Analysis of the Lead Scoring Assignment (Supervised Learning)

Dataset :

Leads

Data dictionary :

Leads Data Dictionary

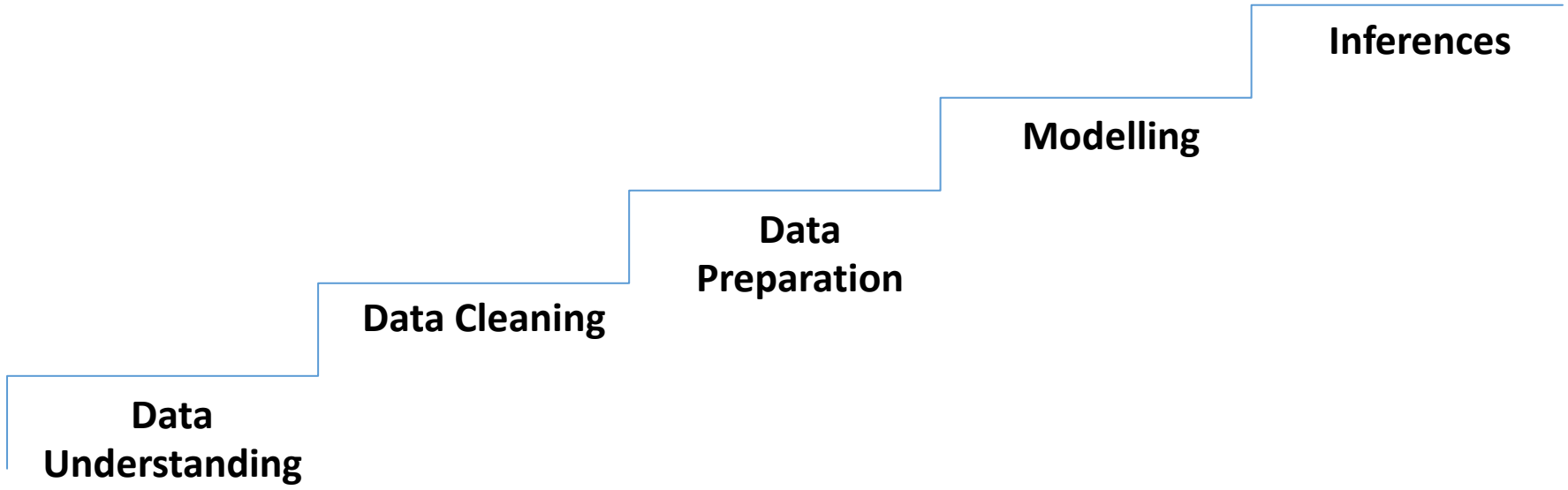
Problem Statement :

An education company named X Education sells online courses to industry professionals through 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(which are classified as lead).Although X Education gets a lot of leads, its lead conversion rate is very poor.(Roughly about 30%)

Expected Solution:-

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%.

All the Steps Involved in the Analysis



Data Understanding – Data Stats

- There are 9240 rows and 37 columns present in the dataset.
- There are multiple categorical columns(i.e. 'Specialization', 'How did you hear about X Education', 'Lead Profile', 'City') having select as values.
- Converted these values to null values by using np.nan since these values are default values present from the drop down list(i.e. not selected by the customer)

Data Cleaning – Finding missing values

- Below are the list of columns having high percentage(>35%) of missing values.Droped them

