LEAD SCORE CASE STUDY

GROUP MEMBERS:

Himanshu Chaudhary

Sandeep Poduval

Problem Statement

- X education sells online courses to industry professionals
- Although, X Education gets many leads but the Lead conversion is poor at only 30%
- Goal is to identify Hot leads that have a higher chance of conversion
- Once identified, the lead conversion rate should go up as the Sales team can concentrate on converting the Hot leads

Business Objective:

To build a Logistic Regression model to identify the “Hot Leads” and achieve a lead conversion rate up to 80%

Solution Methodology

- Data reading
- Cleaning the data
  - Checking and removing redundant columns
  - Converting the label “select” to null values
  - Removing columns with >35% null values
  - Other columns with missing values are imputed with the mode
- Data Transforming
  - Changing the labels to dummy variables and “Yes”:1 and “No”:0
  - Removing the duplicate columns
- Test Train Split
  - Splitting the data in Test and Train split and scaling the data
  - Plotting the Heat Map
- Model Building
  - Running ref on 15 variable and checking the linear model regression results.
  - Removing the columns with high P value one by one to and checking the VIF value.
  - Predicting on the train model.
  - Checking the ROC curve and checking the accuracy, specificity and senility.
  - Checking and calculating precision and recall with cut-off of 0.35 and 0.41 on train and test data.

Understanding the data

Total Number of Rows = 9240

Total Number of Columns = 37

Removing Single value features – Magazines, Receive more updates about our Courses, update me on supply

Removing Prospect ID and Lead Number

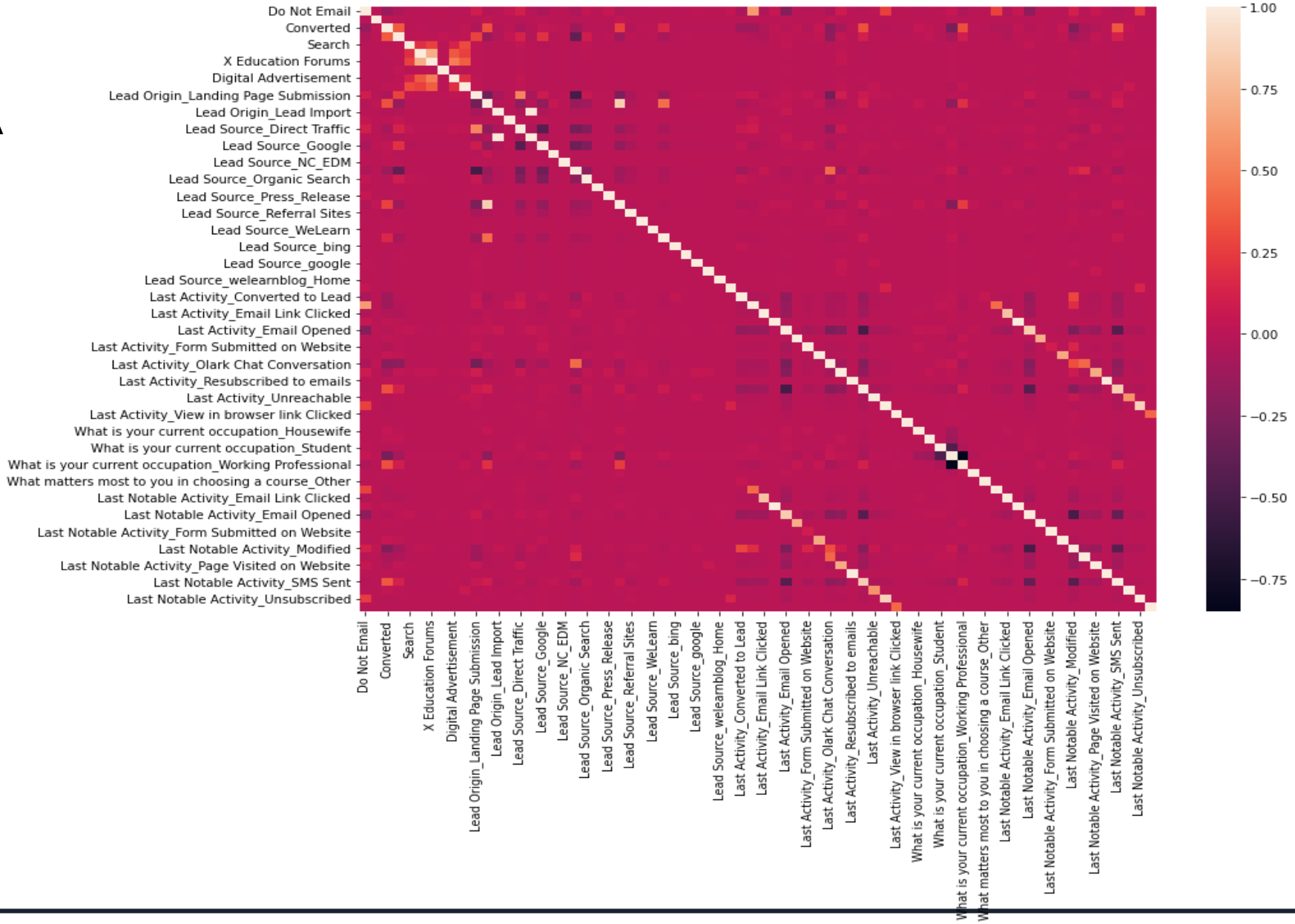
Dropping columns having more than 35% missing values

