LEAD SCORING CASE STUDY

Authors: Ankit Maurya &
Saravanan Kanmani

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals.

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.

PRIMARY GOALS

The company requires to build a model wherein 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

APPROACH FOR THE PROBLEM

DATA PREPARATION

- Read from source
- Clean the data
- Outlier check
- Exploratory Data Analysis

DATA SPLIT AND FEATURE SCALING

- Split data into Test and Train in 70:30 ratio
- Scaling of numerical variables

MODEL BUILDING

- Feature selection using RFE, VIF and p-value.
- Find optimal model using Logistic Regression

MODEL EVALUATION

- Evaluate the final model against test data
- Ensure the model performance on train data

INFERENCE

- Assign the lead score for each leads
- Business descision based on the lead score assigned

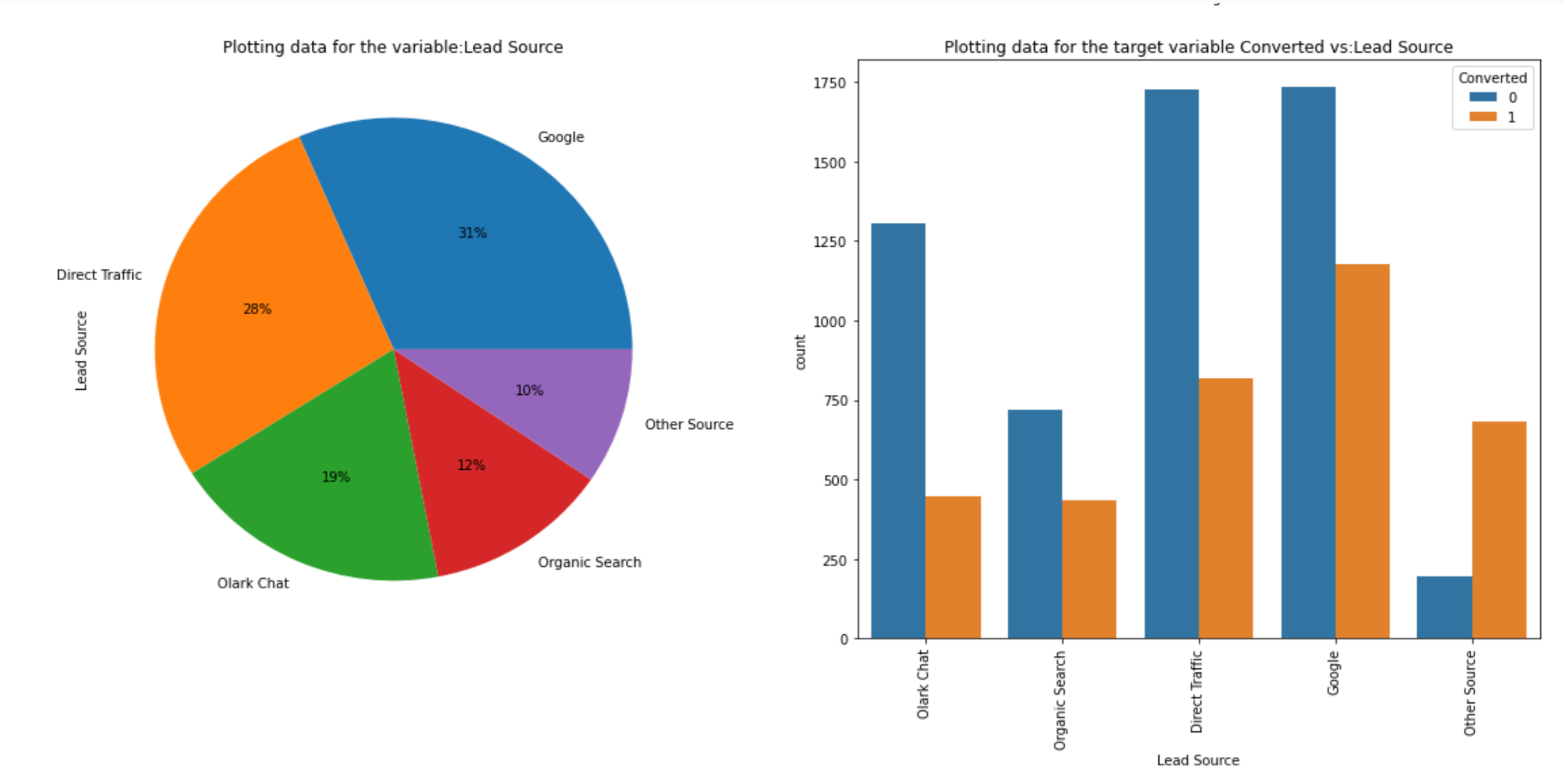
DATA PREPARATION

- Shape of the dataset is 9240 rows and 37 columns.
- Default value 'Select' handled as null.
- Features with more than 35% of null values and contains one unique value are dropped.
- Missing values in few variables are replaced with 'Not Provided.'
- Some of the categorical variables values are bundled together as 'Others'

