Lead Score Case Study

Prepared by:

Subham Jha

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Evaluation of the model with the help of metrics like accuracy, sensitivity, specificity and precision.

Data Loading and Understanding

- Loading the ‘Leads.csv’ file.
- It is having 9240 rows and 37 columns in total.
- There are no duplicate data.
- Total of 7 numerical columns and 30 categorical columns.
- Current conversion rate was 39%.
- Also metadata sheet was provided which helped in understanding the significance of the current data(‘Lead Data Dictionary’).

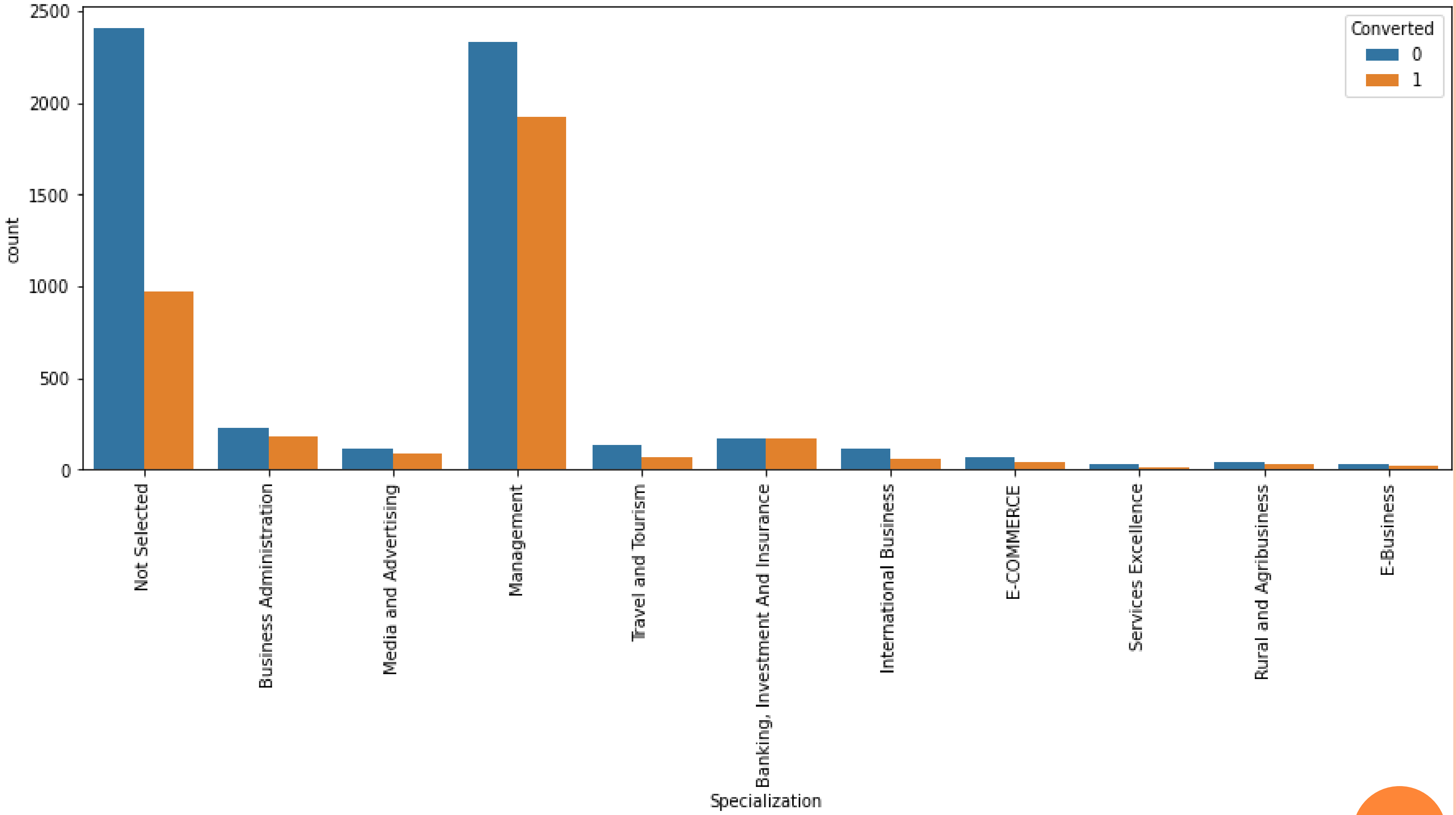
EDA:

- It was seen that the dataset had columns where the value was ‘select’ which was equivalent to null hence converted these values to ‘Null’.
- Then the null percentage value was checked and the columns having null percentage greater than 45 or equal were dropped off and the columns with null percentage less than 45 were treated.
- Categorical columns were imputed with the mode value and continuous columns were imputed with median.
- Also for the categorical analysis if the columns values had low frequency then they were all clubbed in ‘other’ category to reduce the no of levels and make the analysis simpler.

EDA:

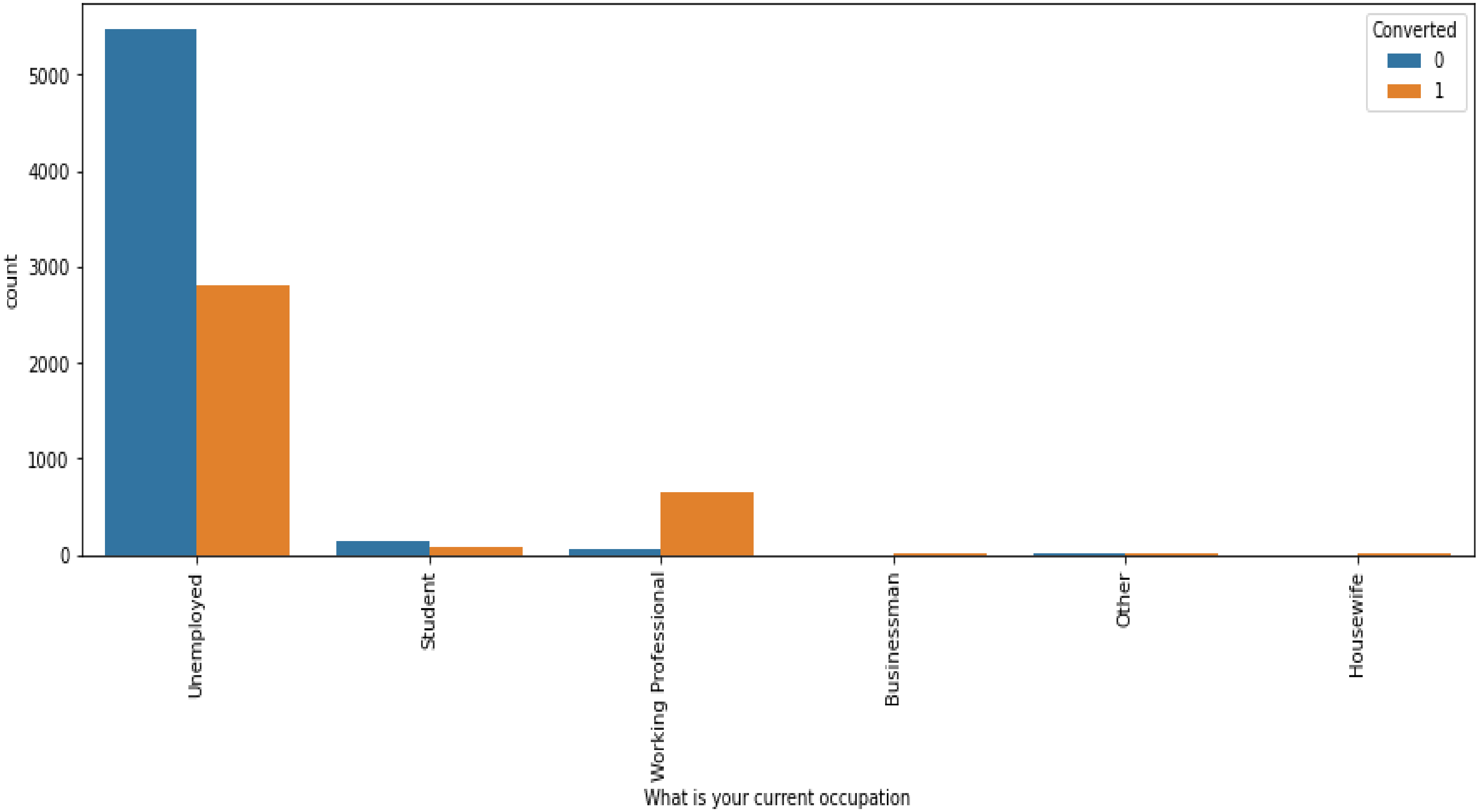
- The columns which was imbalanced or where the data was not proportionate it was dropped off.
- There was separate analysis done for categorical features with respect to ‘converted’ variable.
- And separate analysis for numerical feature with respect to ‘converted’ variable.
- Numerical analysis was carried out by treating the outliers by drawing boxplots and identifying the outliers

