Lead Scoring - Problem Statement

In this lead score case study, we have been given business problem of company named X Education which sells online courses to industry working professionals. On any given day, many professionals who are interested in the courses browse x Education website and browse for courses of study.

X Education markets its courses on several websites and search engines including Google. Once the prospect 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 contact detail such as email address or phone number, they are classified to be a lead. X Education also gets leads through the past referrals and students.

Once these leads are acquired online or via referral, employees from the X Education sales team start making calls, writing emails, teel-calls etc. Through this process, some of the leads get converted while most are not converted.

The typical lead conversion rate at X education is around 30%.

While X Education gets a lot of leads, its lead conversion rate remains 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 identified potential leads rather than making calls to everyone who visited website.

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.

We need to build a logistic regression model to given data-frame to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot and need to be persued meaning it is very likely to convert.

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

Objectives of case study

The goals of this case study are as follows.

To build a logistic regression model to given data-frame to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot and need to be persued meaning it is very likely to convert. On the other hand, a lower score would mean that the lead is cold and shall mostly not get converted.

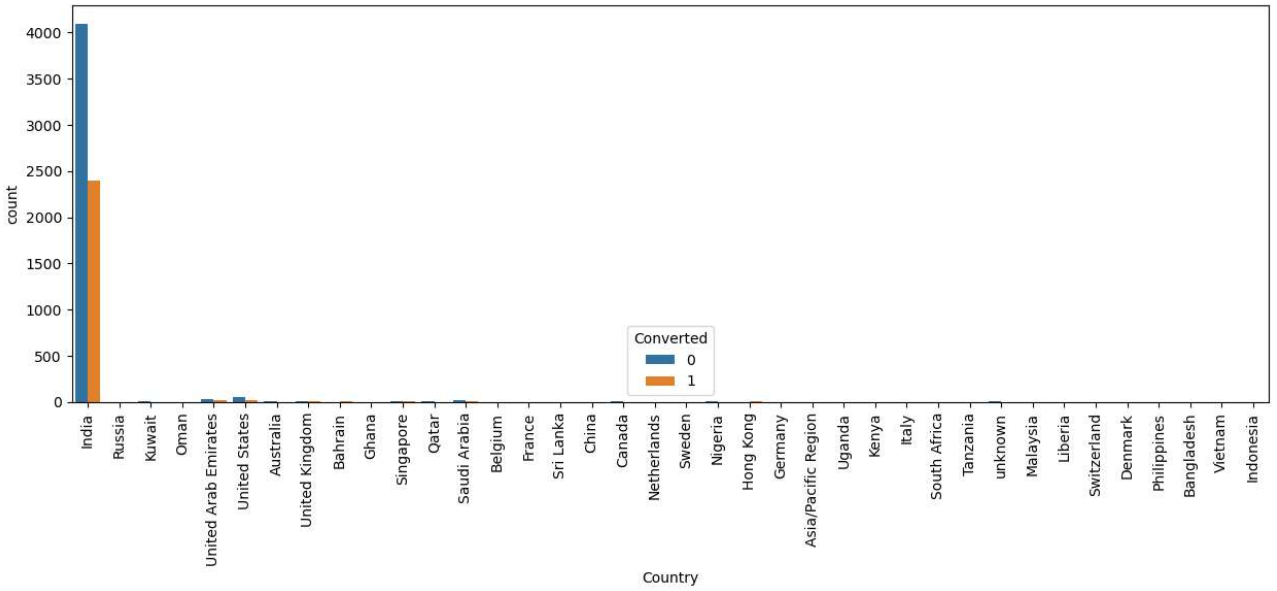
There are some more problems presented by the X Education and our model need to be able to adjust to. if the company's requirement changes in the future so we shall need to handle these issues as well.

