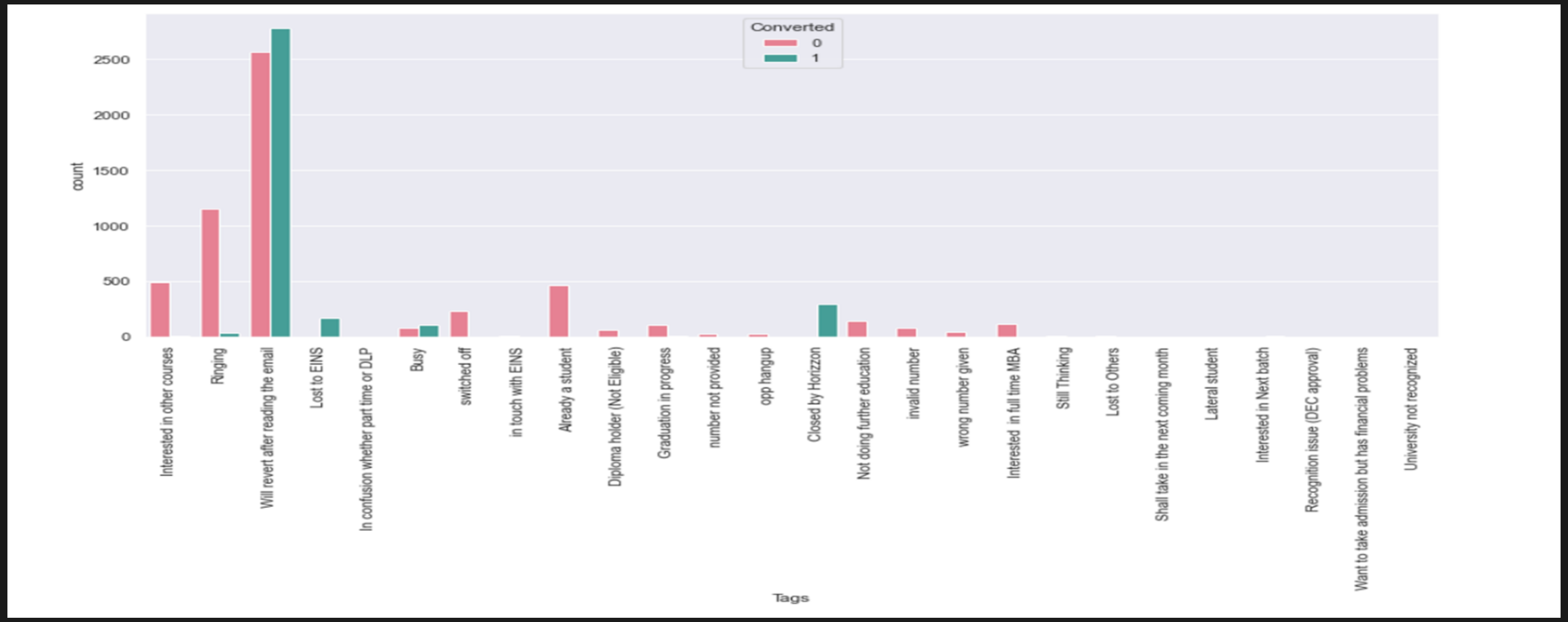
DUE TO THE OVERWHELMING PREVALENCE OF 'NO' ENTRIES IN THIS PARAMETERS, IT IS INCONCLUSIVE AND OFFERS NO MEANINGFUL INFERENCE.