Categorical Analysis



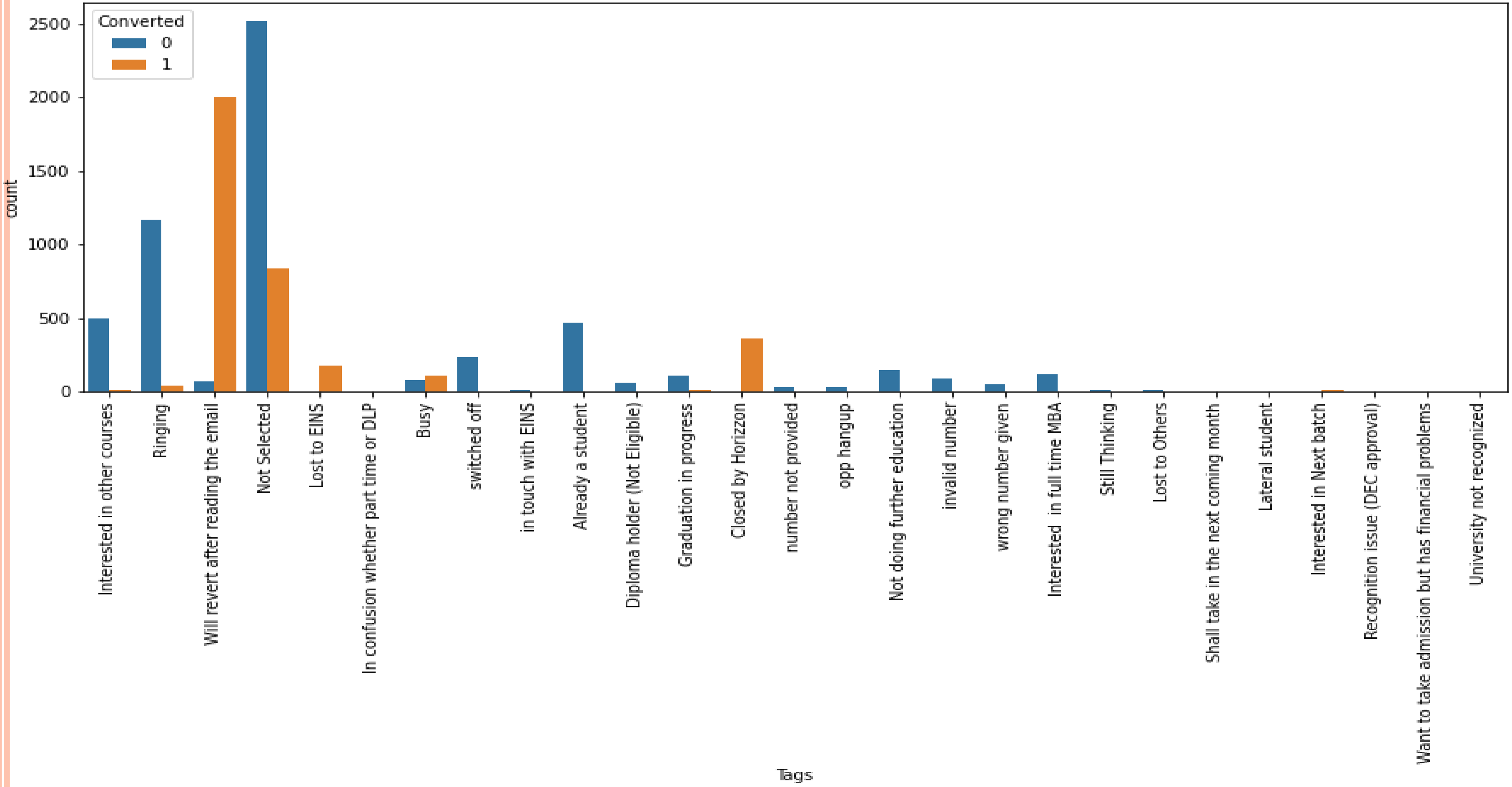
We can see now that Management specialization users are having highest no of leads converted.

Categorical Analysis



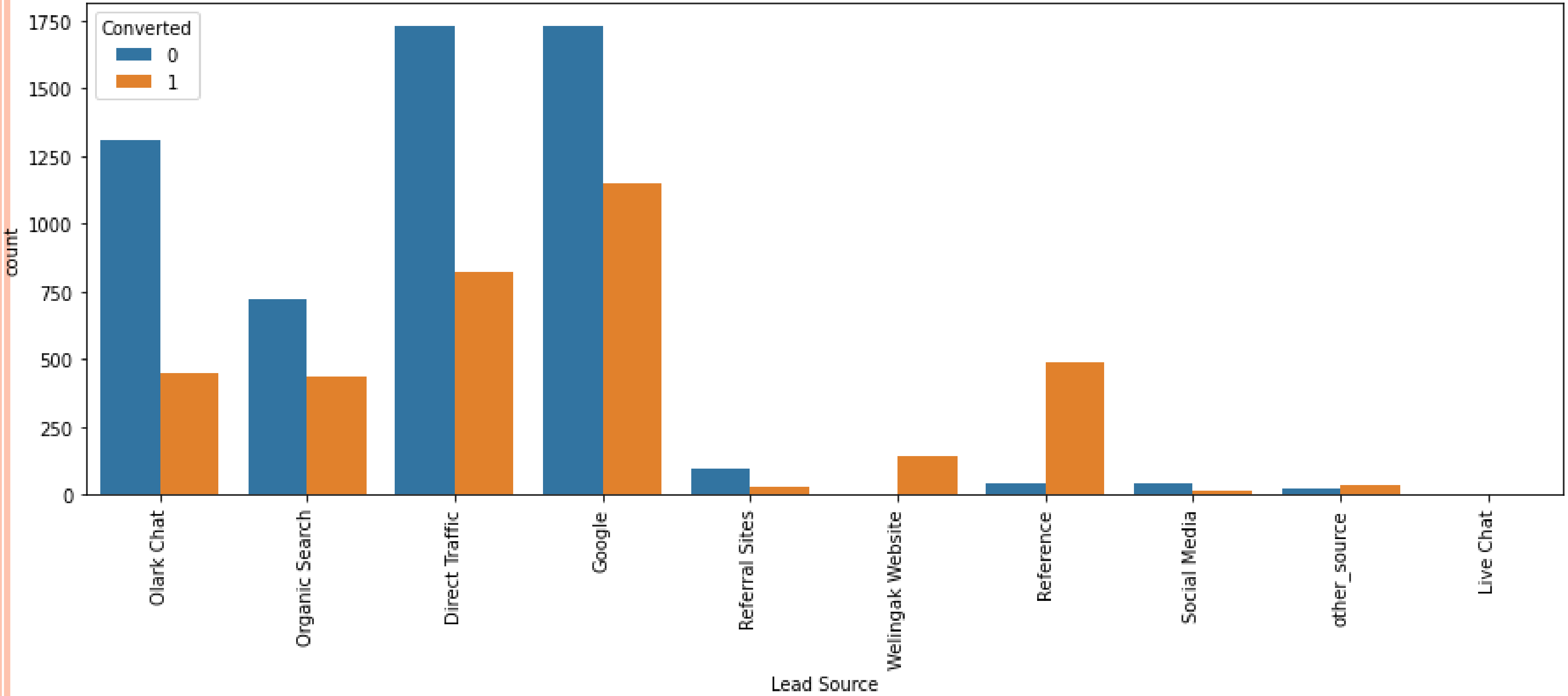
Working prfoessionals have though low no but their conversion rate is higher. Aslo the unemployed are highest in terms of numbers.