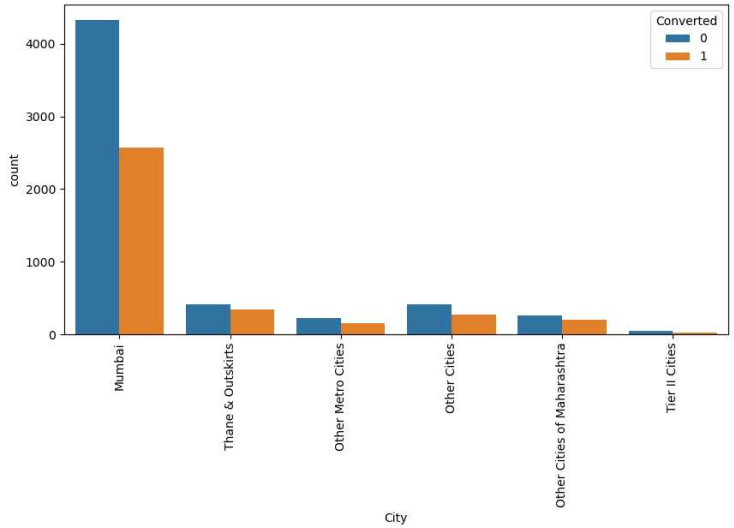
Methodology / Steps Followed

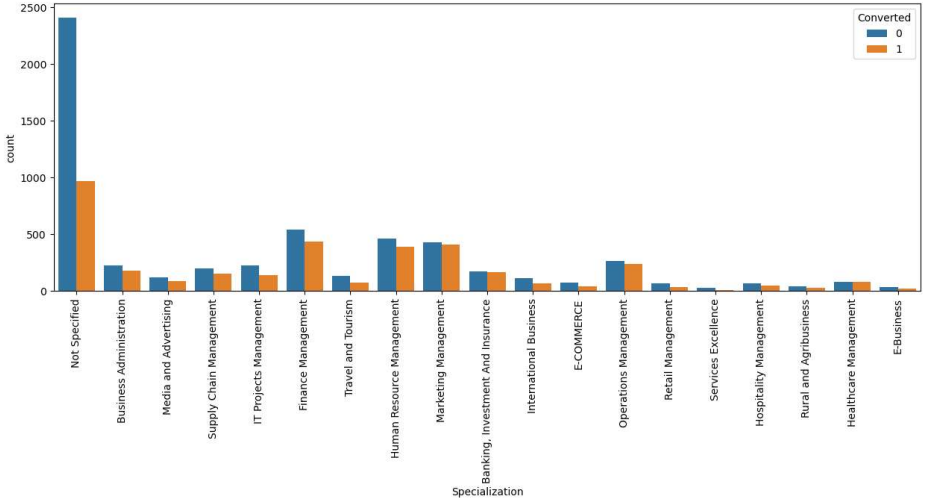
- Reading Data given in this assignment
- Cleaning Data based of use case analysis requirement (handling null values & removing higher null value data
- Removing redundant columns in the dataset
- Imputing numm values
- Undertake EDA and analyze data
- Creating Dummy variables
- Feature standardisation
- Splitting data into train and test set
- Building Model – building strong predictive model
- Making Predictions based on analysis on above steps
- Model Evaluation
- Plot ROC Curve
- Prediction on test set
- Precision - Recall

Exploratory Data Analysis

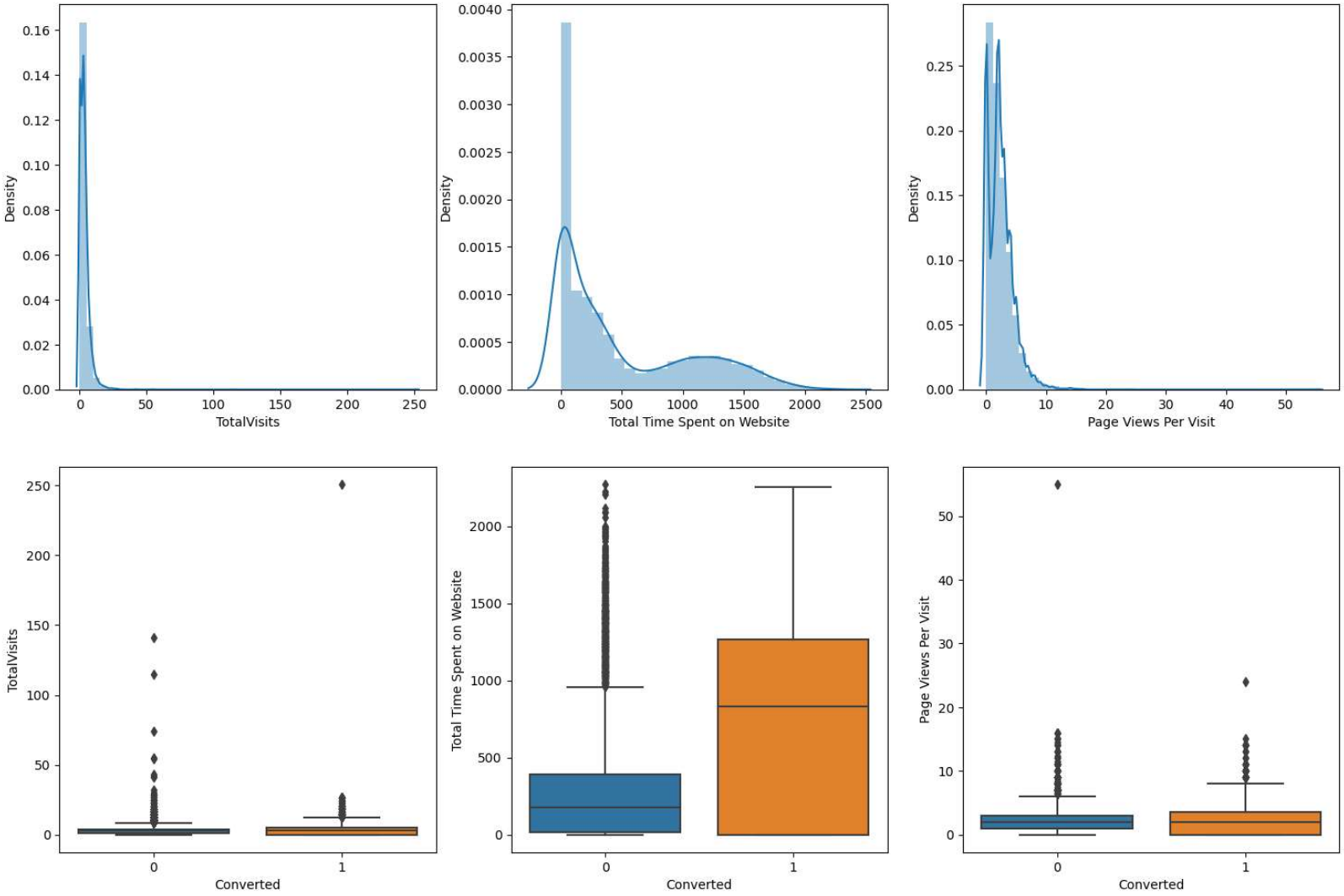
Database dimension : (9240, 37)
Database size : 341880
Number of Row : 9240
Number of Columns : 37



