Lead Scoring Case Study

Presenter : Saravana Kumar R
Sudheer K S
Sonali Bisht
Group : DSC 71
Date : 17-02-2025

Agenda

01 Business Objectives

- ❑ Defining the core business goals and key problem statements driving the case study.

02 Exploratory Data Analysis (EDA)

- ❑ Reviewing data cleaning, handling missing values, detecting outliers, and addressing data imbalance.

03 Model Development & Evaluation

- ❑ Building a logistic regression model for lead scoring and assessing performance using metrics like specificity, sensitivity, precision, and recall.

04 Insights & Recommendations

- ❑ Presenting key findings from the analysis and strategic recommendations to optimize business outcomes.

01 Business Context & Objectives

Overview:

- This case study demonstrates how logistic regression can be used to score and prioritize leads for better conversion.

Company Profile:

- X Education provides online courses for working professionals.
- The company generates leads through multiple sources (website, search engines, referrals).
- The sales team follows up with leads via calls and emails, leading to conversions at a typical rate of 30%.
- The CEO has set a target conversion rate of 80%, making lead prioritization crucial.

Goal:

- Develop a **logistic regression model** to assign a score (0-100) to each lead.
- Higher scores indicate a stronger likelihood of conversion, helping the sales team focus on high-potential leads.

Problem Solving Methodology

Comprehensive Guide to Data Processing and Exploratory Analysis

Exploratory Data Analysis

This pillar involves examining data distributions, identifying correlations, and checking for class imbalances to gain insights.

Feature Engineering

This pillar entails transforming and scaling data to enhance model performance by creating suitable features.

Data Collection & Cleaning

This pillar focuses on gathering data from various sources and preparing it for analysis by addressing missing values and outliers.

Data Splitting

This pillar involves dividing the dataset into training and testing subsets to evaluate model performance.