Categorical Analysis



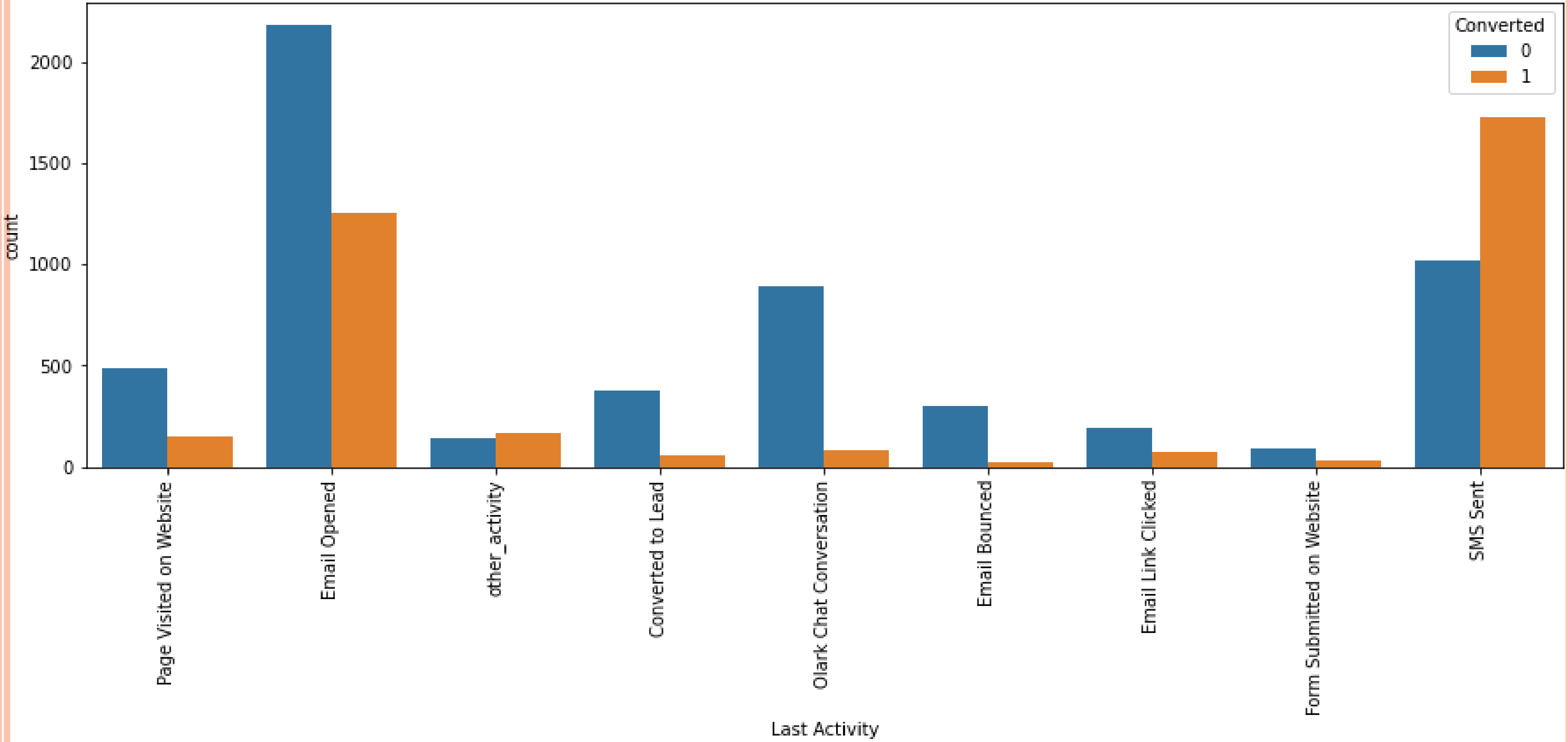
We can see that there are tags which are very low converted rate and it is better to club these tags in other_tags, which will help us to reduce the levels and also help in our analysis.

Categorical Analysis

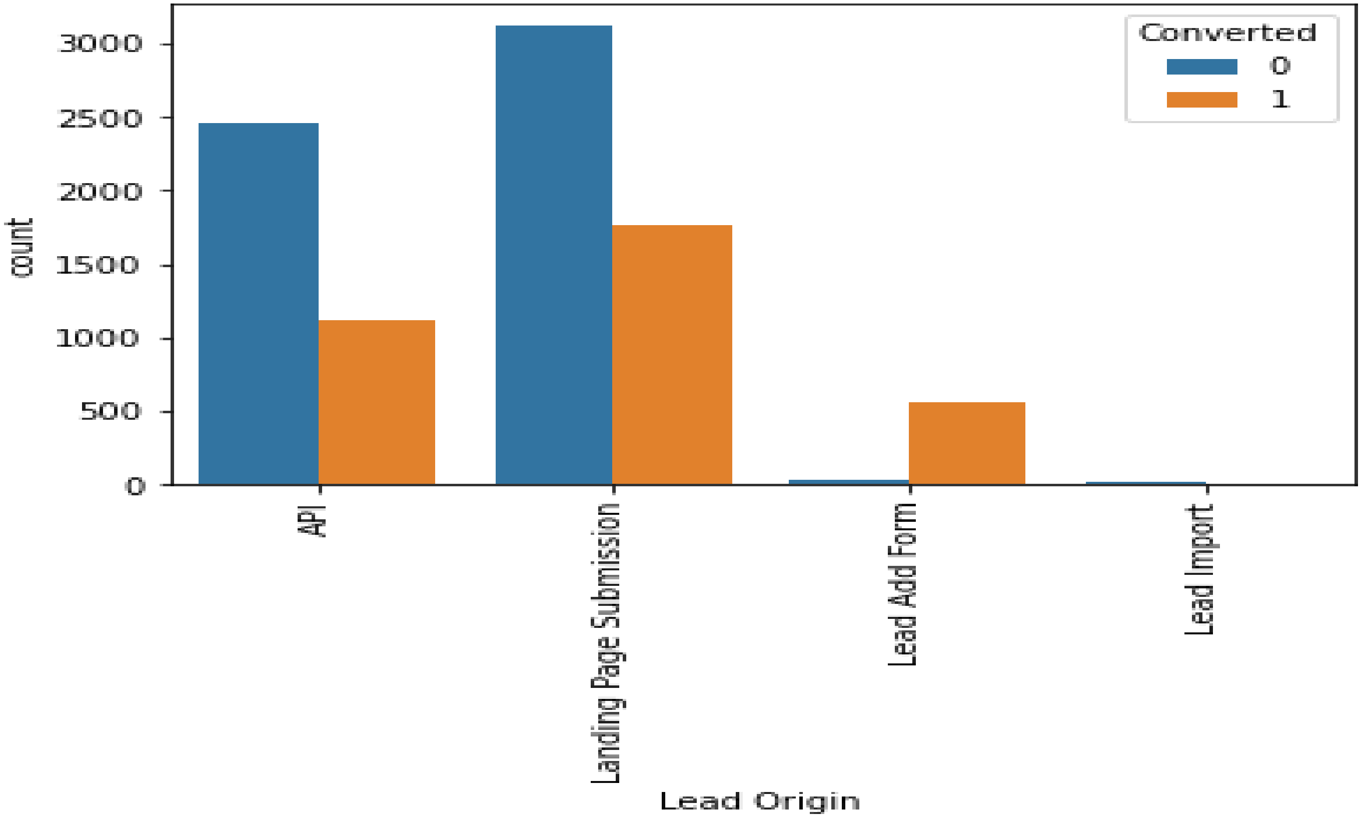


From the above graph it is observed that:
Google and Direct Traffice has the most number of converted users.
The conversion rate through Reference and Welingak Website is high.
The company should focus in it's live chat,referal sites to improve it's conversion rate.

