Lead Scoring Case Study

By
Sachin GR
Vineeth MR
Priya Bakhtani

Problem Statement

- X Education sells online courses to industry professionals.

- 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:

- X education wants to know most promising leads.

- For that they want to build a Model which identifies the hot leads.

- Deployment of the model for the future use

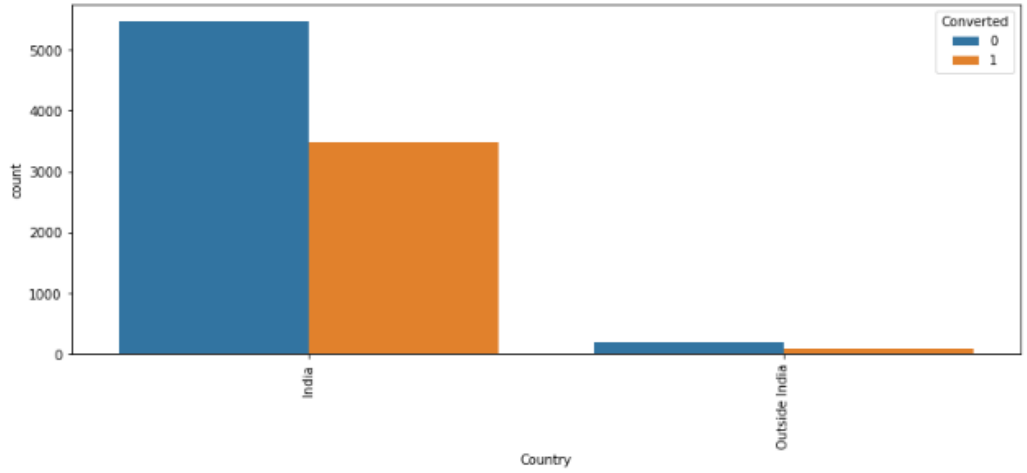
Approach

- Source the data for analysis
- Reading & Understanding the data
- Check and handle duplicate data NA values and missing values.
- EDA
- Data cleaning & Treatment
- Imputing the data with Mode values
- Checking Outliers
- Creating Dummy Variable
- Building Logistic Regression Model
- Scaling of the Data
- Model building
- Model evaluation - Precision and Recall
- Making predictions on the test set
- Final observation

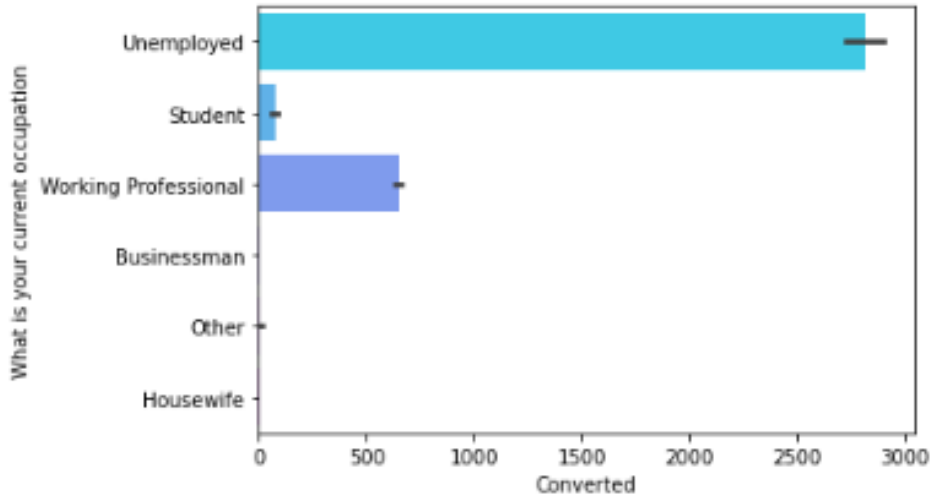
EXPLORATORY DATA ANALYSIS
Data Sourcing, Cleaning & Preparation

