LEADING SCORE CASE STUDY

Improving Lead Conversion through Predictive Modeling

A Report by NGUYEN VAN THUONG, Post Graduate Program in Data Science DS 68

November 2024

The Challenge of Low Conversion Rates at X Education company

- **Current situation:**
  - High lead generation, but low conversion rate (30%)
  - Need to identify "hot leads" to improve efficiency and reach 80% conversion goal.

Using:

```
Versions of Key Libraries and Modules Used in This Report:
Numpy version: 1.26.4
Pandas version: 2.1.4
Matplotlib version: 3.9.2
Seaborn version: 0.13.1
Scipy version: 1.14.1
Tabulate version: 0.9.0
statsmodels.api version: 0.14.2
Scikit-learn version: 1.5.1
```

Data:

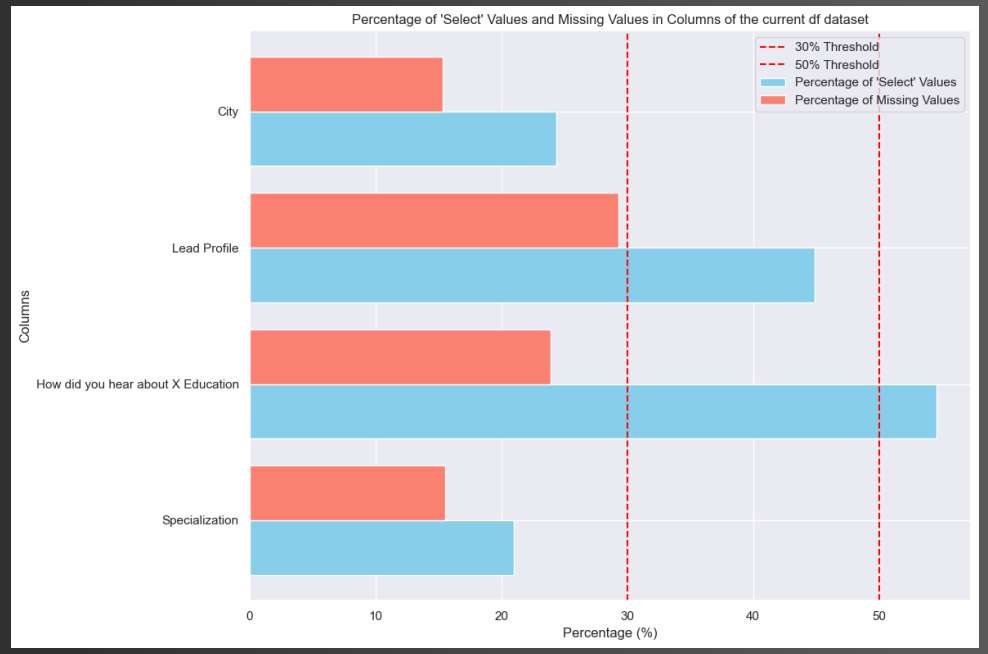
Leads Data Dictionary.csv	4 KB
Leads.csv	2,314 KB

DATA OVERVIEW

- **The preliminary dataset (from 'Leads.csv'):**
  - 9240 rows and 37 columns
  - **several categorical variables that necessitate the creation of dummy variables for analysis**
  - **significant number of NA values and 'Select' present, require appropriate handling to ensure data integrity**
  - **Conversion rate: 38.54%**

```python
# Print the shape of df dataset
print(f'The shape of the dataset is: {df.shape}')
row_initial = df.shape[0] # Initial number of rows in the dataset
```

```
The shape of the dataset is: (9240, 37)
```




```
Missing Values in the current df dataset:
```

Column Name	Missing Values	Percentage Missing (%)	Data Type
Lead Quality	4767	51.59	Categorical
Asymmetrique Activity Index	4218	45.65	Categorical
Asymmetrique Profile Score	4218	45.65	Numerical
Asymmetrique Activity Score	4218	45.65	Numerical
Asymmetrique Profile Index	4218	45.65	Categorical
Tags	3353	36.29	Categorical
Lead Profile	2709	29.32	Categorical
What matters most to you in choosing a course	2709	29.32	Categorical
What is your current occupation	2690	29.11	Categorical
Country	2461	26.63	Categorical
How did you hear about X Education	2207	23.89	Categorical
Specialization	1438	15.56	Categorical
City	1420	15.37	Categorical
Page Views Per Visit	137	1.48	Numerical
TotalVisits	137	1.48	Numerical
Last Activity	103	1.11	Categorical
Lead Source	36	0.39	Categorical
Receive More Updates About Our Courses	0	0.00	Categorical
I agree to pay the amount through cheque	0	0.00	Categorical
Get updates on DM Content	0	0.00	Categorical
Update me on Supply Chain Content	0	0.00	Categorical
A free copy of Mastering The Interview	0	0.00	Categorical
Prospect ID	0	0.00	Categorical
Newspaper Article	0	0.00	Categorical
Through Recommendations	0	0.00	Categorical
Digital Advertisement	0	0.00	Categorical
Newspaper	0	0.00	Categorical
X Education Forums	0	0.00	Categorical
Lead Number	0	0.00	Numerical
Magazine	0	0.00	Categorical
Search	0	0.00	Categorical
Total Time Spent on Website	0	0.00	Numerical
Converted	0	0.00	Numerical
Do Not Call	0	0.00	Categorical
Do Not Email	0	0.00	Categorical
Lead Origin	0	0.00	Categorical
Last Notable Activity	0	0.00	Categorical

```
Number of numerical variables with missing data: 4
Number of categorical variables with missing data: 13
```

DATA CLEANING

- **Irrelevant and incomplete data were removed, resulting in a final dataset with 68.97% of the original rows**
- **An increased Conversion rate from 38.54% to 48.09%.**

```
# Drop above 7 columns from the current df dataset
columns_to_drop = ['Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', '
df.drop(columns_to_drop, axis=1, inplace=True)
```

```
# Drop rows with NA in the above 7 specified columns
columns_to_dropNA = ["Lead Source", "TotalVisits", "Page Views Per Visit", "Last Activity", "Specialization", "What is your cu
df.dropna(subset=columns_to_dropNA, inplace=True)
print(f'The shape of the df dataset now is: {df.shape}')
row_after_data_cleaning = df.shape[0] # Number of rows after data cleaning
```

```
The shape of the df dataset now is: (6373, 12)
```

```
# Print the percentage of remaining rows after data cleaning
print(f'The percentage of remaining rows of df dataset ater data cleaning: {row_after_data_cleaning*100/row_initial:.2f}%')
```

```
The percentage of remaining rows of df dataset ater data cleaning: 68.97%
```

```
# Calculate the Conversion rate (%) based on the 'Converted' column of the current df dataset- after Data Cleaning
conversion_rate_data_cleaning_post = (df['Converted'].sum() / len(df)) * 100
print(f"Conversion Rate (%) after Data Cleaning step: {round(conversion_rate_data_cleaning_post, 2)}%")
```

```
Conversion Rate (%) after Data Cleaning step: 48.09%
```

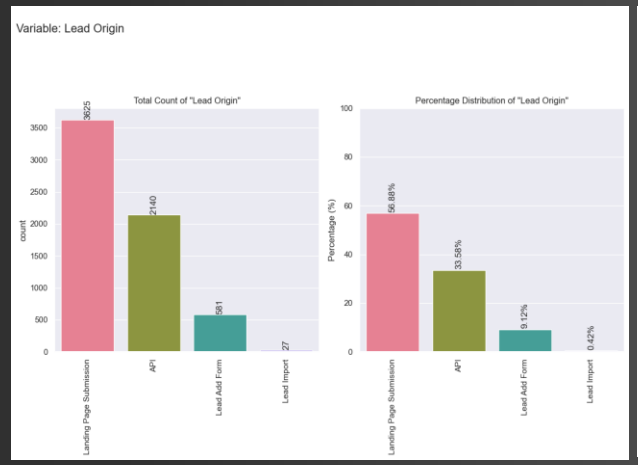
Planed Recommendations for X Education company

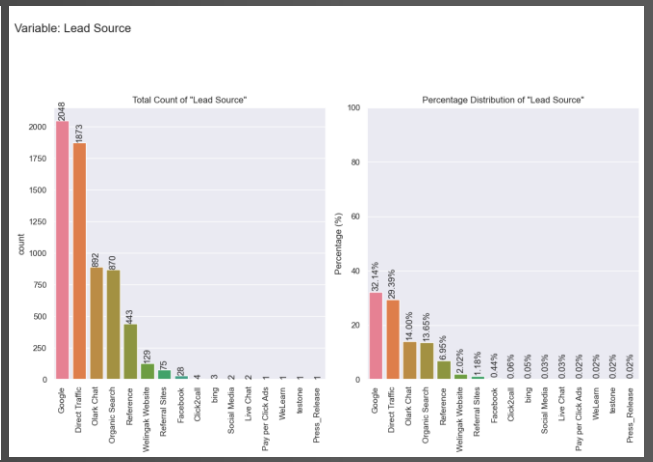
In light of these observations, we recommend that X Education revise their data collection methods and strategies for future endeavors. This revision should focus on the following aspects:

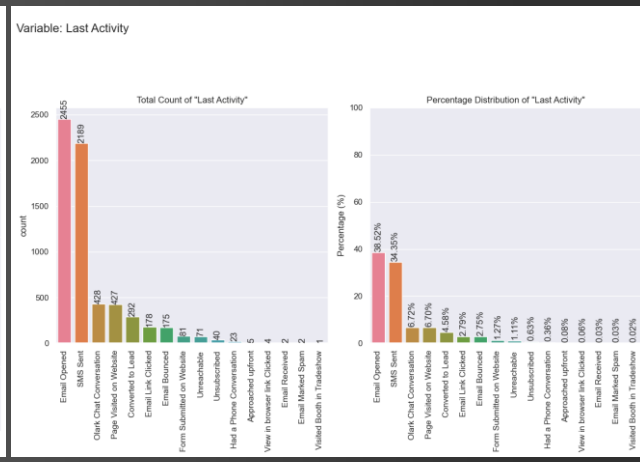
- **Enhanced Data Collection Techniques**: Implement more robust data collection methodologies that minimize the introduction of irrelevant or erroneous entries from the outset.
- **Regular Training for Data Collectors**: Provide ongoing training for personnel involved in data collection to ensure they are aware of best practices and the importance of data quality.
- **Establish Clear Data Quality Metrics**: Develop and utilize specific metrics to evaluate data quality continuously, allowing for timely adjustments in data collection processes.
- **Feedback Loops**: Create mechanisms for feedback on data collection processes to identify areas for improvement and adapt strategies accordingly.

