Lead Scoring Case Study

Presenter : Saroj , Veera & Subha
Group : DSC 45
Date : 3rd Jan 2023

Agenda

01 Business Objectives

The main Business Objectives behind the case study and the requirements of the Business. Problem Statements.

02 EDA Steps & Analysis

The Various steps used, and Analysis done during the EDA. Handling Null Values, Finding Outliers, Finding the Imbalance in Data etc.

03 Model building and evaluation

Building a logistic Regression model and calculating Lead Score. Evaluating the model by using different metrics -Specificity and Sensitivity or Precision and Recall.

04 Insights & Recommendation

The Outcome of the Analysis using the model & the recommendation towards the Business.

Business Objectives

Introduction & Business Requirements

Outline

Introduction

This assignment aims to give you an idea of applying the logistic regression model in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the module, you will also develop a basic understanding of the Lead Scoring technique. Here An education company sells online courses to industry professionals. On any given day, the CEO has given the ballpark of the target lead conversion rate to be around 80%.

Business Understanding

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people 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 email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

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%.

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.

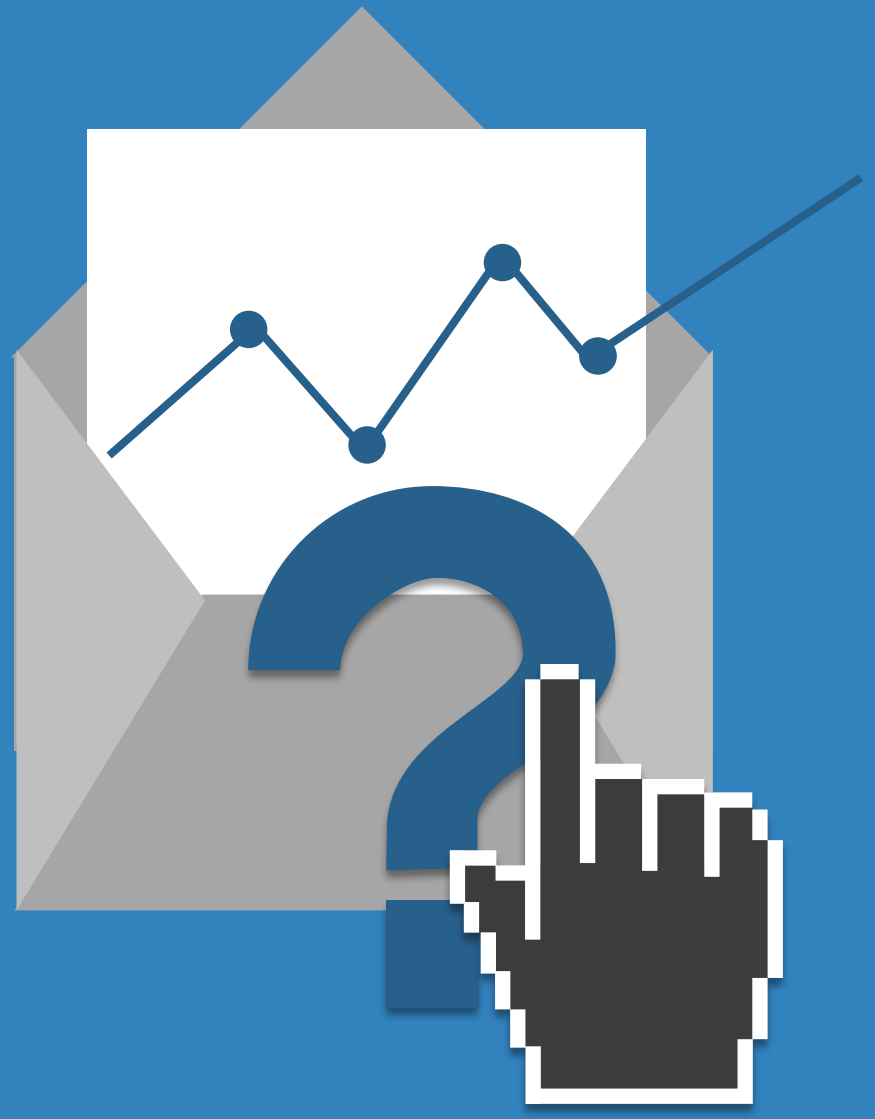
Business Objectives

This case study 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..

Strategy

The Steps and Process Followed to Reach Solution

1. Sourcing the data for analysis
2. Clean and prepare the data
3. Exploratory Data Analysis.
4. Feature Scaling
5. Splitting the data into Test and Train dataset.
6. Building a logistic Regression model and calculating Lead Score.
7. Evaluating the model by using different metrics -Specificity and Sensitivity or Precision and Recall.
8. Applying the best model in Test data based on the Sensitivity and Specificity Metrics.



Problem Solving Methodology

Feature Engineering

Feature Scaling of Numeric data
Splitting data into train and test set.

Feature Scaling and Splitting Train and Test Sets

Model Building

Model Building

Feature Selection using RFE
Determine the optimal model using Logistic Regression
Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Problem Solving Methodology

This allowed us to reach our desired Observation in a processed way.

Data Sourcing, Cleaning and Preparation

Result

Data Preparation

Read the Data from the Source
Convert data into a clean format suitable for analysis
Remove duplicate data
Outlier Treatment
Exploratory Data Analysis
Feature Standardization

Observations

Determine the lead score and check if the target final predictions amount to an 80% conversion rate.

Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Our Team

1

SUBHA SHINI

2

SAROJ KUMAR PRAMANIK

3

VEERAKUMAR M

FLOW

Prepare Data for Modeling

Check the univariate, and bivariate analysis. Identify the missing data and use an appropriate method to deal with it. Impute the null values. Create Dummy Variables and encoding of the data & Feature Scaling

Making Predictions

Before predicting the test set, we need to standardize the test data and have the required features only. Then apply the final model for Prediction.

Data Cleaning, Validation & Prepare for Modeling

Check the Input file type and understand the data thoroughly. Import required python library. Identify if there are outliers in the dataset. Handle Outliers appropriately. Identify data imbalance

Model Building & Evaluation

Split the data, use RFE, and Remove those whose p-value > 0.5 & VIF > 5. Calculate the values for sensitivity, specificity, positive predictive values and negative predictive Values.

Conclusion

Conclude the Analyzed Model and prepare the recommendation to the Business.

Inputs & Data Sets Considered

This Section is to provide the Details description about the Inputs and Data Set received for the Used Cases..

01

Dataset 1

Leads Data

02

Dataset 2

Data Dictionary

03

Document

Business Questions

Reading & Inspecting Datasets

Leads Data (Dataset-1)

Shape of Data

The Shape of the Data set was observed to be **(9240, 37)**

```
# checking the shape of the data 'df'

df.shape
```

```
(9240, 37)
```

Data Types

float64 : 4
int64 : 3
object : 30

```
df.dtypes.value_counts()
```

```
object     30
float64     4
int64       3
dtype: int64
```

Note:
Leads.csv as **df** considered.

Columns present in df

As there are 37 Column present in the df, below represents

```
# Checking the column names

df.columns
```

```
Index(['Prospect ID', 'Lead Number', 'Lead Origin', 'Lead Source',
       'Do Not Email', 'Do Not Call', 'Converted', 'TotalVisits',
       'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity',
       'Country', 'Specialization', 'How did you hear about X Education',
       'What is your current occupation',
       'What matters most to you in choosing a course', 'Search', 'Magazine',
       'Newspaper Article', 'X Education Forums', 'Newspaper',
       'Digital Advertisement', 'Through Recommendations',
       'Receive More Updates About Our Courses', 'Tags', 'Lead Quality',
       'Update me on Supply Chain Content', 'Get updates on DM Content',
       'Lead Profile', 'City', 'Asymmetrique Activity Index',
       'Asymmetrique Profile Index', 'Asymmetrique Activity Score',
       'Asymmetrique Profile Score',
       'I agree to pay the amount through cheque',
       'A free copy of Mastering The Interview', 'Last Notable Activity'],
      dtype='object')
```

Describing the Data Set df

Below is the sample of the Describe of df as difficult to capture in 1 frame.

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Data Understanding :

- ✓ The Data frame is having 9270 rows and 37 columns.
- ✓ 30 columns have Object type, and the rest of the others are either float or integer.
- ✓ Looking into the data the dtype Object is the Date type.
- ✓ Looking into the data few fields seem to be categorical in nature.
- ✓ We can see that there are missing values present in our data.

EDA Steps & Analysis

Data Cleaning

It refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it.

