Optimizing Lead Conversion: Developing a Logistic Regression Model

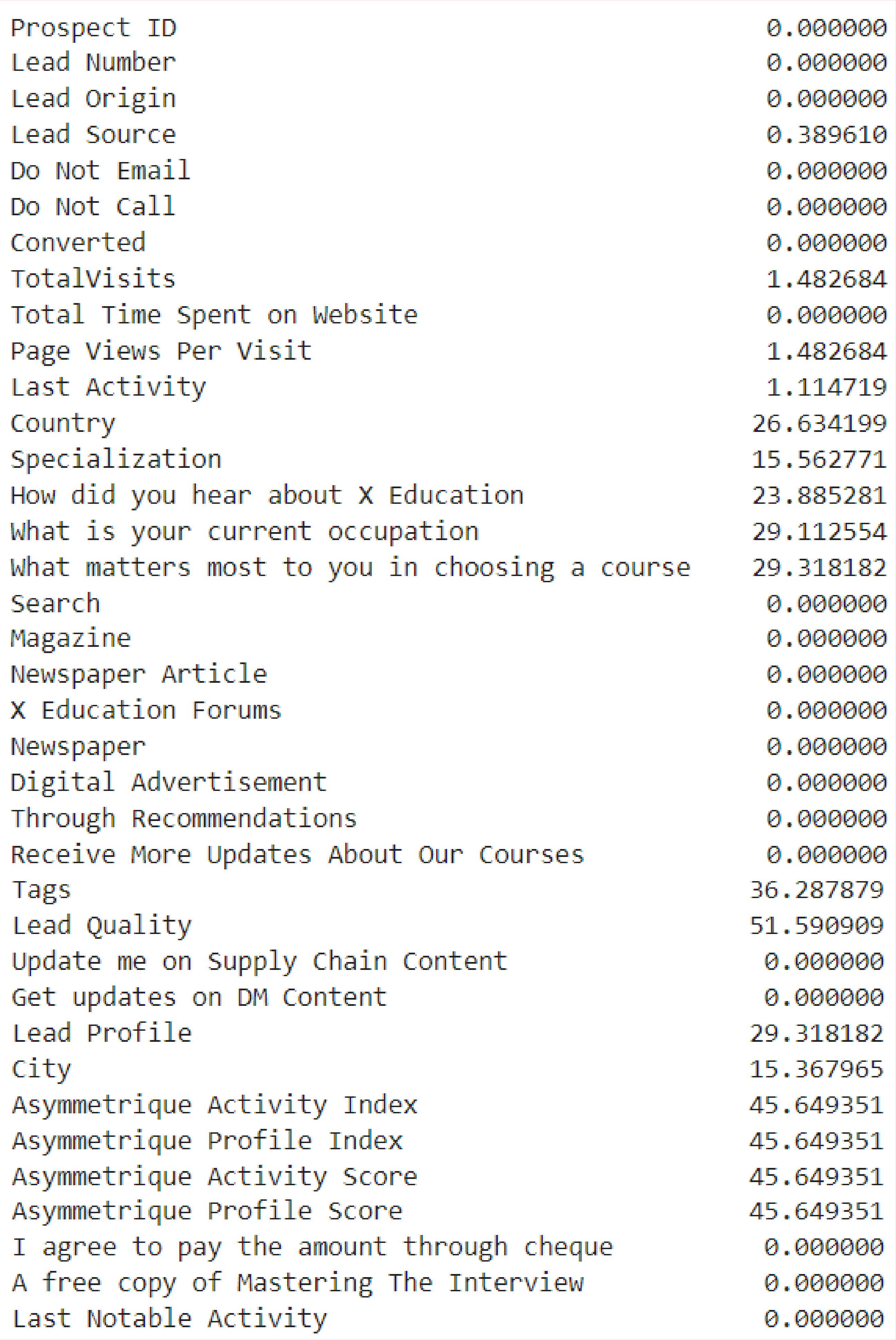
Data Understanding and Preparation

- Data Loading and Initial Checks
- Loaded dataset from 'Leads.xlsx'.
- Checked for missing values and handled them based on a predefined strategy:

Missing Values

Initial Missing Values

- Various columns with missing values ranging from 0% to 51%.
- Applied cleaning strategy to handle these missing values effectively.



```
Prospect ID                                       0.000000
Lead Number                                       0.000000
Lead Origin                                       0.000000
Lead Source                                       0.389610
Do Not Email                                      0.000000
Do Not Call                                       0.000000
Converted                                         0.000000
TotalVisits                                       1.482684
Total Time Spent on Website                       0.000000
Page Views Per Visit                              1.482684
Last Activity                                     1.114719
Country                                          26.634199
Specialization                                   15.562771
How did you hear about X Education               23.885281
What is your current occupation                  29.112554
What matters most to you in choosing a course    29.318182
Search                                            0.000000
Magazine                                          0.000000
Newspaper Article                                 0.000000
X Education Forums                                0.000000
Newspaper                                         0.000000
Digital Advertisement                             0.000000
Through Recommendations                           0.000000
Receive More Updates About Our Courses            0.000000
Tags                                             36.287879
Lead Quality                                     51.590909
Update me on Supply Chain Content                 0.000000
Get updates on DM Content                         0.000000
Lead Profile                                     29.318182
City                                             15.367965
Asymmetrique Activity Index                      45.649351
Asymmetrique Profile Index                       45.649351
Asymmetrique Activity Score                      45.649351
Asymmetrique Profile Score                       45.649351
I agree to pay the amount through cheque          0.000000
A free copy of Mastering The Interview            0.000000
Last Notable Activity                             0.000000
```

Data Cleaning

Handling Missing Values and Outliers

- Deleted columns with >40% missing data.
- Dropped rows with missing values in key columns.
- Filled remaining missing values appropriately.
- Removed duplicates and outliers for key numerical columns.

Standardization and Dummies

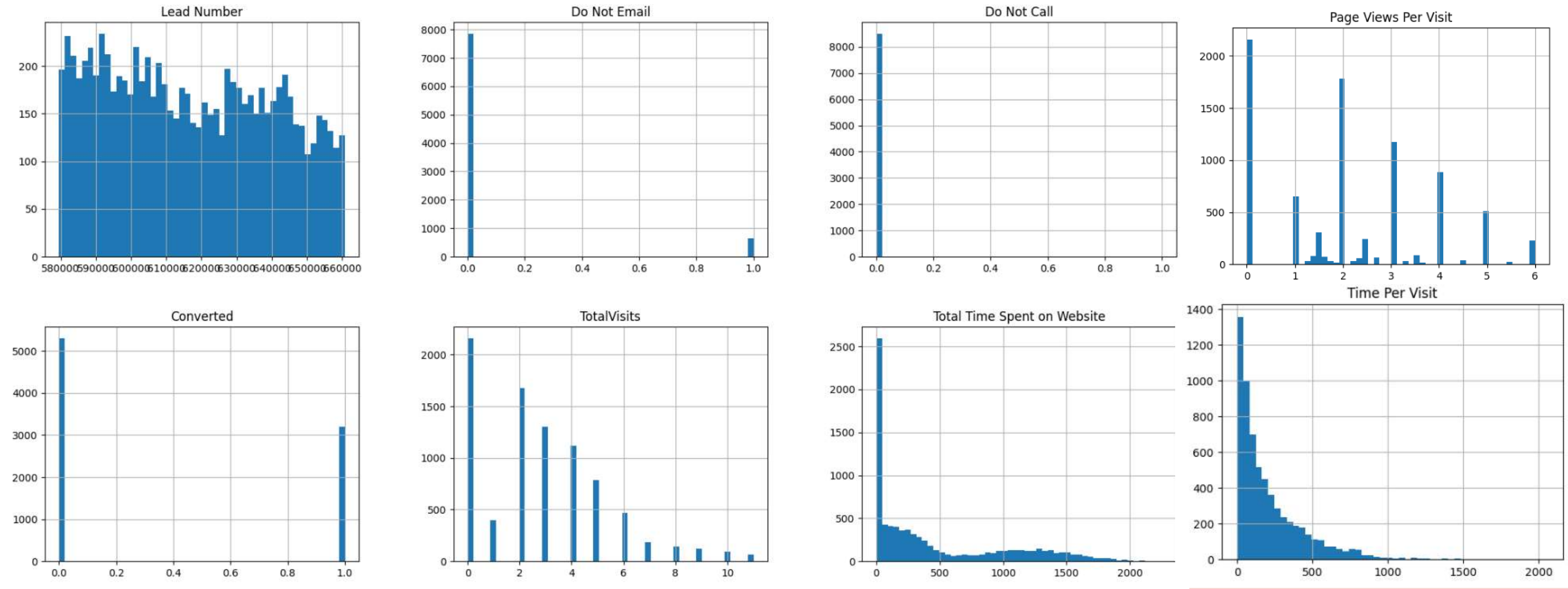
Data Standardization and Creating Dummy Variables

- Standardized 'Do Not Email' and 'Do Not Call' columns.
- Created dummy variables for categorical columns.
- Derived new metrics like 'Time Per Visit'.