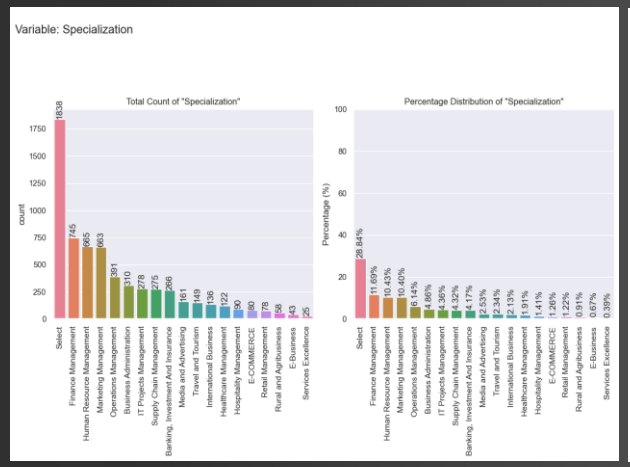
By adopting these recommendations, X Education can enhance the integrity and utility of their datasets, leading to more reliable analyses and informed decision-making in the future.

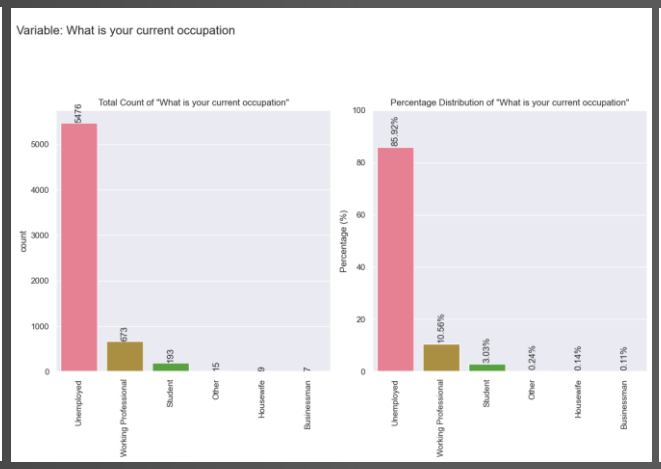
EDA: UNIVARIATE ANALYSIS – CATEGORICAL VARIABLES

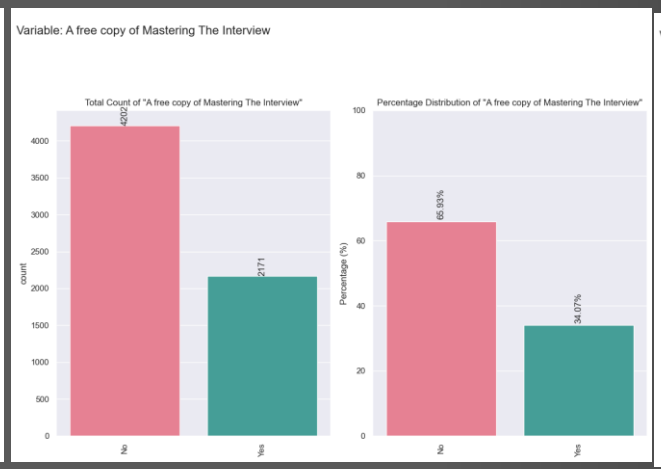


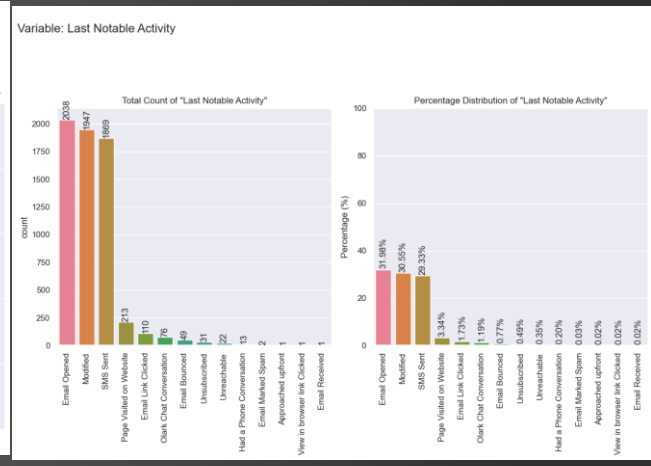












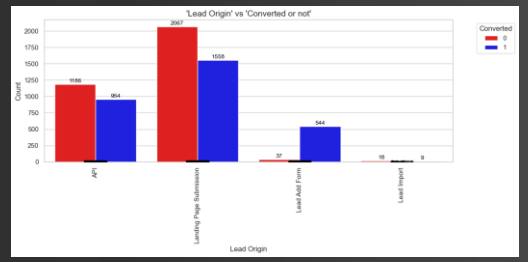
EDA: UNIVARIATE ANALYSIS – NUMERICAL VARIABLES

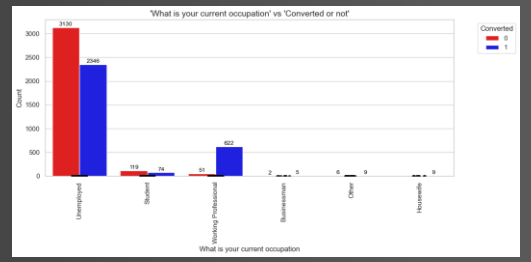

```
No outliers detected in column Converted.
Number of outliers in column TotalVisits: 205
Outlier values: [ 13.  17.  12.  13.  12.  13.  14.  17.  21.  15.  13.  22.  13.  13.
  21.  13.  14.  13.  16.  13.  18.  12.  14.  15.  20.  43.  18.  13.
  16.  12.  14.  13.  13.  14.  14.  22.  30.  16.  15.  13.  14.  13.
  13.  16.  23.  12.  55.  18.  21.  12.  25.  20.  13.  14.  18.  14.
  27.  17.  16.  15.  12.  15.  29.  16.  12.  16.  23.  12.  13.  24.
  18.  18.  14.  16.  17.  13.  14.  14.  18.  14.  13.  14.  16.  19.
  12.  12.  20.  14.  13.  13.  15.  14.  16.  18.  12.  12.  19.  12.
  12.  12.  13.  26.  14.  21.  14.  12.  16.  74.  12.  13.  24.  14.
  23.  19.  12.  12.  14.  13.  16.  19.  14.  18. 115.  18.  21.  22.
  15.  16.  15.  14.  20.  14.  17.  25. 251.  24.  14.  12.  12.  32.
  13.  13.  15.  13.  18.  26.  12.  12.  16.  15.  17.  13.  17.  20.
  13.  12.  14.  20.  16.  28.  20.  17.  20.  13.  15.  15.  27.  27.
  23.  12.  20.  12.  13.  13.  14.  17.  12.  27.  19.  13.  12.  29.
  13.  13.  14.  13.  19.  42.  16.  17.  17.  15.  20.  18.  27.  15.
  12.  13.  23.  12.  25.  17.  14.  13.  13.]
No outliers detected in column Total Time Spent on Website.
Number of outliers in column Page Views Per Visit: 153
Outlier values: [ 8.   11.   13.    8.    8.5   8.   14.   11.    8.   11.   10.   15.
  9.   11.    8.    9.    9.    9.   13.    8.    8.    9.    8.    8.
 10.   11.    8.   10.    9.    9.    9.    9.   11.   10.    8.   11.
  8.    8.   14.    8.    8.   10.   10.    9.    9.   11.    8.    8.
 55.    8.    9.    8.    8.   10.    8.    8.   10.    8.    8.   12.
  9.    8.    8.    8.    8.   15.    8.    8.    8.    8.    8.    8.
  8.   11.   13.   10.    9.    8.   12.33  8.    8.   12.   16.   14.
  9.    8.21  9.   11.    8.    7.5  10.   15.    8.   10.    8.   11.
  9.    8.    8.33  9.    9.    8.    9.    9.    9.   14.   11.    8.
  8.    8.    9.   13.    9.   11.    9.   13.    8.    8.    8.    9.
  9.   11.   10.    9.   10.    8.    9.   10.   10.    9.    8.   14.5
  8.    9.   10.    8.    9.    8.   10.    8.    9.    8.    8.    8.
 15.    8.   11.5   9.    9.   10.    9.    8.    9.  ]
```

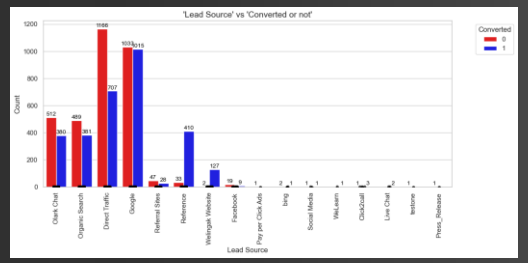
Outlier Handling Decision:

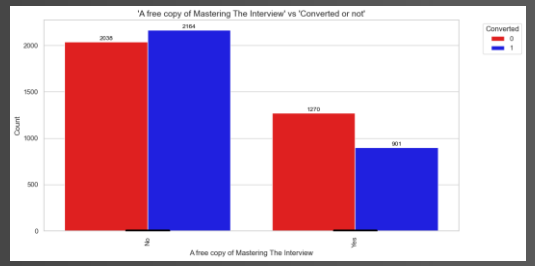
1. For variables like `TotalVisits` and `Page Views Per Visit`, consider investigating the nature of the outliers:
   - If these represent genuine customer behavior (e.g., highly engaged users), retain them for model training.
   - If they are due to data entry errors or bot activity, consider removing them before model training to improve model performance.
   - For now, we will consult with X Education to understand the cause of these outliers. ***Since there is insufficient information at this point, we will not remove the outliers in this analysis and report***.
2. For `Total Time Spent on Website`, no action is needed since no significant outliers were detected.