EXPLORATORY DATA ANALYSIS

UNIVARIATE

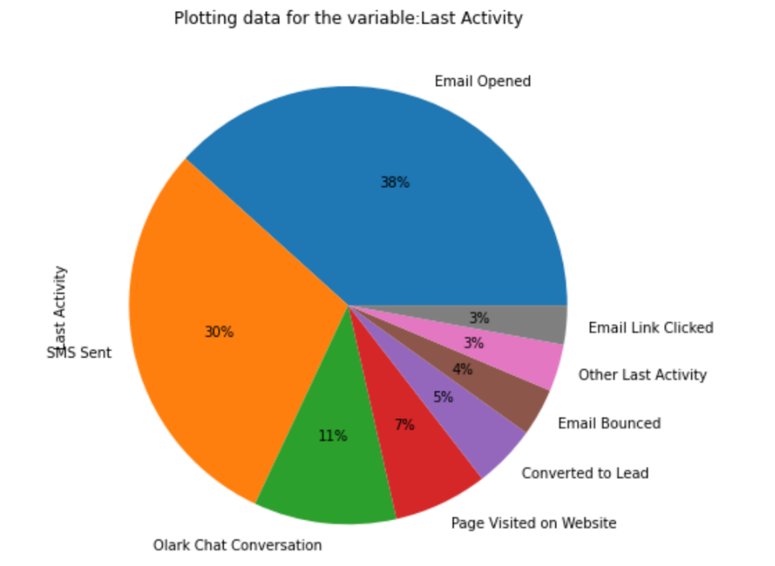
- Some of the important features

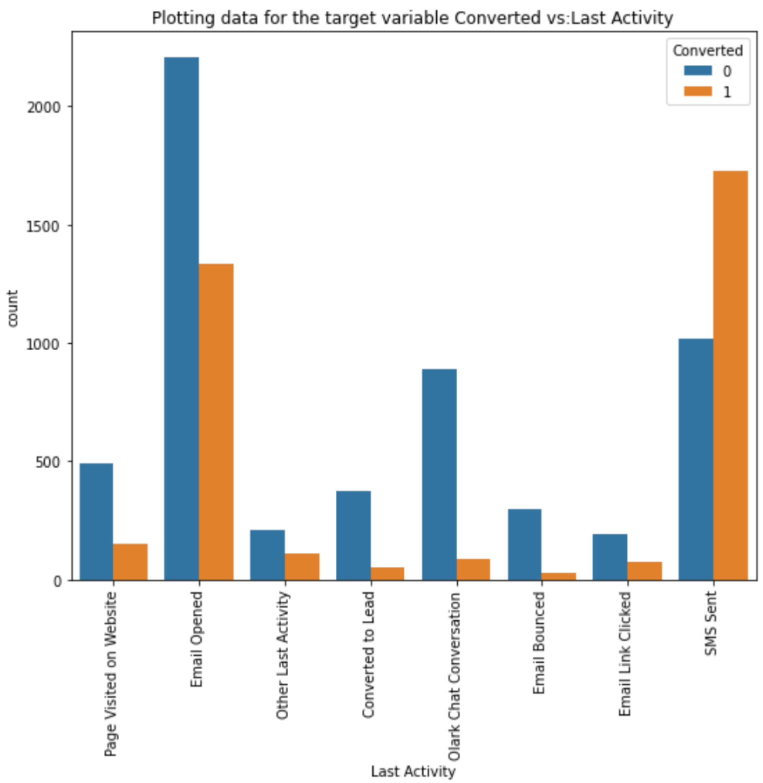


EXPLORATORY DATA ANALYSIS

