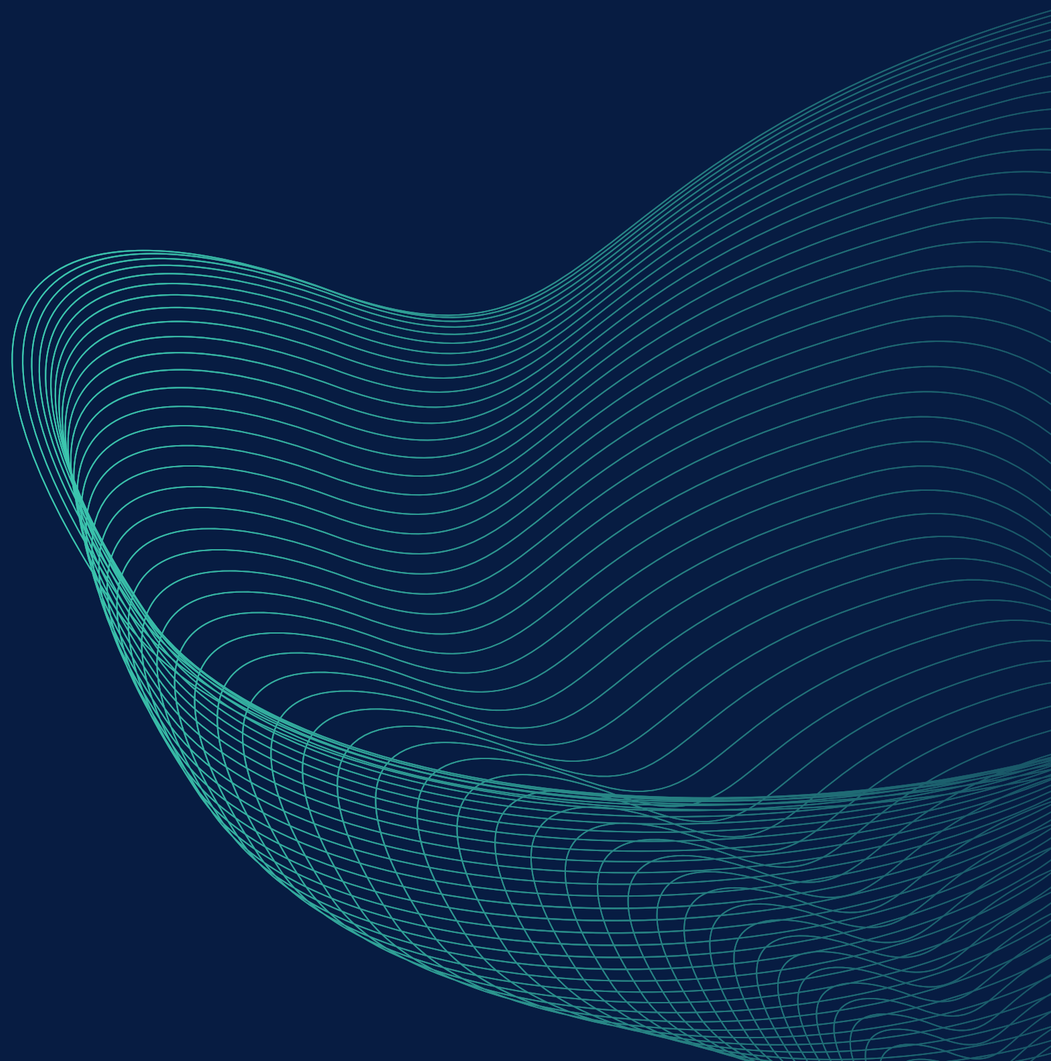
Lead Scoring Case Study

Data Science Classification Project



Problem Statement

X Education online courses

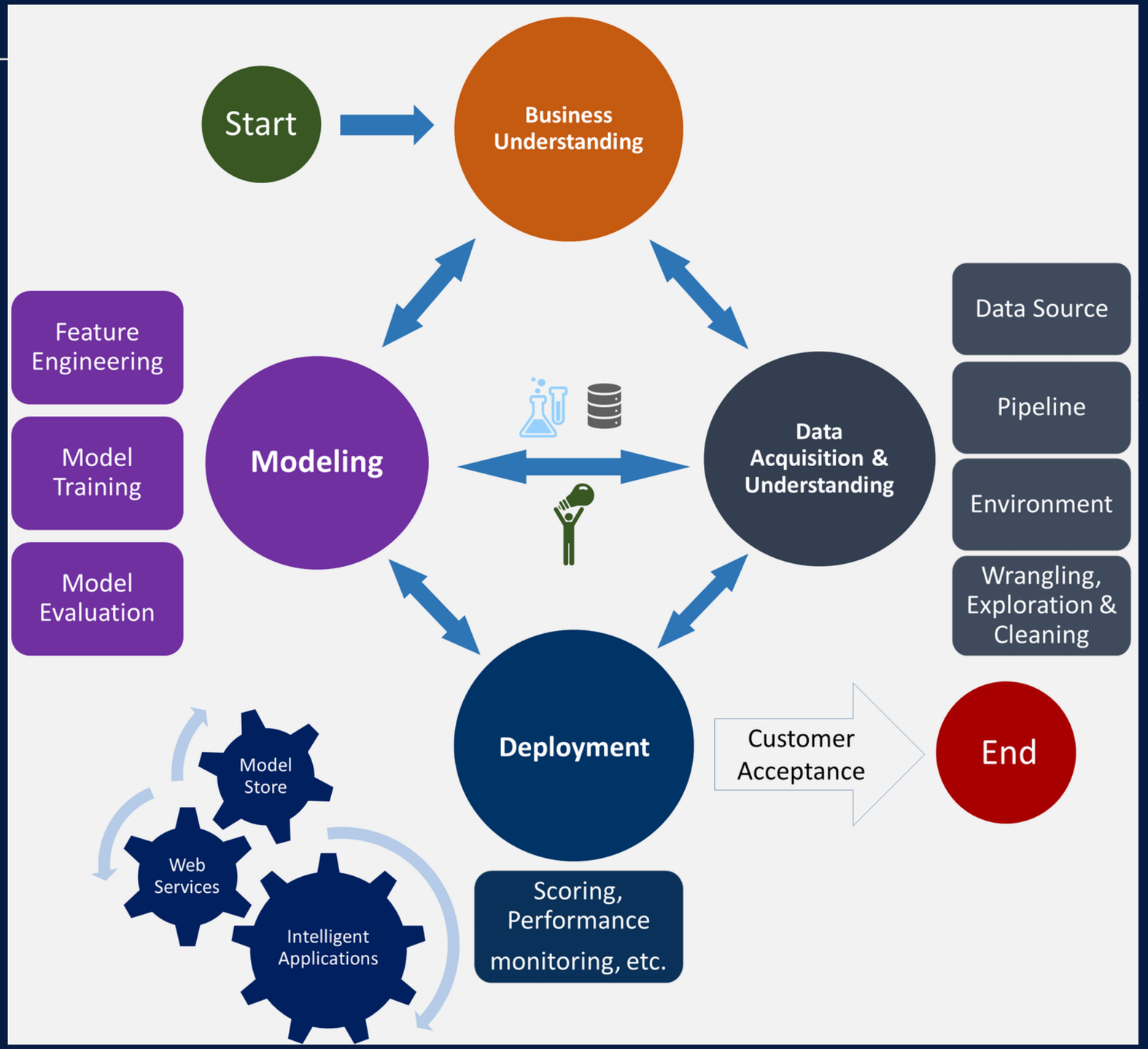
X Education is an online education company selling courses to professionals. Despite generating numerous leads through website visits, form submissions, and referrals, their lead conversion rate is only 30%. To enhance efficiency, the company aims to identify 'Hot Leads' – those most likely to convert into paying customers. The CEO targets an 80% lead conversion rate. The current lead conversion process resembles a funnel, with a significant drop-off from leads to actual customers. The company has tasked you with building a model to assign lead scores, enabling prioritization of leads for more effective sales efforts and ultimately increasing the conversion rate.

Goals of the Case Study

There are quite a few goals for this case study:

- 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Data Science Project Flow



Data

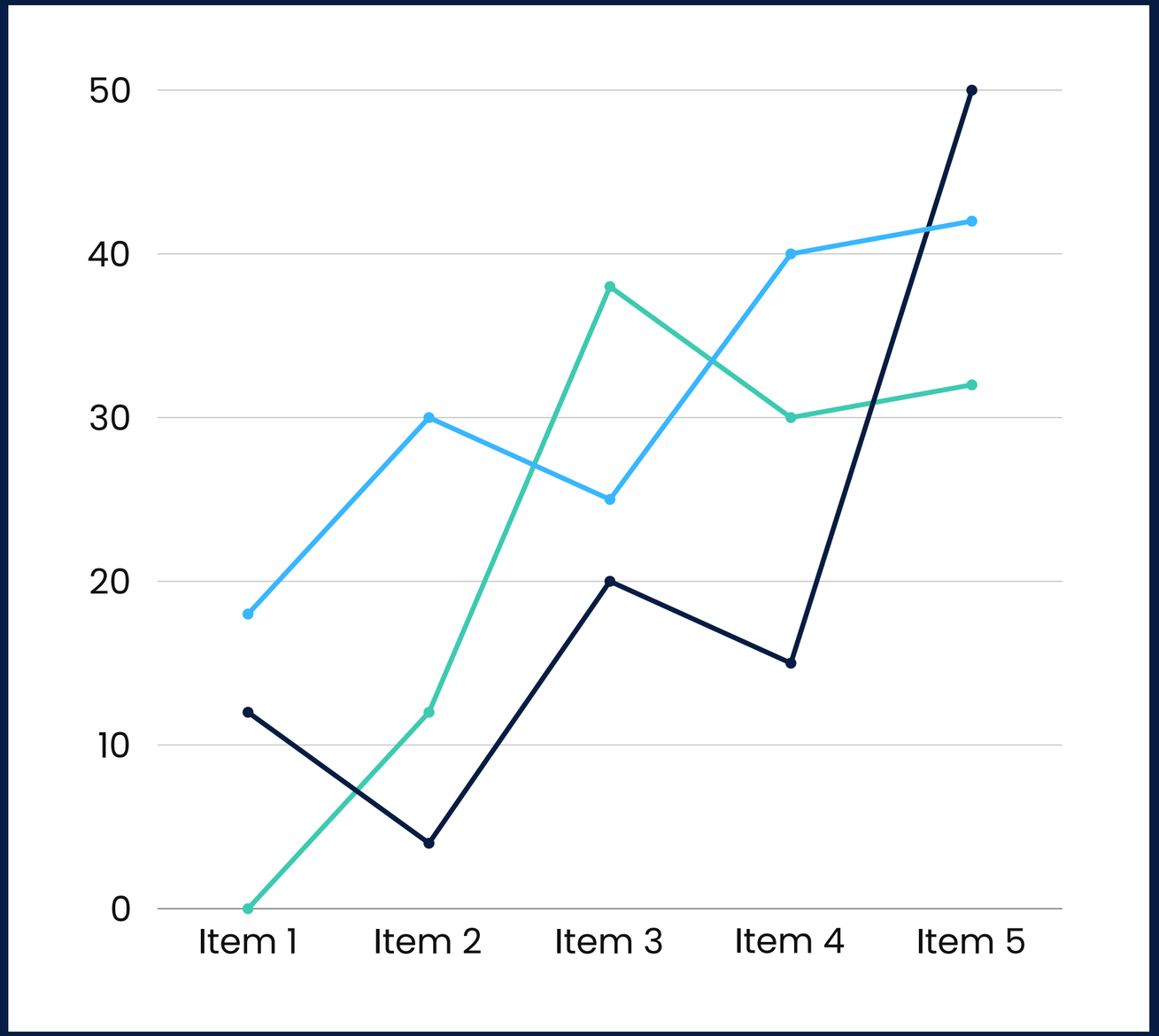
Data Dictionary

	Variables	Description
0	Prospect ID	A unique ID with which the customer is identified.
1	Lead Number	A lead number assigned to each lead procured.
2	Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
3	Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
4	Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
5	Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
6	Converted	The target variable. Indicates whether a lead has been successfully converted or not.
7	TotalVisits	The total number of visits made by the customer on the website.
8	Total Time Spent on Website	The total time spent by the customer on the website.
9	Page Views Per Visit	Average number of pages on the website viewed during the visits.
10	Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
11	Country	The country of the customer.
12	Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.
13	How did you hear about X Education	The source from which the customer heard about X Education.
14	What is your current occupation	Indicates whether the customer is a student, umemployed or employed.
15	What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.
16	Search	Indicating whether the customer had seen the ad in any of the listed items.
17	Magazine	NaN
18	Newspaper Article	NaN
19	X Education Forums	NaN
20	Newspaper	NaN
21	Digital Advertisement	NaN
22	Through Recommendations	Indicates whether the customer came in through recommendations.
23	Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.
24	Tags	Tags assigned to customers indicating the current status of the lead.
25	Lead Quality	Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.
26	Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
27	Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.
28	Lead Profile	A lead level assigned to each customer based on their profile.
29	City	The city of the customer.
30	Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
31	Asymmetrique Profile Index	NaN
32	Asymmetrique Activity Score	NaN
33	Asymmetrique Profile Score	NaN
34	I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.
35	a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
36	Last Notable Activity	The last notable acitivity performed by the student.

Data Sample

```
Prospect ID                                    7927b2df-8bba-4d29-b9a2-b6e0beafe620
Lead Number                                                                 660737
Lead Origin                                                                    API
Lead Source                                                            Olark Chat
Do Not Email                                                                   No
Do Not Call                                                                    No
Converted                                                                       0
TotalVisits                                                                   0.0
Total Time Spent on Website                                                     0
Page Views Per Visit                                                          0.0
Last Activity                                             Page Visited on Website
Country                                                                      NaN
Specialization                                                            Select
How did you hear about X Education                                        Select
What is your current occupation                                       Unemployed
What matters most to you in choosing a course                  Better Career Prospects
Search                                                                         No
Magazine                                                                       No
Newspaper Article                                                              No
X Education Forums                                                             No
Newspaper                                                                      No
Digital Advertisement                                                          No
Through Recommendations                                                        No
Receive More Updates About Our Courses                                         No
Tags                                              Interested in other courses
Lead Quality                                                  Low in Relevance
Update me on Supply Chain Content                                              No
Get updates on DM Content                                                      No
Lead Profile                                                              Select
City                                                                      Select
Asymmetrique Activity Index                                           02.Medium
Asymmetrique Profile Index                                            02.Medium
Asymmetrique Activity Score                                                15.0
Asymmetrique Profile Score                                                 15.0
I agree to pay the amount through cheque                                       No
A free copy of Mastering The Interview                                         No
Last Notable Activity                                                  Modified
Name: 0, dtype: object
```

Exploratory Data Analysis



Initial Analysis

Number of Features:

- Total features: 28
- Numerical features: 3
- Categorical features: 14
- Flag-like features: 8

Identifiers:

- Unique identifiers: 2

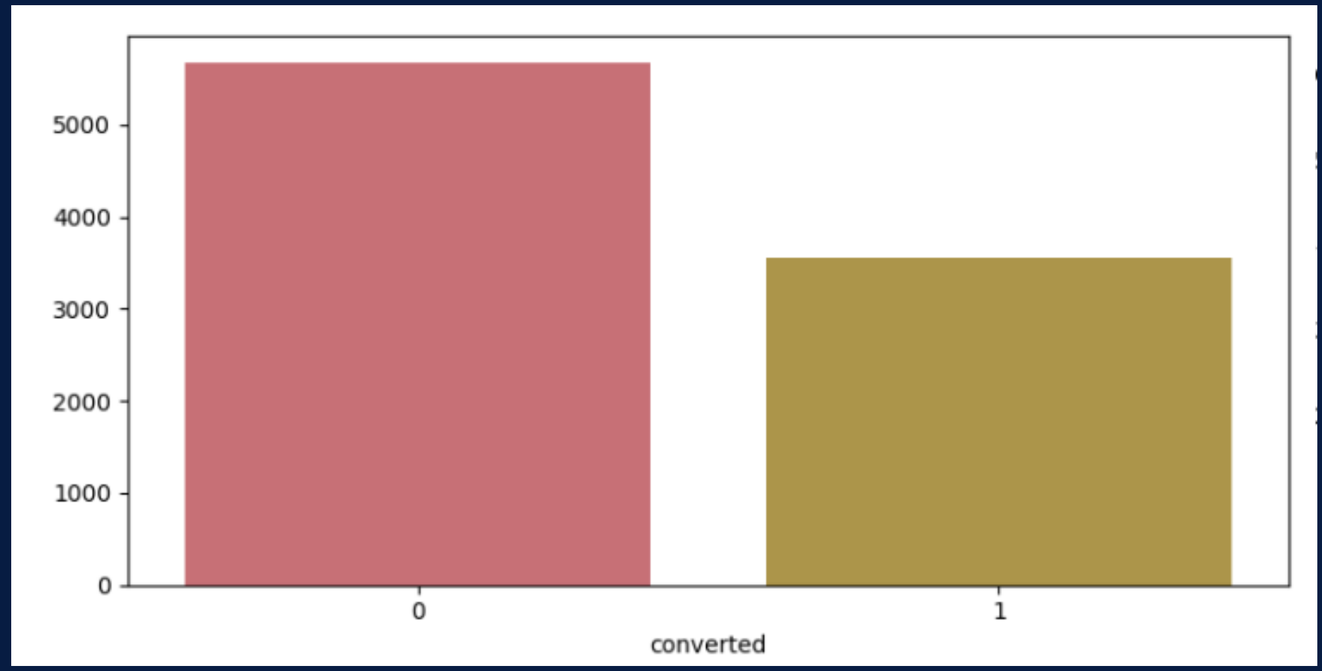
Target Feature:

- Target feature: 'converted'

Percent of Misssing Value

	percent_missing
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

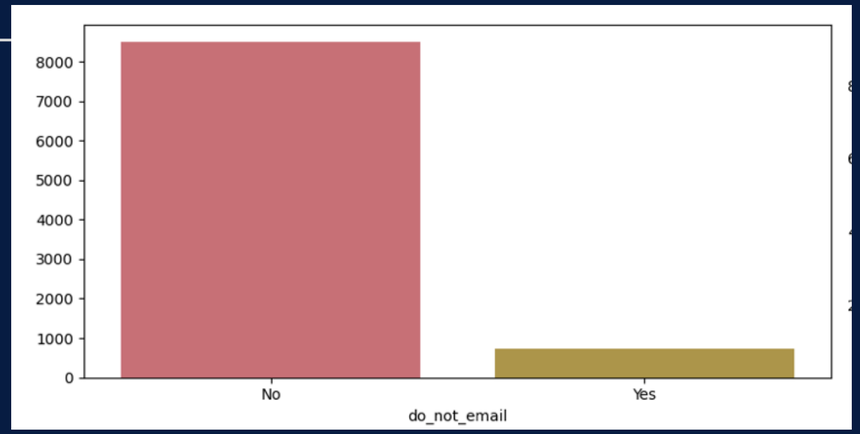
Analysis Of Target Column

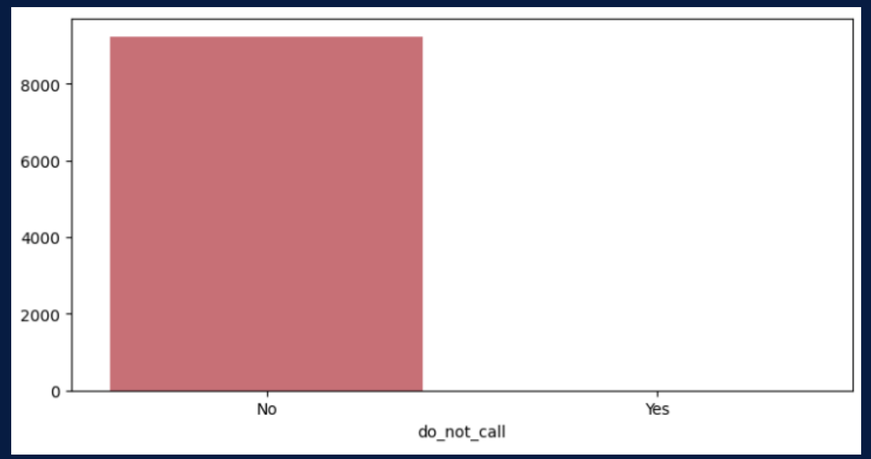


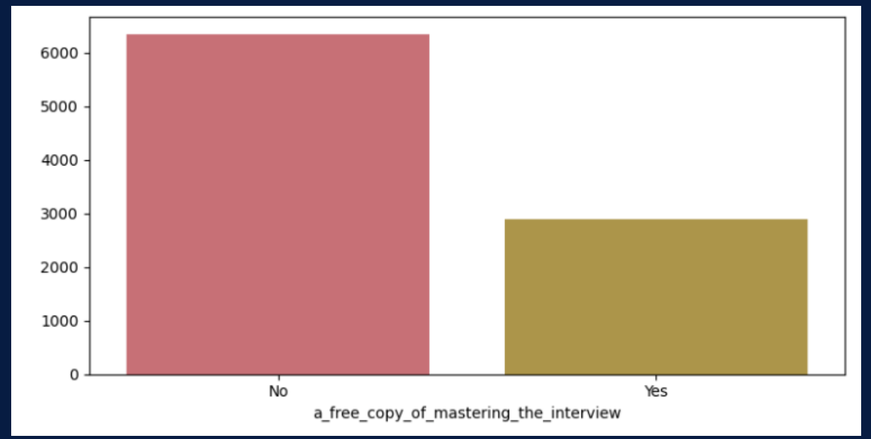
Observations:

- The data is imbalanced in nature
- Most of the data belongs to the non converting class
- There is only `38.54%` chance that a given lead would be converted

Analyzing the flag features



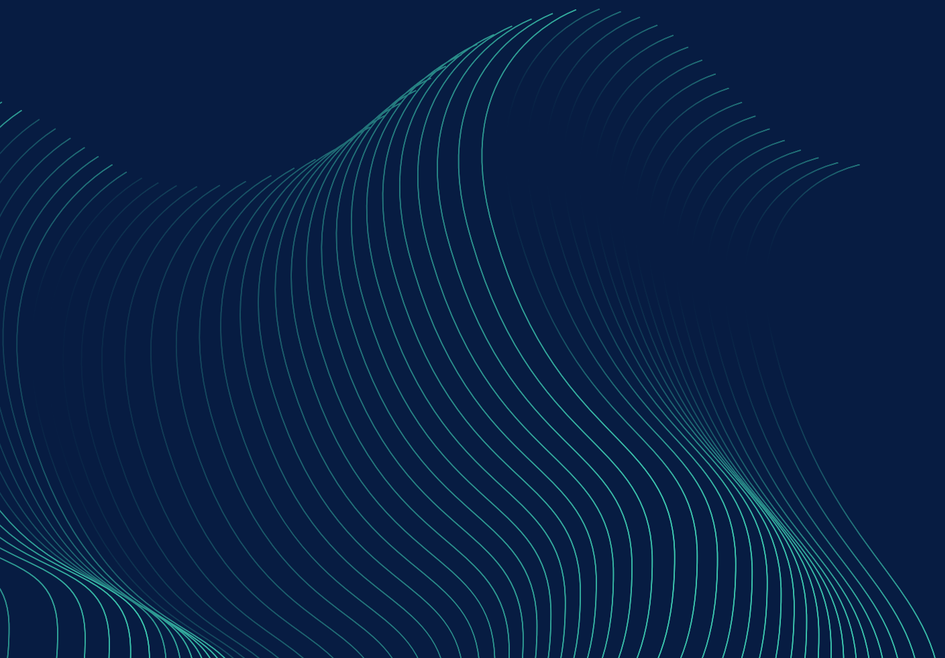




Observations:

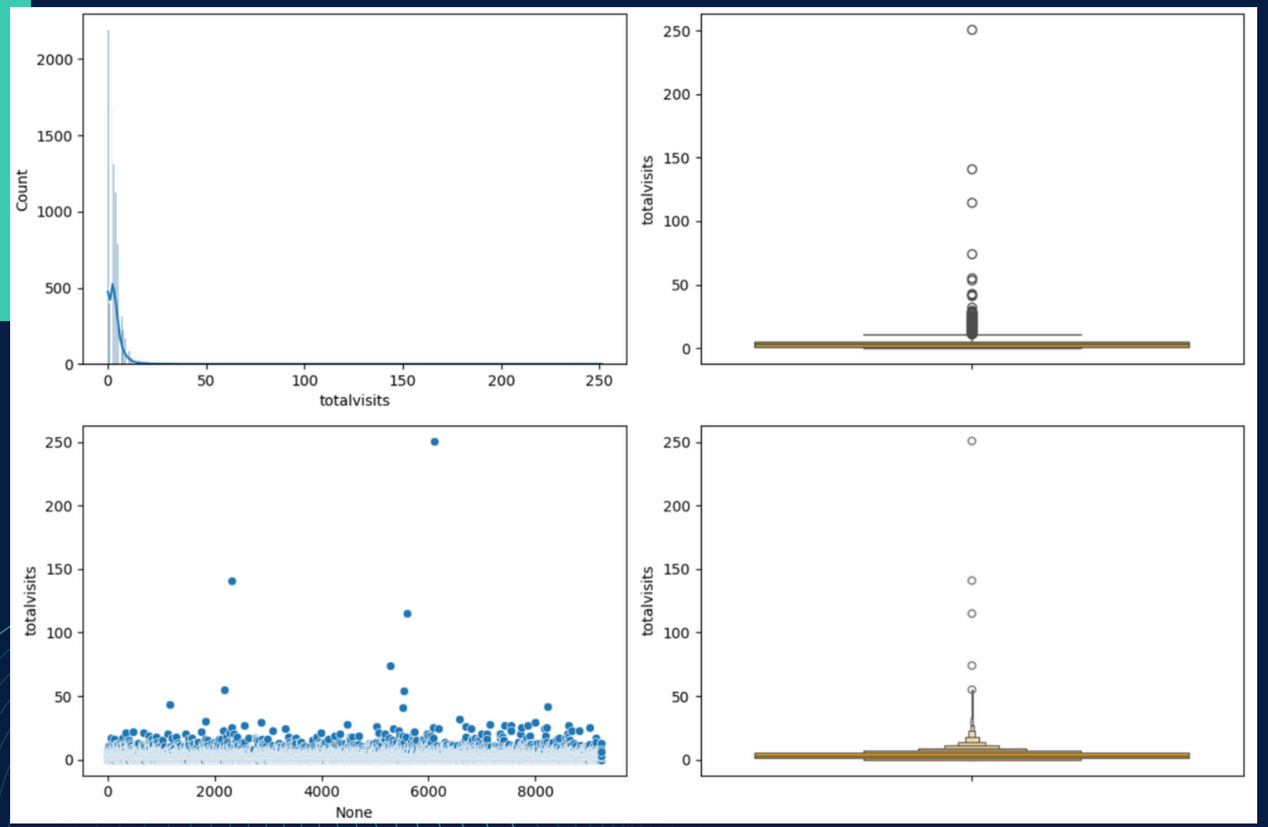
- As we can see from the plots there are numerous flag features that have significantly lower proportional distributions
- There are only one feature that has a good enough distribution to be counted for i.e `a_free_copy_of_mastering_the_interview`
- We can drop all other flag features except this