Exploratory Data Analysis (EDA)

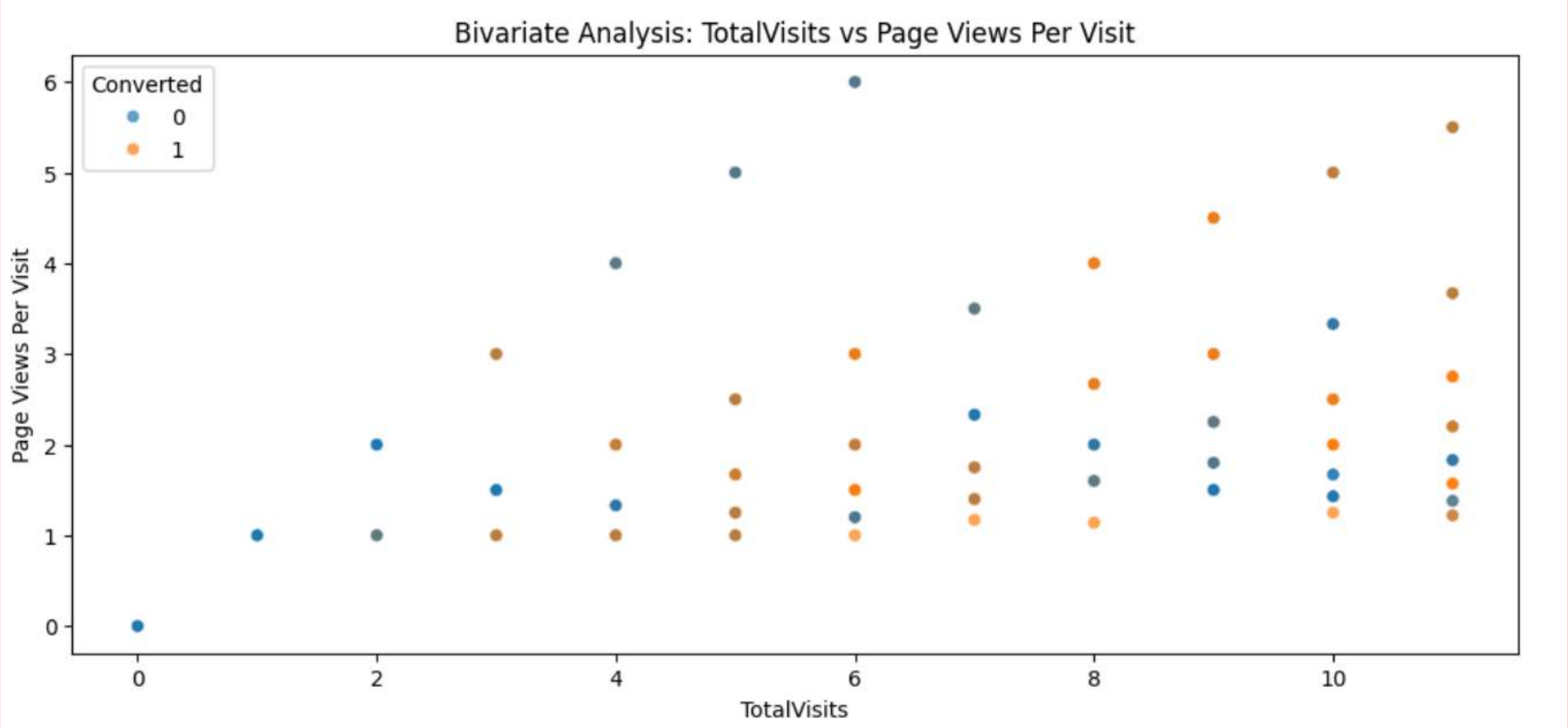
Univariate Analysis