UNIVARIATE

- Some of the important features

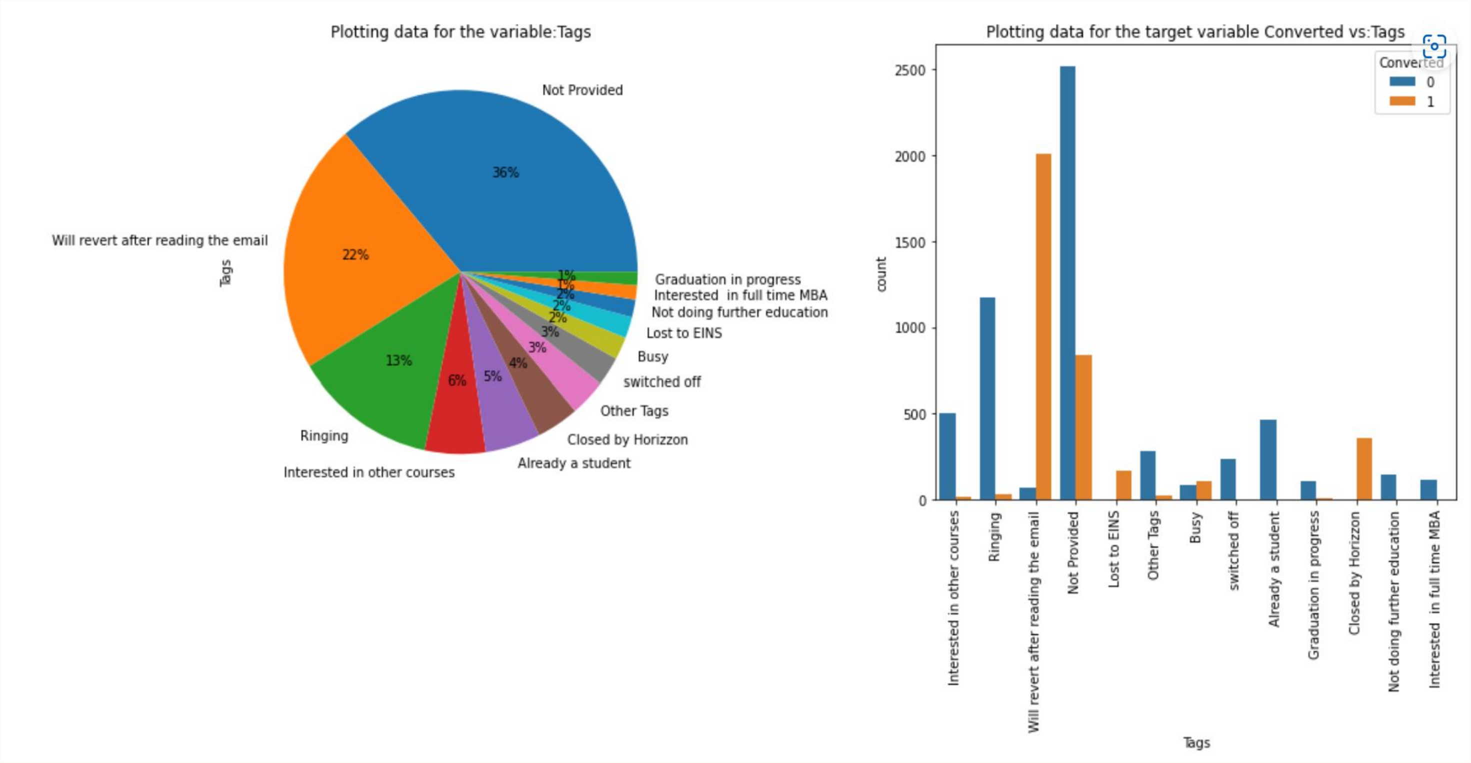

Plotting data for the variable:Last Activity