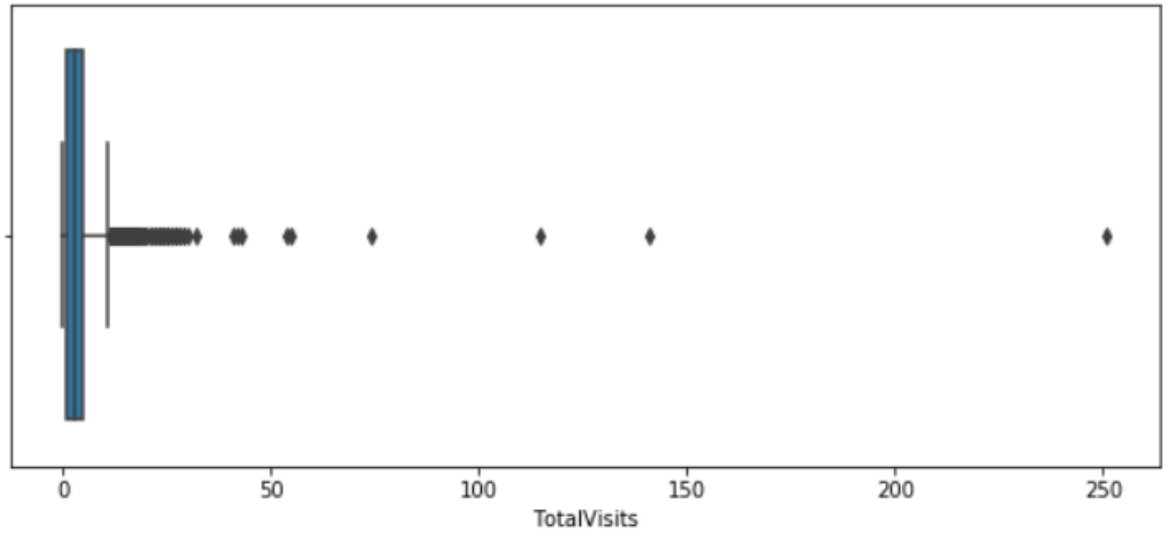
Specialization	36.58
How did you hear about X Education	78.46
Tags	36.29
Lead Quality	51.59
Lead Profile	74.19
City	39.71
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65

- Dropped columns that have only 1 value
  - Update me on Supply Chain Content
  - Get updates on DM Content
  - Magazine
  - I agree to pay the amount through cheque
  - Receive More Updates About Our Courses

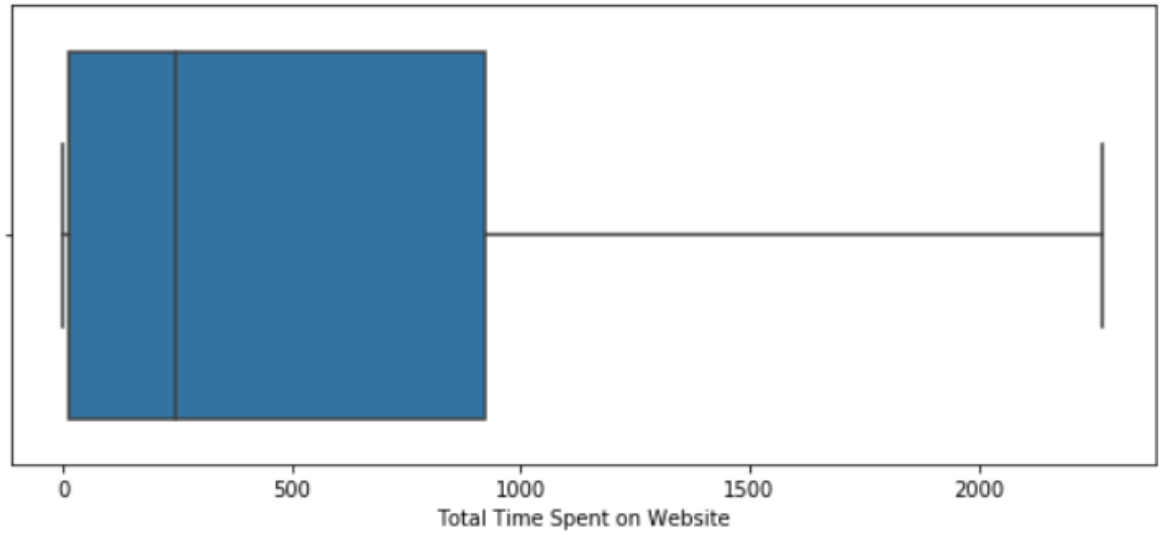
- Dropped columns that were highly skewed

Data Preparation – Checking for Outliers

- Column Name – **TotalVisits**
- Description : The total number of visits made by the customer on the website.
- The plot shows that they are some outliers in the higher range values.