Data Conversion

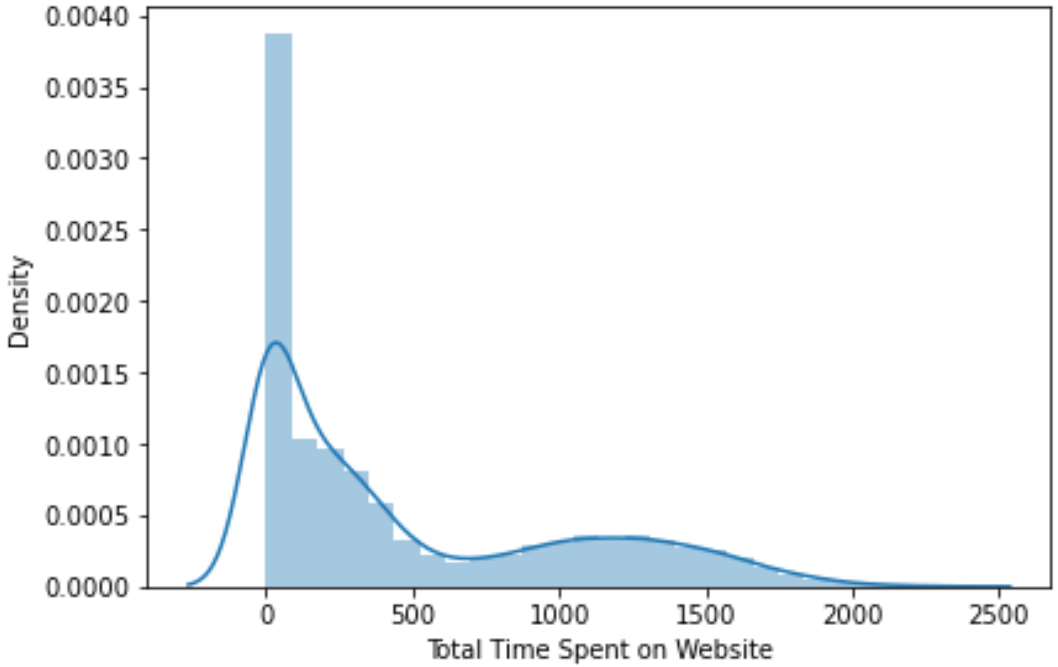
Numerical variables were scaled using MinMax Scaler

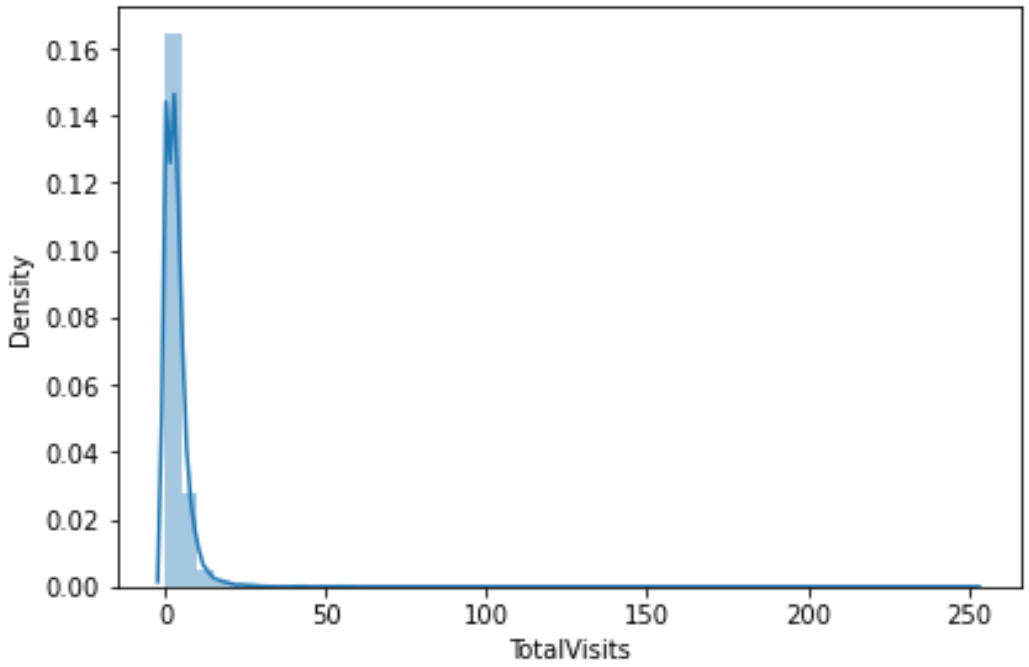
Dummy variables were created for object type variables

