Lead Scoring Assignment

Lead Scoring Assignment

I. Problem statement

II. Analysis Approach

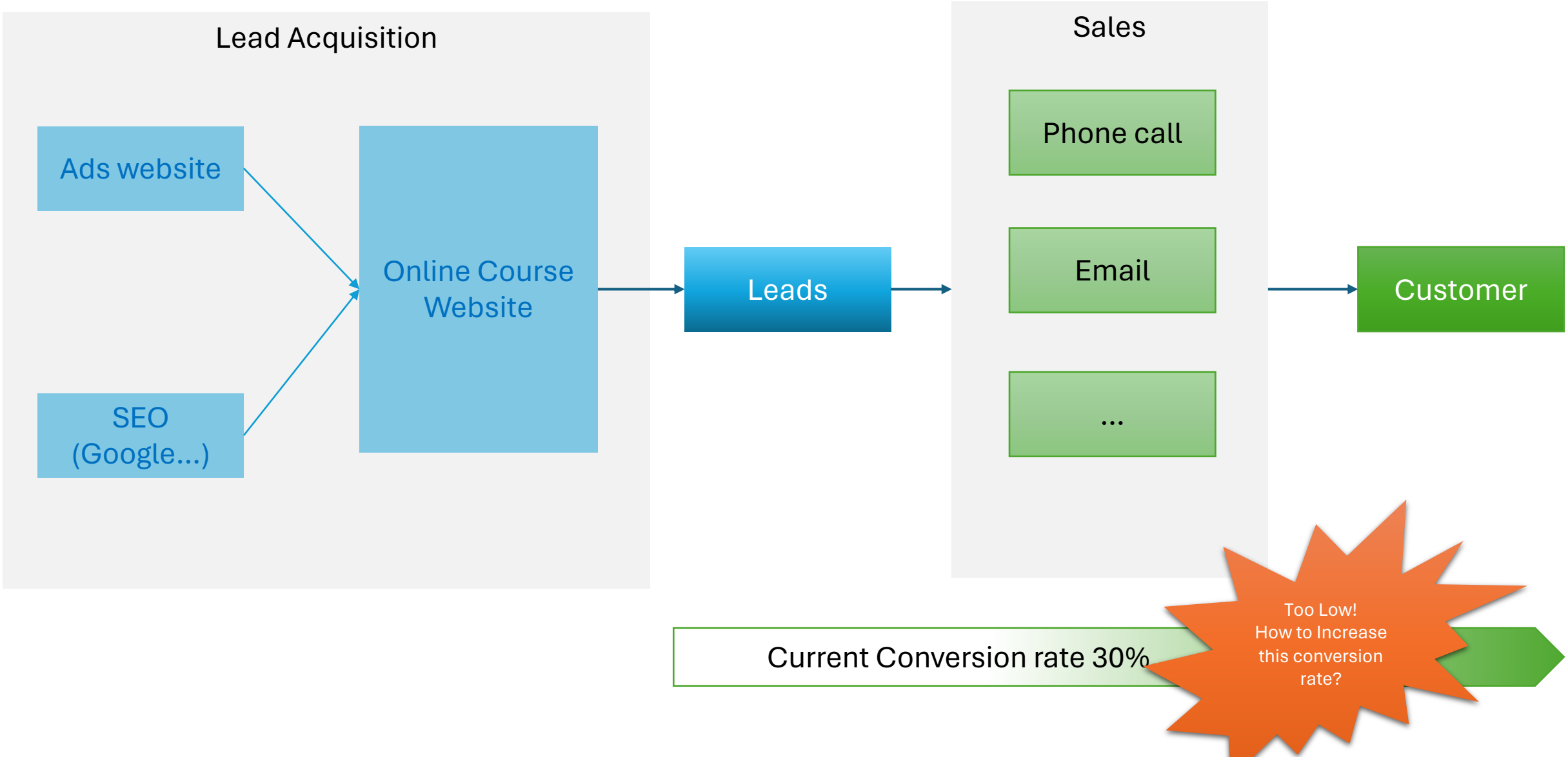
- EDA
- Data Visualization
- Build the logistic regression model
- Model evaluation

III. Interpret the results

I. Problem Statement

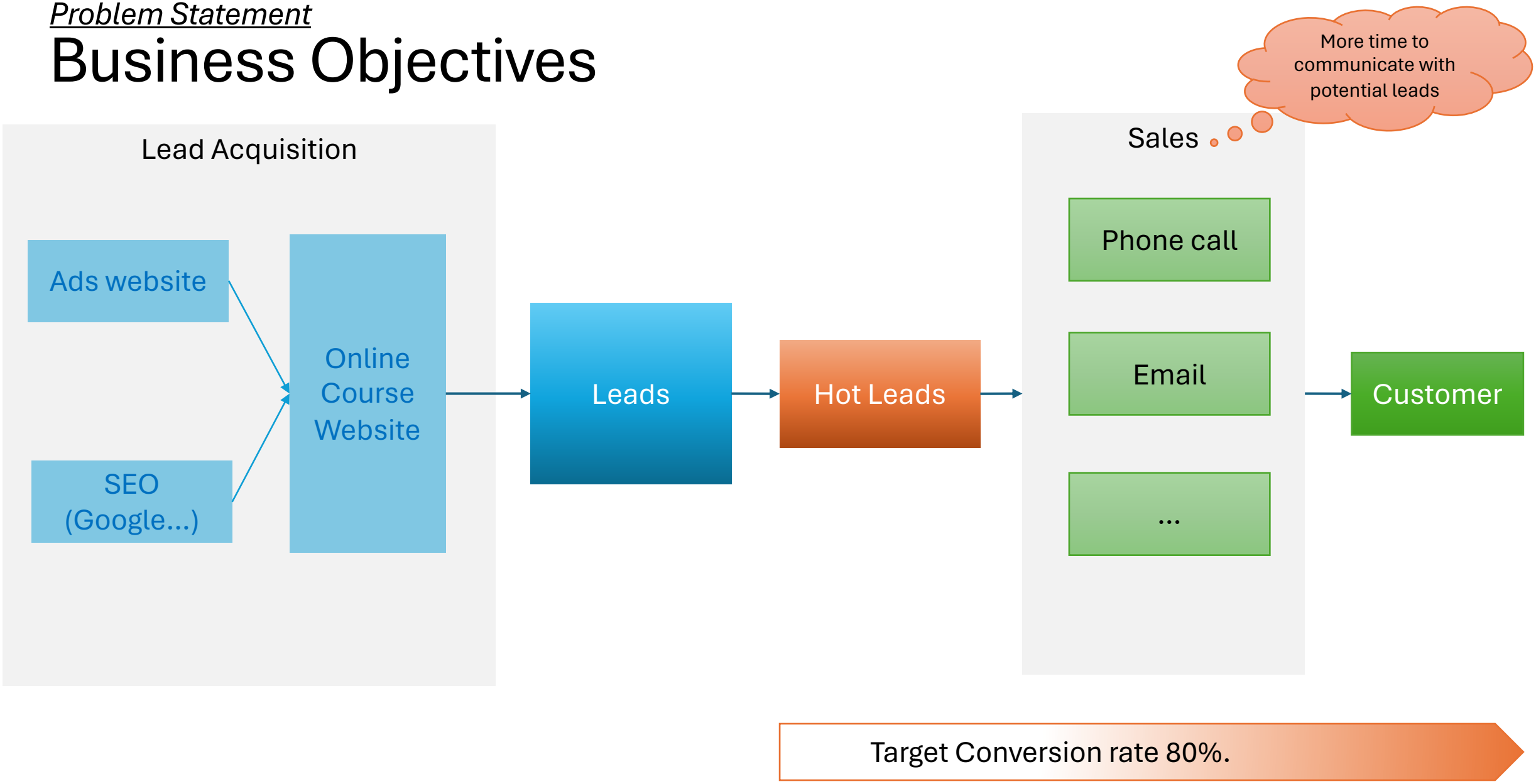
Problem Statement

Current Situation



Problem Statement

Business Objectives

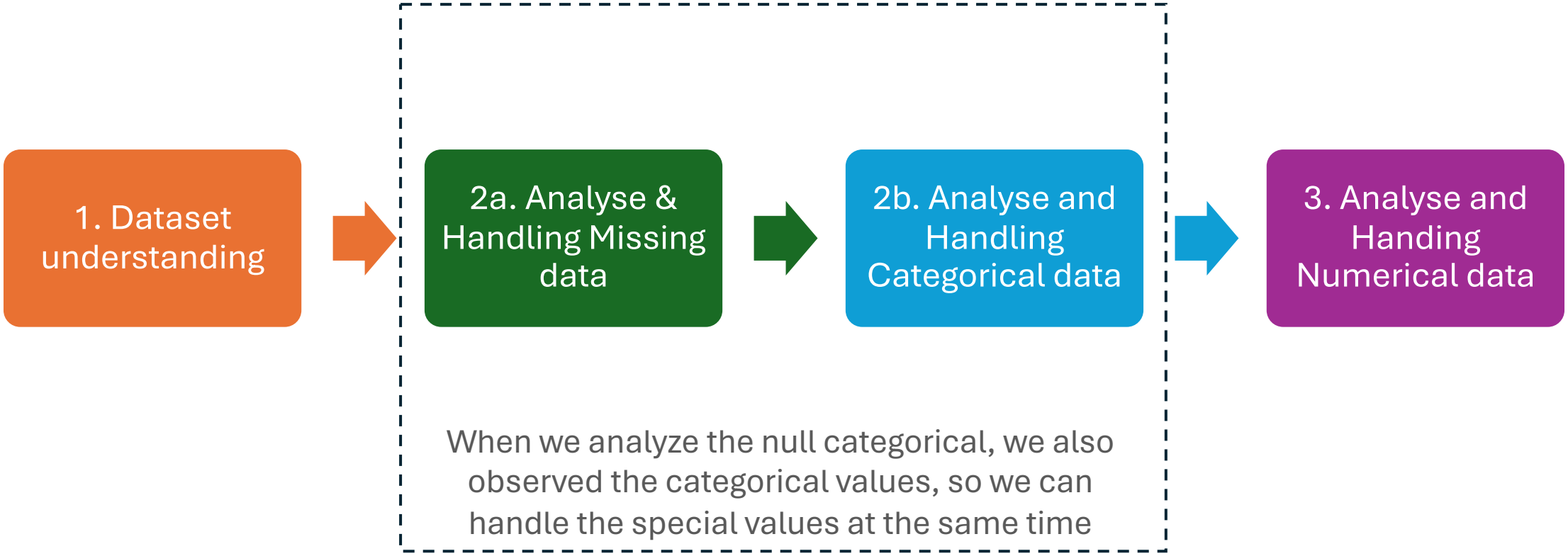


II. Analysis Approach

- EDA
- Data Visualization
- Build the logistic regression model
- Model evaluation

2. Analysis Approach

EDA Steps



2. Analysis Approach - EDA

Step 1: Dataset understanding

- Prospect ID
- Lead Number

Leads Identity

- Country
- City
- Specialization
- What is your current occupation

Demographics

- Lead Origin
- Lead Source
- How did you hear about X Education
- Search
- Magazine
- Newspaper Article
- Newspaper
- X Education Forums
- Digital Advertisement
- Through Recommendations

Lead Sources

- Do Not Email
- Do Not Call
- Receive More Updates About Our Courses
- Update me on Supply Chain Content
- Get updates on DM Content
- I agree to pay the amount through cheque