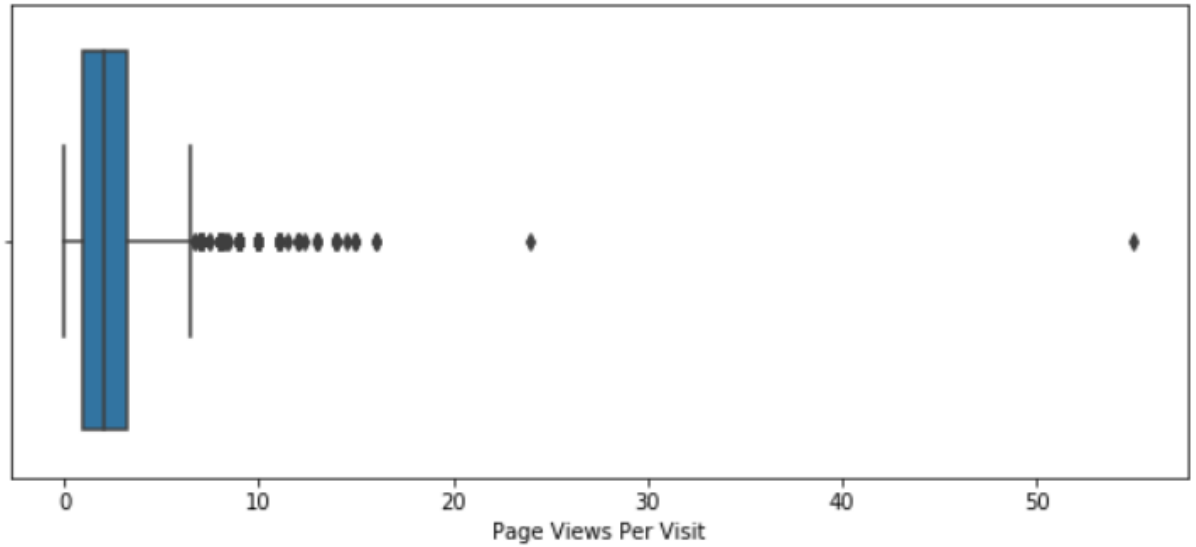


- Column Name – **Total Time Spent on Website**
- Description : The total time spent by the customer on the website.
- The plot indicates that the time spend above 75 quantile is more than 1000 seconds



Data Preparation – Checking for Outliers

- Column Name – **Page Views Per Visit**
- Description : Average number of pages on the website viewed during the visits..
- The plot shows that there is one data point where the visits per page is more than 50. This point needs to be capped.



Data Preparation – Outliers Treatment

Identifying Outliers with Skewness

- Explains the extent to which the data is normally distributed.
- Ideally, the skewness value should be between -1 and +1, and any major deviation from this range indicates the presence of extreme values.