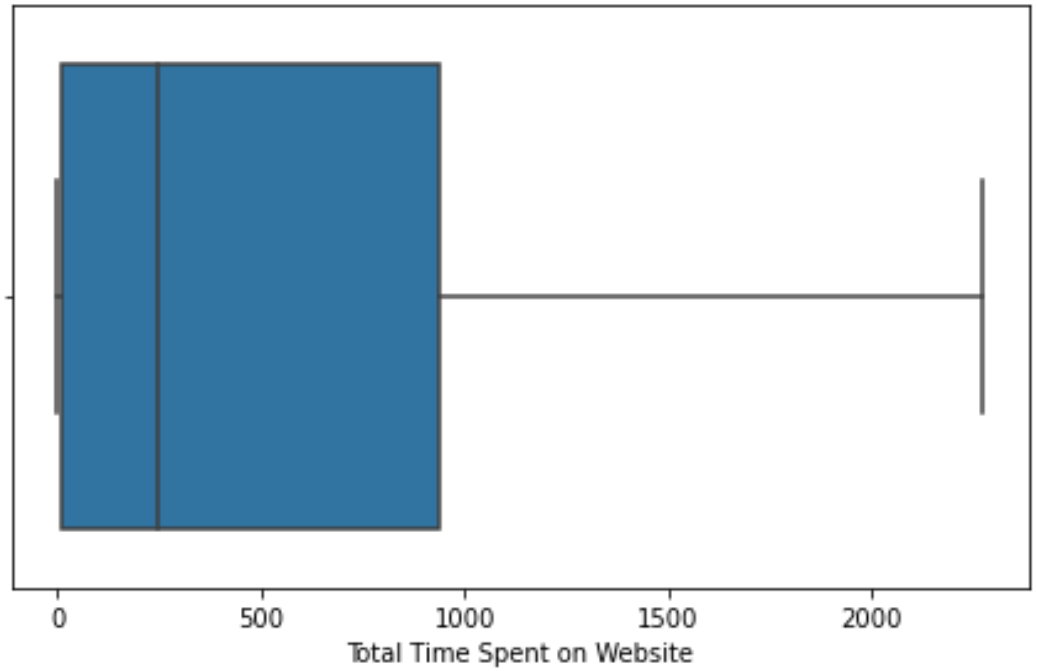
EDA

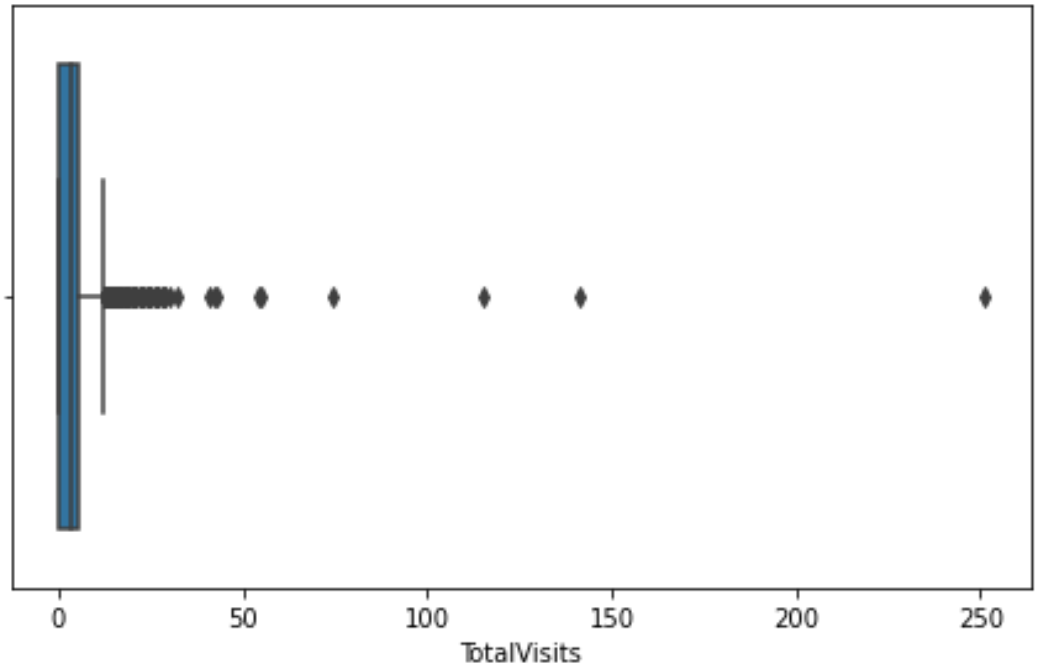

Do Not Email
Converted
Search
X Education Forums
Digital Advertisement
Lead Origin_Landing Page Submission
Lead Origin_Lead Import
Lead Source_Direct Traffic
Lead Source_Google
Lead Source_NC_EDM
Lead Source_Organic Search
Lead Source_Press_Release
Lead Source_Referral Sites
Lead Source_WeLearn
Lead Source_bing
Lead Source_google
Lead Source_welearnblog_Home
Last Activity_Converted to Lead
Last Activity_Email Link Clicked
Last Activity_Email Opened
Last Activity_Form Submitted on Website
Last Activity_Olark Chat Conversation
Last Activity_Resubscribed to emails
Last Activity_Unreachable
Last Activity_View in browser link Clicked
What is your current occupation_Housewife
What is your current occupation_Student
What is your current occupation_Working Professional
What matters most to you in choosing a course_Other
Last Notable Activity_Email Link Clicked
Last Notable Activity_Email Opened
Last Notable Activity_Form Submitted on Website
Last Notable Activity_Modified
Last Notable Activity_Page Visited on Website
Last Notable Activity_SMS Sent
Last Notable Activity_Unsubscribed
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

