LEAD SCORING CASE STUDY

BY: TUSHAR JOSHI & APOORVA BHATLA

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 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%. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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.

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

This summary report includes the steps that were followed to solve the business problem and the conclusions drawn.

Reading and Understanding the data.

We loaded the data in the notebook, and went through all the columns, as per the data dictionary.

Then we do the initial inspection in terms of checking the shape, data types, null values, duplicates and numerical and categorical columns

Data Cleaning and Preprocessing

We first inspect for the ‘select’ word in the data frame. And convert it into NaN

Then column wise inspection of null Values, and drop the columns with more than 30% null values

Drop Highly Skewed Values, and inspect the categorical columns and Numerical columns and impute them appropriately by replacing missing values with mode, or dropping in case of very less missing values

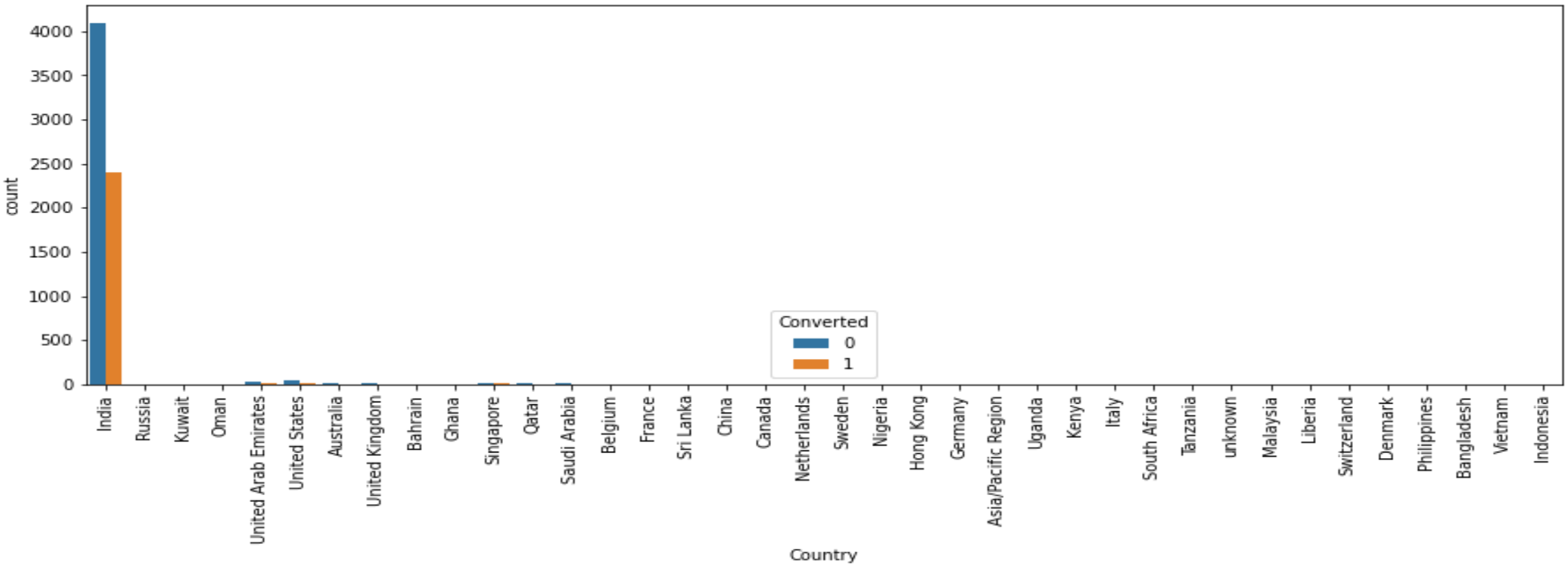
Perform EDA on how each variable plays a part in converting a lead. And Outliers analysis

Drop the columns created by the sales team.

Convert the Yes/No columns with 1/0 and drop the original columns

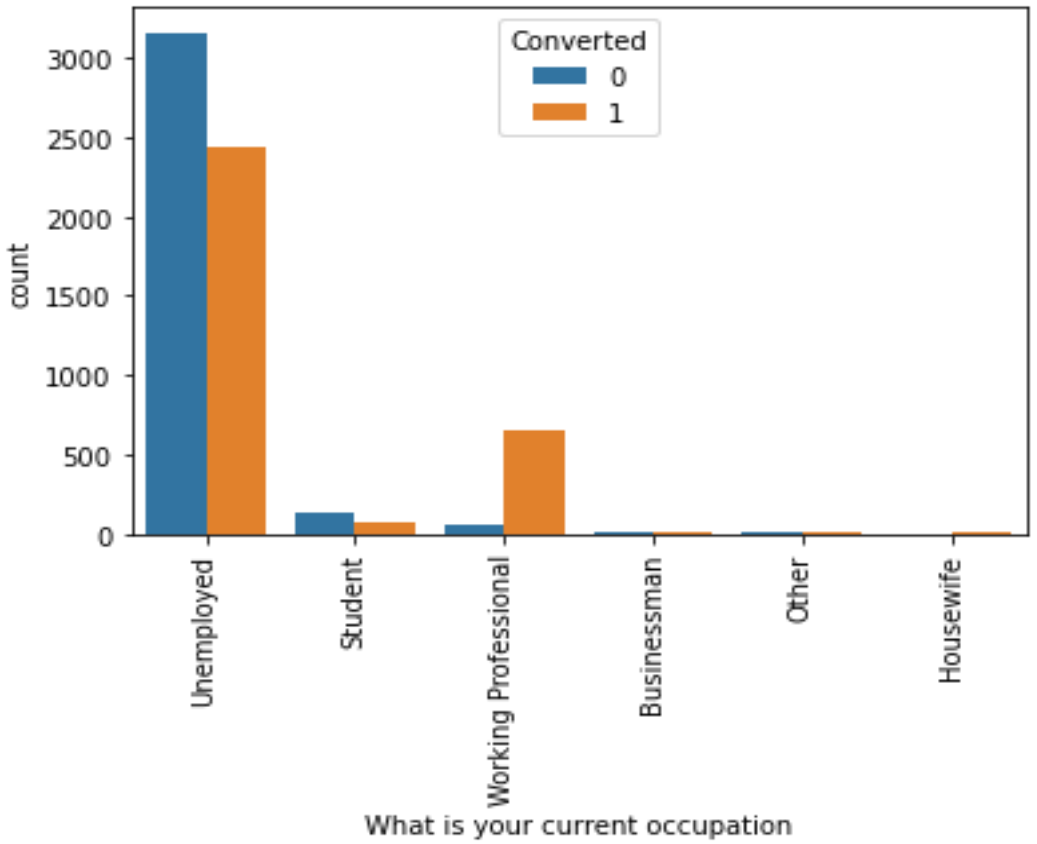
Create Dummy Variables and confirm if the Dataset has Null Values