Consents

- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- Last Activity
- Last Notable Activity

Lead Behaviours

- What matters most to you in choosing a course
- A free copy of Mastering The Interview

Lead Interests

- Tags
- Lead Quality
- Lead Profile
- Asymmetrique Activity Index
- Asymmetrique Profile Index
- Asymmetrique Activity Score
- Asymmetrique Profile Score

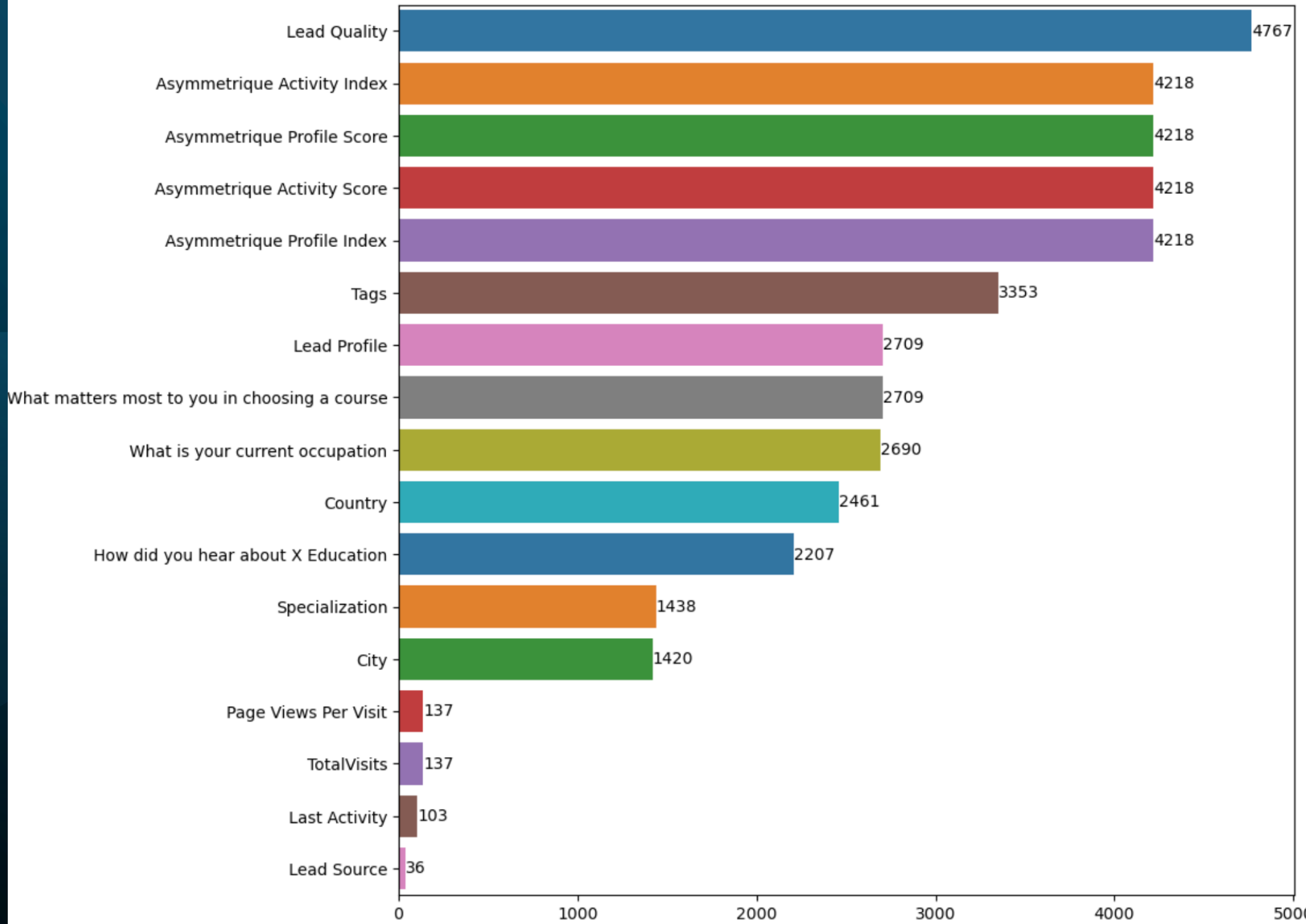
Sales Engagements/Evaluations

- Converted

Target Variable

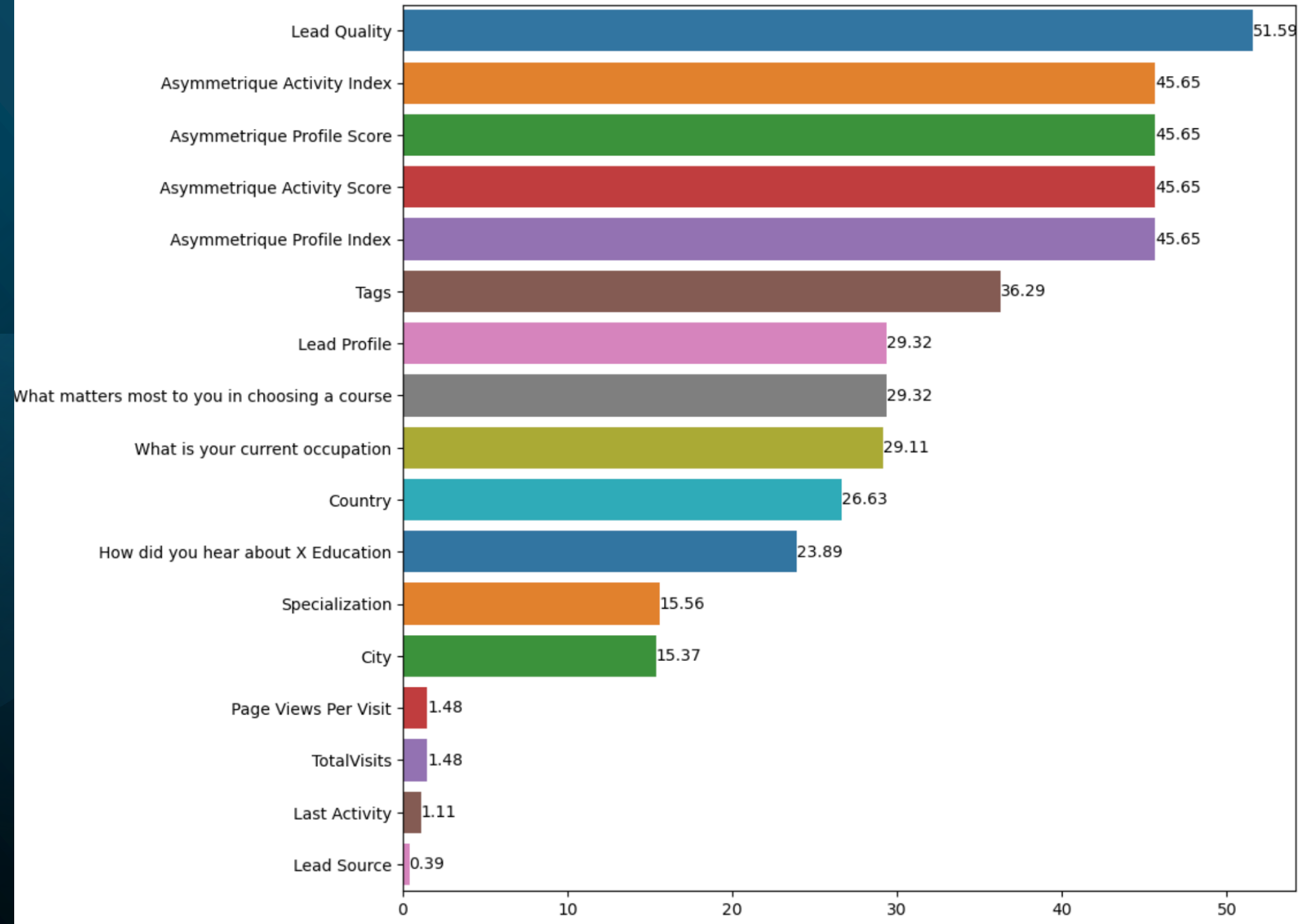
2. Analysis Approach - EDA

Step 1: Dataset understanding - Missing Data



2. Analysis Approach - EDA

Step 1: Dataset understanding - Missing Data



2. Analysis Approach - EDA

Step 2: Handling missing data and categorical data

Missing > 45%
(both categorical & numerical)

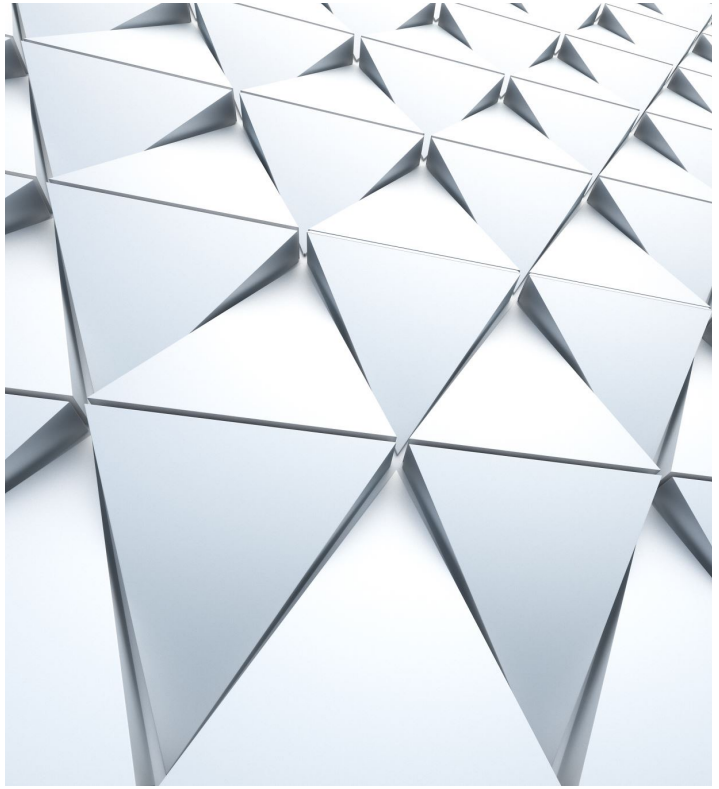
- Missing too many data, impact to model result
- Drop the column

Categorical variables missing >2% (no numerical field)

- Fill null value with “Other”
- Replace “Select” by “Other” (Select value considered as null as user has not picked any value)
- Values with small counts or small percentage should be grouped to a new categorical value (eg: combine with “Other”)

Missing <2%
(both categorical & numerical)

