Lead Scoring Case Study
Group Assignment

Group Members : 1. Tapaswini Dash
2. Priti Singh

Problem Overview

1. X Education sells online courses to industry professionals.
2. 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.
3. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
4. 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:

1. X education wants to know most promising leads.
2. For that they want to build a Model which identifies the hot leads.
3. Deployment of the model for the future use.

Solution Approach

Data Importing and Understanding

Data cleaning and data manipulation

1. Check and handle duplicate data
2. Check and handle NA values and missing values
3. Drop columns, if it contains large amount of missing values and not useful for the analysis
4. Imputation of the values, if necessary
5. Check and handle outliers in data

EDA

1. Univariate data analysis: value count, distribution of variable etc
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc

Feature Scaling & Dummy Variables and encoding of the data

Classification technique: logistic regression used for the model making and prediction

Validation of the model

Model presentation

Conclusions and recommendations

Data Importing and Manipulation

Dataframe has **9240** Rows **37** Columns.

In Dataframe, we have Object, Float & Integer data types.

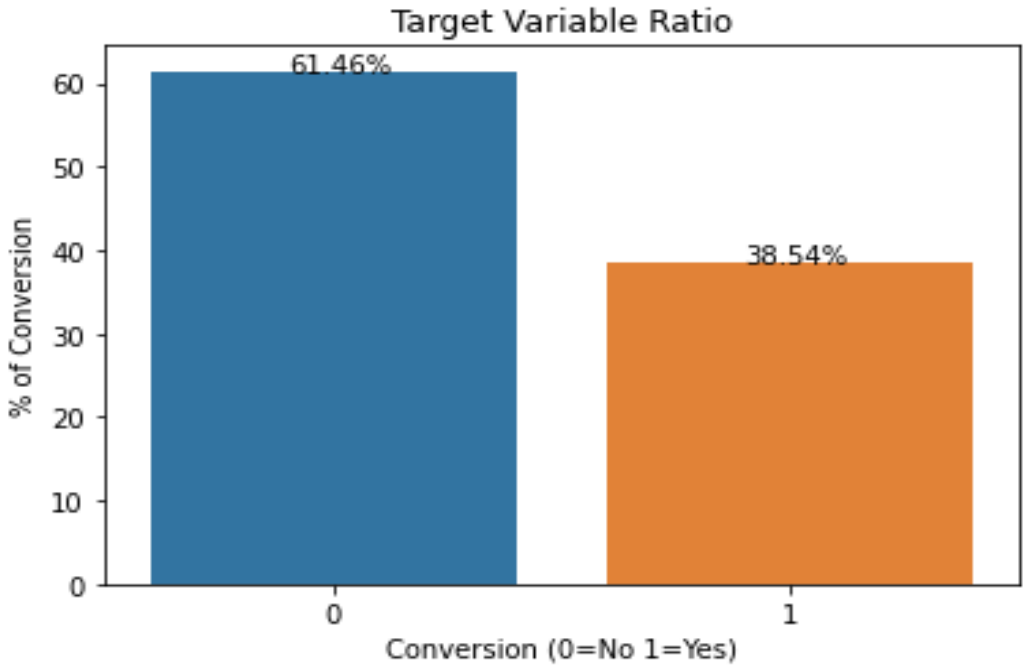
'Prospect ID','Lead Number' seems to be purely unique ID's, and will not make any significant contributions to our model results. So these variables should be dropped.

After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Exploratory Data Analysis

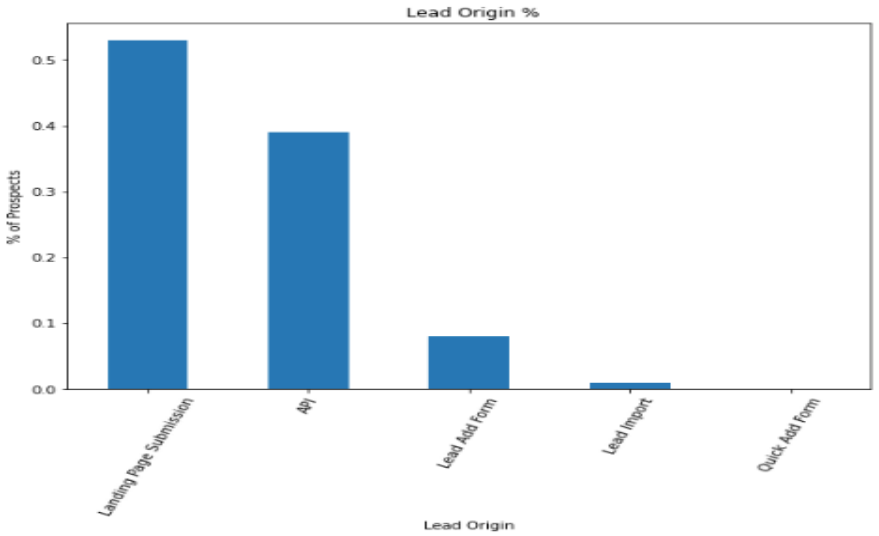
Our Target variable is having approx. 62:38 ratio, and seems to be properly balanced with respect to the conversion ratio.

