

# LEAD SCORE CASE STUDY

Akhila ck

# Problem statement

- 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.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. 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.



# Problem statement

- So we have to find the leads that are most likely to convert into paying customers. The company requires 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.

# About the Data



- We have 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.

# Data preparation

## Null value treatment

- Six columns had null values in excess of 40%. 6 columns are thus immediately dropped.
- Following that, a continuous variable with a median is imputed, and a categorical variable with a mode.
- Unique values in columns have no impact on lead conversion. Therefore, these columns are removed.
- Some unnecessary columns were present in the Lead Source, Last Activity, Tags, and Last Notable Activity columns. These are classified under other categories.

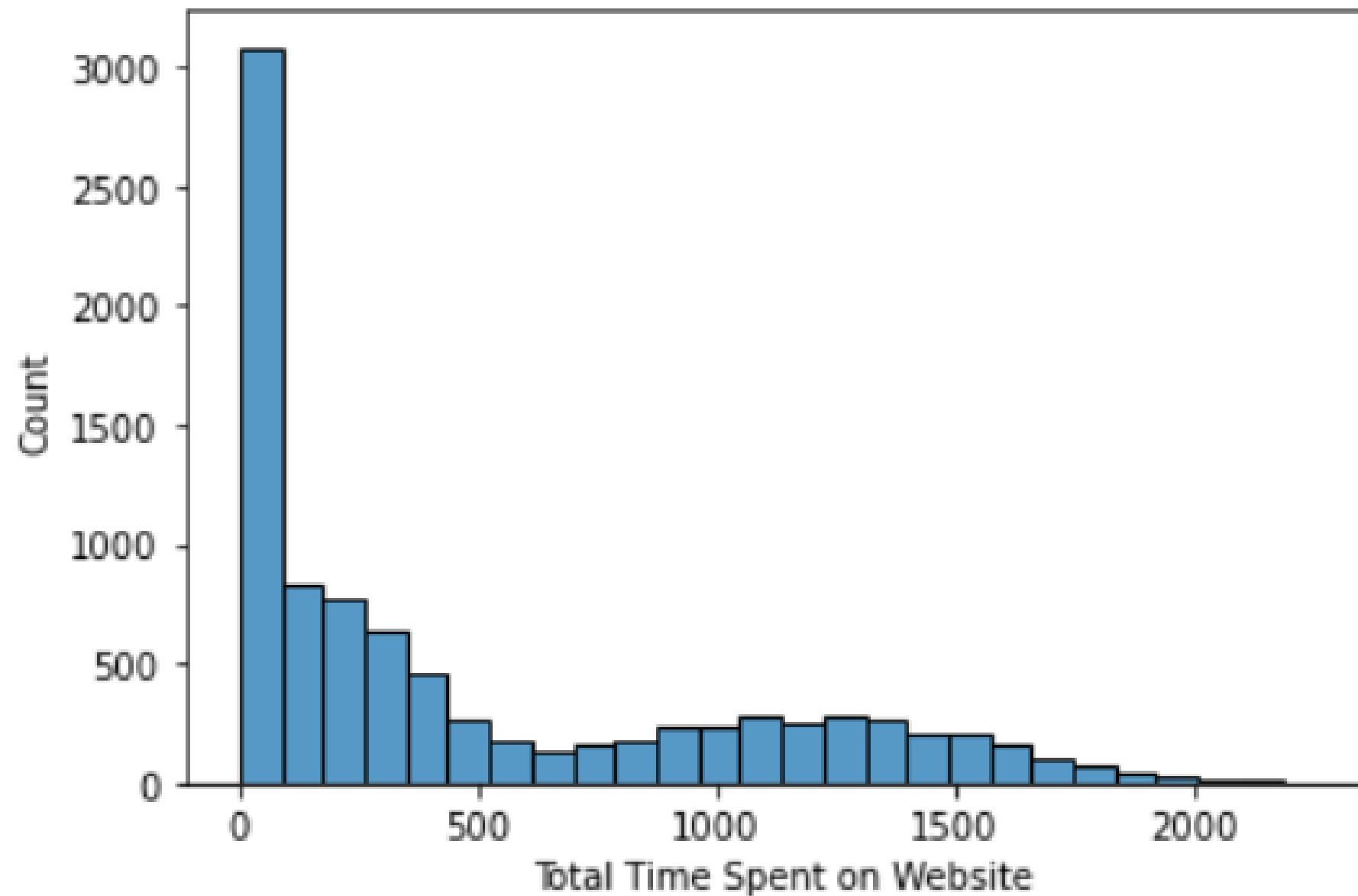


# EDA

In EDA graph between different variable are plotted to understand the relationship