VISUALISATION OF COUNTRY



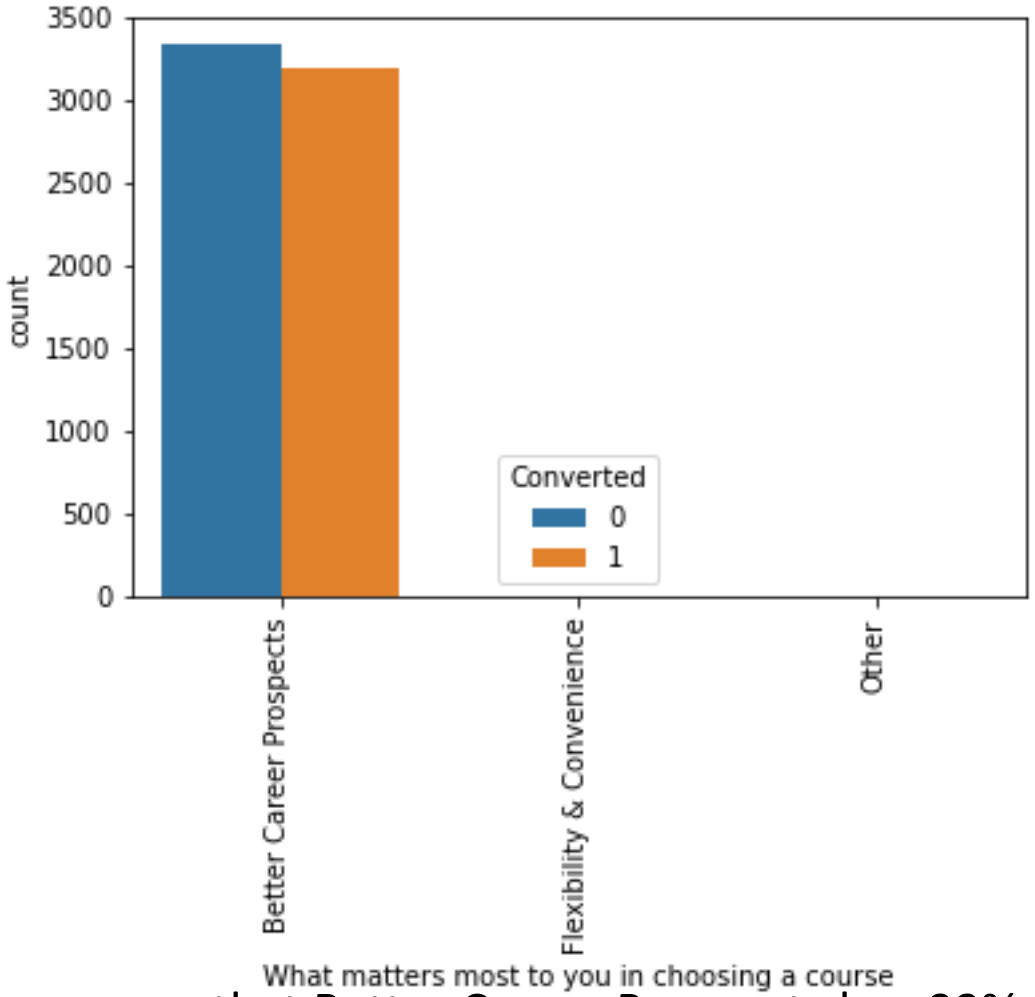
- We see that the country column has most of the values around 95% as India, Country is a highly skewed column, we can drop them, but we can also group the other countries into variable "outside India."