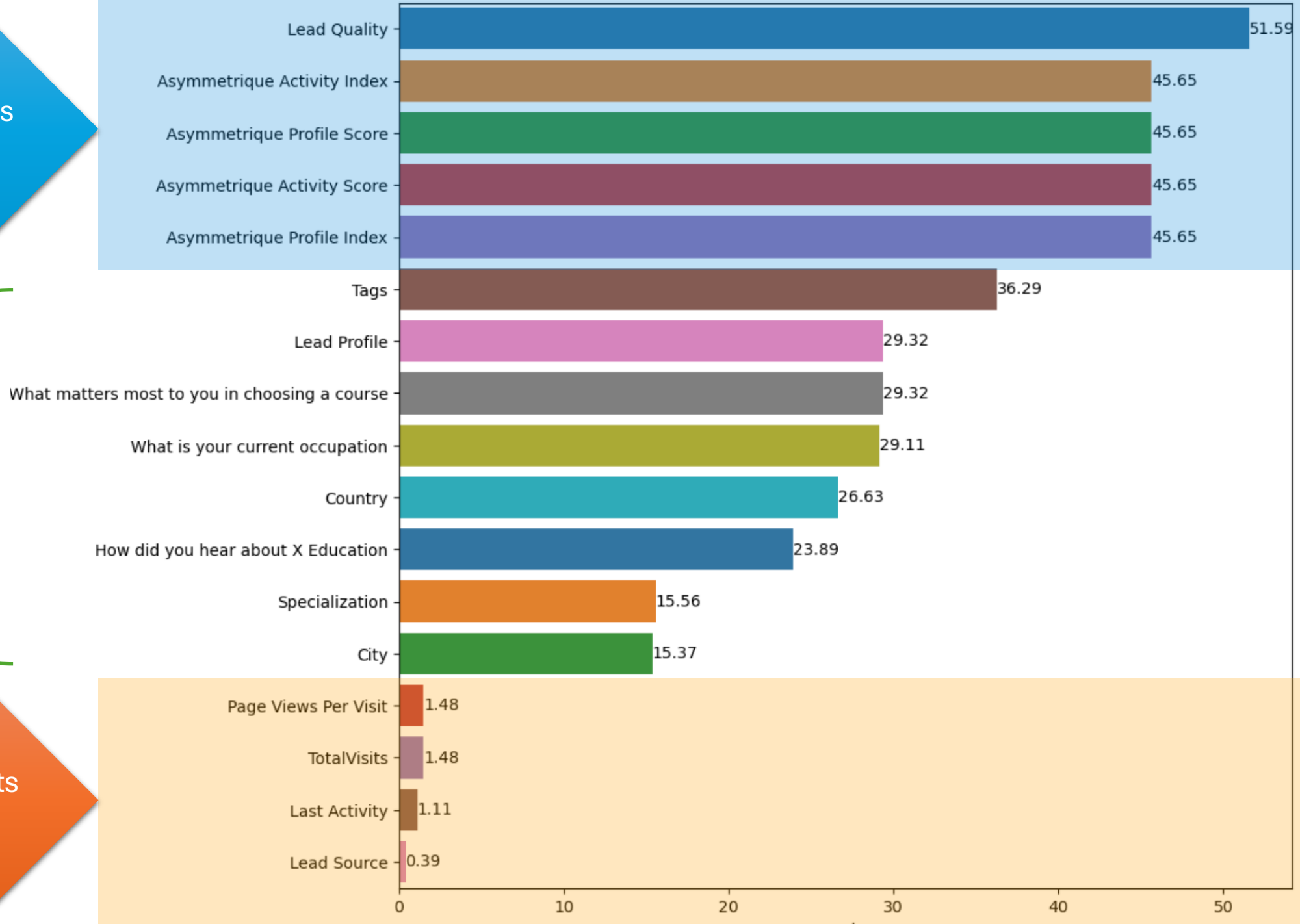
- The missing count is too small, we can drop these datapoint/row without impact to ML result
- Drop the row



Drop the fields/ columns which missing >45%

Handling one by one

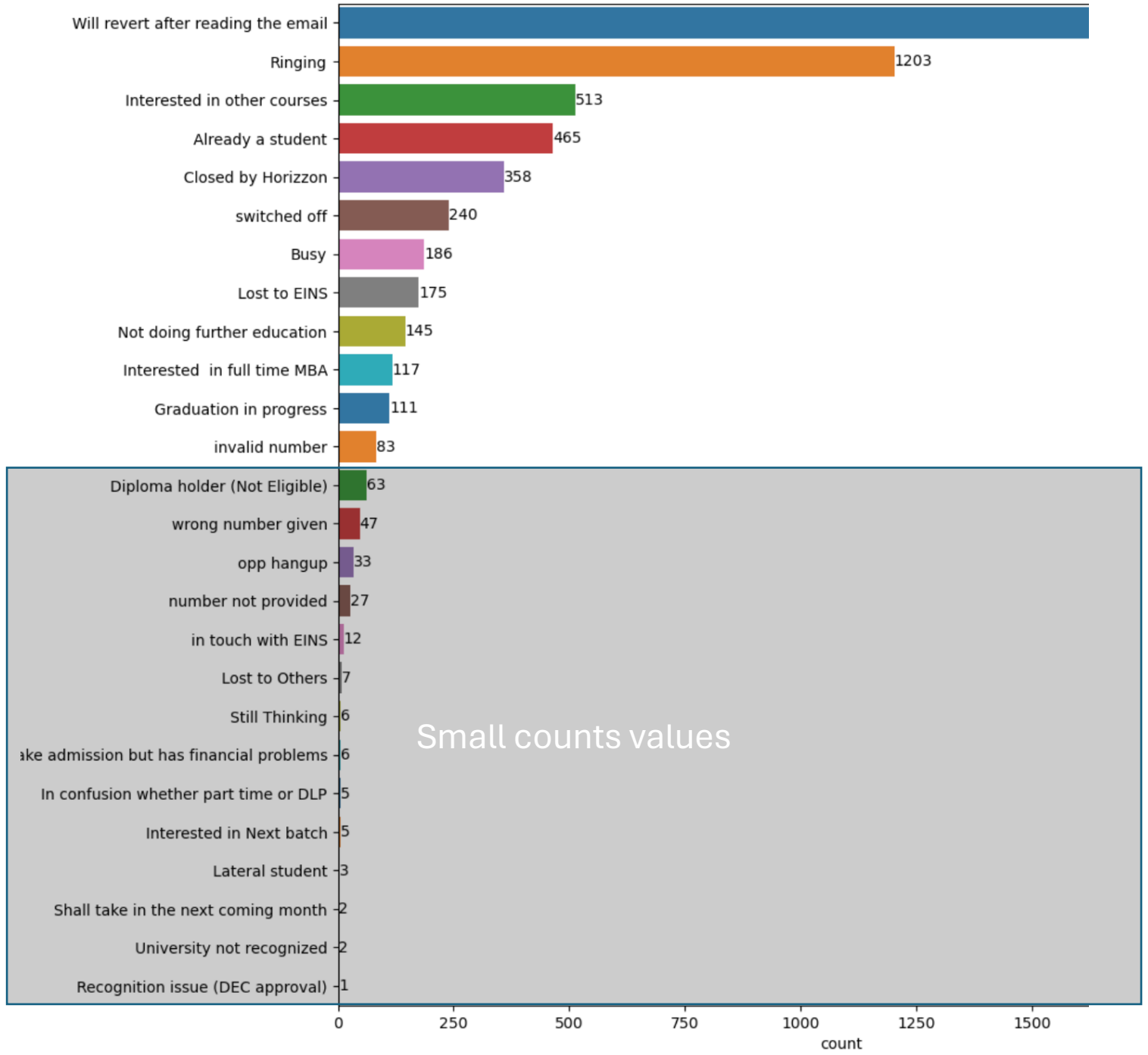
Drop the row/data points which missing <2%



2. Analysis Approach - EDA

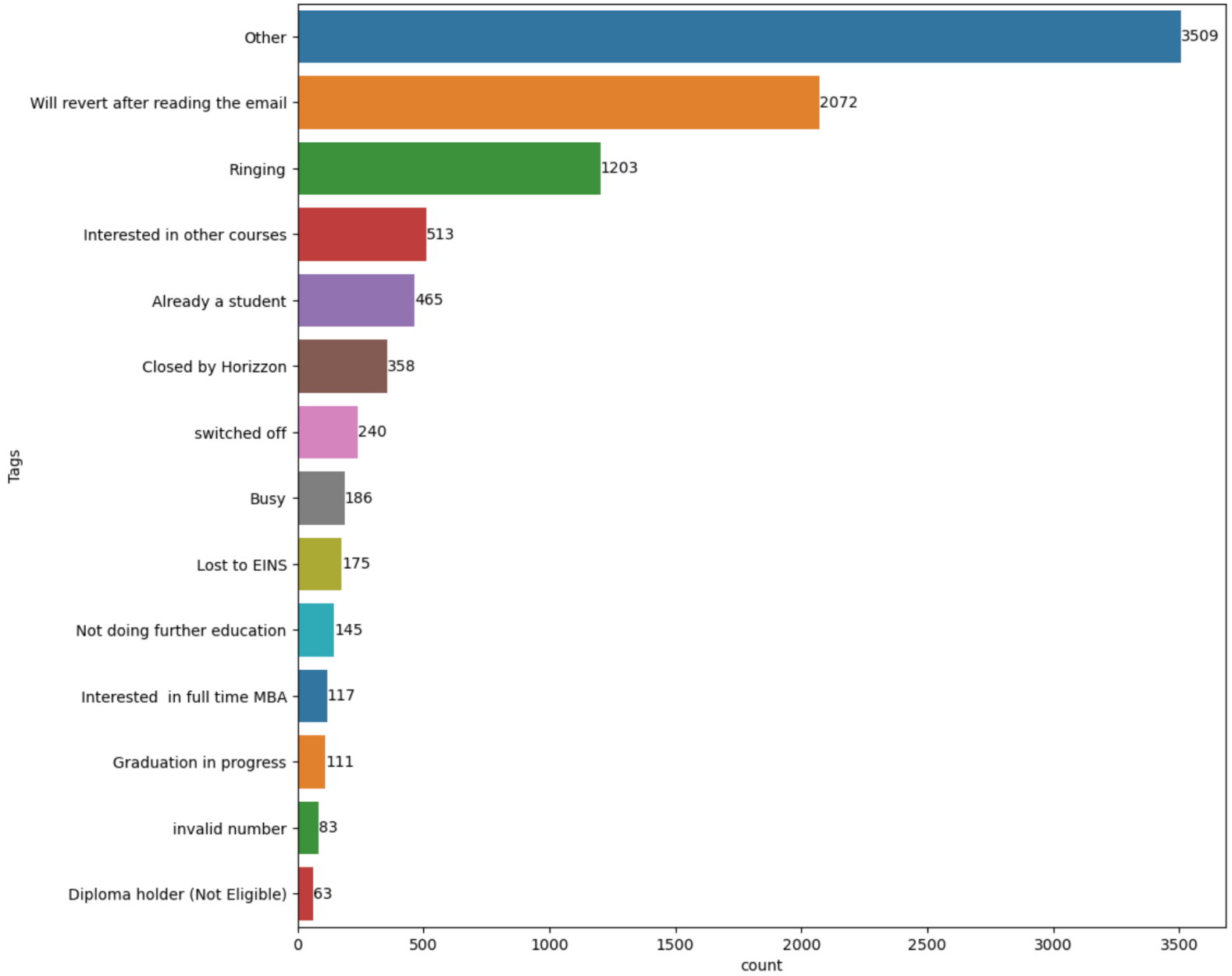
Step 2: Handling categorical field “Tags” (missing 36%)

- 36% missing is high percentage, if we drop these datapoints, it may impact to the overall ML result. So we will **replace NULL by “Other”** for further analysis
- The small count value (<1%) can be **grouped together to one group**. As the percentage is too small, we can group to above “Other” group, it’s not impact to the overall ML