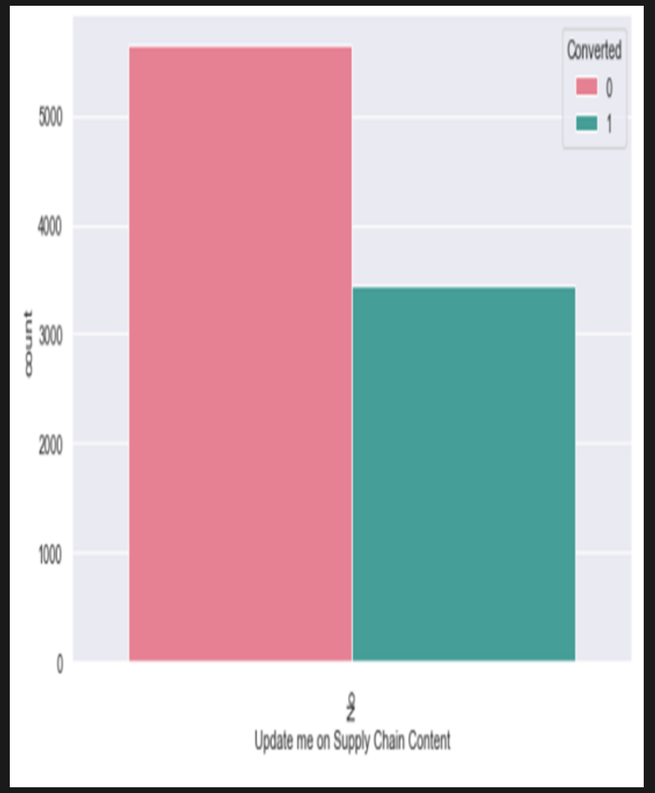


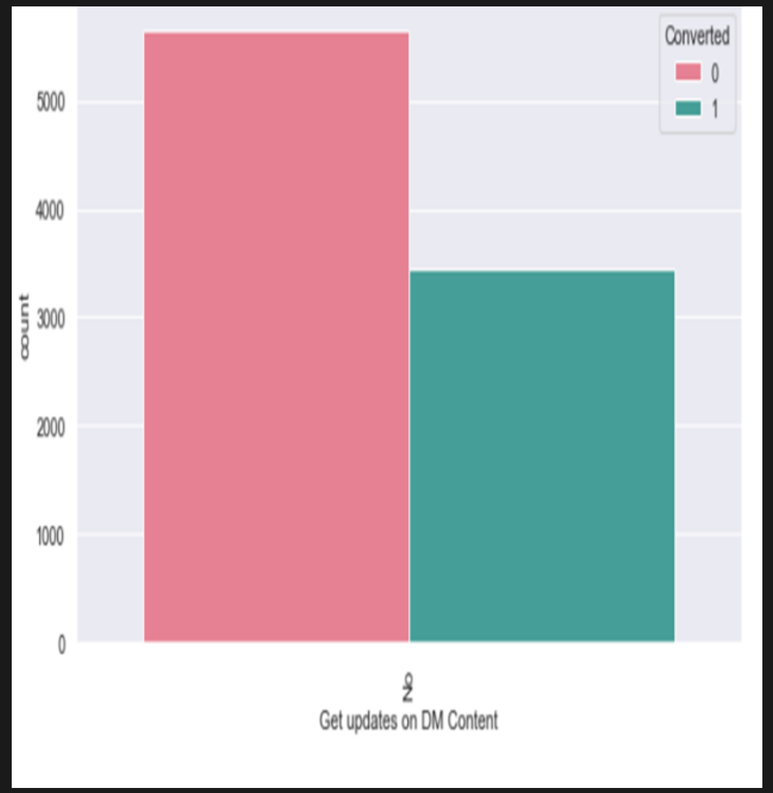


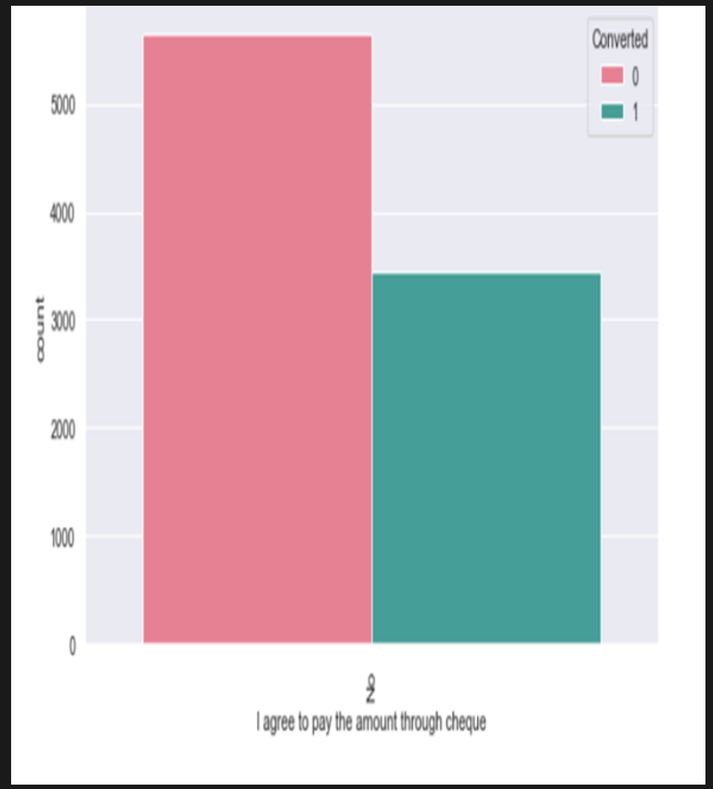
DUE TO THE OVERWHELMING PREVALENCE OF 'NO' ENTRIES IN THIS PARAMETERS, IT IS INCONCLUSIVE AND OFFERS NO MEANINGFUL INFERENCE.