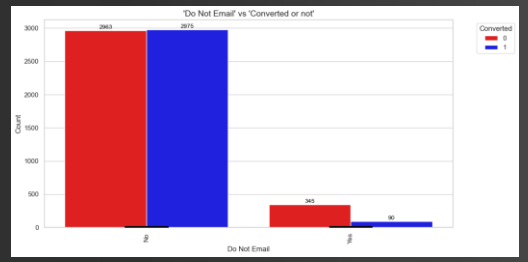
EDA: BIVARIATE ANALYSIS – CATEGORICAL VARIABLES- RELATED TO THE TARGET 'CONVERTED' VARIABLE

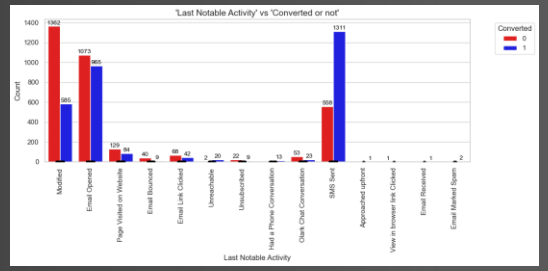


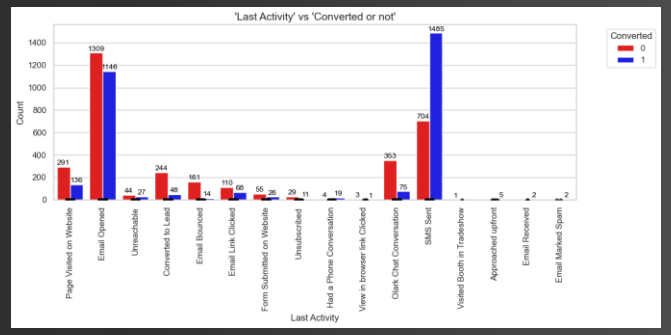


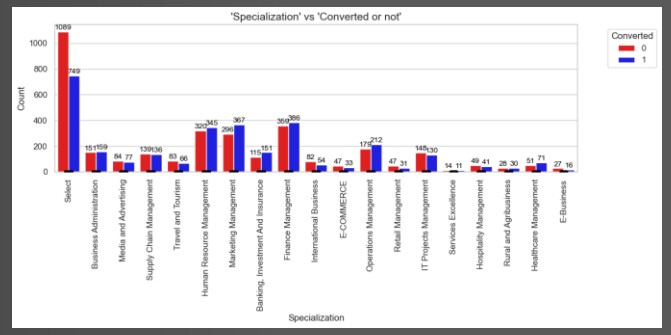












1. **Lead Origin vs. Converted-or-not**

- **Distribution:**
  - Landing Page Submission: Higher conversion rate compared to other origins.
  - API: Moderate conversion rate.
  - Lead Add Form and Lead Import: Lower conversion rates.
- **Insight:** Focus on optimizing landing page submissions as they show higher conversion rates.

2. **Lead Source vs. Converted-or-not**

- **Distribution:**
  - Google and Direct Traffic: Higher conversion rates.
  - Other sources: Lower conversion rates.
- **Insight:** Continue leveraging Google and Direct Traffic for lead generation.

3. **Do Not Email vs. Converted-or-not**

- **Distribution:**
  - No: Higher conversion rate.
  - Yes: Lower conversion rate.
- **Insight:** Email communication is effective; consider alternative strategies for those who opt out.

4. **Last Activity vs. Converted-or-not**

- **Distribution:**
  - Email Opened and SMS Sent: Higher conversion rates.
  - Other activities: Lower conversion rates.
- **Insight:** Email and SMS campaigns are effective; focus on these channels.

5. **Specialization vs. Converted-or-not**

- **Distribution:**
  - Finance Management, Human Resource Management, Marketing Management: Higher conversion rates.
  - Other specializations: Lower conversion rates.
- **Insight:** Tailor marketing efforts towards popular specializations.

6. **What is Your Current Occupation vs. Converted-or-not**

- **Distribution:**
  - Working Professional and Student: Higher conversion rates.
  - Unemployed: Lower conversion rate.
- **Insight:** Target working professionals and students for higher conversion potential.

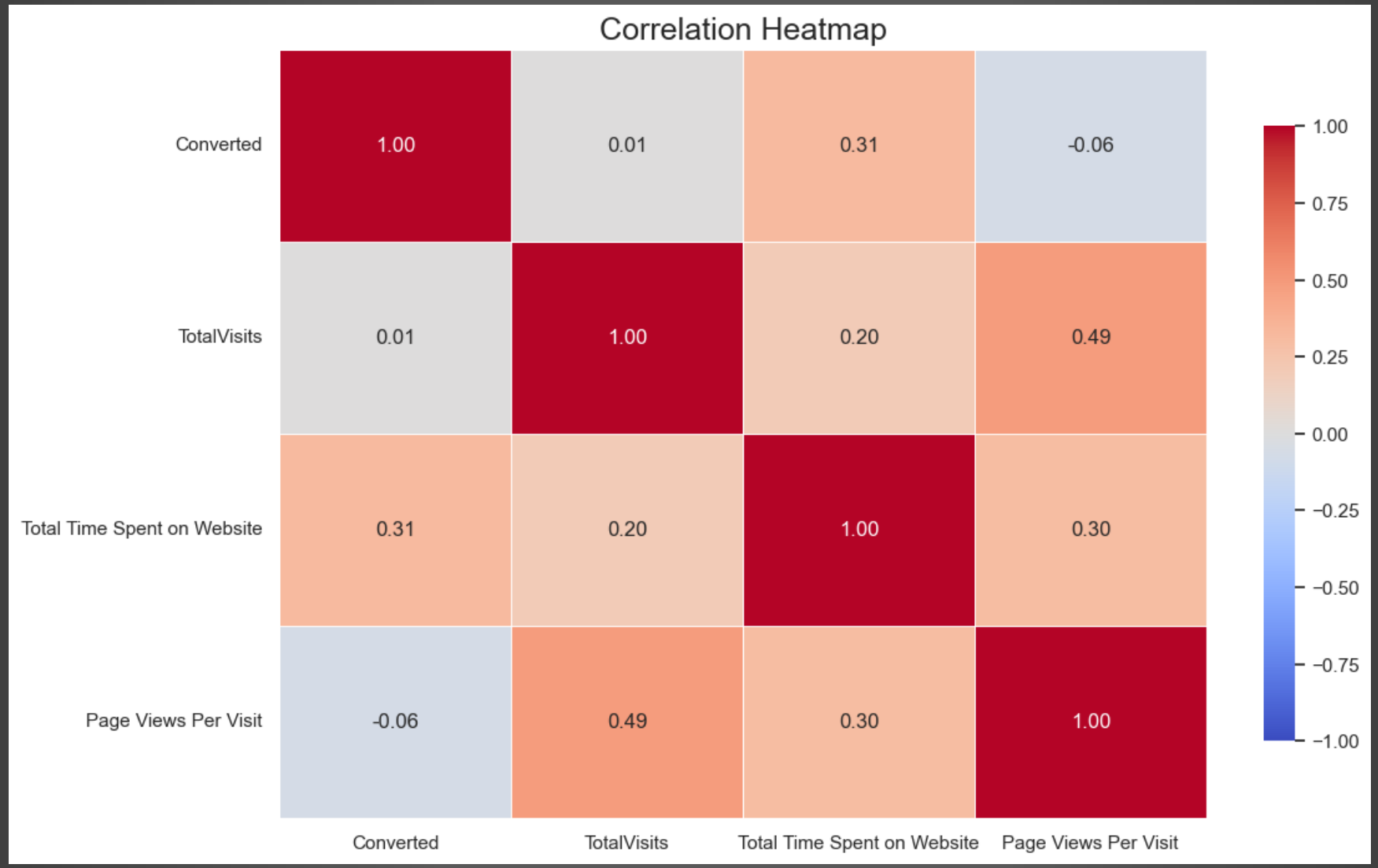
7. **A Free Copy of Mastering The Interview vs. Converted-or-not**

- **Distribution:**
  - Yes: Higher conversion rate.
  - No: Lower conversion rate.
- **Insight:** Offering a free copy of "Mastering The Interview" is an effective incentive.

8. **Last Notable Activity vs. Converted-or-not**

- **Distribution:**
  - Email Opened and SMS Sent: Higher conversion rates.
  - Other activities: Lower conversion rates.
- **Insight:** Focus on email and SMS follow-ups to increase conversions.

EDA: BIVARIATE ANALYSIS – NUMERICAL VARIABLES- RELATED TO THE TARGET 'CONVERTED' VARIABLE



- The **Total Time Spent on Website** shows a ***moderate positive correlation*** with the conversion rate, suggesting that ***it is an important factor in predicting lead conversion***.
- Both **TotalVisits** and **Page Views Per Visit** have ***very weak correlations*** with the conversion rate, indicating that these variables alone ***may not be strong predictors of conversion or not***. However, they could still provide valuable information when combined with other variables in a predictive model.