Plotting data for the target variable Converted vs:Last Activity

EXPLORATORY DATA ANALYSIS

UNIVARIATE

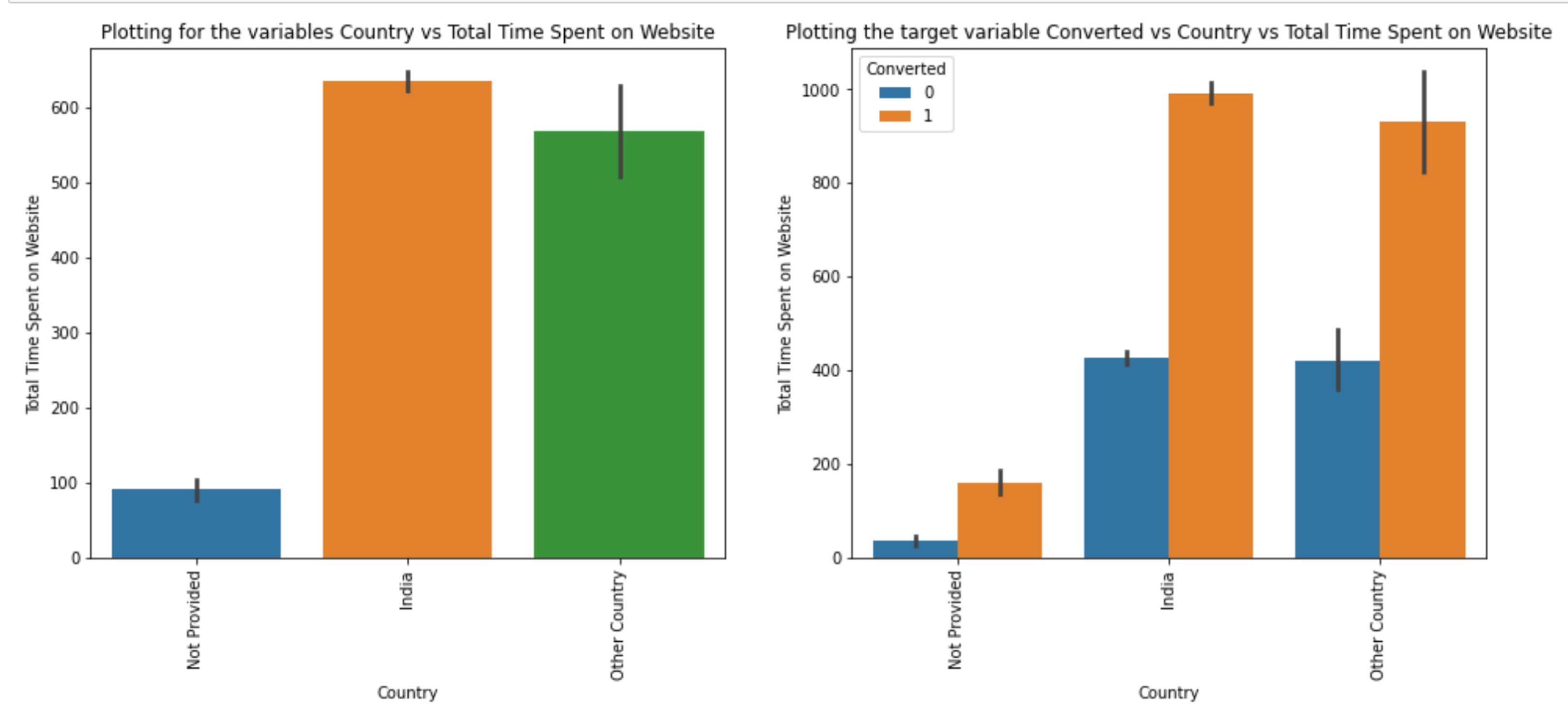
- Some of the important features