## univariate analysis

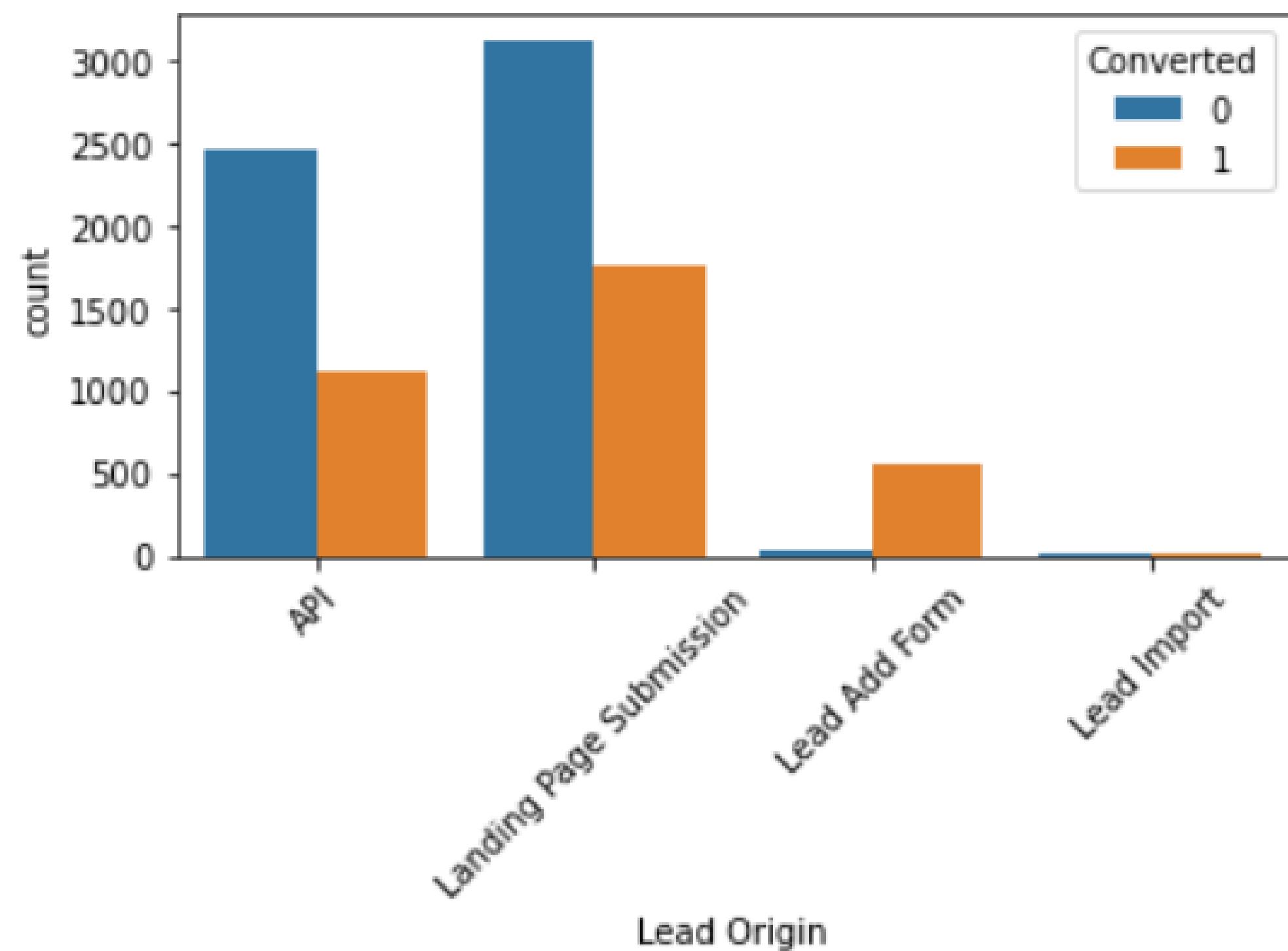
Total Time Spent on Website



- More people are spent less time on website .As the amount of time increases count of people decreases
- That is major part of people visiting the site are spending less time on the website