Categorical Analysis

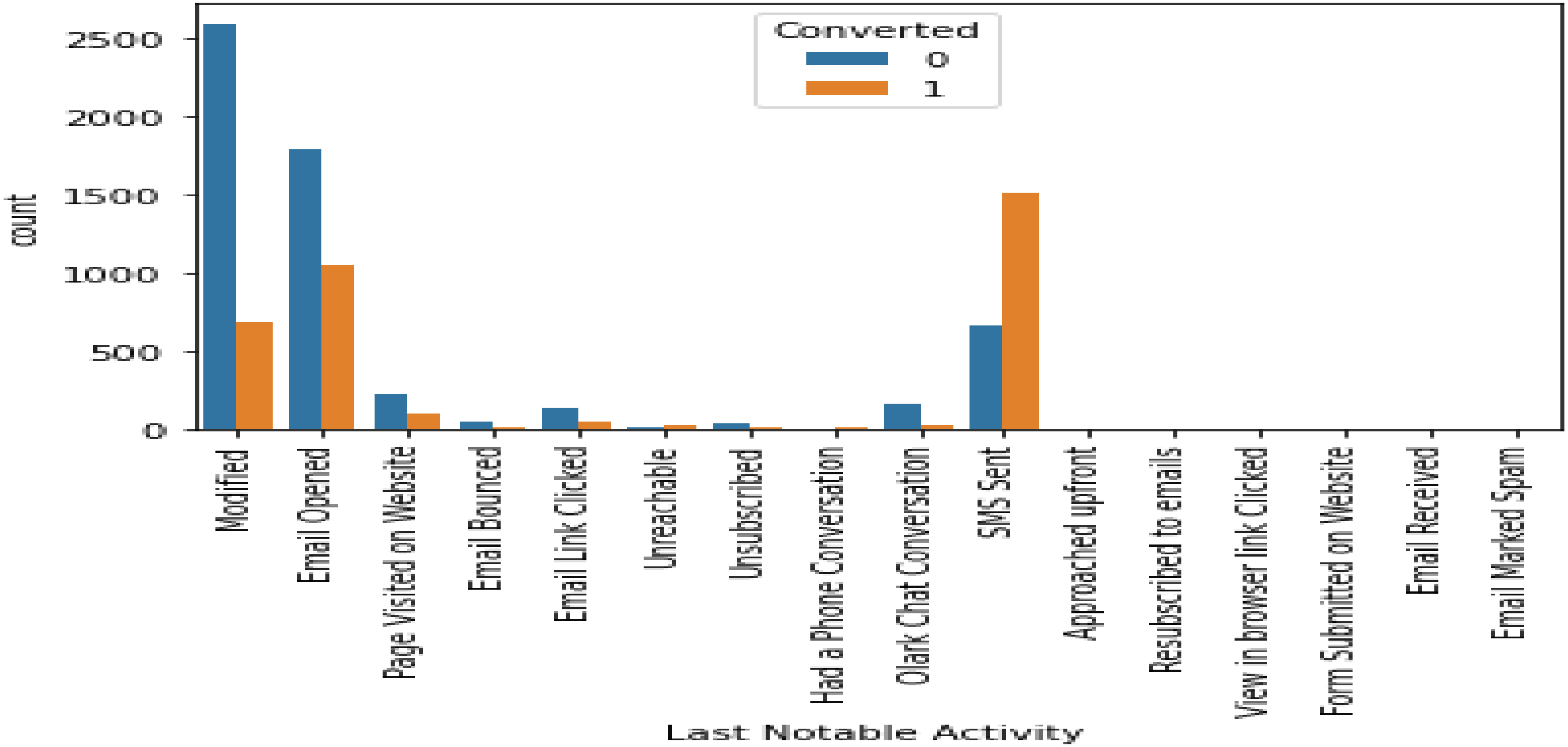


Categorical Analysis



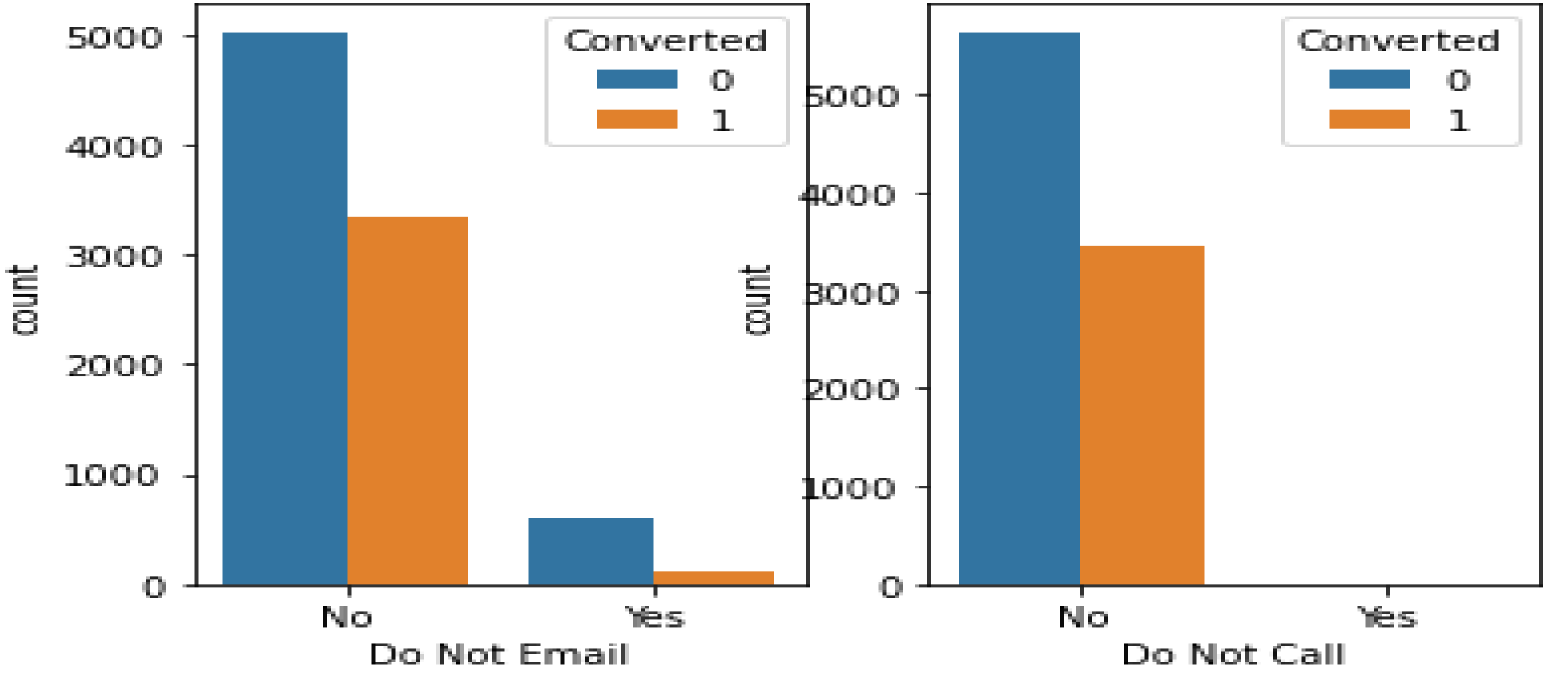
From the above graph we can see that:
Api and Landing Page Submission has the highest no of leads adn also higher rate of conversion. Lead Add Form has low no of leads comparitvely however the concersion rate seems to be pretty high. Hence compan should focus on the above factors to have even higher conversion rates.

Categorical Analysis



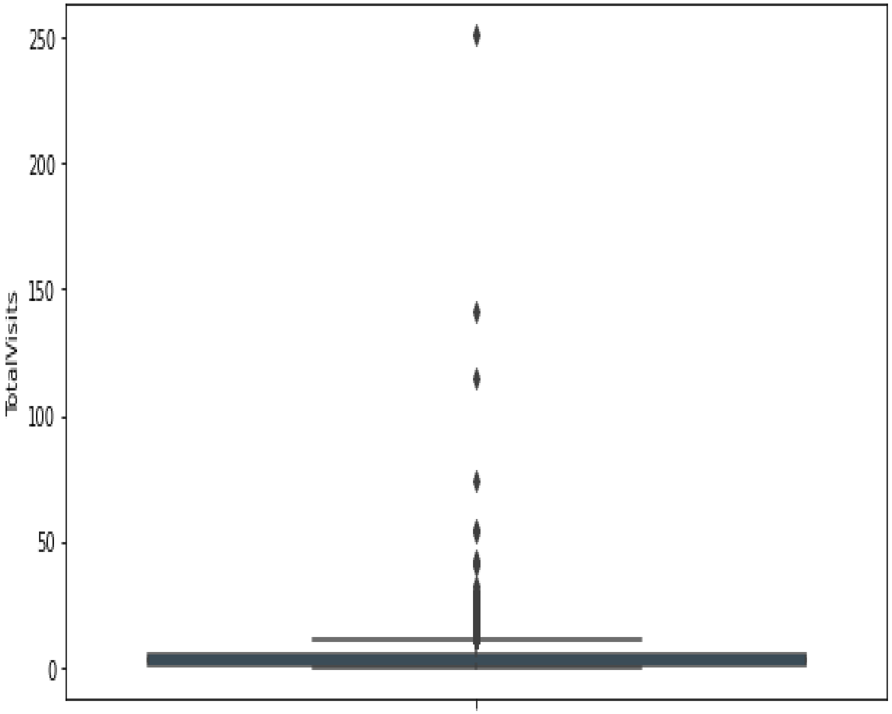
We see that SMS sent has higher conversion rate when compared with other activity.