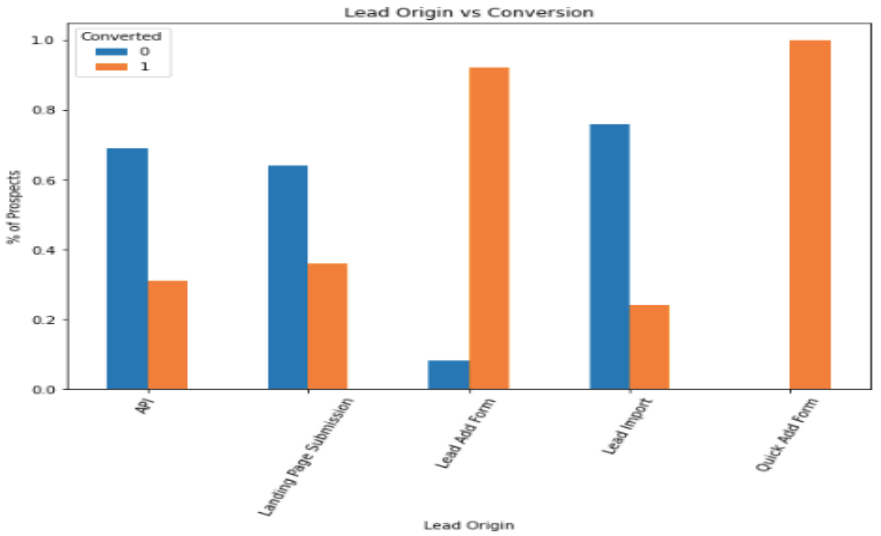


Exploratory Data Analysis (Categorical Variable)

Performed Univariate, Bivariate analysis and get the variable impact on lead conversion.

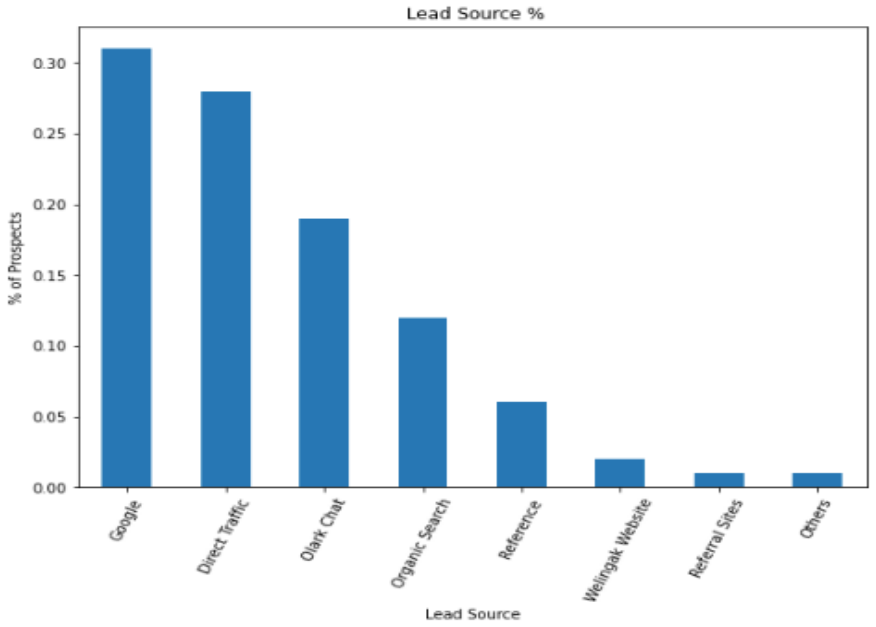
1. Univariate Analysis states that approx. 53% of the Lead Origin is from 'Landing Page Submission' followed by ~39% from API.
2. Bivariate Analysis states that 'Landing Page Submission' has 36% of Conversion and 'API' has 31% of Conversion.