FEATURE ENGINEERING-CREATING DUMMY VARIABLES AND ENCODING FOR BINARY VARIABLES

```python
# Print the shape of the current df dataset
print(f'The shape of the current df dataset is: {df.shape}')
```

```
The shape of the current df dataset is: (6373, 75)
```

```python
# Encode for Binary Categorical columns
def binary_map(x):
    return x.map({'Yes': 1, "No": 0})
df[binary_categorical_columns] = df[binary_categorical_columns].apply(binary_map)
# Display the first few rows of the current df dataset
df.head()
```

	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview	Specialization_Banking, Investment And Insurance	Specialization_Business Administration	Specialization_E-Business	Specialization_E-COMMERCE	Specialization_Finance Management
0	0	0	0.0	0	0.0	0	0	0	0	0	0
1	0	0	5.0	674	2.5	0	0	0	0	0	0
2	0	1	2.0	1532	2.0	1	0	1	0	0	0
3	0	0	1.0	305	1.0	0	0	0	0	0	0
4	0	1	2.0	1428	1.0	0	0	0	0	0	0

```python
# Print the shape of the current df dataset
print(f'The shape of the current df dataset is: {df.shape}')
```

```
The shape of the current df dataset is: (6373, 75)
```

```python
# Calculate the conversion rate (%) based on the 'Converted' column- after Dummy Variable Creation step
conversion_rate_dum_post = (df['Converted'].sum() / len(df)) * 100
print(f"Conversion Rate (%) after Dummy Variable Creation: {round(conversion_rate_dum_post, 2)}%")
```

```
Conversion Rate (%) after Dummy Variable Creation: 48.09%
```

Categorical features were converted into dummy variables to optimize model performance, resulting in 74 features

PREPARING DATA FOR MODELLING

- The dataset was split into training (70%) and testing (30%) subsets.
- Robust scaling was applied to mitigate the impact of potential outliers in numerical variables.

```python
# Split the data into training and testing sets (70/30 split)
np.random.seed(42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
print('Shape of the X_train: ', X_train.shape)
print('Shape of the X_test: ', X_test.shape)
print('Shape of the y_train: ', y_train.shape)
print('Shape of the y_test: ', y_test.shape)
```

```
Shape of the X_train:  (4461, 74)
Shape of the X_test:  (1912, 74)
Shape of the y_train:  (4461,)
Shape of the y_test:  (1912,)
```

Training Set:

- **Total Samples:** 4461
- **Average "Do Not Email":** 6.75%
- **Average "Total Visits":** 3.57
- **Average "Total Time Spent on Website":** 536.57 seconds
- **Average "Page Views Per Visit":** 2.47
- **Conversion Rate (y_train):** 47.79%

Testing Set:

- **Total Samples:** 1912
- **Average "Do Not Email":** 7.01%
- **Average "Total Visits":** 3.74
- **Average "Total Time Spent on Website":** 532.08 seconds
- **Average "Page Views Per Visit":** 2.51
- **Conversion Rate (y_test):** 48.80%

The distribution of the target variable (`Converted`) is ***balanced across both sets***. Most features exhibit similar statistical characteristics; however, `TotalVisits` and `Page Views Per Visit` show ***higher maximum values in the testing set***, indicating potential outliers. As discussed in `Note 7`, we will consult with X Education to understand the cause of these outliers. ***Due to insufficient information, we would not remove them from this analysis***.

This discrepancy in outliers may influence the performance of the logistic regression model. Fortunately, as analyzed in `Note 9`, ***these two features have very weak correlations with the conversion rate, indicating they may not be strong predictors of conversion***.

```python
# Initialize the the RobustScaler
scaler = RobustScaler()
# Apply scaling to the Scaled Features in the Training dataset
X_train[scale_vars] = scaler.fit_transform(X_train[scale_vars])
# Apply scaling to the Scaled Features in the Testing dataset
X_test[scale_vars] = scaler.transform(X_test[scale_vars])
```

Sorted Features by Top Absolute Correlation Values with respect to the 'Converted' target variable in the Training set:

	Features	Correlation Values
17	Converted	1.000000
14	What is your current occupation_Working Professional	0.312891
1	Total Time Spent on Website	0.310994
12	Last Activity_SMS Sent	0.285331
16	Last Notable Activity_SMS Sent	0.284160
4	Lead Origin_Lead Add Form	0.281969
13	What is your current occupation_Unemployed	-0.267919
6	Lead Source_Reference	0.235275
15	Last Notable Activity_Modified	-0.232125
10	Last Activity_Olark Chat Conversation	-0.169113
0	Do Not Email	-0.150019
7	Lead Source_Welingak Website	0.147957
8	Last Activity_Converted to Lead	-0.140002
9	Last Activity_Email Bounced	-0.137563
5	Lead Source_Direct Traffic	-0.127006
3	Lead Origin_Landing Page Submission	-0.104778
2	A free copy of Mastering The Interview	-0.088701
11	Last Activity_Page Visited on Website	-0.078552

==========

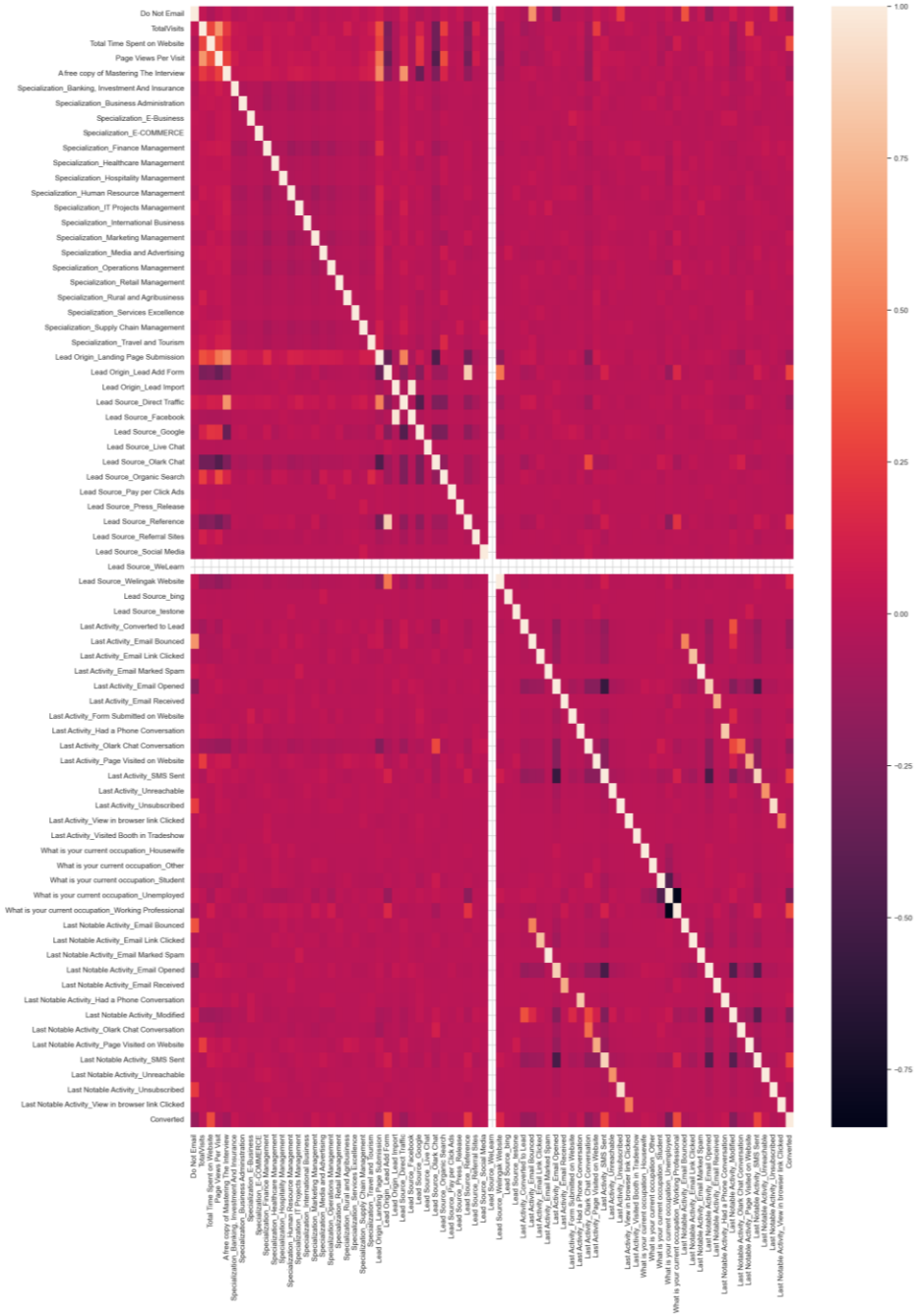
Top Positive Correlations with respect to the 'Converted' target variable in the Training set:

	Features	Correlation Values
17	Converted	1.000000
14	What is your current occupation_Working Professional	0.312891
1	Total Time Spent on Website	0.310994
12	Last Activity_SMS Sent	0.285331
16	Last Notable Activity_SMS Sent	0.284160
4	Lead Origin_Lead Add Form	0.281969
6	Lead Source_Reference	0.235275
7	Lead Source_Welingak Website	0.147957

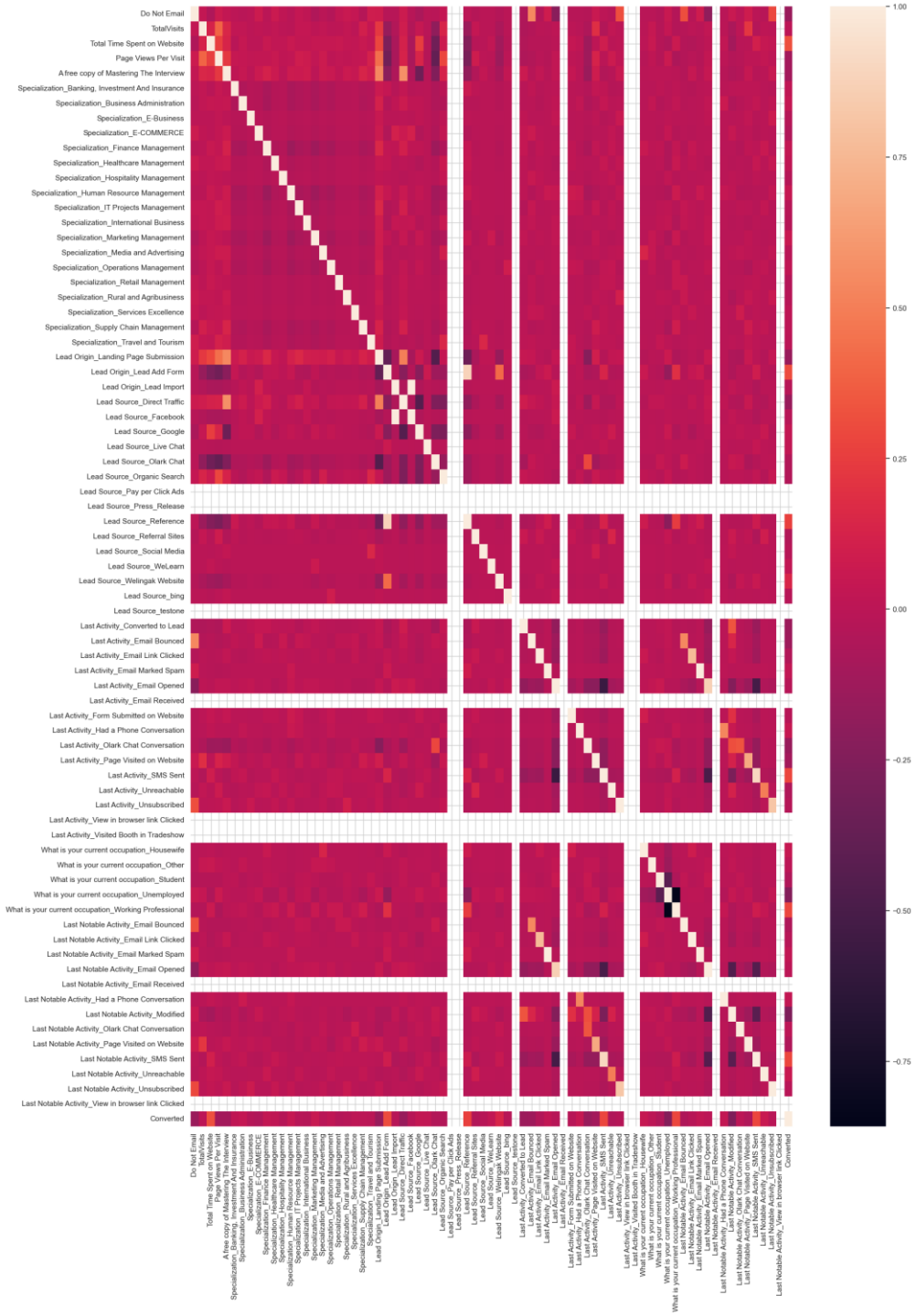
==========

Top Negative Correlations with respect to the 'Converted' target variable in the Training set:

	Features	Correlation Values
13	What is your current occupation_Unemployed	-0.267919
15	Last Notable Activity_Modified	-0.232125
10	Last Activity_Olark Chat Conversation	-0.169113
0	Do Not Email	-0.150019
8	Last Activity_Converted to Lead	-0.140002
9	Last Activity_Email Bounced	-0.137563
5	Lead Source_Direct Traffic	-0.127006
3	Lead Origin_Landing Page Submission	-0.104778
2	A free copy of Mastering The Interview	-0.088701
11	Last Activity_Page Visited on Website	-0.078552



Training set



```
Sorted Features by Top Absolute Correlation Values with respect to the 'Converted' target variable in the Testing set:
                                          Features  Correlation Values
18                                        Converted            1.000000
1                           Total Time Spent on Website            0.319011
5                             Lead Origin_Lead Add Form            0.303726
13                               Last Activity_SMS Sent            0.287580
15  What is your current occupation_Working Professional            0.286231
17                        Last Notable Activity_SMS Sent            0.284736
7                                Lead Source_Reference            0.260622
16                        Last Notable Activity_Modified           -0.256947
14            What is your current occupation_Unemployed           -0.240625
11                 Last Activity_Olark Chat Conversation           -0.151673
6                             Lead Source_Direct Traffic           -0.149236
4                   Lead Origin_Landing Page Submission           -0.147689
0                                           Do Not Email           -0.145042
8                         Lead Source_Welingak Website            0.137830
9                       Last Activity_Converted to Lead           -0.136233
10                        Last Activity_Email Bounced           -0.127945
3                 A free copy of Mastering The Interview           -0.109040
12                 Last Activity_Page Visited on Website           -0.107628
2                                   Page Views Per Visit           -0.104929
==========================================================================================

Top Positive Correlations with respect to the 'Converted' target variable in the Testing set:
                                          Features  Correlation Values
18                                        Converted            1.000000
1                           Total Time Spent on Website            0.319011
5                             Lead Origin_Lead Add Form            0.303726
13                               Last Activity_SMS Sent            0.287580
15  What is your current occupation_Working Professional            0.286231
17                        Last Notable Activity_SMS Sent            0.284736
7                                Lead Source_Reference            0.260622
8                         Lead Source_Welingak Website            0.137830
==========================================================================================

Top Negative Correlations with respect to the 'Converted' target variable in the Testing set:
                                        Features  Correlation Values
16                 Last Notable Activity_Modified           -0.256947
14  What is your current occupation_Unemployed           -0.240625
11       Last Activity_Olark Chat Conversation           -0.151673
6                   Lead Source_Direct Traffic           -0.149236
4          Lead Origin_Landing Page Submission           -0.147689
0                                   Do Not Email           -0.145042
9                 Last Activity_Converted to Lead           -0.136233
10                    Last Activity_Email Bounced           -0.127945
3       A free copy of Mastering The Interview           -0.109040
12       Last Activity_Page Visited on Website           -0.107628
2                           Page Views Per Visit           -0.104929
```

Testing set

Note 12: Common Trends and Differences in Top Correlation Anaysis between Features and 'Converted'in Training and Testing Sets

Positive Correlations:

- Both datasets reveal that features such as:
  - **Total Time Spent on Website**
  - **Lead Origin_Lead Add Form**
  - **Last Activity_SMS Sent**
  - **What is your current occupation_Working Professional**
  - **Lead Source_Reference**

  exhibit strong positive correlations with the target variable **Converted**. This indicates that these factors positively influence the likelihood of lead conversion.

Negative Correlations:

- Similarly, features like:
  - **Last Notable Activity_Modified**
  - **What is your current occupation_Unemployed**
  - **Last Activity_Olark Chat Conversation**
  - **Do Not Email**

  show negative correlations with **Converted**. This suggests that these factors negatively impact the chances of conversion.

Correlation Consistency:

- There are no features that demonstrate negative correlations in one dataset while exhibiting positive correlations in the other. This consistency reinforces the reliability of the findings across both datasets.

Differences Between the Two Dataset:

Order and Correlation Values:

- Although the most important features are similar, their order and correlation values differ between the two datasets. For instance, **Total Time Spent on Website** has a higher correlation value in the testing set compared to the training set.

Emergence of New Features:

- Certain features appear in the top features list of one dataset but not in the other. For example, **Page Views Per Visit** is identified as a top feature in the testing set but is absent from the training set.

MODEL BUILDING

- Recursive Feature Elimination (RFE) and stepwise elimination based on Variance Inflation Factor (VIF) and p-values were used to select the most relevant features.
- The final model included 12 features at the 4th step.

```
The Estimated Parameters (Coefficients) of the 4th Logistic fitted model (GLM):
 const                                                 0.073658
Do Not Email                                          -1.536722
Total Time Spent on Website                            1.907709
Lead Source_Olark Chat                                 1.424428
Lead Source_Reference                                  3.457119
Lead Source_Welingak Website                           5.455480
Last Activity_Converted to Lead                       -1.246210
Last Activity_Olark Chat Conversation                 -1.336028
Last Activity_SMS Sent                                 1.046418
What is your current occupation_Student               -1.427618
What is your current occupation_Unemployed            -1.505898
What is your current occupation_Working Professional   1.275306
Last Notable Activity_Unreachable                      2.552012
dtype: float64
***************************
The Summary of the 4th Logistic fitted model (GLM):
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 4461
Model:                            GLM   Df Residuals:                     4448
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2067.3
Date:                Mon, 25 Nov 2024   Deviance:                       4134.7
Time:                        12:06:56   Pearson chi2:                 4.50e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3671
Covariance Type:            nonrobust
==========================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                             0.0737      0.583      0.126      0.899      -1.069       1.216
Do Not Email                                     -1.5367      0.191     -8.028      0.000      -1.912      -1.162
Total Time Spent on Website                       1.9077      0.080     23.709      0.000       1.750       2.065
Lead Source_Olark Chat                            1.4244      0.118     12.112      0.000       1.194       1.655
Lead Source_Reference                             3.4571      0.229     15.109      0.000       3.009       3.906
Lead Source_Welingak Website                      5.4555      0.725      7.527      0.000       4.035       6.876
Last Activity_Converted to Lead                  -1.2462      0.236     -5.284      0.000      -1.708      -0.784
Last Activity_Olark Chat Conversation            -1.3360      0.184     -7.276      0.000      -1.696      -0.976
Last Activity_SMS Sent                            1.0464      0.083     12.536      0.000       0.883       1.210
What is your current occupation_Student          -1.4276      0.620     -2.304      0.021      -2.642      -0.213
What is your current occupation_Unemployed       -1.5059      0.584     -2.579      0.010      -2.650      -0.362
What is your current occupation_Working Professional  1.2753      0.615      2.072      0.038       0.069       2.481
Last Notable Activity_Unreachable                 2.5520      0.815      3.133      0.002       0.955       4.149
==========================================================================================================
***************************
Features with p-value > 0.05:
(const, 0.8995)
```

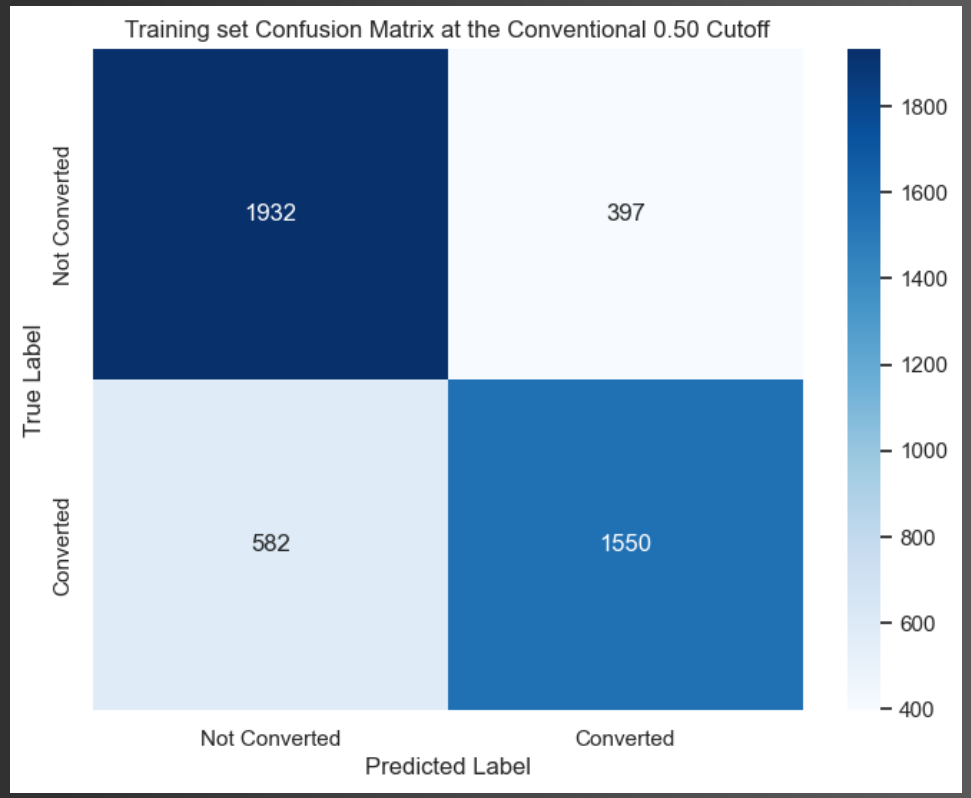
```
Shape of the Updated X_train in this step: (4461, 15)
We use 15 Features Using in this 1st Logistic fitted model (GLM), including:
 Index(['Do Not Email', 'Total Time Spent on Website',
        'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
        'Lead Source_Reference', 'Lead Source_Welingak Website',
        'Last Activity_Converted to Lead',
        'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent',
        'What is your current occupation_Student',
        'What is your current occupation_Unemployed',
        'What is your current occupation_Working Professional',
        'Last Notable Activity_Email Bounced',
        'Last Notable Activity_Had a Phone Conversation',
        'Last Notable Activity_Unreachable'],
       dtype='object')
```