EXPLORATORY DATA ANALYSIS

MULTIVARIATE

- Some of the important features



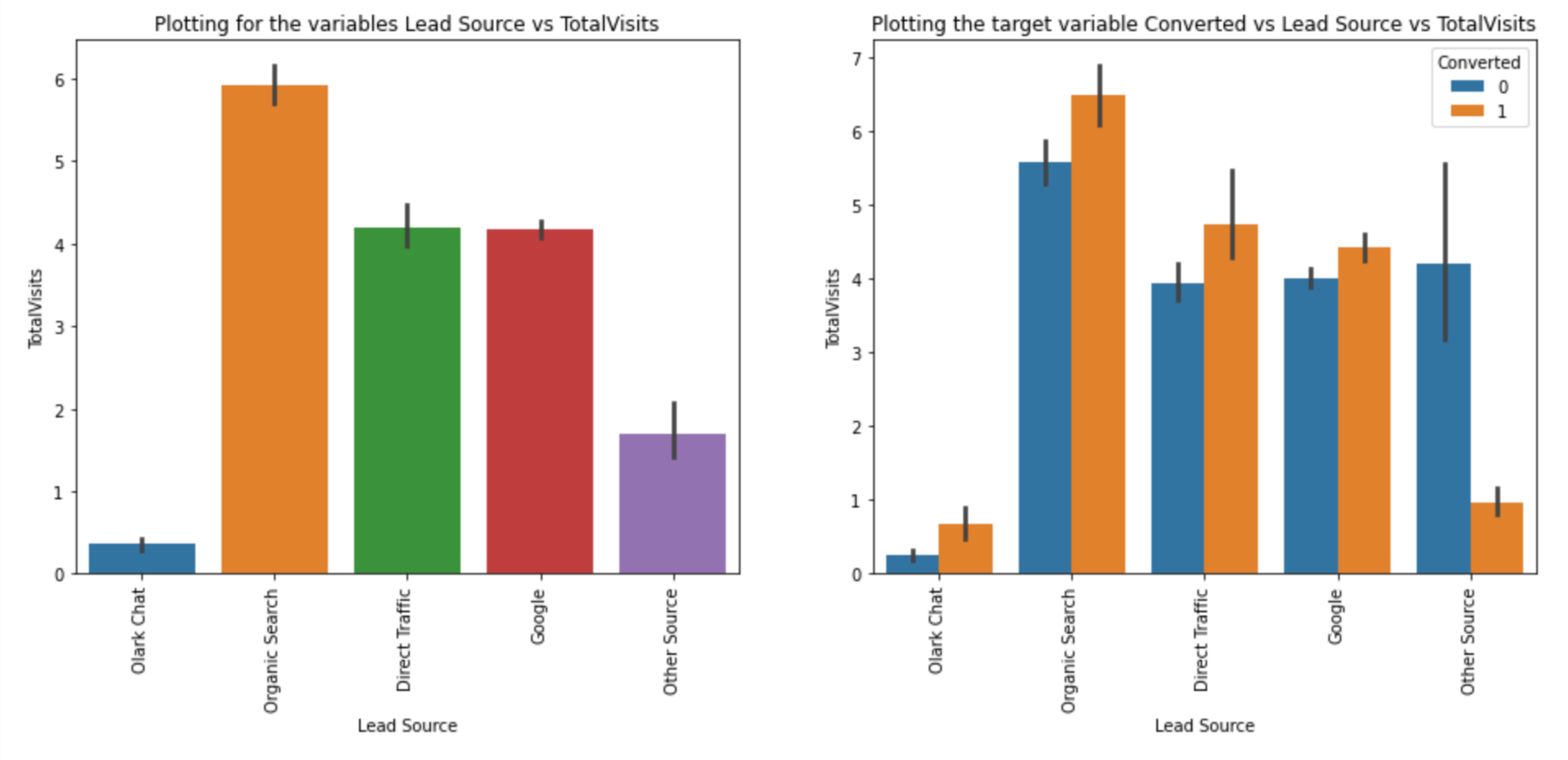
Plotting for the variables Country vs Total Time Spent on Website