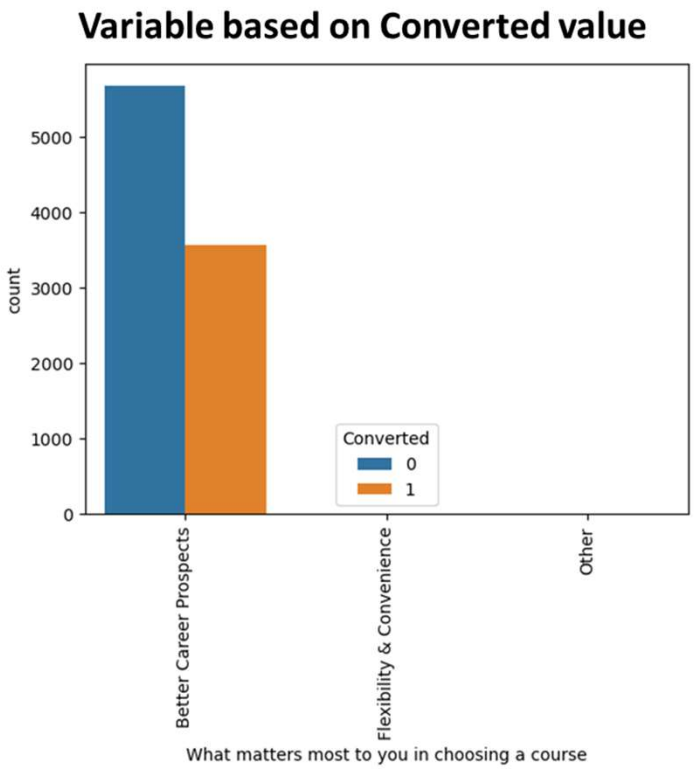


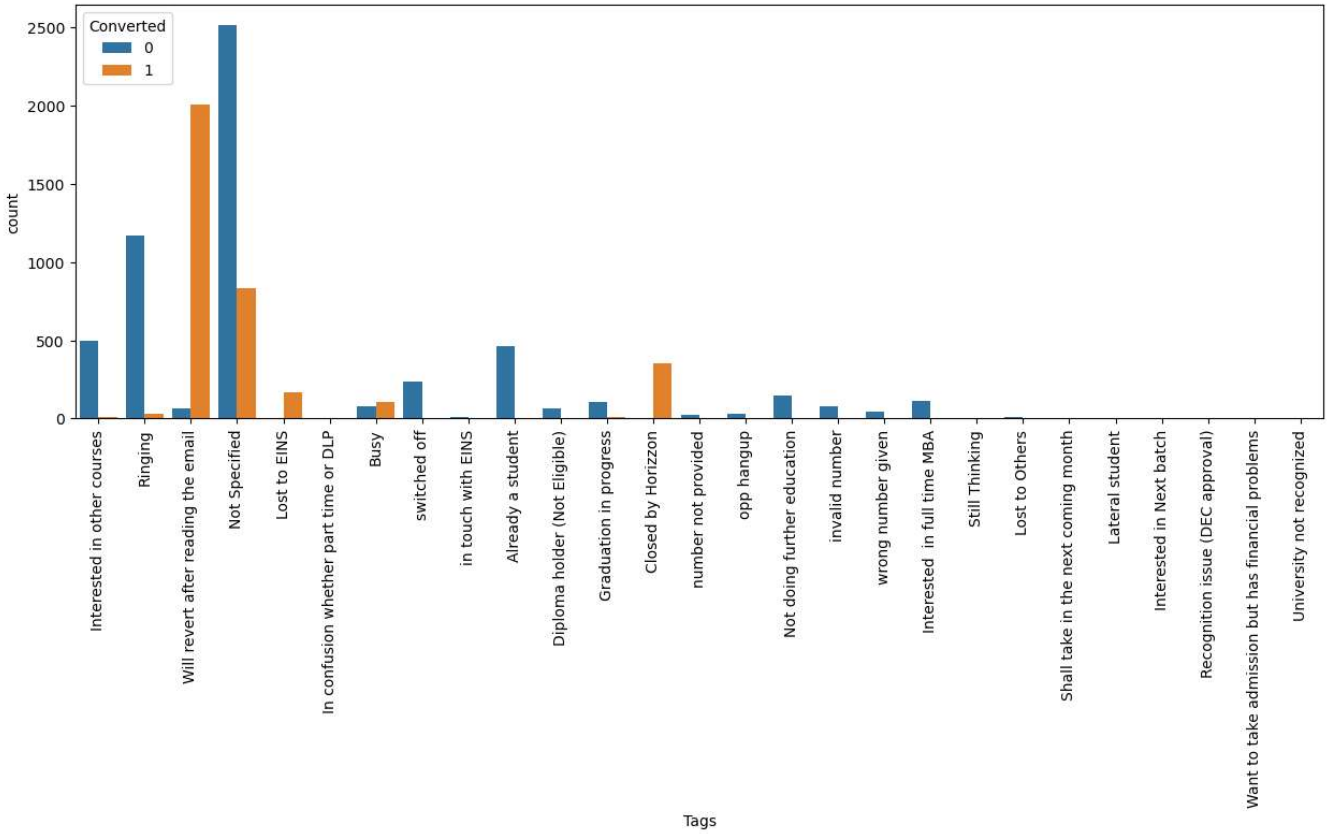


Exploratory Data Analysis