Analyzing the numerical features

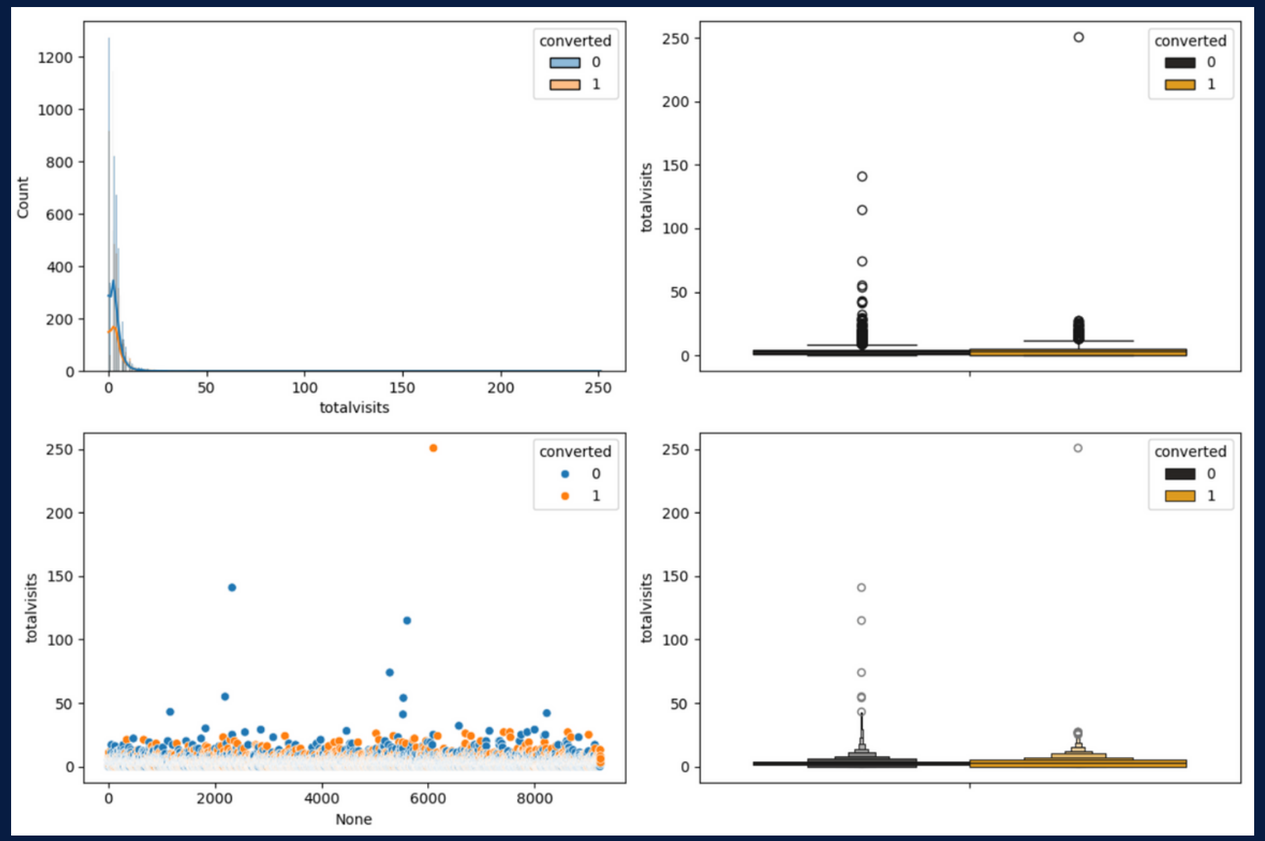


totalvisits

Univariate Analysis



Bivariate Analysis



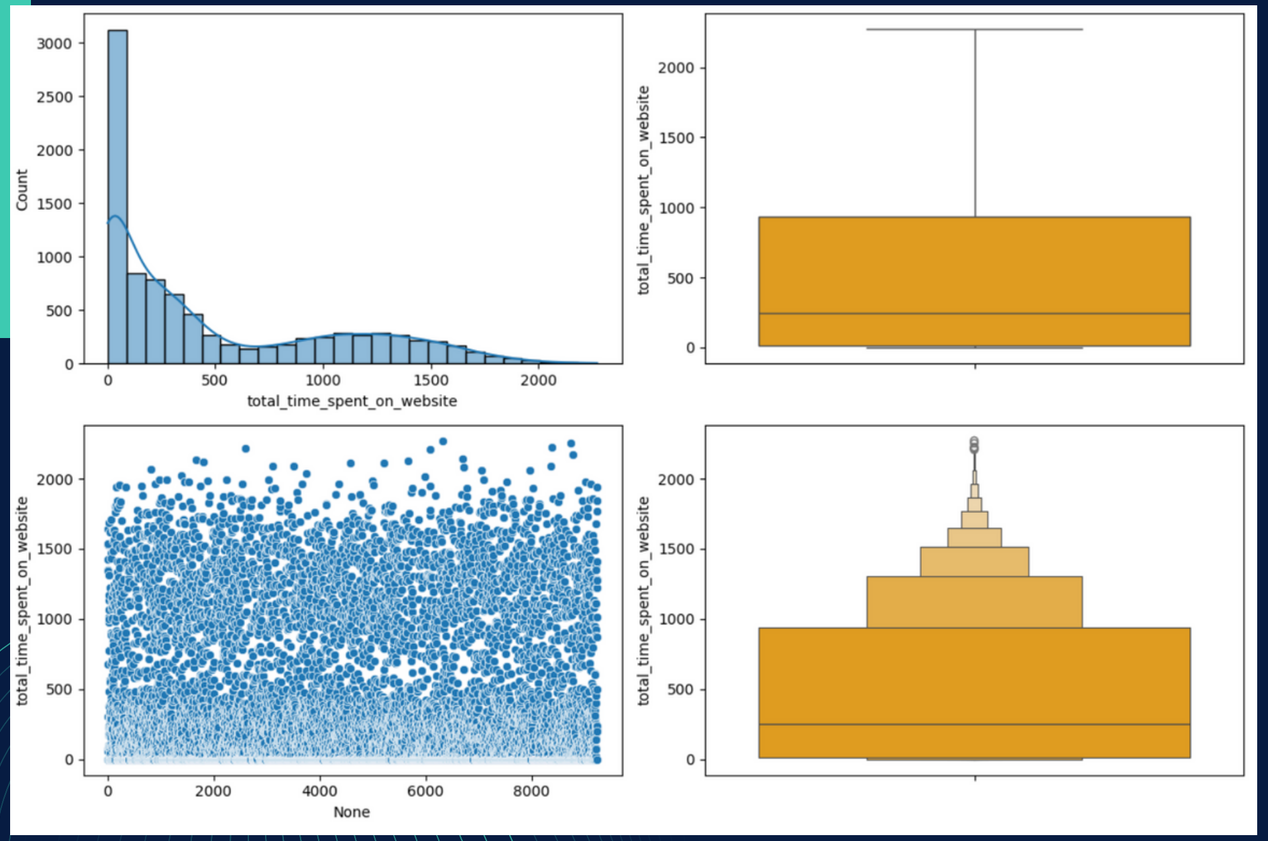
Observations: Significant statistical relationship between totalvisits and converted, with evident outliers and highly skewed data. Non-converted segment has more outliers, and means across target variable groups are similar.

Conclusions: Converted segment typically has lower totalvisits, while non-converted shows higher values, potentially indicating web-crawlers or indecisive users.

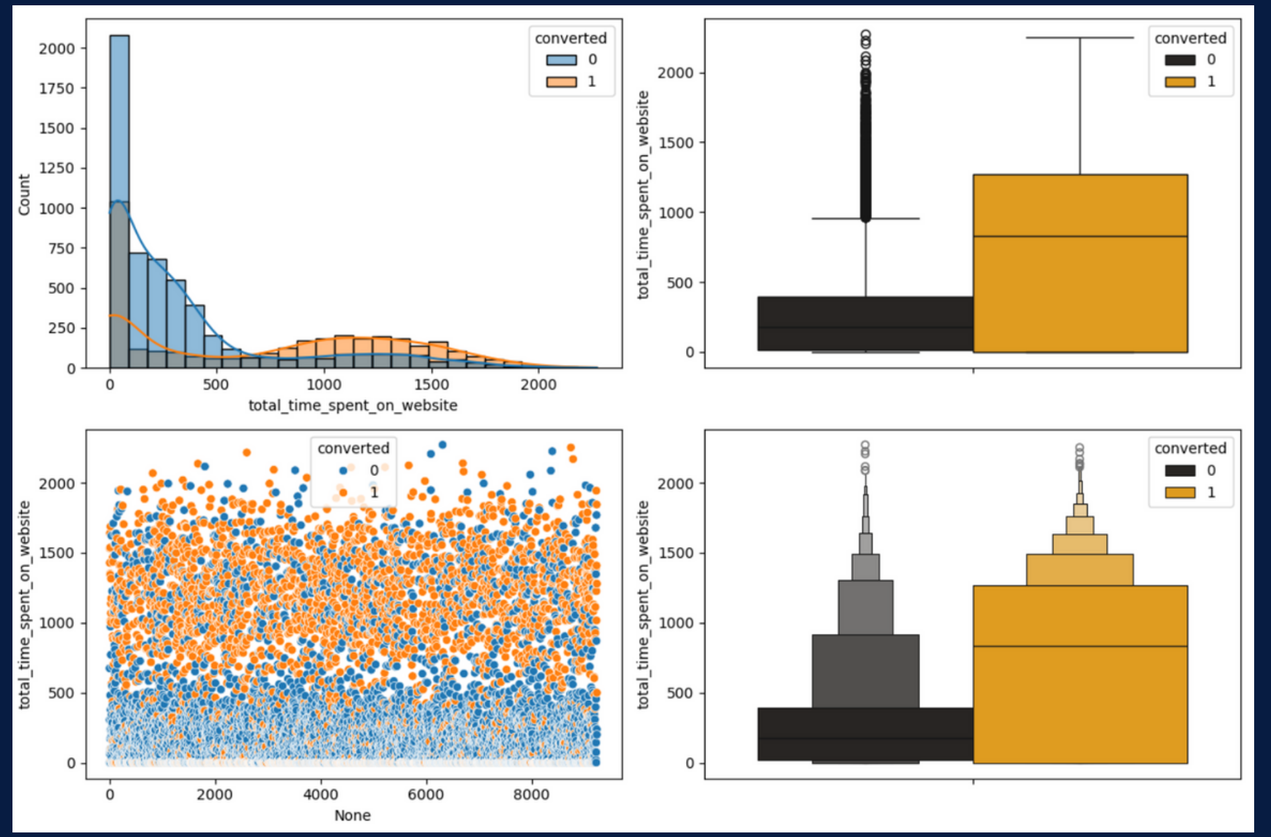
Possible Solutions: Address outliers using IQR and consider replacing values with WOE obtained through discretization for enhanced analysis.

total time spent on website

Univariate Analysis



Bivariate Analysis



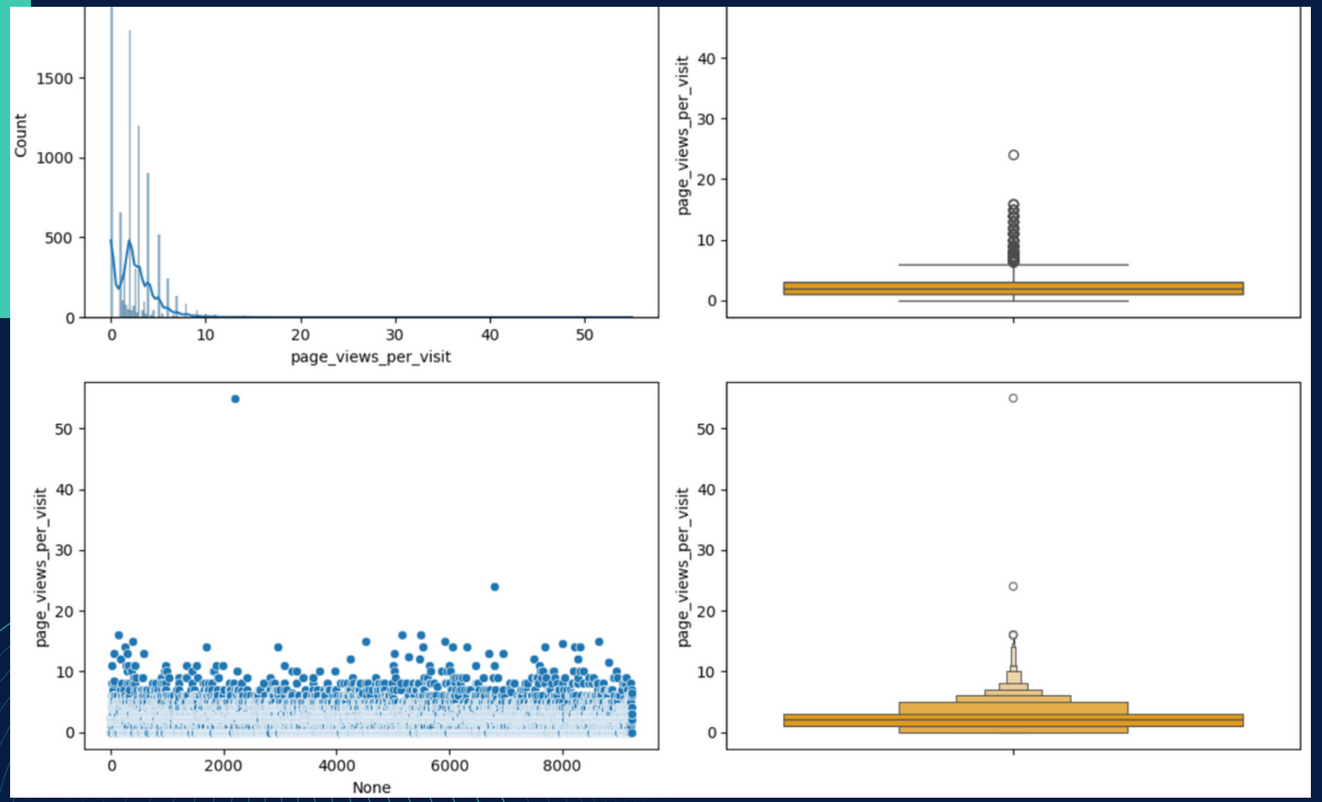
Observations: Despite data skewness, identifiable patterns exist.

Conclusions: Converted candidates spend significantly more time on the website, while most non-converted candidates spend no time. Notably, a substantial non-converted group spends extensively, potentially indicating web crawlers or indecisive behavior.