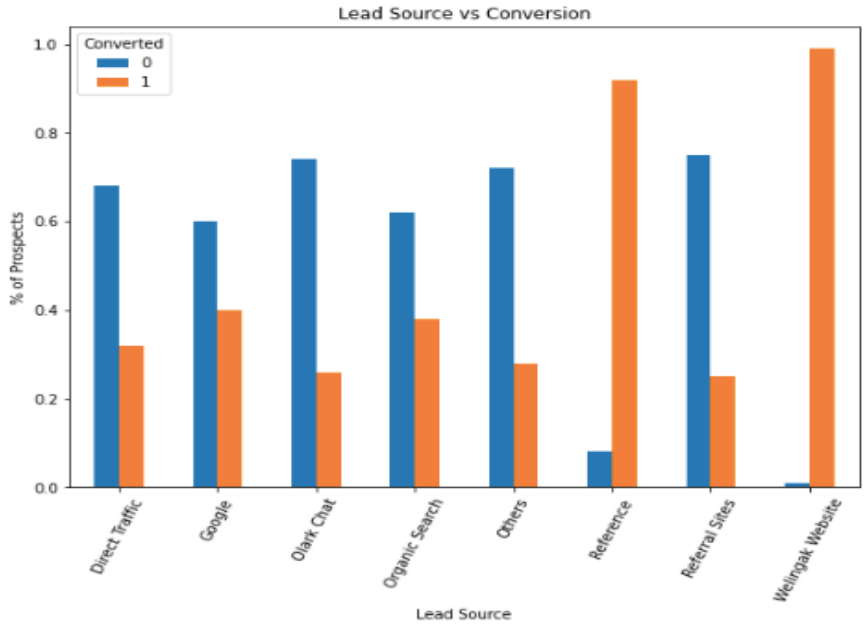




Exploratory Data Analysis (Categorical Variable)

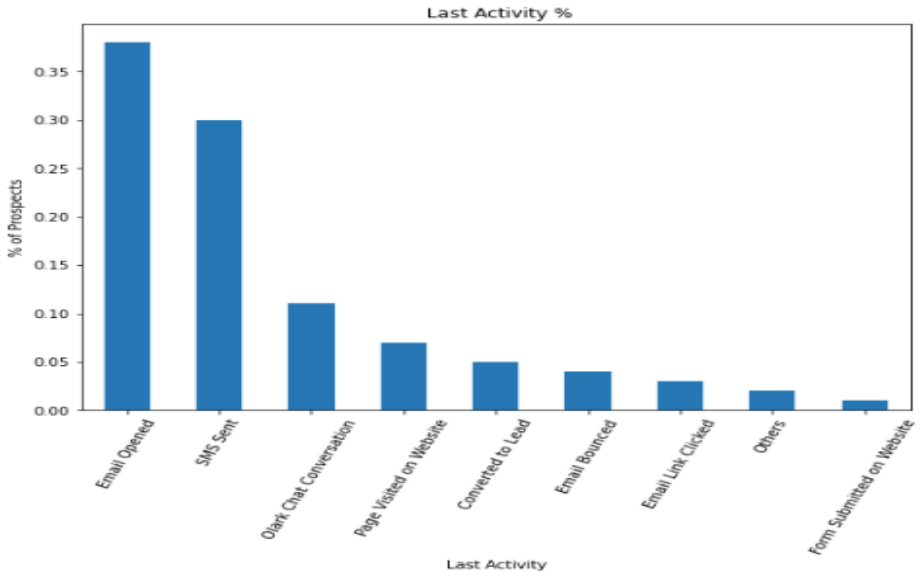
1. Univariate Analysis states that ~31% of the Lead Source is from 'Google' followed by ~28% from 'Direct Traffic'.
2. Bivariate Analysis states that 'Google' as a Lead Source has 40% of Conversion and 'Direct Traffic' has 32% of Conversion.

