Presentation title

Lead Scoring Case Study

Prepared by -
AYUSH SAXENA
BHRANTI DESAI
AVNEET SETIA

Agenda

The Purpose of the case study is to create a Logistic Machine Learning Model which 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.

Business Problem

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 markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although 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 Objective

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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance

Dataset on which the analysis is performed

We have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

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

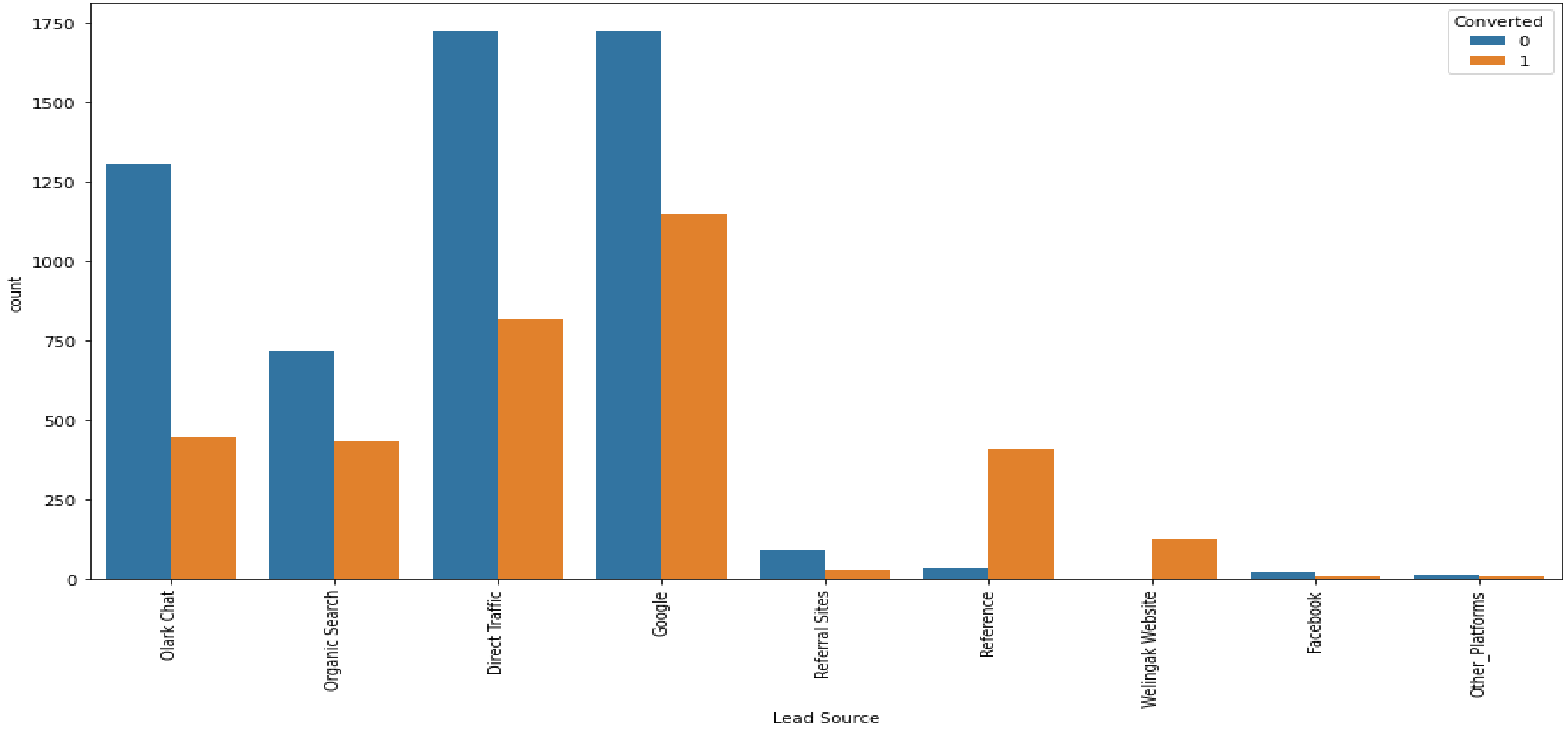
Approach taken to do Modelling

1. Dataset is loaded using Pandas libraries.
2. Data cleaning Performed on Application dataset.
3. Null values percentage is calculated and columns having more than 60% Null values is removed
4. Columns which doesn’t seems to be important and not at all related to TARGET variable are also removed
5. Imputation of null Values is done:
6. Exploratory Data Analysis is then performed to see the importance of various features and how they are related to target. Relationship between of features with respect to target column has been done and inferences have been taken out.
7. Features that doesn’t seems to be correlated with the target has been removed after this EDA.
8. Data Preparation for the model is then done. Finally the data is then divided into train and test data in a 70:30 ratio.
9. Scaling is done on Training Data set.
10. Feature selection using RFE is done and then the model is build.
11. the model is rebuild after eliminating features based on high P-value and VIF.
12. ROC curve is also build and cut off threshold is decided based on that.
13. Model accuracy along with confusion metrics and metrics beyond accuracy such as sensitivity , specificity etc. is calculated on training data and the same has been calculated on test data after doing the prediction on test data
14. Three key features that should be kept in mind are noted after the complete model building and analysis.

Bivariate Analysis on following Columns with respect to TARGET variable:

- Lead Source
- Do not Email and Do not call
- What is your current occupation
- Tags
- Last Notable Activity

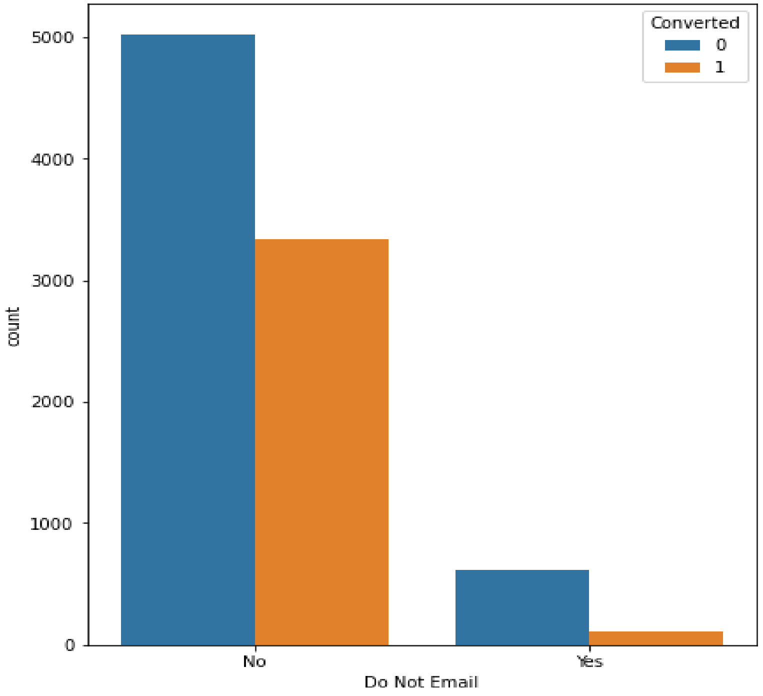
Relationship between Lead Source and Target Variable

