X-Education
Lead Conversion Case Study

By- Chinmay Kumar Sahu
Date- 26/08/2023

AGENDA

1. Problem Statement
2. Objective
3. EDA Approach
4. Model Building
5. Final Observation
6. Conclusion

1. Problem Statement

- Company named X Education gets a lot of leads. However, its lead conversion rate is very poor. 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.

2. Objective

X Education has appointed me to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The objective is to build a model wherein I 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.

3. EDA Approach

3.1. Importing Necessary libraries for EDA. :

- Import the libraries which will help for analysis such as Numpy, Pandas, Seaborn, matplotlib & Warnings.

3.2. Dataset Loading. :

- *Leads.csv* : contains all the relevant information of the customers who are enquiring about the courses.

3.3. Dataset Understanding. :

- Shape (Rows:9240, Features:37)
- Using.describe() function to get the statistical information & distribution
- Data types available:
  - float64(4)
  - int64(3)
  - object(30)

3. EDA Approach

3.4. Dataset Wrangling:

- Data has been inspected & all sanity check has been performed.
- Handling of Null/ NA values has been done on the data set.
- Upon further analysis, it was found that 5 features has more than 40% data missing from the dataset.
  (The number 40 % was chosen by doing trails on the data)
- Hence only 4 out of 5 features has been dropped to make the dataset viable. This was achieved by using the "*Leads Data Dictionary.xlsx*" file.
- Again Null values were checked & were imputed based on the type of data [numerical (MODE) & categorical (MEDIAN)]for 13 features.
- 'Select' was replaced with a null value since it did not give us much information & was instructed before to be taken care off.

3. EDA Approach

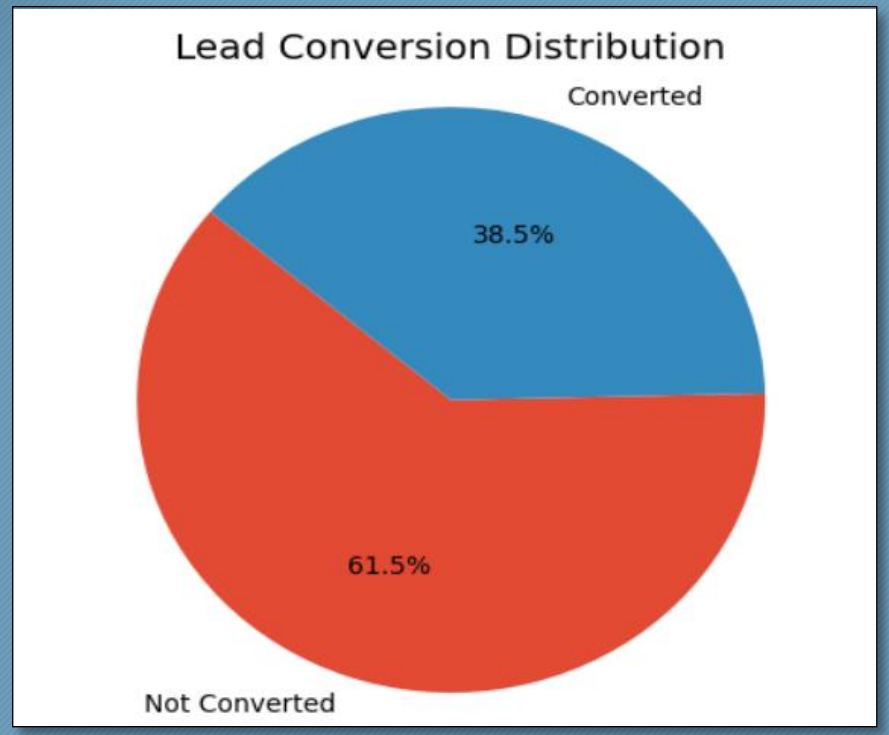
3.5. Data Visualization :

- *Dataset was checked for data imbalance at first.*
- *The numeric values seem good but had outliers this was found with box plot.*
- *Performed Univariate Analysis for both Continuous and Categorical variables.*
- *Bivariate Analysis with respect to Target variable.*
- *Correlation Matrix was made for quick check of trend & correlation.*
- *As per the findings few columns were dropped before moving to next phase.*