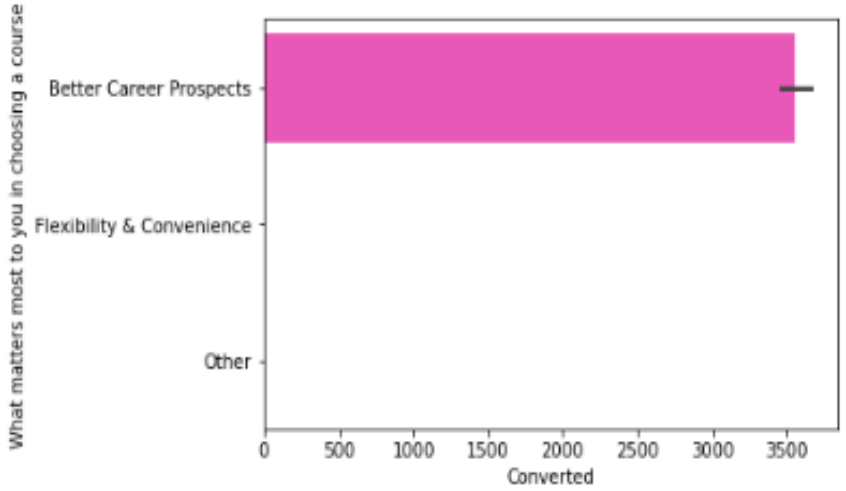
- Read the data from CSV File
- Data cleaning – Handling and removing the null values
- Removing unwanted columns in the data
- Imputing null values and mode values
- Imbalanced Variables that can be dropped
- Treating Outliers