Reading & understanding 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 represe nts

```
# 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 i n 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 9240 rows and 37 columns.
- ✓ 30 columns have Object type, and the rest of the others are either float or int eger.
- ✓ 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.

02 Exploratory Data Analysis (EDA)

Data Preparation and Analysis Process

Data Cleaning

Process of removing unwanted variables and values

EDA Steps & Analysis

Identifying and handling missing values in columns

Outliers

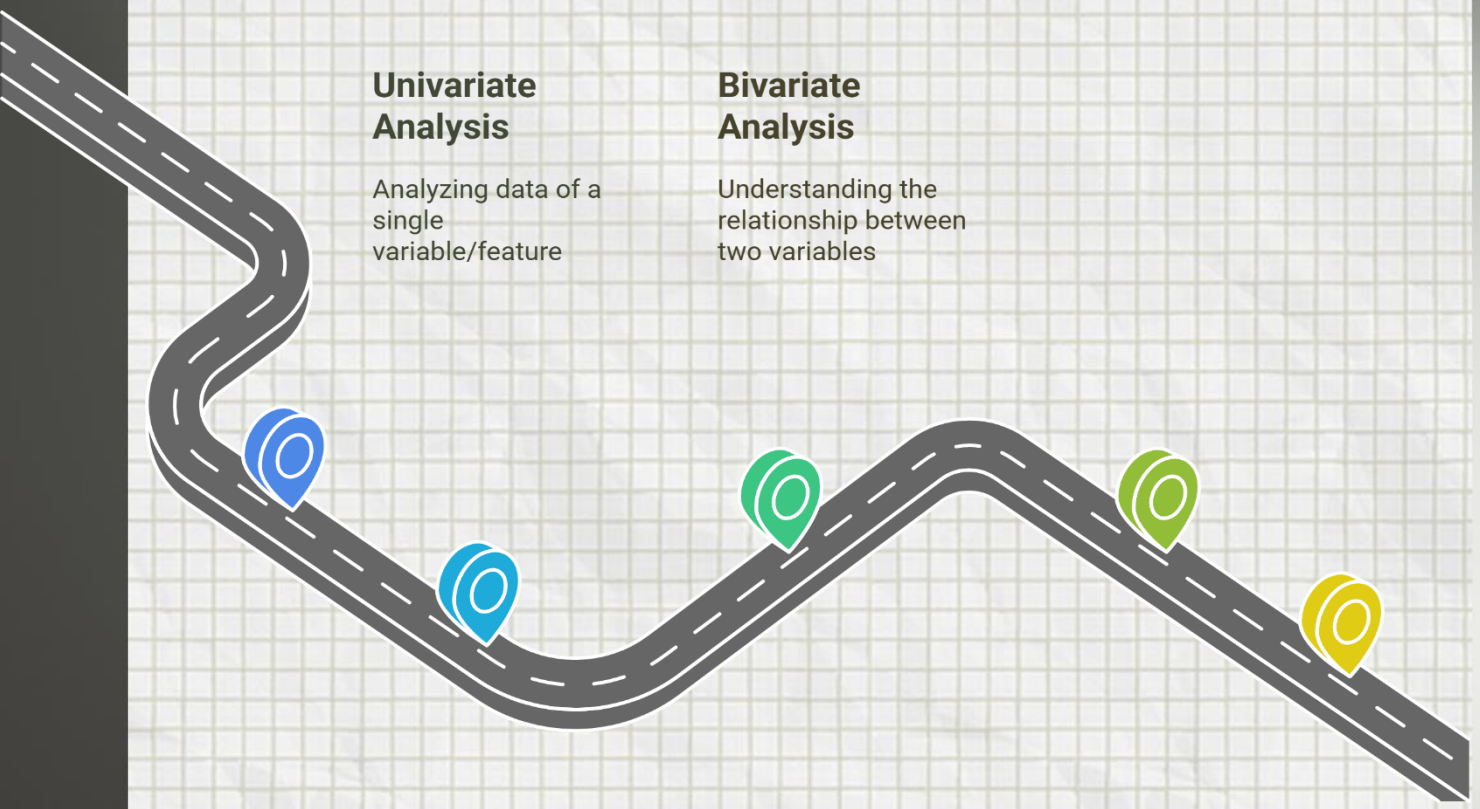
Addressing univariate and multivariate outliers

Univariate Analysis

Analyzing data of a single variable/feature

Bivariate Analysis

Understanding the relationship between two variables



Data Cleaning & Handling Missing Values

Data Cleaning

- Identified and removed columns with more than 45% null values in the leads.csv file.

Identifying Categorical and Numerical Columns

- Since the describe() function operates on numerical data, we identified and listed numerical columns from the dataset to distinguish them from categorical columns.

Handling Missing Values and Imputation

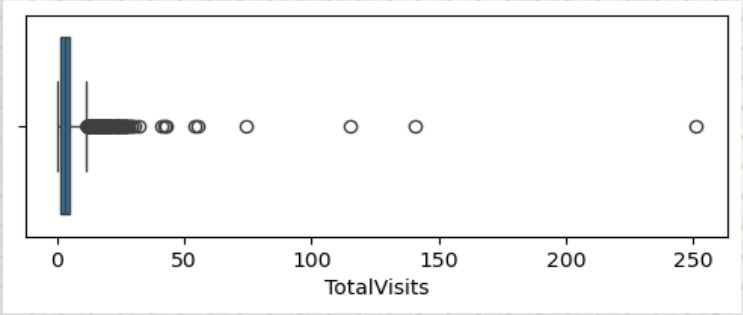
- Detected outliers and applied the most suitable imputation methods.
- Noted significant null values in certain columns. Removing rows with nulls would result in substantial data loss, especially in crucial column.