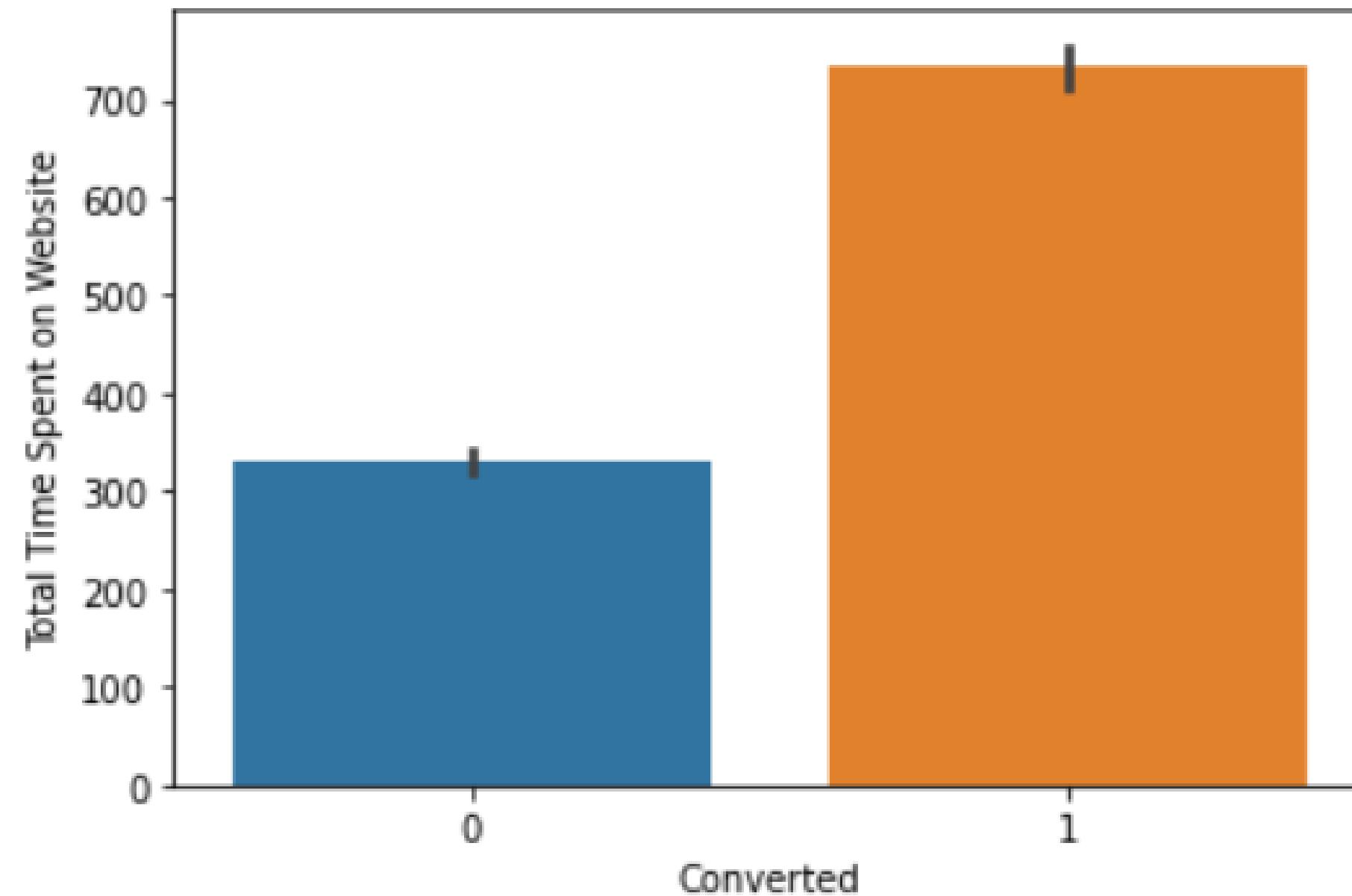
# Bivariate analysis

## Lead Origin



- Major portion of leads are from landing page submission and APIs
- In that major portion of leads are converted