VISUALISATION OF CURRENT OCCUPATION



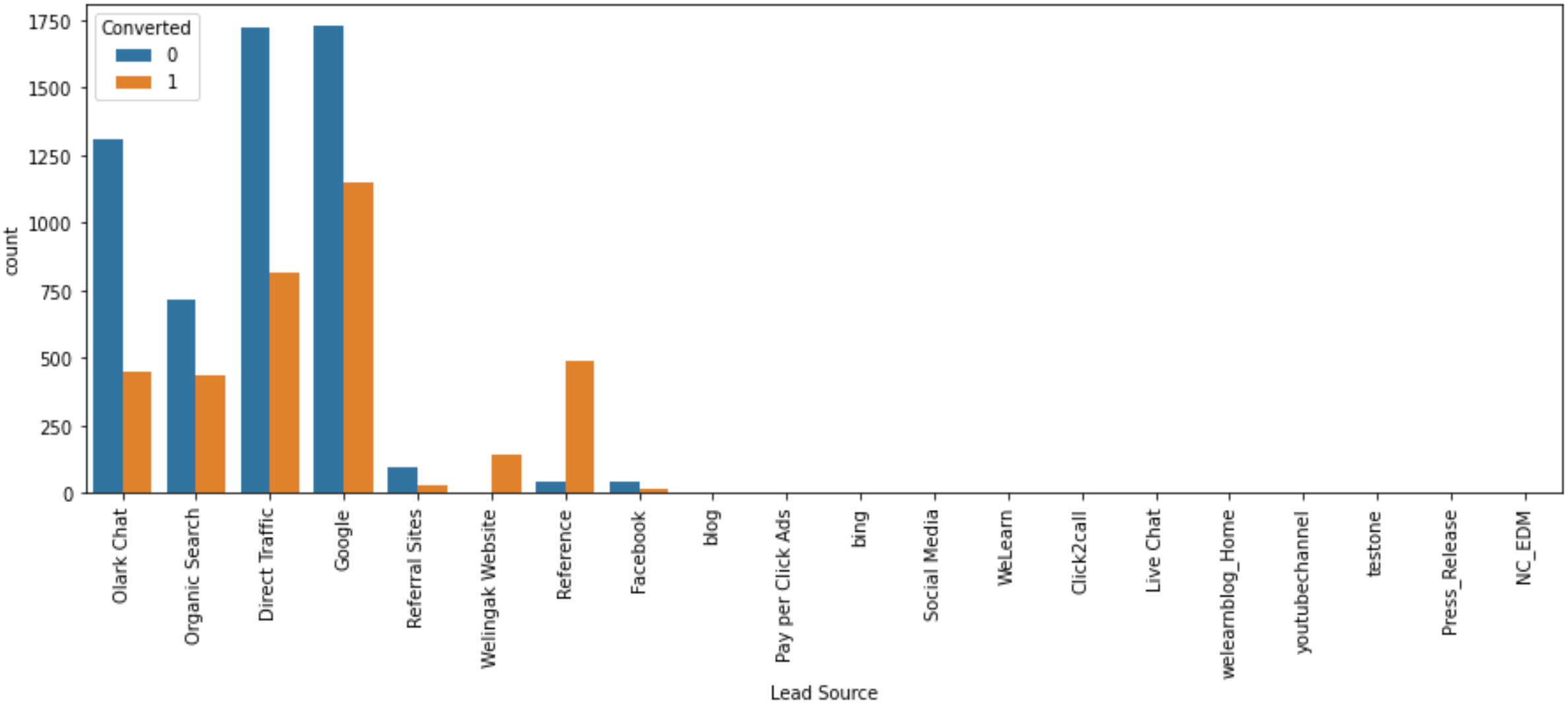
- We see that the Unemployed and Working Professional are significant in number while other features are very less. So grouping the other features to a new group Other Professionals.

VISUALISATION OF EMPLOYMENT



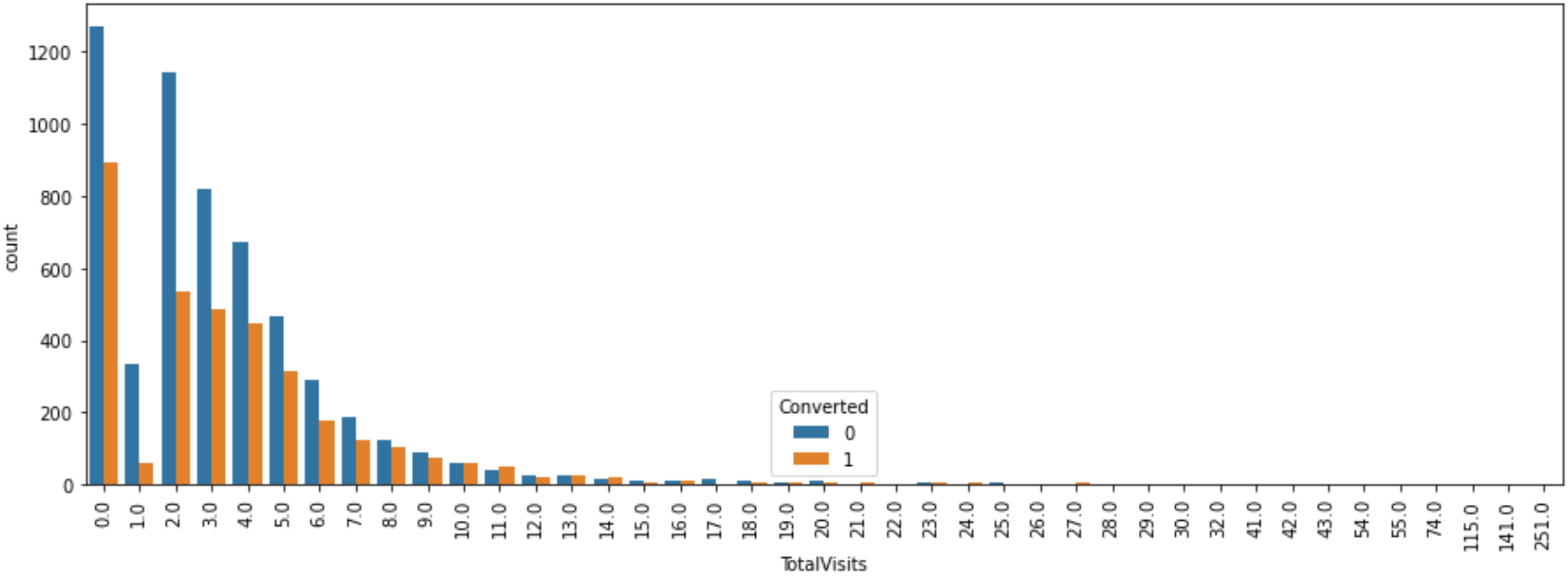
- Inspecting the column we see that Better Career Prospects has 99% presence in the column, thus making it a highly skewed column. So, we can drop it.