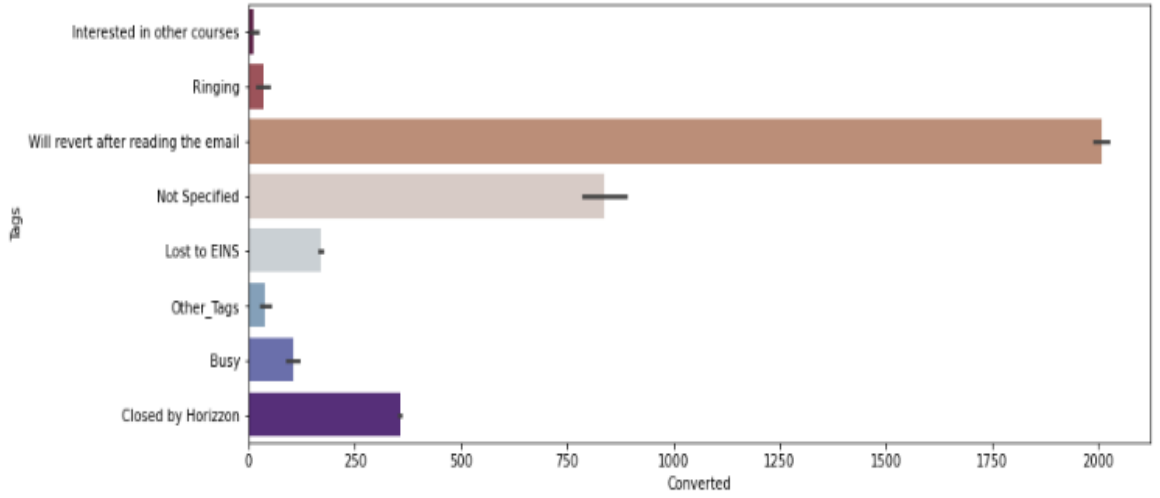
Imputing the data with Mode values

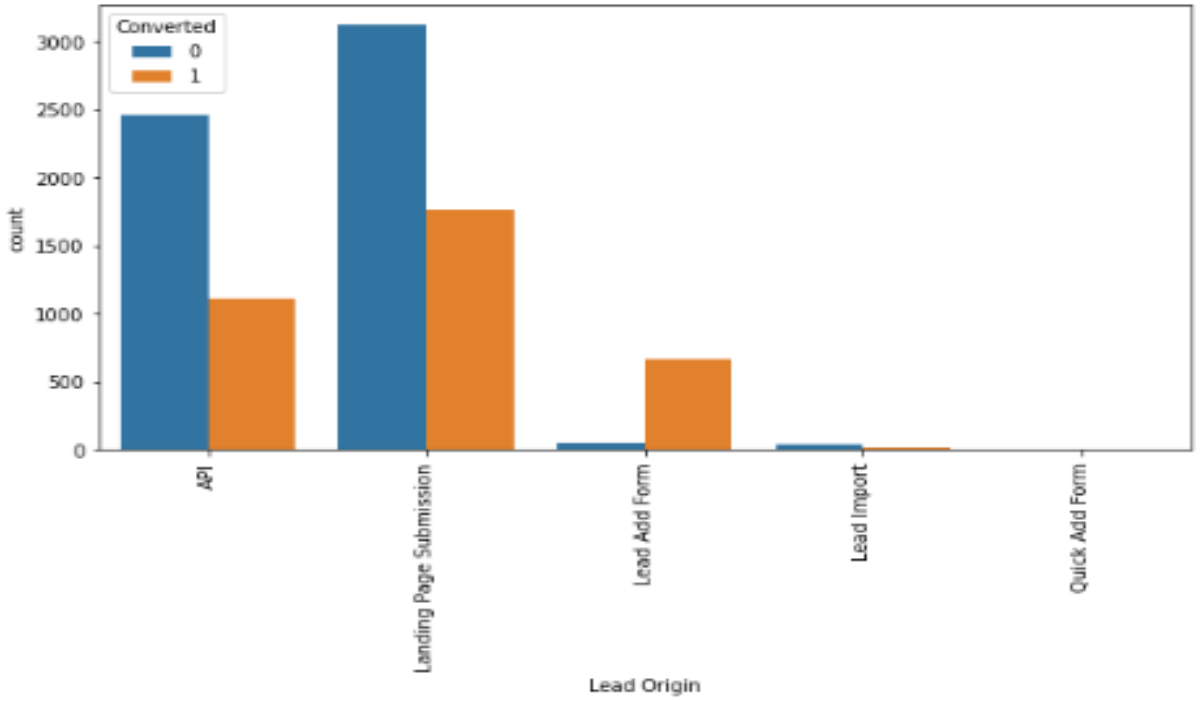


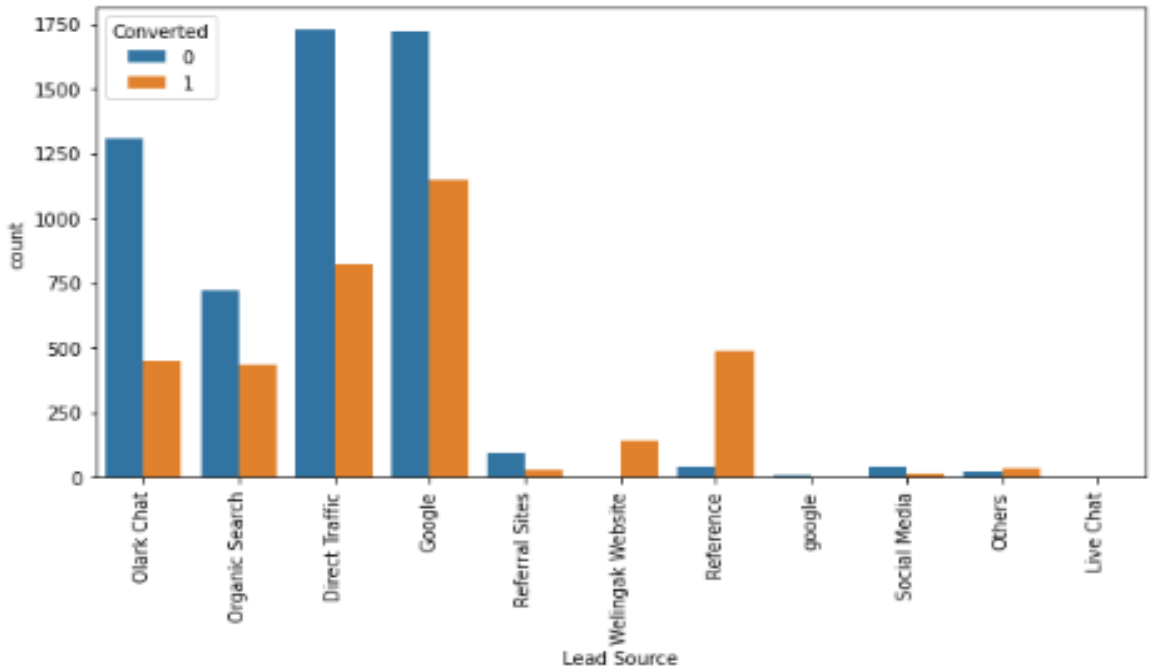
As we can see the Number of Values for India are quite high i.e nearly 97% of the Data and these columns can be dropped