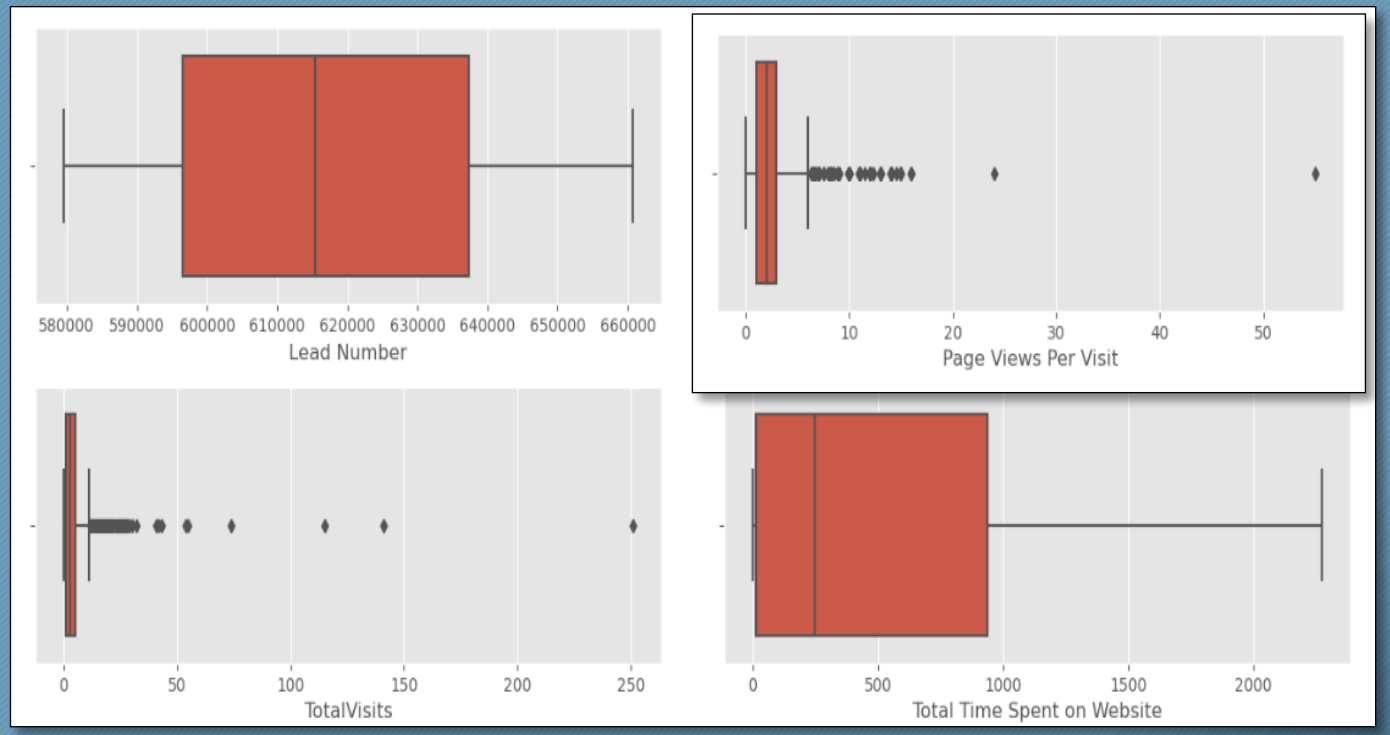
Some of the major graphs will be shown in the next coming slides.