- Table 1 – Indicates skewness before capping the outliers

- Table 2 – Indicates skewness after capping the outliers

TABLE 1

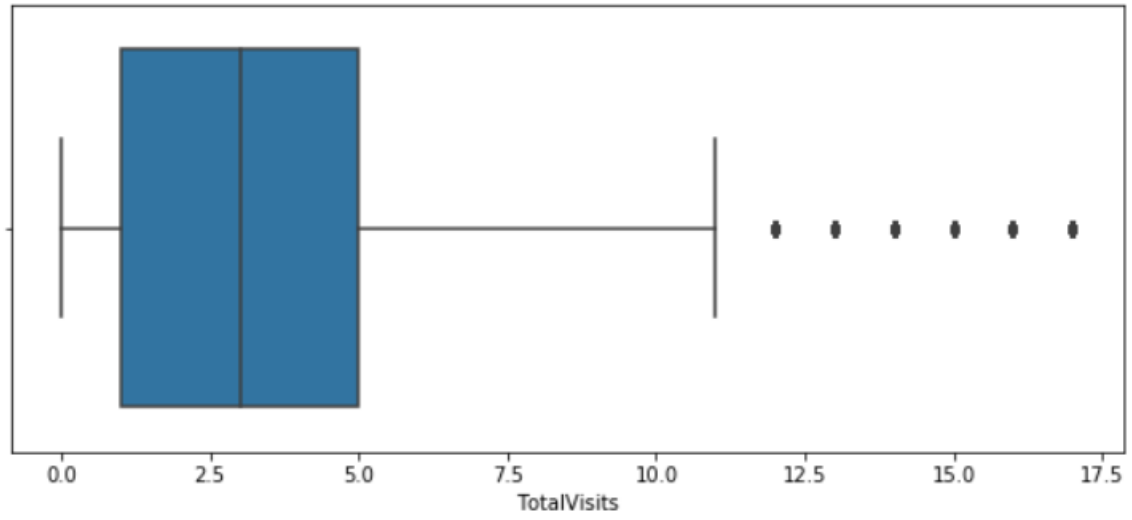
Column Names	Skew Value
Converted	0.500863
TotalVisits	19.921091
Total Time Spent on Website	0.970703
Page Views Per Visit	2.877019
A free copy of Mastering The Interview	0.780403