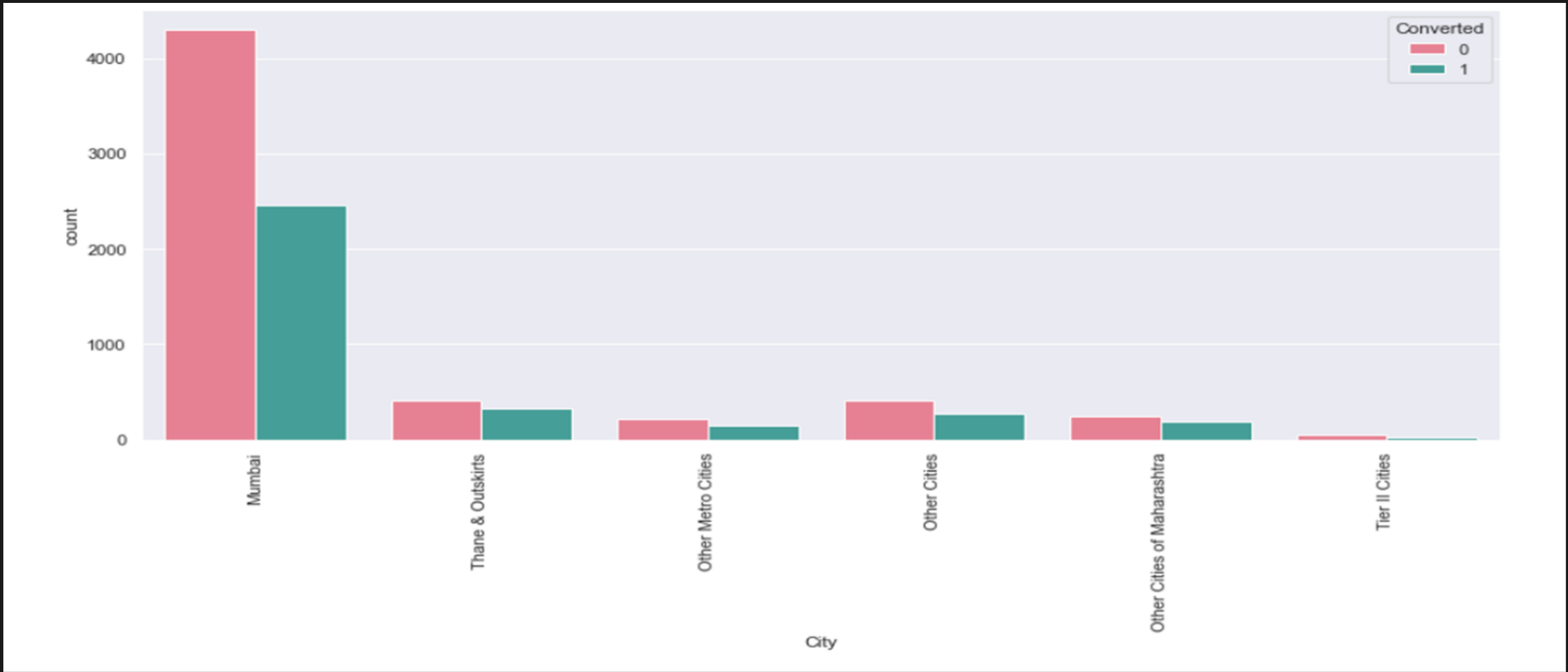
AS THE "TAGS" COLUMN IS GENERATED SOLELY FOR SALES TEAM ANALYSIS AND IS NOT APPLICABLE FOR MODEL BUILDING PURPOSES, IT IS NECESSARY TO EXCLUDE THIS COLUMN FROM THE DATASET PRIOR TO CONSTRUCTING THE MODEL.



