LEAD SCORING CASE STUDY

BY

MONALISA BANERJEE

SWETHA MENNDE

AMIT KUMAR

Student of IIIT - Bangalore and UpGrad
Batch : DS-C70

INTRODUCTION

This assignment aims to build a Logistic Regression model 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, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

BUSINESS UNDERSTANDING

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%. Although X Education gets a lot of leads, 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.

BUSINESS UNDERSTANDING - 2

- A typical lead conversion process can be represented using the funnel you can see at the right-hand side.

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

BUSINESS OBJECTIVES

- X Education has appointed us 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

DATA UNDERSTANDING

There are 3 files given for this EDA as explained below:

1. 'Leads.csv' contains all the information of leads dataset from the past with around 9000 data points.
   The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

2. Leads Data Dictionary.xlsx' is data dictionary which describes the meaning of the variables.

3. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

INITIAL IMPUTATIONS

- The total shape of Leads.csv is (9240, 37)
- Many of the categorical variables have a level called 'Select' which we updated as null value as its as good as a NULL Value.
- We have identified below columns with more than 40% missing values which is relatively high w.r.t a column. So we have dropped them

```
['How did you hear about X Education', 'Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique
Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score']
```

- We have identified below columns with more than 30% missing values and we decided to treat them

  ['Specialization', 'Tags', 'City']

- Also dropped columns exists with only one unique value

  ['Magazine' , 'Receive More Updates About Our Courses' ,'Update me on Supply Chain Content' , 'Get updates on DM Content', 'I agree to pay the amount through cheque']

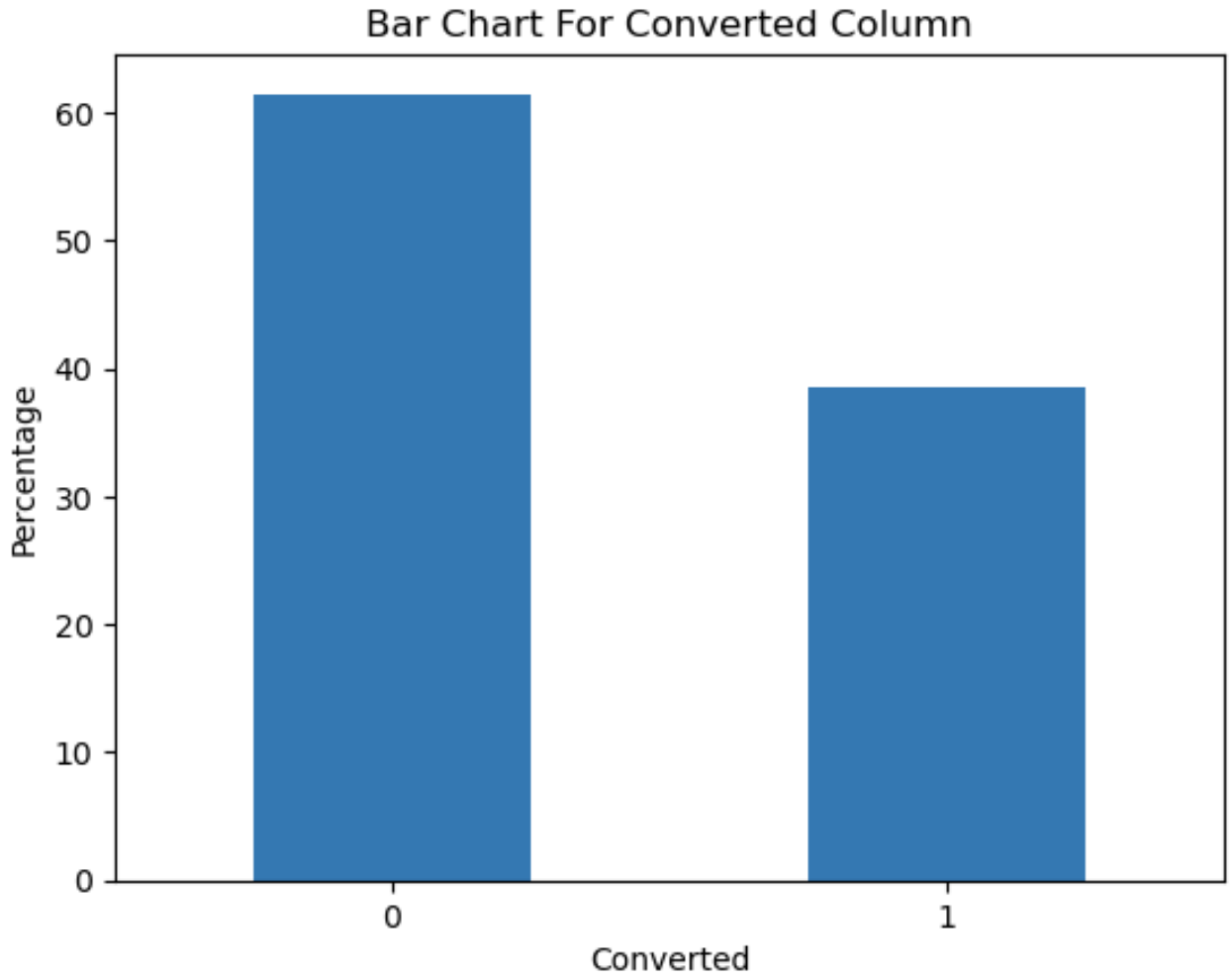
- Replacing null values with "UNKNOWN" as this column is important for analysis and we cant drop it.
- Analyzed 'Prospect ID' , 'Lead Number', and found having unique values for each row. So, we can drop these columns to proceed with our analysis