VISUALISATION OF EMPLOYMENT



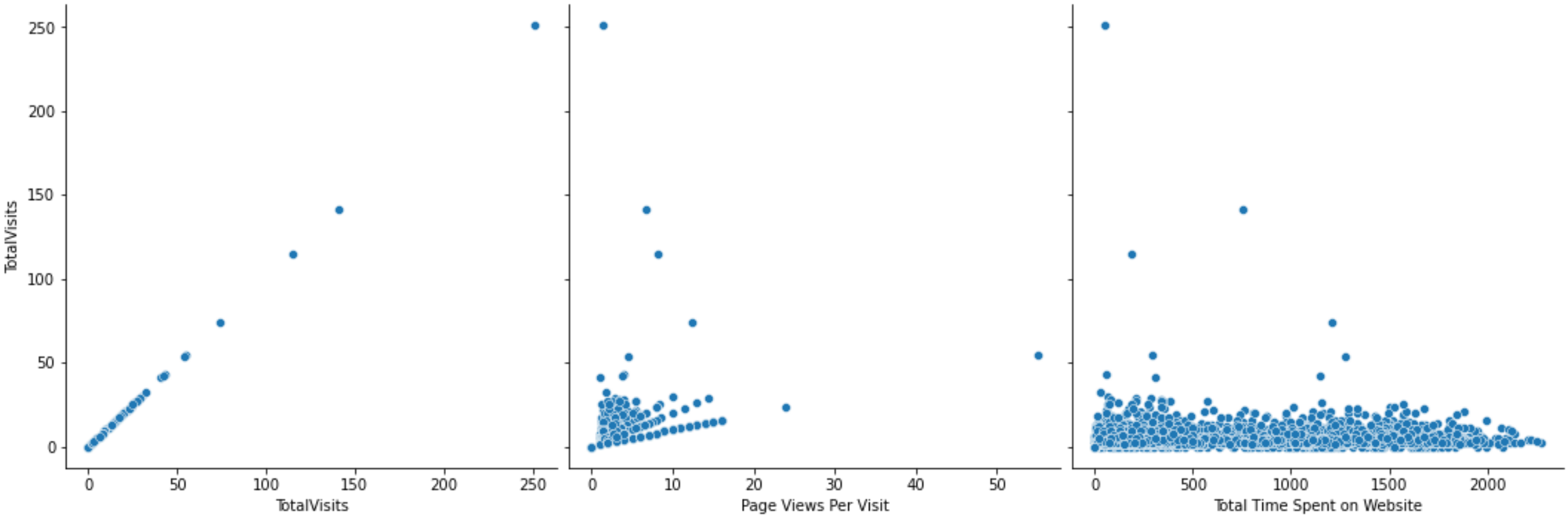
Inspecting the graph we see that Google, Olark Chat and Direct Traffic brings in the most visitors, followed by facebook and Refernces

VISUALISATION OF VISITS TO THE SITE

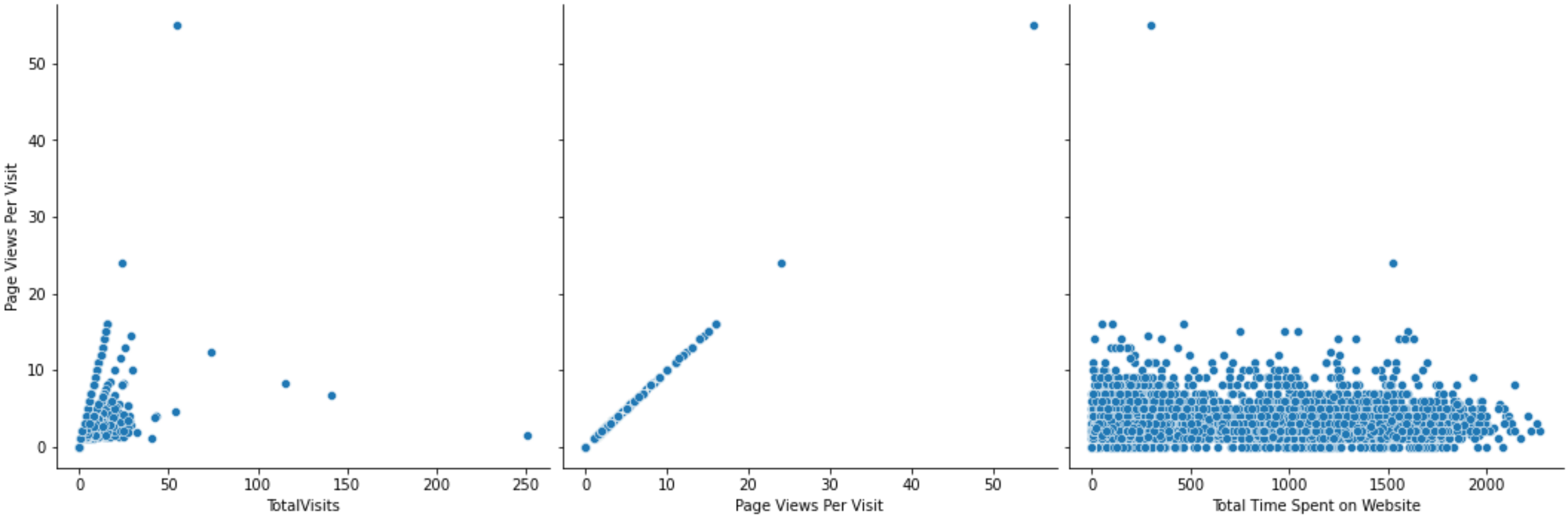


We see that a lot of visitors that got converted without even visiting the website, and a significant number of visitors were converted who visted between 2-7 times.