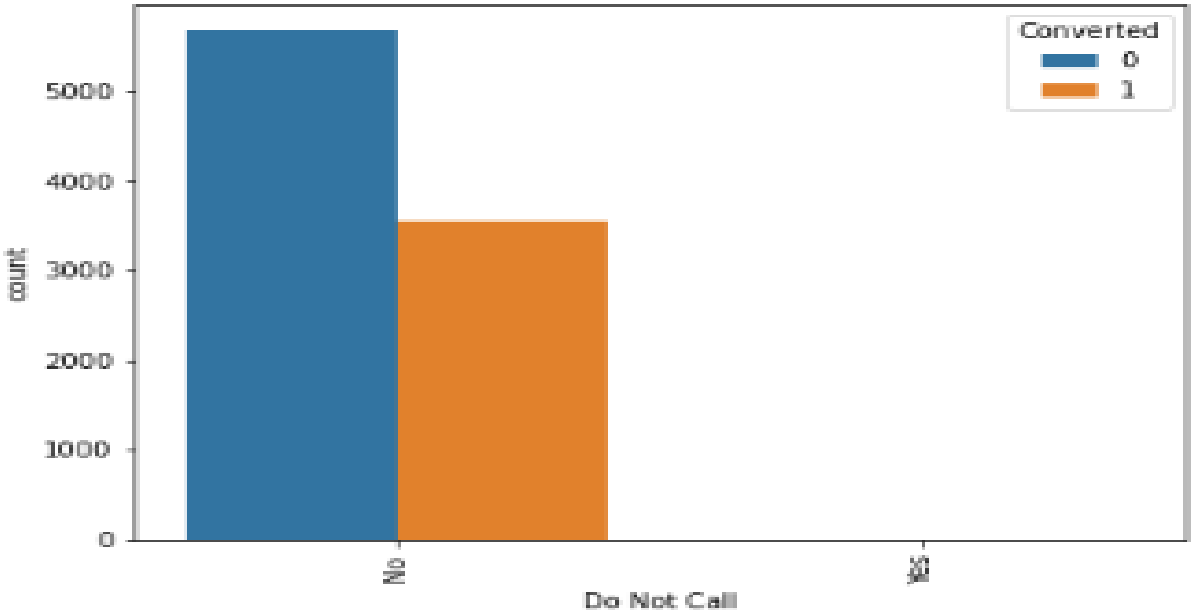


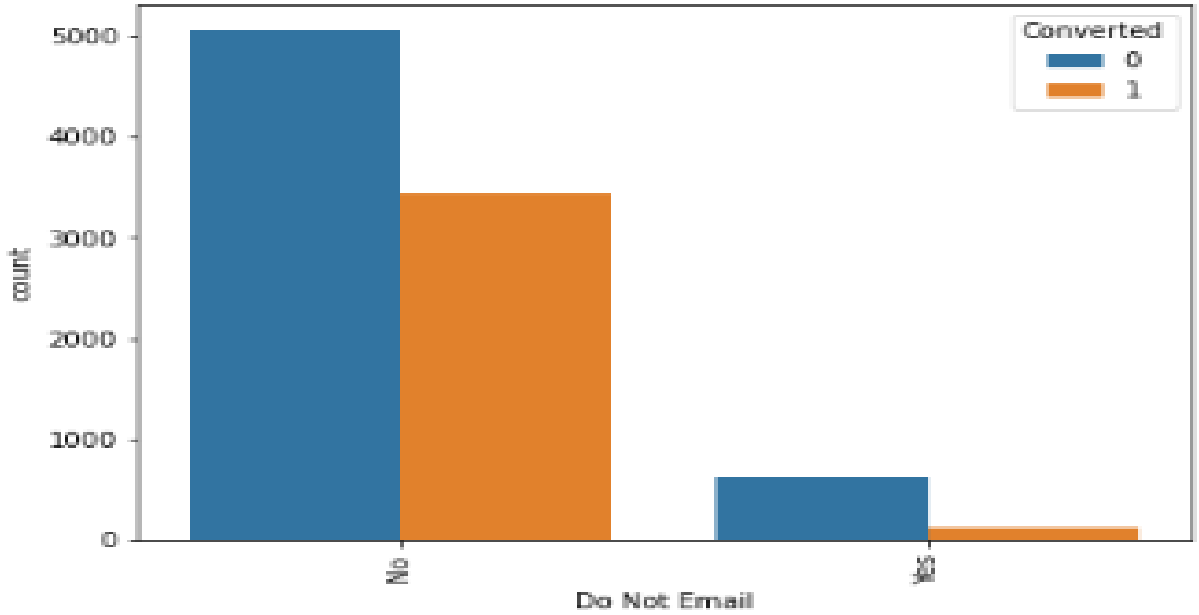




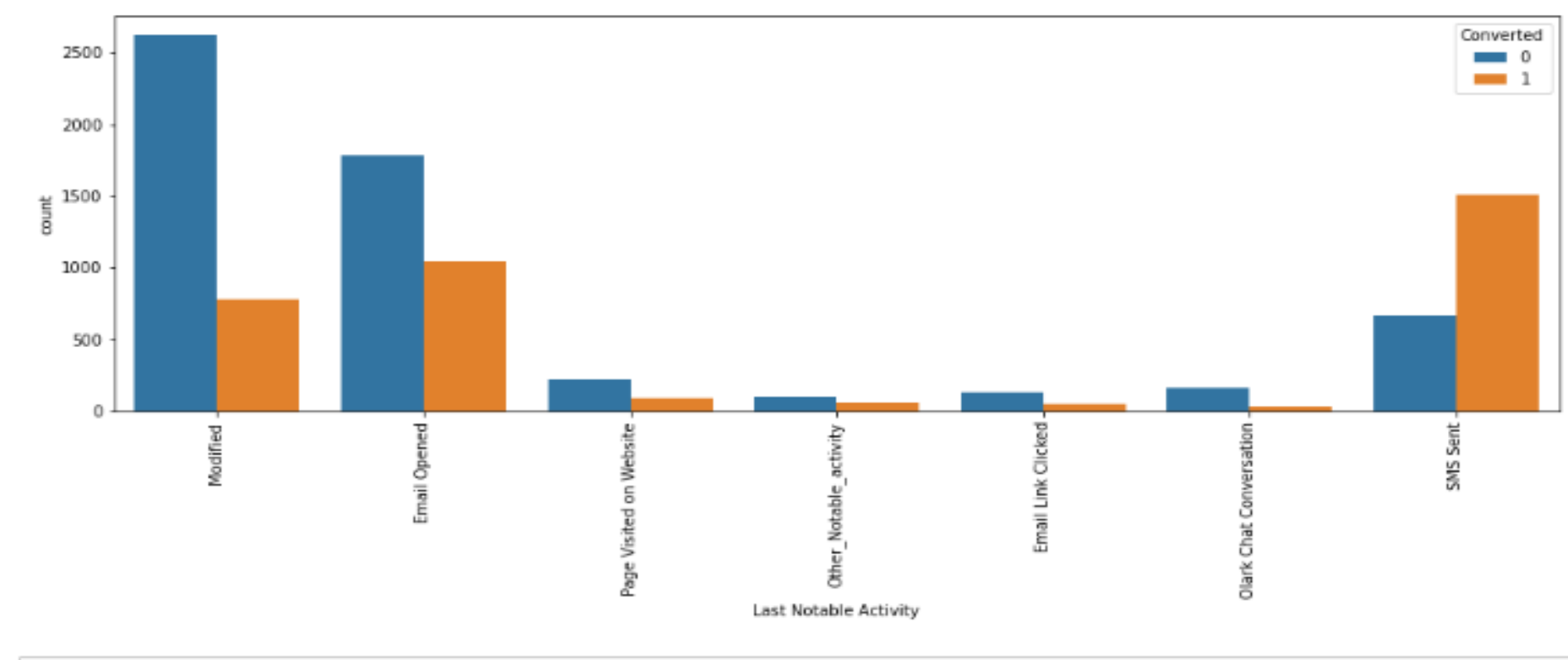

Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Social Media
Others
Live Chat
Lead Source