UNIVARIATE ANALYSIS FOR 'CONVERTED' COLUMN

- Analysis for 'Converted' column

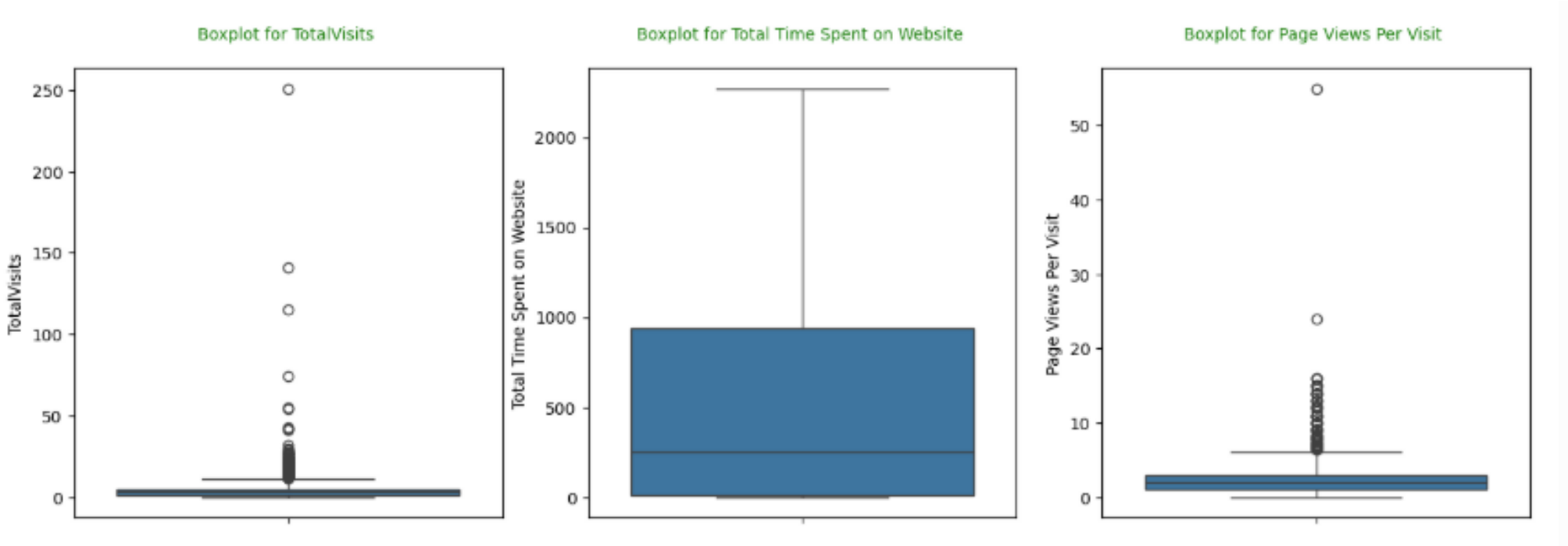
Meaning of the Converted Values :-

1 → Converted
0 → Non-Converted



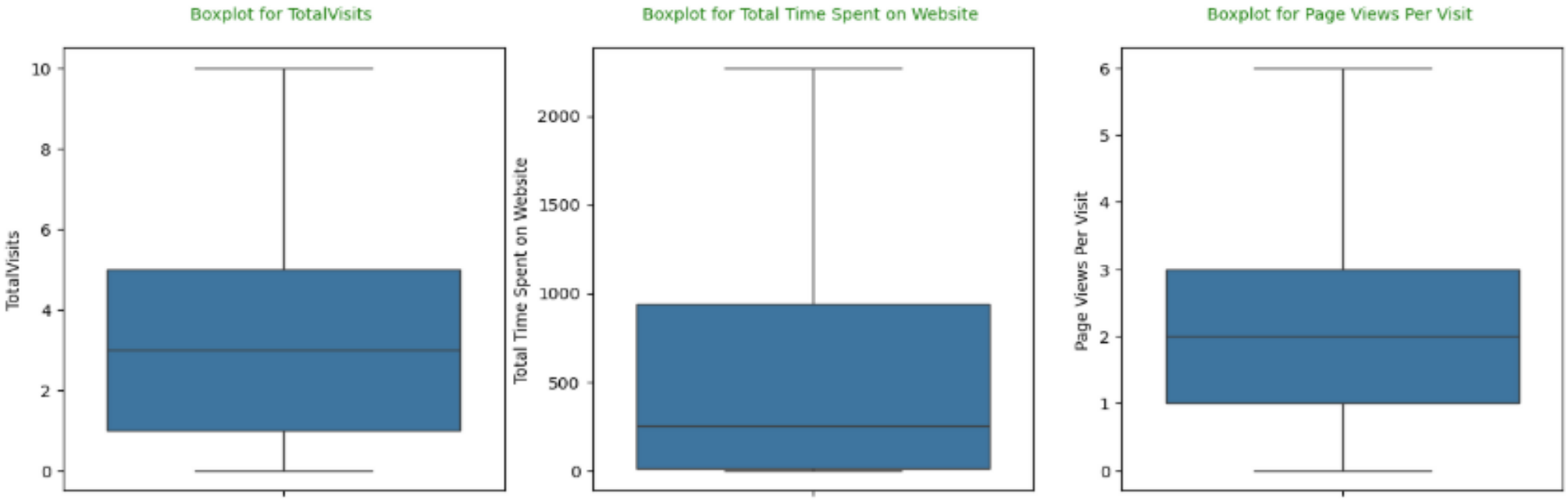
OUTLIER ANALYSIS : BEFORE OUTLIER TREATMENT

- In the Boxplot distribution we can clearly see that 'TotalVisits' and 'Page Views Per Visit' has Outliers, and we need to treat it before we progress with Model Building