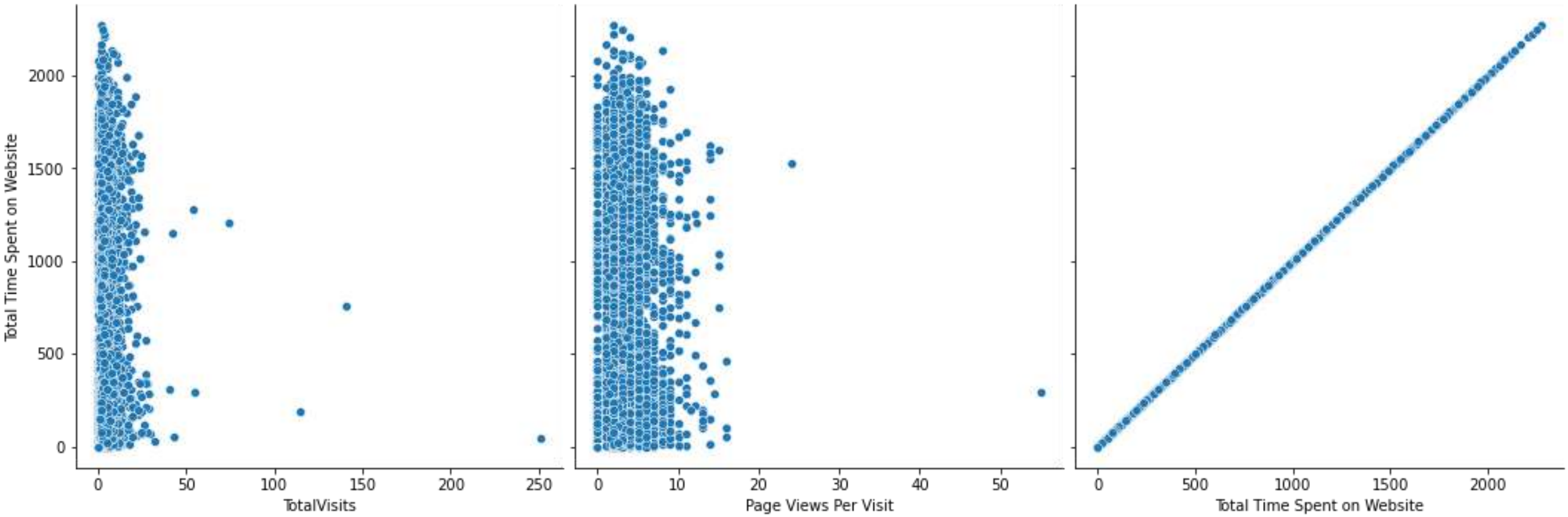
BIVARIATE ANALYSIS OF NUMERICAL VALUES



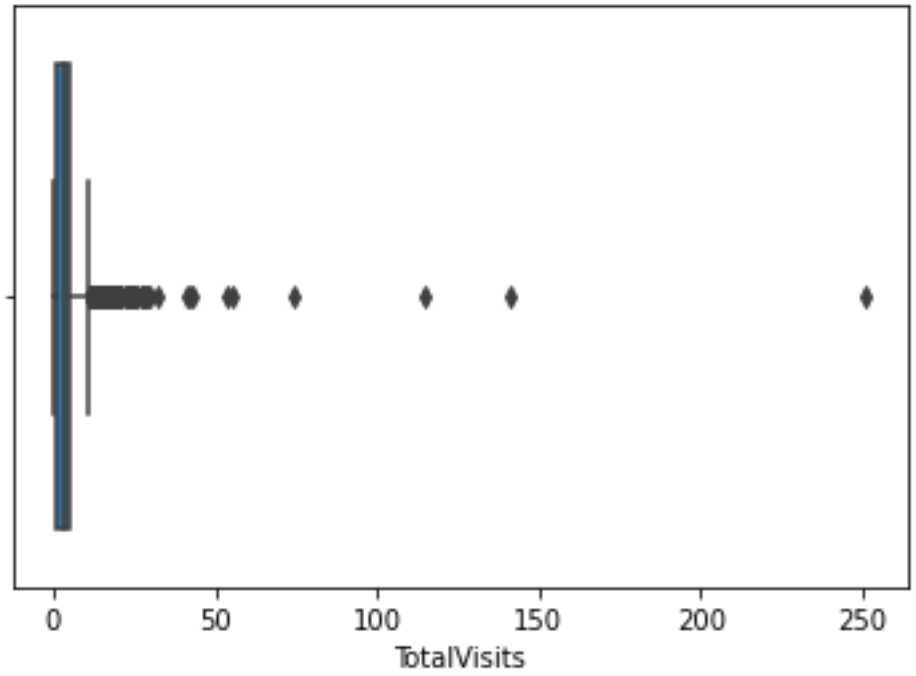
BIVARIATE ANALYSIS OF NUMERICAL VALUES



BIVARIATE ANALYSIS OF NUMERICAL VALUES

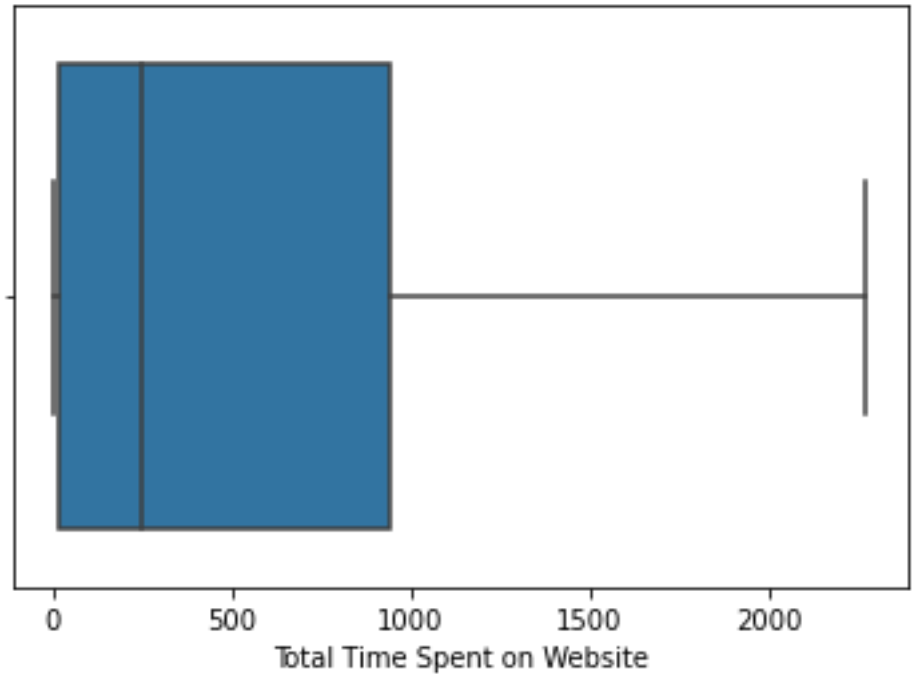


OUTLIERS ANALYSIS



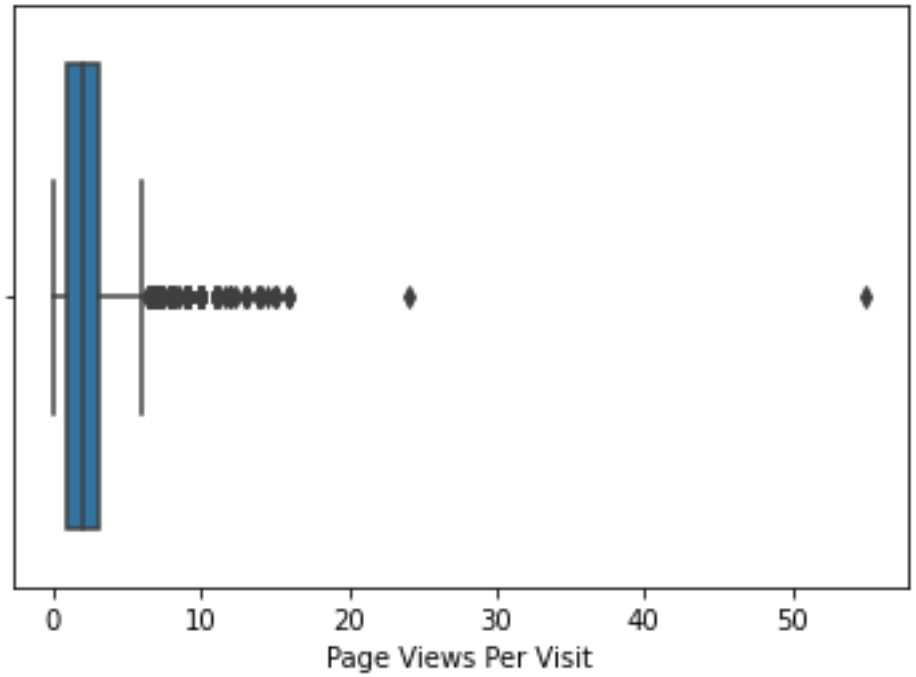
BOXPLOT FOR TOTAL VISISTS

OUTLIERS ANALYSIS

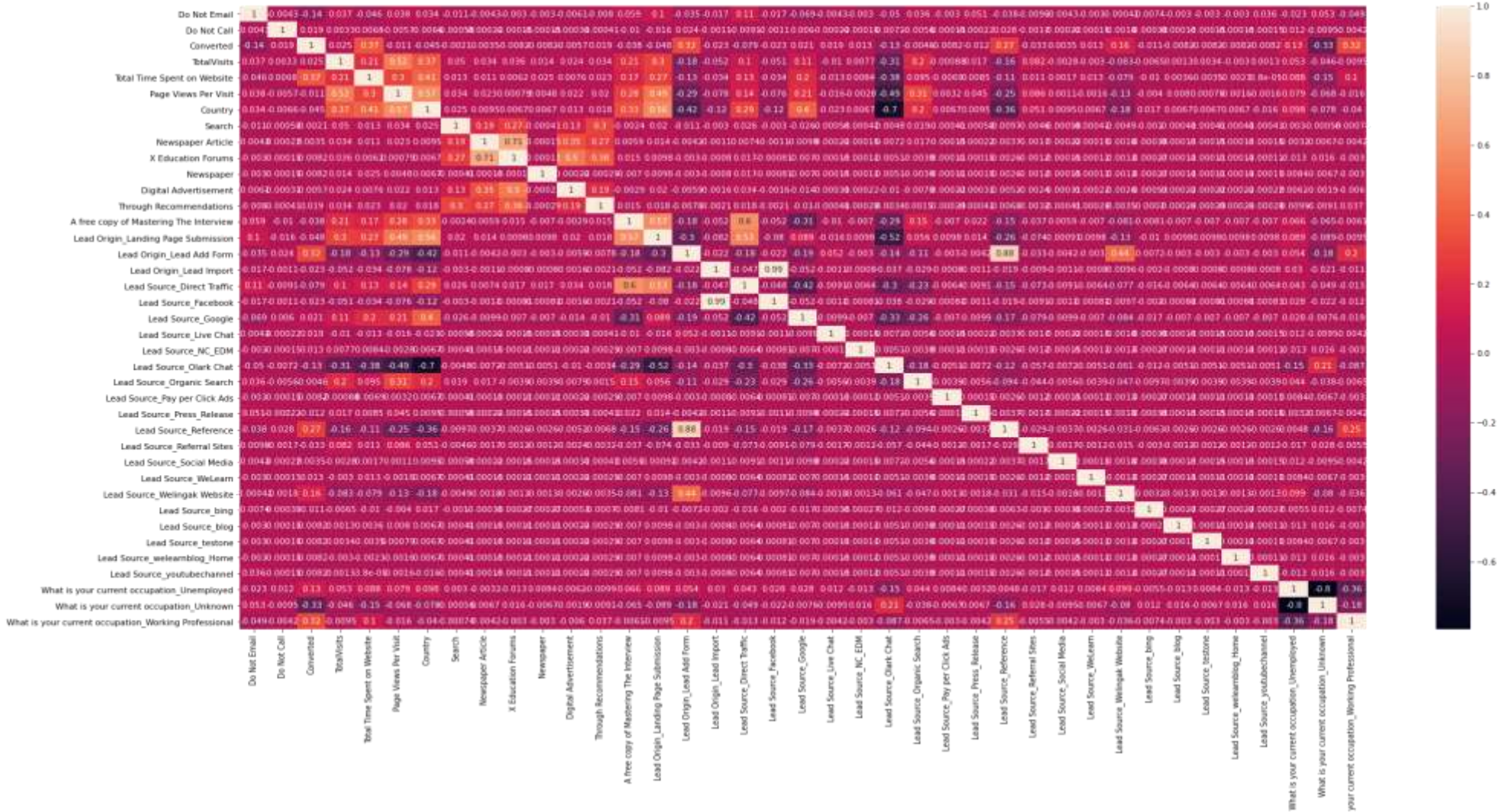


BOXPLOT FOR TOTAL TIME SPENT ON WEBSITE

OUTLIERS ANALYSIS



BOXPLOT FOR TOTAL PAGE VIEWS PER VISIT



From the Heatmap "Lead Origin_Lead Import" and "Lead Source_Facebook" have a correlation of 0.99, so we drop them.

Train-Test Split.

- Perform the Train-Test Split and assign the variables to X, y sets
- Perform Feature Scaling of Numerical Variables and check the conversion Rate
- Analyze the correlations
- Build a Logistic Regression Model, using GLM function and assign the constant

Model Building using RFE and Manual Feature Elimination

- Do feature scaling using RFE and select top 15 features
- Build our model iteratively until we get the optimum P-Value<0.05 and VIF value
- We found the Y-Value and Conversion Probabilities based on our Train Data Set
- We assigned a random value of 0.5 to predict if the conversion is 1 or 0

Confusion Matrix and Parameters.