Missing Values

The data has some missing values in its columns. There are three major categories of missing values are MCAR, MAR & MNAR

Handling Outliers

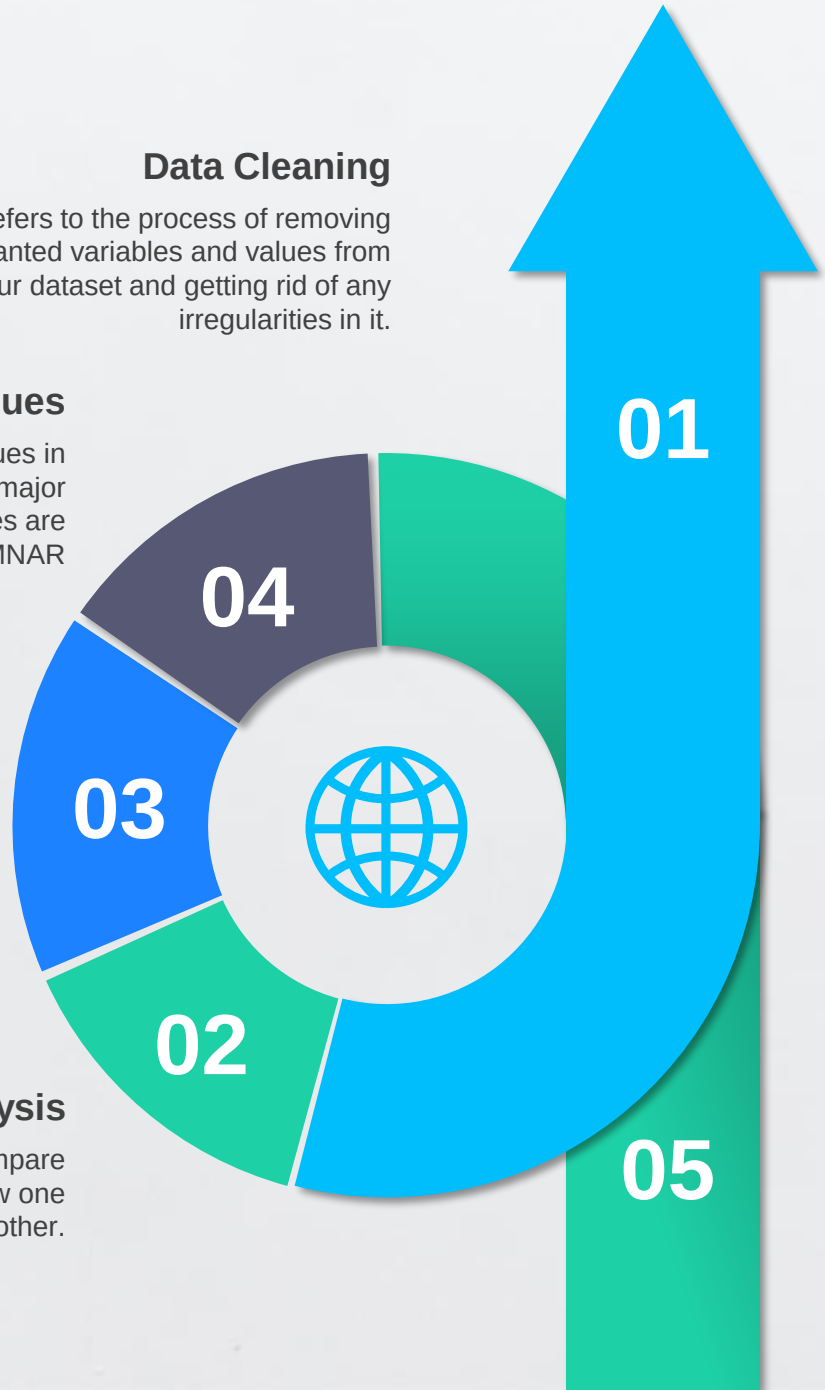
There are two types of outliers: Univariate outliers &Multivariate outliers

Univariate Analysis

In Univariate Analysis, we analyse data of just one variable. A variable in our dataset refers to a single feature/ column.

Bivariate Analysis

Here, you use two variables and compare them. This way, you can find how one feature affects the other.



Data Cleaning & Handling Missing Values

Data Cleaning

Checked Null values for all the columns and eliminated columns which had null values of more than **45%** in the leads.csv.

Finding the Categorical & Numerical Column

As we know Describe() only works with the numerical column, so Identifying the numerical column & taking out the list of the numerical column from the total column to find out the categorical column.

Handling Missing Values & Imputation

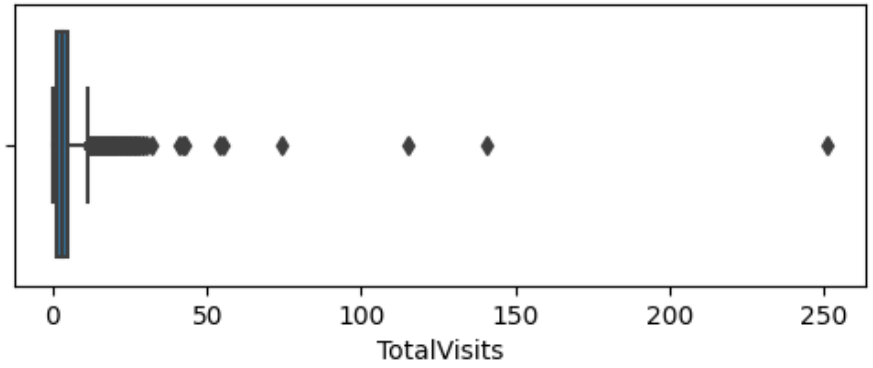
Identified outliers and imputed using the best approach available. There is a huge value of null variables in some columns as seen above. But removing the rows with the null value will cost us a lot of data and they are important columns. So, instead, we are going to replace the NaN values with 'not provided'. This way we have all the data and almost no null values. In case these come up in the model, it will be of no use, and we can drop it off then.

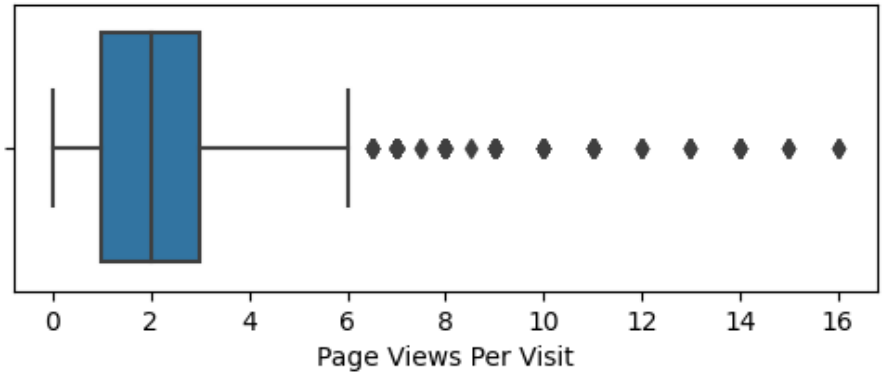
We observe that there are 'Select' values in many columns. It may be because the customer did not select any option from the list, hence it shows 'Select'. 'Select' values are as good as NULL. So, we can convert these values to null values.

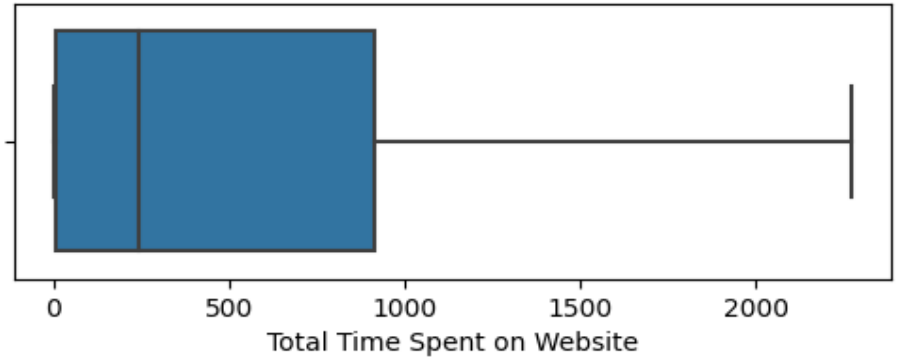
```
# Sorting % of null valaues of each column in decending order

x = (df.isnull().sum()/len(df.index)*100)
x.sort_values(ascending = False).head(30)
```

```
How did you hear about X Education              78.463203
Lead Profile                                    74.188312
Lead Quality                                    51.590909
Asymmetrique Activity Score                     45.649351
Asymmetrique Profile Score                      45.649351
Asymmetrique Profile Index                      45.649351
Asymmetrique Activity Index                     45.649351
City                                            39.707792
Specialization                                  36.580087
Tags                                            36.287879
What matters most to you in choosing a course   29.318182
What is your current occupation                 29.112554
Country                                         26.634199
TotalVisits                                      1.482684
Page Views Per Visit                             1.482684
Last Activity                                    1.114719
Lead Source                                      0.389610
Get updates on DM Content                        0.000000
Update me on Supply Chain Content                0.000000
I agree to pay the amount through cheque         0.000000
A free copy of Mastering The Interview           0.000000
Lead Origin                                      0.000000
X Education Forums                               0.000000
Receive More Updates About Our Courses           0.000000
Through Recommendations                          0.000000
Digital Advertisement                            0.000000
Newspaper                                        0.000000
Newspaper Article                                0.000000
Magazine                                         0.000000
Search                                           0.000000
dtype: float64
```