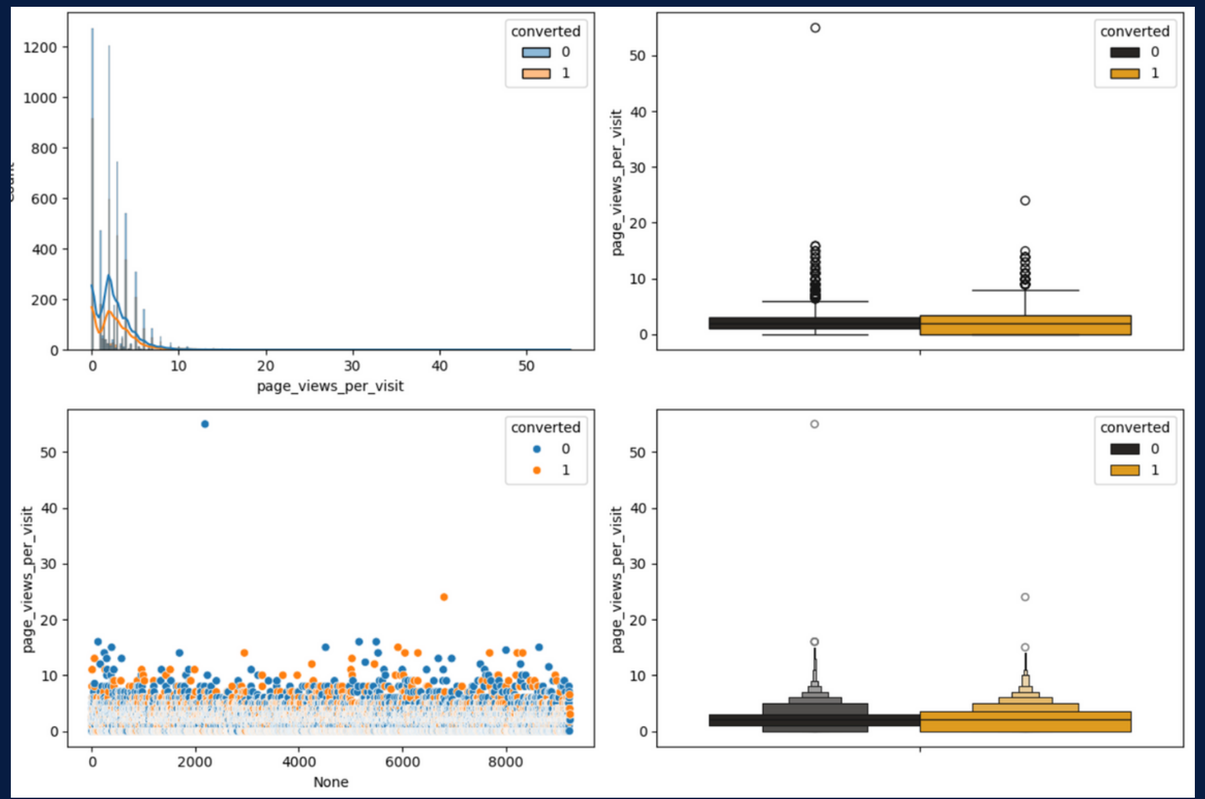
Possible Solutions: Address skewness through outlier treatment and consider either discretization or Weight of Evidence (WOE) replacement for improved analysis.

page_views_per_visit

Univariate Analysis



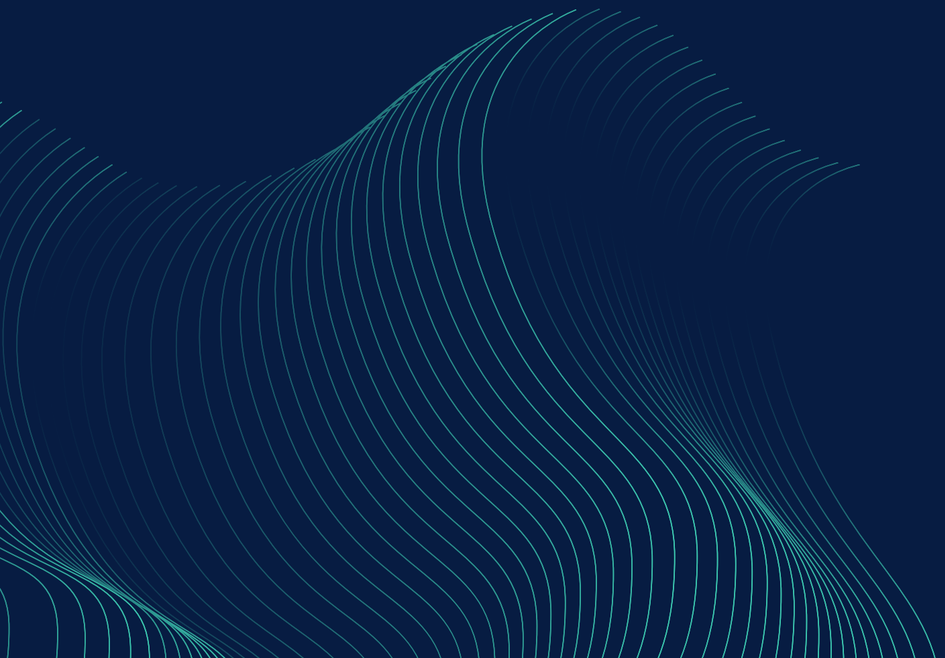
Bivariate Analysis



Observations: The distribution appears random, with means for converted and non-converted segments almost identical, indicating minimal differences. The point-biserial correlation further suggests an insignificant relationship.

Possible Solutions: Consider data cleanup to enhance correlation with the target variable, aiming to discern meaningful patterns in the seemingly random distribution.

Analyzing the categorical features



Initial Analysis

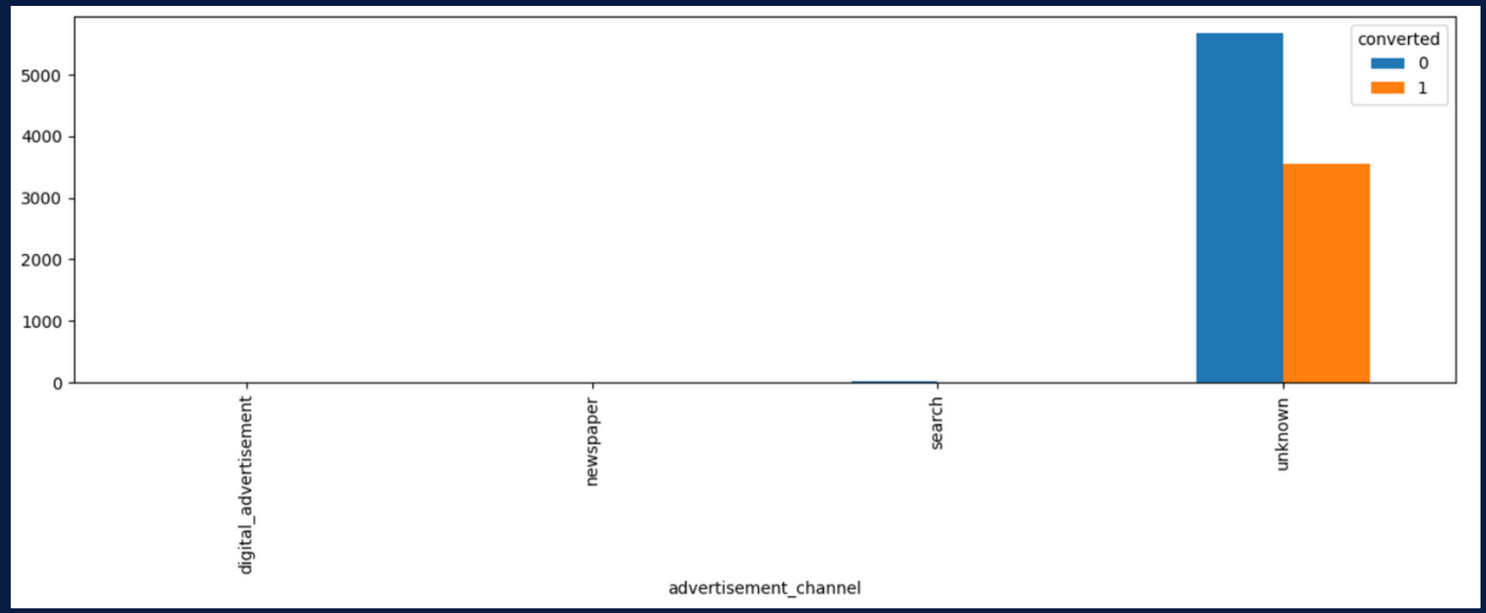
Percentage of missing data in Categorical Variable

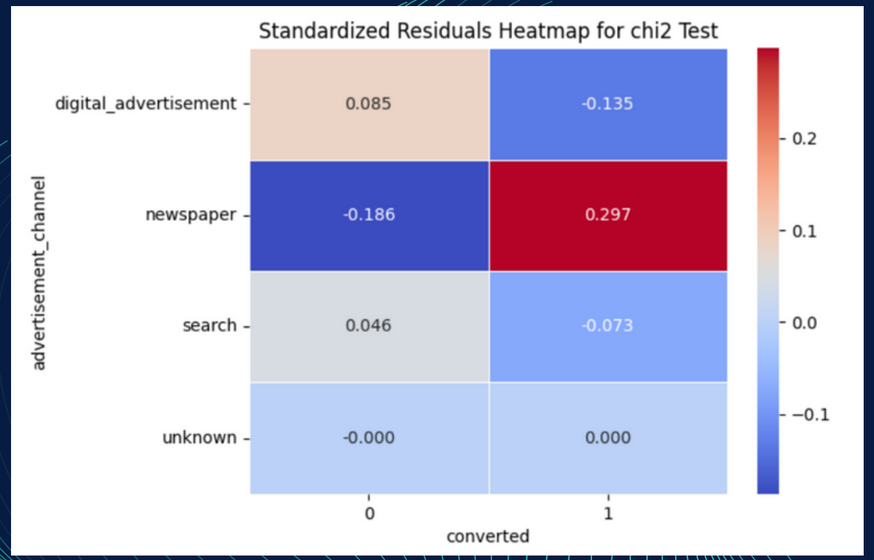
```
Data with missing data less than 25%

prospect_id                              0.00
lead_number                              0.00
converted                                0.00
advertisement_channel                    0.00
lead_origin                              0.00
lead_source                              0.39
last_activity                            1.11
specialization                          15.56
how_did_you_hear_about_x_education      23.89
last_notable_activity                    0.00
city                                    15.37
dtype: float64
**********************************************************

Data with missing data more than 25%

country                                         26.63
what_is_your_current_occupation                 29.11
what_matters_most_to_you_in_choosing_a_course   29.32
tags                                            36.29
lead_quality                                    51.59
lead_profile                                    29.32
```