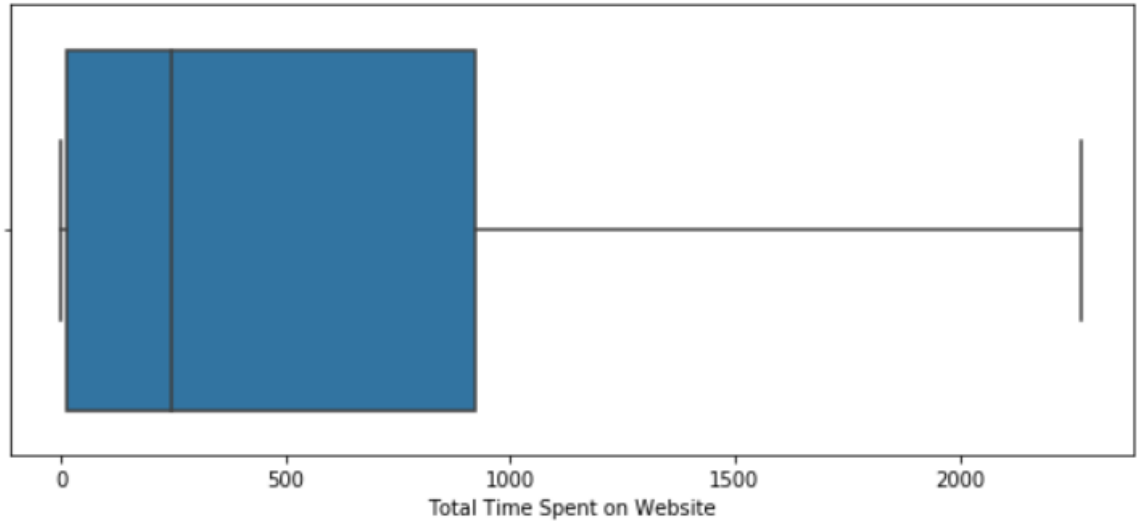
TABLE 2

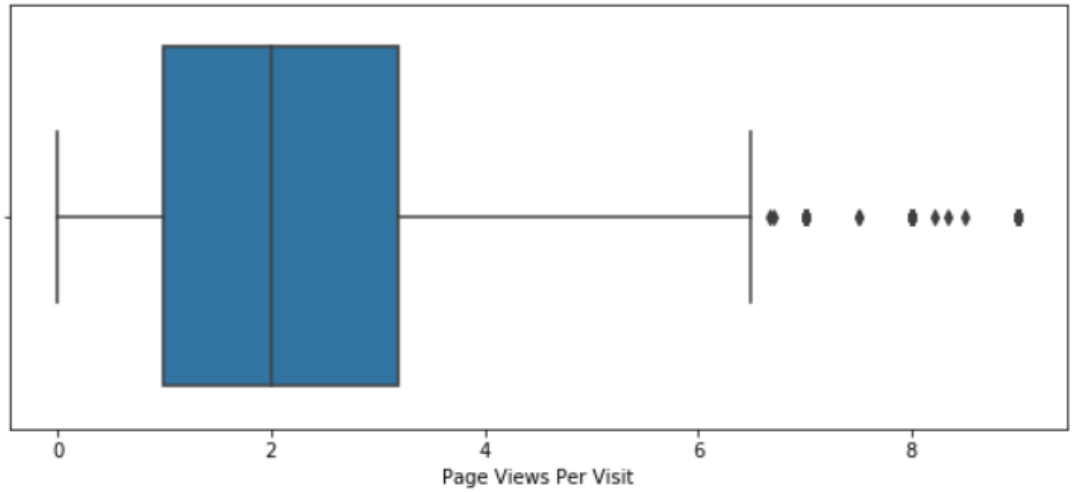
Column Names	Skew Value
Converted	0.500863
TotalVisits	1.607299
Total Time Spent on Website	0.970703
Page Views Per Visit	0.912265
A free copy of Mastering The Interview	0.780403

Data Preparation– Outliers Treatment

Plots after capping the outliers







Data Preparation – Binning Variables

- **Lead Source** - Binning values having less percentage of records as 'Other‘
  - Google
  - Direct Traffic
  - Olark Chat
  - Organic Search

  All other values that the not mentioned above are binned as Other

- **What is your current occupation** - Binned occupation into 3 classes
  - Unemployed
  - Working Professional
  - Student

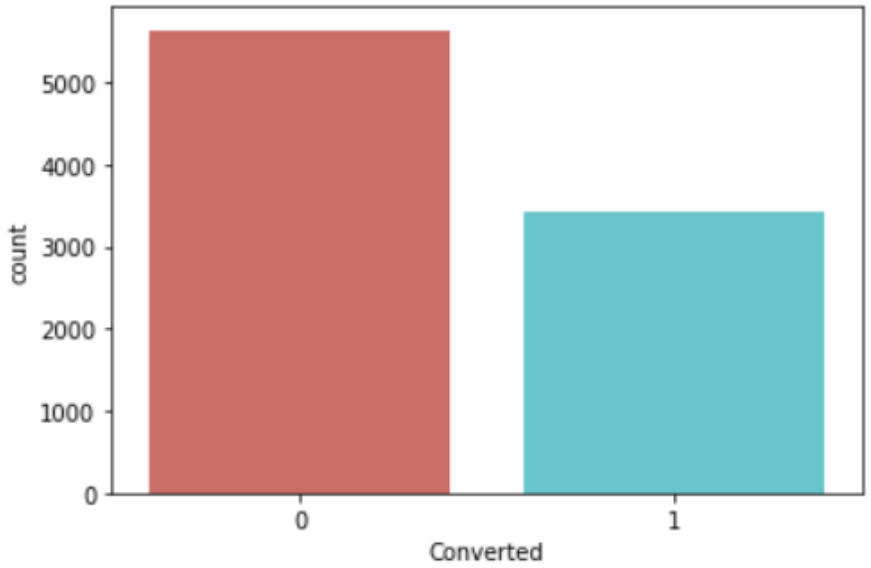
- **Last Activity** (Below list of values are having more number of records,while other values are binned as Other)
  - Email Opened
  - SMS Sent
  - Olark Chat Conversation
  - Page Visited on Website

Data Preparation – Creating Dummies

- Created dummies for the below list of Categorical columns
  - Lead Origin
  - Last Activity
  - Lead Source
  - What is your current occupation