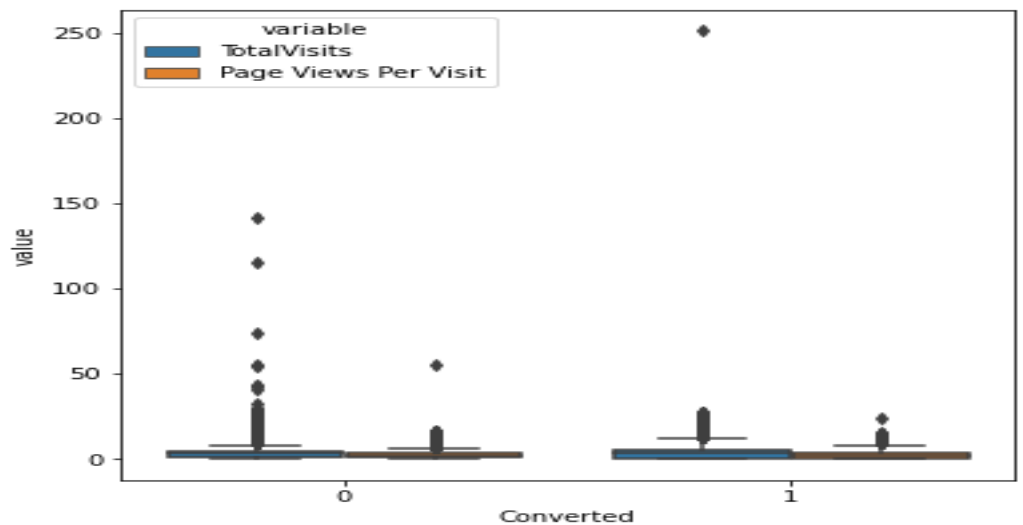
Univariate Analysis



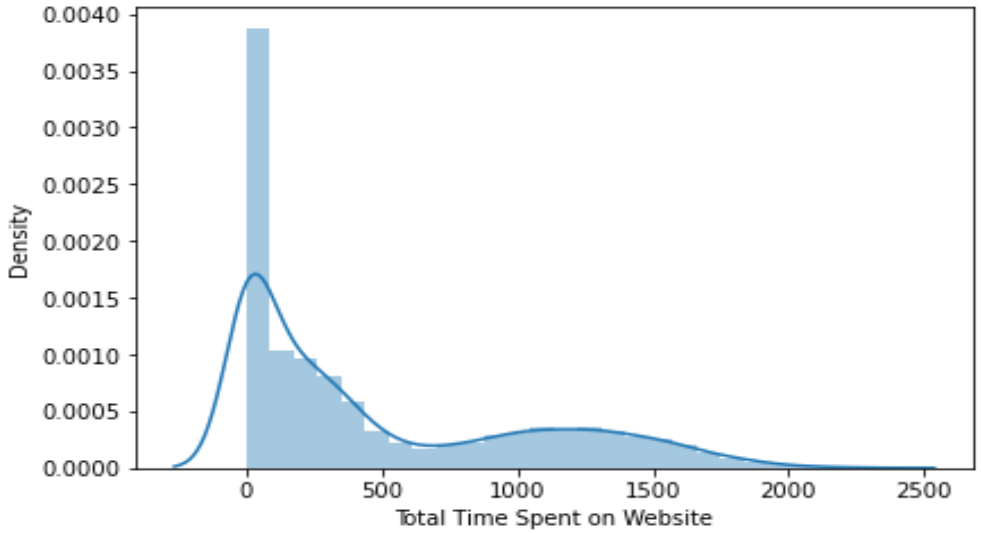


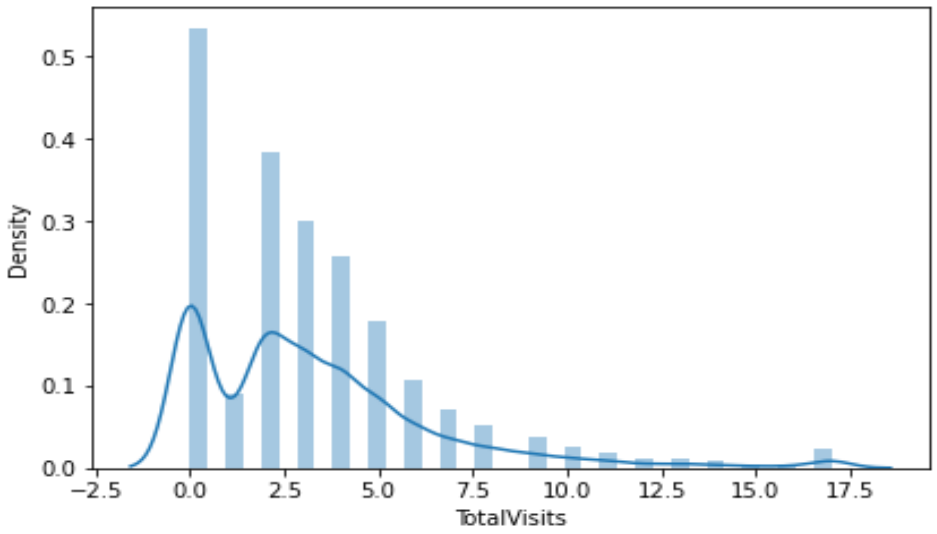

0
500
1000
1500
2000
Total Time Spent on Website