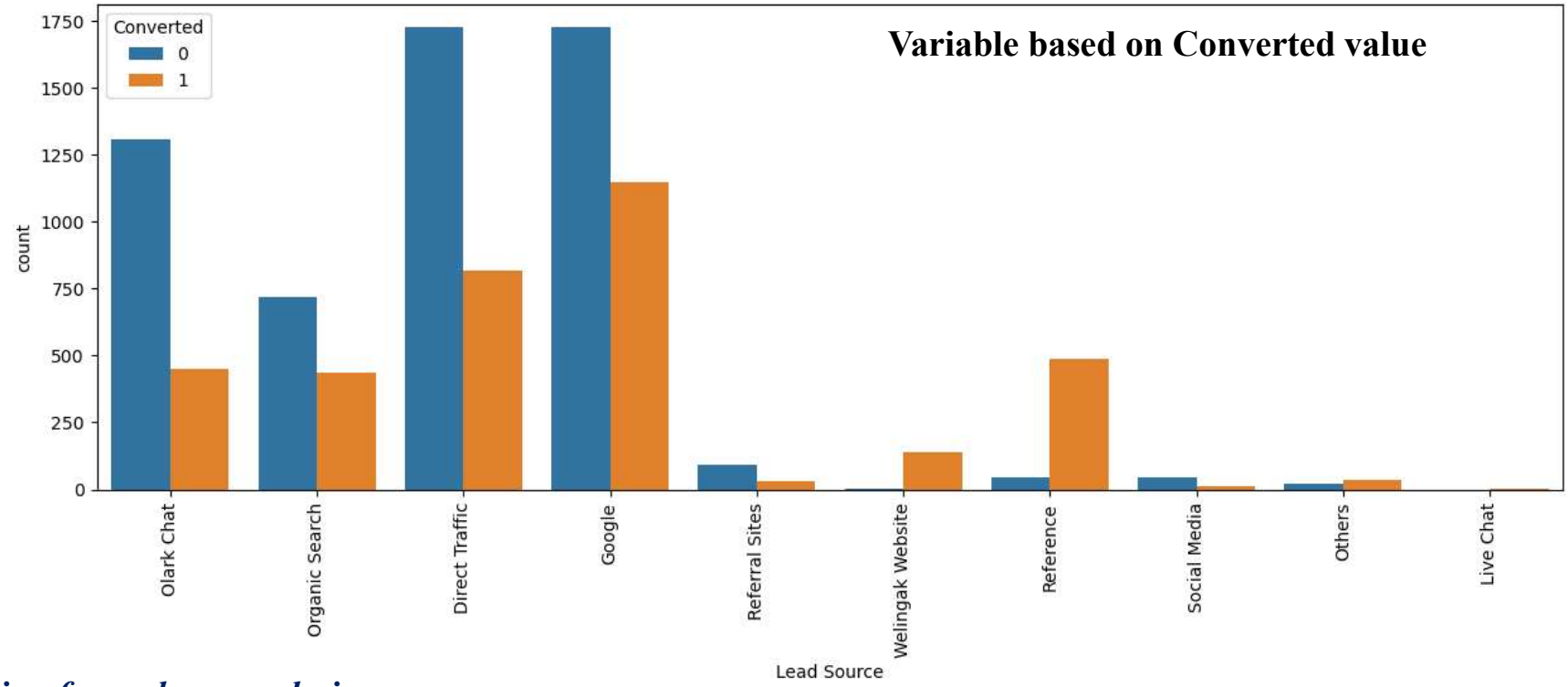


Exploratory Data Analysis





Exploratory Data Analysis

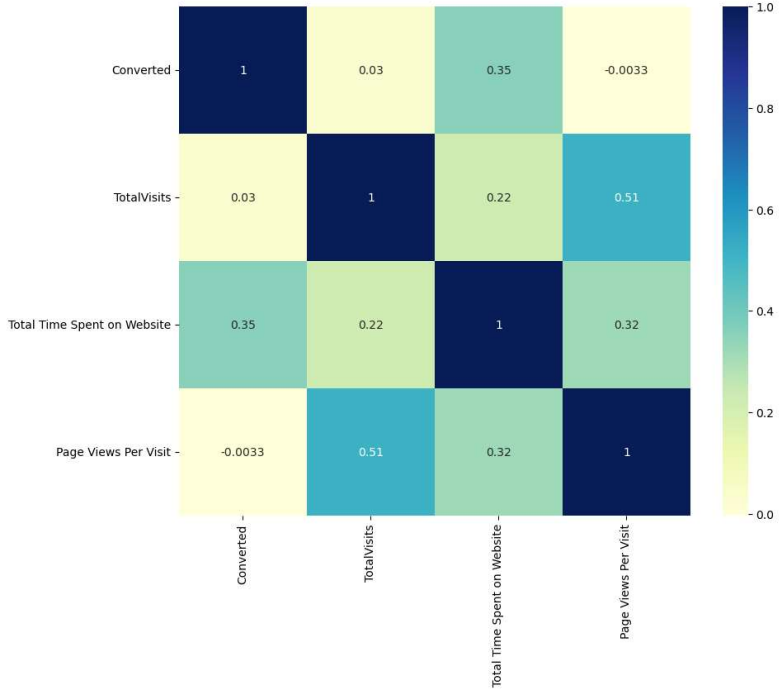