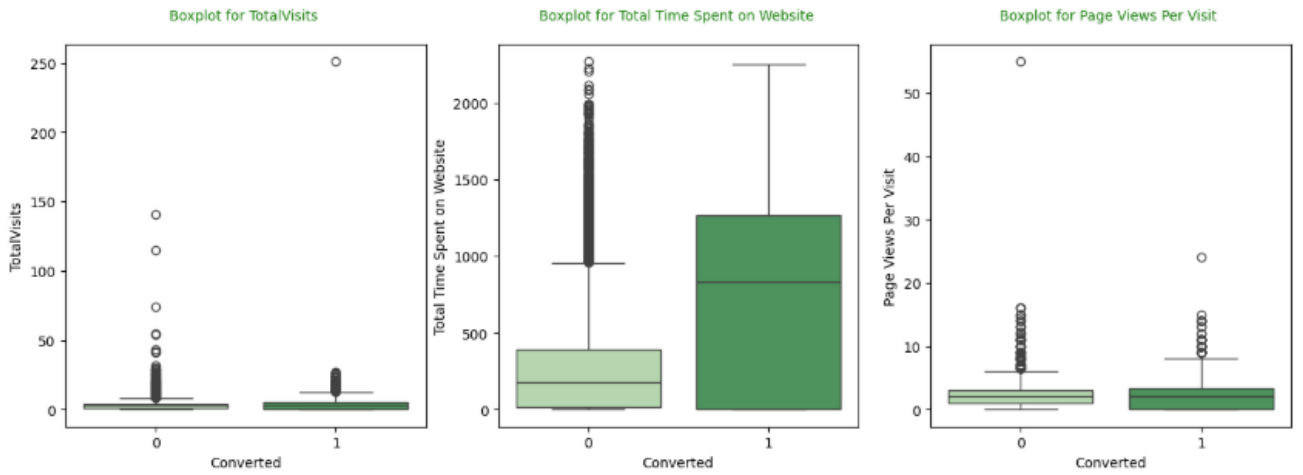
OUTLIER ANALYSIS :AFTER OUTLIER TREATMENT

- After Outlier Treatment we can now see the data is in a good shape.

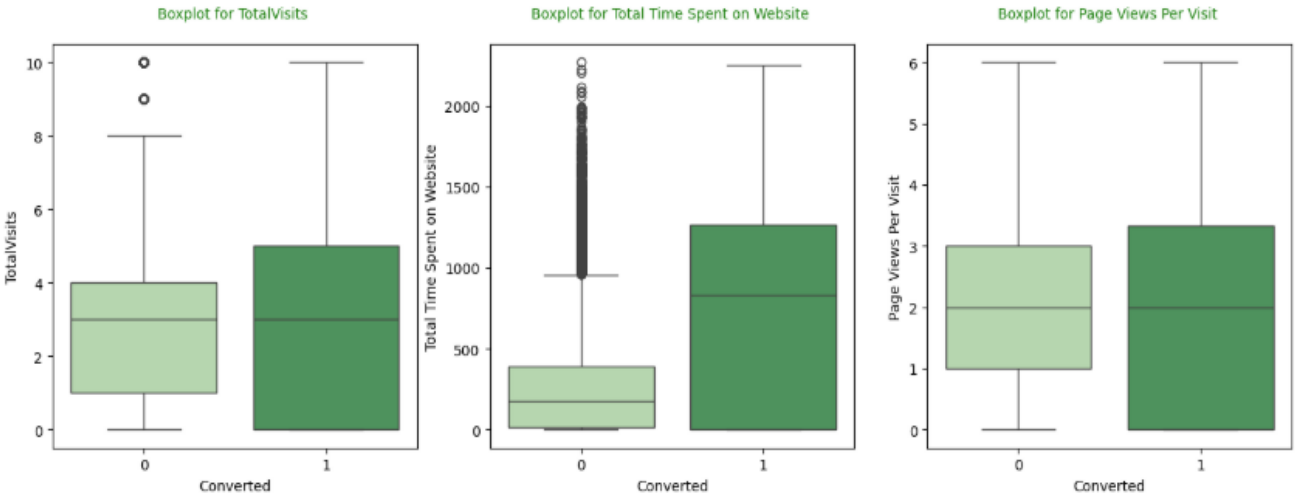


BIVARIATE ANALYSIS : BEFORE AND AFTER OUTLIER TREATMENT

- Before Outlier Treatment we can see the data distribution and can easily recognize these features has outliers