Plotting the target variable Converted vs Country vs Total Time Spent on Website

EXPLORATORY DATA ANALYSIS

MULTIVARIATE

- Some of the important features



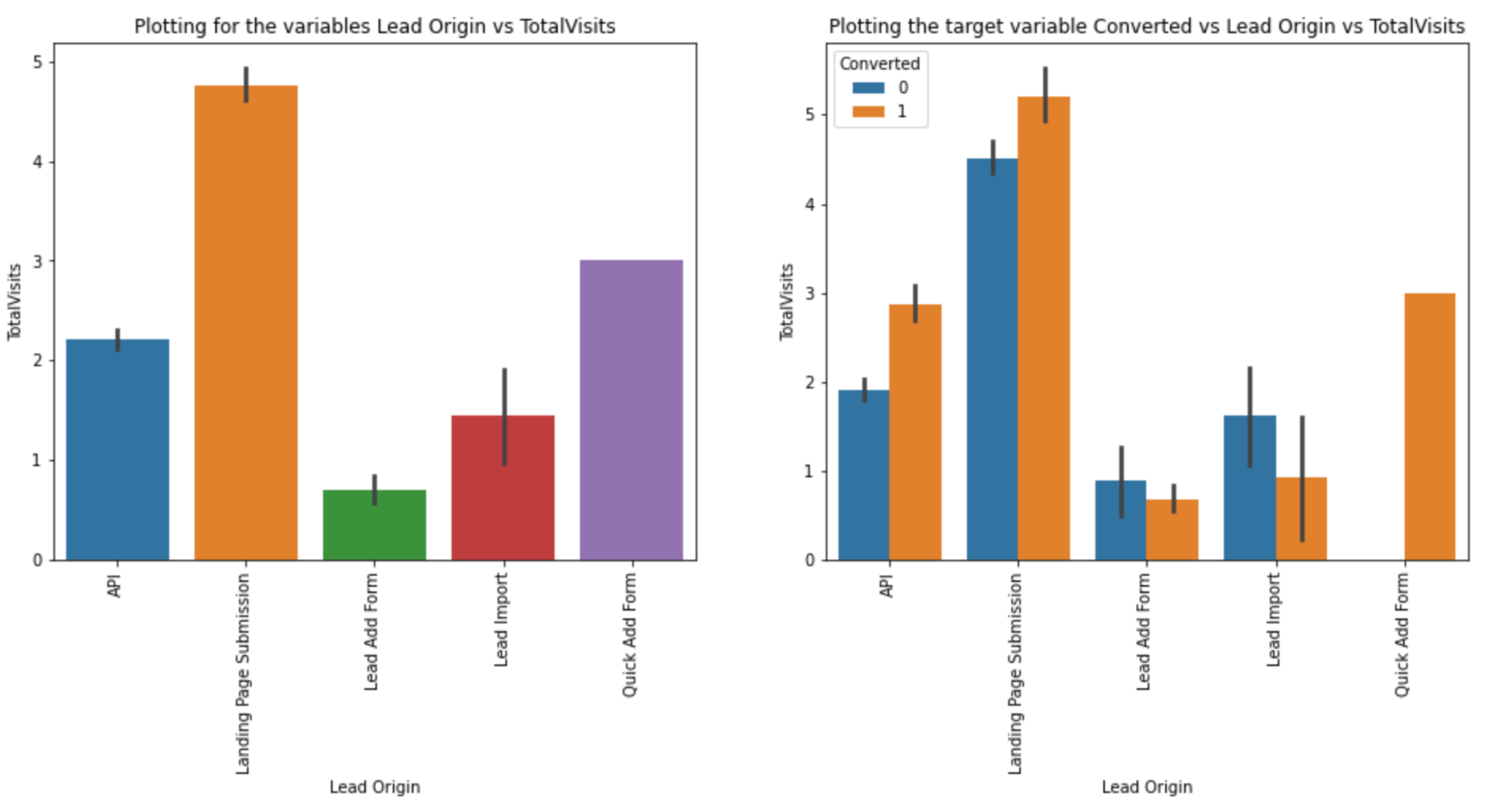
Plotting for the variables Lead Source vs TotalVisits

Plotting the target variable Converted vs Lead Source vs TotalVisits

EXPLORATORY DATA ANALYSIS

MULTIVARIATE

- Some of the important features



DATA CONVERSION

- Normalized numerical variables.
- Dummy variables are created for the categorical variables.
- Test-Train data splitted based on 70:30 ratio.

MODEL BUILDING

- Used RFE for feature elimination.
- Running RFE with 15 variables as output.
- Building model by iteratively by eliminating the features that have high p-value and VIF greater than 5.
- Evaluated the model against test data.
- Overall accuracy is approximately 90%.

SUMMARY

X-Education should focus on few aspects to convert the leads. They are:

- Total Time Spent on Website
- Tags_Will revert after reading the email
- Last Activity_SMS Sent
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_switched off

THANK YOU