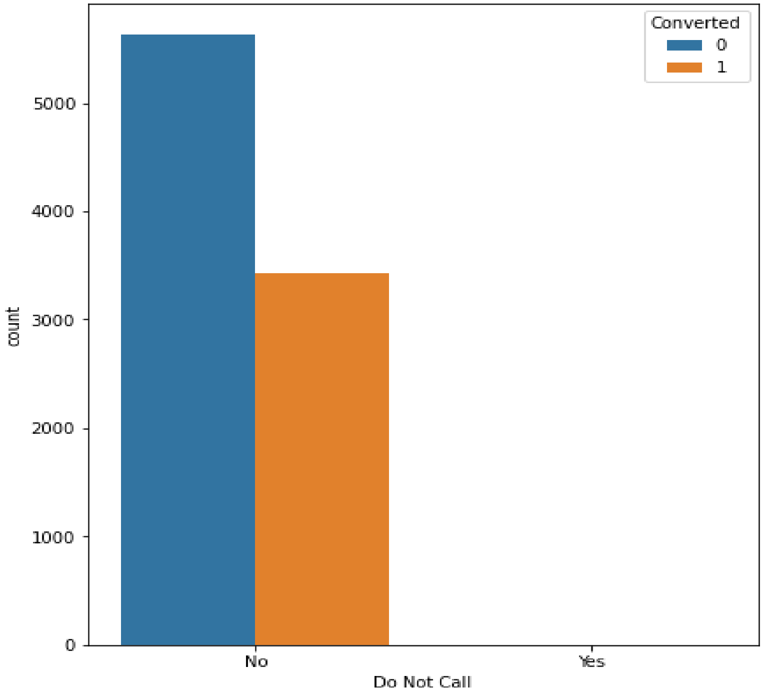


We can conclude from the above graph that –

•Google and Direct Traffic generate the maximum leads

Relationship between Do not Email and Do not call and Target Variable

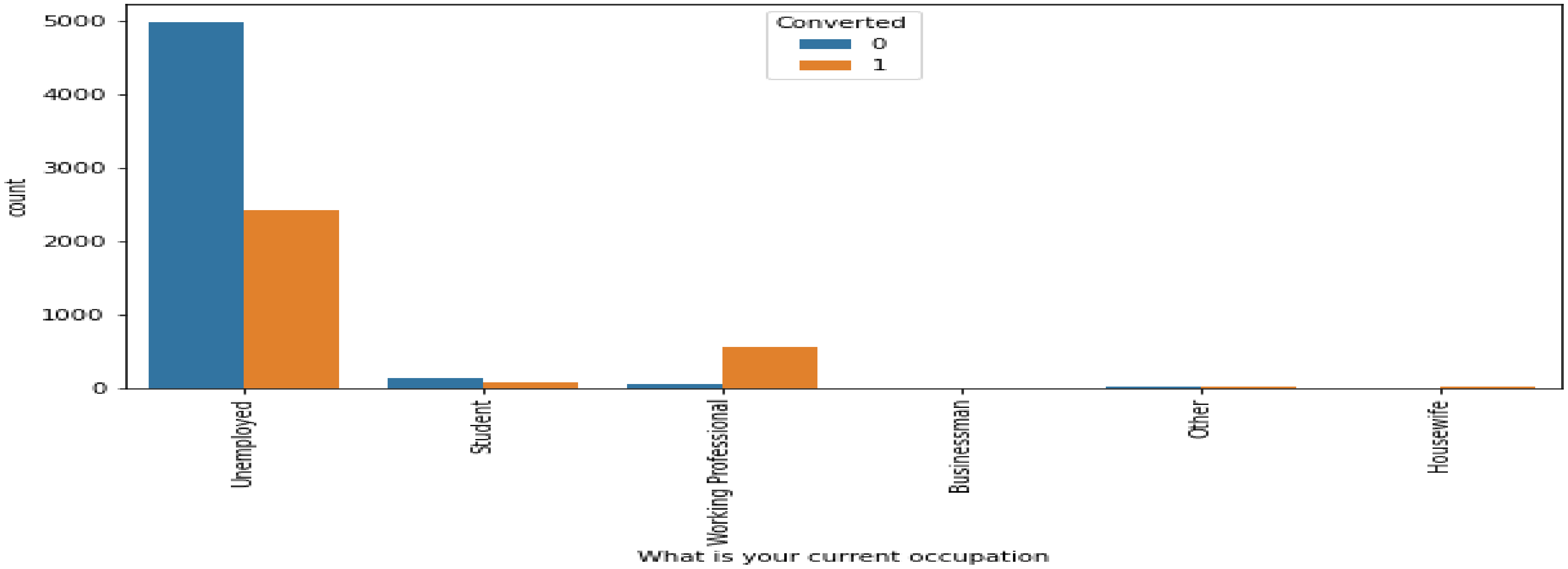




We can conclude from the above graph that –

Mostly people who choose to get contacted by an email and call and there are fair chances of those getting converted