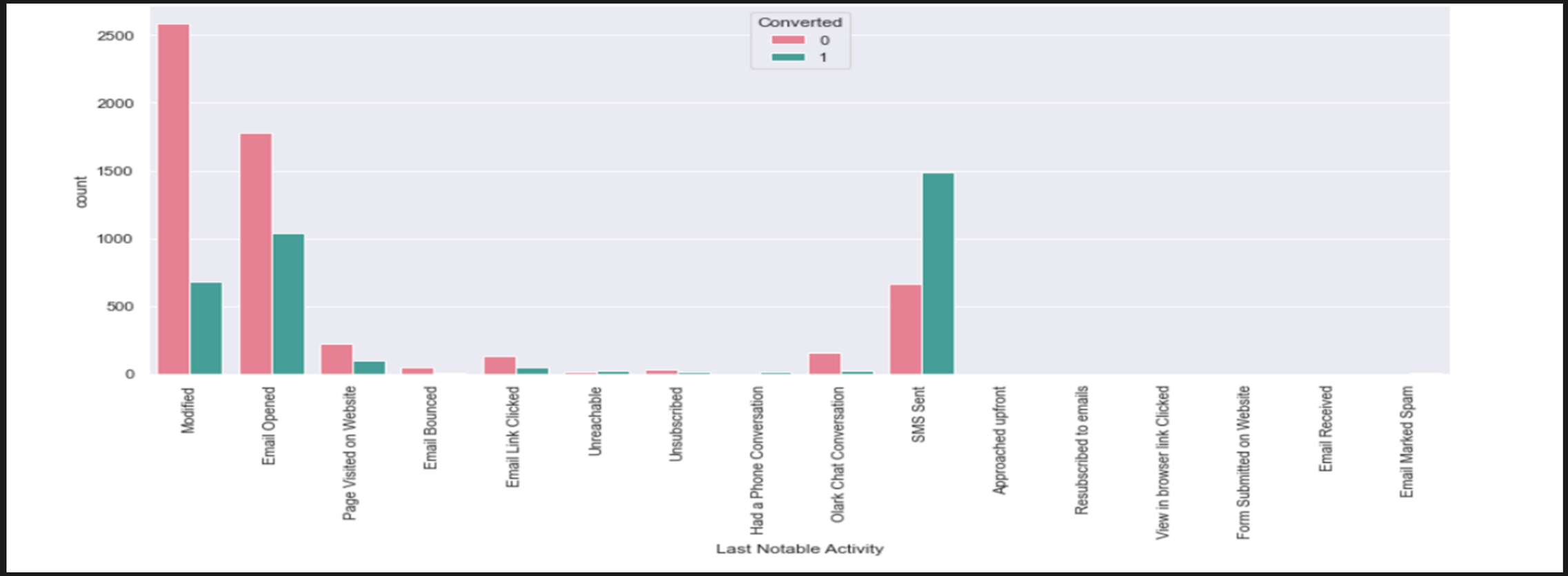




DUE TO THE OVERWHELMING PREVALENCE OF 'NO' ENTRIES IN THIS PARAMETERS, IT IS INCONCLUSIVE AND OFFERS NO MEANINGFUL INFERENCE.



A NOTABLE OBSERVATION REVEALS THAT A SIGNIFICANT MAJORITY OF LEADS ORIGINATE FROM MUMBAI, ACCOMPANIED BY AN APPROXIMATE 50% CONVERSION RATE.