- Replaced NaN values with 'not provided' to preserve data integrity w hile minimizing null entries. These can be dropped later if they prove i rrelevant to the model.
- Recognized 'Select' values in multiple columns, likely due to customer s not choosing an option. These 'Select' entries were treated as null v alues and converted accordingly.

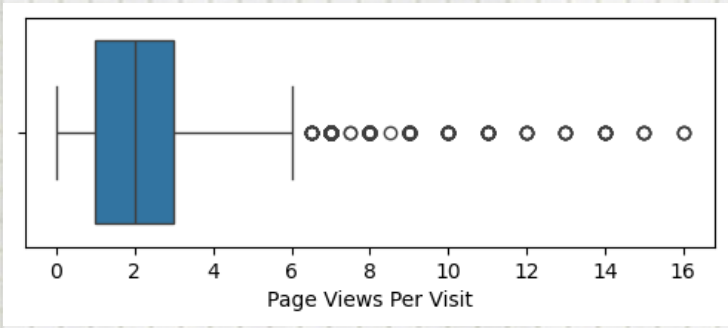
```
# Now again Checking for Null value
round(100*(leads.isnull().sum()/len(leads.index)),2).sort_values(ascending = False)
```

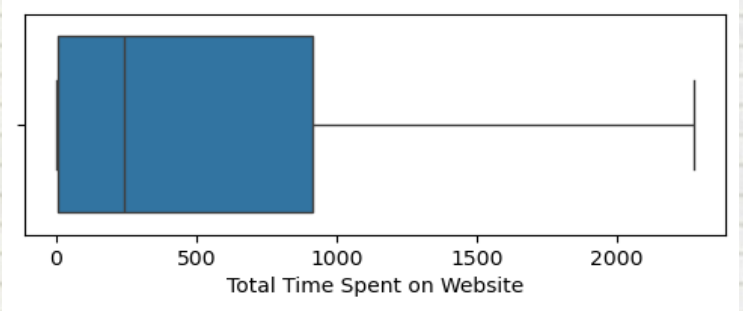
```
How did you hear about X Education              78.46
Lead Profile                                    74.19
Lead Quality                                    51.59
Asymmetrique Activity Score                     45.65
Asymmetrique Profile Score                      45.65
Asymmetrique Profile Index                      45.65
Asymmetrique Activity Index                     45.65
City                                            39.71
Specialization                                  36.58
Tags                                            36.29
What matters most to you in choosing a course   29.32
What is your current occupation                 29.11
Country                                         26.63
TotalVisits                                      1.48
Page Views Per Visit                             1.48
Last Activity                                    1.11
Lead Source                                      0.39
Get updates on DM Content                        0.00
```

Outliers Analysis