- Lead add form is providing less number of people.but its lead conversion rate is higher

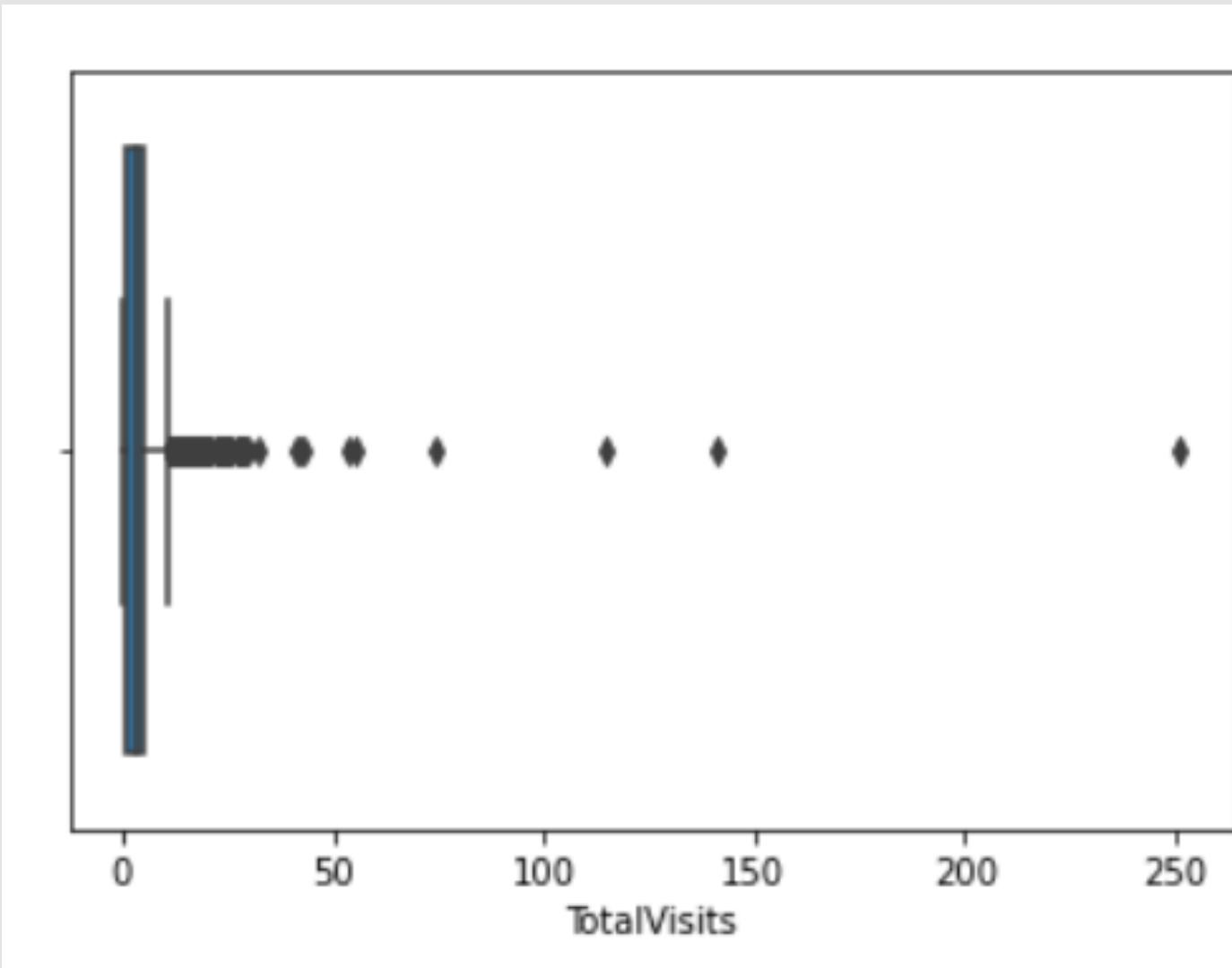
## BarPlot Converted Vs Total Time Spent on Website