3. EDA Approach

3.5. Data Visualization :



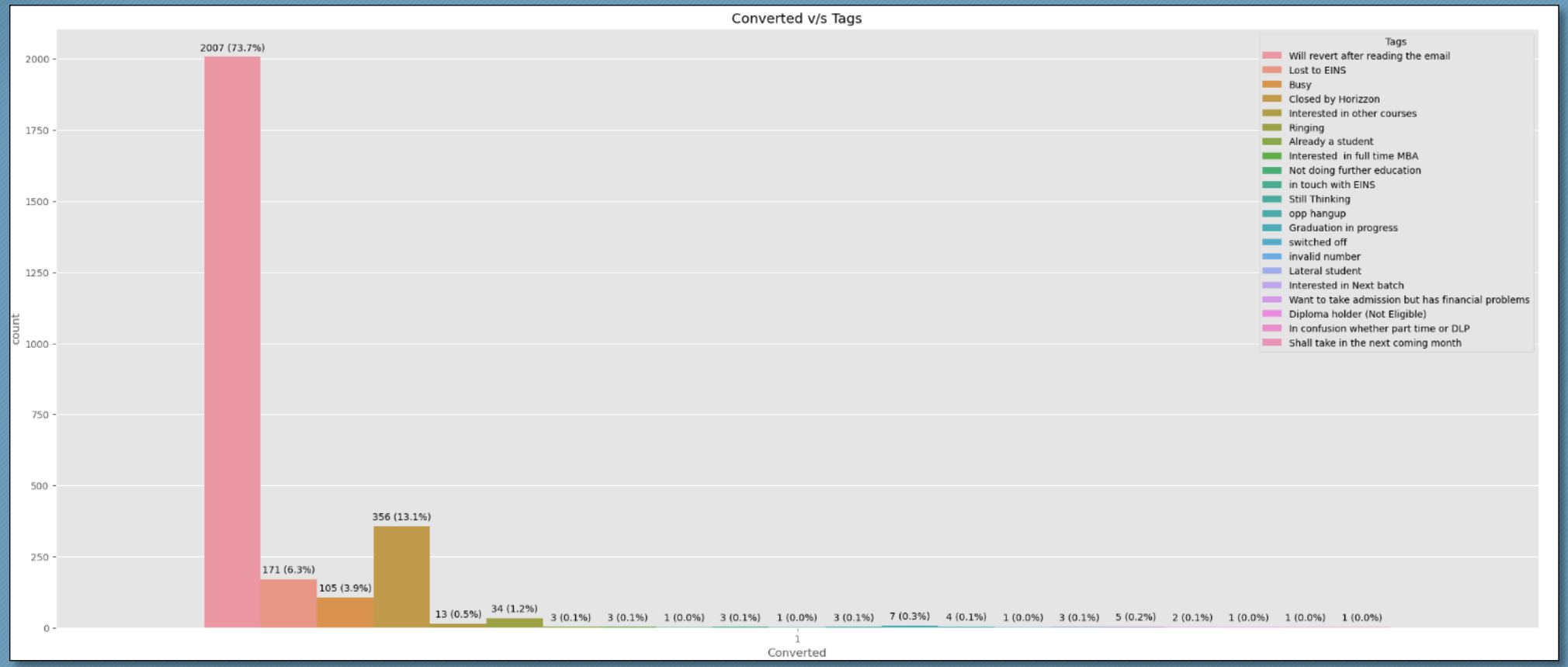
As the target feature **Converted** value count indicates a good distribution of conversion(1) & non conversion(0). So, it is concluded that data is not highly imbalance.



Only two features has outliers which is **TotalVIsits** & Page **View Per Visit**.