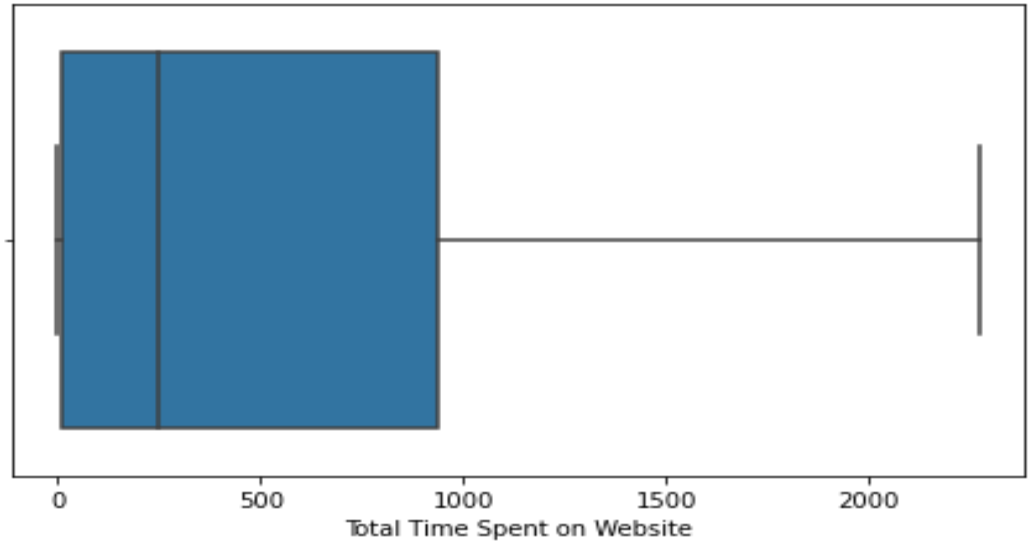

0
50
100
150
200
250
TotalVisits

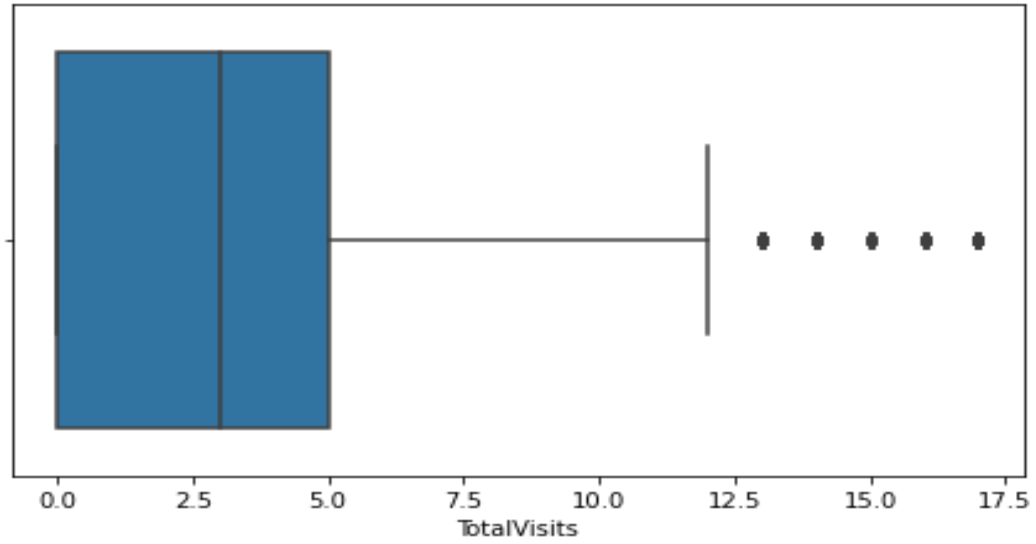

variable
TotalVisits
Page Views Per Visit
250
200
150
100
50
0
value
0
1
Converted