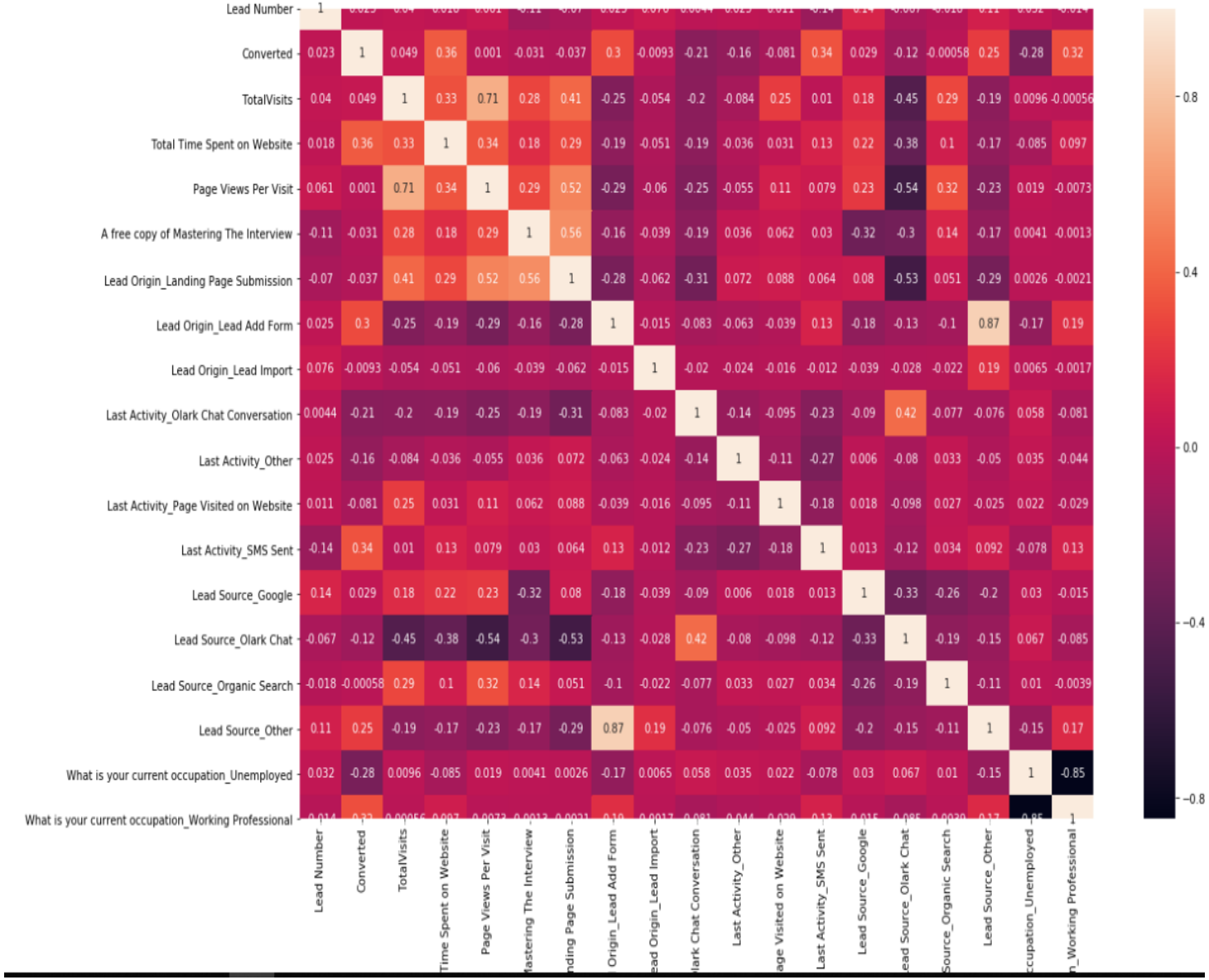
- Dropped the original Columns for which dummies have been created
- The lead conversion rate of the dataset = 38%

- Analysis of the target variable (i.e. Converted)