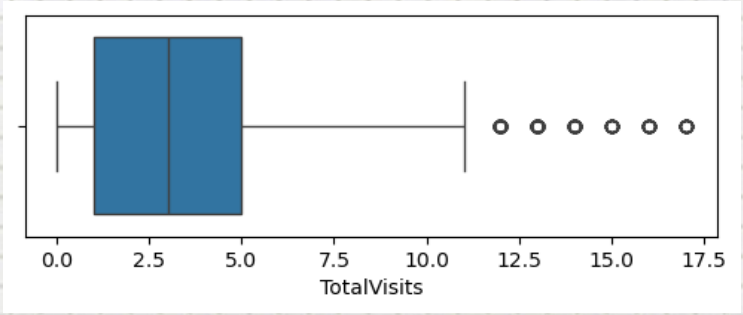
Before outlier Removal

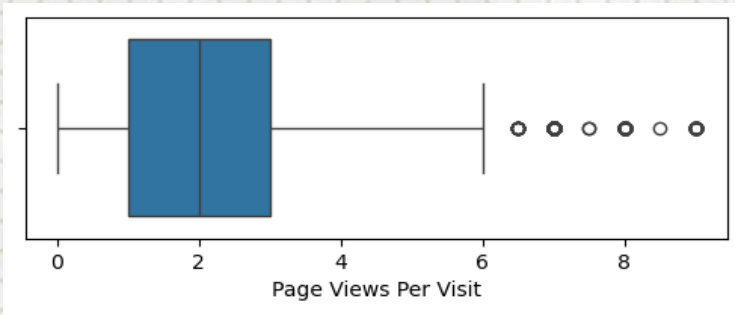






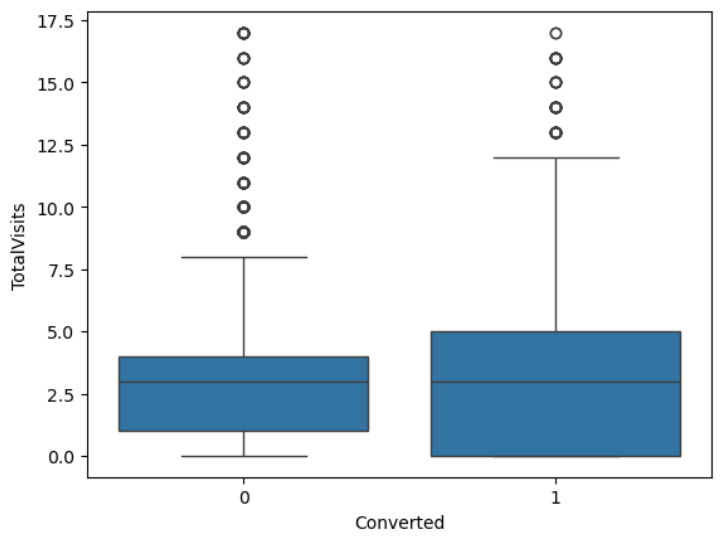
After Outlier Removal



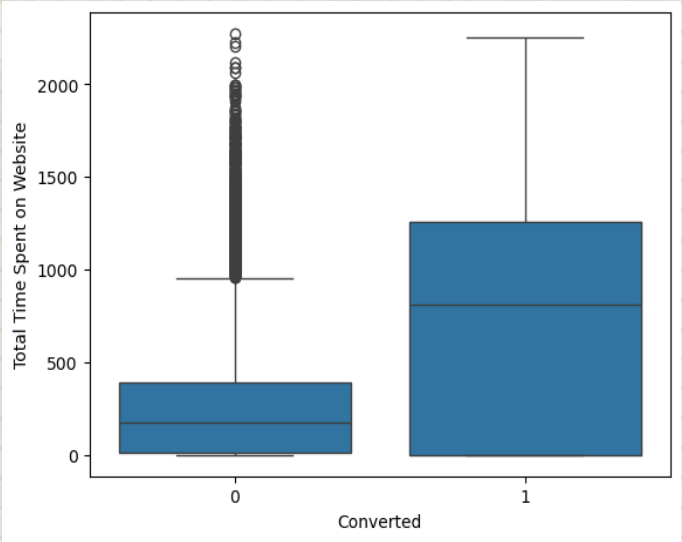


- **Outliers were detected** in the **"Total Visits"** and **"Page Views per Visit"** columns.
- To address this, the **top and bottom 1% of outlier values** were removed from both variables.
- The **"Total Time Spent on the Website"** showed a consistent distribution with **no significant outliers** observed.

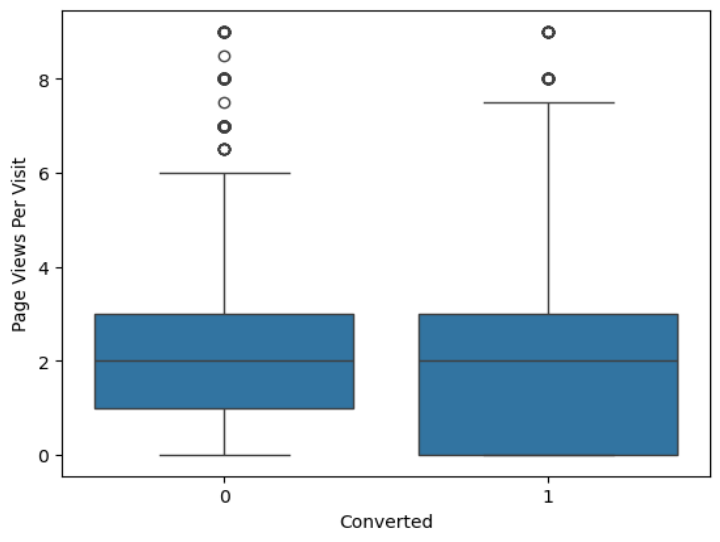
Outliers Analysis – Checking with Converted variable



- The median values for con verted and non-converted leads are similar
- No clear insights can be dr awn from Total visits.



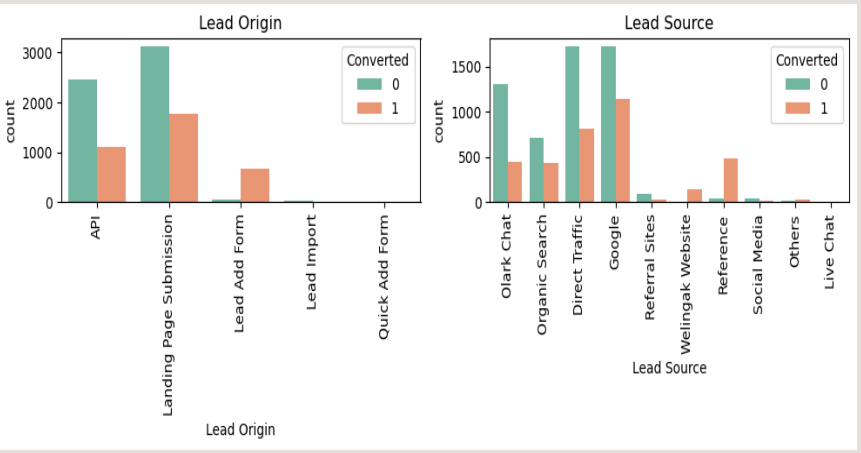
- Leads spending more time o n the website are more likely to be converted.
- Website should be made mo re engaging to make leads sp end more time.

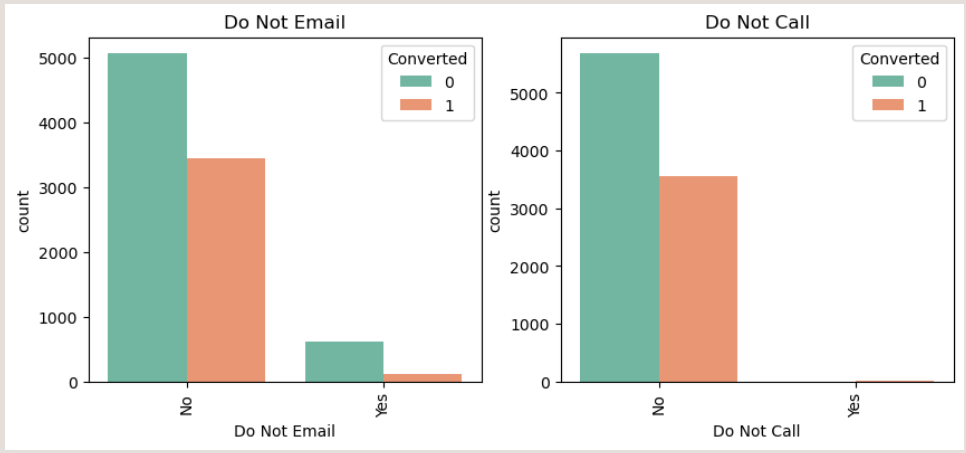


- The median values for con verted and non-converted leads are similar
- No clear insights can be dr awn from Total visits

Univariate Analysis

(Categorical columns)





Inference from Lead Origin:

•API and Landing Page Submissions have a 30-35% co nversion rate with a significant lead count.

•Lead Add Form has a 90%+ conversion rate, but few er leads.

•Lead Import and Quick Add Form generate very few leads.

Inference from Lead Source:

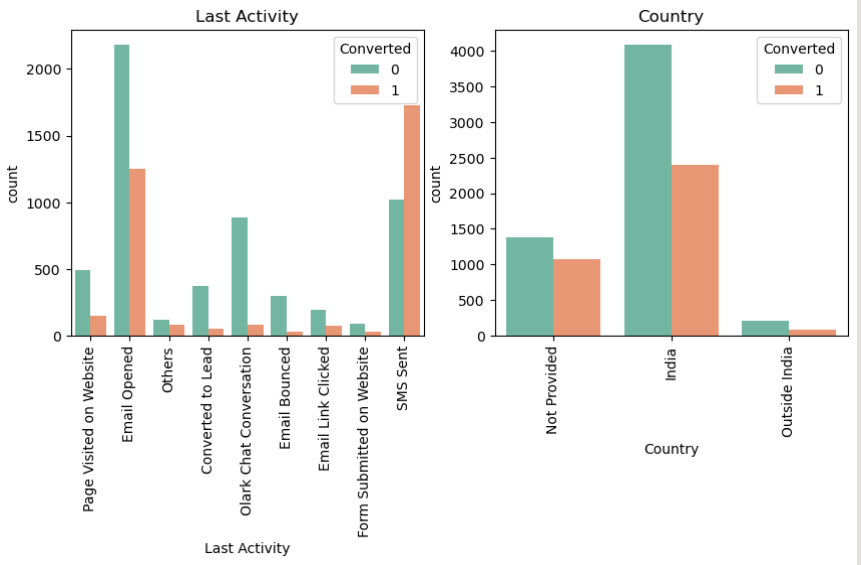
•Most leads come from Google and Direct Traffic.