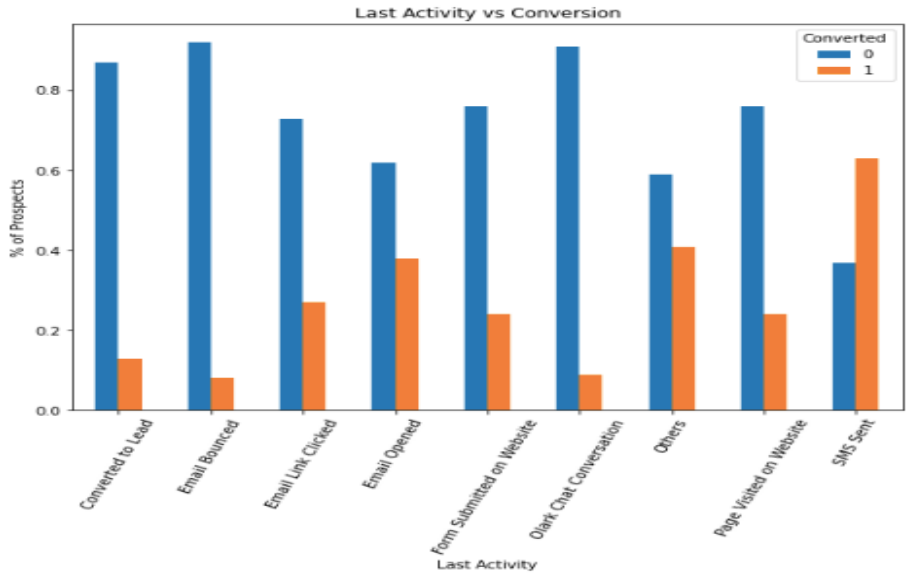




Exploratory Data Analysis (Categorical Variable)

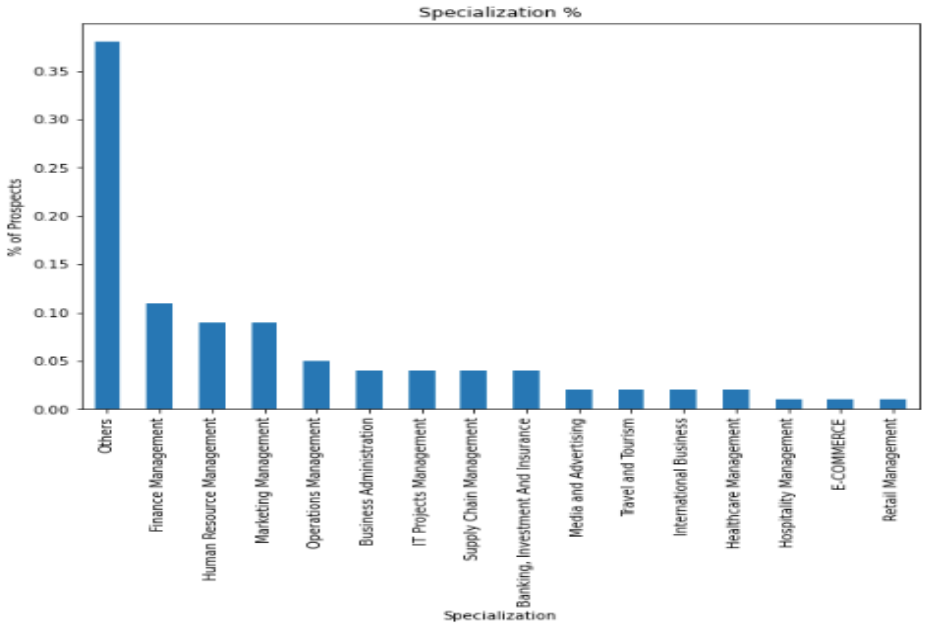
1. Univariate Analysis states that ~38% of the Last Activity is from 'Email Opened' followed by ~28% from 'SMS Sent'
2. Bivariate Analysis states that 'Email Opened' as a Last Activity has 38% of Conversion and 'SMS Sent' has 63% of Conversion. All calculations shown above.

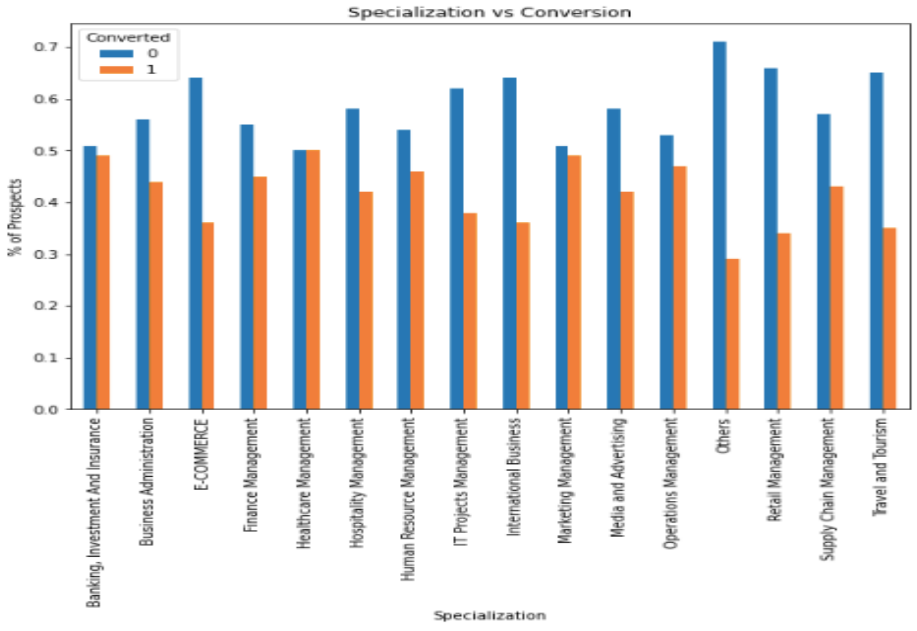




Exploratory Data Analysis (Categorical Variable)