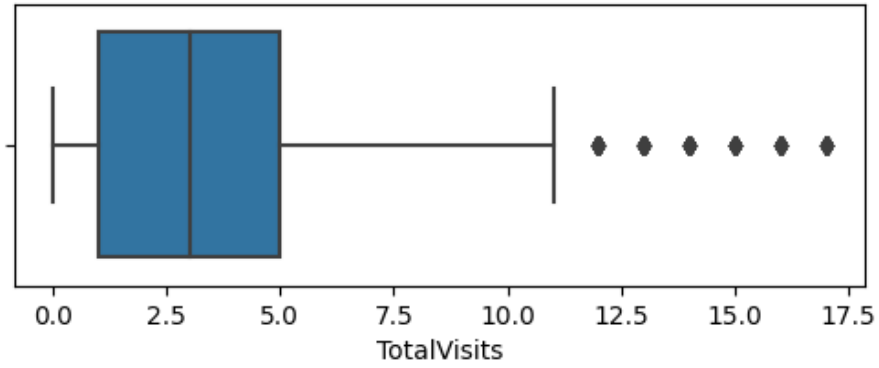
Outliers Analysis

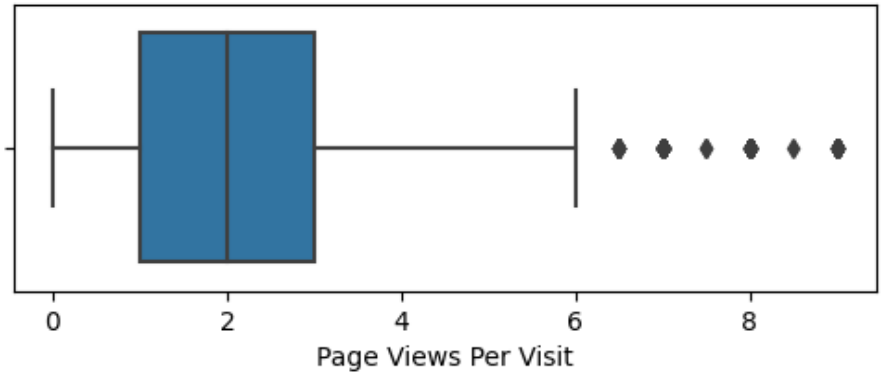






After Outlier Removal



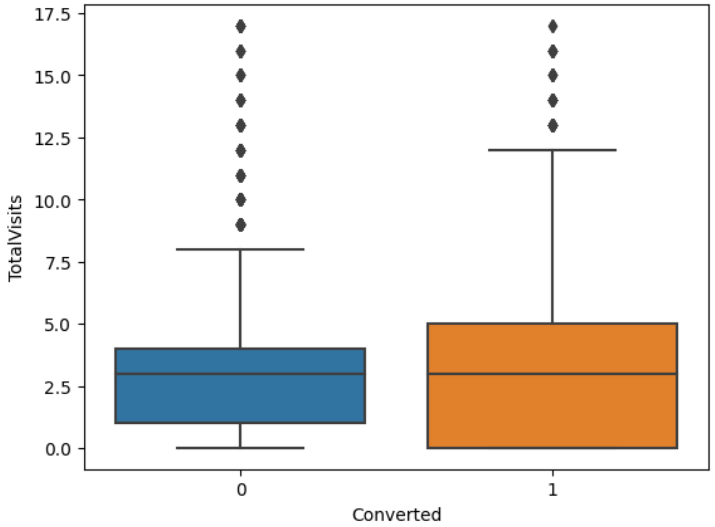


There are outliers observed in the “Total Visits” & “Page Views per Visit”.

Remove top & bottom 1% of the Column Outlier values from both the variable

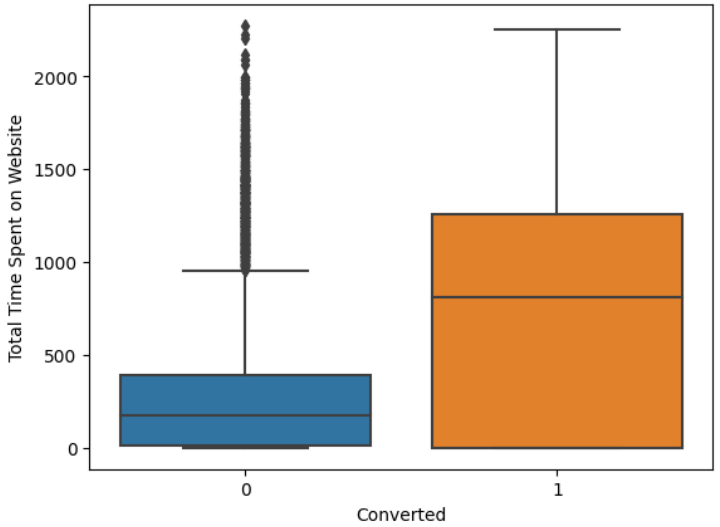
The total Time Spent on the Website was observed to be having no Outliers.

Outliers Analysis – Checking with Converted variable



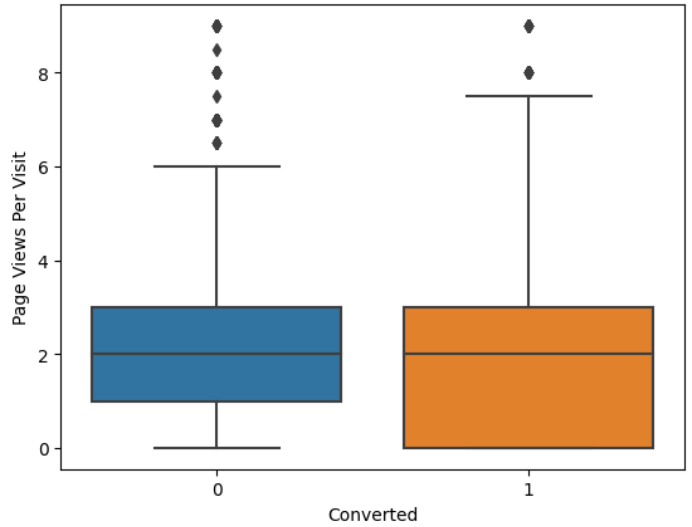
The Median for converted and not converted leads are close.

Nothing conclusive can be said on the basis of Total Visits



Leads spending more time on the website are more likely to be converted.

The website should be made more engaging to make leads spend more time

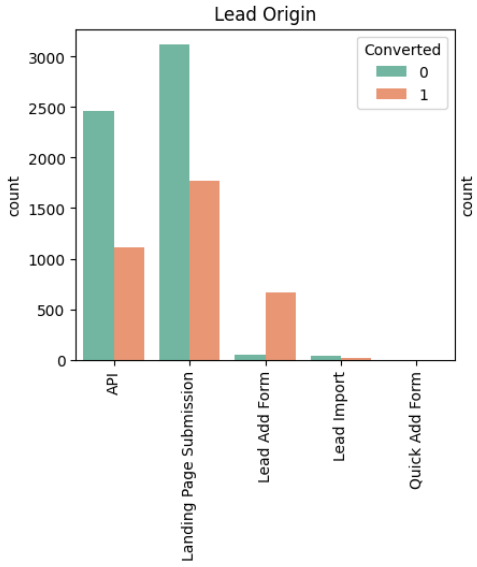


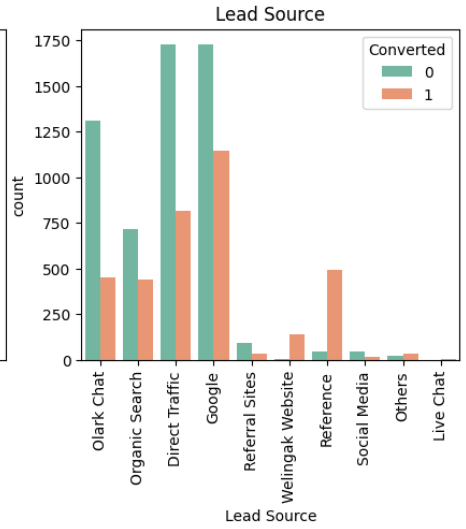
The median for converted and unconverted leads is the same.