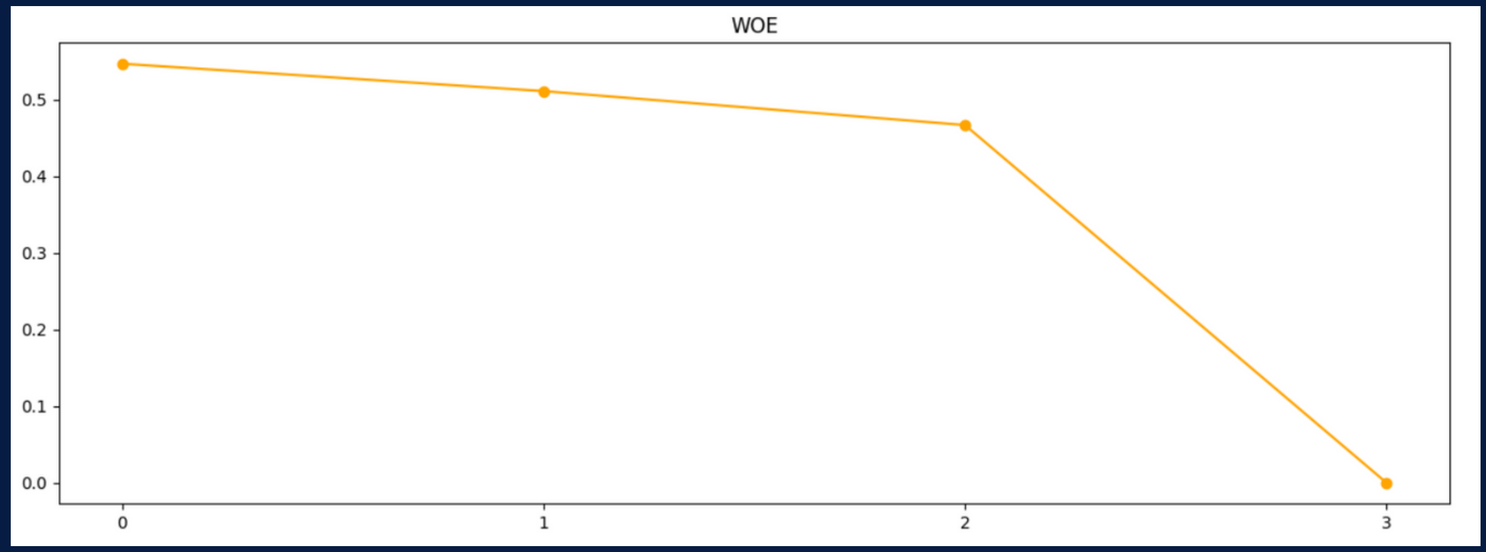
- Observations: Numerous features have over 50% missing data, prompting consideration of a threshold for permissible missing values to enhance dataset quality.
- Possible Actions: Features surpassing the threshold can either be dropped or imputed based on requirements, addressing the challenge of high proportions of missing data in the dataset.

Advertisements channel

- Observations: A substantial imbalance exists in advertisement channels, with 99.79% of observations having an unknown source. Statistical tests (chi2, Cramér's V, Theil's uncertainty) and point-biserial correlation indicate a very weak association between the advertisement channel and the target variable.

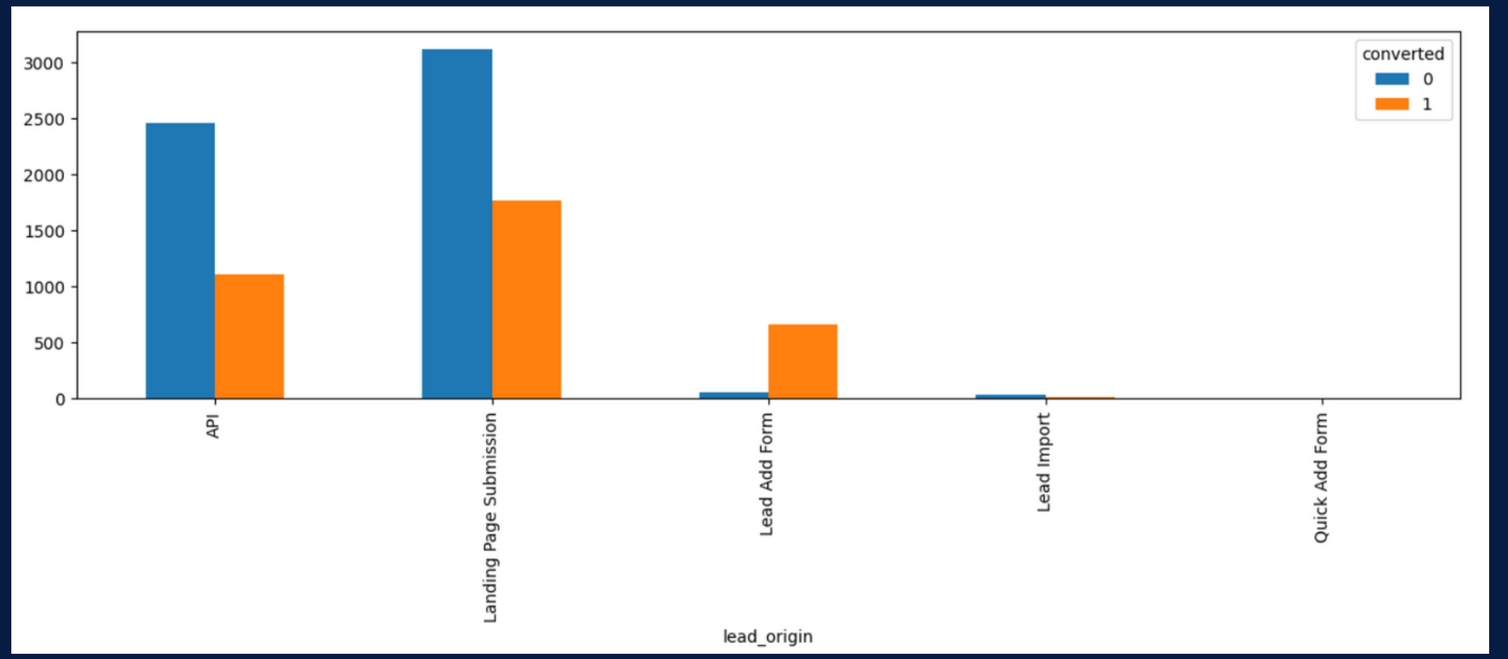


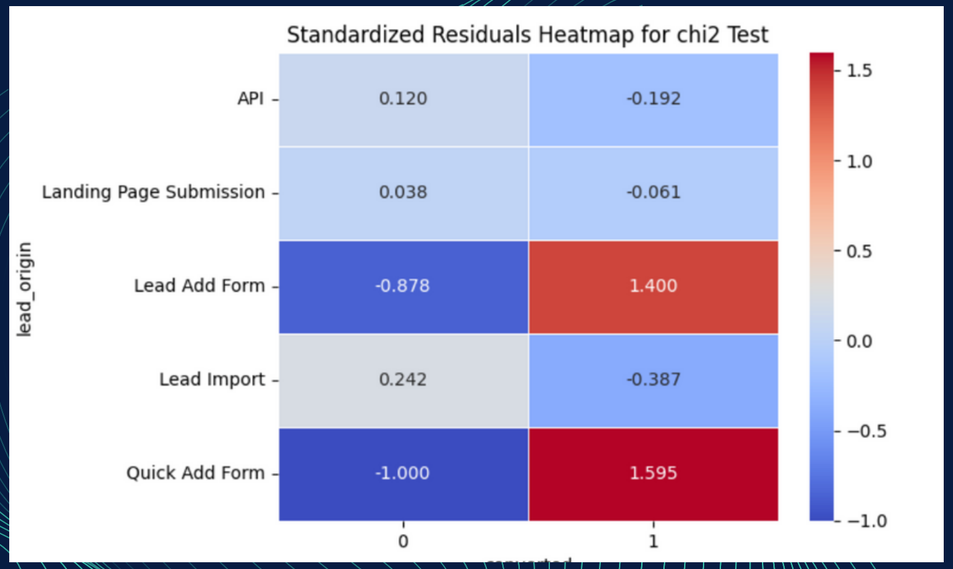


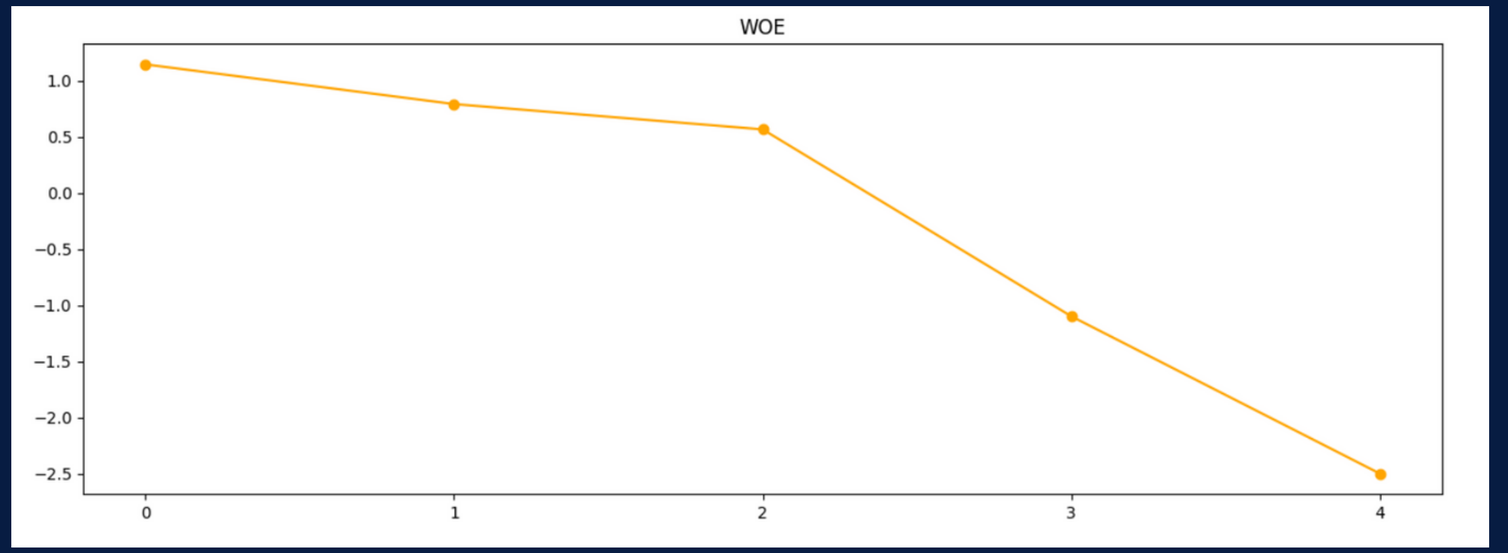


lead origin

- Observations: Landing page submission and API are dominant lead sources, but their conversion rates are not high. Although Lead Add Form has a high conversion rate, it constitutes only 7% of the data. Lead origin shows statistically significant, moderate correlation with the target; combining API and landing page submissions slightly improves correlations, but differences are not substantial.