	Features	VIF
9	What is your current occupation_Unemployed	2.49
7	Last Activity_SMS Sent	1.70
1	Total Time Spent on Website	1.51
2	Lead Source_Olark Chat	1.49
10	What is your current occupation_Working Professional	1.36
3	Lead Source_Reference	1.25
6	Last Activity_Olark Chat Conversation	1.24
5	Last Activity_Converted to Lead	1.11
0	Do Not Email	1.10
4	Lead Source_Welingak Website	1.07
8	What is your current occupation_Student	1.07
11	Last Notable Activity_Unreachable	1.01

The Effectively Evaluated Logistic Regression Model Equation is:
Logit(p) = Log(p/1-p)=5.455*Lead Source_Welingak Website + 3.457*Lead Source_Reference + 2.552*Last Notable Activity_Unreachable + 1.908*Total Time Spent on Website + 1.424*Lead Source_Olark Chat + 1.275*What is your current occupation_Working Professional + 1.046*Last Activity_SMS Sent + 0.074*const -1.537*Do Not Email -1.506*What is your current occupation_Unemployed -1.428*What is your current occupation_Student -1.336*Last Activity_Olark Chat Conversation -1.246*Last Activity_Converted to Lead
(with p as the Probability of Conversion)

INITIAL MODEL EVALUATION

- A conventional 0.5 cutoff was used for classification.
- The model achieved an accuracy of 78.05% on the training set but showed a high number of false negatives.



Overall Accuracy Score:

The overall accuracy score of the logistic regression model is **78.05%**. This indicates that approximately 78 out of every 100 predictions made by the model are correct. While this score suggests reasonably good performance, it is essential to delve deeper into the confusion matrix to understand the model's strengths and weaknesses.

Confusion Matrix:

The confusion matrix is as follows:

	Predicted: Not Converted	Predicted: Converted
Actual: Not Converted	1932	397
Actual: Converted	582	1550

Interpretation of the Confusion Matrix:

- **True Negatives (TN)**: 1932 instances were correctly predicted as "Not Converted."
- **False Positives (FP)**: 397 instances were incorrectly predicted as "Converted" when they were actually "Not Converted."
- **False Negatives (FN)**: 582 instances were incorrectly predicted as "Not Converted" when they were actually "Converted."
- **True Positives (TP)**: 1550 instances were correctly predicted as "Converted."

Training Set Metrics:

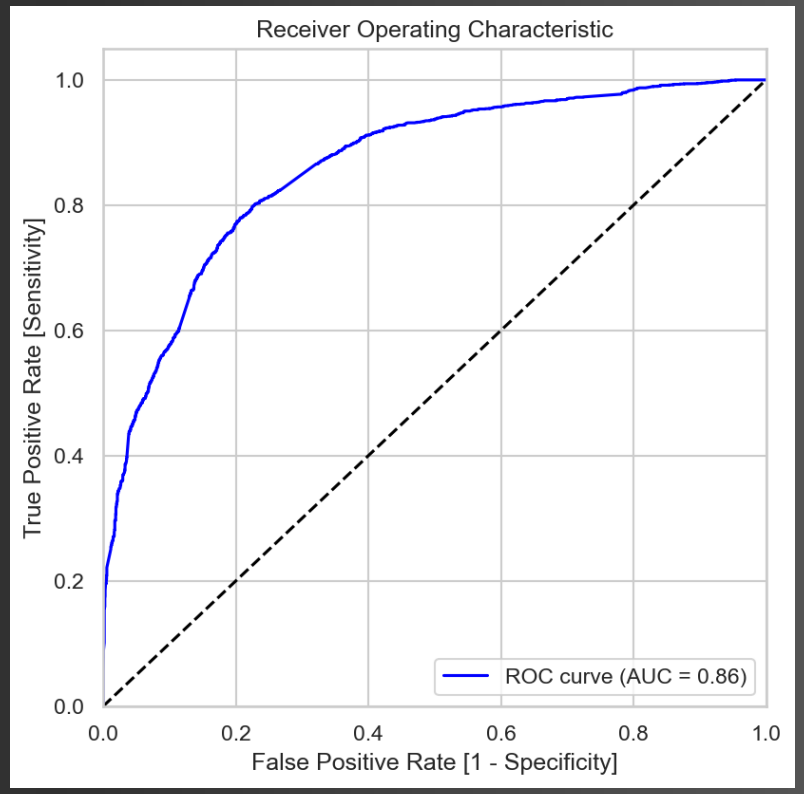
- **Sensitivity (True Positive Rate)** using the Conventional 0.50 Cutoff: **72.70%**
- **Specificity (True Negative Rate)** using the Conventional 0.50 Cutoff: **82.95%**
- **Precision (Positive Predictive Value)** using the Conventional 0.50 Cutoff: **79.61%**
- **Recall (Sensitivity)** using the Conventional 0.50 Cutoff: **72.70%**
- **F1 Score** using the Conventional 0.50 Cutoff:
  The F1 Score is calculated as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.7961 \times 0.7270}{0.7961 + 0.7270} = 0.7597 \quad (\approx 75.97\%)$$

Insights from the Confusion Matrix:

1. **Class Imbalance**: The number of True Negatives (1932) is significantly higher than False Positives (397), indicating that the model performs better at predicting the "Not Converted" class. This could suggest a class imbalance in the dataset, which may require addressing through techniques such as resampling or using class weights.
2. **Error Analysis**:
   - The model has a relatively high number of False Negatives (582), indicating that there are many actual positive cases (Converted) that the model fails to identify. This could lead to missed opportunities in scenarios where identifying positive cases is critical.
   - The False Positive rate is also notable, with 397 instances incorrectly classified as Converted. This may lead to unnecessary actions or costs associated with false conversions.

ROC CURVE OPTIMIZATION (1)

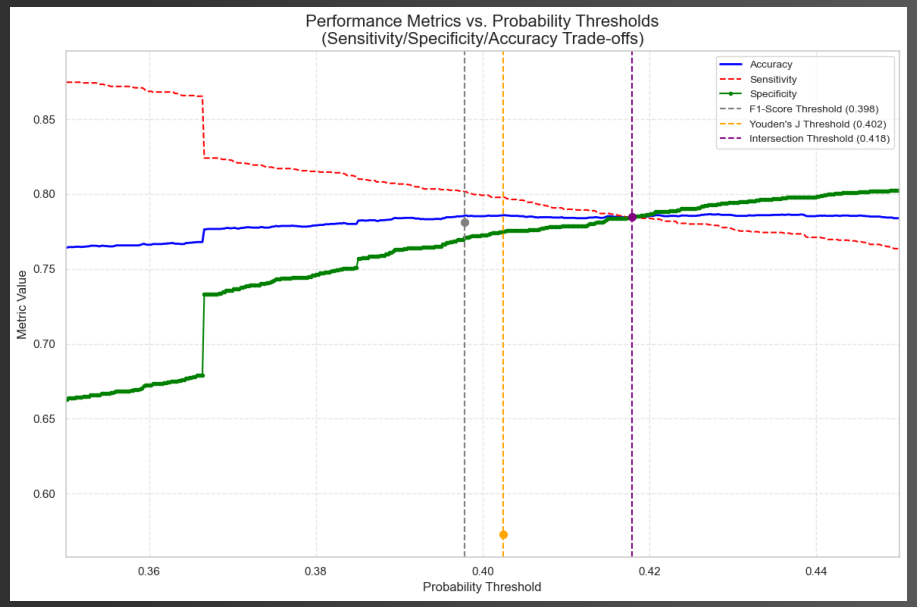


The ROC curve here indicates that the model possesses **good discriminatory power in the Training set**. Below is a detailed breakdown of the evaluation:

- **Shape**:
  - The curve arches significantly towards the upper left corner of the plot. This is a desirable characteristic, as it signifies that the model is effectively separating the positive and negative classes.
- **AUC (Area Under the Curve)**:
  - The AUC is **0.86**, which is a solid value (generally, an AUC above **0.7** is considered acceptable, while an AUC above **0.8** is regarded as ***good***). This suggests that the model has a high probability of ranking a randomly chosen positive instance higher than a randomly chosen negative instance.
- **Alignment with Metrics**:
  The ROC curve summarizes the trade-off between Sensitivity (True Positive Rate) and Specificity (1 - False Positive Rate) across various thresholds, providing a holistic view of the performance of the Last Fitted Model here.

Overall, the ROC curve and its corresponding AUC indicate that the Last Fitted Model demonstrates strong performance in distinguishing between classes, **making it a reliable choice for predictive tasks**.