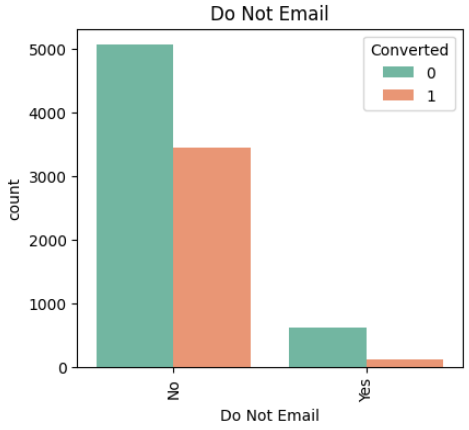
Nothing can be said specifically for lead conversion from Page Views Per Visit

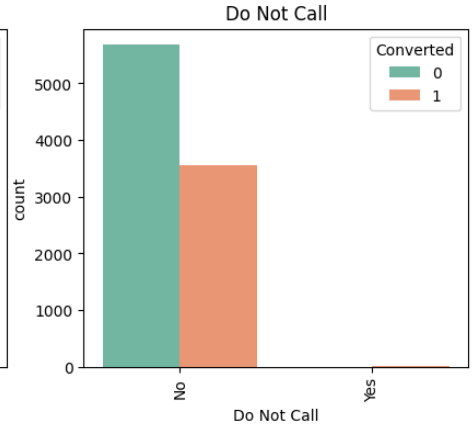
Univariate Analysis:

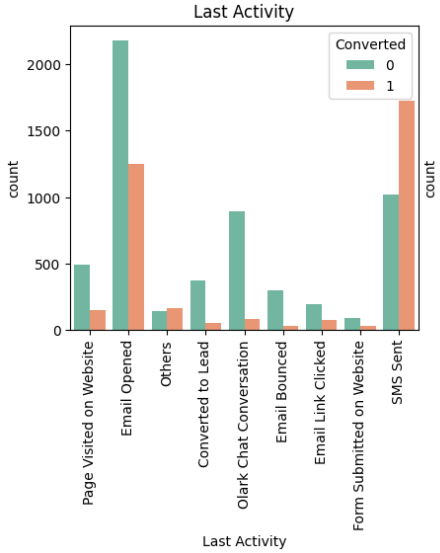
Analyzing all the categorical variables with the "Converted" Y Variable.

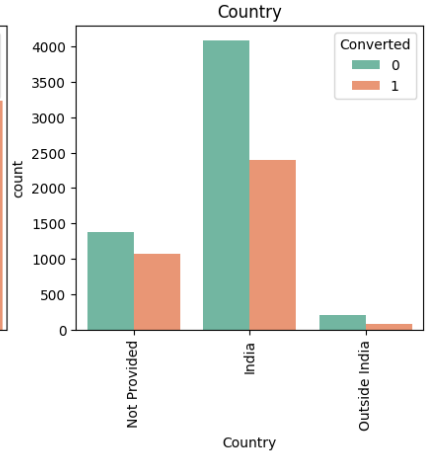


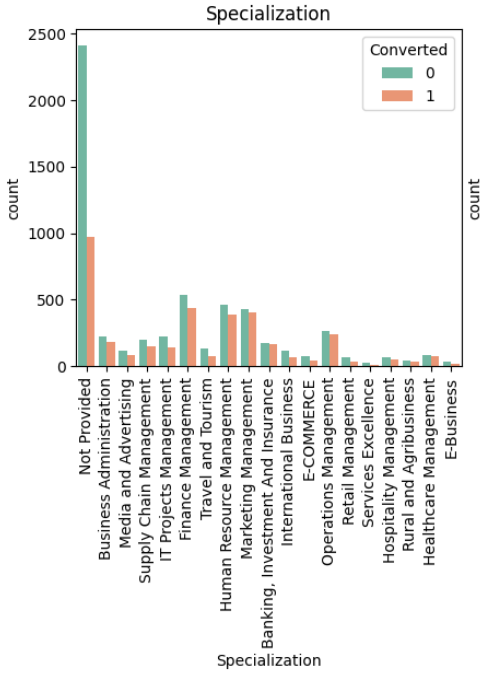


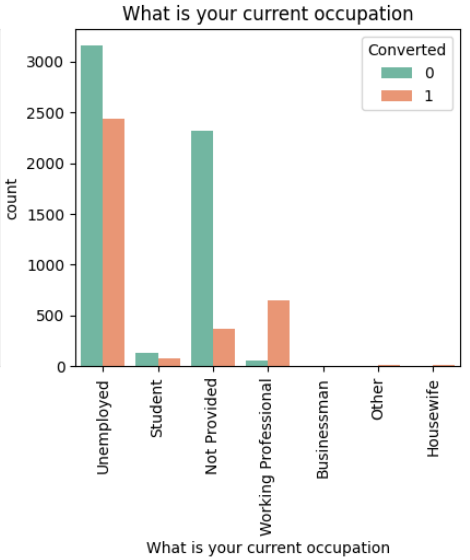












- API and Landing Page Submission have a 30-35% conversion rate but the count of lead originating from them are considerable.
- Lead Add Form has a more than 90% conversion rate, but the count of lead is not very high.
- Lead Imports are very less in the count.
- Lead Import and Quick Add Form get very few leads.

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

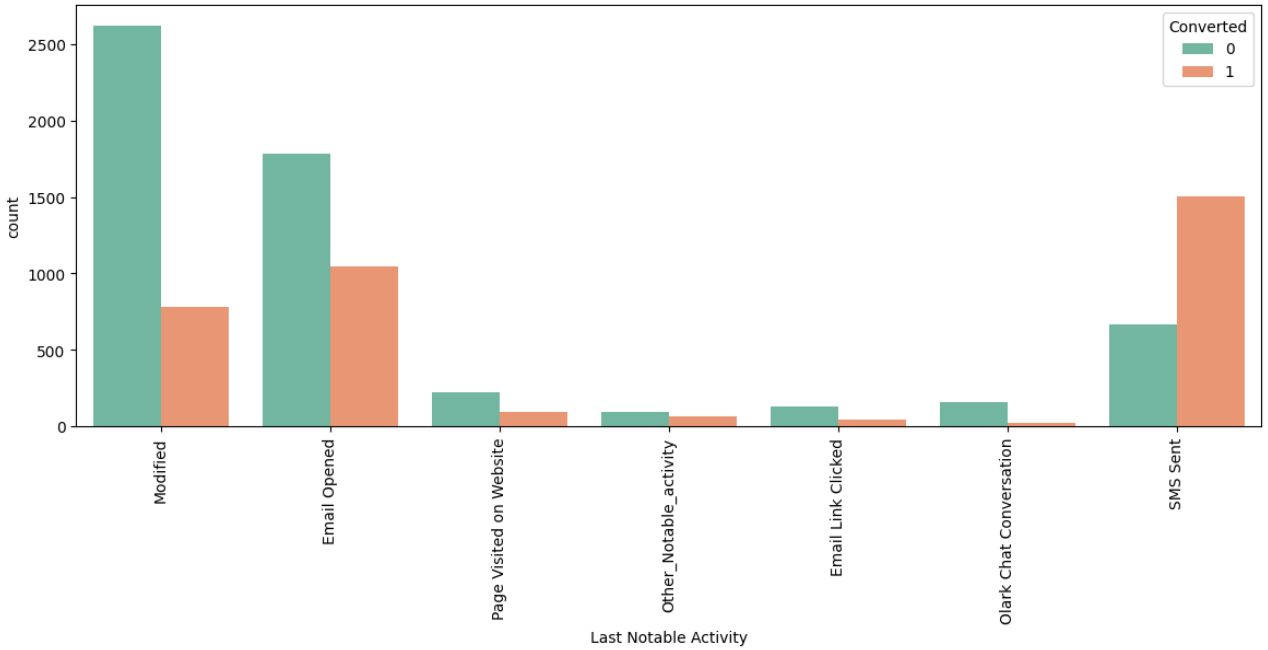
- Most entries are 'No'. No Inference can be drawn with this parameter and can be removed this feature.