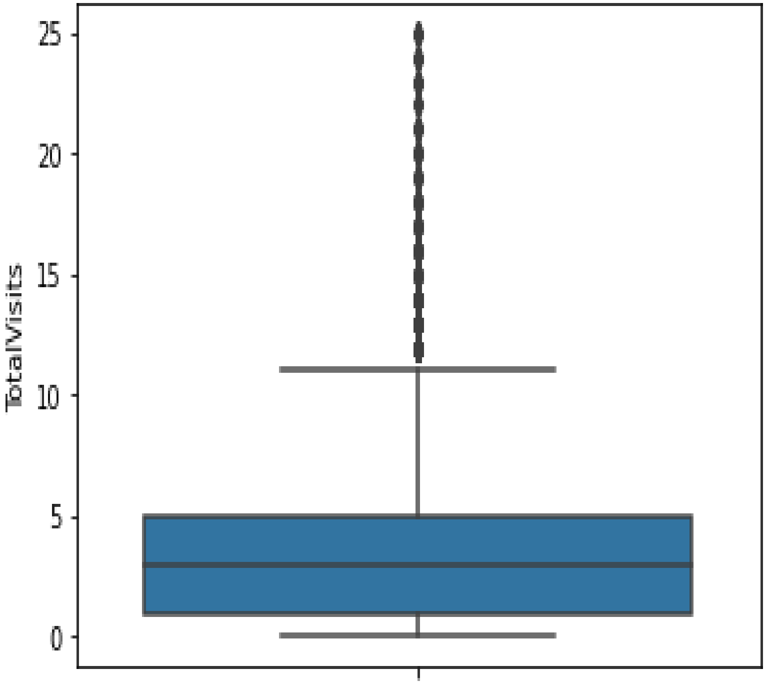
Categorical Analysis



From the above graph we can see that percentage of 'yes' in 'Do Not Call' is significantly low and it makes the variable imbalance hence it is bette to add this variable in drop_col list

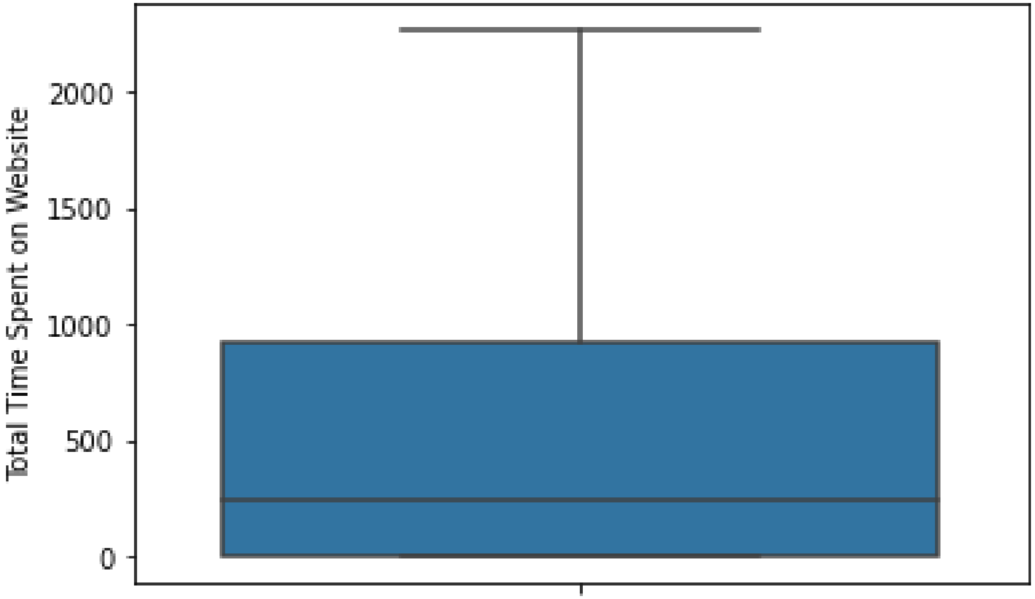
Numerical analysis for checking outliers





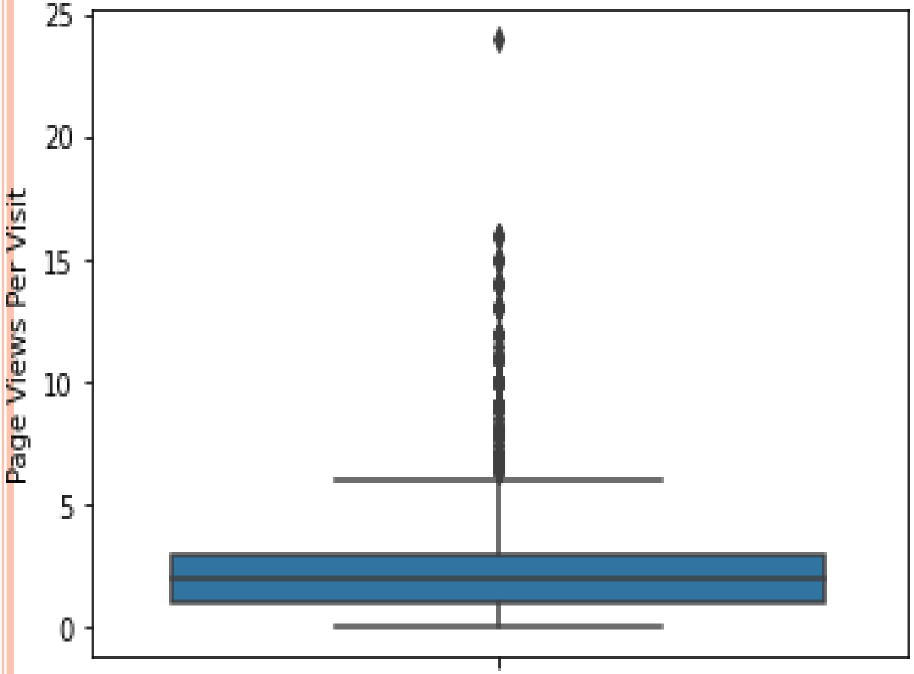
For ‘Total Visit’ feature there were outlier present hence it was treated so that no outlier values should be passed into the model for predicting.

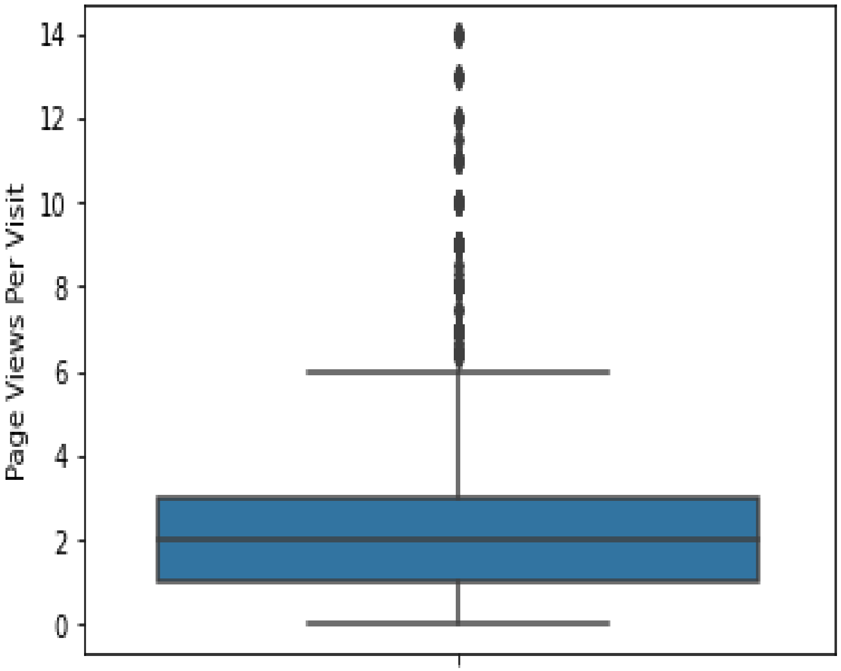
NUMERICAL ANALYSIS FOR CHECKING OUTLIERS



We don't see any major outlier present in the 'Total Time Spent on Website' cols as the value consistently. Hence no outlier treatmentis requireed for this variable