3. EDA Approach

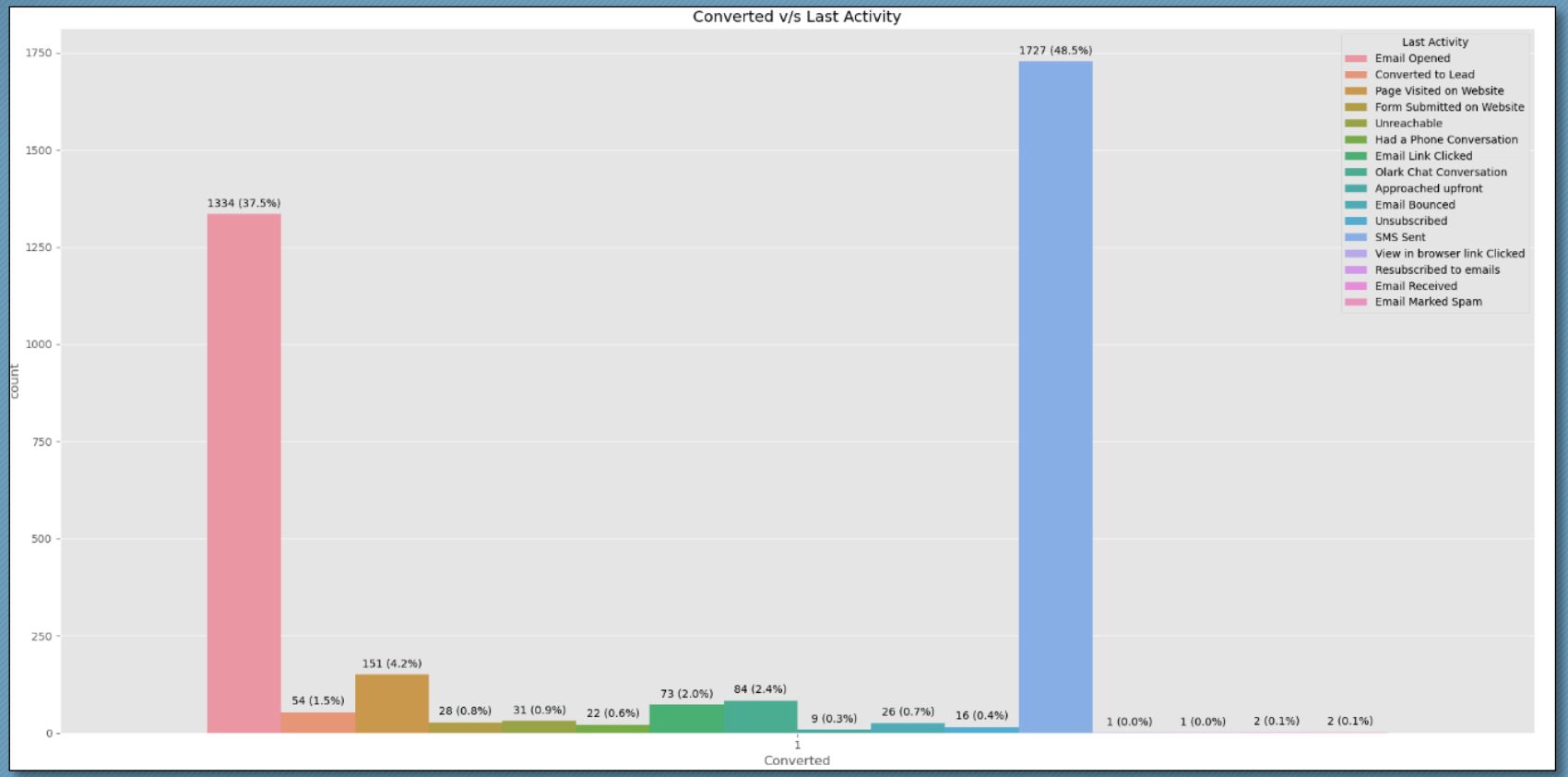
3.5. Data Visualization :



In **Tags** feature “will revert back after reading the email ” was found to be one of the major conversions..

3. EDA Approach

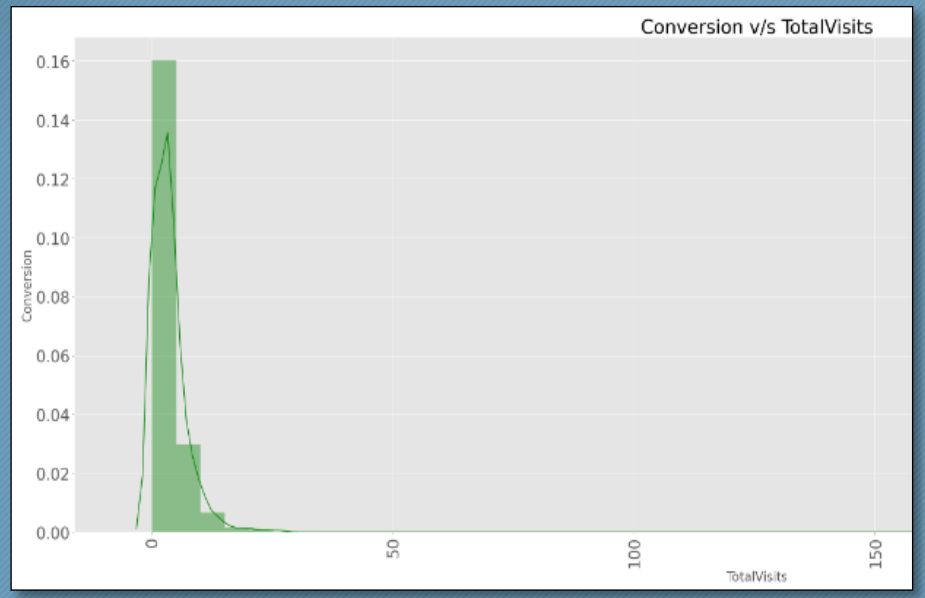
3.5. Data Visualization :