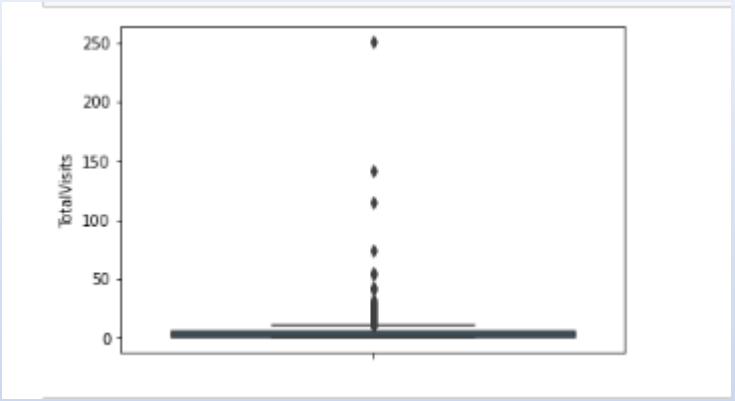

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call

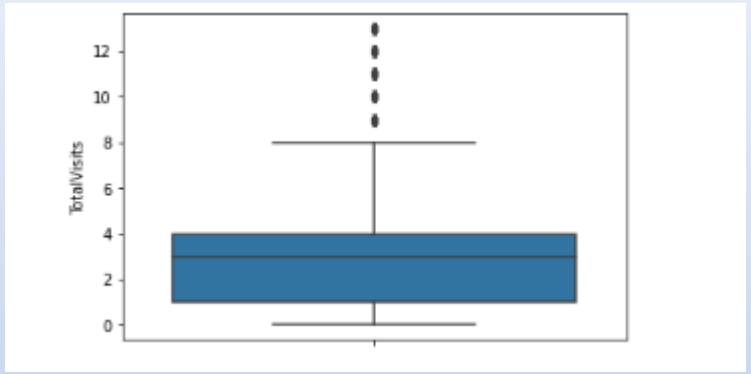

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

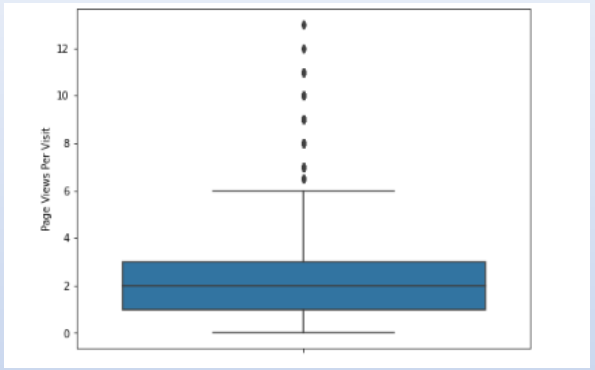
IMBALANCED VARIABLES THAT CAN BE DROPPED