Observation from above analysis

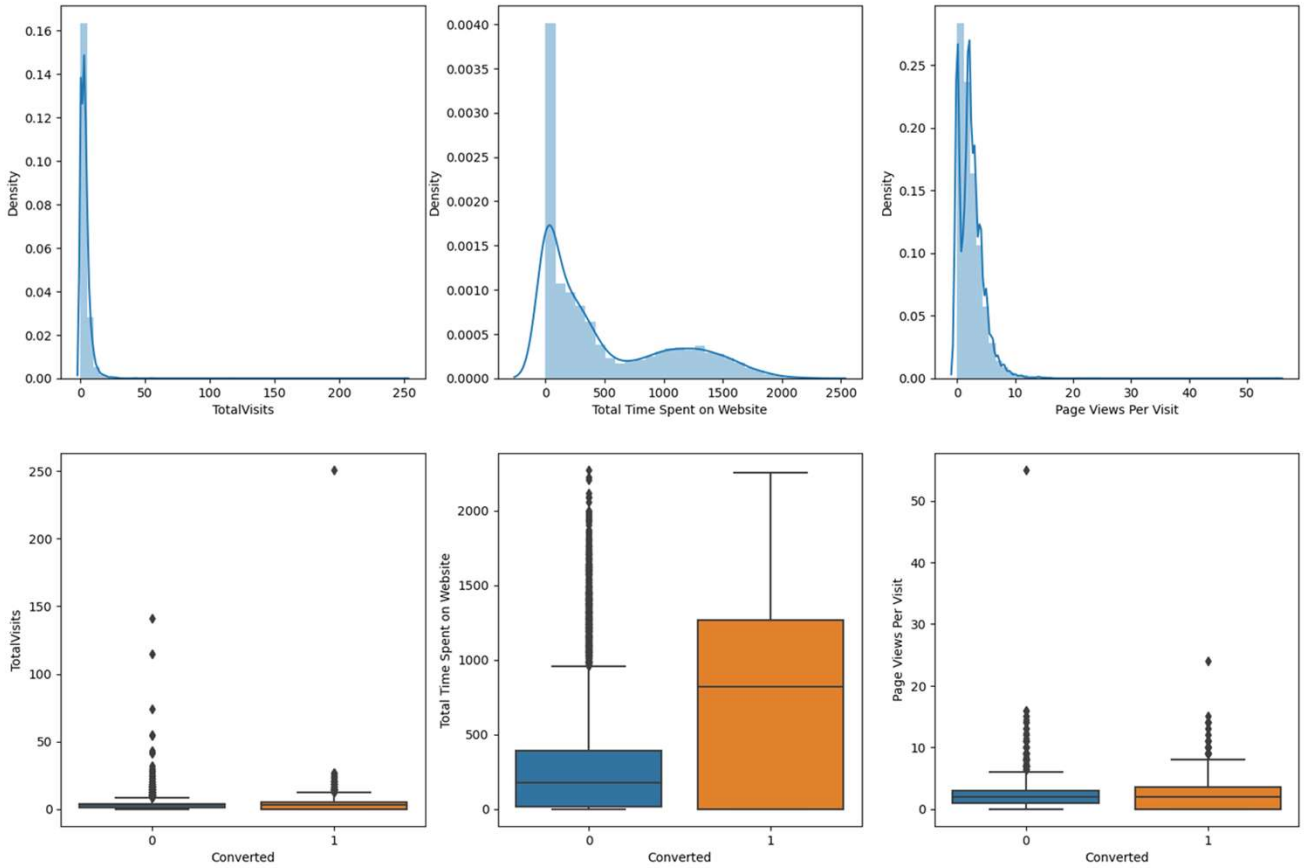
- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