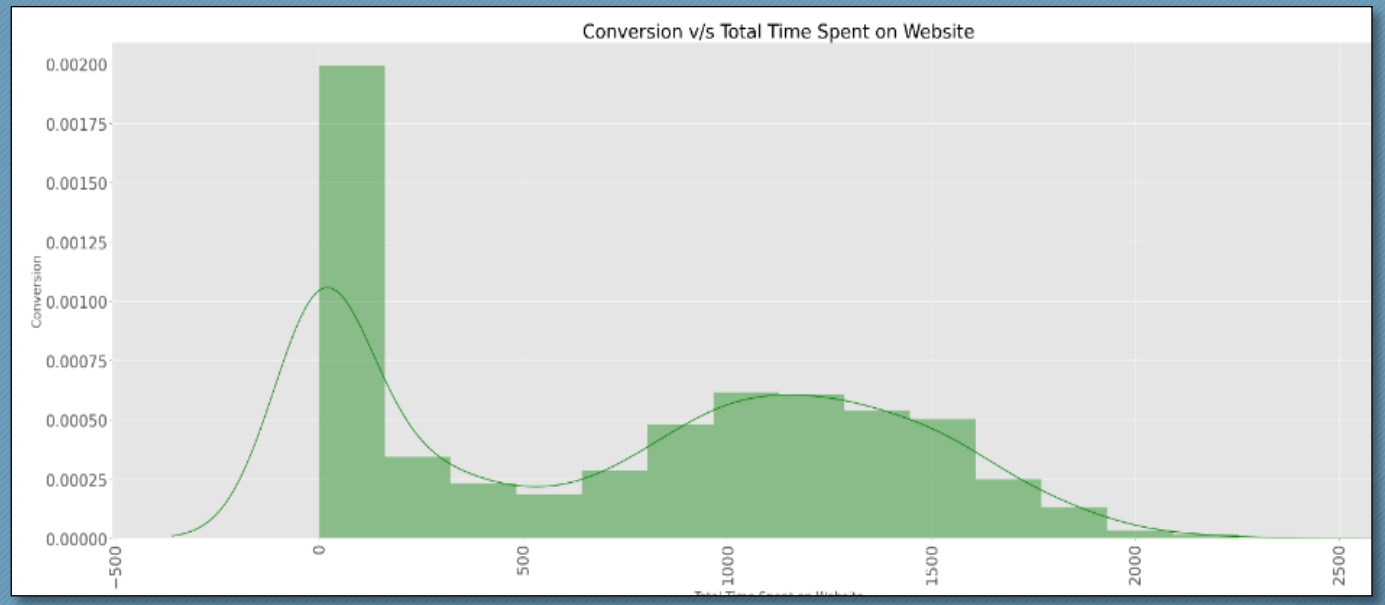
In **Last Activity** feature “SMS sent” was found to be one of the major conversions.