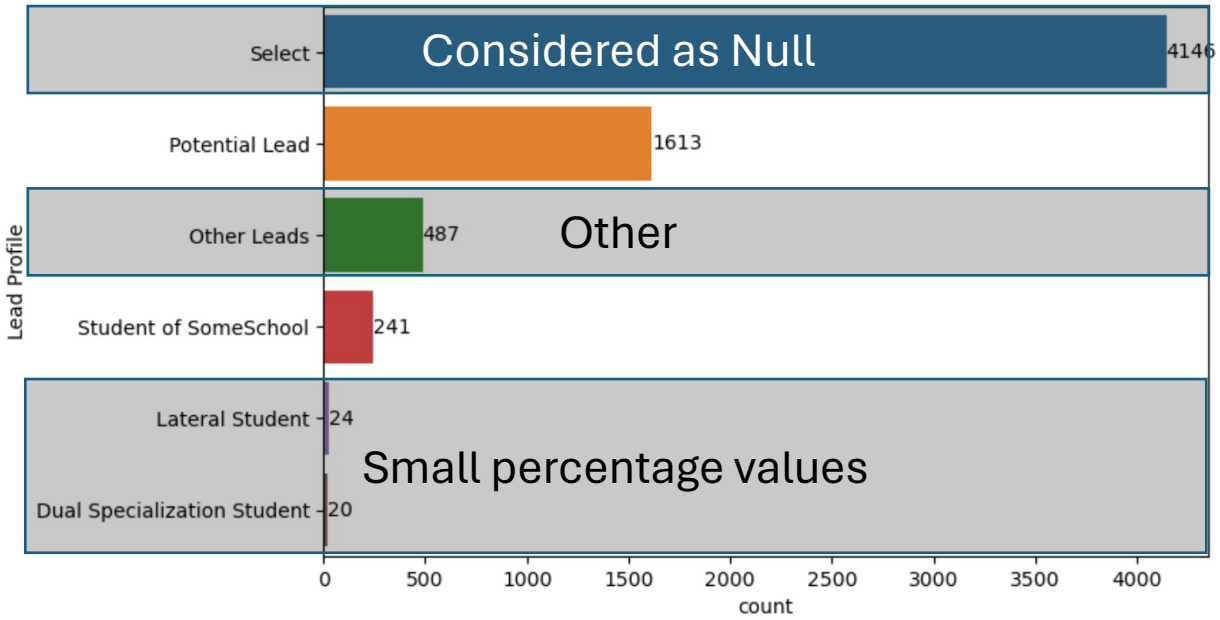


2. Analysis Approach - EDA

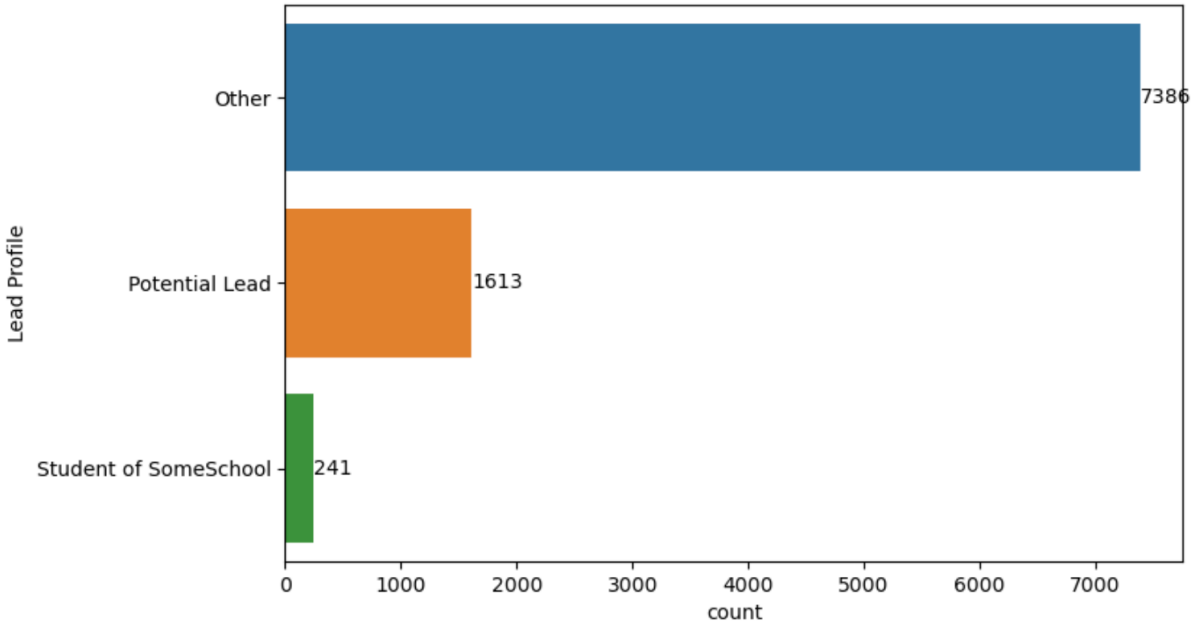
Step 2: Handling categorical field “Tags” (After handling)



Step 2: Handling categorical field “Lead Profile” (missing 29%)



Before handling missing values and cleaning the data



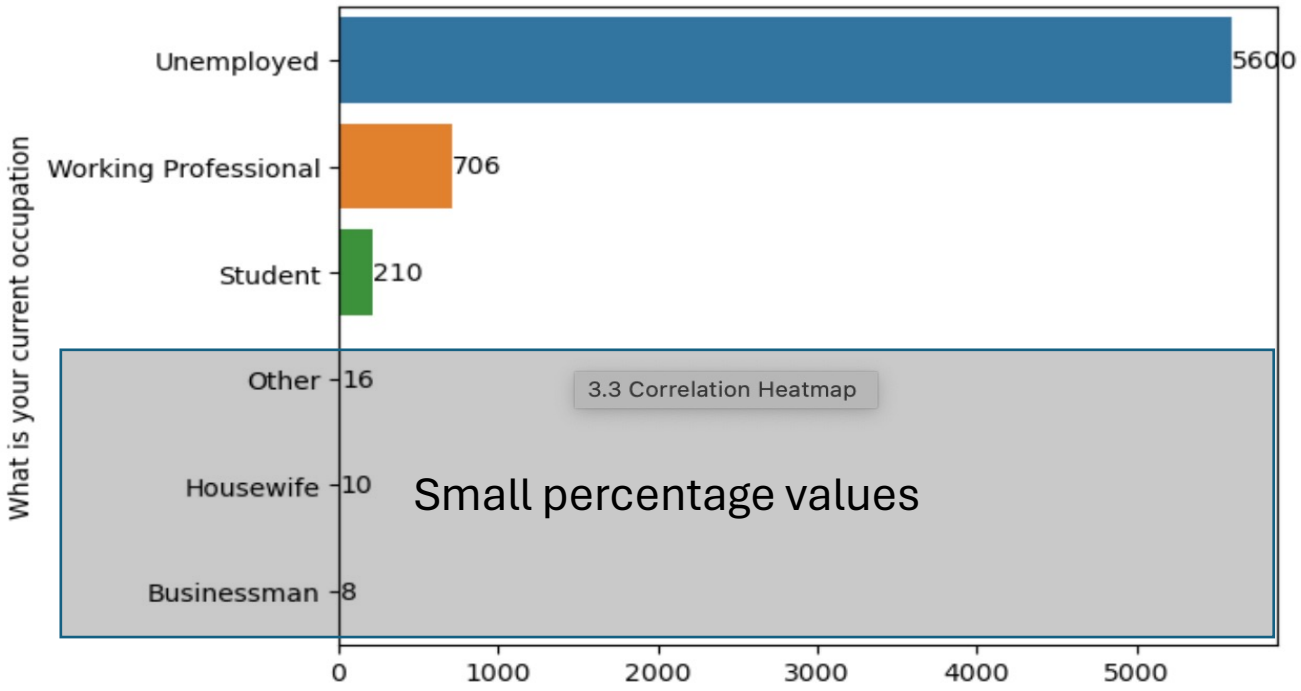
After handling missing values and cleaning the data

There are some problems with this categorical field:

- Null data 29%
- “Select” value means: user has not chosen any values for this field. This is considered as null value
- “Other Leads”: no specific meaning in this case.
- Other very small percentage values has no impact to the ML model.

→All of above cases can be replaced as “Other” so that we can reduce the dummy variables for models

Step 2: Handling categorical field “What is your current occupation”(missing 29%)



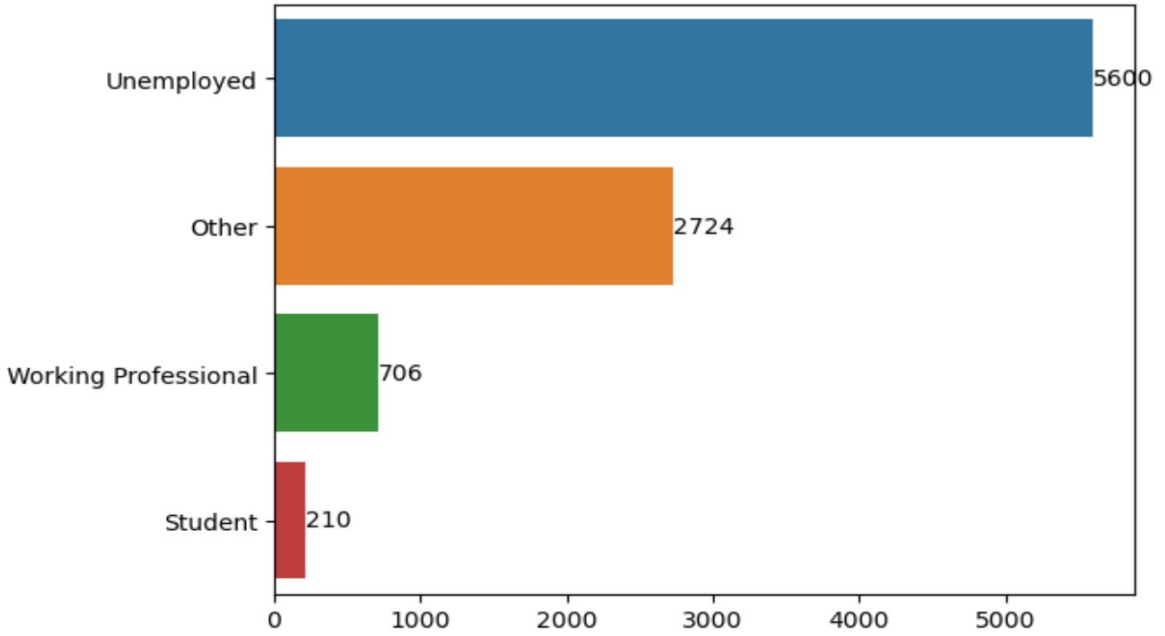
Before handling missing values and cleaning the data

There are some problems with this categorical field:

- Null data 29%

- Small percentage values (Businessman, Housewife, Other) can be group to “Other” as the value count is too small

→All of above cases can be replaced as “Other” so that we can reduce the dummy variables for models