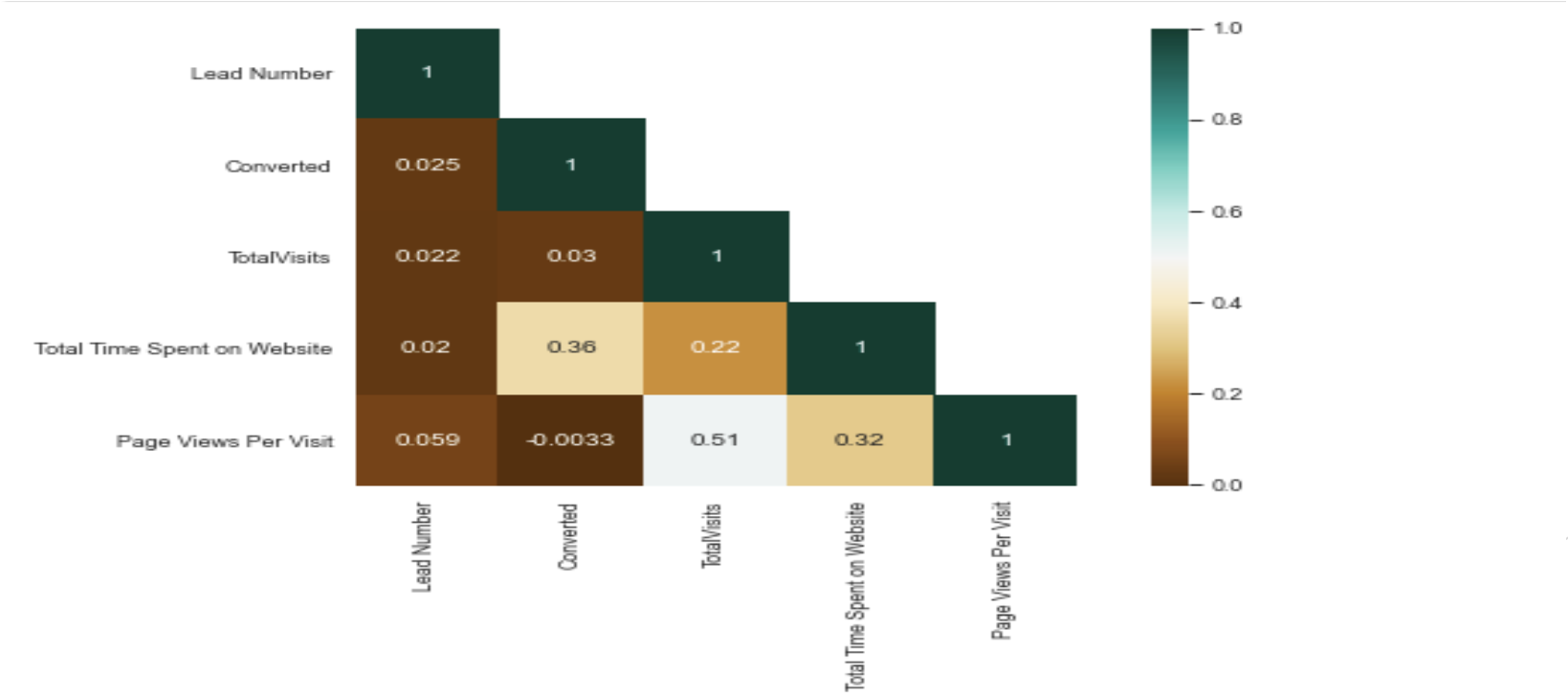
1. Univariate Analysis states that ~40% of the Specialization is from 'Others' category, followed by ~10% from 'Finance Management'
2. Bivariate Analysis states that 'Finance Management' as a Specialization has 45% of Conversion and 'Human Resource Management' has 46% of Conversion.





Exploratory Data Analysis (Numerical Variables)

Outlier Analysis and Treatment: The Heat Map tells us that there is a strong correlation between 'TotalVisits' & 'Page View Per Visit'. We will handle this during our Multi Collinearity check.