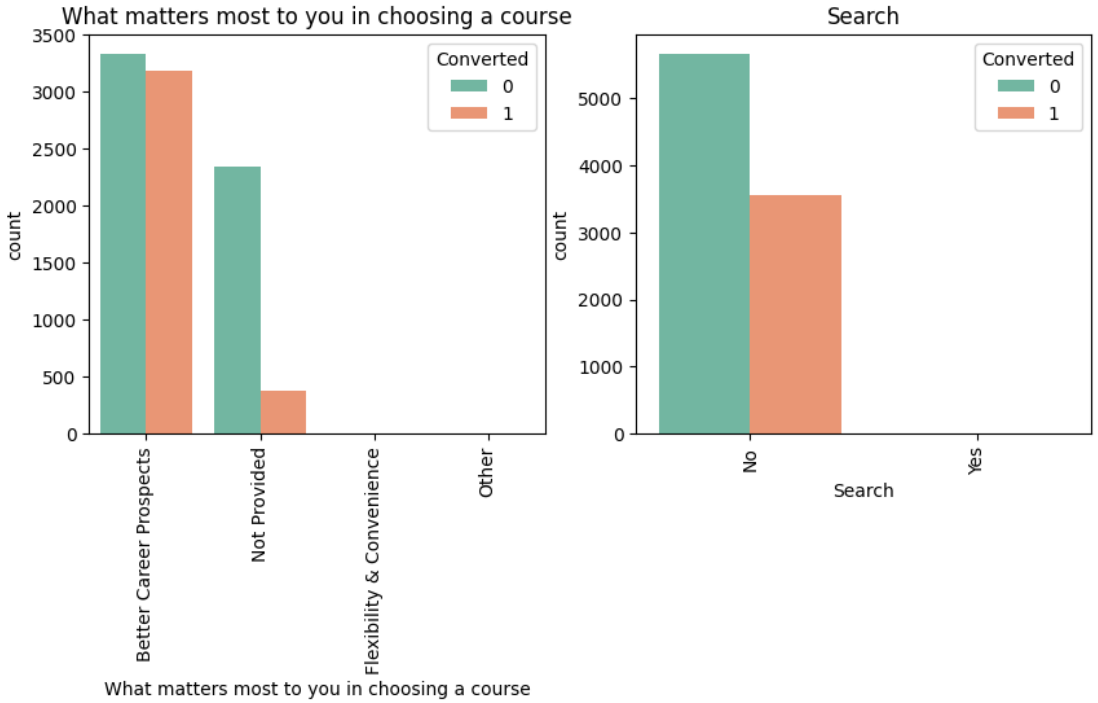
- Most of the leads have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent is almost 60%.
- For the Country most are from India so no such inferences can be done.

- Focus should be more on the Specialization with a high conversion rate.
- Working Professionals going for the course have a high chance of joining it.
- Unemployed leads are the most in numbers.

Univariate Analysis:

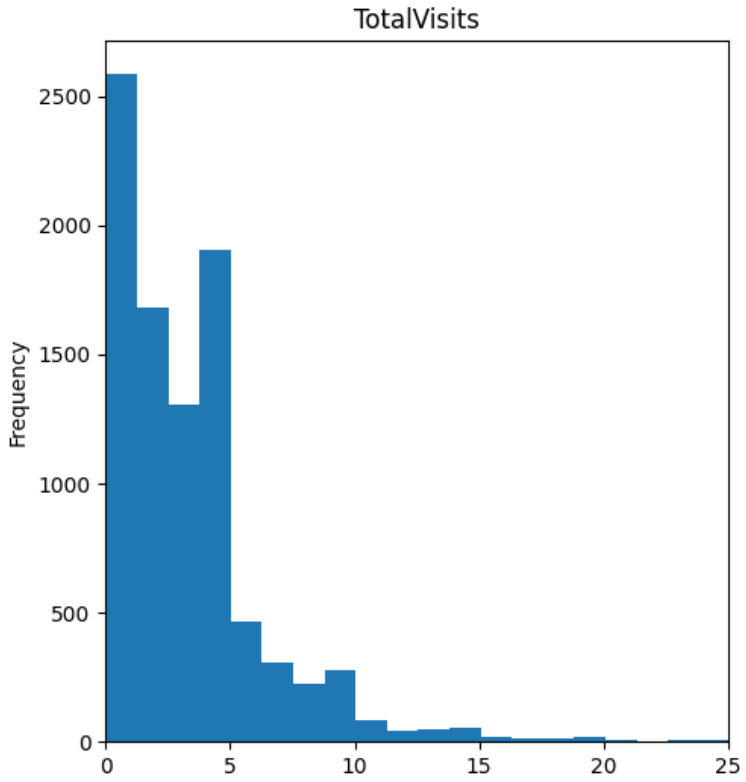
Analyzing all the categorical variables with the "Converted" Y Variable.

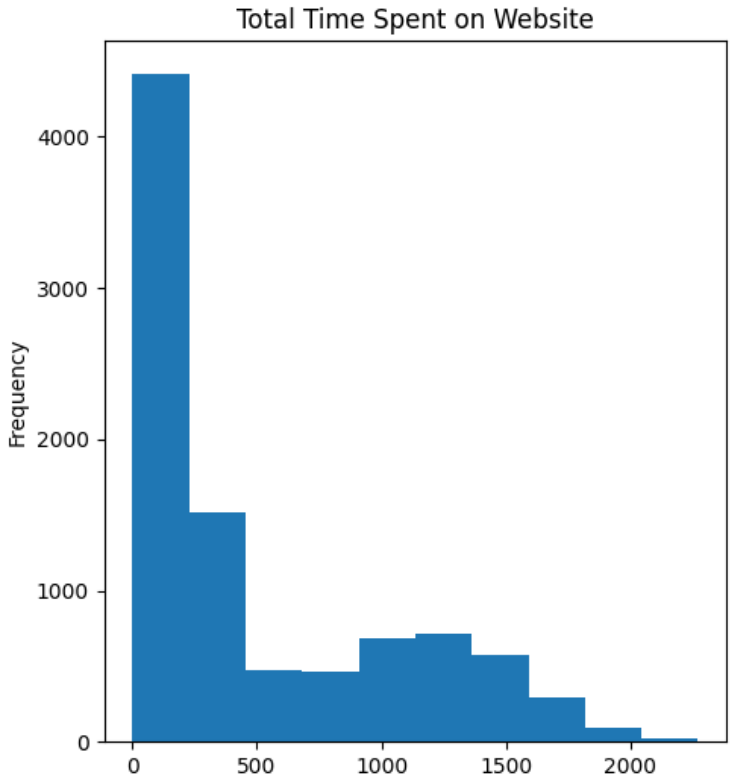


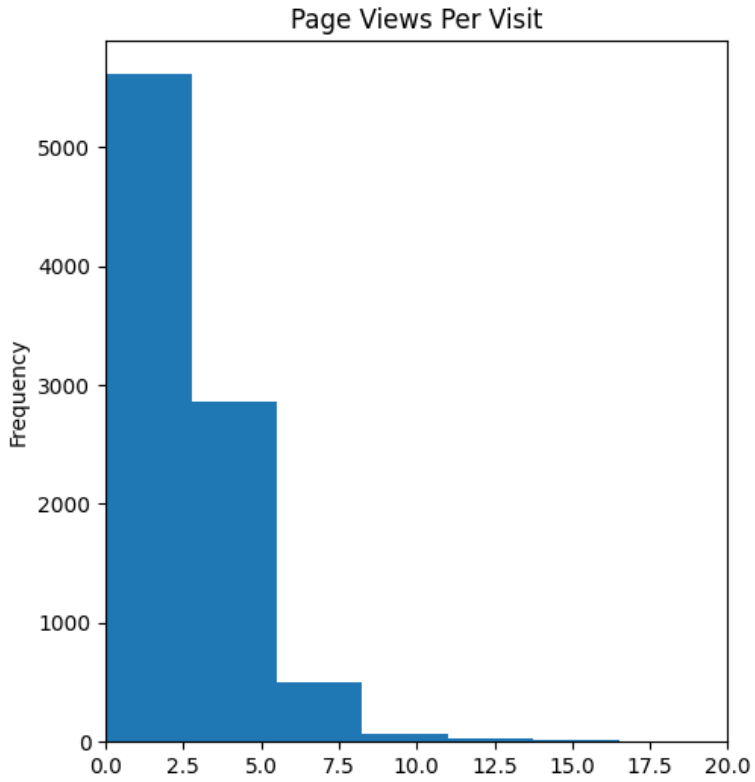


- We can see that that highly skewed column so we can remove this column.

Numerical Attributes Analysis:

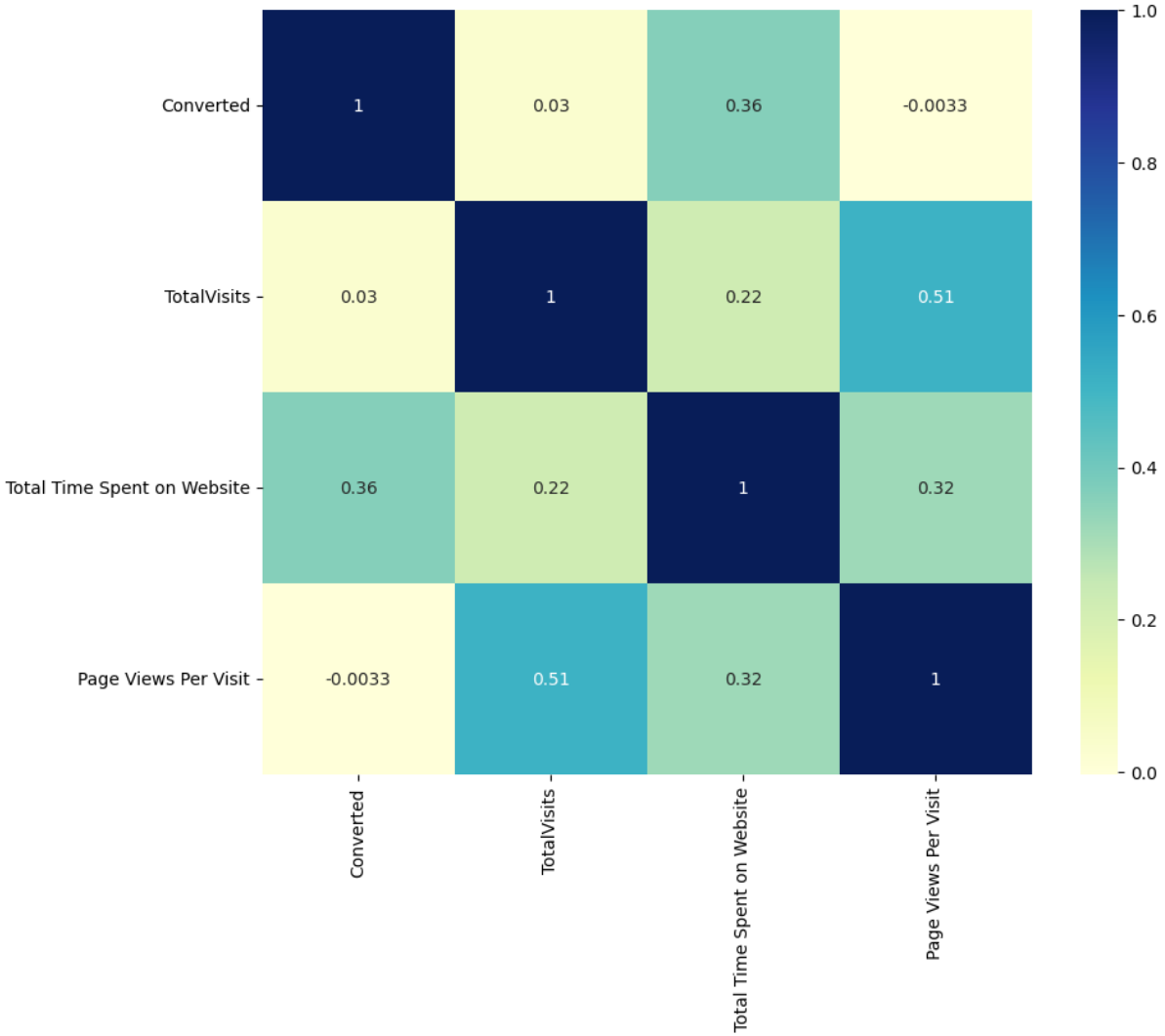


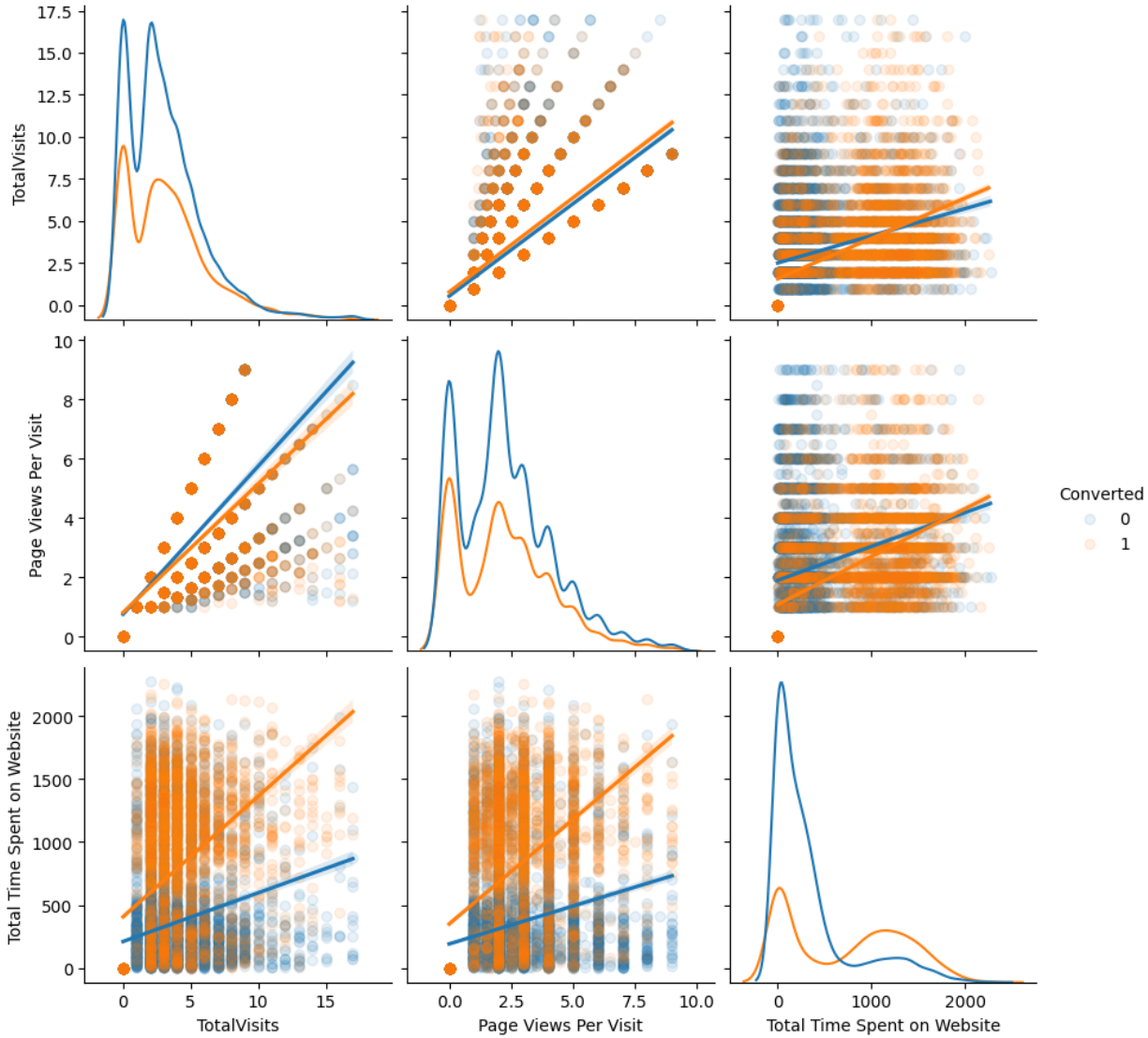




Bivariate Analysis:

Numerical & Categorical for Target variables:





Dummy Variable Creation

Dummy Creation
Encoding & Feature Scaling

Created dummy features for categorical variables with multiple levels (one hot encoded).

Creating a dummy variable for the categorical variables and dropping the first one.

'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'City', 'Last Notable Activity'.

Dropped the columns for which dummies were created above.

Model Building

To implement Logistic Regression, we will use the Scikit-learn library. We'll start by building a base model with default parameters, then look at how to improve it with Hyperparameter Tuning.

We have split the data into 70:30 Ratio as train and test sets. Then called the numerical variable using MinMaxScaller. Used RFE for Feature Selection and selected the 15 features by using the parameter n_features_to_select=15

Recursive Feature Elimination

Calculated precision and recall scores.

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

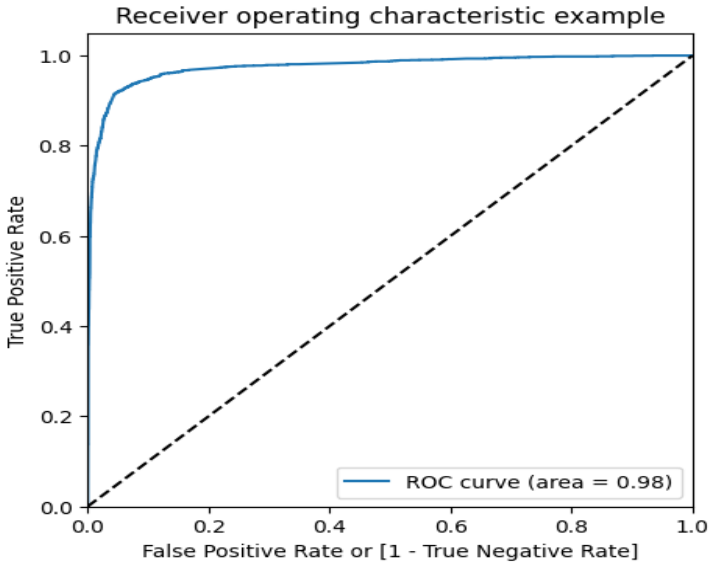
Calculated the values for sensitivity, specificity, positive predictive values and negative predictive values.