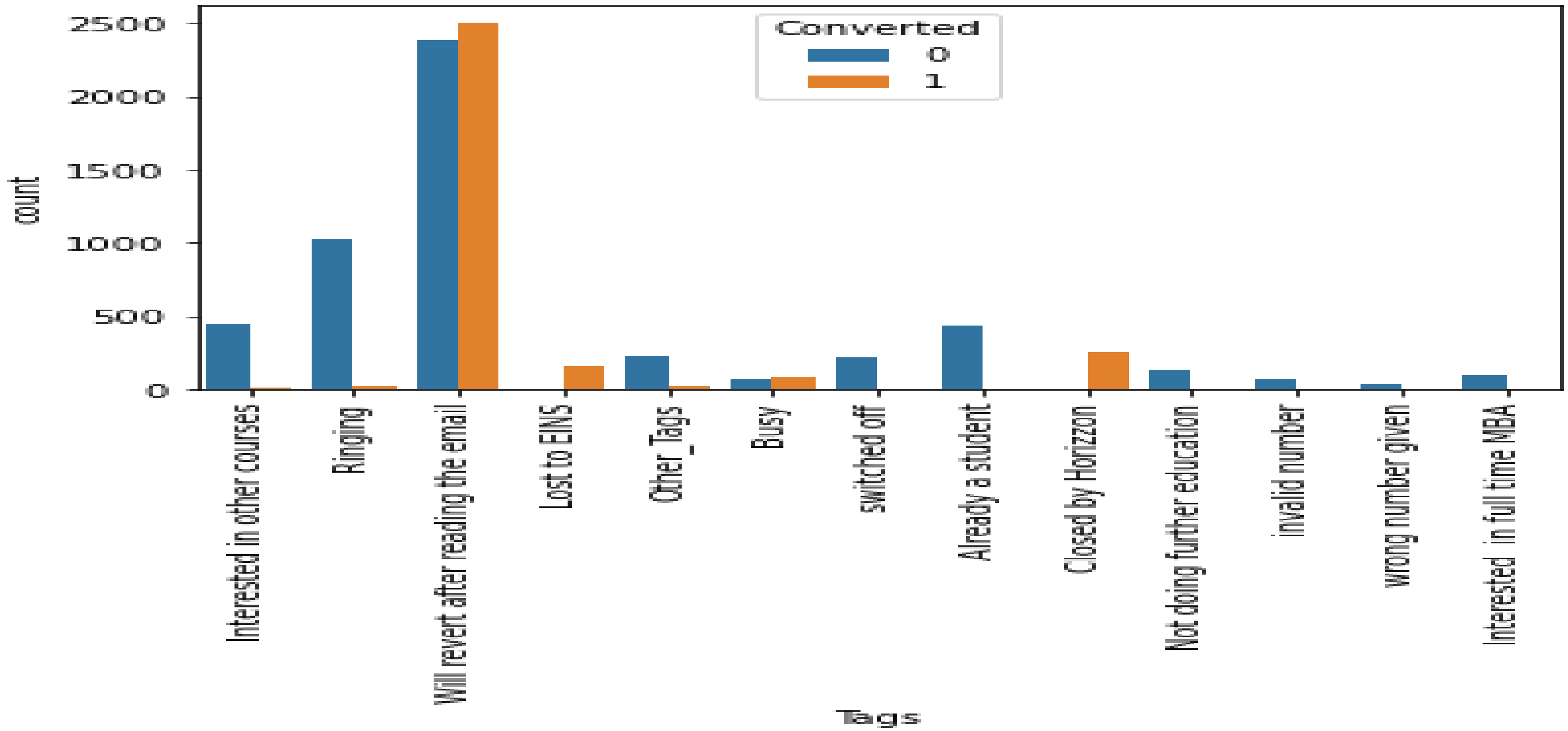
Relationship between Occupation and Target Variable



We can conclude from the above graph that –

Mostly the people those are coming to coming to website are unemployed and willing to do some course for getting placed