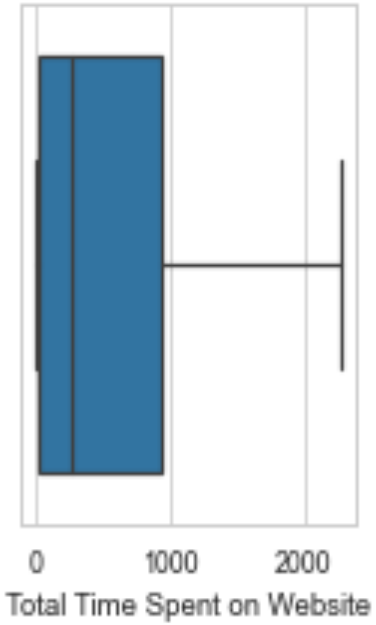


Exploratory Data Analysis (Numerical Variables)

We could see that TotalVisits and Page Views Per Visits has Outliers. As mentioned in the Outlier handling Approach above, we will impute it using IQR*1.5.

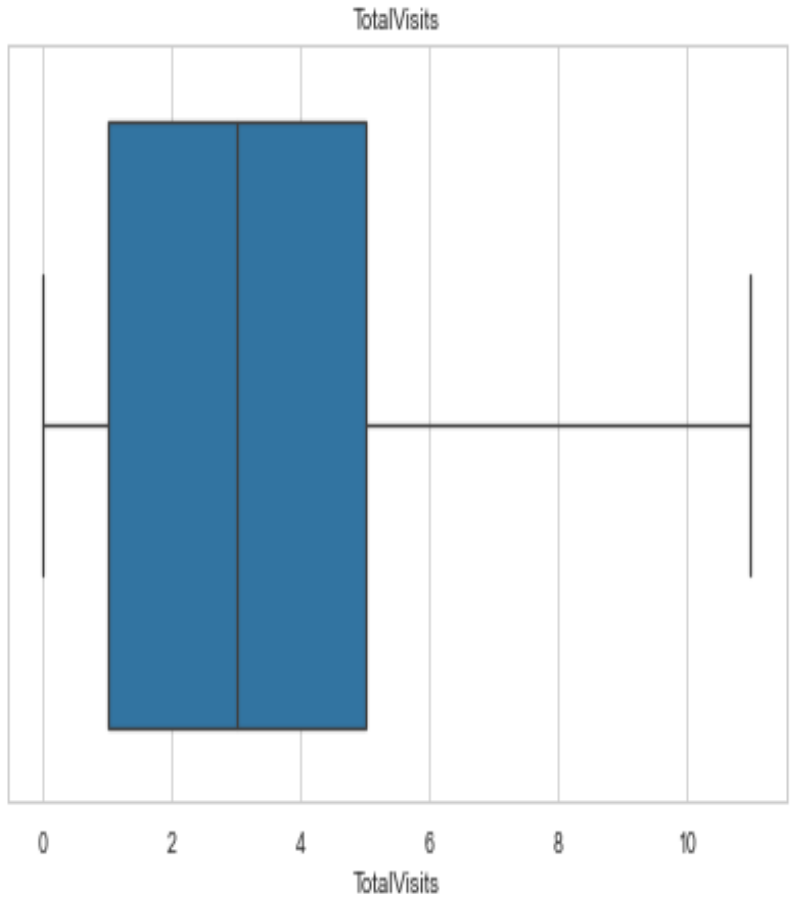
Plot 1: TotalVisits

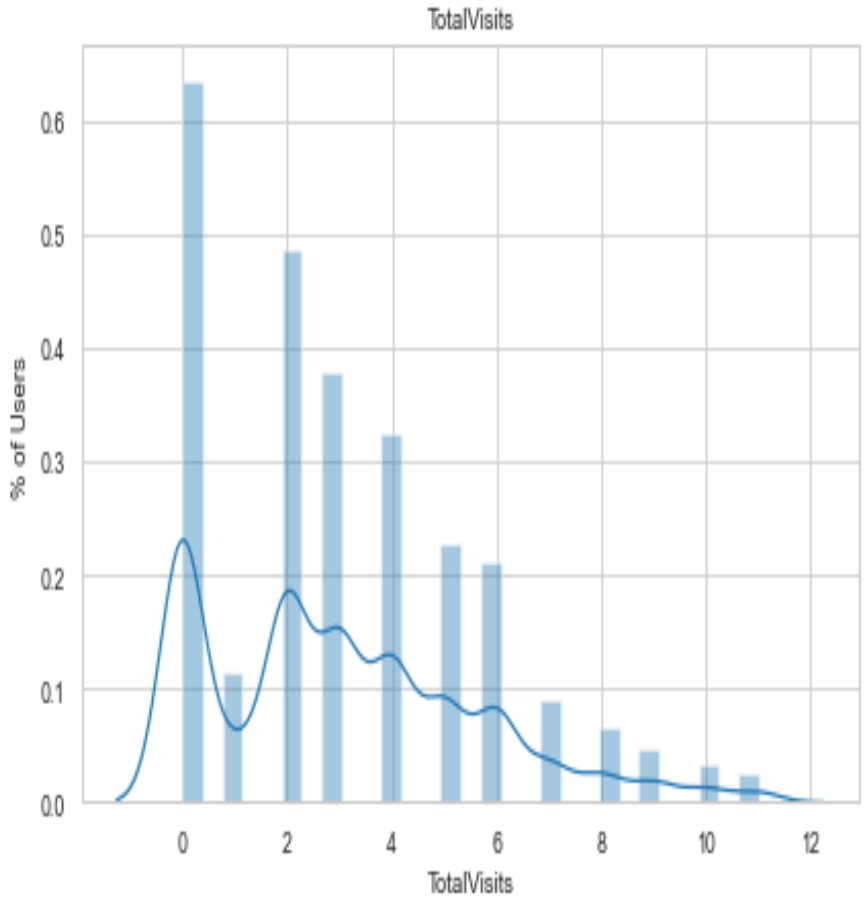
Plot 2: Total Time Spent on Website



Plot 3: Page Views Per Visit

Exploratory Data Analysis (Numerical Variables)





Model Building