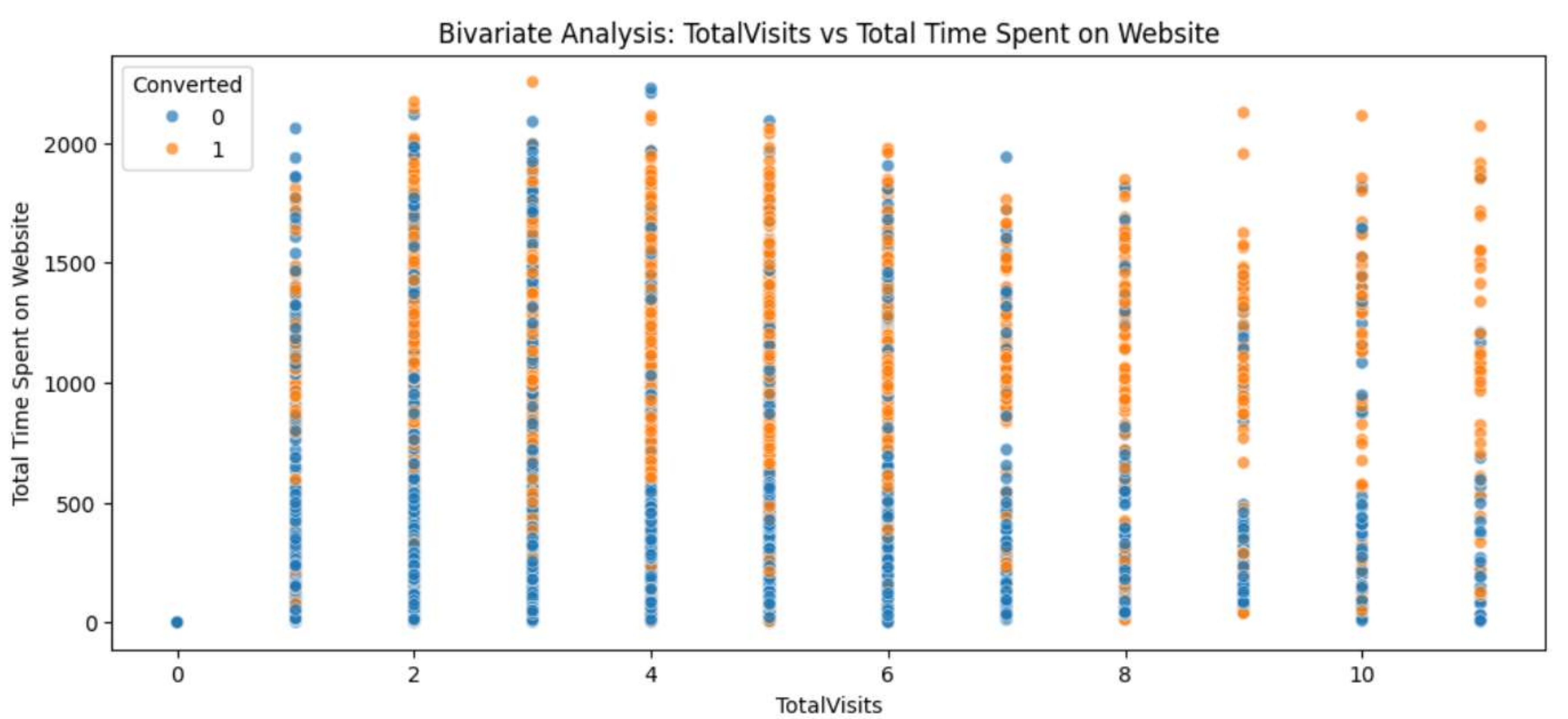
Plotted histograms for all variables to understand distribution.