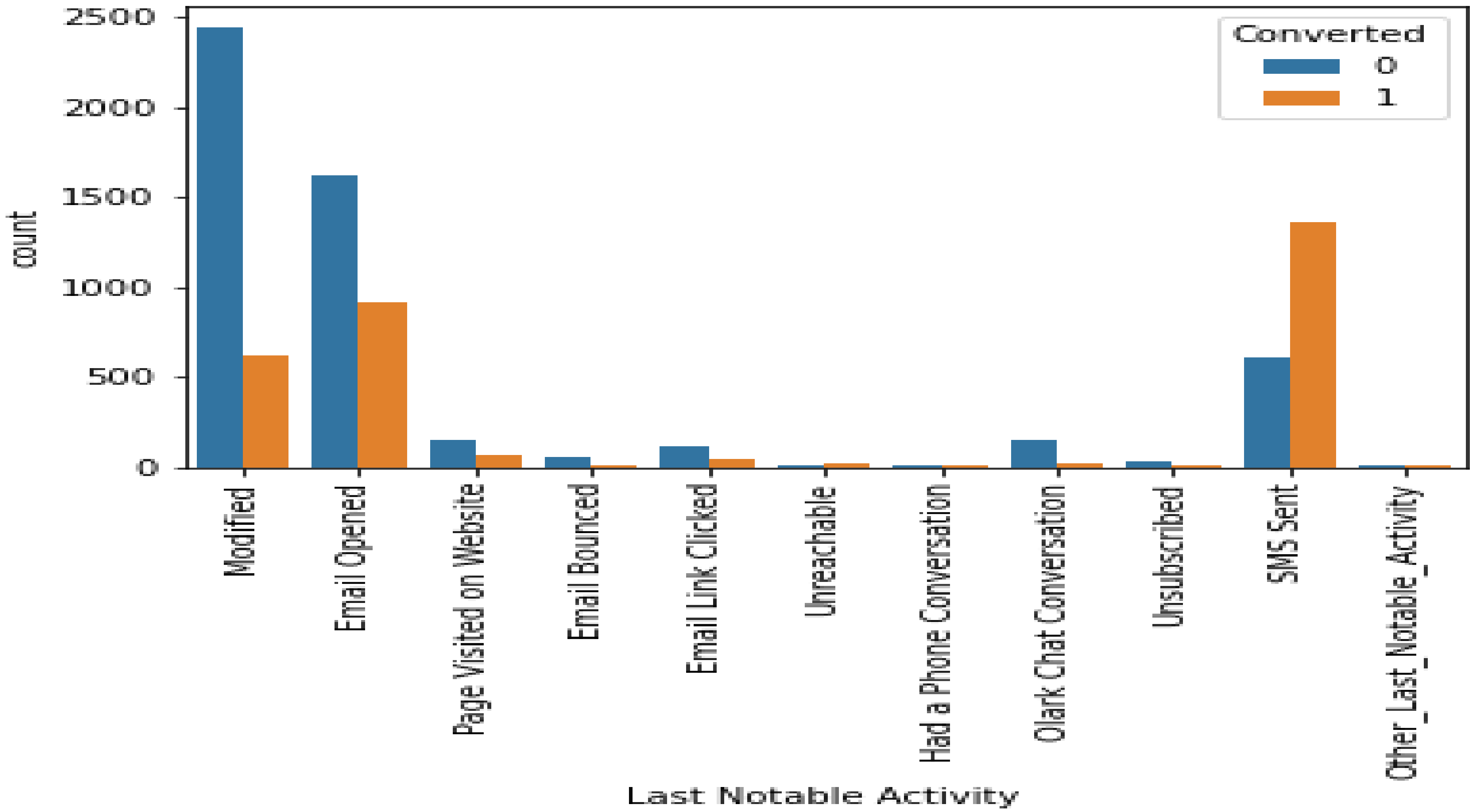
Relationship between Tags and Target Variable