1. Splitting the Data into Training and Testing Sets

2. The first basic step for regression is performing a train-test split, we have chosen 80:20 ratio.

3. Use RFE for Feature Selection.

4. Running RFE with 15 variables as output.

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

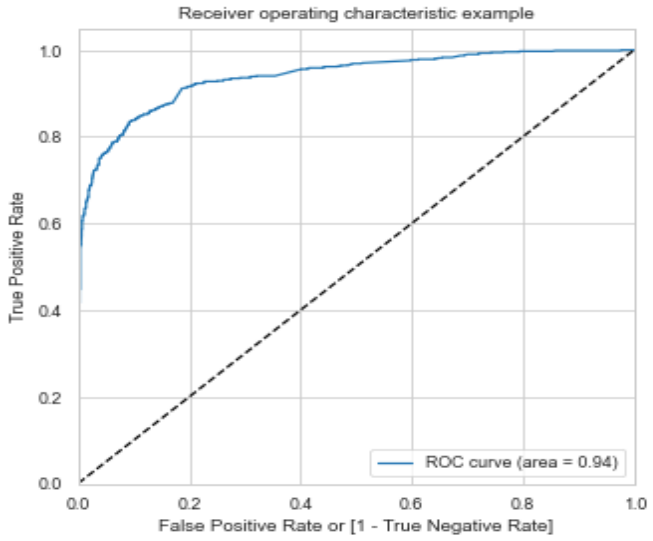
6. Predictions on test data set.

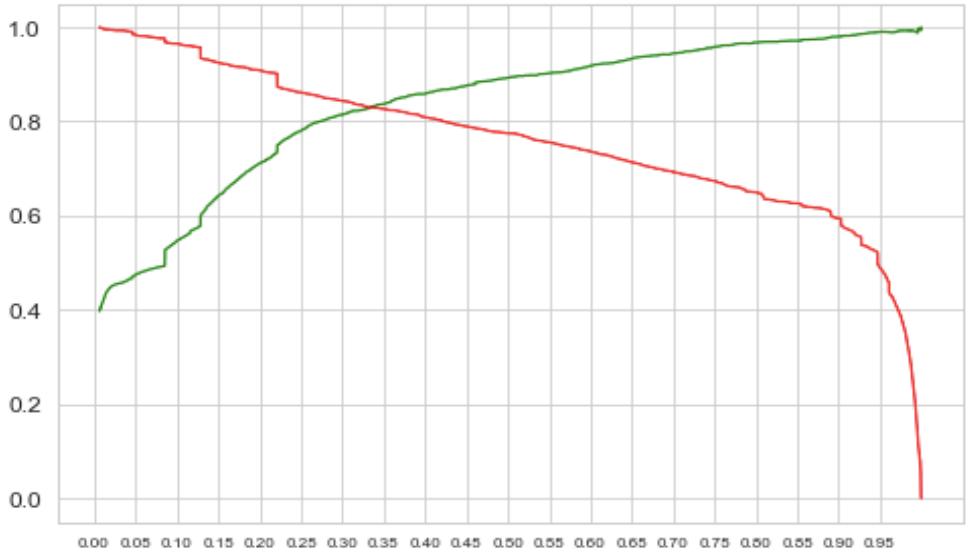
Model Summary

From the below curve, 0.34 is the optimal cut-off point to be taken as threshold.