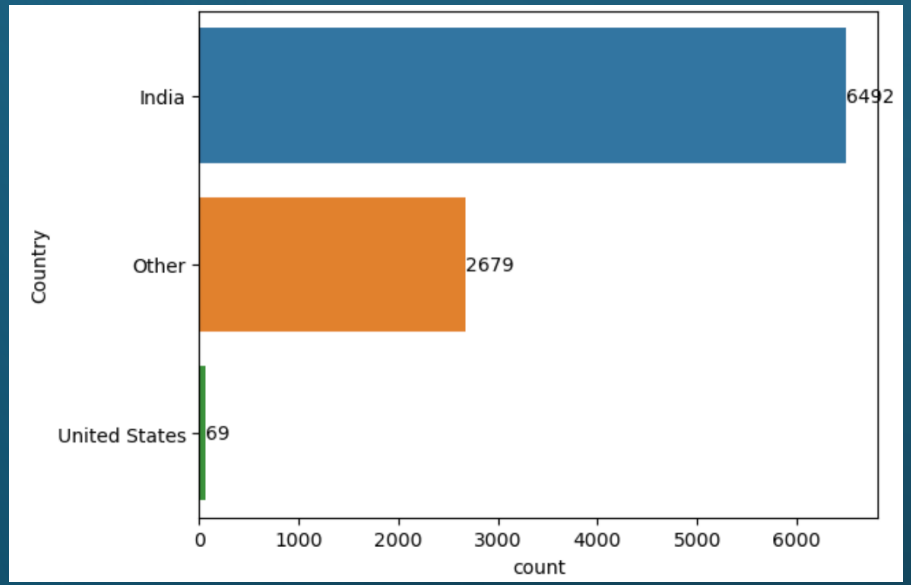
After handling missing values and cleaning the data

Country	
India	6492
United States	69
United Arab Emirates	53
Singapore	24
Saudi Arabia	21
United Kingdom	15
Australia	13
Qatar	10
Hong Kong	7
Bahrain	7
Oman	6
France	6
unknown	5
South Africa	4
Nigeria	4
Germany	4
Kuwait	4
Canada	4
Sweden	3
China	2
Asia/Pacific Region	2
Uganda	2
Bangladesh	2
Italy	2
Belgium	2
Netherlands	2
Ghana	2
Philippines	2
Russia	1
Switzerland	1
Vietnam	1
Denmark	1
Tanzania	1
Liberia	1
Malaysia	1
Kenya	1
Sri Lanka	1
Indonesia	1

Drop these values

Step 2: Handling categorical field “Country” categorical data (missing 27%)

- The null values are 27% which is quite high. If we drop these rows, we may lose many significant information.
  - We will replace Null with “Other” to keep these datapoints for ML
- There are many countries, most of them are ”India”. The remaining countries have small number of leads
  - Opt 1: We can drop this column
  - Opt 2: We can keep Top 2 countries, group these values to one group “Other” (same as Null’s replacement group) to reduce the number of dummy variables and keep the possible important information for ML
- For now we follow option 2. And below chart are the country data after handling



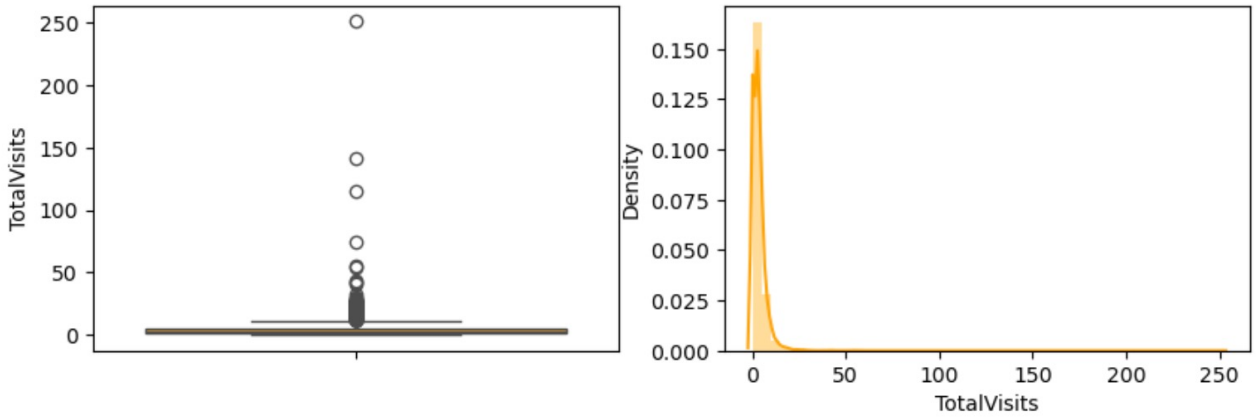
Step 2: Handling remained categorical fields

We do similar for remaining missing fields:

No	Field name	Analysis	Actions
1	What matters most to you in choosing a course	- Missing 29% which is quite big. If we drop these rows, we may lose important information	Replace Null by “Other” categorical
2	How did you hear about X Education	- Missing 24%, cover ¼ data set. - “Select” value mean that user has not selected the value from DropDown control. - Some small percentage values (eg: email, sms) can be consider dropping or keep it without impact. For now we will keep it	Replace Null and “Select” by “Other”
3	Specialization	- Missing 15%. This percentage is in average (not too much, too less). We can drop it or impute Null as “Other” - “Select” value mean that user has not selected the value from DropDown control.	Replace Null and “Select” by “Other”
4	City		

2. Analysis Approach - EDA
Step 3: Handling numerical data - TotalVisits

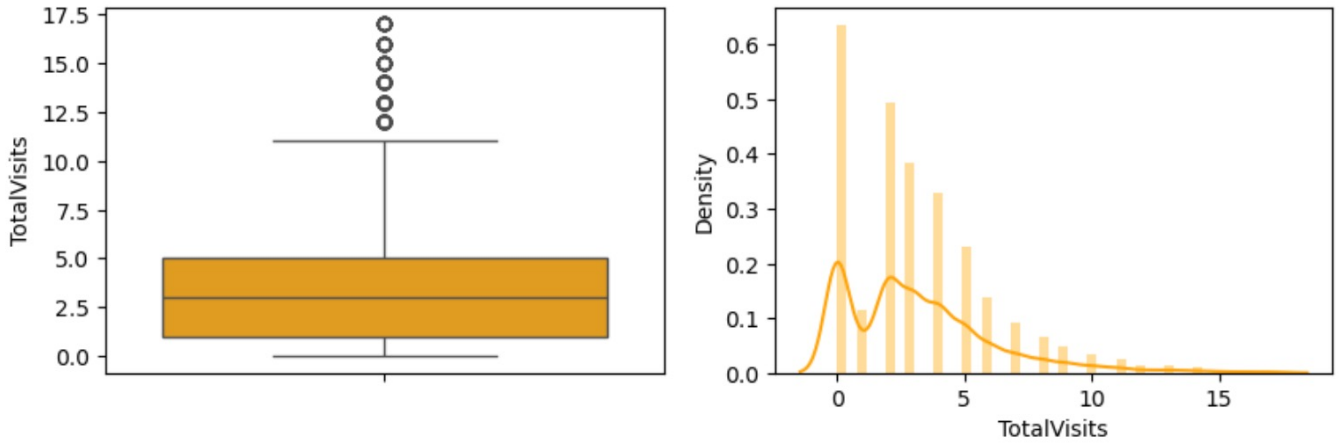
Boxplot showing that there're many outliers datapoints



Quantile 99th is very faraway (17) from max value (251)

Quantile	Value
0.95	10
0.98	13
0.99	17
1	251

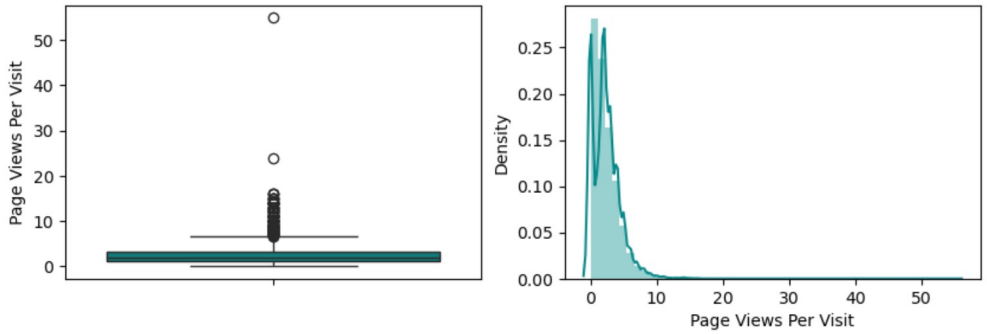
After remove outliers > 99th percentile



Some datapoints seem be outliers but it's not too far vs. original datapoints. This is acceptable!

2. Analysis Approach - EDA
Step 3: Handling numerical data – Page Views Per Visit

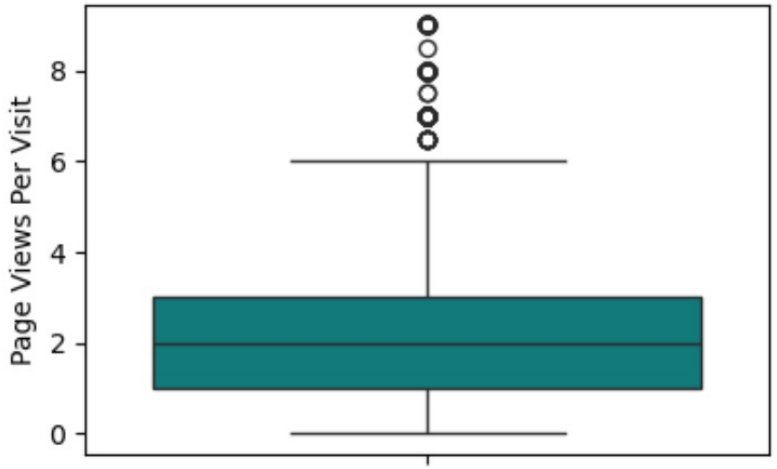
Boxplot showing that there're many outliers datapoints

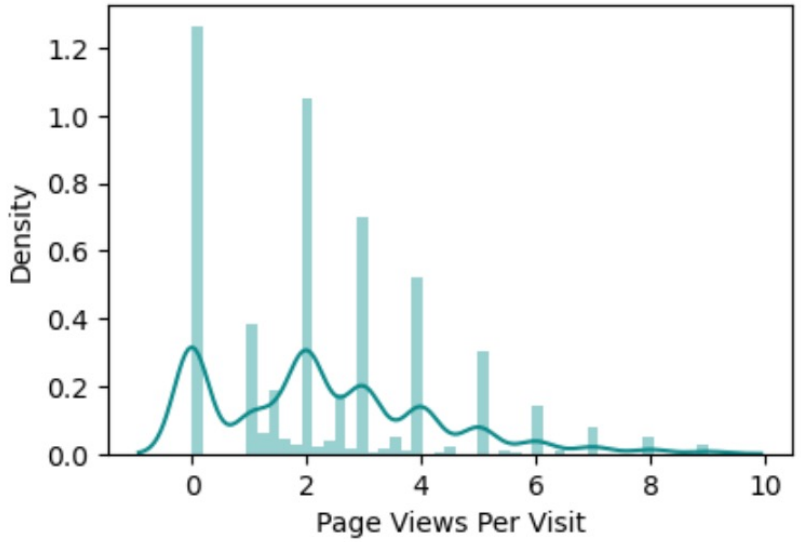


Quantile 99th is very faraway (9) from max value (55)