We can conclude from the above graph that –

‘Will revert after reading the email' has higher conversion rate

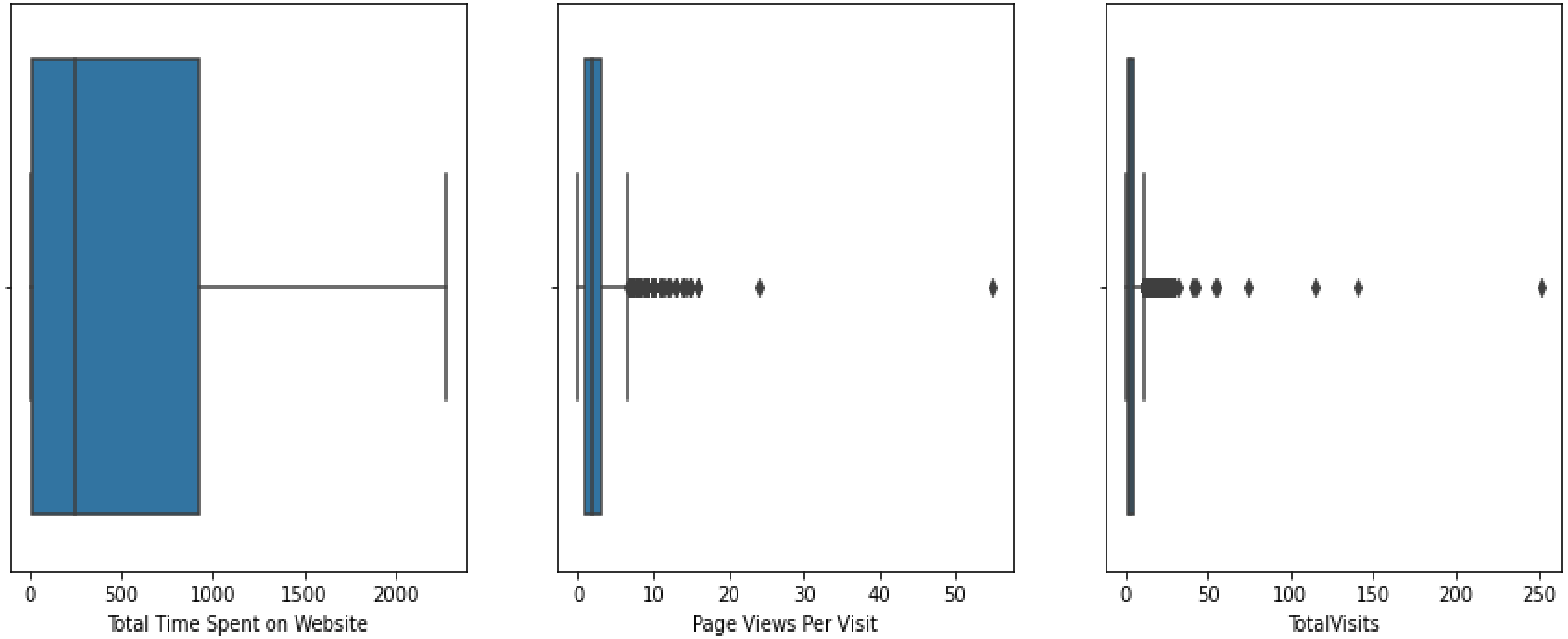
Relationship between Last Notable Activity and Target Variable



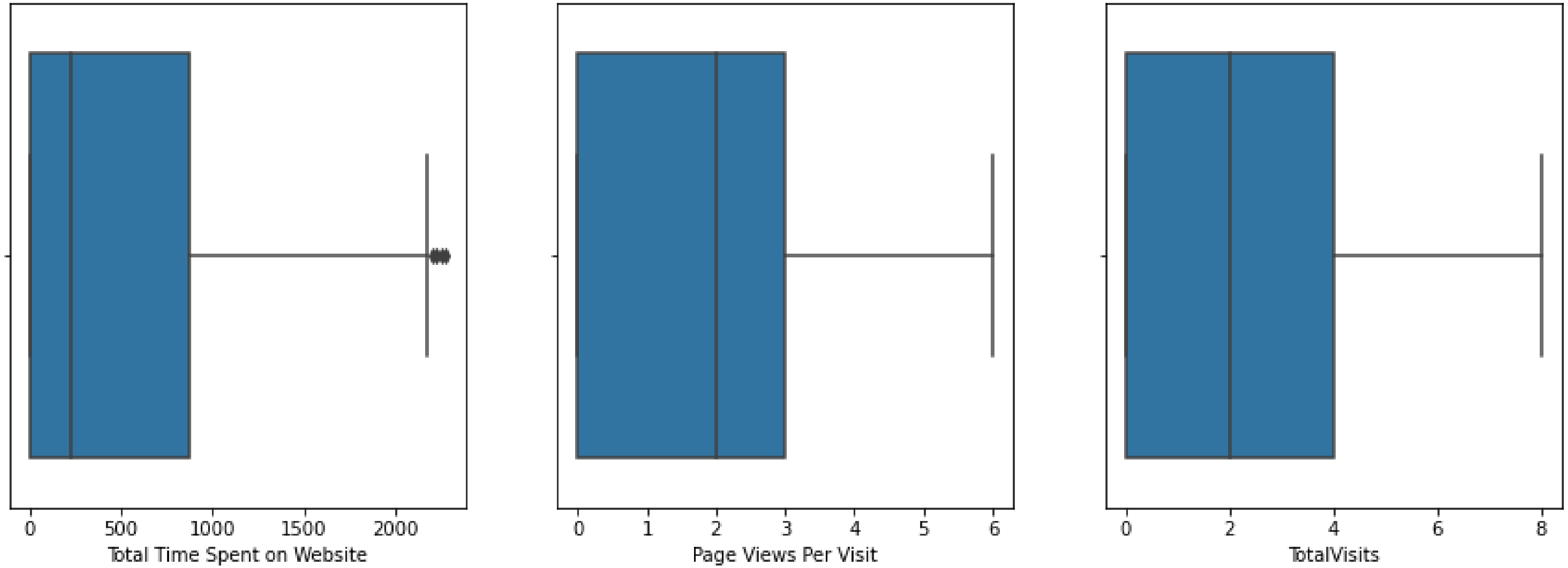
We can conclude from the above graph that –