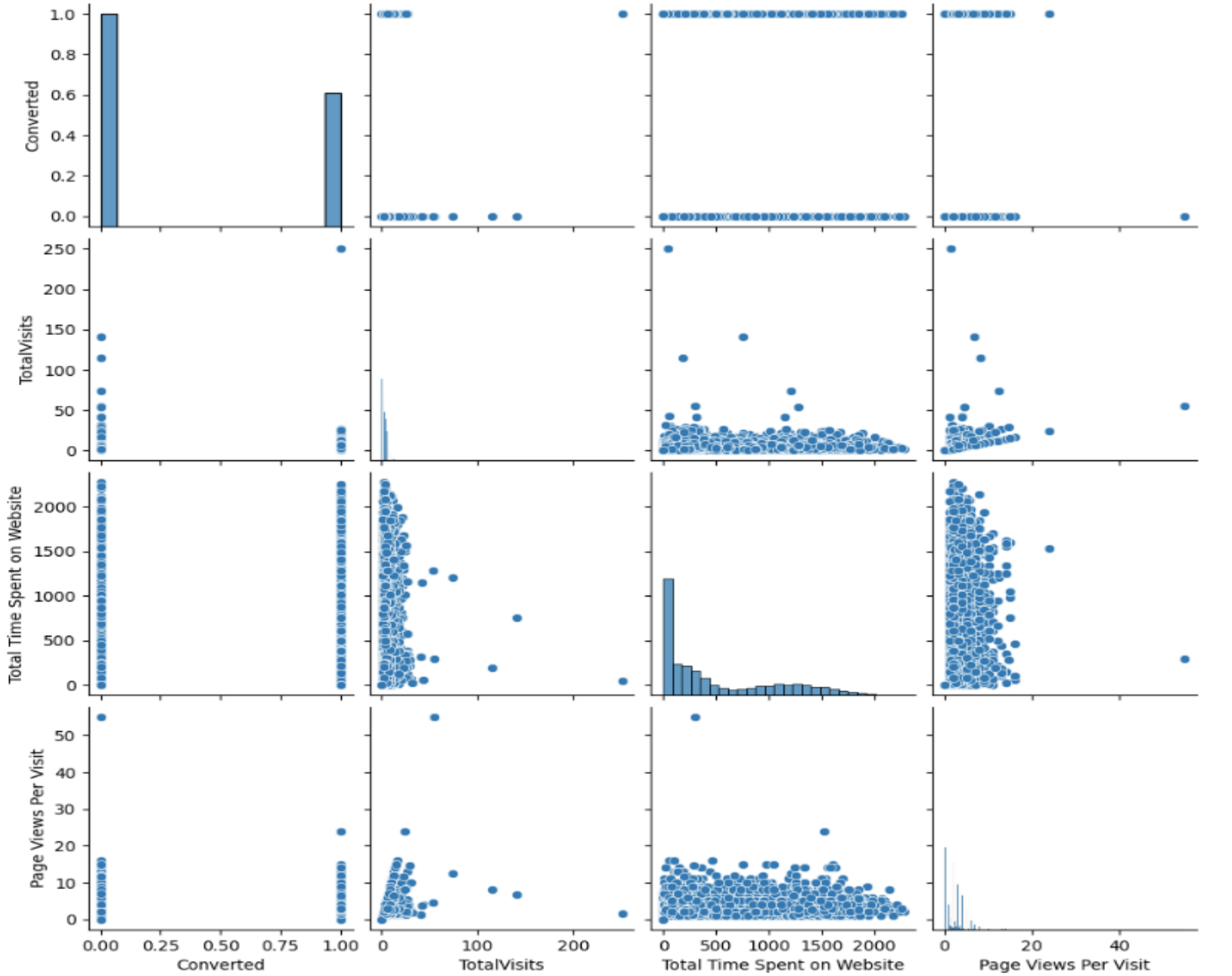


- After Outlier Treatment we can now see the data is in a good shape.

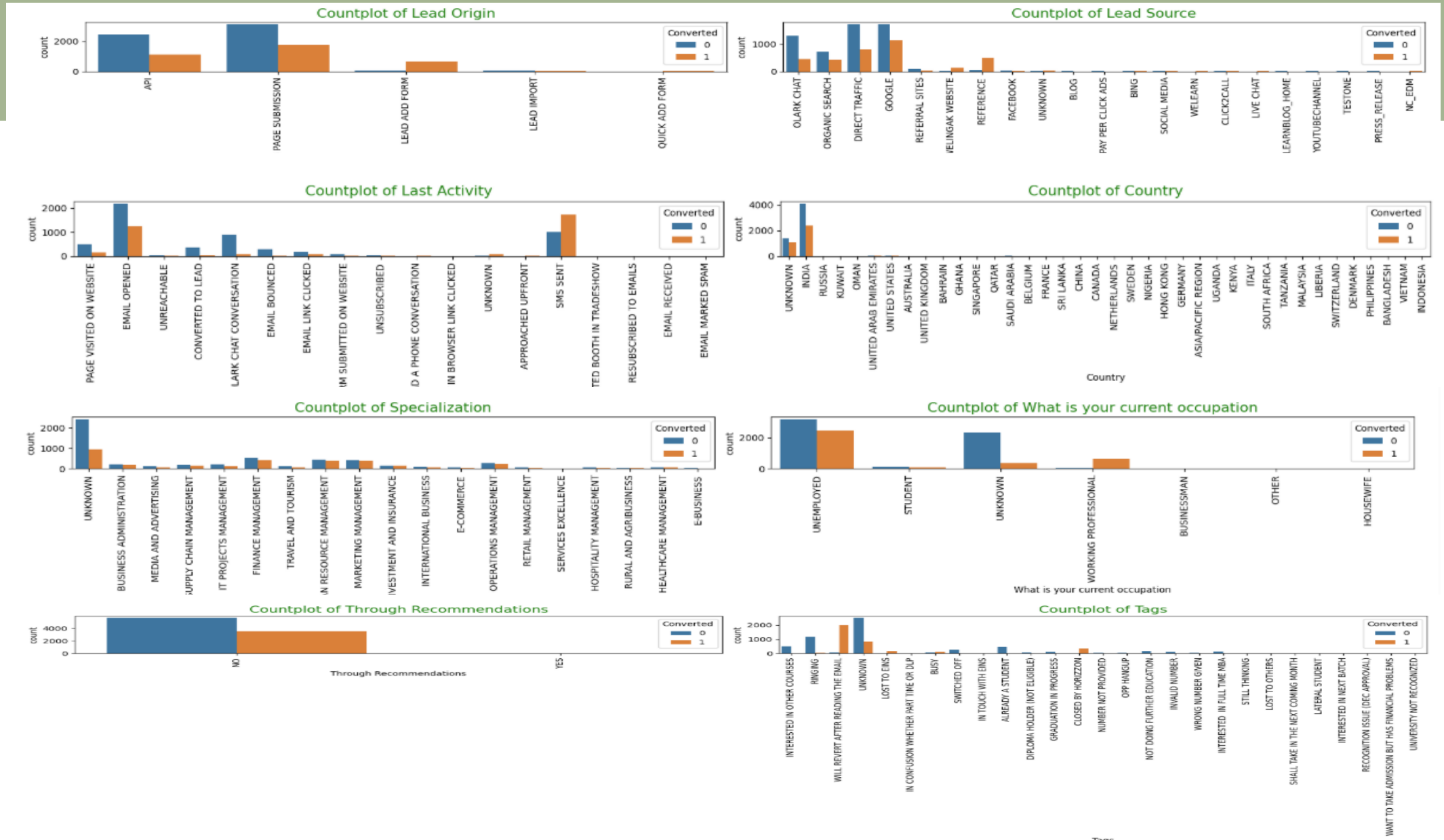


PAIRPLOT TO VISUALIZE NUMERICAL VARIABLE

- We have plotted 3 numerical variables along with our Target variable 'Converted'