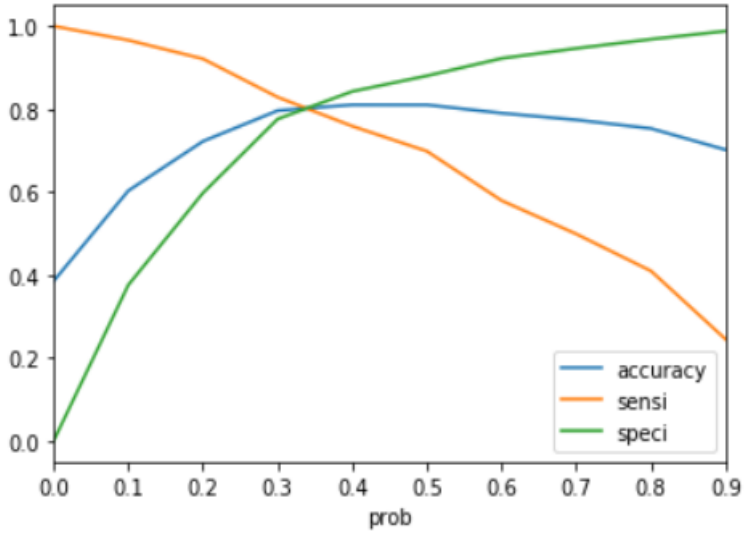
Modelling – Rescaling Columns

- Divided the dataset into Test-Train split of 70-30 ratio
- Rescaling the columns Using Standard Scalar , so that all the columns are scaled before modelling.
- Created the heat map to understand the correlation between predictor variables and the target variable.

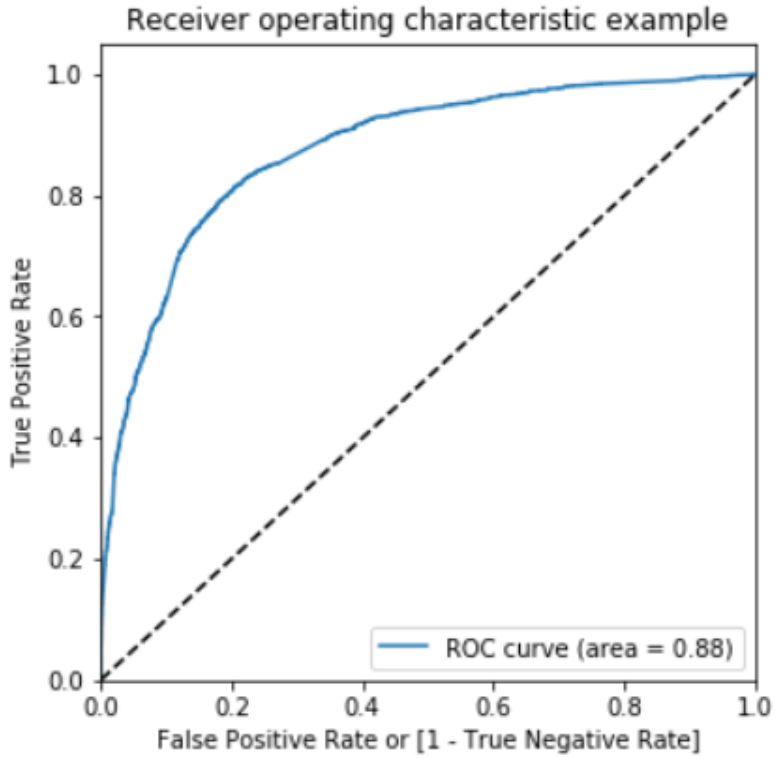


Modelling – Model Building

- Used RFE(Recursive Feature Elimination) to select the top 15 parameters/features.
- Build the Logistic model and checked for p-value and VIF value.
- Removed the insignificant variables and built the model.
- Plotted the ROC Curve
- Found the optimal cutoff Range (Between Accuracy ,sensitivity and specificity)



Optimal Cutoff Range



ROC Curve

From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Modelling – Confusion Matrix (Train Dataset)

Actual/Predicted	Not Converted	Converted
Not Converted	True Negative	False Positive
Converted	False Negative	True Positive