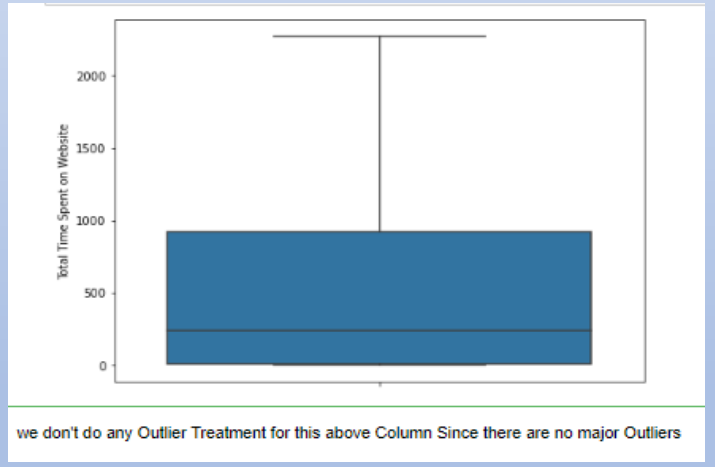


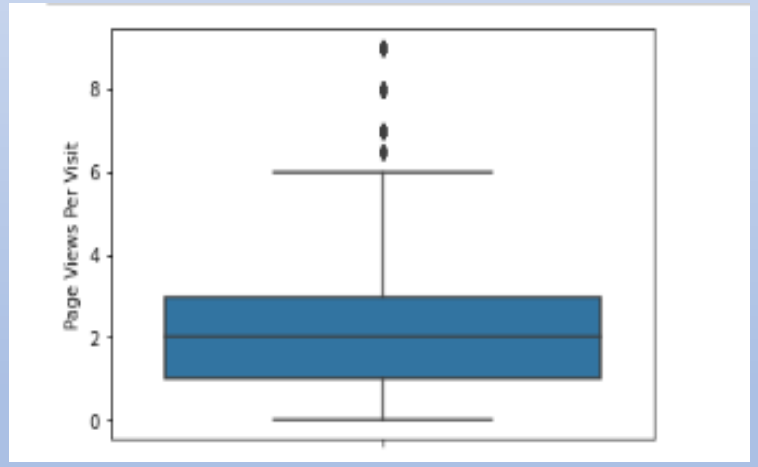
Treating Outliers







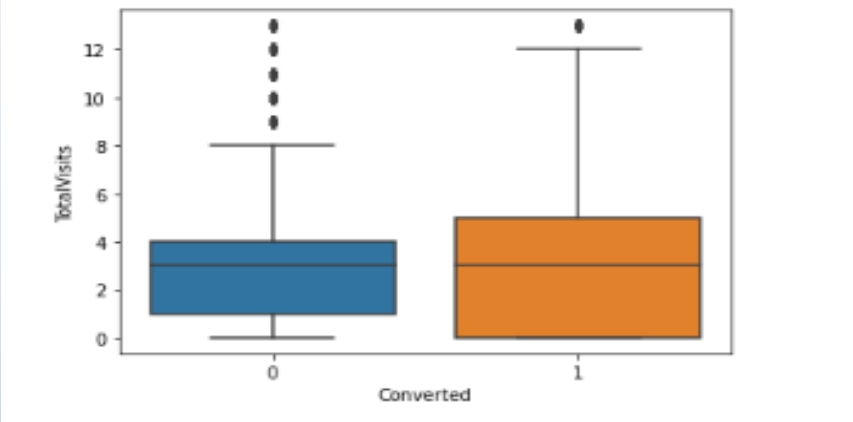




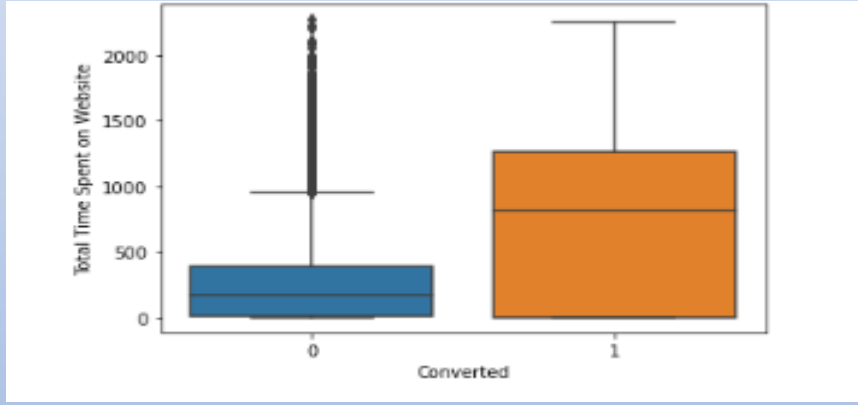
The below three have outliers

- Total Visits
- Total Time Spent on Website
- Page Views Per Visit

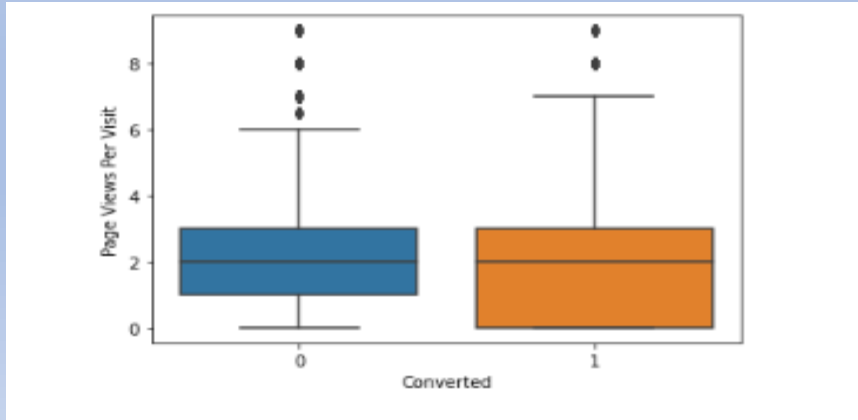
Analysis of Numerical Variable



- Median for converted and not converted leads are the close.
- other conclusive can be said based on Total Visits



- Leads spending more time on the website are more likely to be converted
- Website should be made more engaging to make leads spend more time