BIVARIATE ANALYSIS FOR ALL CATEGORICAL VARIABLES



DATA PREPARATION

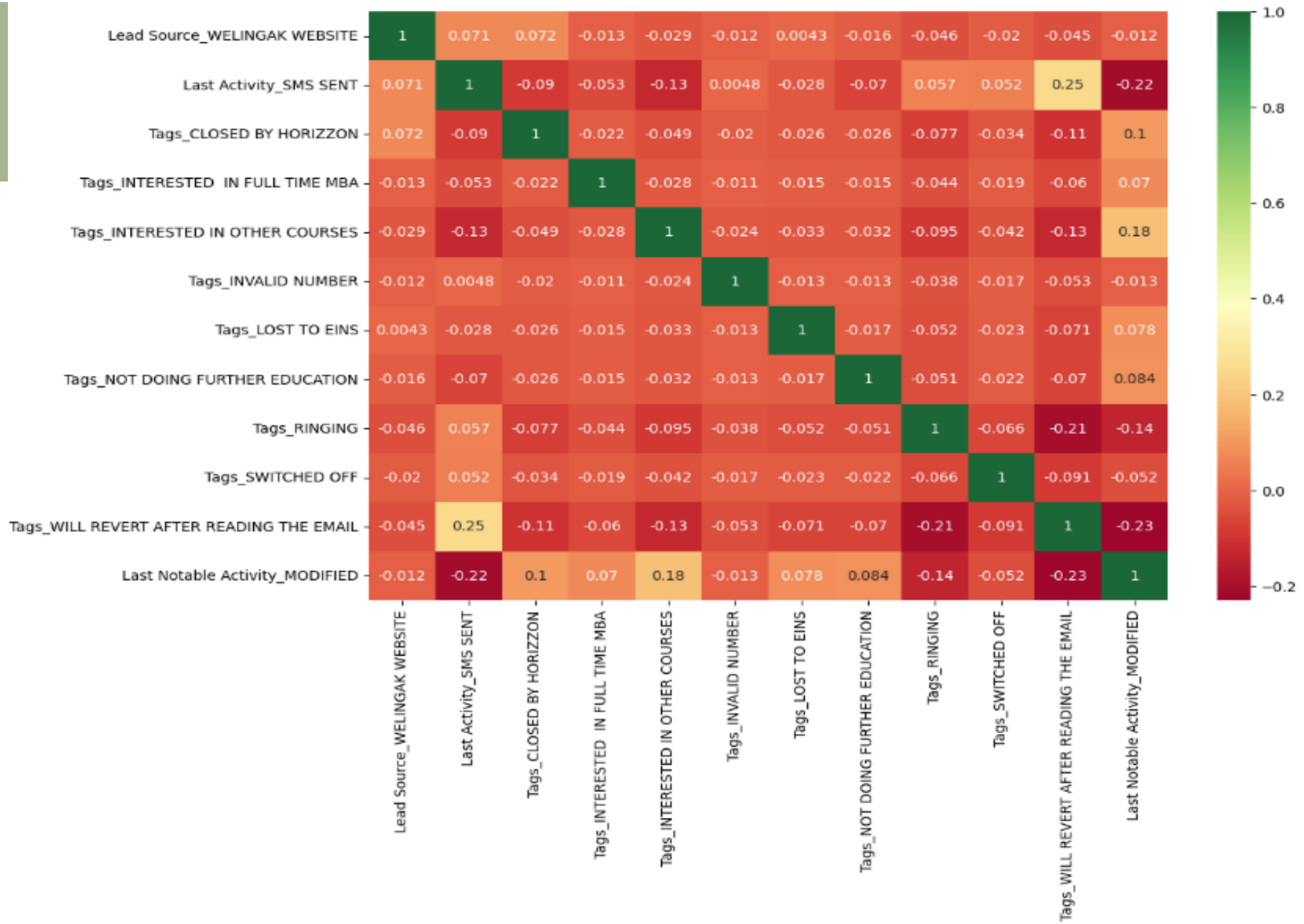
- Numerical Variables are Normalised by Scaling
- Dummy Variables are created for object type variables by get_dummies
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 155

MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- The first basic step for Logistic Regression is performing a train-test split, we have chosen 70:30 ratio. 70% data as Train data and 30% data as Test data
- Use RFE (Recursive Feature Elimination) for Feature Selection and selecting 15 features out of 155 features
- Running RFE with 15 variables as output
- Building Models one by one, by removing the features one by one whose p-value is greater than 0.05 and VIF is greater than 5
- Final Model is ready with p-value < 0.05 and VIF < 5.
- Check for Accuracy, Sensitivity, Specificity on Train Dataset
- Draw ROC Curve and find the optimal threshold
- Check for Precision and Recall and also find the threshold from there.
- Predictions on test data set based on the probability
- Overall accuracy is 92.13%

CORRELATION MATRIX

- This is the correlation matrix with the 12 final features which we finally obtained after building 5 models.
- After rejecting initial + 4 models we have finalized the 5th Model as our Final model



GLIMPSE OF FINAL MODEL

- As we can see the p-values in this model are all less than 0.05
- Final 12 Features has been identified

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1511.9
Date:	Tue, 21 Jan 2025	Deviance:	3023.8
Time:	03:09:01	Pearson chi2:	1.60e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.5776
Covariance Type:	nonrobust		