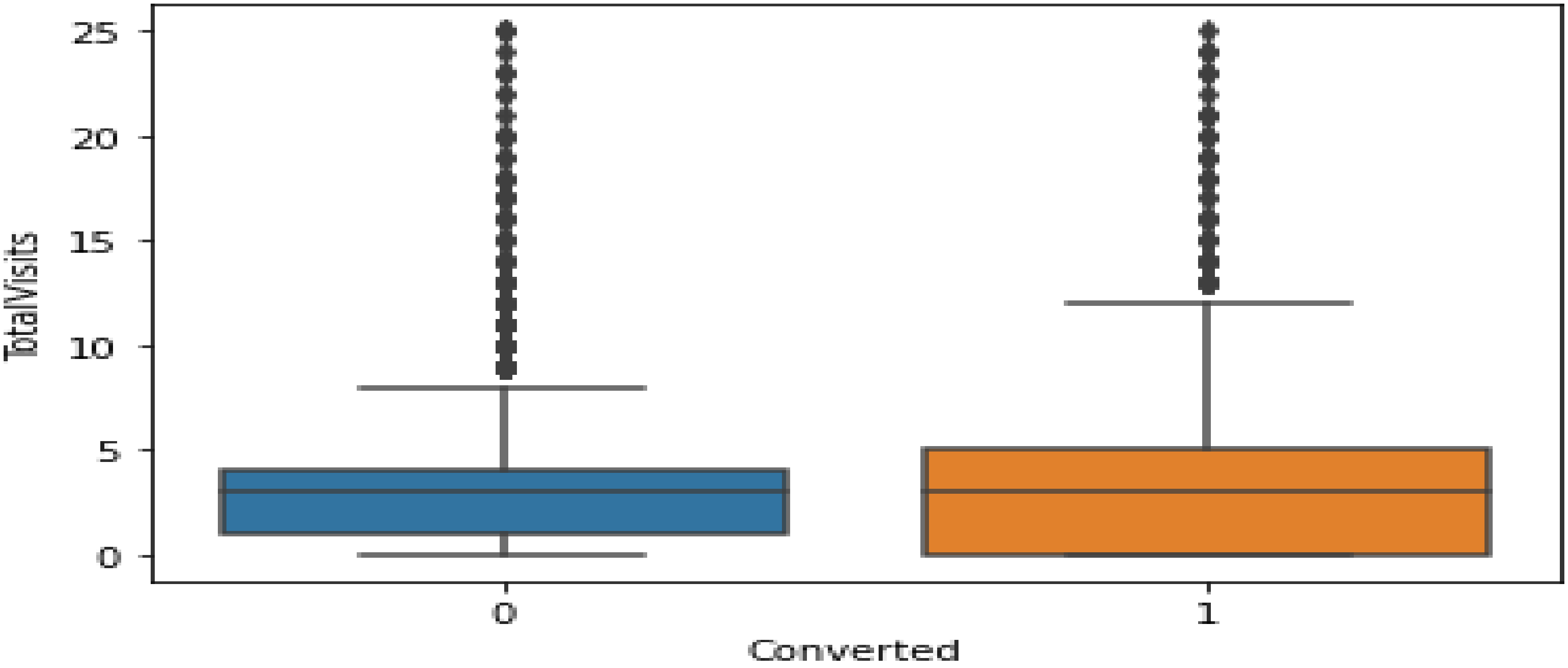
Numerical analysis for checking outliers





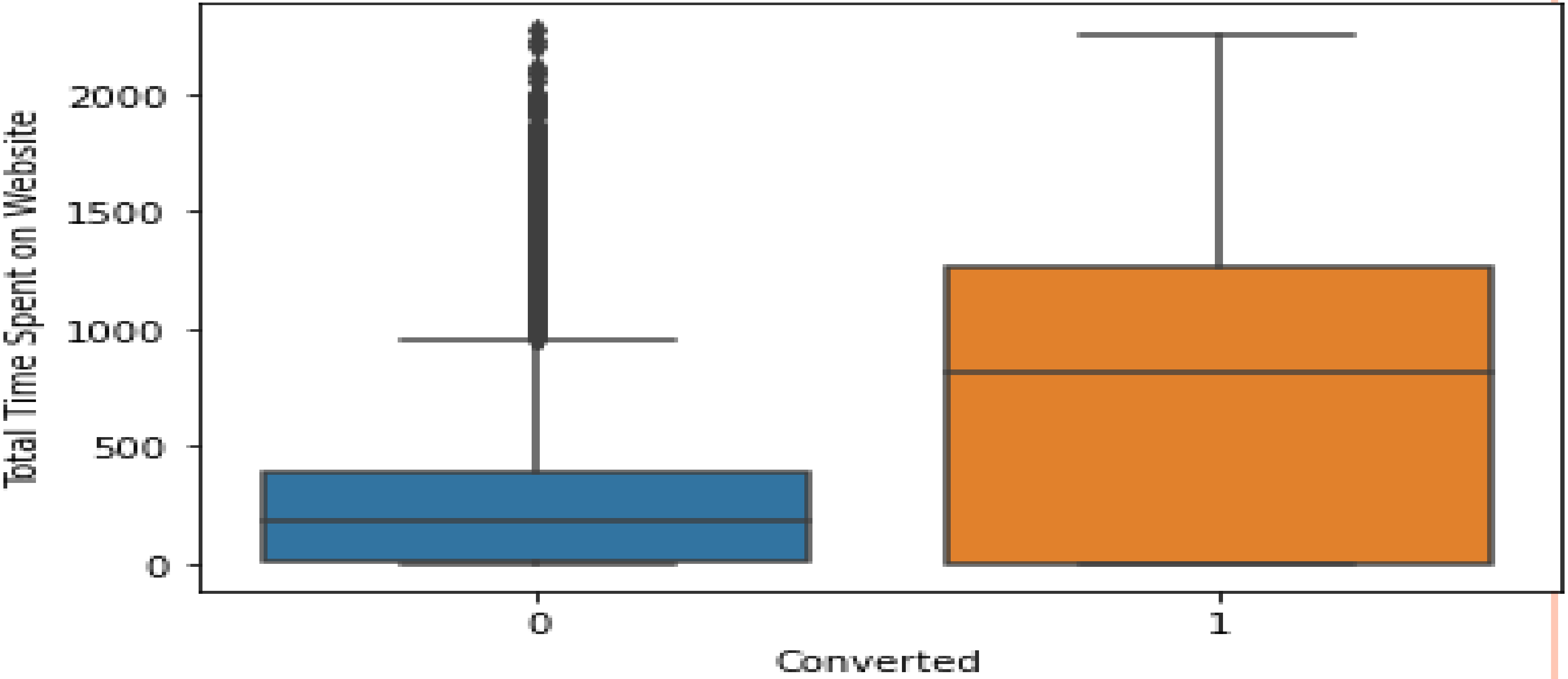
For ‘Total Views Per Visit’ feature there were outlier present hence it was treated so that no outlier values should be passed into the model for predicting.

Bivariate-Analysis



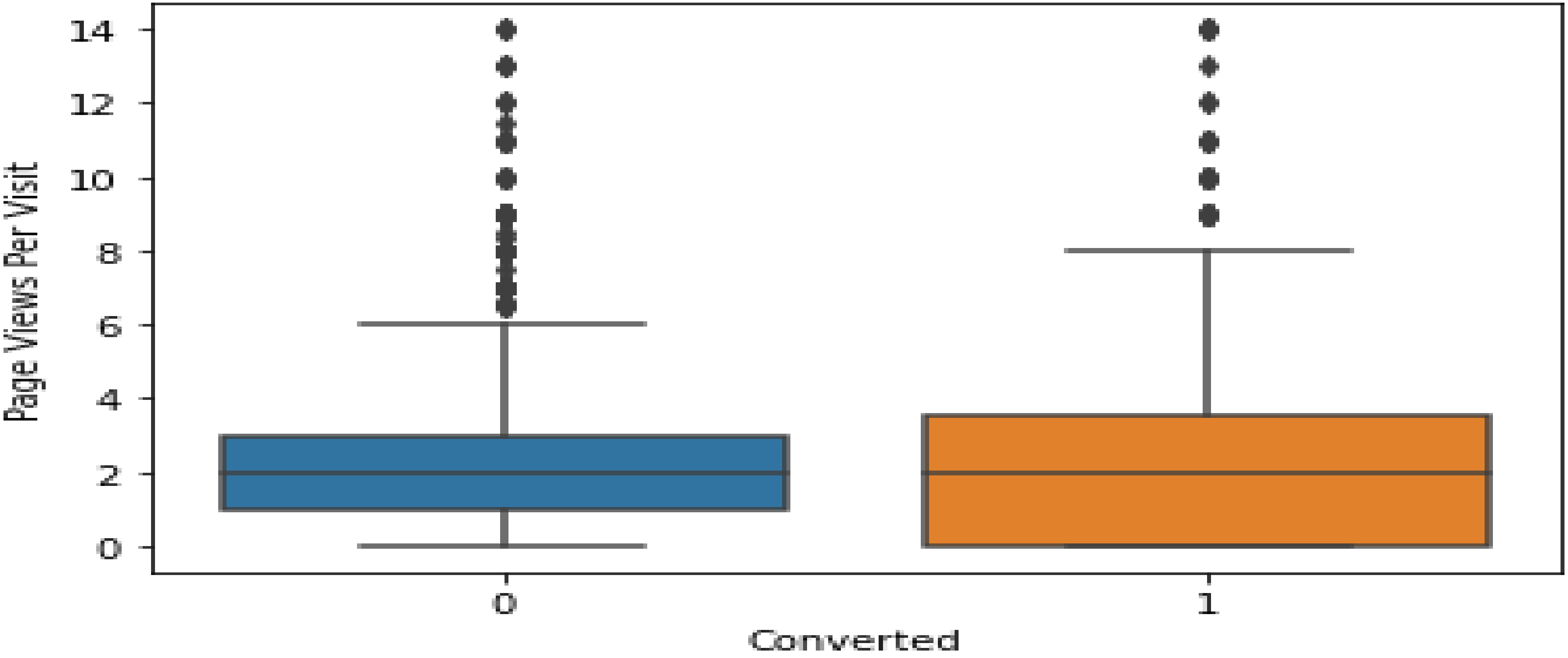
We can see that median or 50% of the converted leads and non-converted leads are same