UPON CONDUCTING UNIVARIATE ANALYSIS, IT HAS BECOME APPARENT THAT NUMEROUS COLUMNS DO NOT CONTRIBUTE ANY VALUABLE INFORMATION TO THE MODEL. THEREFORE, IT IS ADVISABLE TO EXCLUDE THESE COLUMNS FOR FURTHER ANALYSIS. SO WE DROP 'LEAD NUMBER','TAGS','COUNTRY','SEARCH','MAGAZINE','NEWSPAPER ARTICLE','X EDUCATION FORUMS', 'NEWSPAPER','DIGITAL ADVERTISEMENT','THROUGH RECOMMENDATIONS','RECEIVE MORE UPDATES ABOUT OUR COURSES‘, 'UPDATE ME ON SUPPLY CHAIN CONTENT','GET UPDATES ON DM CONTENT','I AGREE TO PAY THE AMOUNT THROUGH CHEQUE', 'A FREE COPY OF MASTERING THE INTERVIEW'

Data Preparation

1. Converting some binary variables (Yes/No) to 1/0
2. Creating Dummy variables for the categorical features: 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation',‘ City',‘ Last Notable Activity
3. Splitting the data into train and test sets. (70,30 ratio is chosen for the split)
4. Feature scaling using the standardization method
5. Feature Selection Using RFE(Recursive Feature Elimination)

Model Building

- Assessing the model with Stats Models
- Model 1: Considering the significantly high p-value for the column 'What is your current occupation Housewife', it is recommended to remove this column from further analysis.
- Model 2 : As the p-value associated with the 'Last Notable Activity Had a Phone Conversation' column is considerably high, it is recommended to remove this column from further analysis.
- Model 3: Considering the significantly high p-value associated with the 'What is your current occupation Student' column, it is reasonable to omit this column from further analysis.
- Model 4: Given the considerably high p-value associated with the 'Lead Origin Lead Add Form' column, it is advisable to remove this column from further analysis.
- After careful analysis, it has been determined that all variables in model-9 exhibit p-values of 0 and possess low VIF values. Hence, model-9 with a total of 12 variables is considered the final model for our analysis.

Making Predictions on the Train set

1. Creating a data frame with the actual Converted flag and the predicted probabilities
2. Choosing an arbitrary cut-off probability point of 0.5 to find the predicted labels
3. Creating a new column 'predicted' with 1 if Converted_Prob > 0.5 else 0
4. Making the Confusion matrix
5. Metrics beyond simply accuracy

Upon analysis, it was observed that our specificity exhibited a satisfactory level of approximately 88%, whereas our sensitivity was relatively lower at 70%. Therefore, addressing and improving the sensitivity aspect became a crucial requirement.

The obtained sensitivity value of 70% can be primarily attributed to the arbitrary selection of the 0.5 cut-off point. In order to achieve a more desirable sensitivity value, optimization of this cut-off point is necessary, which can be accomplished through the utilization of the ROC curve.

Plotting the ROC Curve:

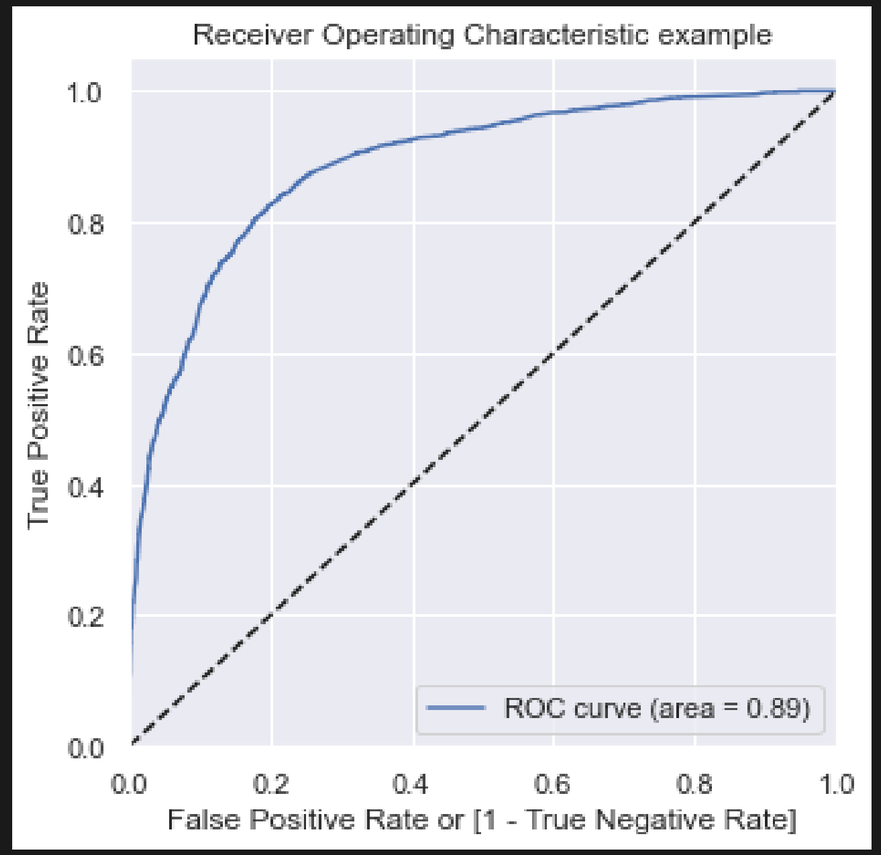
Here are three key insights that can be derived from a ROC curve:

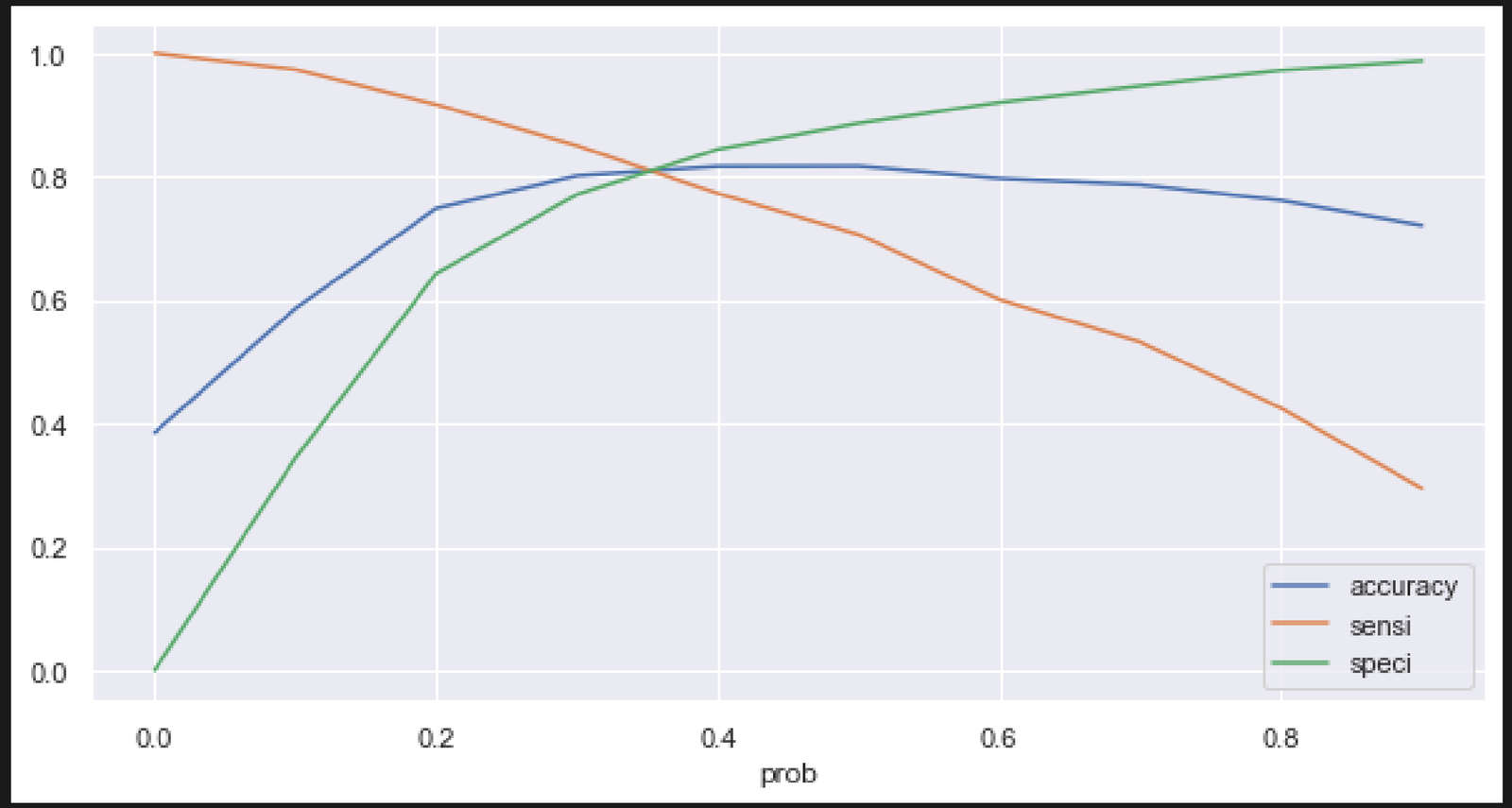
1. Sensitivity and specificity tradeoff: The curve illustrates the relationship between sensitivity (true positive rate) and specificity (true negative rate). If sensitivity increases, it typically leads to a decrease in specificity, and vice versa.