Train Dataset

- Accuracy : 81%
- Sensitivity : 80%
- Specificity : 81%
- Positive predictive value : 73%
- Negative predictive value : 86%
- Precision : 78%
- Recall : 70%
- F1 Score : 74%

Not Converted	Converted
3174	731
500	1946

Modelling – Confusion Matrix (Test Dataset)

Actual/Predicted	Not Converted	Converted
Not Converted	True Negative	False Positive
Converted	False Negative	True Positive

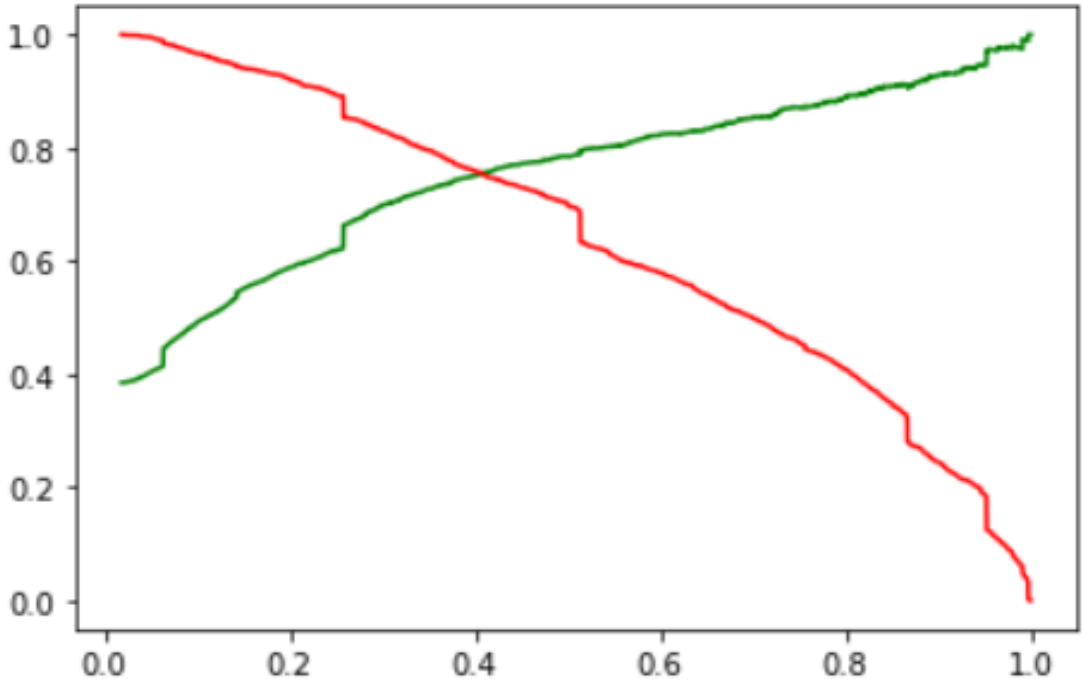
Test Dataset

- Accuracy : 81%
- Sensitivity : 75%
- Specificity : 84%
- Positive predictive value : 73%
- Negative predictive value : 86%

Not Converted	Converted
1460	274
246	743

Modelling – Precision & Tradeoff

Precision Trade Off Curve



Legend

Green – Precision

Red - Recall

Selecting the value of 0.4 from the Precision Recall Trade-off Curve for the cut-off probability.

Inferences

The top 3 variables that contribute to the probability of a lead getting converted are:-

- Lead Origin
- What is your current occupation
- Total Time Spent on Website

Reason:-

- Since these variables have a positive coefficient with respect to the target variable and are highly significant.
- Also theses variables have a low VIF value indicating that the influence of one variable doesn’t affect the other variable. They contribute to the lead getting converted (i.e. to hot lead)

Target Strategy:-

- Target Working professionals since they have a source of income and they want to learn and grow more in the organization in which they are working.
- Focus on the customers who spend more time on the website browsing the courses and watching the videos for more information
- Target the lead origin which landed in this page while filling the add form

Thank You