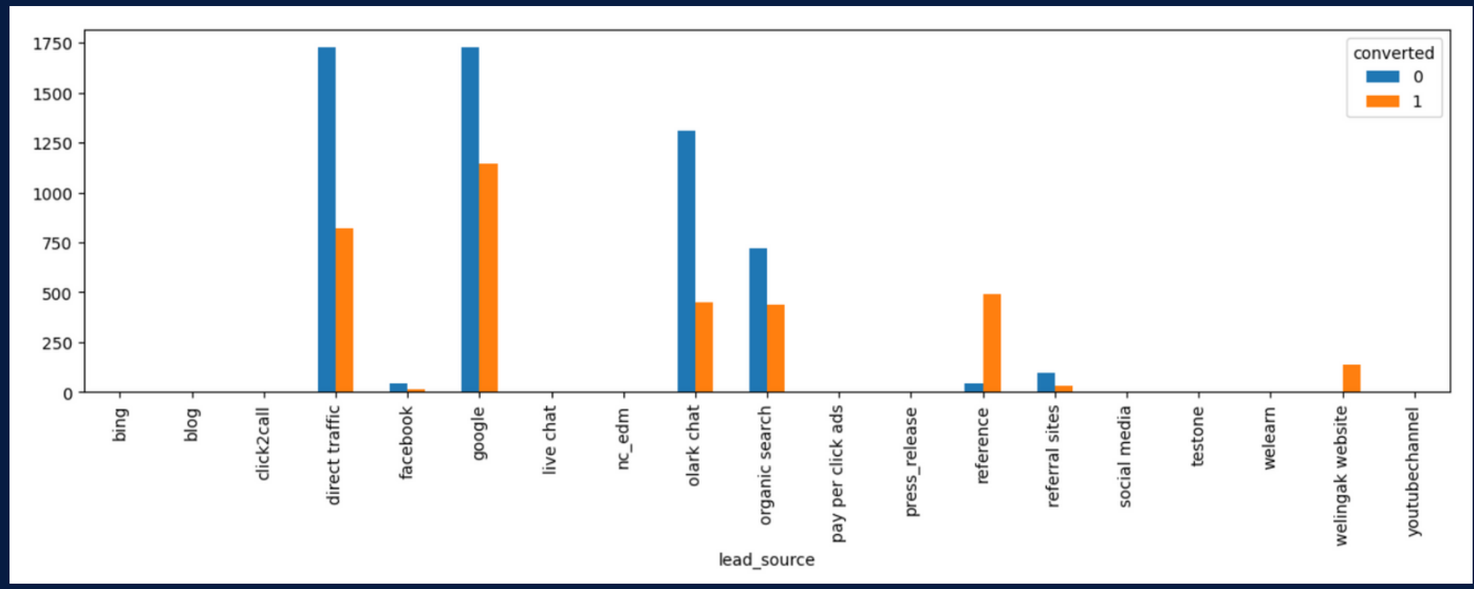


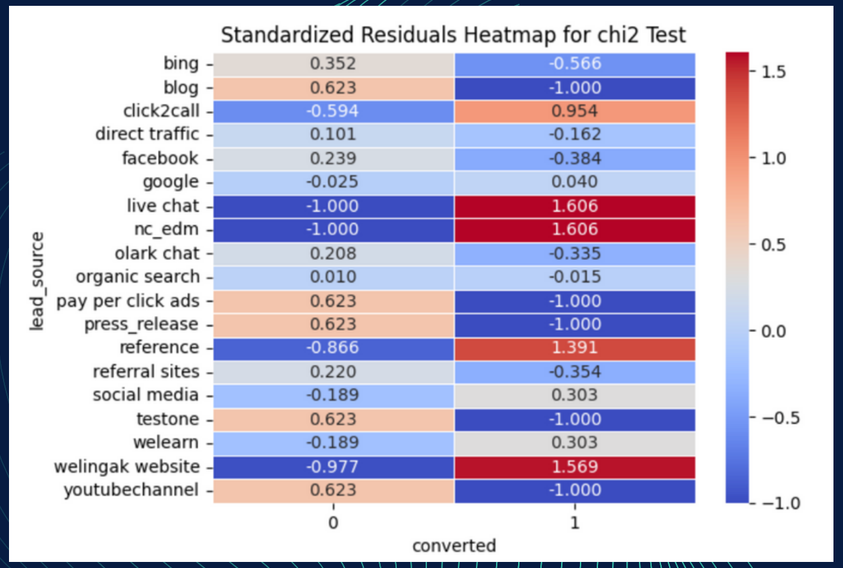


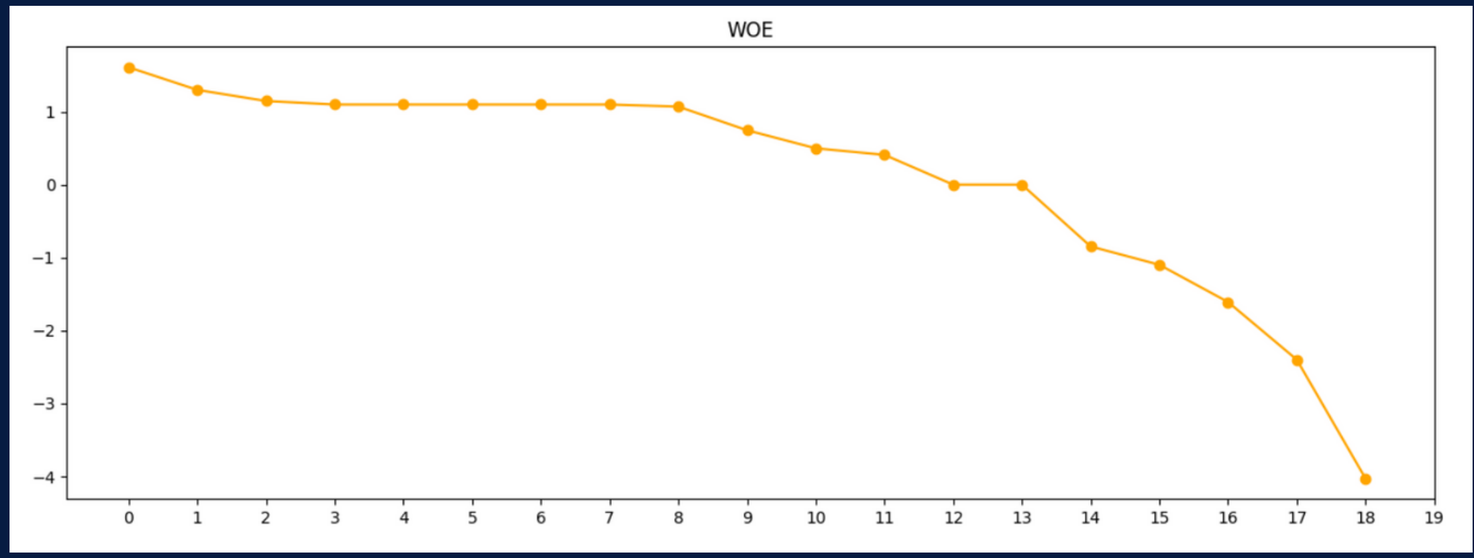
lead source

Observations: Predominantly, data points are in ['google','direct traffic','olark chat','organic search']. To reduce cardinality, consider grouping less frequent categories as 'others' or using Weight of Evidence (WOE) for categorization based on similarity. The feature exhibits a statistically significant, moderately correlated relationship with the target variable..

- The feature lead_source is moderately associated to the target
- The feature can better be transformed using iv values of the most suitable categories