Exploratory Data Analysis

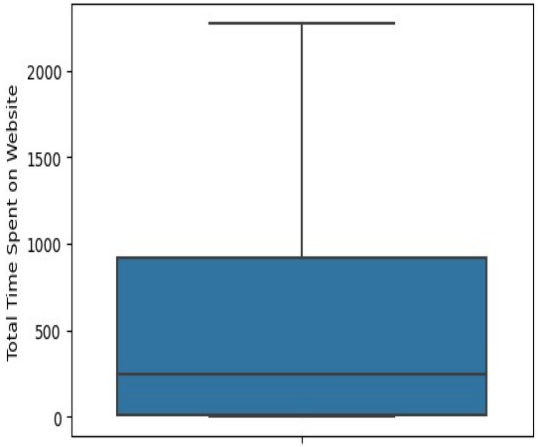
Correlations of numeric values



'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'

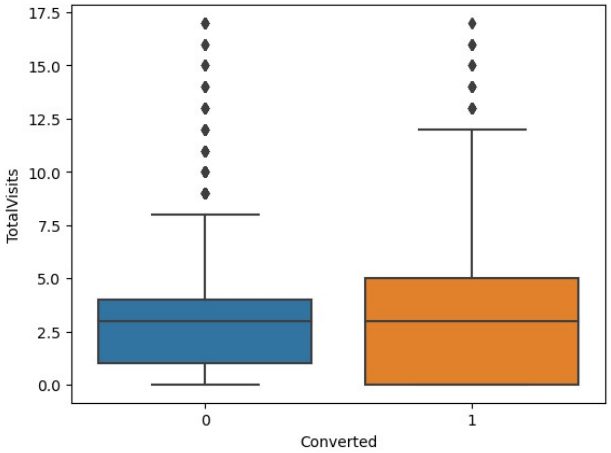


Exploratory Data Analysis



After outlier treatment

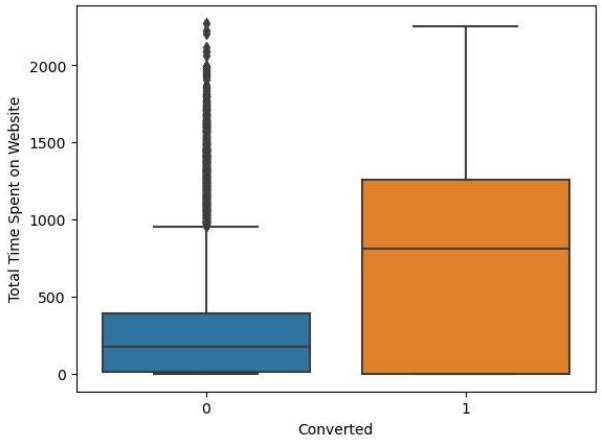
Since there are no major Outliers for the above variable we don't do any Outlier Treatment for this above Column



Observation from above analysis

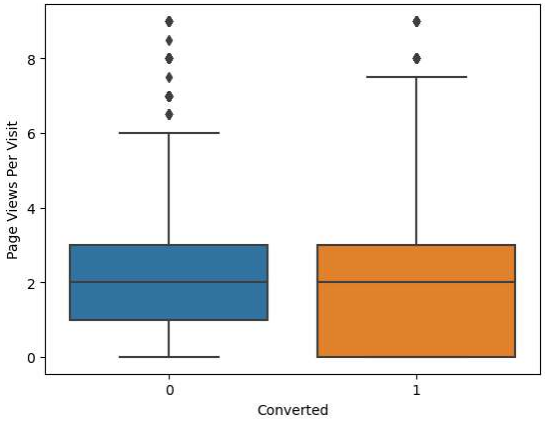
•Median for converted and not converted leads are the close.