Quantile	Value
0.9	5
0.95	6
0.98	8
0.99	9
1	55

After remove outliers > 99th percentile

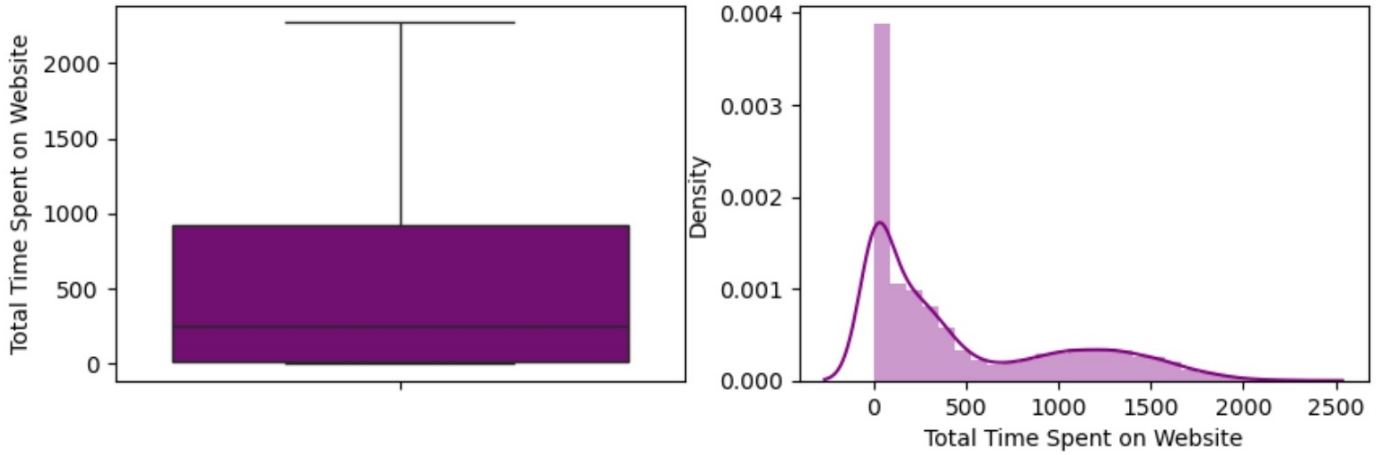




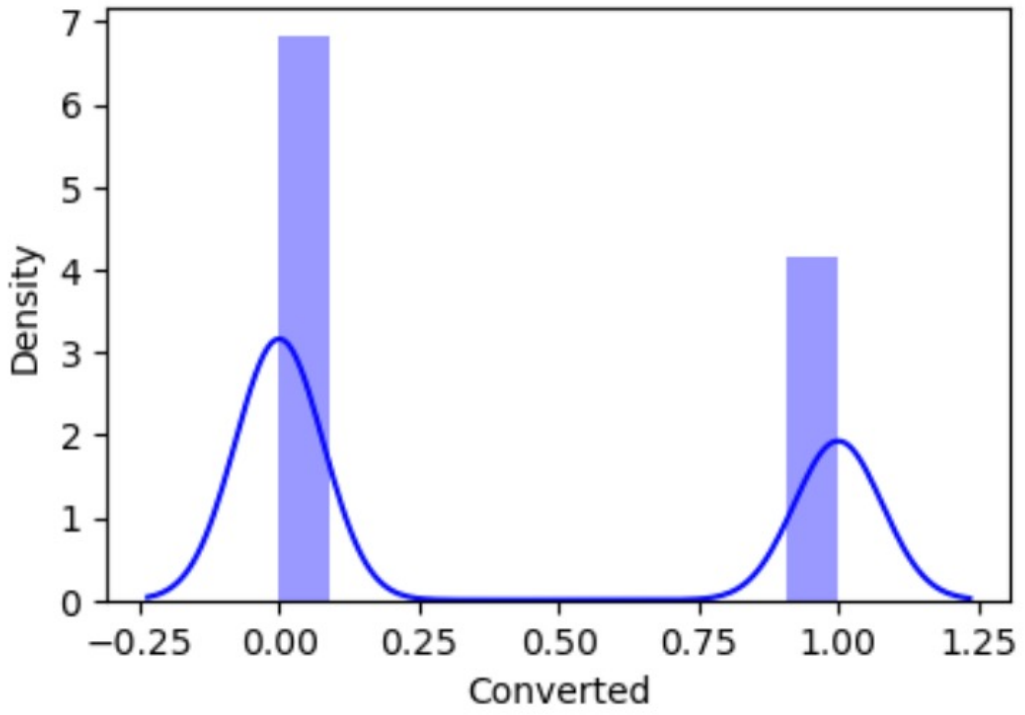
Some datapoints seem be outliers but it's not too far vs. original datapoints. This is acceptable!

2. Analysis Approach - EDA
Step 3: Handling remained numerical fields

"Total Time Spent on Website" has no outliers



Converted field is actually categorical data (1/0). There is no outliers

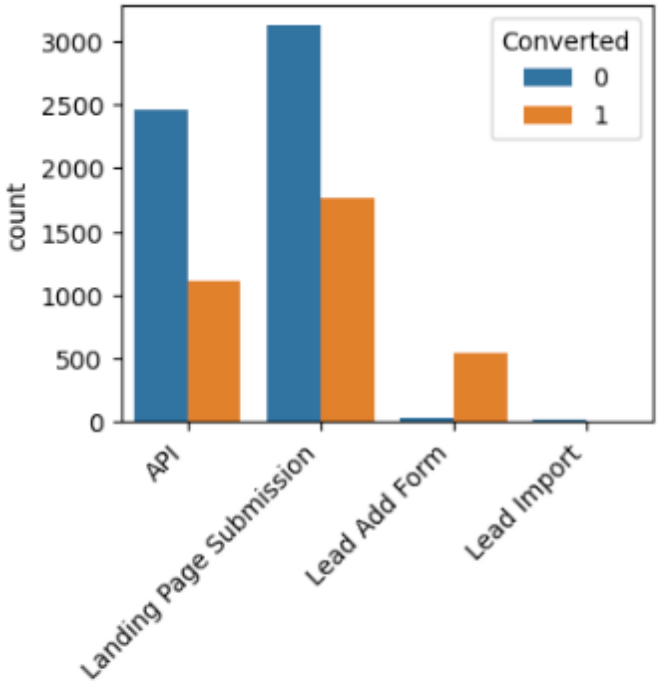


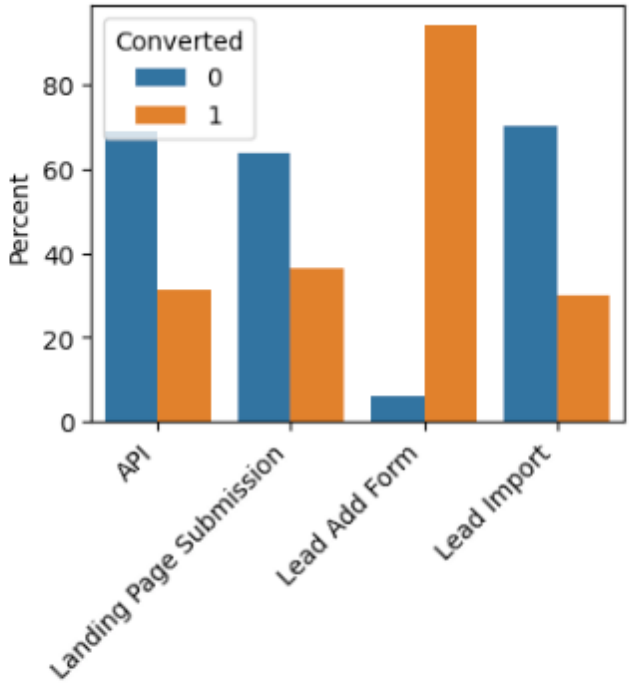
II. Analysis Approach

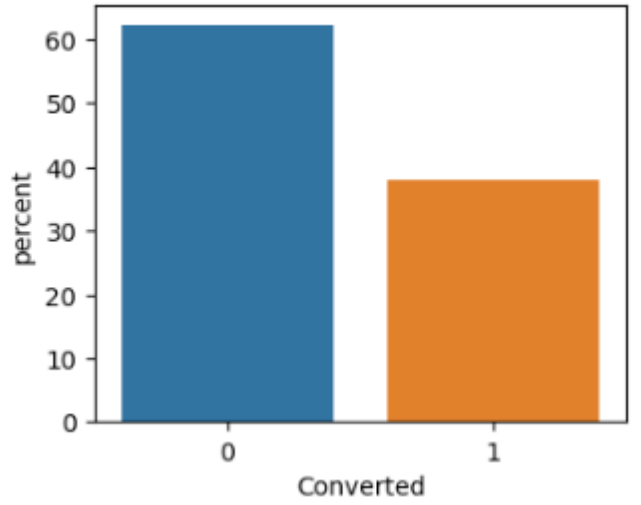
- EDA
- Data Visualization
- Build the logistic regression model
- Model evaluation