2. Accuracy indication: The proximity of the curve to the left and upper borders of the ROC space indicates the accuracy of the test. The closer the curve follows these borders, the more accurate the test is considered to be.

3. Diagnostic performance indication: The proximity of the curve to the 45-degree diagonal line in the ROC space reflects the test's accuracy. When the curve is closer to this diagonal, it suggests that the test is less accurate in distinguishing between true positive and true negative cases.

Plotting the ROC Curve:





Model Evaluation

Based on the evaluation of the model using the test data, the following observations were made:

- Accuracy: The model achieved an accuracy rate of 80.4%. This indicates that approximately 80.4% of the predictions made by the model were correct.
- Sensitivity: The sensitivity, also known as the true positive rate, was also found to be 80.4%. This means that the model correctly identified 80.4% of the positive instances in the test data.
- Specificity: The specificity, also referred to as the true negative rate, was determined to be 80.5%. This implies that the model accurately recognized 80.5% of the negative instances in the test data.
- These results provide insights into the model's performance in terms of overall accuracy and its ability to correctly classify both positive and negative cases.

Here are the results obtained when comparing the performance metrics for the train and test data:

For the train data:

- Accuracy: 81.0%
- Sensitivity: 81.7%
- Specificity: 80.6%

For the test data:

- Accuracy: 80.4%
- Sensitivity: 80.4%
- Specificity: 80.5%

These results indicate the performance of the model on both the train and test datasets. The accuracy represents the overall correctness of the model's predictions. Sensitivity refers to the model's ability to correctly identify positive instances, while specificity measures its capability to correctly identify negative instances.

As a result, our objective of obtaining an estimated lead conversion rate of approximately 80% has been accomplished. The model demonstrates a strong ability to predict the conversion rate, instilling confidence in utilizing it as a tool for making informed decisions. With this model, we can provide the CEO with the assurance of making well-informed choices to achieve a higher lead conversion rate of 80%.

Recommendations

1. To increase the chances of conversion, the company should prioritize making calls to leads that originate from the lead sources "Welingak Websites" and "Reference". These sources have shown a higher likelihood of conversion based on available data.
2. To enhance the chances of conversion, it is recommended that the company prioritize making calls to leads categorized as "working professionals." This particular segment exhibits a higher likelihood of successful conversions.
3. To improve the conversion rate, it is advisable for the company to prioritize making calls to leads who have spent a significant amount of time on the websites. These leads have demonstrated a higher level of engagement, suggesting a greater likelihood of conversion. By focusing on these leads, the company can potentially increase its chances of successful conversions.
4. To increase the chances of lead conversion, it is advisable for the company to prioritize making calls to leads originating from the lead source "Olark Chat." These leads have a higher likelihood of being successfully converted into customers.
5. To increase the likelihood of conversions, it is advisable for the company to prioritize making calls to leads whose last activity recorded in the dataset was "SMS Sent." This strategy takes into account the observation that leads who have recently received an SMS message are more likely to respond positively to phone calls and potentially result in conversions.
6. To increase the chances of successful conversions, it is advisable for the company to avoid making calls to leads whose most recent activity is recorded as "Olark Chat Conversation." These leads are considered less likely to be converted, and therefore, it would be more effective for the company to focus their efforts on leads with different or more promising recent activities.
7. It is recommended that the company refrains from making calls to leads whose lead origin is "Landing Page Submission" since these leads are less likely to result in conversions.
8. It is recommended that the company refrain from making calls to leads whose Specialization is categorized as "Others," as these leads are considered unlikely to convert.
9. It is recommended that the company refrain from making phone calls to leads who have indicated their preference as "Do not Email" set to "yes." These leads are unlikely to be converted, so it would be more prudent to avoid contacting them via phone calls.