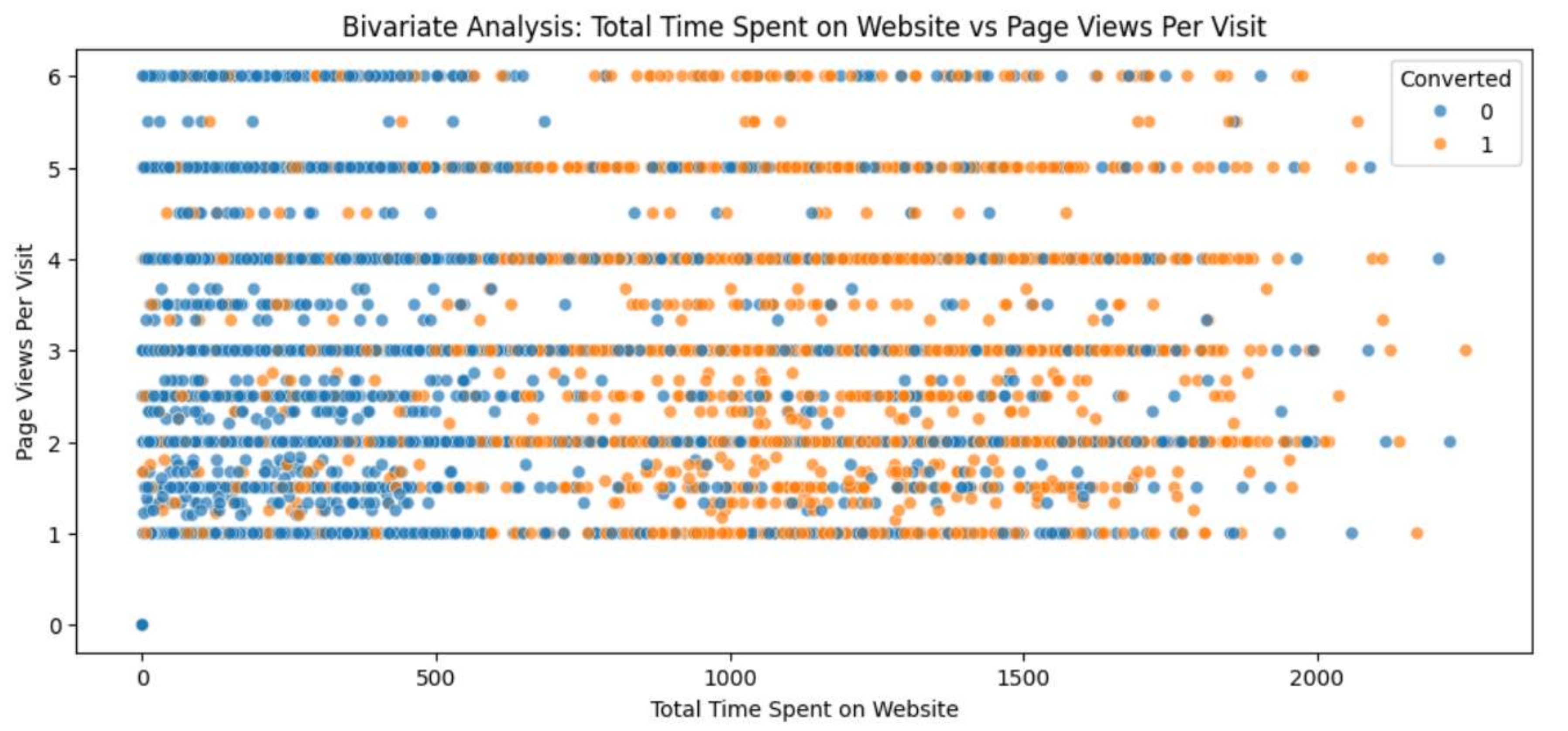


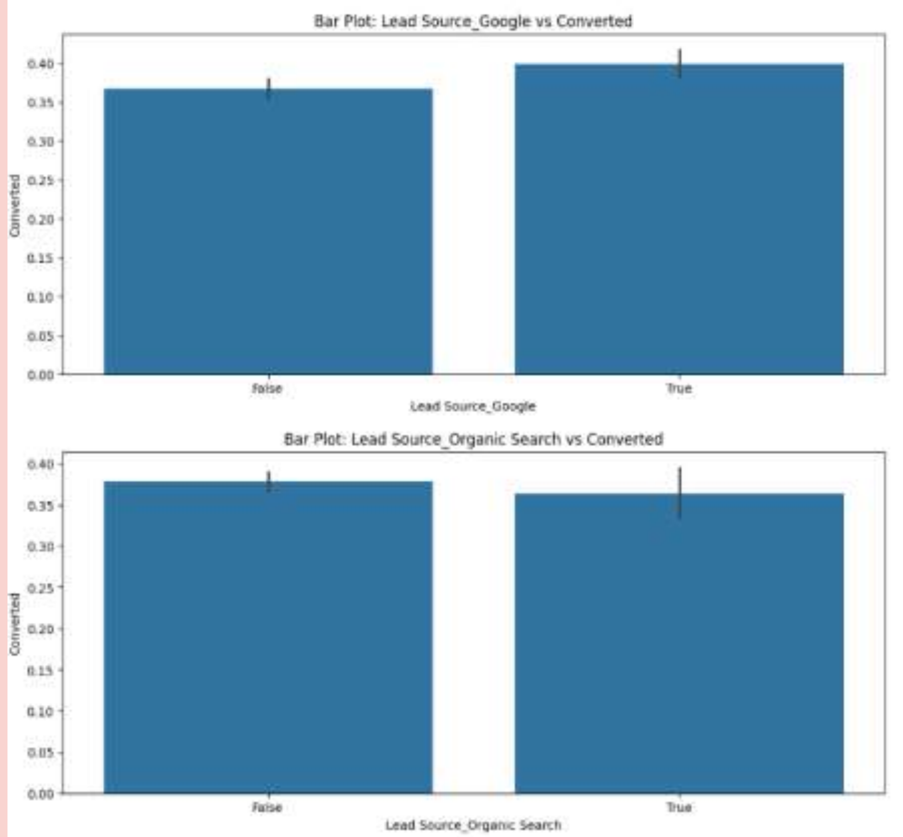
Bivariate Analysis

Scatter Plots and Box Plots



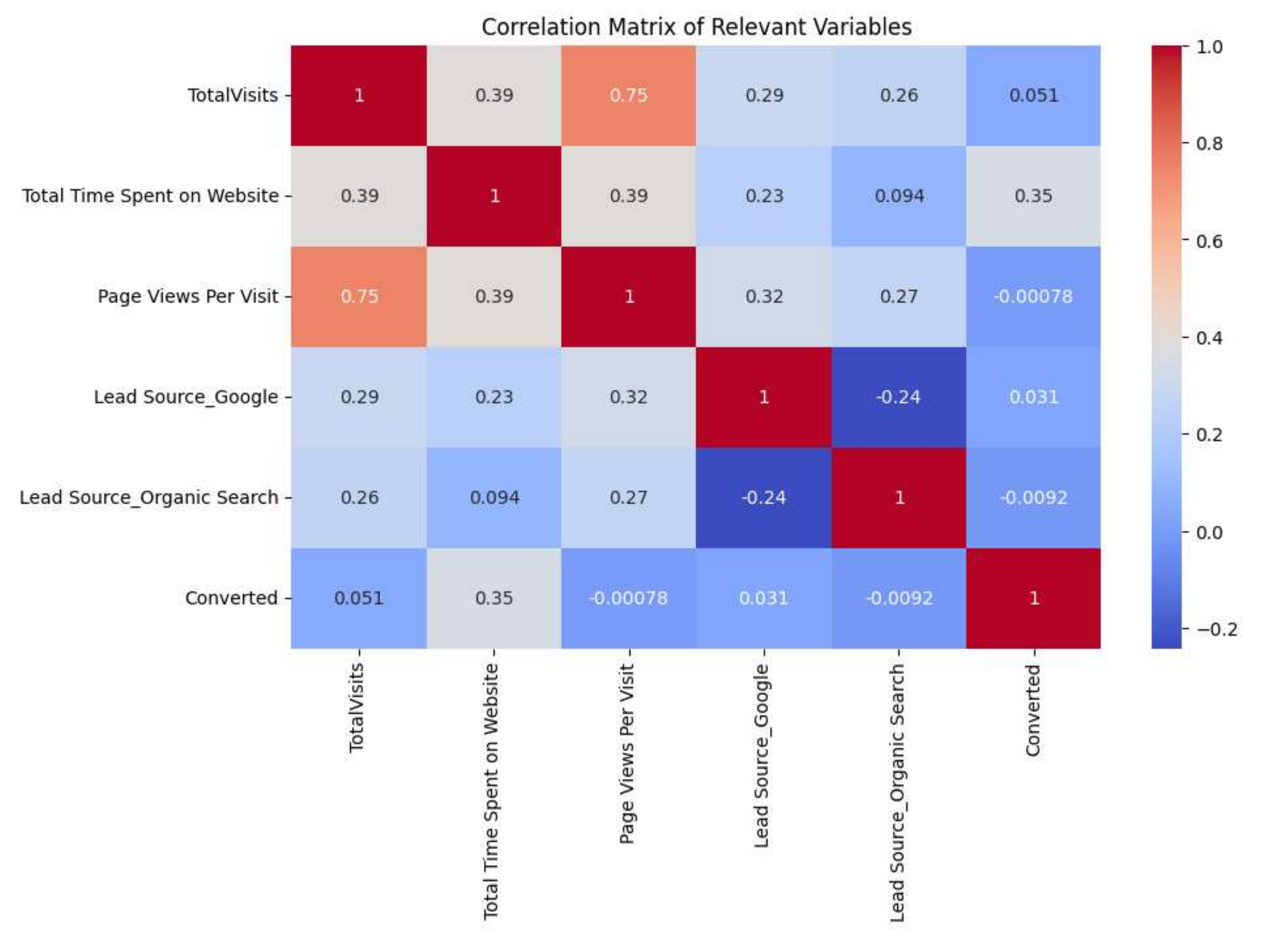






Correlation Matrix

Heat Map



	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Source_Google	Lead Source_Organic Search	Converted
TotalVisits	1	0.39	0.75	0.29	0.26	0.051
Total Time Spent on Website	0.39	1	0.39	0.23	0.094	0.35
Page Views Per Visit	0.75	0.39	1	0.32	0.27	-0.00078
Lead Source_Google	0.29	0.23	0.32	1	-0.24	0.031
Lead Source_Organic Search	0.26	0.094	0.27	-0.24	1	-0.0092
Converted	0.051	0.35	-0.00078	0.031	-0.0092	1

Model Building

Logistic Regression Model

- Split data into training and testing sets.
- Handled missing values with SimpleImputer.
- Fitted a logistic regression model with max_iter=1000.