3. EDA Approach

3.5. Data Visualization :



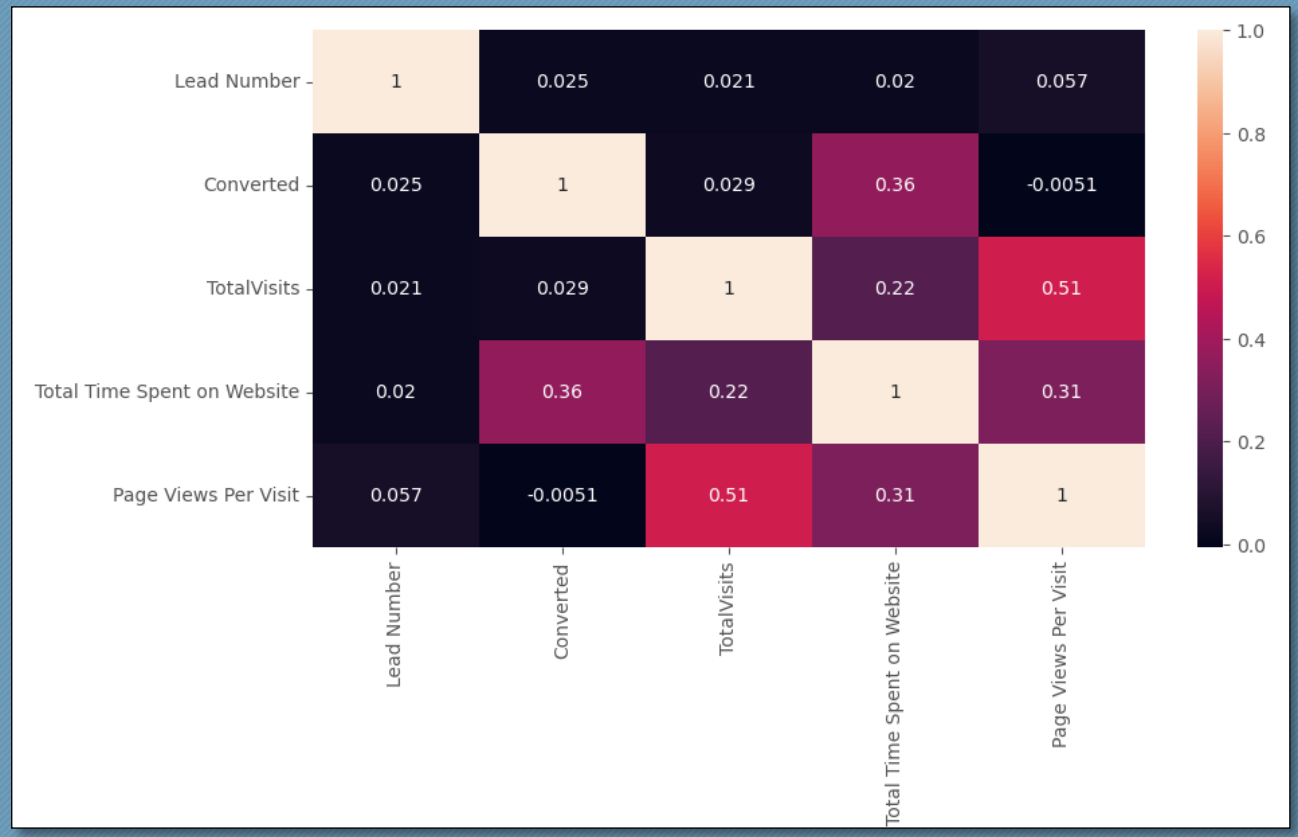
Distribution of continuous variables **TotalVisits** shows a narrow distribution with highest frequency of 3-7 times.



Distribution of continuous variables **Total Time Spent on Website** shows a distribution spread of 0 – 2000 minutes. Showing a mean 100 minutes.

3. EDA Approach

3.5. Data Visualization :



Hardly any strong correlation exists as per the matrix. Only 1 relation TotalVisits & Page Views per visit has good score but it is obvious.

4. Model Building

4.1. Data Preparation:

- Conversion of some binary variables were performed from Yes/No to 1/0.
- The dummy variables are created for all the categorical columns. (One-hot-encoding).

4.2. Model Building:

- The Spit was done at 70% and 30% for train and test the data respectively with random state value as 100
- Feature scaling was done for Continuous variables using StandardScaler.
- By using GLM & RFE with provided 15 variables. 15 features were selected by the model.

4. Model Building

4.2. Model Building:

- A total of 6 iteration was done to reach the final model with all values being in limit for p-values & VIF. (The variables with VIF < 5 and p-value 0.05 were kept).

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6456
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1489.4
Date:	Sat, 26 Aug 2023	Deviance:	2978.8
Time:	18:49:41	Pearson chi2:	1.60e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.5806
Covariance Type:	nonrobust		