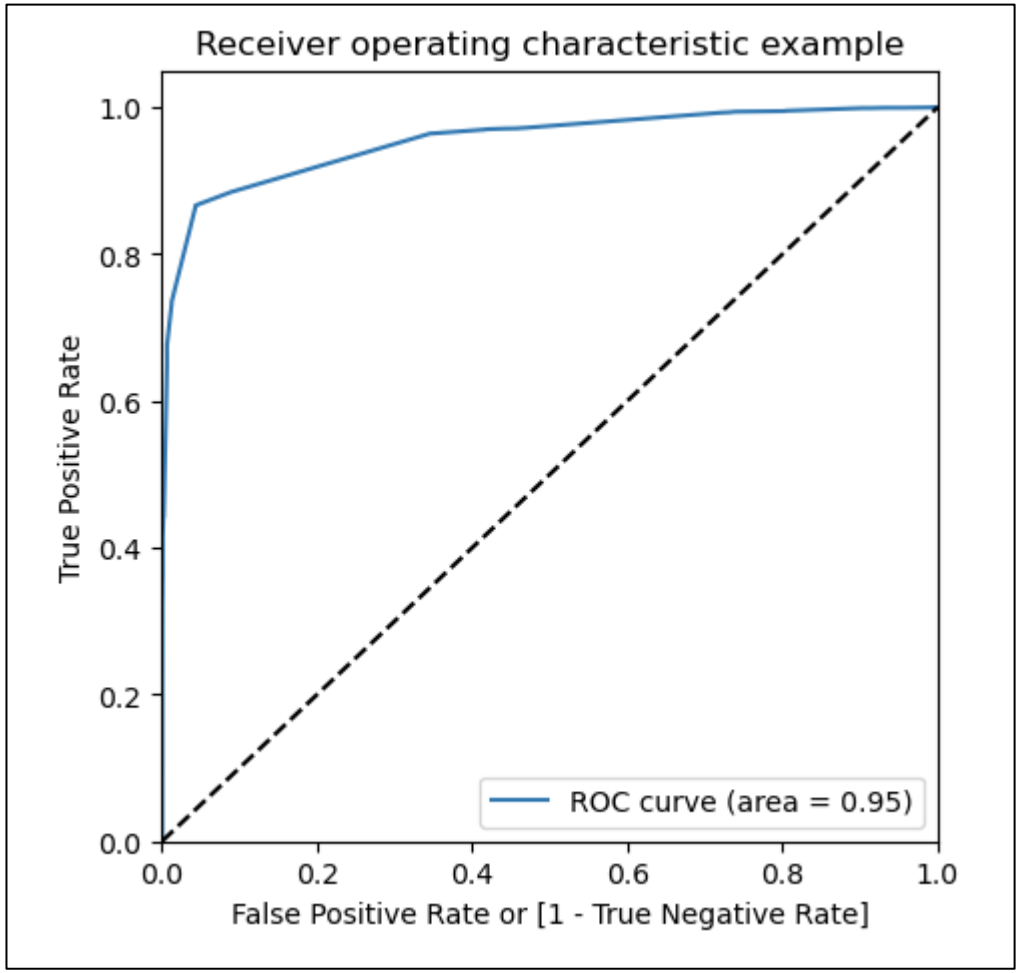
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5349	0.070	-21.997	0.000	-1.672	-1.398
Lead Source_WELINGAK WEBSITE	4.8911	0.752	6.505	0.000	3.417	6.365
Last Activity_SMS SENT	2.2383	0.103	21.638	0.000	2.036	2.441
Tags_CLOSED BY HORIZZON	8.2070	1.009	8.138	0.000	6.230	10.184
Tags_INTERESTED IN FULL TIME MBA	-1.5751	0.725	-2.172	0.030	-2.996	-0.154
Tags_INTERESTED IN OTHER COURSES	-1.3587	0.326	-4.169	0.000	-1.997	-0.720
Tags_INVALID NUMBER	-3.7576	1.023	-3.673	0.000	-5.763	-1.752
Tags_LOST TO EINS	6.6272	0.725	9.142	0.000	5.206	8.048
Tags_NOT DOING FURTHER EDUCATION	-2.6810	1.022	-2.624	0.009	-4.683	-0.679
Tags_RINGING	-3.2177	0.220	-14.610	0.000	-3.649	-2.786
Tags_SWITCHED OFF	-3.6895	0.517	-7.131	0.000	-4.704	-2.675
Tags_WILL REVERT AFTER READING THE EMAIL	4.9710	0.174	28.640	0.000	4.631	5.311
Last Notable Activity_MODIFIED	-1.7358	0.115	-15.139	0.000	-1.960	-1.511

ROC(RECEIVER OPERATING CHARACTERISTIC) CURVE

An ROC curve demonstrates several things:

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

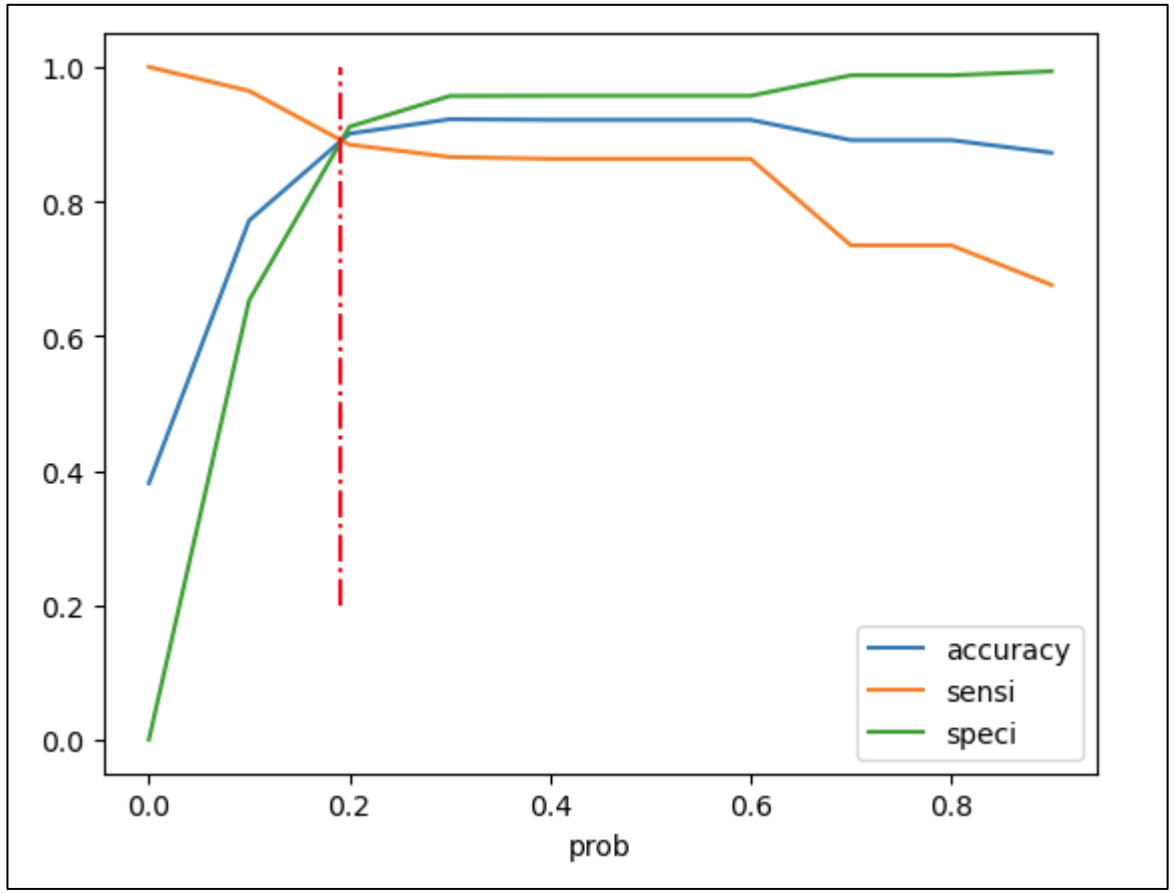
In our case we have received the AUC (Area Under the Curve) as 95%



PLOT FOR ACCURACY, SENSITIVITY AND SPECIFICITY FOR VARIOUS PROBABILITIES

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

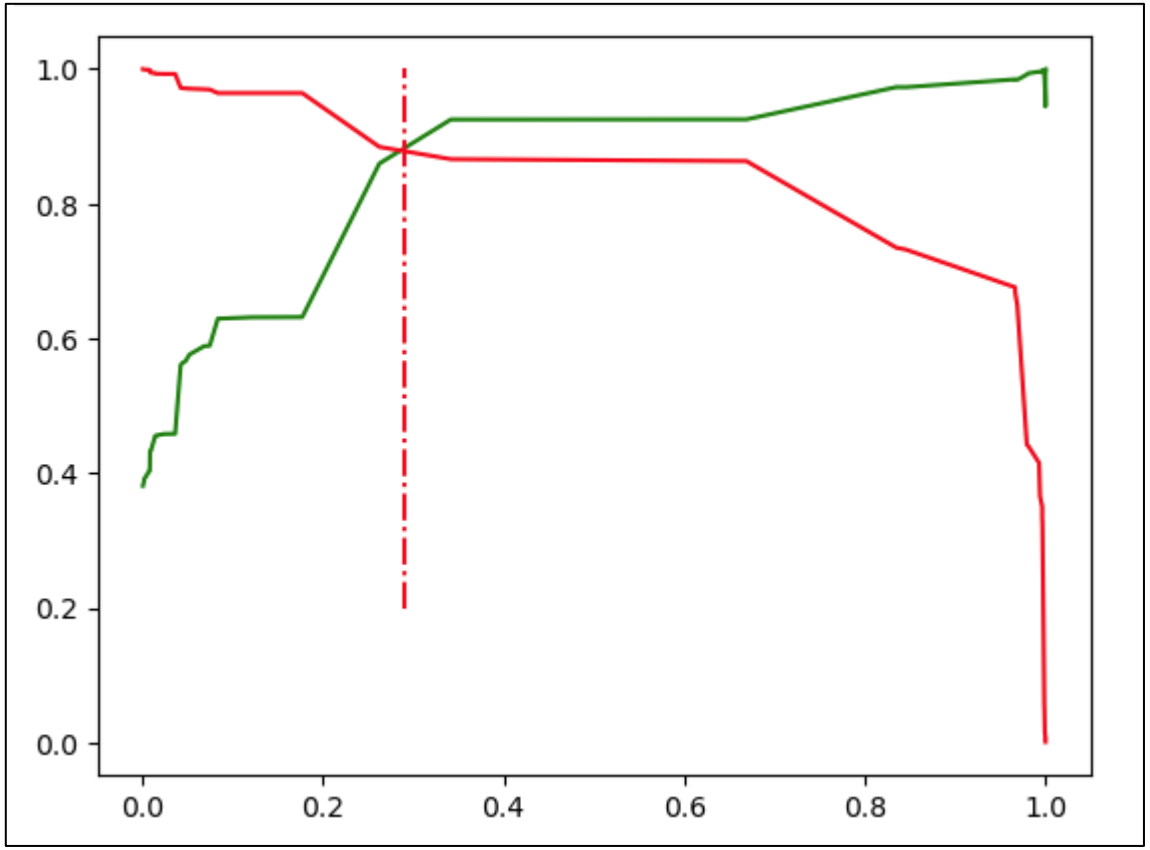
- From this plot we got 0.19 as the optimum point to take it as a cutoff probability. This probability will be used further in Predicting the Test dataset