Our final model gave us below performance:

1. ~84% of Recall value indicates that our model is able to predict 84% of actual conversion cases correctly.
2. ~85% of Precision value indicates that 85% of the conversions that our model predicted is actually converted.

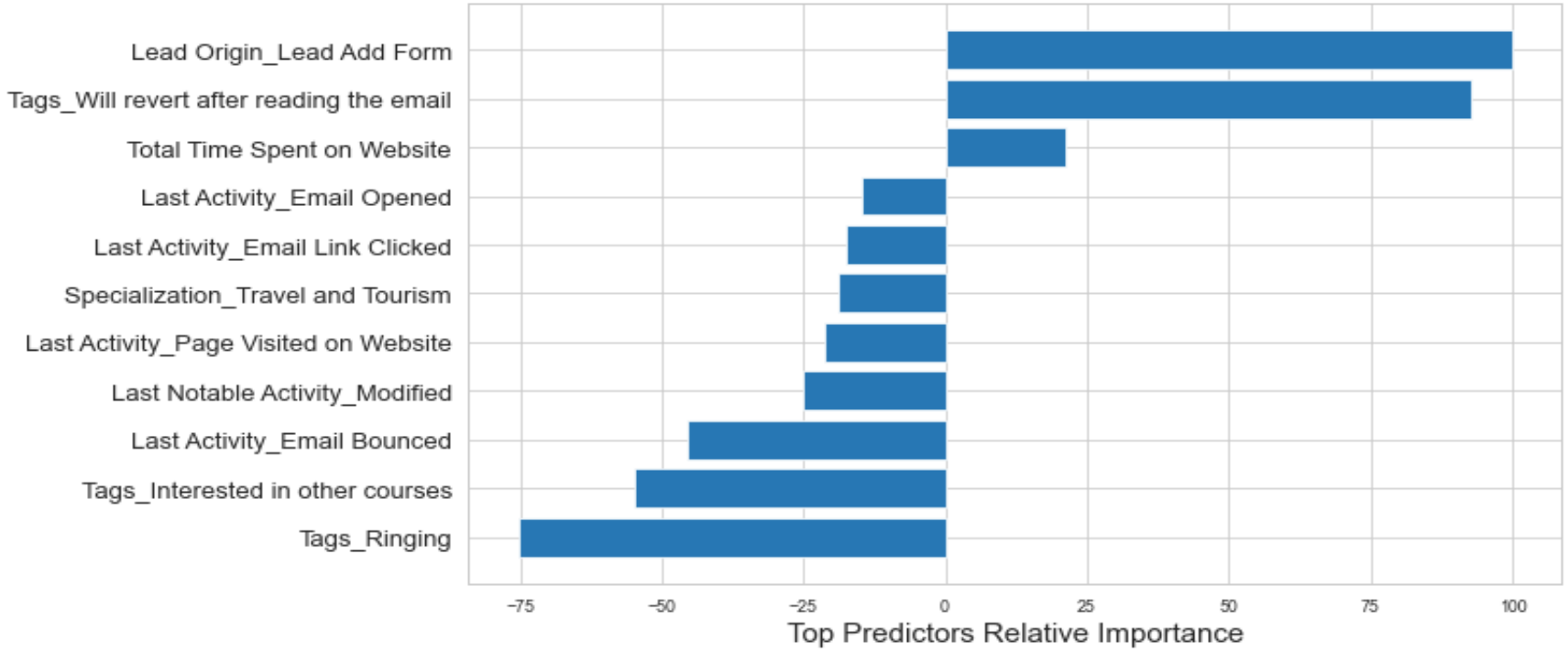




Top Three Features contribution in Lead conversion

Top three variables which contribute most towards the probability of a lead getting converted.

1. Lead Origin_Lead Add Form
2. Tags_Will revert after reading the email
3. Total Time Spent on Website



Interpretability using log odds

To find if any new lead will be converted or not, we could use our log odds formula along with our coefficient values of our final model.

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

Where ß0, ß1, ß3 are coefficient values and x1, x2, x3 are respective variable values.

	coef
const	-0.4190
Total Time Spent on Website	0.9544
Lead Origin_Lead Add Form	4.4630
Last Activity_Email Bounced	-2.0255
Last Activity_Email Link Clicked	-0.7772
Last Activity_Email Opened	-0.6567
Last Activity_Page Visited on Website	-0.9504
Specialization_Travel and Tourism	-0.8347
Tags_Interested in other courses	-2.4422
Tags_Ringing	-3.3614
Tags_Will revert after reading the email	4.1464
Last Notable Activity_Modified	-1.1154

Coefficient values from our final model

Thank You