Using these values, we derived the confusion Matrix. And Accuracy, sensitivity and specificity

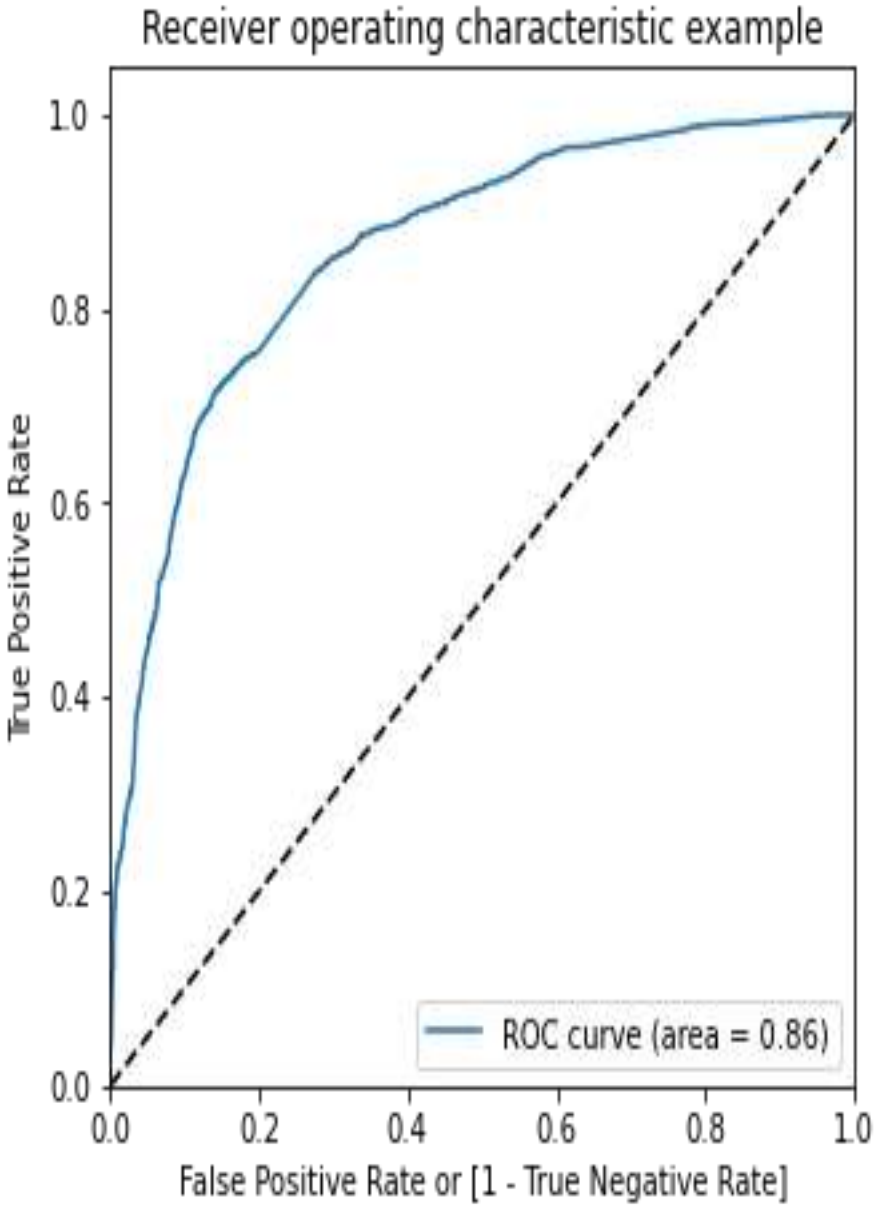
Based on this we plotted a graph to find the cut-off point, which in this case was

Based on the cutoff point We created another confusion matrix, and analyzed the sensitivity, specificity value as well as Precision and Recall

Compared these values to Business requirements to make decisions

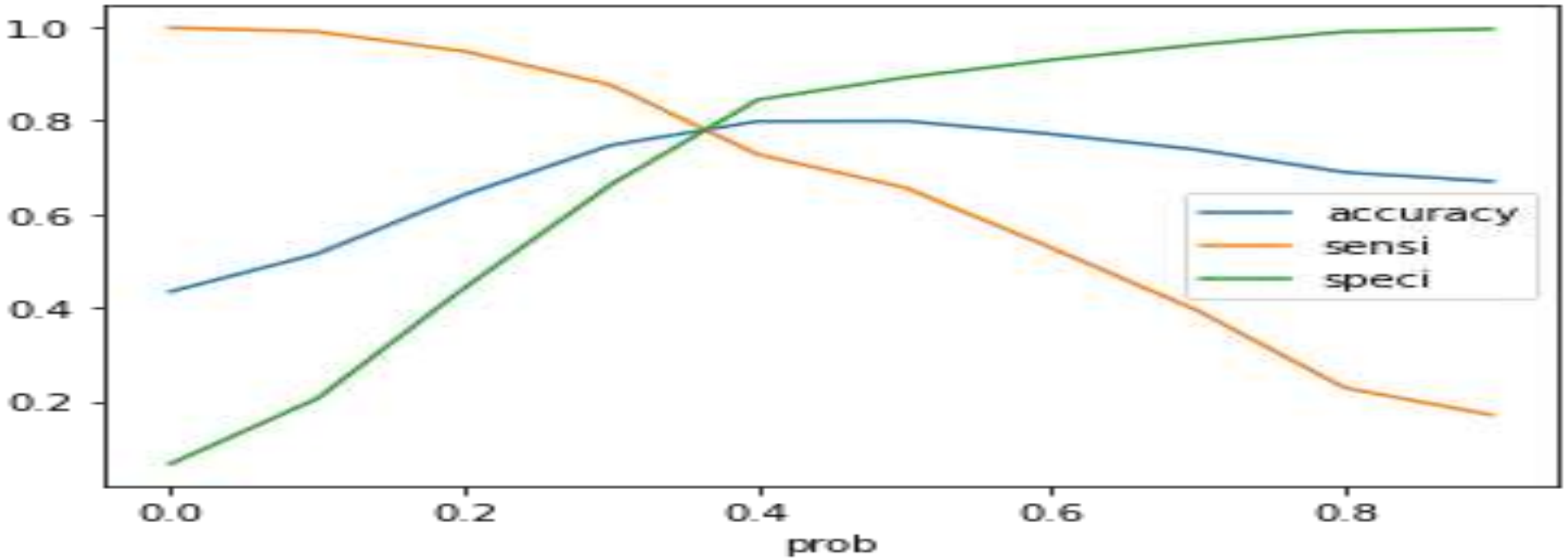
Then we plot a graph for precision and Recall Tradeoff