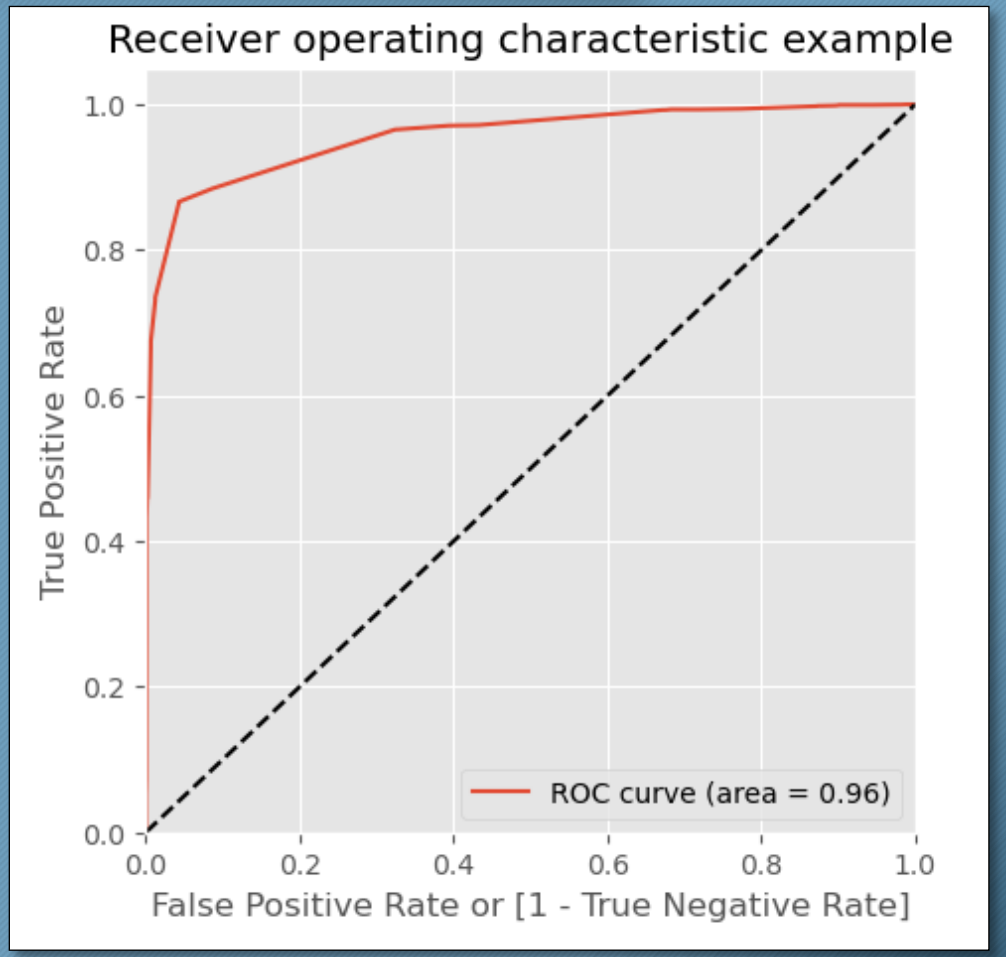
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4568	0.070	-20.715	0.000	-1.595	-1.319
Lead Source_Welingak Website	4.8225	0.751	6.421	0.000	3.350	6.295
Last Activity_SMS Sent	2.1701	0.104	20.948	0.000	1.967	2.373
Lead Quality_Worst	-2.9125	0.549	-5.302	0.000	-3.989	-1.836
Last Notable Activity_Modified	-1.7340	0.115	-15.120	0.000	-1.959	-1.509
Tags_Closed by Horizzon	8.1425	1.008	8.074	0.000	6.166	10.119
Tags_Interested in other courses	-1.2573	0.329	-3.823	0.000	-1.902	-0.613
Tags_Lost to EINS	7.1710	0.792	9.060	0.000	5.620	8.722
Tags_Ringing	-3.2385	0.220	-14.724	0.000	-3.670	-2.807
Tags_Will revert after reading the email	4.9133	0.174	28.224	0.000	4.572	5.254
Tags_invalid number	-3.4944	1.028	-3.399	0.001	-5.510	-1.479
Tags_switched off	-3.6826	0.517	-7.118	0.000	-4.697	-2.668

	Features	VIF
4	Tags_Closed by Horizzon	1.06
0	Lead Source_Welingak Website	1.03
6	Tags_Lost to EINS	1.03
10	Tags_switched off	1.03
9	Tags_invalid number	1.01
2	Lead Quality_Worst	0.41
5	Tags_Interested in other courses	0.34
8	Tags_Will revert after reading the email	0.14
1	Last Activity_SMS Sent	0.11
7	Tags_Ringing	0.10
3	Last Notable Activity_Modified	0.04

4. Model Building

4.3. Model Evaluation & Prediction:

- 1st Train dataset split prediction was performed for converted variable using default cut-off value 0.5.
- Later on, ROC curve (area = 0.96) to check the relation between sensitivity & specificity.