Model Evaluation

Model Performance Metrics

Evaluated model using:

- ○ Accuracy: 90%
- ○ Precision: 91%
- ○ Recall: 82%
- ○ F1 Score: 87%
- ○ ROC-AUC Score: 96%

Model Interpretation

Top 5 Important Features

- Identified top 5 features contributing to the model:
- Tags_Will revert after reading the email
- Tags_Lost to EINS
- Tags_Closed by Horizzon
- City_Select
- Lead Origin_Lead Add Form

Recommendations

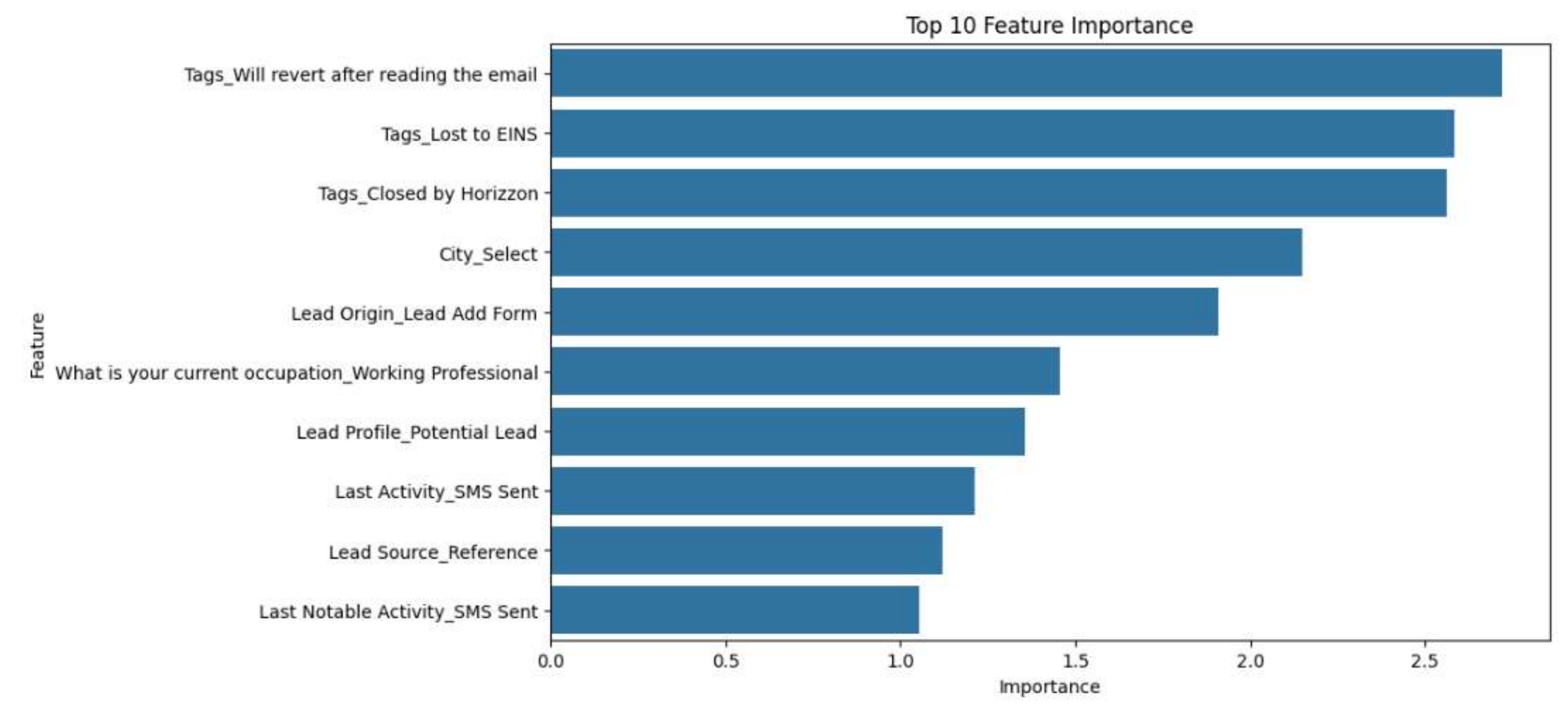
Aggressive Strategy with Interns

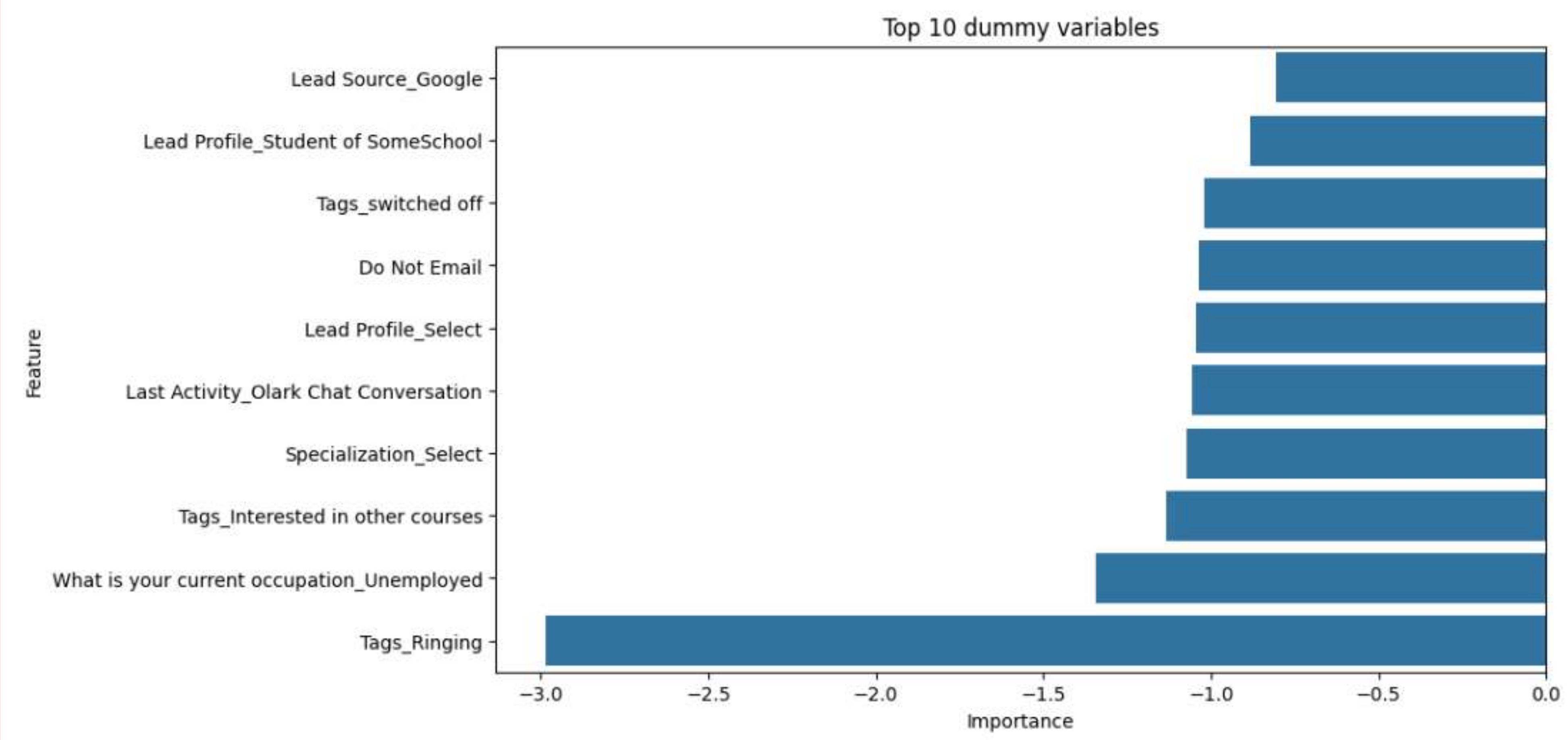
- Prioritize leads predicted as 1.
- Increase follow-up calls and emails.
- Offer promotions and provide intern training.

Minimizing Calls Strategy

- Focus on leads with high conversion probability (>0.8).
- Reduce follow-up calls and use automated emails.
- Reallocate resources to other tasks.

Graphs


Top 10 Feature Importance
Tags_Will revert after reading the email
Tags_Lost to EINS
Tags_Closed by Horizzon
City_Select
Lead Origin_Lead Add Form
What is your current occupation_Working Professional
Lead Profile_Potential Lead
Last Activity_SMS Sent
Lead Source_Reference
Last Notable Activity_SMS Sent
Feature
0.0
0.5
1.0
1.5
2.0
2.5
Importance


Top 10 dummy variables
Lead Source_Google
Lead Profile_Student of SomeSchool
Tags_switched off
Do Not Email
Lead Profile_Select
Last Activity_Olark Chat Conversation
Specialization_Select
Tags_Interested in other courses
What is your current occupation_Unemployed
Tags_Ringing
Feature
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
Importance

Conclusion

Summary of the Analysis and Model

- Effective data cleaning and preparation.
- Built a robust logistic regression model.
- Provided actionable recommendations based on model insights.

Thanks!