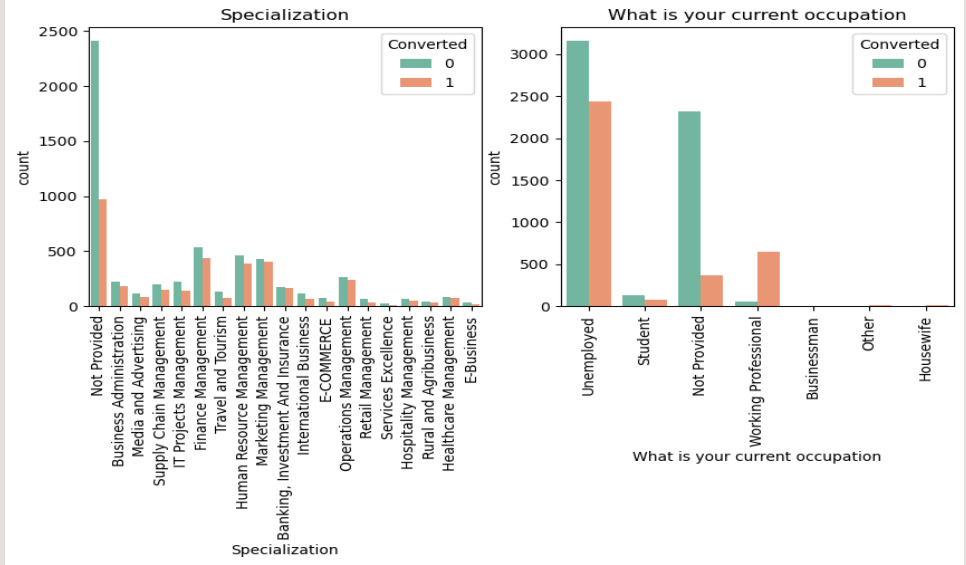
•Reference leads and Welingak website have high co nversion rates.

•Focus on improving conversion for Olark Chat, Orga nic Search, Direct Traffic, and Google, while increasin g Reference and Welingak website leads.

Inference- Do Not Email & Do Not Call

Most entries are 'No'. No Inference can be drawn with t his parameter and can be removed this feature.





Inference -Last activity and country

•Most leads had their last activity as "Email Opened."

•Leads with "SMS Sent" have a ~60% conversion rate.

•Since most leads are from India, country-based insight are limited.

Inference - Specialization and What is your current occ upation

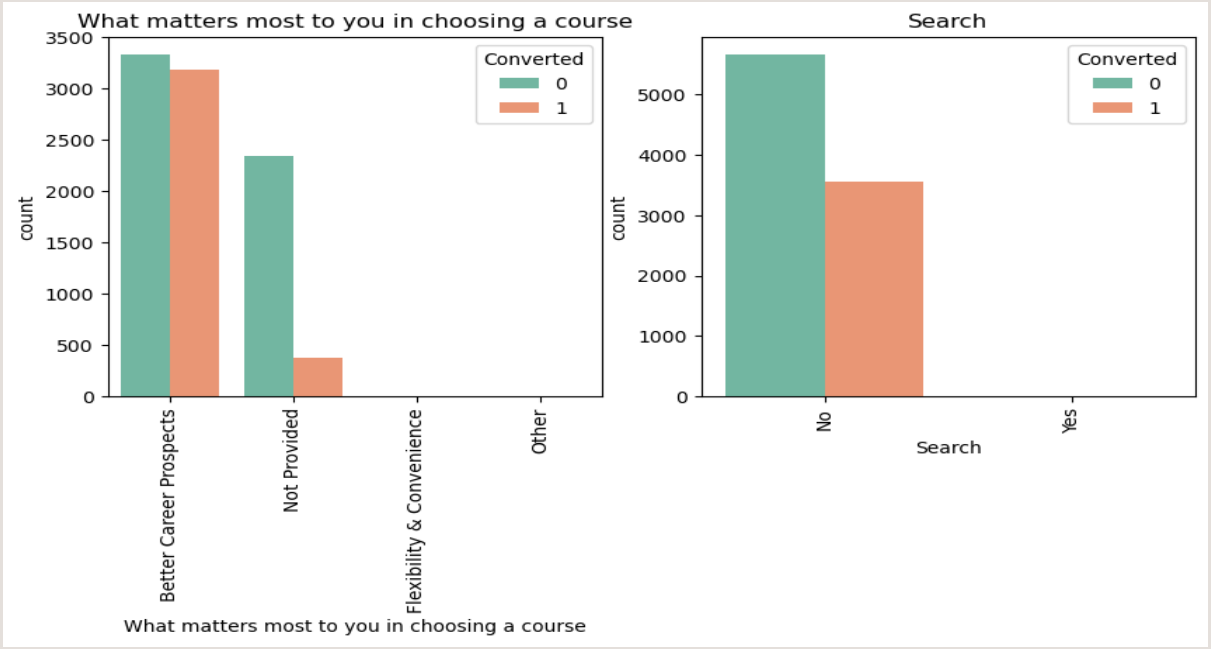
•Focus should be more on the Specialization with high conversion rate.

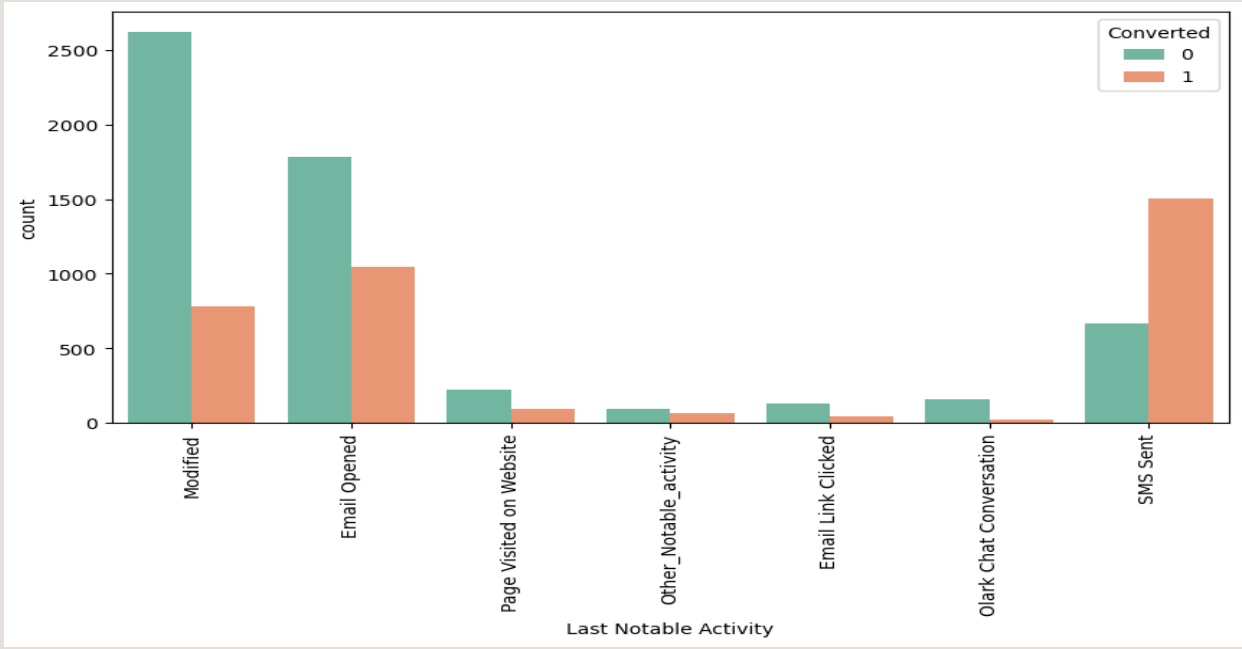
•Working Professionals going for the course have high chances of joining it.

•Unemployed leads are the most in numbers.

•We see that specialization with Management in them have higher number of leads as well as leads converted So this is definitely a significant variable and should not be dropped

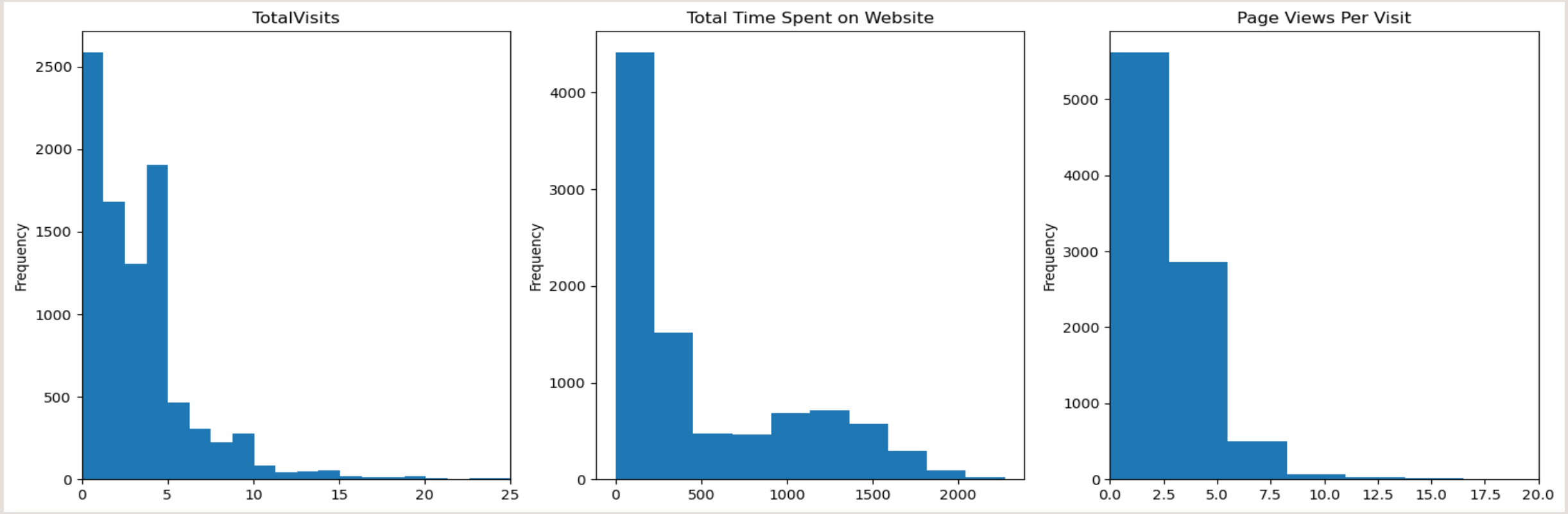
Univariate Analysis





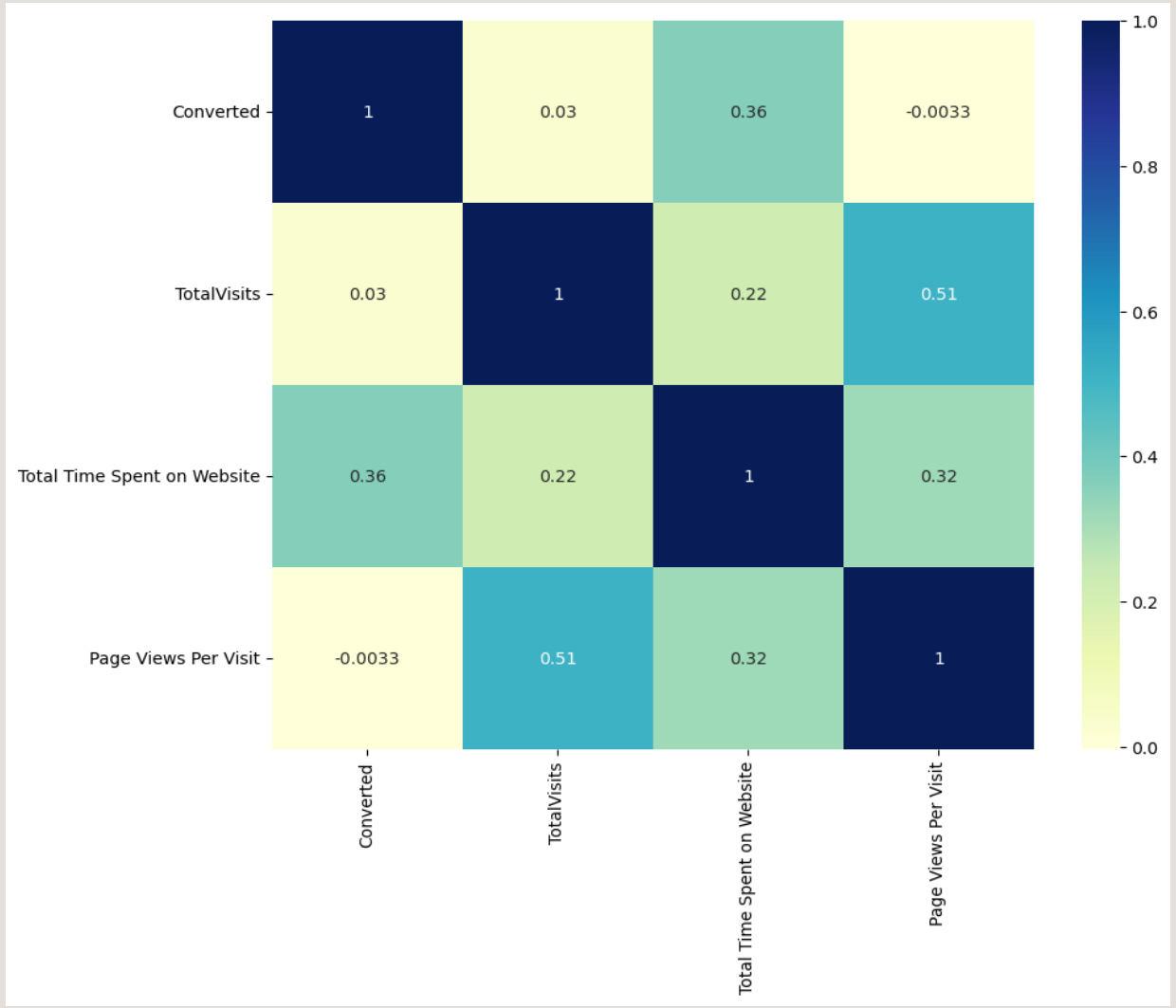
- We can see that that highly skewed column so we can remove this column.

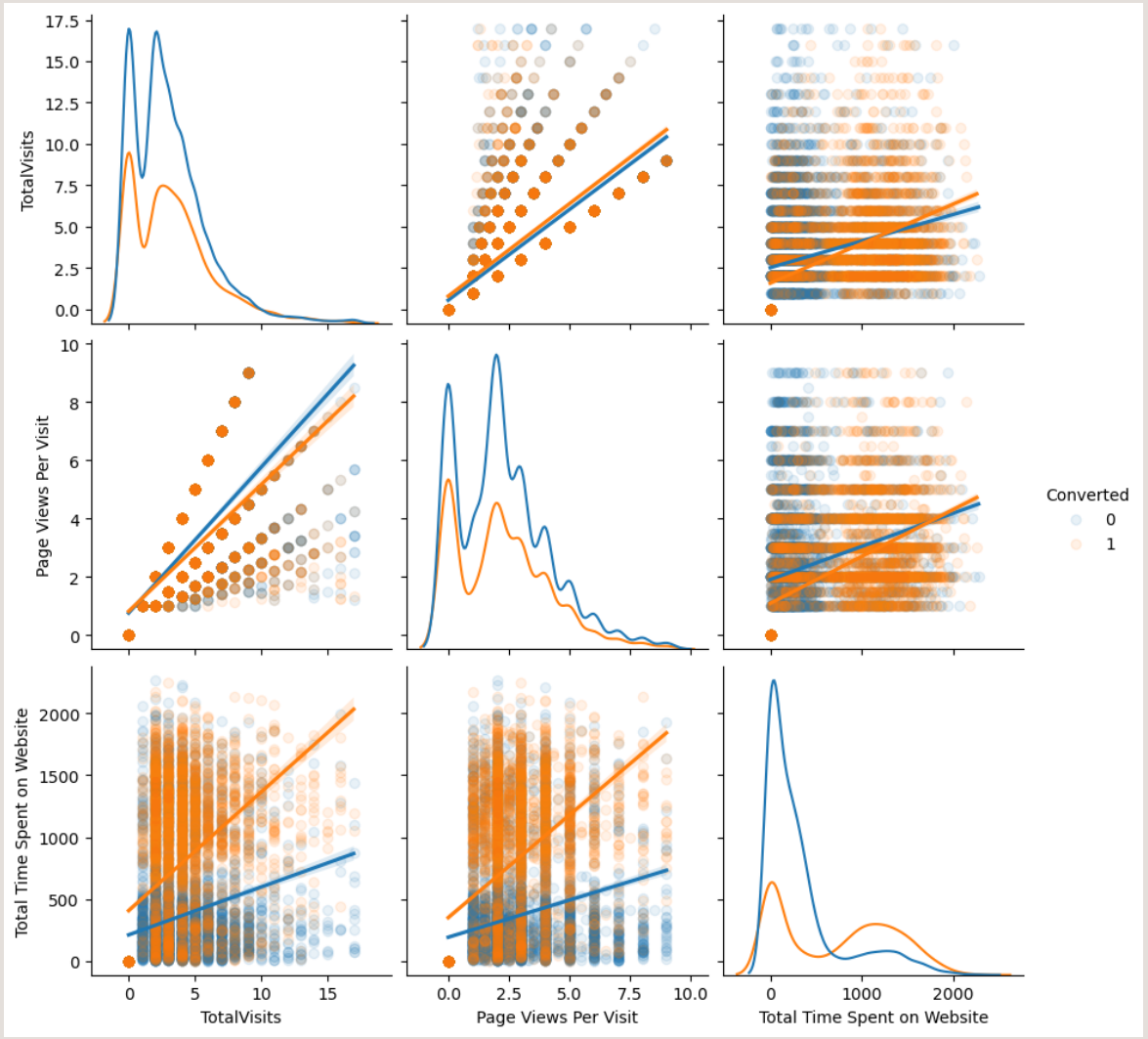
Numerical value Analysis



Bivariate Analysis

(Numerical & Categorical for Target variables)





Dummy Creation, Encoding & Feature Scaling

Dummy Variables:

- Created for categorical features with multiple levels using one-hot encoding.

Columns Encoded:

- Lead Origin
- Lead Source
- Last Activity
- Specialization
- What is your current occupation
- Tags
- City
- Last Notable Activity

Post-Encoding:

- Dropped the original columns after dummy creation to streamline the dataset.

Feature Scaling:

- Applied to ensure consistent scale across features for modeling.

03 Model Development & Evaluation

Model Building

•Utilize the Scikit-learn library to implement Logistic Regression.

•Begin by constructing a base model using default parameters.

•Enhance the model's performance through Hyperparameter Tuning.

Data Splitting and Feature Selection

•Split the dataset into a 70:30 ratio for training and testing sets.

•Applied MinMaxScaler to normalize the numerical variables.

•Utilized Recursive Feature Elimination (RFE) for feature selection.

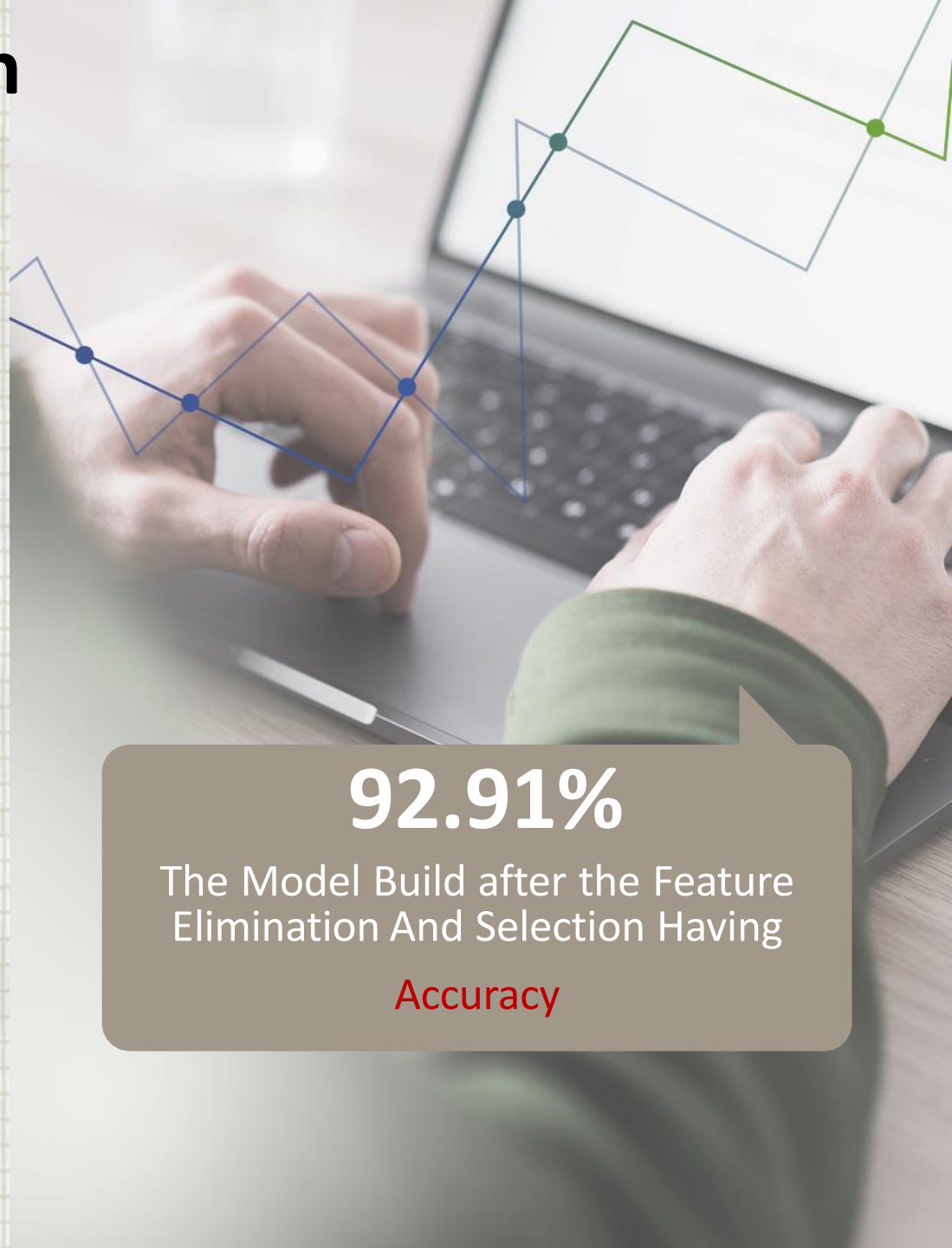
•Selected the top 15 features by setting the parameter n_features_to_select=15.

Model Performance After Feature Elimination and Selection

•Achieved an accuracy of **92.91%** post-feature elimination and selection.

•The model was built using the top 15 features selected through Recursive Feature Elimination (RFE).

•This high accuracy reflects the effectiveness of the feature selection process in im proving model performance.



92.91%

The Model Build after the Feature Elimination And Selection Having

Accuracy

Final Model Parameter & Features