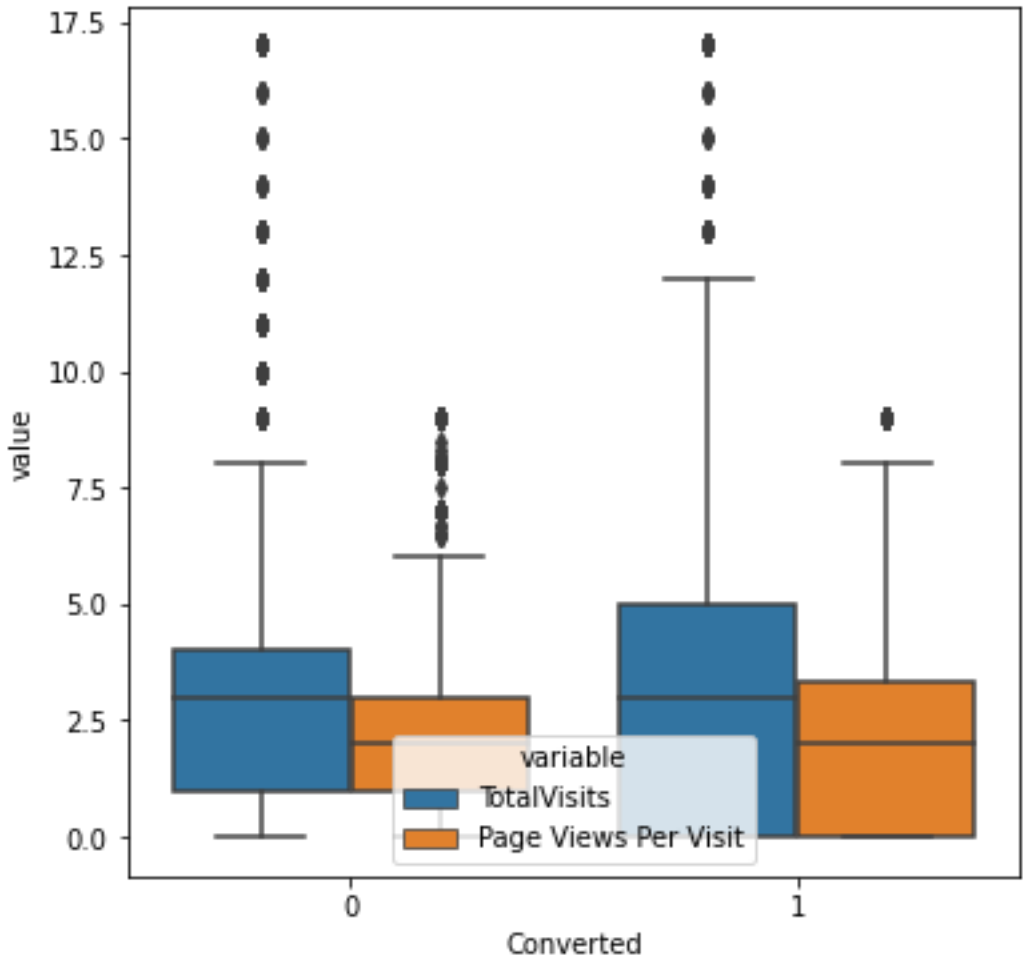
After Removing Outliers



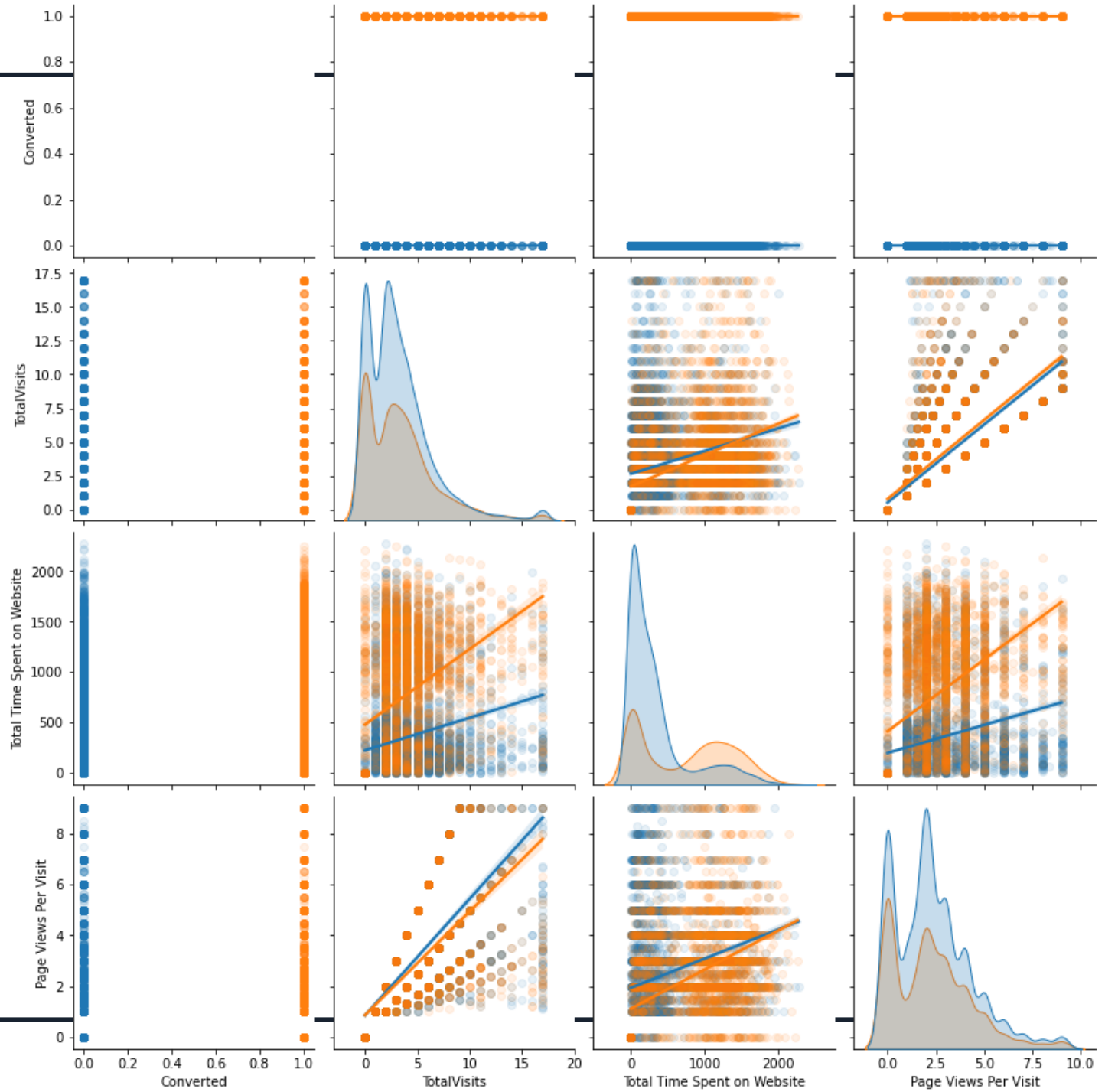







17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
value
variable
TotalVisits
Page Views Per Visit
0
1
Converted