ROC CURVE OPTIMIZATION (2)- DETERMINE FOR THE OPTIMIZED CUTOFF

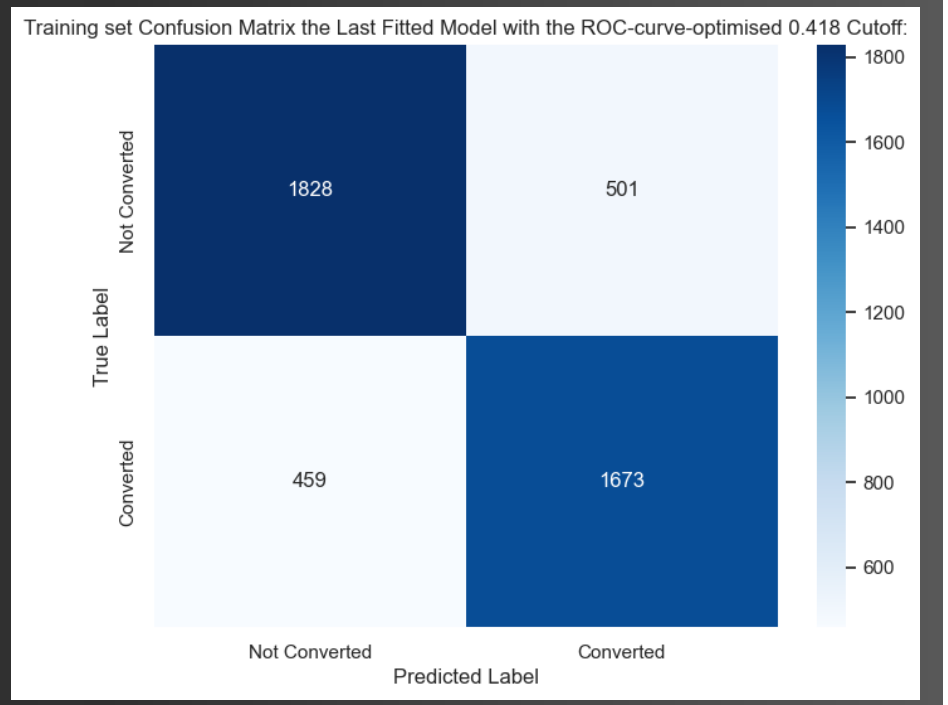


The plot above illustrates how key performance metrics—**Accuracy**, **Sensitivity**, and **Specificity**—vary across different probability cutoffs. The goal is to identify the optimal cutoff where these metrics are balanced.

1. **Optimal Cutoff Selection**:
   - The optimal cutoff based on the intersection of Accuracy, Sensitivity, and Specificity curves is **0.418**, as determined by minimizing the total distance between these metrics.
   - This cutoff provides a balance between correctly identifying positive cases (**Sensitivity**) and avoiding false positives (**Specificity**).
2. **F1-score and Youden's J Statistic**:
   - At a cutoff of **0.398** and **0.402**, respectively, the F1-score and Youden's J statistic reach their respective peaks:
     - The F1-score optimally balances Precision and Sensitivity.
     - Youden's J statistic maximizes the sum of Sensitivity and Specificity minus one ($J = TPR + TNR - 1$).
3. **Comparison of Methods**:
   - While both F1-score and Youden's J statistic suggest a cutoff of nearly **0.40**, using the intersection method ensures alignment with the requirement to balance Accuracy, Sensitivity, and Specificity at a cuttoff **0.418**, very close with the above cutoffs.
   - So, ***we select an optimal cutoff of 0.418, as it achieves the best balance between key performance metrics based on ROC curve analysis***

This cutoff will be used for further evaluation on both training and test datasets to validate model performance.

ROC CURVE OPTIMIZATION (3)- EVALUATION IN THE TRAINING SET

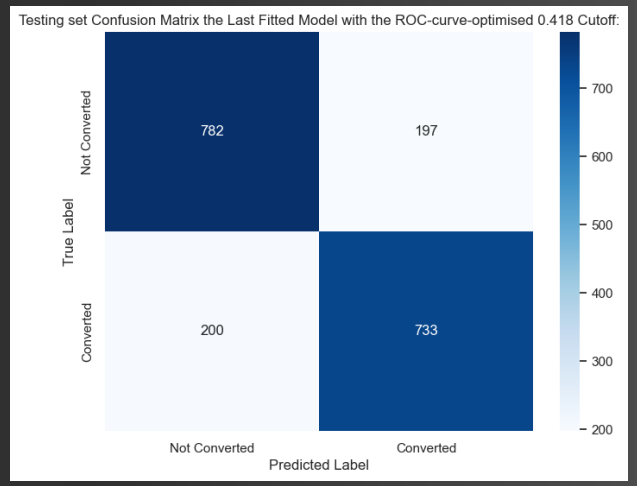


Metric	Conventional Cutoff (0.50)	ROC-curve-Optimized Cutoff (0.418)
Accuracy	78.05%	78.48%
Sensitivity	72.70%	78.47%
Specificity	82.95%	78.49%
Precision	79.61%	76.95%
F1 Score	75.97%	77.71%

1. **Improved Balance Between Sensitivity and Specificity**:
   - The optimized cutoff of **0.418** achieves a near-perfect balance between sensitivity and specificity in the Training set, with both metrics close to **78.5%**.
   - Sensitivity improved significantly (+5.77%) compared to the Conventional Cutoff (e.g., 0.50), reducing false negatives and identifying more positive cases correctly.
   - Specificity slightly decreased compared to the Conventional cutoff (-4.46%), resulting in a small increase in false positives, but this trade-off is acceptable given the improvement in sensitivity.
2. **Marginal Improvement in Overall Accuracy**:
   - The overall accuracy increased slightly to **78.48%**, reflecting a balanced improvement across both classes.
3. **Alignment with Business Objectives**:
   - This optimized cutoff is particularly suitable for scenarios prioritizing the identification of positive cases (high sensitivity) while maintaining reasonable performance for negative cases (specificity).
   - The trade-off ensures that fewer positive cases are missed, which is critical in applications where false negatives carry significant costs or risks.
4. **Importance of Precision and Recall**:
   - The precision of **76.95%** indicates that when a positive prediction is made, there is a high likelihood it is correct, which is essential for applications where the cost of false positives needs to be minimized.
   - Recall, being equivalent to sensitivity in this context, reinforces our focus on capturing as many true positive cases as possible.

This comprehensive evaluation highlights the effectiveness of using ROC curve analysis to optimize model performance on the training set while aligning with strategic business goals.

ROC CURVE OPTIMIZATION (4)- PREDICTION IN THE TESTING SET

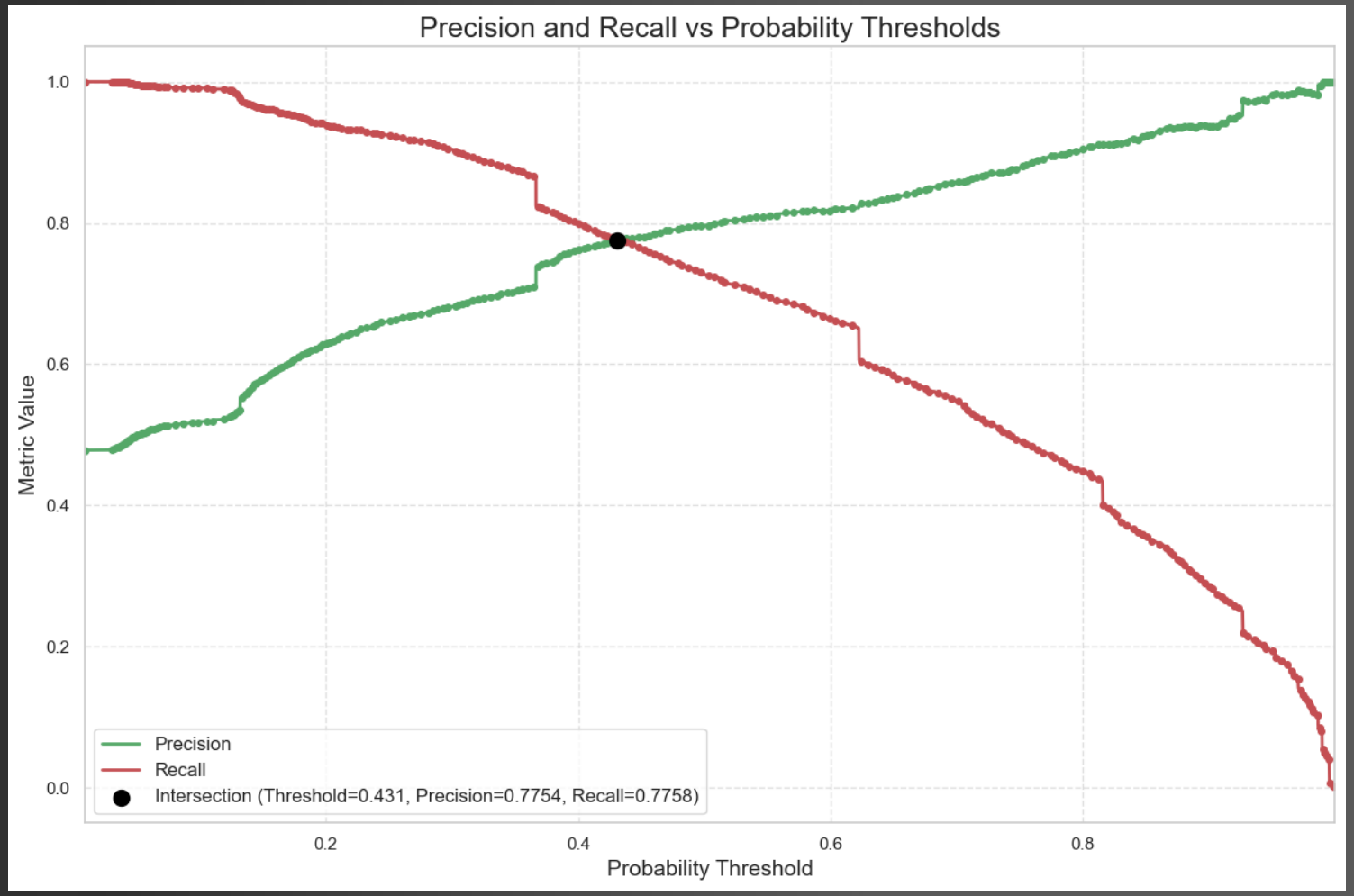


Insights

1. **Generalization Performance**:
   - The model demonstrates consistent performance across both training and testing sets, with metrics showing minimal variation.
   - The testing set accuracy of **79.24%** is slightly higher than the training set accuracy of **78.48%**, indicating good generalization to unseen data.
2. **Sensitivity and Specificity Trade-Offs**:
   - Sensitivity remains stable between the training (**78.47%**) and testing (**78.56%**) sets, ensuring that most positive cases are correctly identified.
   - Specificity improves on the testing set (**79.88%**) compared to the training set (**78.49%**), reflecting better control over false positives in unseen data.
3. **Balanced Performance with Optimized Cutoff**:
   - The ROC-curve-optimized cutoff of **0.418** achieves a strong balance between sensitivity and specificity, aligning well with scenarios where both false positives and false negatives need to be minimized.
4. **Model Robustness**:
   - The consistent metrics across datasets suggest that the model is neither overfitting nor underfitting, making it suitable for deployment in real-world applications.
5. **Visual Confirmation of Optimal Threshold**:
   - The intersection of Accuracy, Sensitivity, and Specificity curves at the cutoff value of **0.418**, as shown in the threshold optimization plot, validates its selection as an effective threshold.
6. **F1 Score Analysis**:
   - The F1 score on the testing set is calculated as approximately **78.68%**, indicating a balanced performance between precision and recall, which is crucial for applications where both false positives and false negatives carry significant costs.