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6254
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1284.4
Date:	Sun, 16 Feb 2025	**Deviance:**	2568.8
Time:	20:01:51	**Pearson chi2:**	1.00e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6010
Covariance Type:	nonrobust		

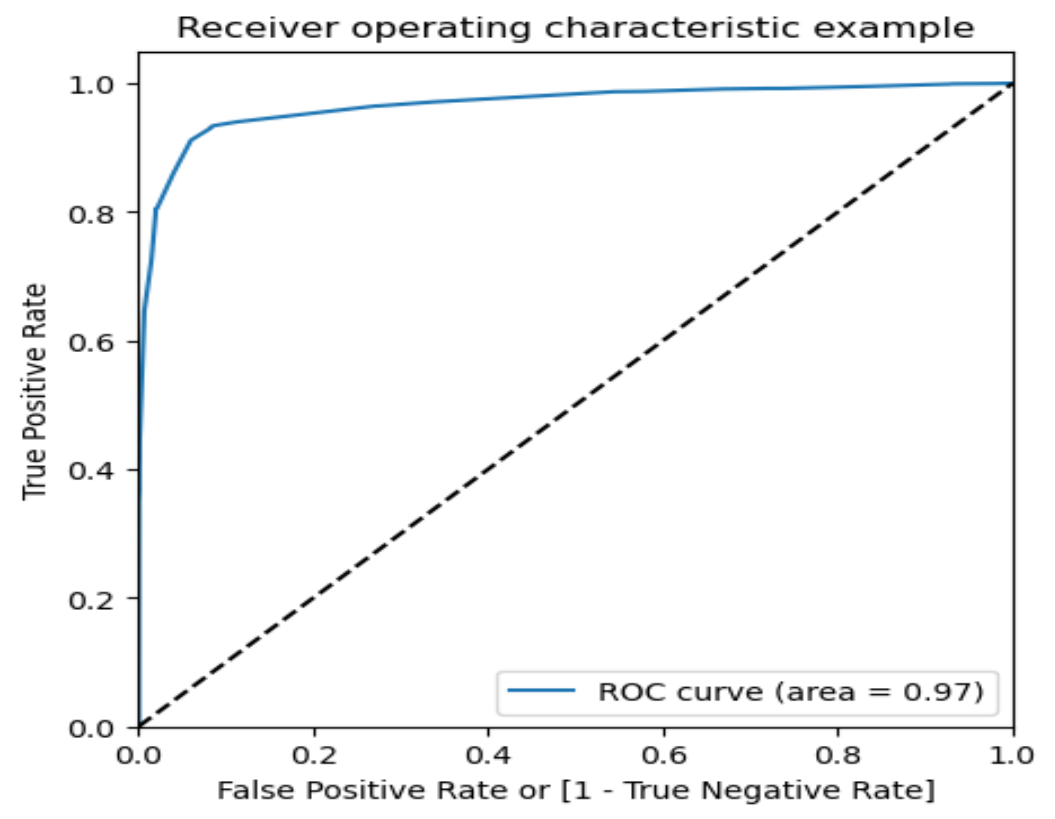
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7524	0.135	-5.567	0.000	-1.017	-0.488
Lead Source_Welingak Website	3.7037	1.021	3.628	0.000	1.703	5.705
Last Activity_Email Bounced	-1.5579	0.471	-3.307	0.001	-2.481	-0.635
Last Activity_Email Opened	0.8801	0.132	6.661	0.000	0.621	1.139
What is your current occupation_Not Provided	-2.1870	0.129	-16.931	0.000	-2.440	-1.934
Tags_Closed by Horizzon	5.6060	1.010	5.552	0.000	3.627	7.585
Tags_Interested in other courses	-3.5073	0.379	-9.245	0.000	-4.251	-2.764
Tags_Lost to EINS	5.0489	0.617	8.177	0.000	3.839	6.259
Tags_Other_Tags	-3.9034	0.218	-17.946	0.000	-4.330	-3.477
Tags_Ringing	-4.8438	0.249	-19.486	0.000	-5.331	-4.357
Tags_Will revert after reading the email	3.0728	0.190	16.168	0.000	2.700	3.445
Last Notable Activity_Other_Notable_activity	1.4194	0.419	3.389	0.001	0.599	2.240
Last Notable Activity_SMS Sent	3.1164	0.152	20.474	0.000	2.818	3.415

VIF

	Features	VIF
2	Last Activity_Email Opened	1.86
11	Last Notable Activity_SMS Sent	1.70
9	Tags_Will revert after reading the email	1.69
3	What is your current occupation_Not Provided	1.32
8	Tags_Ringing	1.31
7	Tags_Other_Tags	1.20
1	Last Activity_Email Bounced	1.15
10	Last Notable Activity_Other_Notable_activity	1.11