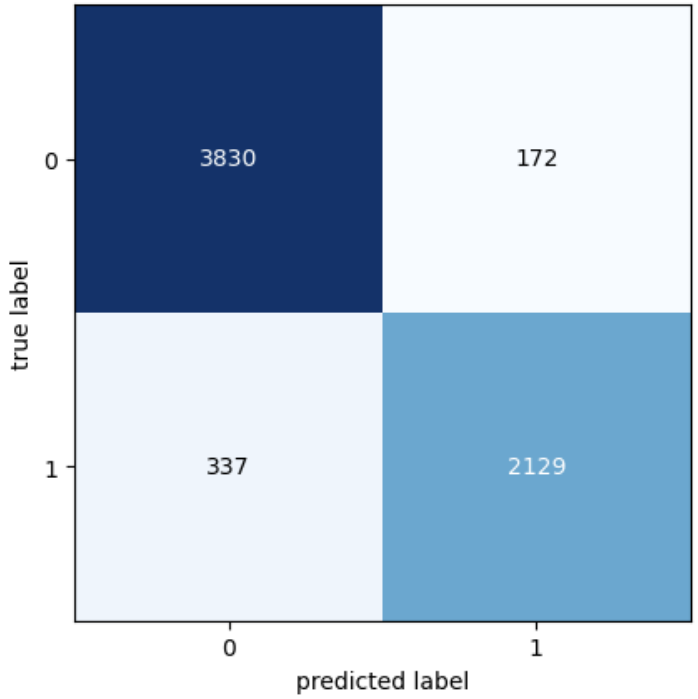
PRECISION AND RECALL TRADEOFF PLOT

- From Precision-Recall Plotting also we can get the Optimal cutoff probability. And from this also we can get balanced sensitivity and specificity

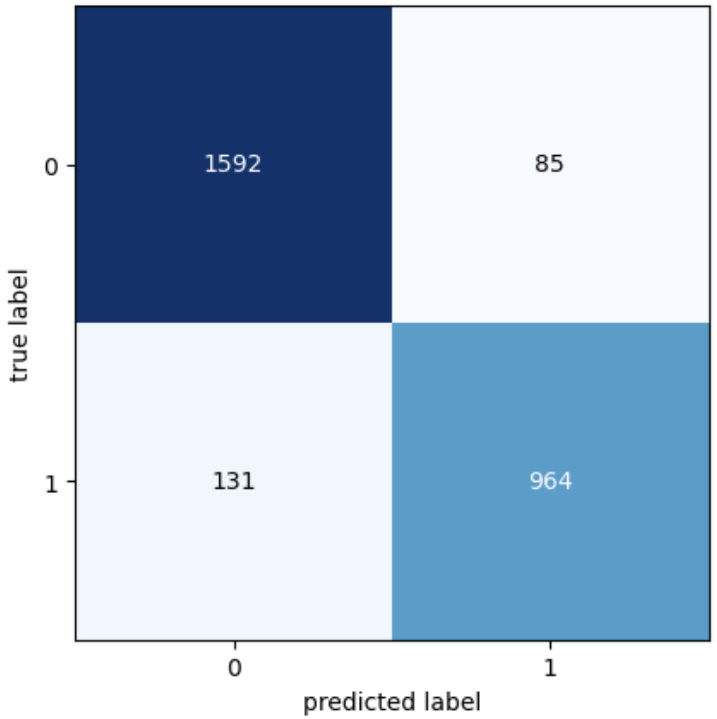
- From this plot we got 0.29 as the optimum point to take it as a cutoff probability. This probability will be used further in Predicting the Test dataset



CONFUSION MATRIX



Train Set



Test Set

FINAL RESULTS

Evaluation Rubrics	Train Set (in %)	Test Set in (in %)
Accuracy	92.22	92.20
Sensitivity	86.61	88.03
Specificity	95.67	94.93
False Positive Rate	4.32	5.06
Positive Predictive Value	92.5	91.89
Negative Predictive Value	92.06	92.39
AUC(Area Under Curve)	95.0	95.81

RELATIVE FEATURE IMPORTANCE GRAPH

INFERENCES : FEATURE IMPORTANCE

- From the Relative Feature Importance Graph 3 variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:
  i. Tags
  ii. Lead Source
  iii. Last Activity

- From the Relative Feature Importance Graph we have got 3 best categorical/dummy variables with Positive coefficient as below,
  i. Tags_CLOSED BY HORIZZON
  ii. Tags_LOST TO EINS
  iii. Tags_WILL REVERT AFTER READING THE EMAIL

- All three contribute positively towards the probability of a lead conversion.
- These results indicate that the company should focus more on the leads with these 3 Tags

Situation and Recommendations :-

Situation-1 : X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Recommendations :

Sensitivity tells us how well our model avoids missing actual conversions. A high sensitivity means the model is good at identifying potential customers.

- Sensitivity = TP / (TP + FN)

1. High sensitivity implies that our model will correctly predict almost all leads who are likely to convert. At the same time, it may overestimate and misclassify some of the non-conversions as conversions
2. As the company has extra man-power for two months and wants to make the lead conversion more aggressive, it is a good strategy to go for high sensitivity. To achieve high sensitivity, we need to choose a low threshold value.

Hence, it is a good strategy to go with high Sensitivity.

Situation and Recommendations :-

Situation- 2 : Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Recommendations :

In the last strategy we used sensitivity to increase the conversion rate. Now as the target has been achieved, and the man-power is also going to less during this time so, the concept of specificity will be applicable here.

- **Specificity** = TN / (TN + FP)

1. In the given situation, we'll use a high specificity because high specificity means that the model will correctly predict almost all leads which are not likely to convert.
2. At the same time, it may misclassify some of the conversions as non-conversions. But that's okay as the company has already reached its target for a quarter and doesn't want to make phone calls unless it's extremely urgent.

Hence, it is a good strategy to go with high Specificity.

CONCLUSION

This model will help to identify the hot leads which would enhance speed-to-lead and the response rate.

- Approaching only to hot lead would result in:

  - Shorter sales cycle through intuitive prioritization.
  - Better opportunity-to-deal ratio
  - Increase marketing effectiveness
  - Better sales forecasting
  - Minimize opportunities loss
  - Increase in revenue