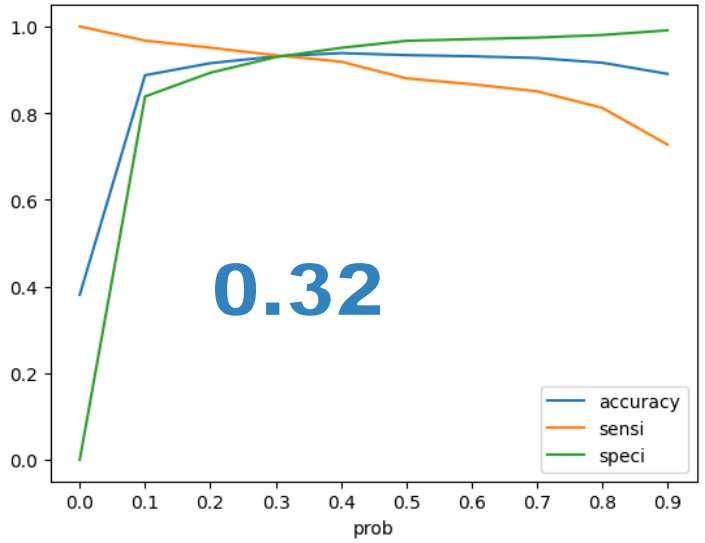
93.38%

The Model Build after the Feature Elimination And Selection Having

Accuracy

ROC CURVE





Sensitivity	Specificity	False Positive Rate	Positive Predictive	Negative predictive
88%	97%	0.03	94%	93%

Sensitivity
Sensitivity is ok But can be improved

88%

97%

Specificity
Specificity is good

ROC

The receiver operating characteristic (ROC) curve is another common tool used with binary classifiers. The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).

Observed Results

We found out that our specificity is good (~97%) but our sensitivity is 88%. Hence, this needs to be taken care of. We have got sensitivity of 88% and this is mainly because of the cut-off point of 0.5 that we had arbitrarily chosen. Now, this cut-off point had to be optimized in order to get a decent value of sensitivity and for this, we will use the ROC curve. Let's verify the cutoff point using the ROC Curve

Final Model Parameter & Features

Generalized Linear Model Regression Results

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:**	-1136.9
Date:	Thu, 29 Dec 2022	**Deviance:**	2273.7
Time:	15:28:36	**Pearson chi2:**	7.86e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6193
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1930	0.151	-7.926	0.000	-1.488	-0.898
Total Time Spent on Website	1.0634	0.064	16.682	0.000	0.938	1.188
Lead Origin_Lead Add Form	2.0156	0.447	4.514	0.000	1.140	2.891
Lead Source_Olark Chat	1.0487	0.153	6.839	0.000	0.748	1.349
Lead Source_Welingak Website	2.9752	1.111	2.678	0.007	0.797	5.153
Last Activity_Email Opened	0.9061	0.136	6.682	0.000	0.640	1.172
What is your current occupation_Not Provided	-2.1124	0.138	-15.356	0.000	-2.382	-1.843
Tags_Closed by Horizzon	5.2530	1.020	5.151	0.000	3.254	7.252
Tags_Interested in other courses	-3.7011	0.409	-9.041	0.000	-4.504	-2.899
Tags_Lost to EINS	4.6706	0.618	7.558	0.000	3.459	5.882
Tags_Other_Tags	-3.8392	0.231	-16.623	0.000	-4.292	-3.387
Tags_Ringing	-4.8986	0.264	-18.550	0.000	-5.416	-4.381
Tags_Will revert after reading the email	3.0476	0.203	15.023	0.000	2.650	3.445
Last Notable Activity_SMS Sent	3.0827	0.158	19.466	0.000	2.772	3.393

VIF

	Features	VIF
11	Tags_Will revert after reading the email	1.87
1	Lead Origin_Lead Add Form	1.85
4	Last Activity_Email Opened	1.80
12	Last Notable Activity_SMS Sent	1.71
2	Lead Source_Olark Chat	1.51
5	What is your current occupation_Not Provided	1.45
0	Total Time Spent on Website	1.44
3	Lead Source_Welingak Website	1.35
10	Tags_Ringing	1.30
6	Tags_Closed by Horizzon	1.20
9	Tags_Other_Tags	1.17
7	Tags_Interested in other courses	1.04
8	Tags_Lost to EINS	1.03

Model Evaluation-Precision and Recall

93%

Accuracy is defined as the ratio of correctly predicted examples to the total examples.

93%

Sensitivity (true positive rate) refers to the probability of a positive test, conditioned on truly being positive.

93%

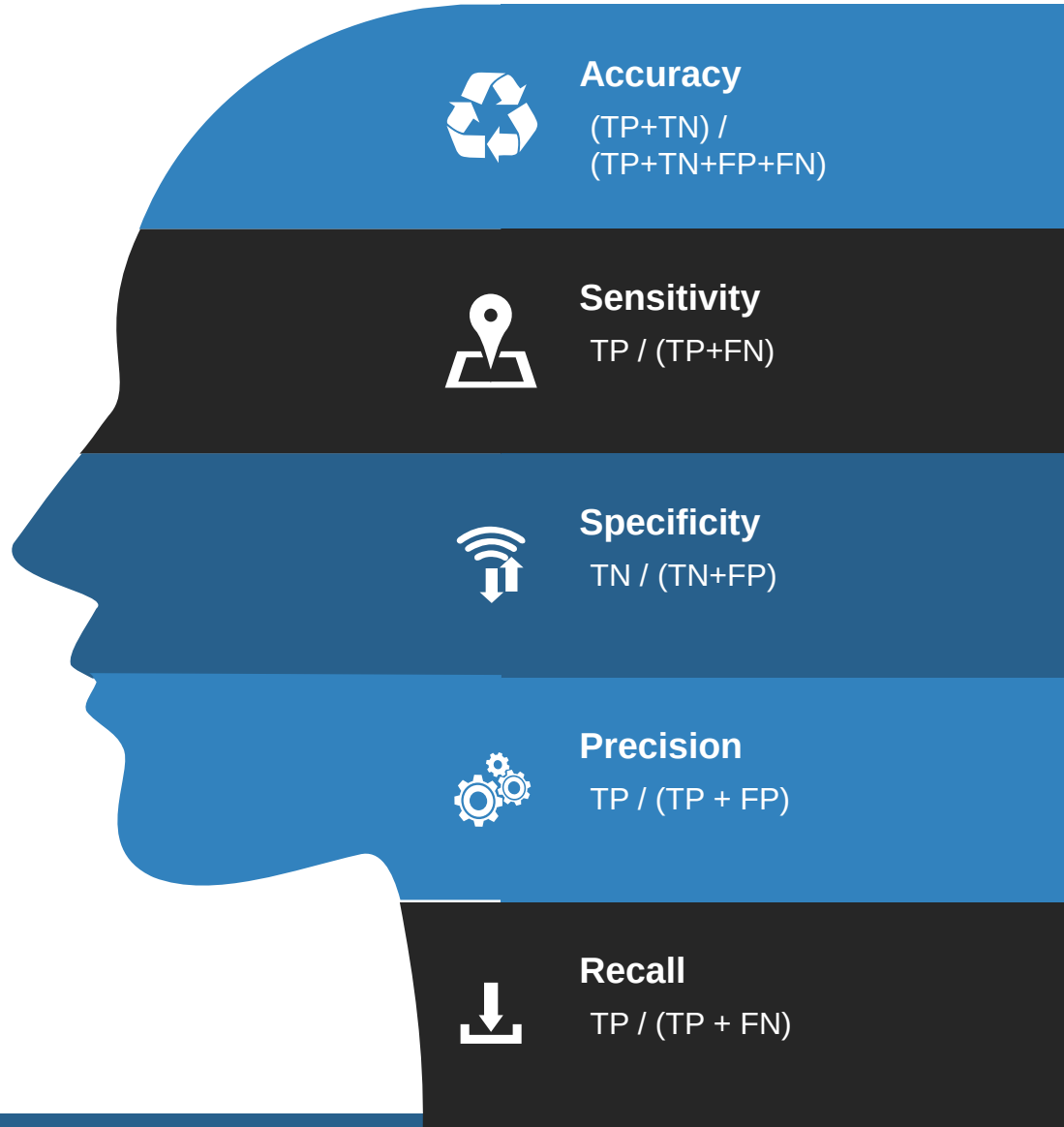
Specificity (true negative rate) refers to the probability of a negative test, conditioned on truly being negative.