- Nothing specifically can be said for lead conversion from Page Views Per Visit
- Median for converted and unconverted leads is the same

Data Preparation

- Created dummy variables for categorical variables
- Converted Binary Variables into 0 & 1

Splitting Train & Test set
AND
Data scaling

- Splitting data into Train & Test set.
- Scaling of Data

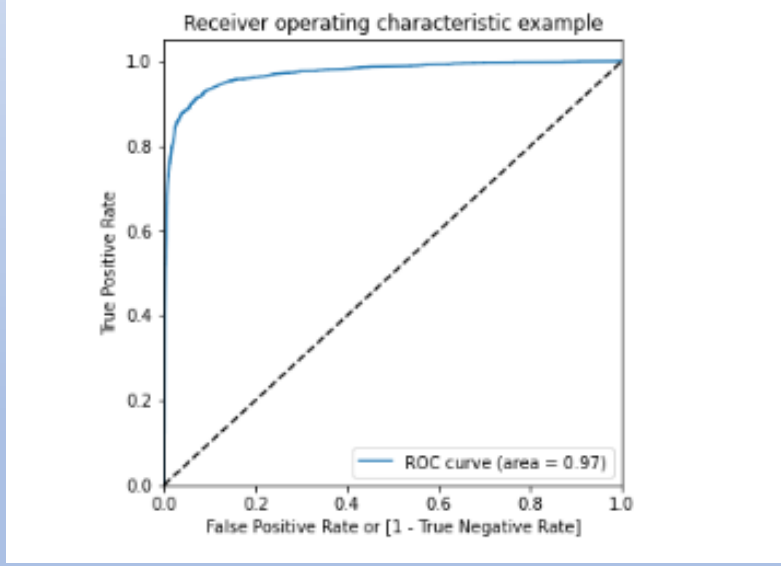
Model Building

- Feature Selection using RFE
- Plotting ROC curve
- Determined Optimal Model using Logistic Regression
- Calculated accuracy, sensitivity specificity, precision & recall