ROC CURVE



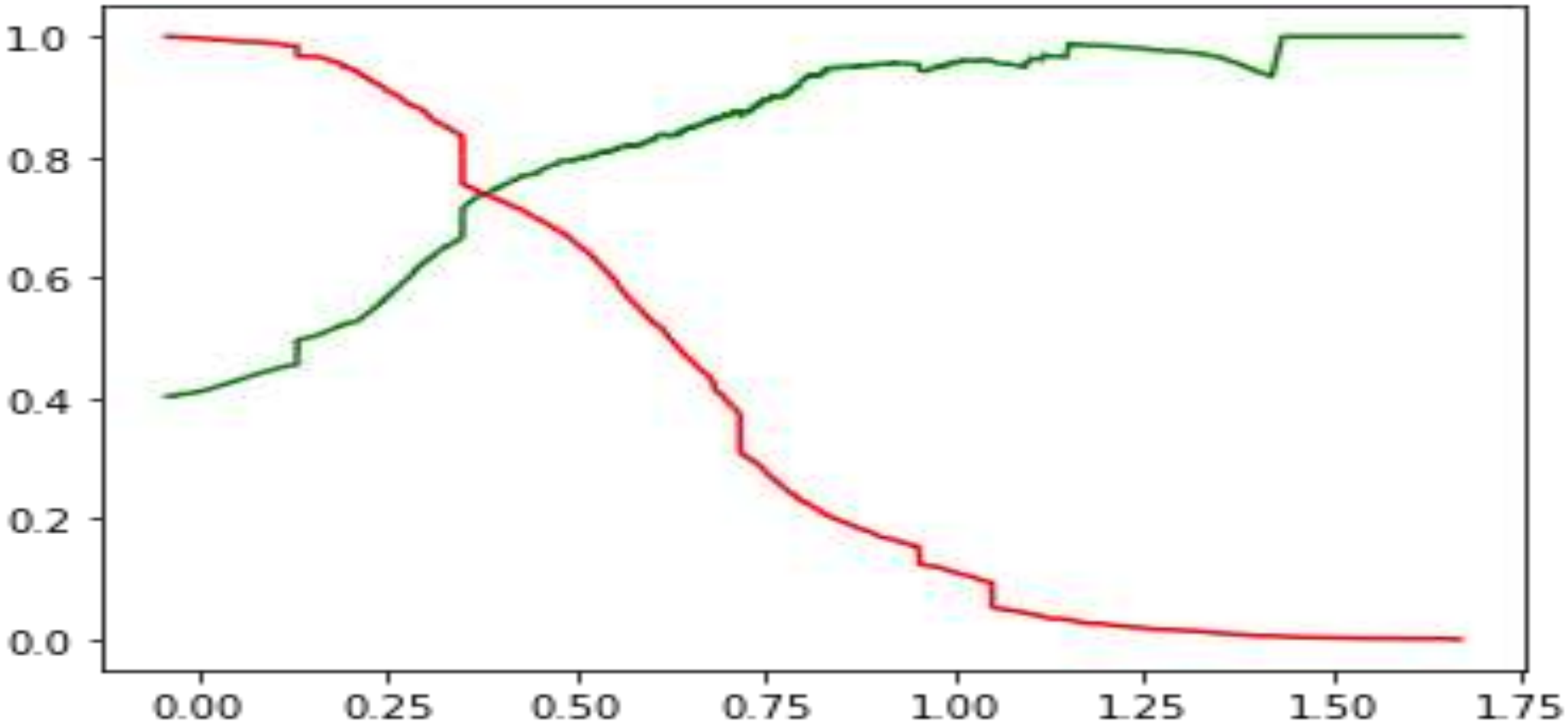
- Points to be concluded from above roc curve -
- The curve is closer to the left side of the border than to the right side hence our model is having great accuracy.
- The area under the curve is 86% of the total area.

GRAPH TO FIND THE CUTT OFF POINT



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

RECALL AND PRECISION CURVE



Making Predictions on the Test Set.

- We Scale and Transform the test data
- Derive the test predicted values for Y set
- Create a dataset of predicted values and conversion probability
- Create a Confusion Matrix and derive the parameters, accuracy, sensitivity and specificity
- Compare on the train data and find that all the values are satisfactory
- Create a Lead score, to associate the potential lead, if he will convert or not

Conclusions.

Train set data

- Accuracy =74.65%
- Sensitivity=87.52%
- Specificity=66.25

Test Set Data

- Accuracy =73.96%
- Sensitivity=87.27%
- Specificity=66.53%

Inference

- The CEO expected a conversion rate of 80%, we see that our sensitivity, which is the conversion rate is 87%, so, its a very good model
- The Accuracy, Precision and Sensitiviy for both train and test set are within range.
- In business terms, this model has the ability to adapt to the company's needs of the future.
- The most Important features for positive lead conversion are:

  Lead Origin_Lead Add Form

  What is your current occupation_Working Professional

  Lead Source_Welingak Website

BUSINESS RECOMMENDATIONS

- The CEO expected a conversion rate of 80%, we see that our sensitivity, which is the conversion rate is 87%, so, its a very good model
- Target People who are working professionals for high conversion rate
- Leads from Olark Chat and Lead Add forms are good.
- More time the visitor spends on website, better the chances of conversion, improve the website.
- News Papers and Emails do not seem to give the required results for conversion.
- Welingak Website acts as a good source of lead generator, target the audiense that visit the website
- Avoid people whose professional background is not known, identify the profession and then target the lead.