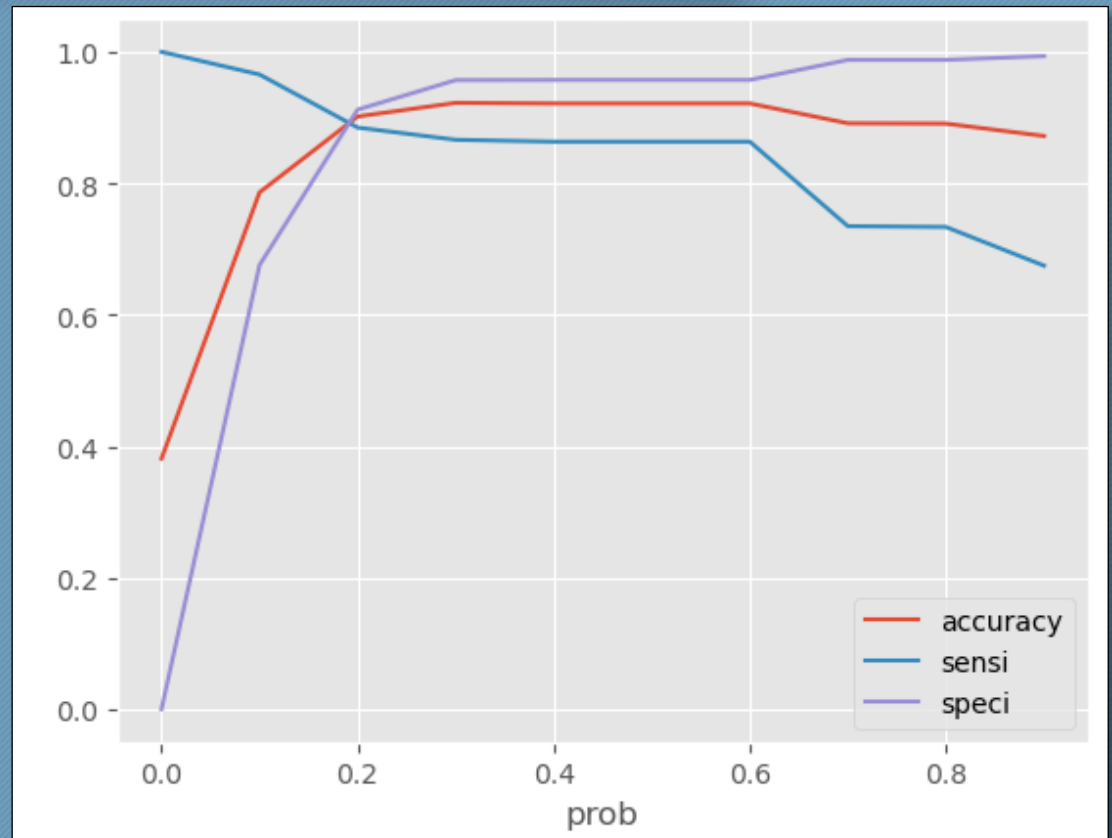
4. Model Building

4.3. Model Evaluation & Prediction:

- The optimal cut off is at 0.195 & was used to calculate metrics again.
- Afterwards with this value prediction was done the test dataset split with previous selected features & cut-off value.

4.4. Final Observation:

Metrics	Train Data	Test Data
Accuracy (%)	90.16	90.25
Sensitivity (%)	88.48	89.40
Specificity (%)	91.20	90.81



From the graph it is visible that the optimal cut off is at 0.195

4. Conclusion

- As per the final model only 12 features were selected for the model & prediction.
- Top variables were **Tags**, **Last Activity** & **Lead Source**.
- The dummy variables which contributes most to the prediction are **Tags_Closed by Horizzon, Tags_Lost to EINS, Tags_Will revert after reading the email, Lead Source_Welingak Website** & **Last Activity_SMS Sent.**
- Also the accuracy, sensitivity & specificity for the both Test & Train data split came very close which ultimately gives confidence in the generated lead scores.

I urge the stakeholders to keep these points in mind before taking any major decision going forward.

Thank You