BIVARIATE-ANALYSIS



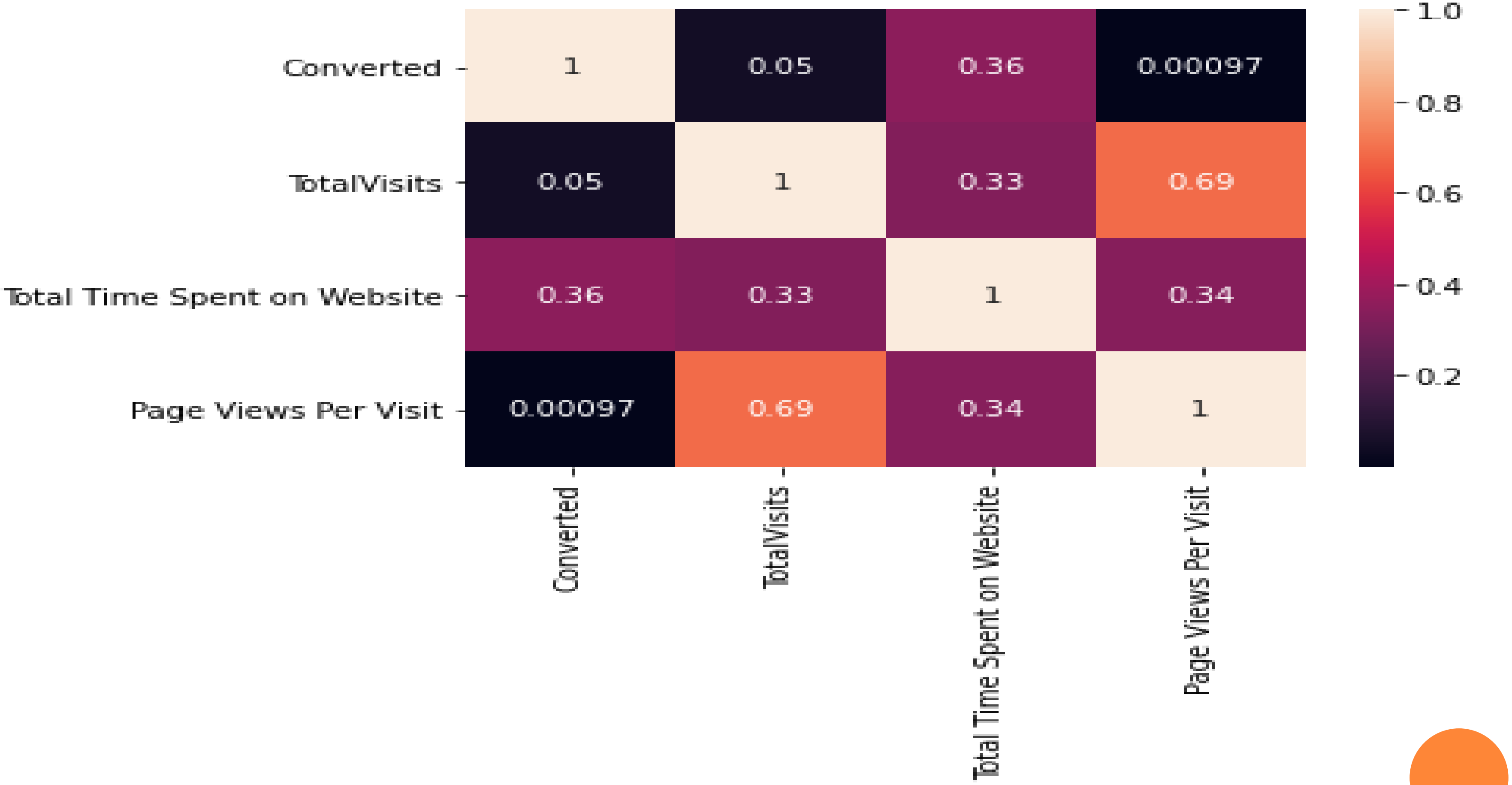
From the above plot we can see that median of the leads who are sending more time on the websites are getting converted more

Bivariate-Analysis



From the above plot we can see that median of the converted leads and non-converted leads due to 'Page Views Per Visit' is same.

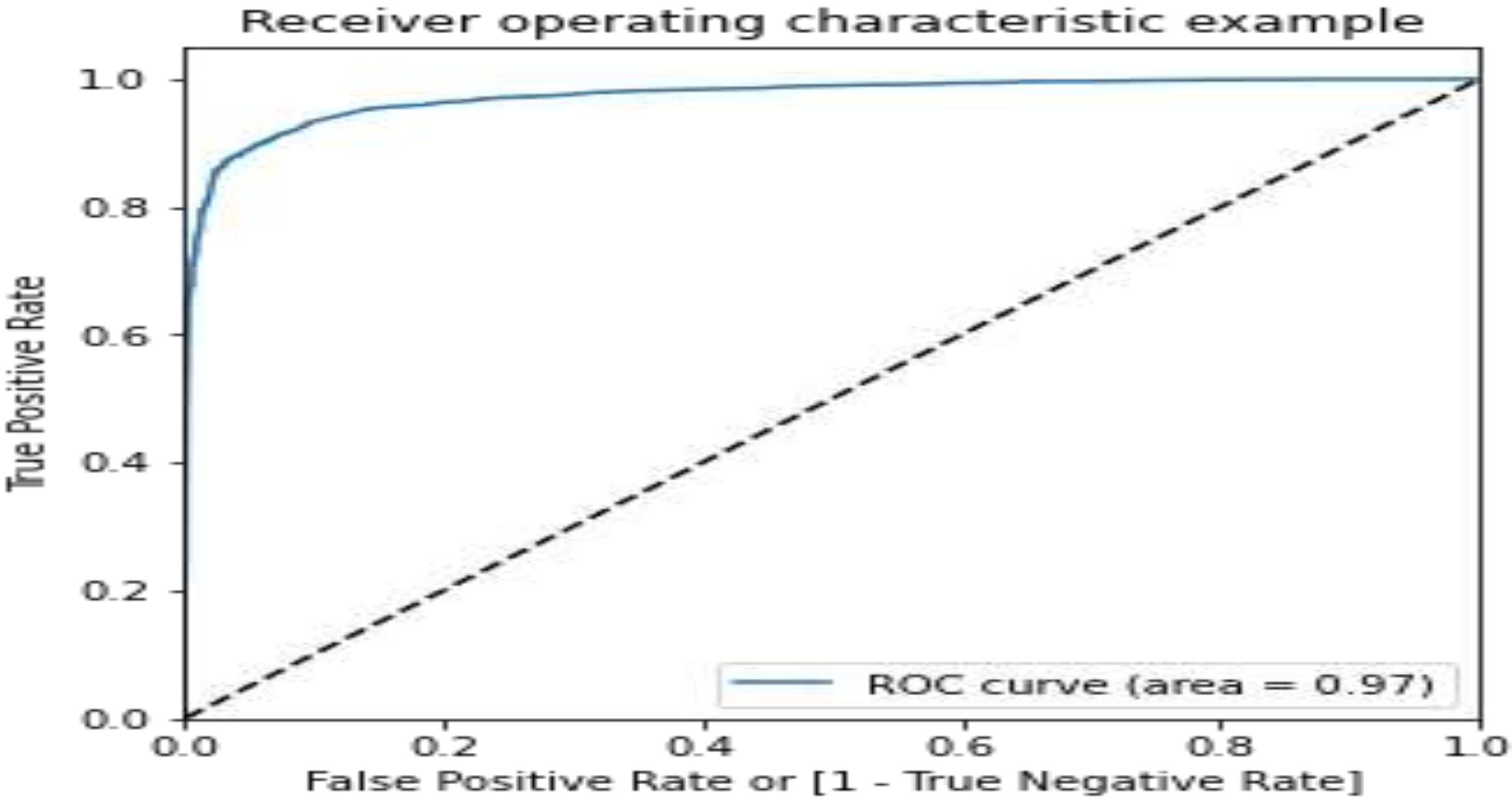
BIVARIATE-ANALYSIS(HEAT MAP)