SMS Sent has a high conversion rayte when compared to other

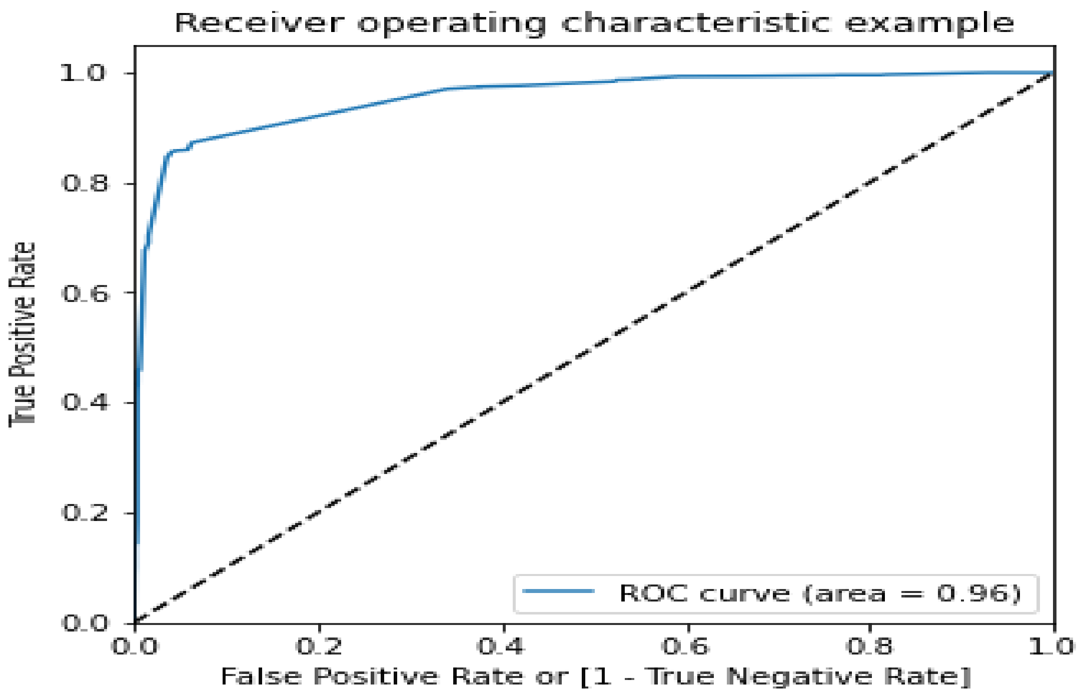
Outlier Analysis :



After Handling Outlier



ROC Curve



- 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.

Logistic Regression Model is Built and following conclusions were drawn:

Our Model has achieved the overall accuracy of 91% and the sensitivity of 86%, precision of around 89% and a recall of around 85% on the test data.

We should focus on following type of customer:

- Who revert after reading the email.
- Who are working professional.
- Who spent time good amount of time on website.

Thank you