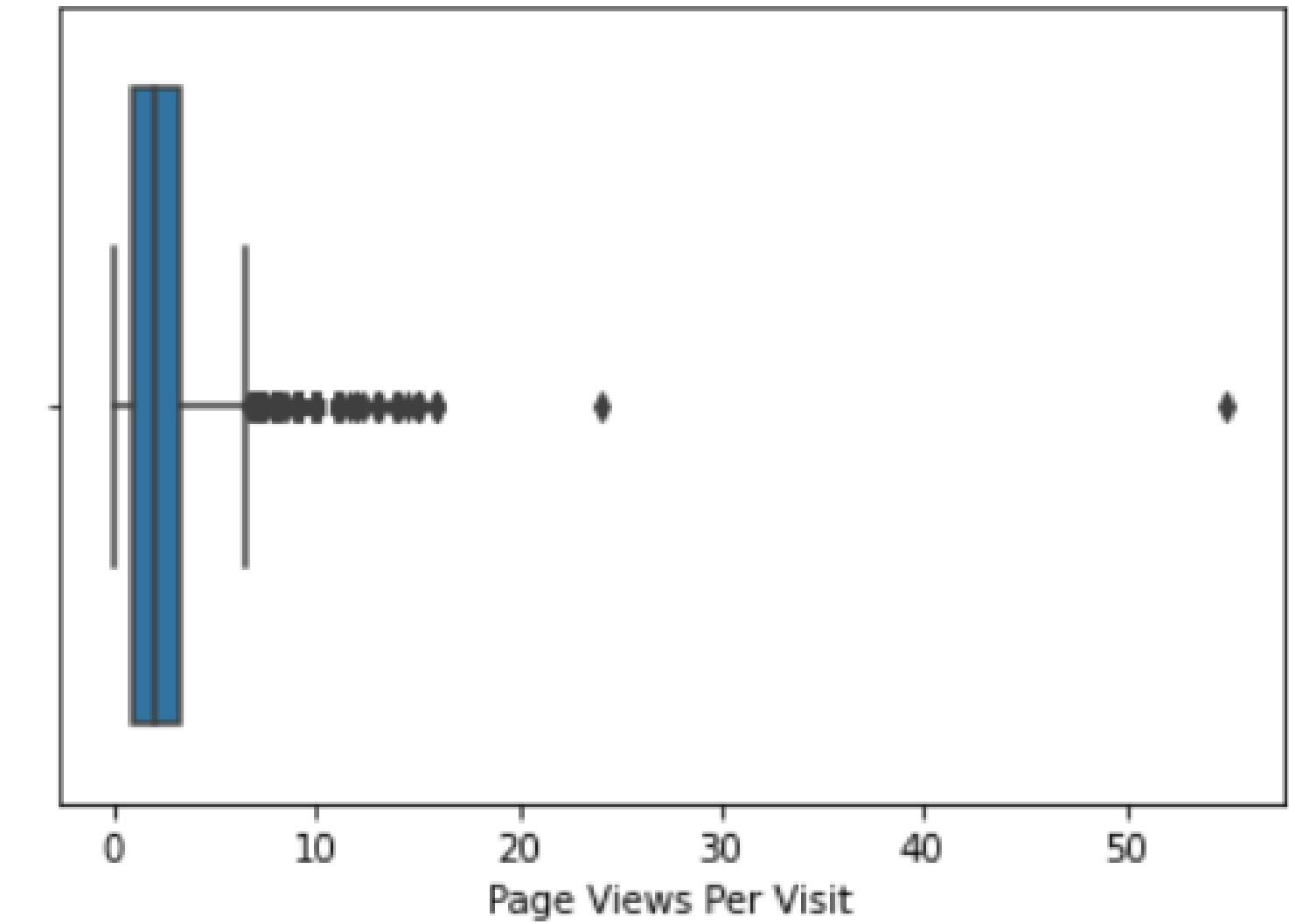
- Lead that spent more time on website is having more conversion rate .So this variable is having major effect on lead conversion