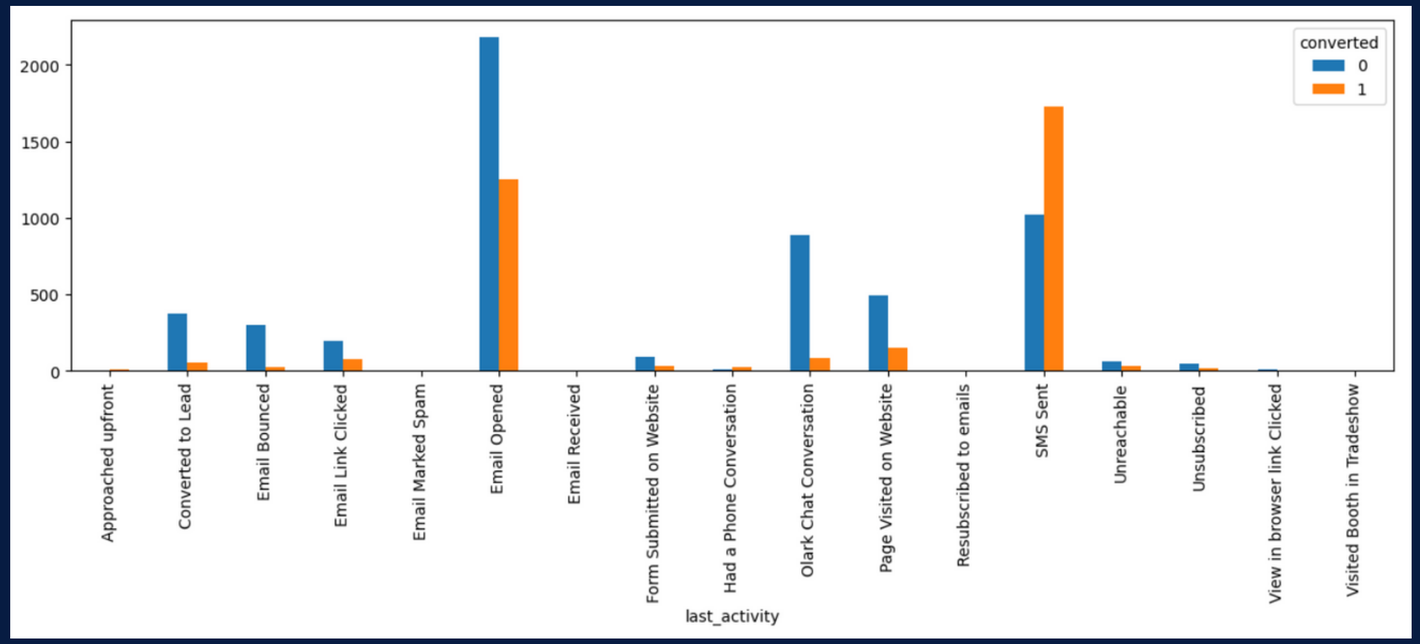


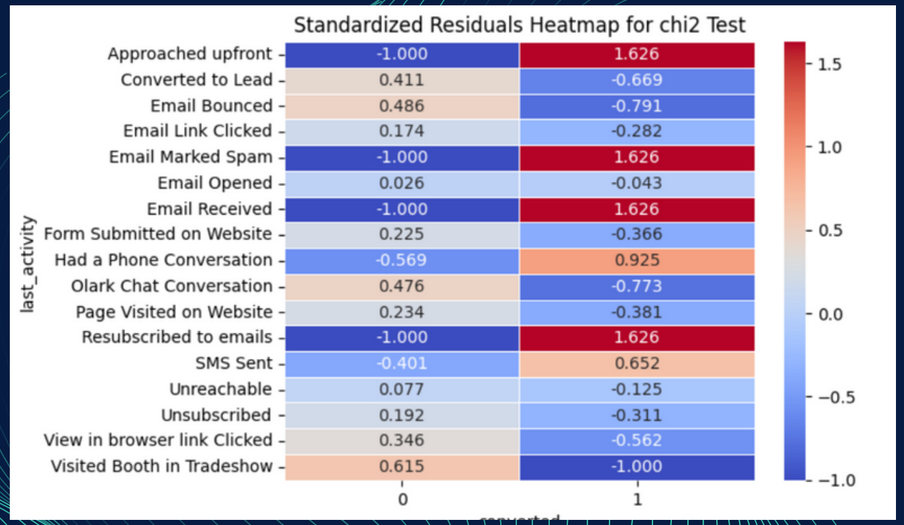


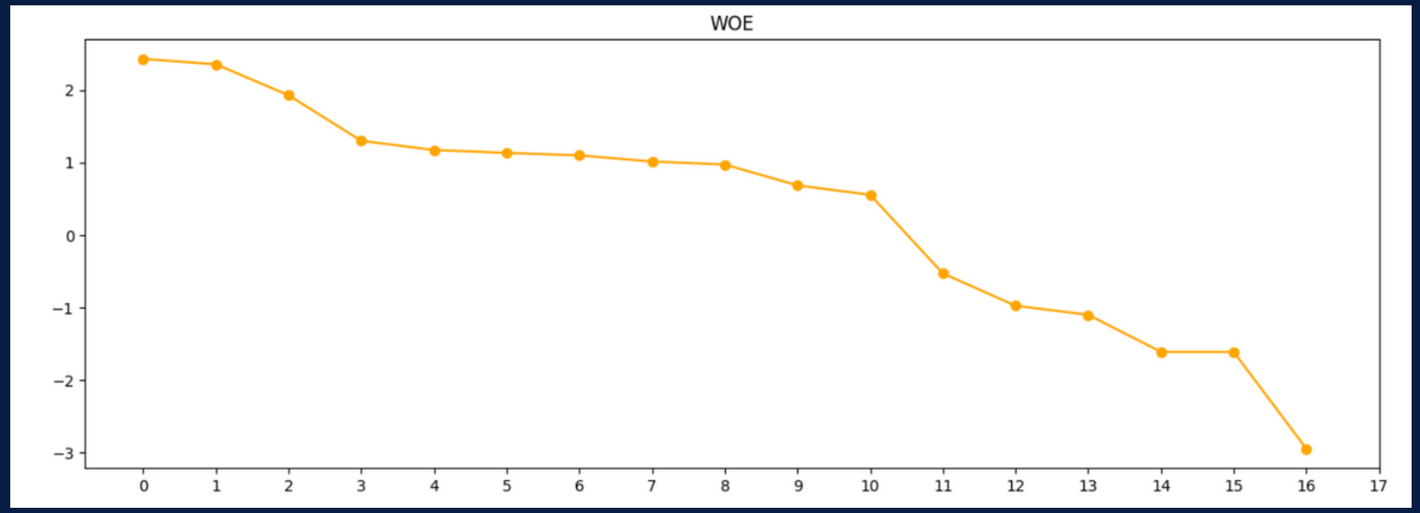


last activity

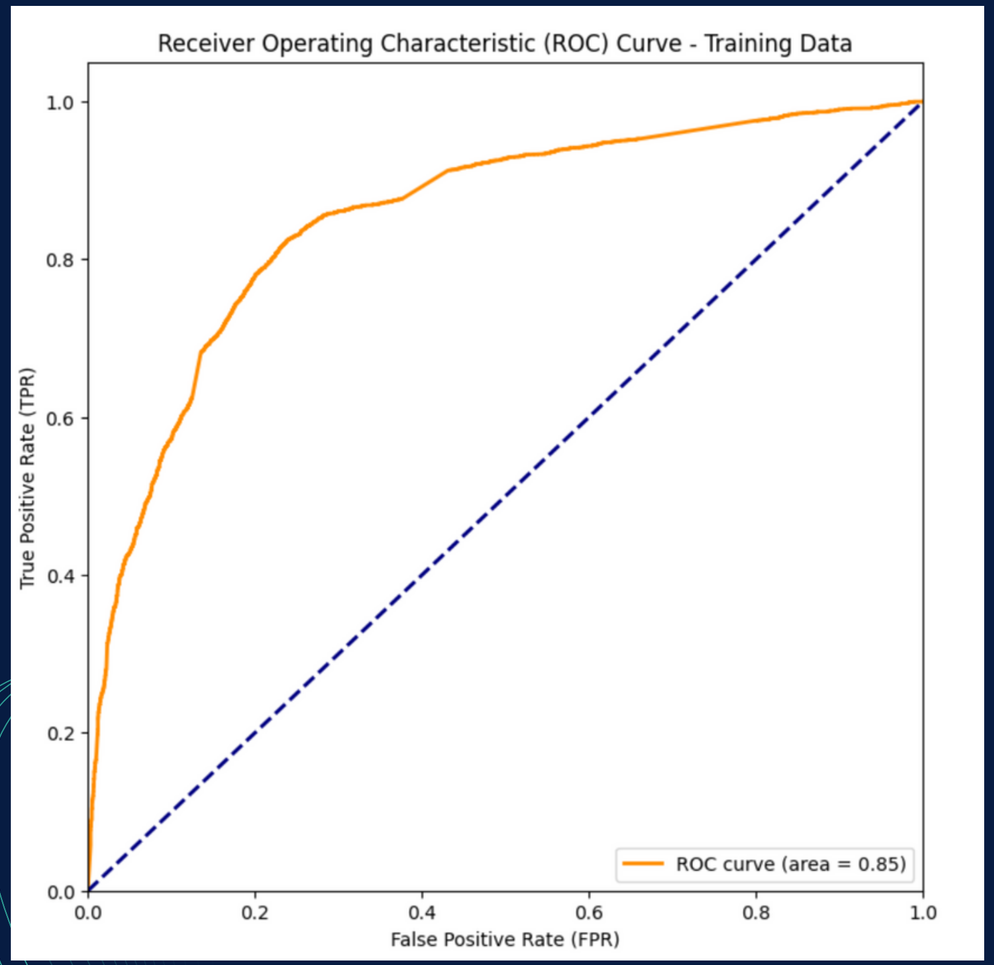
- Observations: Key contributors to the feature are 'last_activity' and 'Email Opened,' leading to high cardinality and abnormal chi2 distribution. Despite this, a moderately strong association with the target is observed. To address high cardinality, consider grouping less frequent categories as 'others' or using Weight of Evidence (WOE) for cardinality reduction.





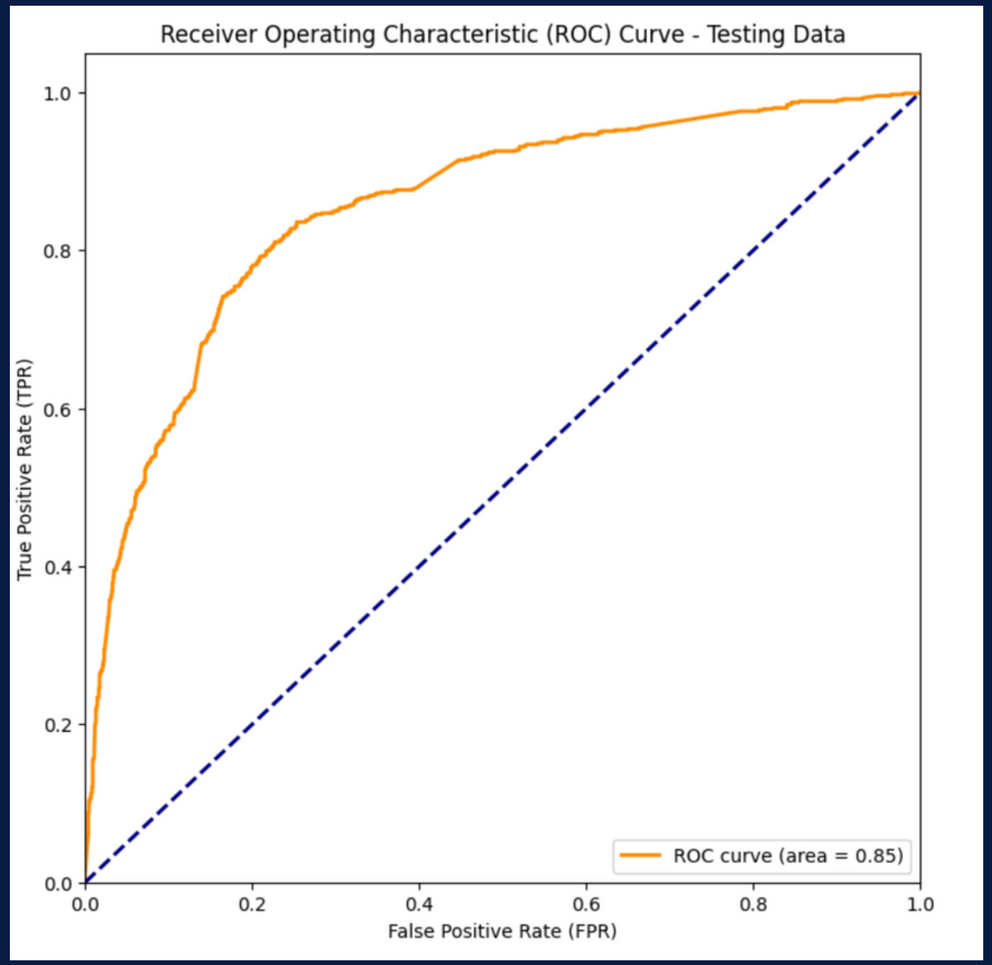


Model Evaluation



- Testing model evaluation...

The accuracy score is : 0.79
The recall score is : 0.79
The f1-score is : 0.75



- Final Model observation

Training model evaluation...
The accuracy score is : 0.79
The recall score is : 0.78
The f1-score is : 0.79

Our Team

Our team thrives on shared goals, with members actively engaging, pooling diverse skills, and fostering open communication for collective success. Collaboration ensures a dynamic synergy where each member's strengths contribute to a unified and effective outcome.

Nilotpal Malakar

Data Scientist

Prajakta Joshi

Data Scientist

Manisha

Data Scientist