•Nothng conclusive can be said on the basis of Total Visits



Observation

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.



Observation

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

Model Building

- ❑ Splitting into train & Test sets
- ❑ Scale variables in train data set
- ❑ Build first model
- ❑ Use RFE to eliminate less relevant variables
- ❑ Build the next model
- ❑ Eliminate variables based on high p-values
- ❑ Check VIF value for all the existing columns
- ❑ Predict using train data set
- ❑ Evaluate accuracy & other metric
- ❑ Predict using test set
- ❑ Precision and recall analysis on the test prediction

BUILDING MODEL 3 (Final Model after treatment)

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6253
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1263.3
Date:	Mon, 23 Jan 2023	**Deviance:**	2526.6
Time:	23:01:35	**Pearson chi2:**	8.51e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6037
Covariance Type:	nonrobust		

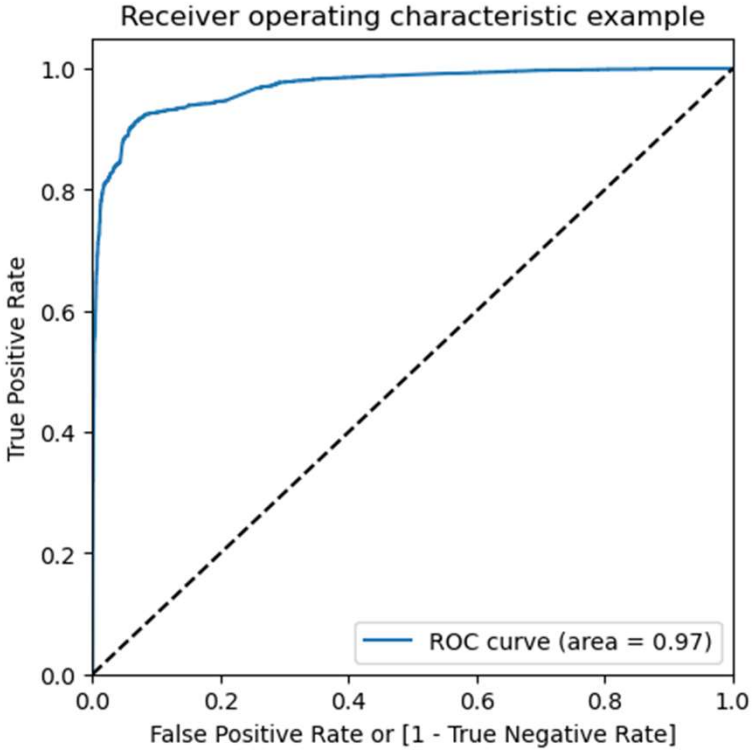
	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

Coefficients

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1179	0.084	-13.382	0.000	-1.282	-0.954
Total Time Spent on Website	0.8896	0.053	16.907	0.000	0.786	0.993
Lead Origin_Lead Add Form	1.6630	0.455	3.657	0.000	0.772	2.554
Lead Source_Direct Traffic	-0.8212	0.127	-6.471	0.000	-1.070	-0.572
Lead Source_Welingak Website	3.8845	1.114	3.488	0.000	1.701	6.068
Last Activity_SMS Sent	1.9981	0.113	17.718	0.000	1.777	2.219
Last Notable Activity_Modified	-1.6525	0.124	-13.279	0.000	-1.896	-1.409
Last Notable Activity_Olark Chat Conversation	-1.8023	0.491	-3.669	0.000	-2.765	-0.839
Tags_Closed by Horizzon	7.1955	1.020	7.053	0.000	5.196	9.195
Tags_Interested in other courses	-2.1318	0.406	-5.253	0.000	-2.927	-1.336
Tags_Lost to EINS	5.9177	0.611	9.689	0.000	4.721	7.115
Tags_Other_Tags	-2.3737	0.206	-11.507	0.000	-2.778	-1.969
Tags_Ringing	-3.4531	0.238	-14.532	0.000	-3.919	-2.987
Tags_Will revert after reading the email	4.5070	0.188	24.002	0.000	4.139	4.875

All Values are in order so now, Moving on to derive the Probabilities, Lead Score, Predictions on Train Data:

Train Data Set



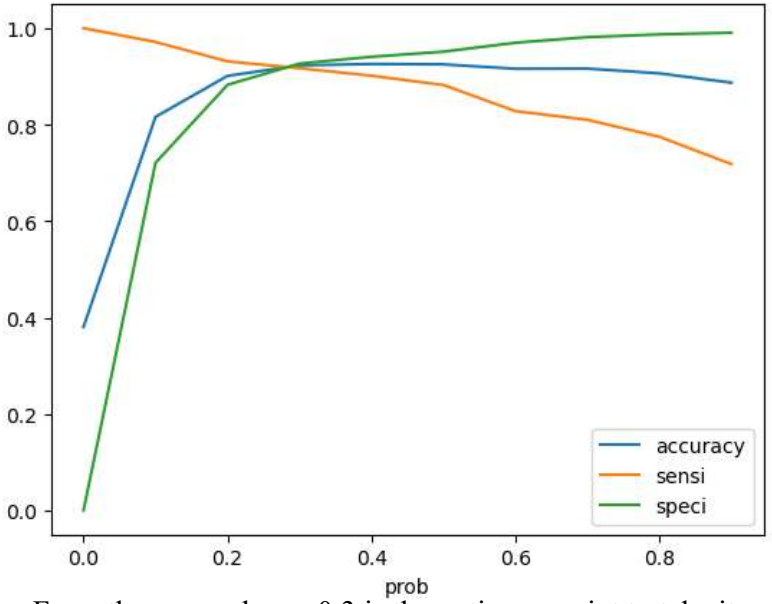
The ROC Curve should be a value close to 1.

From our analysis above we are getting a good value of 0.97 indicating a good predictive model.

Finding Optimal Cut-off Point

Above we had chosen an arbitrary cut-off value of 0.5. We need to determine the best cut-off value and the below section deals with that:

Plotting accuracy sensitivity and specificity for various probabilities.



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

Observation based on above analysis:

Based on analysis undertaken in steps above, the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:

•Accuracy : 92.29%
•Sensitivity : 91.70%
•Specificity : 92.66%

Findings from Case Study: After running the model on the Test Data our finding of key figures are as follows:

Final Observation:

Let us compare the values obtained for Train & Test:

Train Data:
Accuracy : 92.29% Sensitivity : 91.70% Specificity : 92.66%

Test Data:
Accuracy : 92.78% Sensitivity : 91.98% Specificity : 93.26%

- Precision Score : 89.92%
- Recall Score : 91.98%

Interpretation Logistic regression model with multiple predictor variables

In general, we can have multiple predictor variables in a logistic regression model as below:

$$logit(p) = log(p/(1-p)) = \beta 0 + \beta 1 * X1 + \ldots + \beta n * Xn$$

•Applying model to our example dataset, each estimated coefficient is the expected change in the log odds of being a potential lead for a unit increase in the corresponding predictor variable holding the other predictor variables constant at a certain value.

•Each exponentiated coefficient is the ratio of two odds, or the change in odds in the multiplicative scale for a unit increase in the corresponding predictor variable holding other variables at a certain value.

The magnitude and sign of the coefficients loaded in the logit function (Model 3):

*logit(p) = log(p/(1-p) = (7.2 * Tags_Closed by Horizzon) + (5.92 * Tags_Lost to EINS) + (4.51 * Tags_Will revert after reading the email) + (3.88 * Lead Source_Welingak Website) + (2.00 * Last Activity_SMS Sent) + (1.66 * Lead Origin_Lead Add Form) + (0.89 * Total Time Spent on Website) - (0.82 * Lead Source_Direct Traffic) - (1.65 * Last Notable Activity_Modified) - (1.80 * Last Notable Activity_Olark Chat Conversation) - (2.13 * Tags_Interested in other courses) - (2.37 * Tags_Other_Tags) - (3.45 * Tags_Ringing) - 1.12*

Conclusions of Study

Exploratory Data Analysis

- People spending higher than average time are promising leads so X Education can target them to help higher conversion
- SMS messages have been seen to have high impact to conversion of lead
- Landing page submission can help finding out more leads
- Marketing management and HR have high conversion rates. X Education should focus on these people as from these specializations can be promising leads
- Offers for referencing and references can be good source for higher rate of conversions
- Alert messages or information has been seen to have high lead conversion rate and X Education can treat these as action points

Logistics regression model

- The model shows high accuracy score > 90%
- The threshold has been selected from accuracy, sensitivity, specificity measures & precision, recall curves
- The model shows high sentivity and specificity

Train Data: Accuracy : 92.29% Sensitivity : 91.70% Specificity : 92.66%
Test Data: Accuracy : 92.78% Sensitivity : 91.98% Specificity : 93.26%

- The model finds correct promising leads that have less chances of getting converted
- We observe overall prediction of this model proves to be highly accurate

Thanks for attention