Bivariate Analysis



Model Building

Splitting the data into Training and Test slit with 70:30 ratio

Using RFE for feature selection

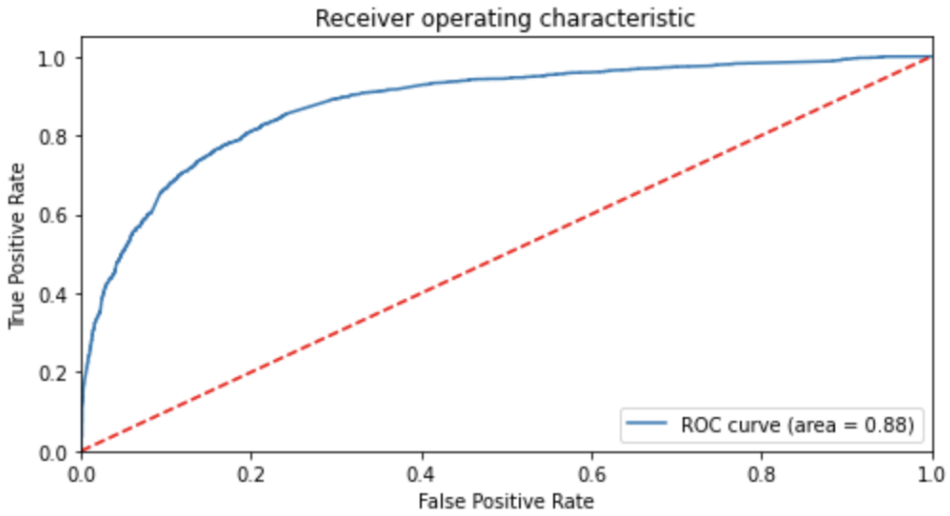
Running RFE with 15 variables

Building Model by removing the variable whose p-value is greater than 0.05 and VIF is greater than 5

Predicting the Test Data

Overall accuracy is 81%

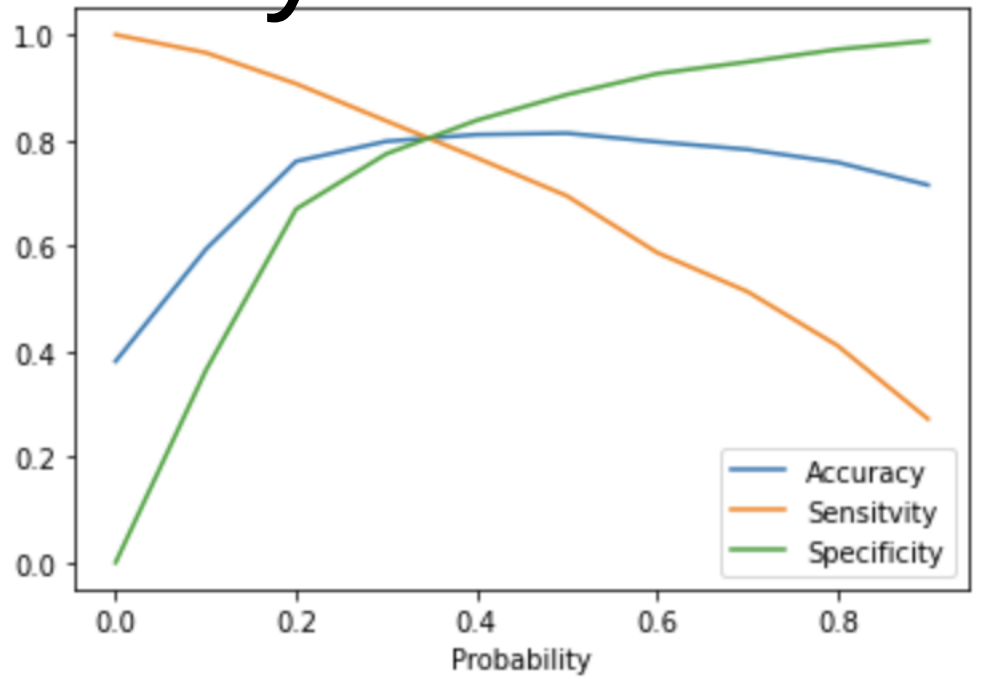
ROC Curve



Finding the Optimal Cut off point where we have balanced Sensitivity and Specificity.

Optimal Cut off is **0.35**

Accuracy, Sensitivity, Specificity



Conclusion

The variables that mattered the most in Potential buyers:

For a good conversion ratio:

- Lead Origin is through a Lead Add form that is the lead has entered his details in the form provided.
- Last Notable Activity is Had a phone conversation, the customer has had a phone conversation with the Sales team.
- And the Lead is a working professional