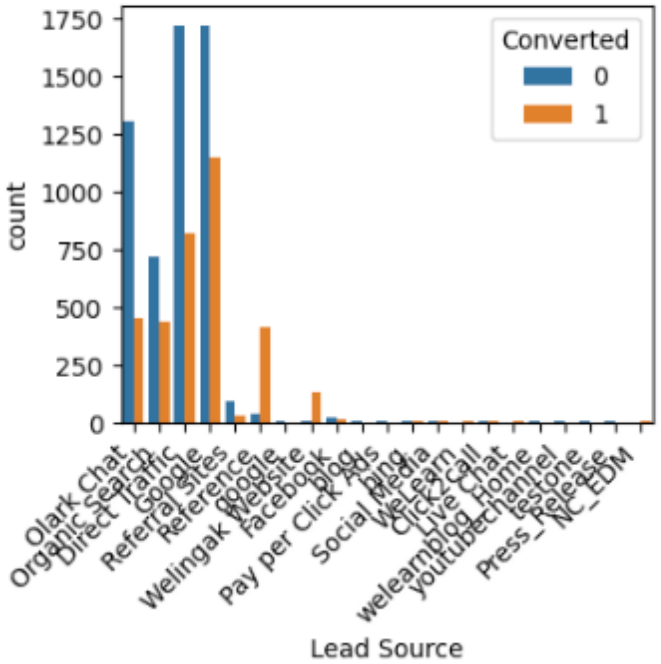
Visualize Categorical Data

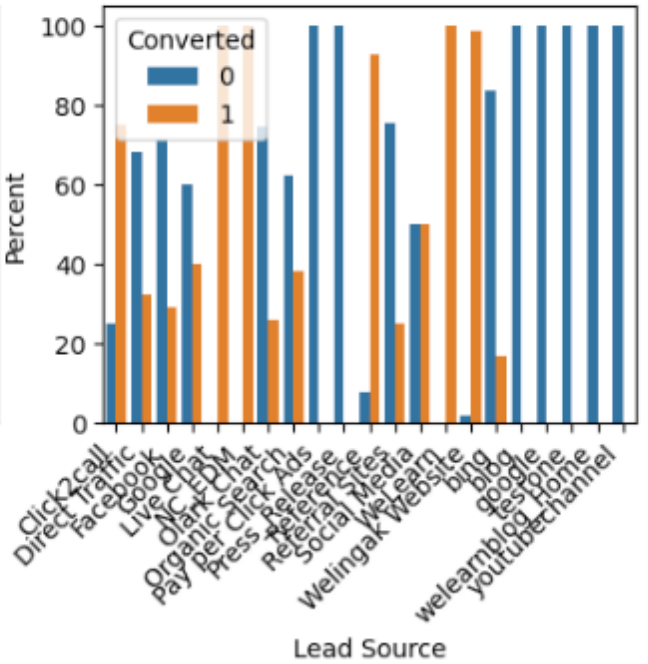
- Lead Origin
- Lead Source

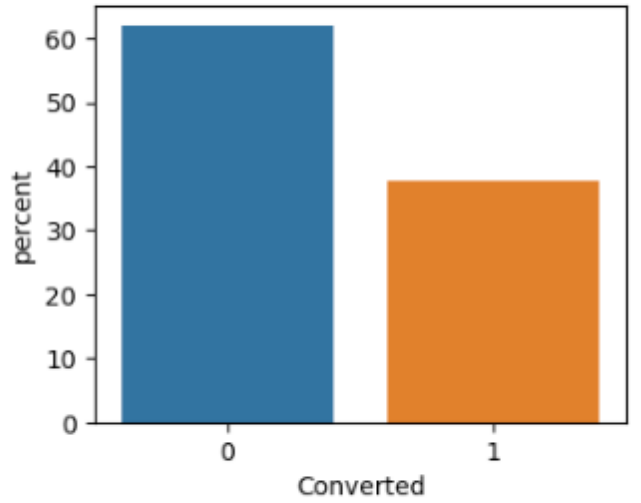






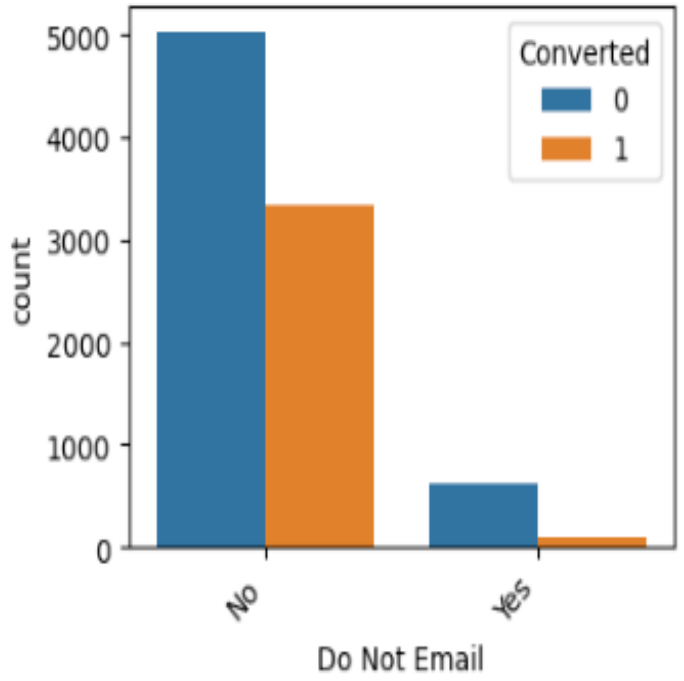


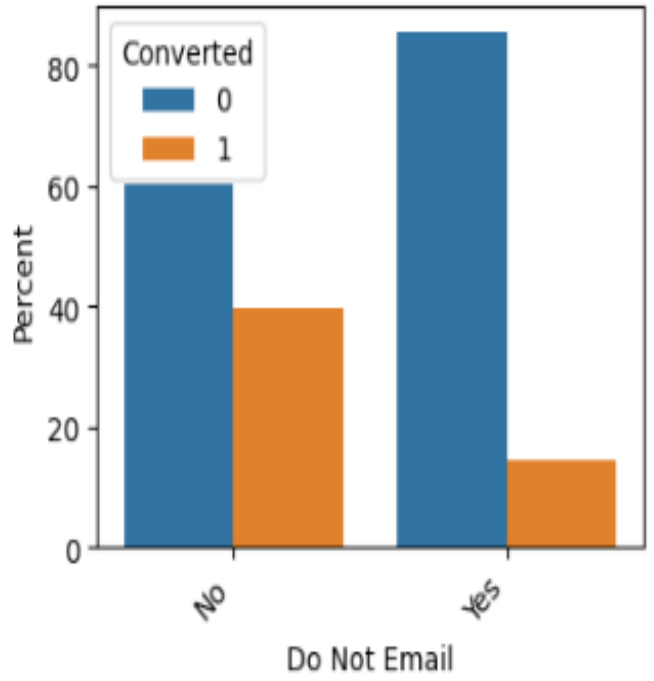


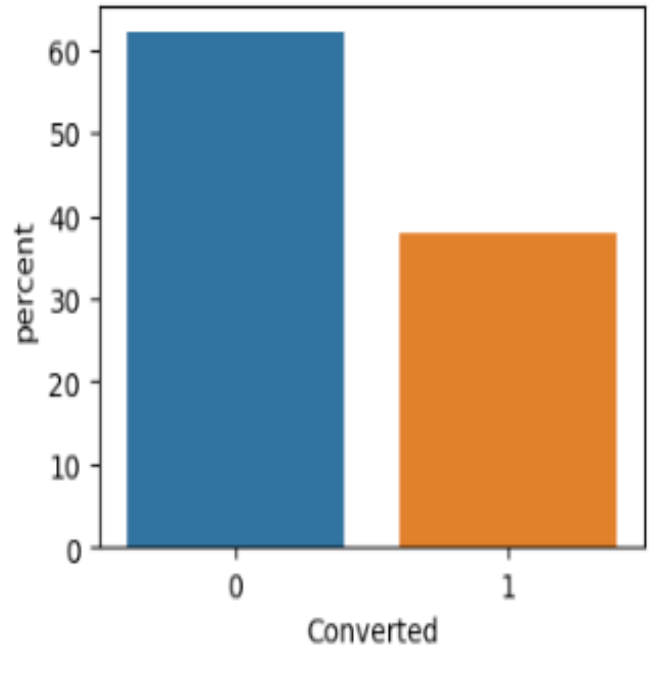


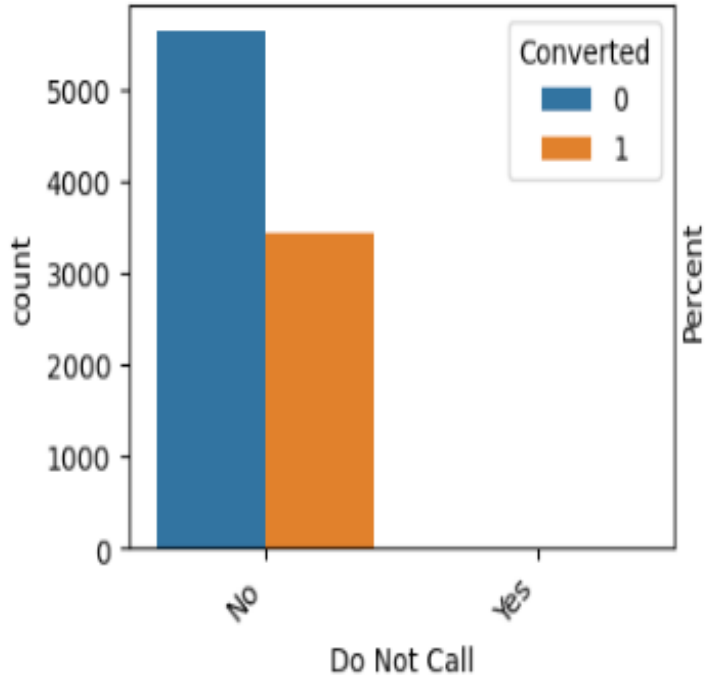
Visualize Categorical Data

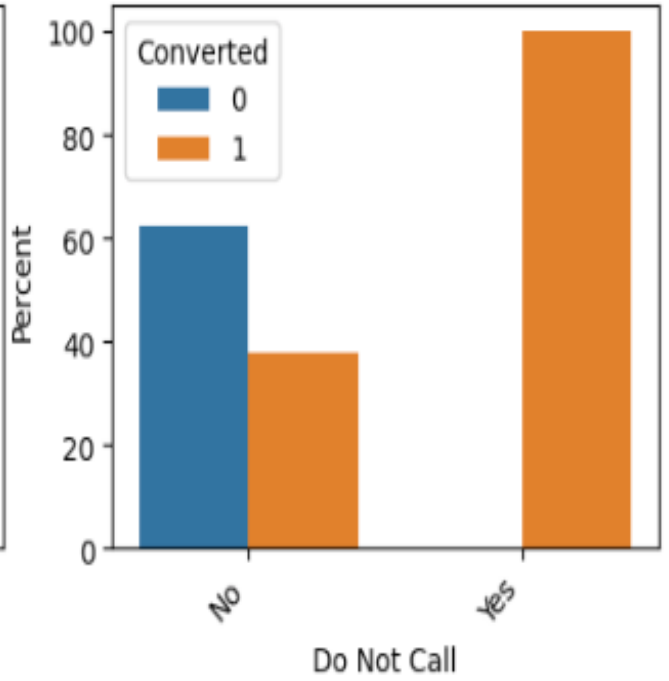
- Do Not Email
- Do Not Call