Plotting ROC Curve

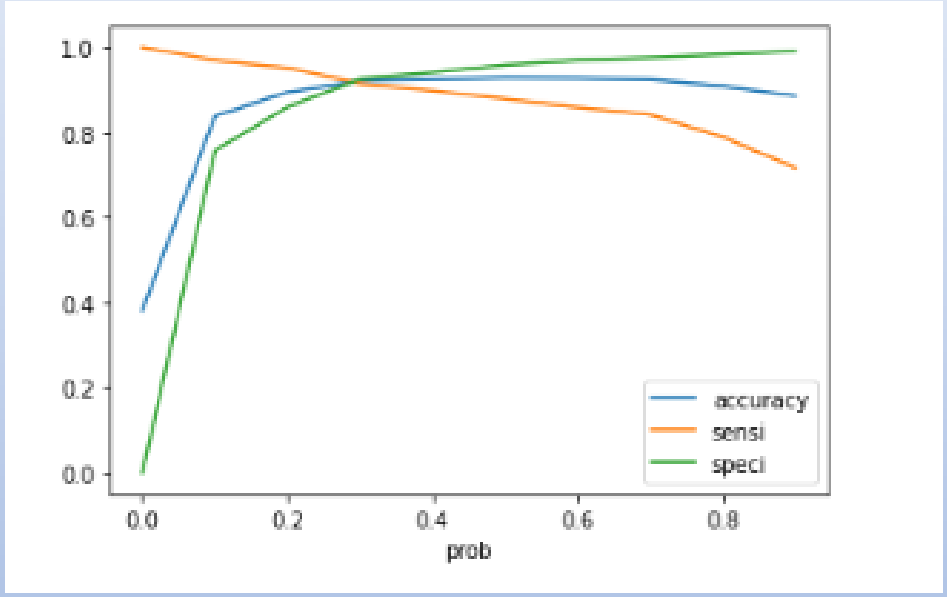
An ROC curve demonstrates several things:

- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).



- The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.

Finding Optimal Cut-off Point
AND
Observation(Train Data)



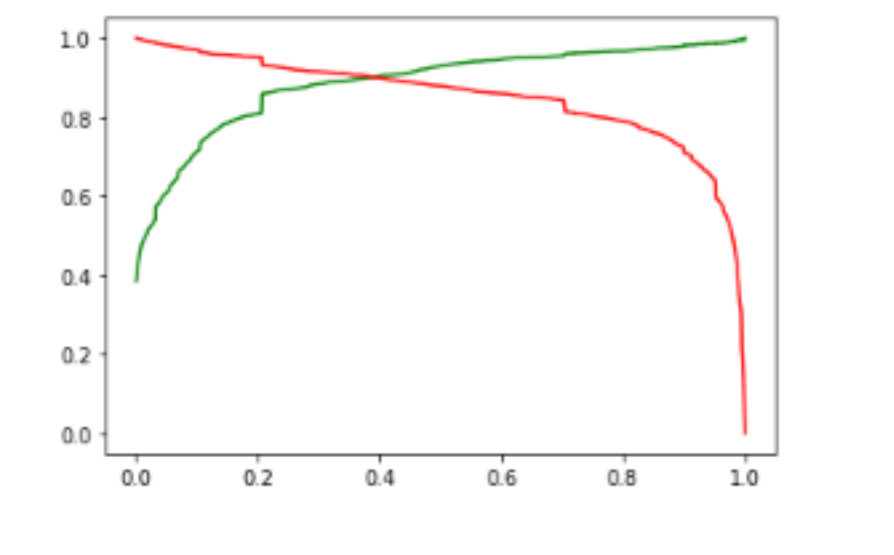
We have the following values for the Train Data:

- Accuracy : 92.14%
- Sensitivity : 91.49%
- Specificity : 92.54%

- From the curve above, we can get to know that 0.3 is the optimum point to take it as a cutoff probability.

- So as we can see above the model seems to be performing well. The ROC curve has a value of 0.97, which is very good.

Precision and Recall



- Precision = 88%
- Recall = 91%
- And the current cut-off point is 0.3

Final Observation

After running the model on the Test Data these are the figures we obtain:
Train Data:

- Accuracy : 92.14%
- Sensitivity : 91.49%
- Specificity : 92.54%

Test Data:

- Accuracy : 92.57%
- Sensitivity : 91.18%
- Specificity : 93.46%

The model seems to be performing well. which gives us confidence in recommending this model for making good calls