DATA PREPARATION FOR MODEL

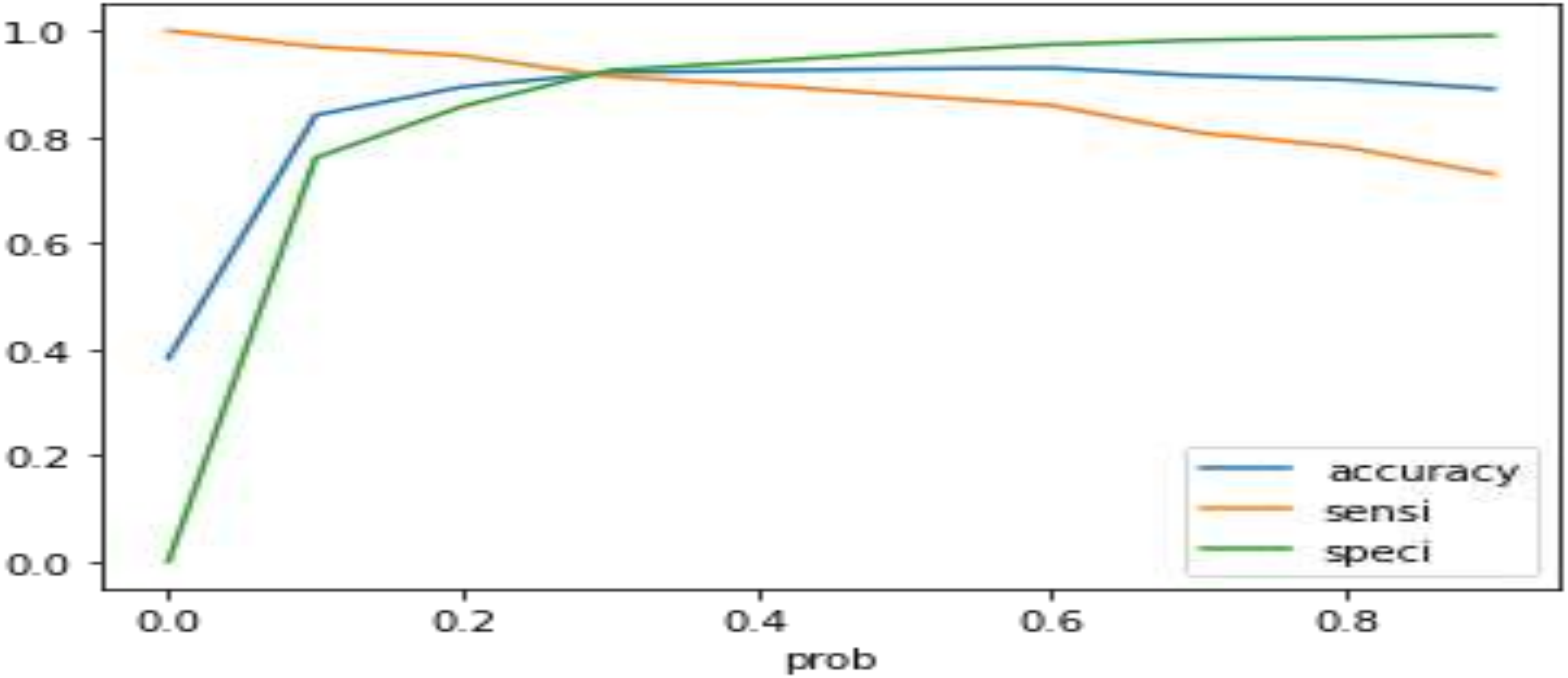
- Since the logistic model can be built only with numerical data, we used dummy variable creation.
- After the data preparation and cleaning we moved to the step of model building. We first started with label encoding process where the binary value yes/no were converted into 1's and 0's.
- Then splitting the dataset into train and test.
- Then standard scaling was performed to bring all the numerical value on the same scale.
- To reduce the no of columns used RFE.

ROC CURVE



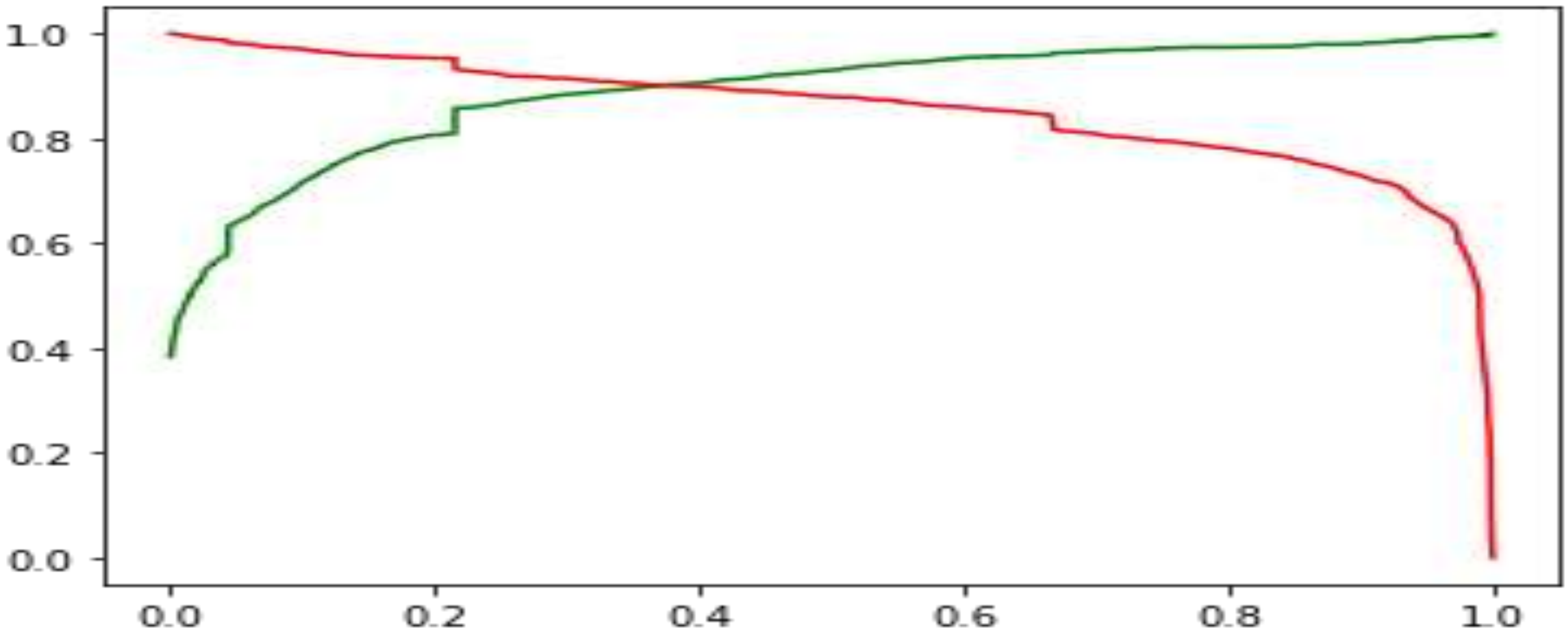
An ROC curve demonstrates several things:
It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test

FINDING THE OPTIMAL CUT-OFF



From the curve above, 0.3 is the optimum point to take it as a cutoff probability

Precision and recall tradeoff



Model Evaluation:

- Once the model was built it was evaluated against the unseen or test dataset and with the help of below metrics it was confirmed that the model built was good enough to help X company to have more lead conversions.
- Accuracy:92%
- Sensitivity:90%
- Specificity:93%
- Precision:88%
- Recall:90%

Conclusions and Recommendations

- Though we have plotted both sensitivity-specificity and precision-recall but we have considered sensitivity-specificity to calculate the final prediction .
- Accuracy, sensitivity and specificity calculated on the trained dataset are 92%,91% and 92% respectively after taking the cut-off as 0.3.
- Accuracy, sensitivity and specificity calculated on the trained dataset are 92%,90% and 93% respectively.
- Since the values are close we can say that model built is efficient enough.

Conclusions and Recommendations

- Top 3 variables contributing most towards leads based on the model built are:
  - Total Time Spent on Website
  - Tags
  - Lead Source
- Top 3 categorical/dummy variables in the model and based on the EDA /visulaizations which should be focused the most on in order to increase the probability of lead conversion are:
  - Lead Origin_Lead_add_form
  - What is your current occupation_working_professional
  - Last activity_sms_sent