# Outlier Treatment

Outlier treatment is done on continuous variable



Page Views Per Visit



Variable named Total visit and page views per visit is having large amount of outlier .so for this two variable outlier treatment is done using IQR Method

# Dummy variable creation

- Column "Do Not Email" and "A free copy of Mastering The Interview" is having yes/no .theses are converted in to Binary variable
- After that dummy variables are created for some of the categorical variables and first one is dropped
- The categorical variable are 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization','Tags','Last Notable Activity','What is your current occupation"
- After dummy variable creation we have got 74 columns



# Modelling

**Train Test split:** After Data preparation Data is splitted in to train and test in ratio 7:2. Training set is used for train the machine learning model and test set, on the other hand, is used for evaluate the performance of the trained model.

---

**Scalling:** To make the value of continuous variable in range of 0 and 1 scalling step is performed on the data

---

**Feature selection:** Using "sklearn.feature\_selection" 20 variables are sorted out from the all variable. This all steps are repeated in the test data.

---

- Using GLM logistic regression is done on the model
- variable having Pvalue greater than significance level 0.05 are dropped
- variable having VIF greater than 4 are dropped
- probability according to each customer id is calculated
- The value > 0.5 is predicted as converted and value < 0.5 is predicted as not converted

For evaluation confusion matrix is plotted. The value got after evaluation is given below



## **Train\_data**

Overall accuracy:89.84

sensitivity:84.05

specificity:93.42

False positive rate:6.57

positive predictive value:88.79

Negative predictive value:90.43

## **Test\_data**

Overall accuracy:88.69

sensitivity:85.38

specificity:90.86

## **Top three variables in the model which contribute most towards the probability**

- Tags\_Closed by Horizzon
- Tags\_Lost to EINS
- Total Time Spent on Website

## **Top 3 variables that need improvement to convert a lead**

- Tags\_will revert after reading the email
- Lead Origin\_Lead Add Form
- Tags\_Busy

- First three variables are having higher coefficients when compared to the remaining variables.
- The higher the coefficient, contribution towards probability will also be greater.
- According to the Time Spent on Website by the lead, we understand whether they are interested in the course or not. So this will help largely upon the prediction
- By concentrating on the next 3 variables we can make improvements in the conversion rate .



# Reccomendations

- Concentrate on the Top three or five variables in the model which contribute most toward the probability
- Use the predictive model to identify potential leads that have been predicted as 1, indicating a higher probability of conversion. Prioritize these leads as they are more likely to convert.
- Encourage the interns to start calling leads from the highest-potential segment first. Once they have exhausted the calls in that segment, they can move on to the next segment. This approach ensures that the most valuable leads are contacted first and increases the chances of conversions.

- Emphasize the importance of prompt follow-ups. If a lead expresses interest or requests more information, ensure that the interns follow up quickly and efficiently. Set up a system to track follow-up actions and monitor their progress.
- Instead of making immediate phone calls, shift the focus to lead nurturing activities.
- Implement a well-designed email marketing campaign or automated follow-up sequence to provide valuable information, address potential concerns, and build a stronger relationship with the leads.
- Continuously monitor key performance indicators (KPIs) and track the progress of lead nurturing activities and customer support efforts



**Thank you**