4	Tags_Closed by Horizzon	1.05
0	Lead Source_Welingak Website	1.04
5	Tags_Interested in other courses	1.03
6	Tags_Lost to EINS	1.02

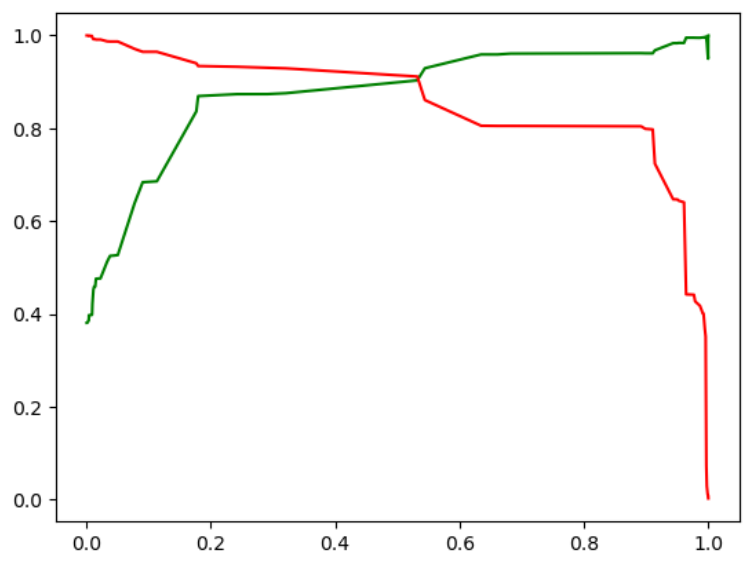
ROC Curve

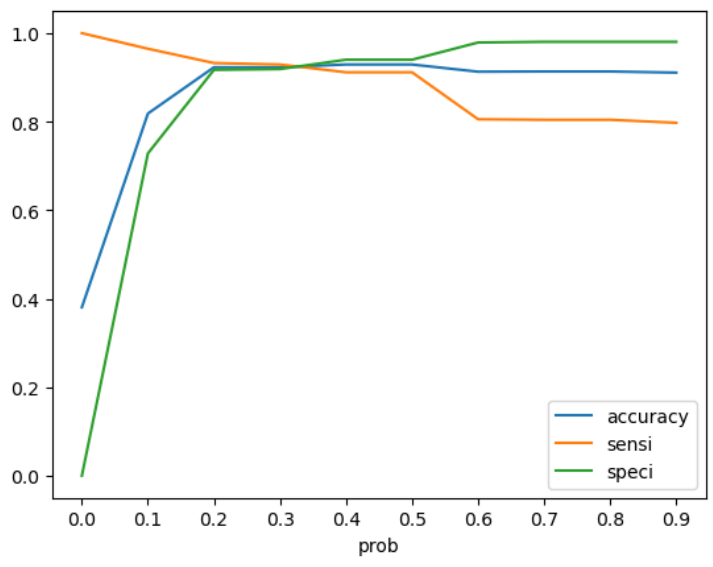
- The receiver operating characteristic (ROC) curve is another common too l used with binary classifiers. The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).



- The ROC Curve should be a value close to 1. We are getting a good value of 0.9 7 indicating a good predictive model.

Model Evaluation-Precision and Recall Tradeoff





TRAIN SET

Sensitivity: 92.91
Specificity: 91.89
False Positive Rate: 0.08
Positive Predictive Value: 0.88
Negative predictive value: 0.95

Precision: 87.55
Recall: 92.91

TEST SET

Sensitivity: 94.55
Specificity: 92.42
False Positive Rate: 0.08
Positive Predictive Value: 0.88
Negative predictive value: 0.97

Precision: 88.26
Recall: 94.55

Confusion Matrix

TRAIN SET

3567	315
169	2216

TEST SET

1549	127
55	955

04 Recommendations

Should Make Calls (Leads Likely to Convert):

- Call leads tagged as:
  - "Closed by Horizzon"
  - "Lost to EINS"
  - "Will revert after reading the email"
- Call leads from the "Welingak Website" lead source.
- Call leads with Last Notable Activity: "SMS Sent".
- Call leads with Last Activity: "Email Opened".

Should Not Make Calls (Leads Unlikely to Convert):

- Avoid calling leads with:
  - "Not Provided" current occupation.
  - Tags: "Ringing", "Other_Tags", or "Interested in other courses".


THANK YOU