Conclusion The logistic regression model with the ROC-curve-optimized cutoff of **0.418** ***performs reliably across both training and testing sets, achieving a balanced trade-off between sensitivity and specificity while maintaining high accuracy. This confirms its suitability for deployment in scenarios requiring robust classification with minimal false positives and false negatives***.

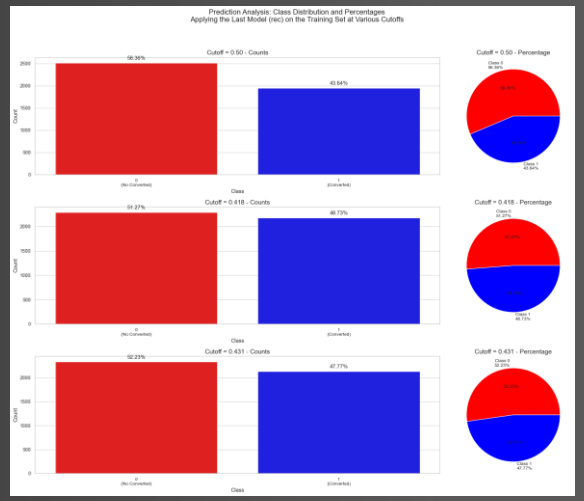
PRECISION-RECALL TRADEOFF OPTIMIZATION (1)

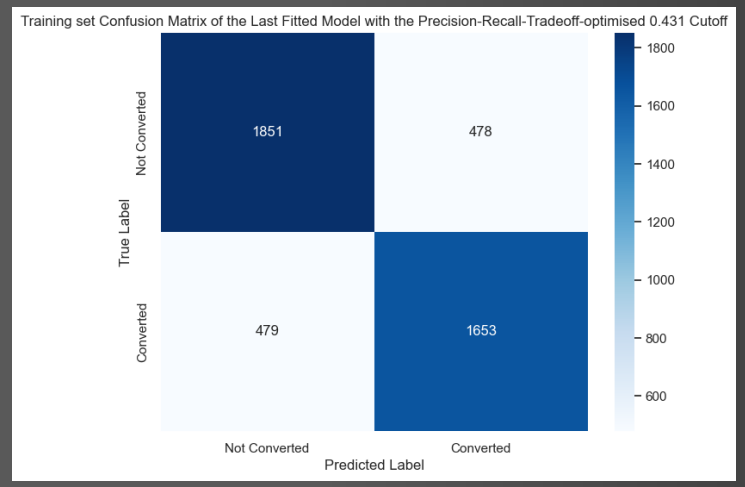


Key Results on Training Set:

1. **At the Conventional Cutoff (0.50):**
   - **Precision**: 79.61%
   - **Recall**: 72.70%
2. **At the Intersection Point (Precision = Recall):**
   - **Intersection Threshold**: **0.431**
   - **Precision at Intersection**: 77.54% (↓ **2.07** compared to the Cutoff 0.50)
   - **Recall at Intersection**: 77.58% (↑ **4.88%** compared to the Cutoff 0.50)

PRECISION-RECALL TRADEOFF OPTIMIZATION (2)-EVALUATION IN THE TRAINING SET





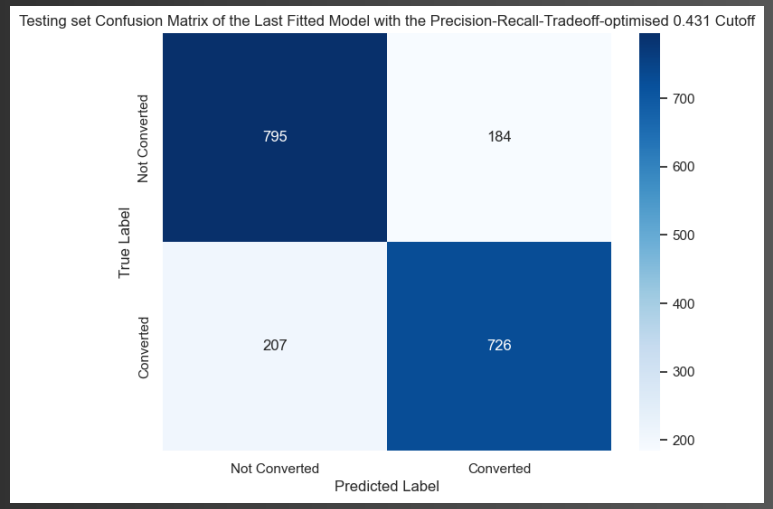
Key Metrics on Training Set:

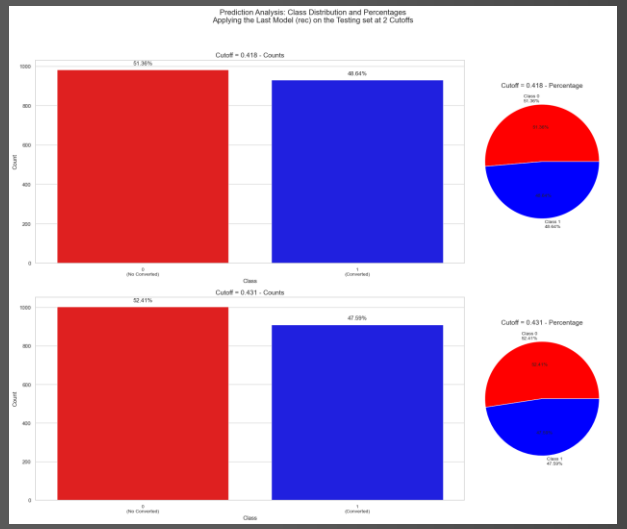
Metric	Conventional Cutoff (0.50)	Optimized Cutoff (0.431)
Overall Accuracy	78.05%	78.55%
Precision	79.61%	77.57%
Recall/ Sensitivity	72.70%	77.53%
Specificity	82.95%	79.48%

Insights:

1. **Accuracy Improvement**:
   - The overall accuracy increased from **78.05%** to **78.55%**, reflecting a slight improvement in model performance with the optimized cutoff.
2. **Precision vs Recall Tradeoff**:
   - While Precision decreased slightly (↓ **2.04%**), Recall improved significantly (↑ **4.83%**) at the optimized cutoff.
   - This tradeoff aligns with scenarios where ***identifying positive cases is more critical than minimizing false positives***.
3. **Confusion Matrix Analysis**:
   - At the optimized cutoff (**0.431**), ***false negatives decreased compared to the conventional cutoff, indicating better identification of positive cases***.
4. **Prediction Distribution Across Cutoffs**:
   - The proportion of predictions (`mean`) shifted slightly from **43.64% (cutoff = 0.50)** to **47.77% (cutoff = 0.431)**, reflecting a more balanced classification.

PRECISION-RECALL TRADEOFF OPTIMIZATION (3)-MAKING PREDICTION IN THE TESTING SET





Performance at Cutoff = 0.431 (Precision-Recall Tradeoff):

- On the ***Testing set***, the model achieves an `Accuracy` of **79.55%**, `Precision` of **79.78%**, `Recall/Sensitivity` of **77.81%**, and `Specificity` of **81.21%**.
- On the ***Training set***, the model achieves an `Accuracy` of **78.55%**, `Precision` of **77.57%**, `Recall/Sensitivity` of **77.53%**, and `Specificity` of **79.48%**.
- The ***Testing set performs slightly better than the Training set, indicating good generalization***.

Recommendations in Practice:

1. ***Use the ROC Curve optimized cutoff (0.418)*** if `minimizing false negatives is the primary objective`.
2. ***Use the Precision-Recall Tradeoff optimized cutoff (0.431)*** if `a balanced tradeoff between precision and recall is more critical (e.g., marketing campaigns or customer retention analysis)`.
3. ***Continuously monitor model performance in production to ensure that the selected cutoff aligns with changing business objectives.***

List of Features with the Descending Order of Coefficients in The Effectively Evaluated Logistic Regression Model Equation:

	Feature	Coefficient	P-value
5	Lead Source_Welingak Website	5.455	5.205e-14
4	Lead Source_Reference	3.457	1.404e-51
12	Last Notable Activity_Unreachable	2.552	1.732e-03
2	Total Time Spent on Website	1.908	2.889e-124
1	Do Not Email	-1.537	9.869e-16
10	What is your current occupation_Unemployed	-1.506	9.906e-03
9	What is your current occupation_Student	-1.428	2.124e-02
3	Lead Source_Olark Chat	1.424	9.176e-34
7	Last Activity_Olark Chat Conversation	-1.336	3.436e-13
11	What is your current occupation_Working Professional	1.275	3.824e-02
6	Last Activity_Converted to Lead	-1.246	1.265e-07
8	Last Activity_SMS Sent	1.046	4.751e-36
0	const	0.074	8.995e-01

Top 3 Features Positively Impacting Conversion:

	Feature	Coefficient	P-value
5	Lead Source_Welingak Website	5.455	5.205e-14
4	Lead Source_Reference	3.457	1.404e-51
12	Last Notable Activity_Unreachable	2.552	1.732e-03

Top 3 Features Negatively Impacting Conversion:

	Feature	Coefficient	P-value
1	Do Not Email	-1.537	9.869e-16
10	What is your current occupation_Unemployed	-1.506	9.906e-03
9	What is your current occupation_Student	-1.428	2.124e-02

Suggest future work: Model refinement, data updates, A/B testing of strategies, …

Thank you