90%

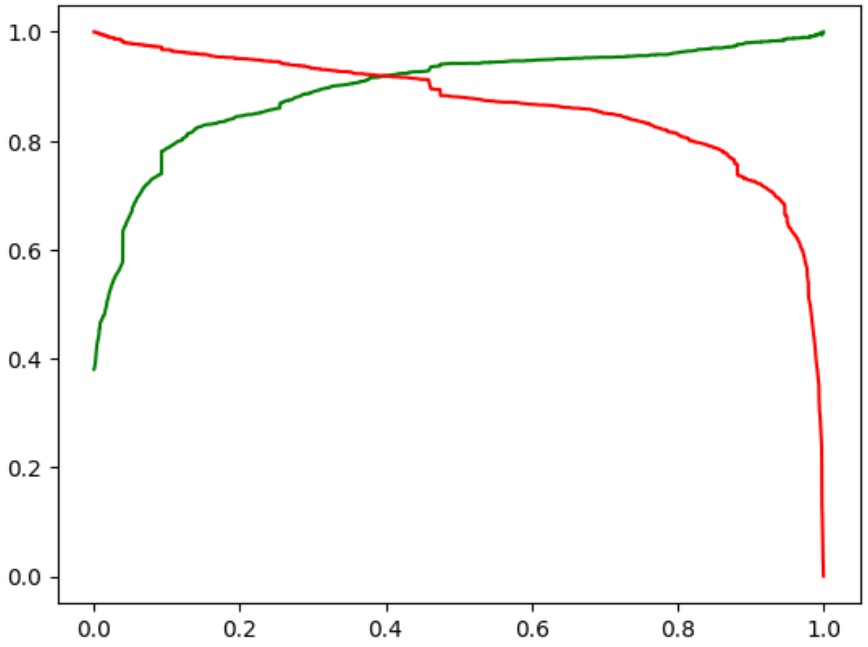
Precision also known as Positive Predictive Value, refers to the percentage of the results which are relevant.

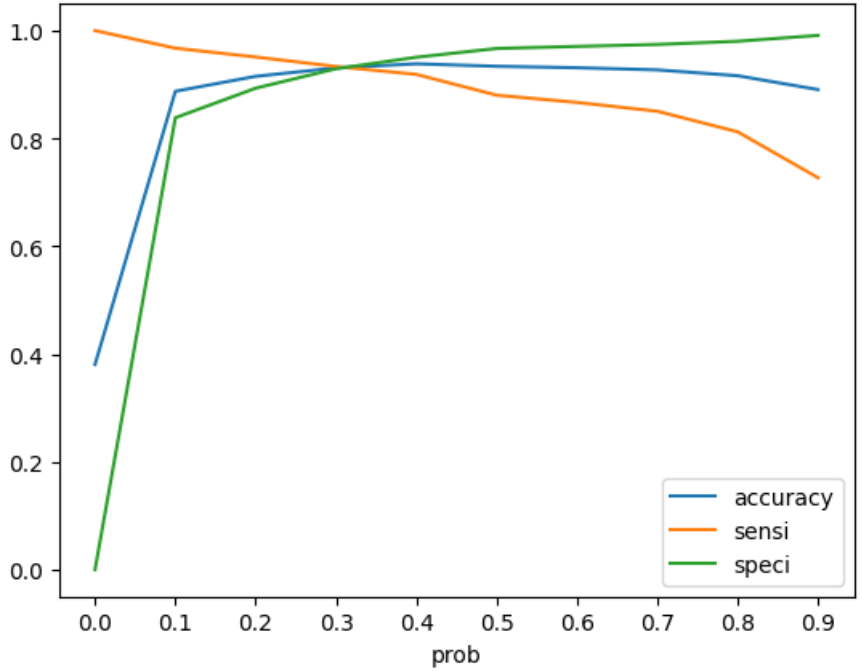
93%

Recall also known as Sensitivity, it refers to the percentage of total relevant results correctly classified by the algorithm.



Model Evaluation-Precision and Recall Tradeoff





Confusion Matrix

TRAIN SET

3628	254
166	2219

TEST SET

1581	95
63	947

TRAIN SET

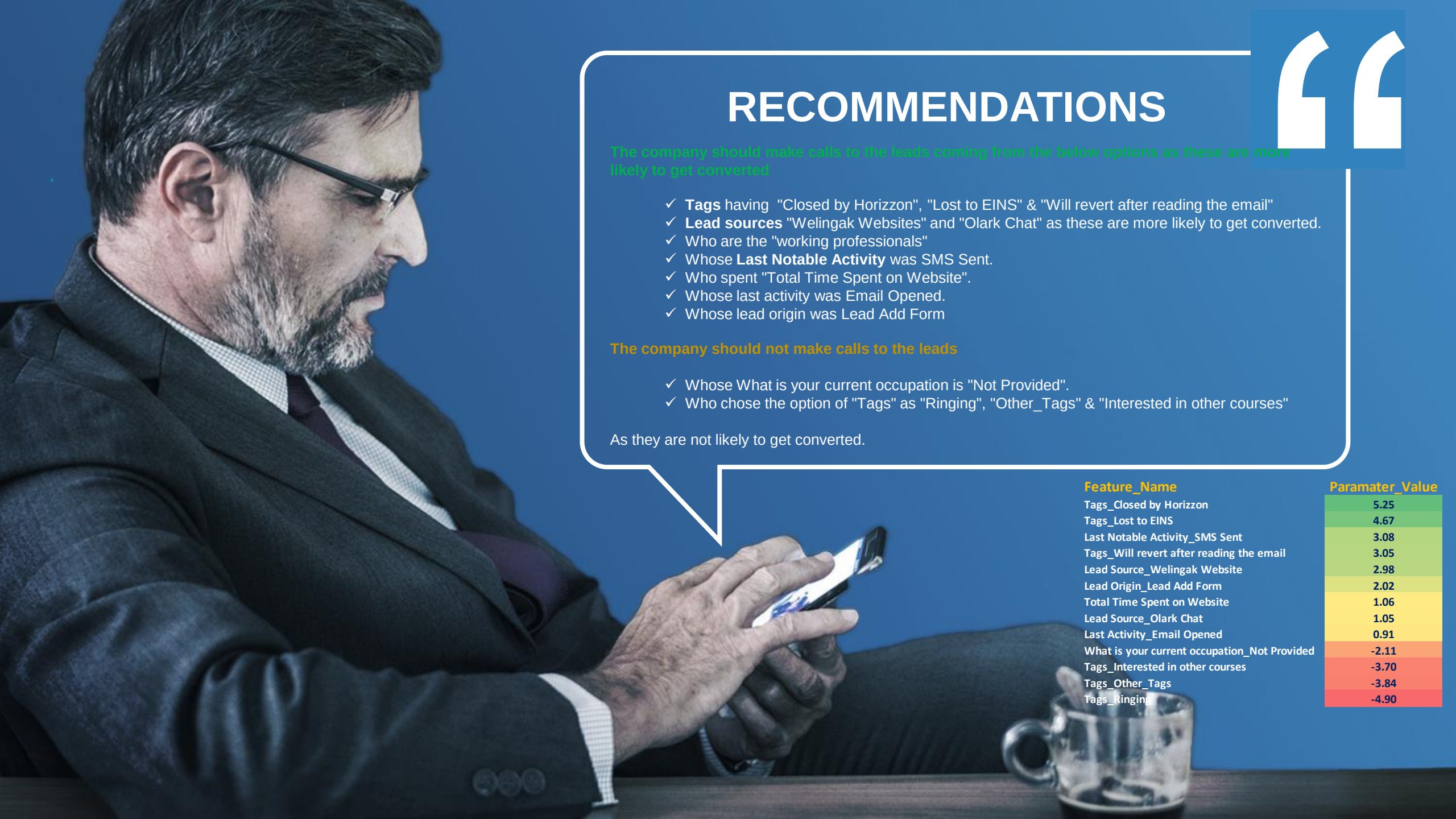
Sensitivity: 93.04
Specificity: 93.46
False Positive Rate: 0.07
Positive Predictive Value: 0.9
Negative predictive value: 0.96

Precision: 89.73
Recall: 93.04

TEST SET

Sensitivity: 93.76
Specificity: 94.33
False Positive Rate: 0.06
Positive Predictive Value: 0.91
Negative predictive value: 0.96

Precision: 90.88
Recall: 93.76



RECOMMENDATIONS

The company should make calls to the leads coming from the below options as these are more likely to get converted

- ✓ **Tags** having "Closed by Horizzon", "Lost to EINS" & "Will revert after reading the email"
- ✓ **Lead sources** "Welingak Websites" and "Olark Chat" as these are more likely to get converted.
- ✓ Who are the "working professionals"
- ✓ Whose **Last Notable Activity** was SMS Sent.
- ✓ Who spent "Total Time Spent on Website".
- ✓ Whose last activity was Email Opened.
- ✓ Whose lead origin was Lead Add Form

The company should not make calls to the leads

- ✓ Whose What is your current occupation is "Not Provided".
- ✓ Who chose the option of "Tags" as "Ringing", "Other_Tags" & "Interested in other courses"

As they are not likely to get converted.

Feature_Name	Paramater_Value
Tags_Closed by Horizzon	5.25
Tags_Lost to EINS	4.67
Last Notable Activity_SMS Sent	3.08
Tags_Will revert after reading the email	3.05
Lead Source_Welingak Website	2.98
Lead Origin_Lead Add Form	2.02
Total Time Spent on Website	1.06
Lead Source_Olark Chat	1.05
Last Activity_Email Opened	0.91
What is your current occupation_Not Provided	-2.11
Tags_Interested in other courses	-3.70
Tags_Other_Tags	-3.84
Tags_Ringing	-4.90

Thank You

Adapt it to your needs!

Thank you!