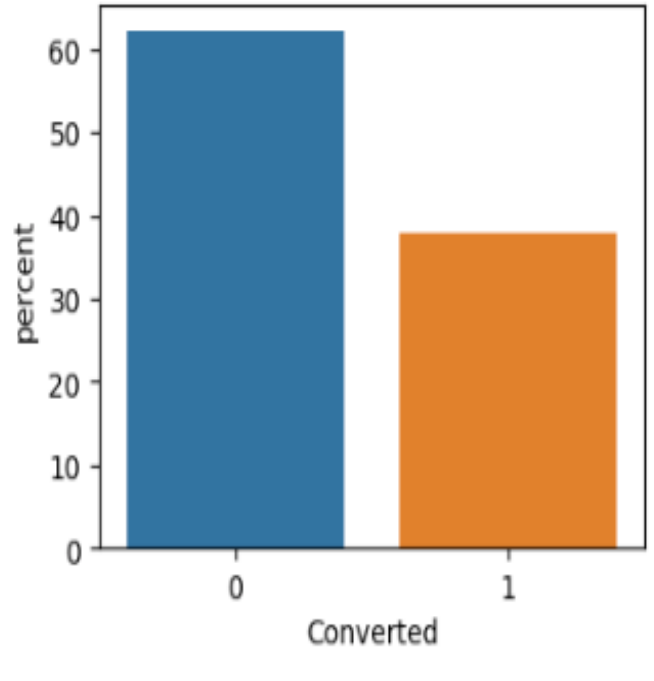






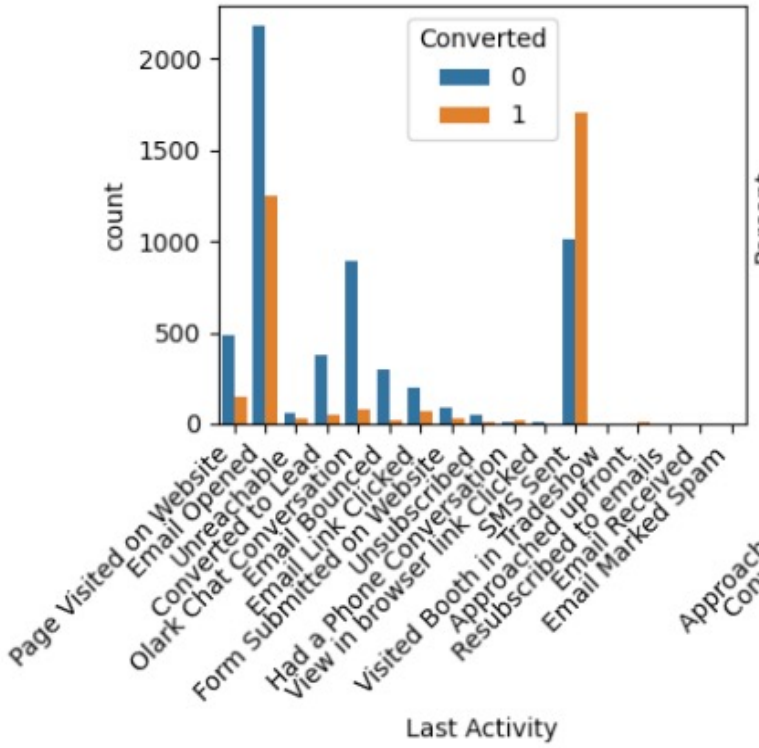


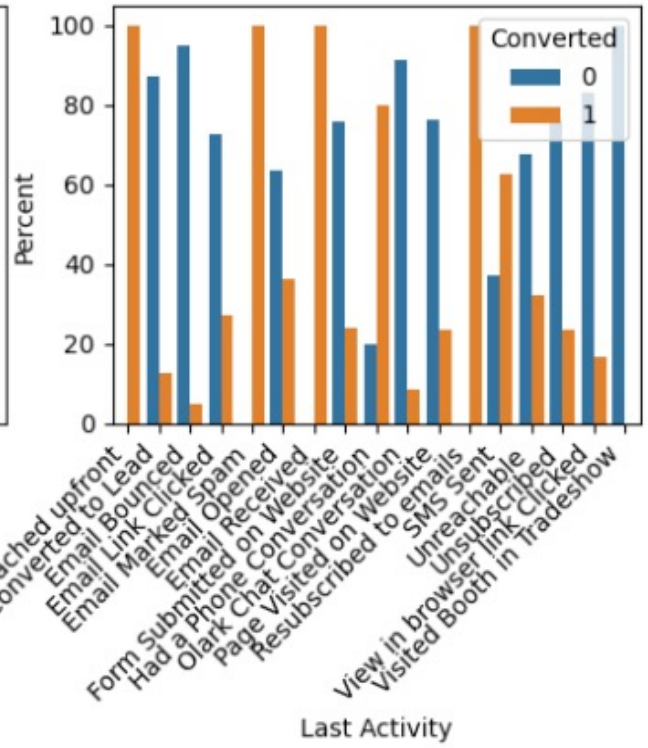


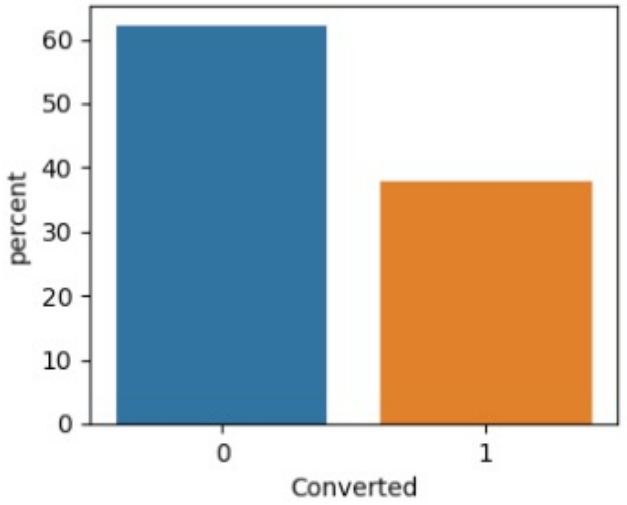


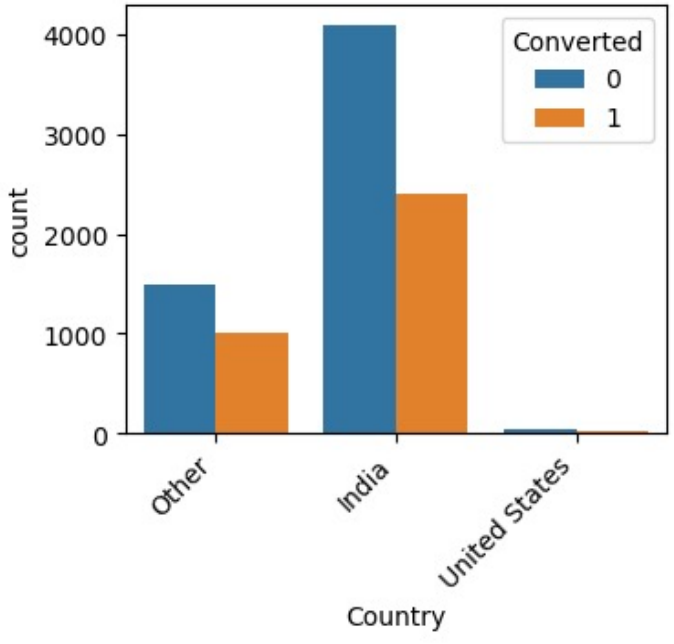
Visualize Categorical Data

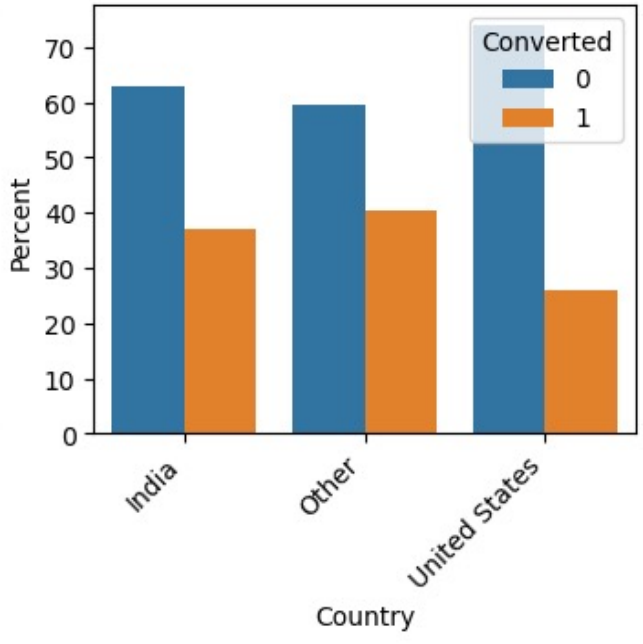
- Last Activity
- Country

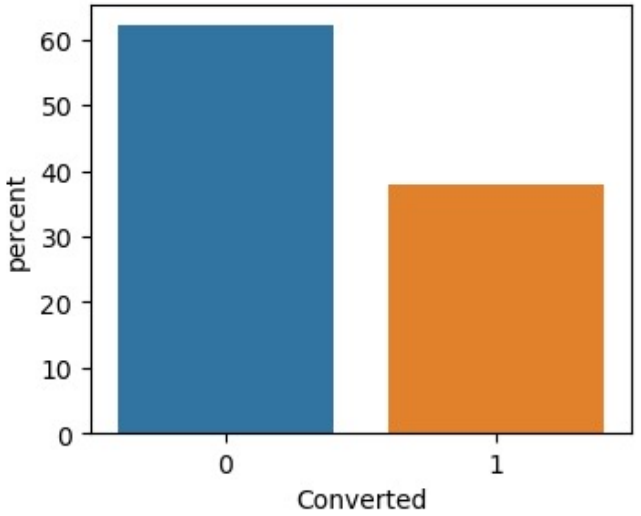






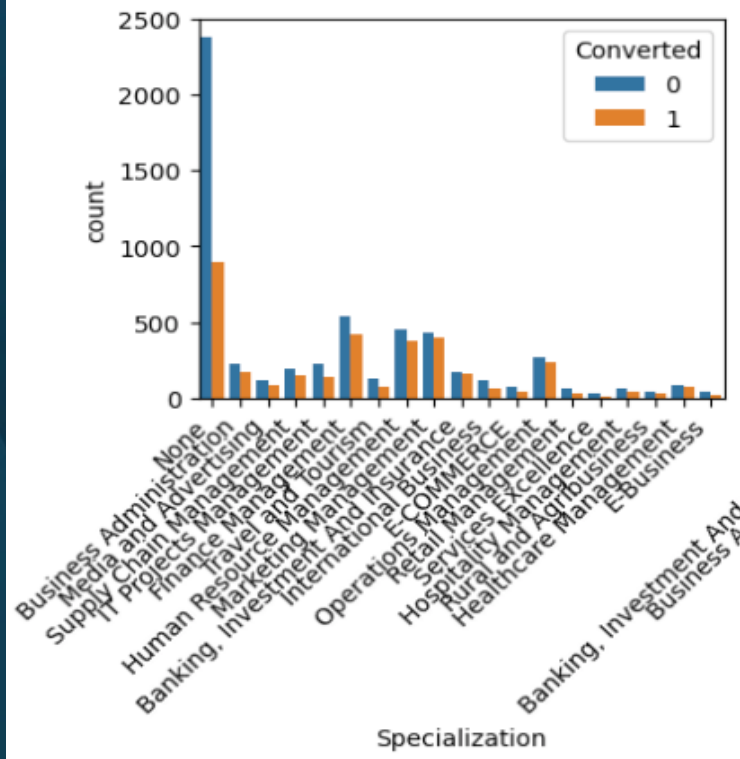


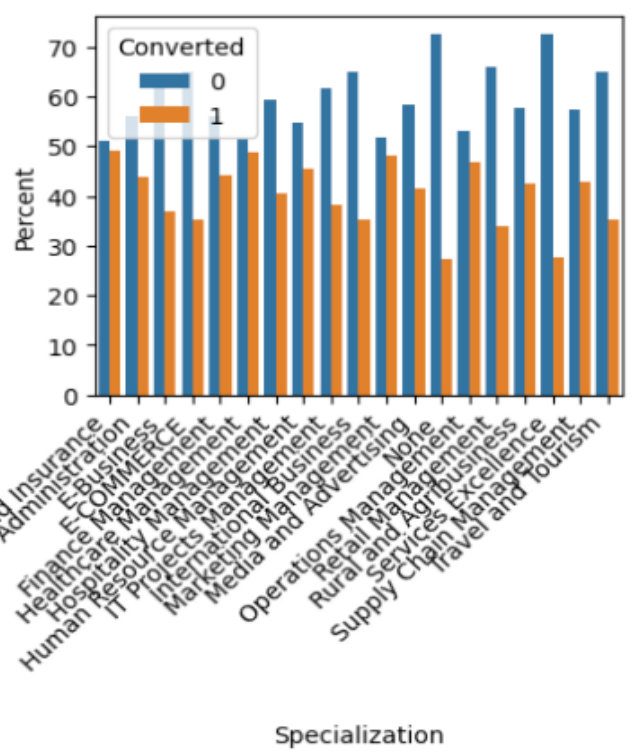


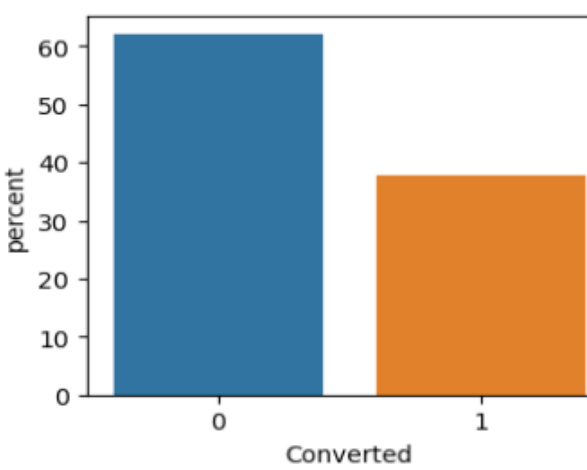


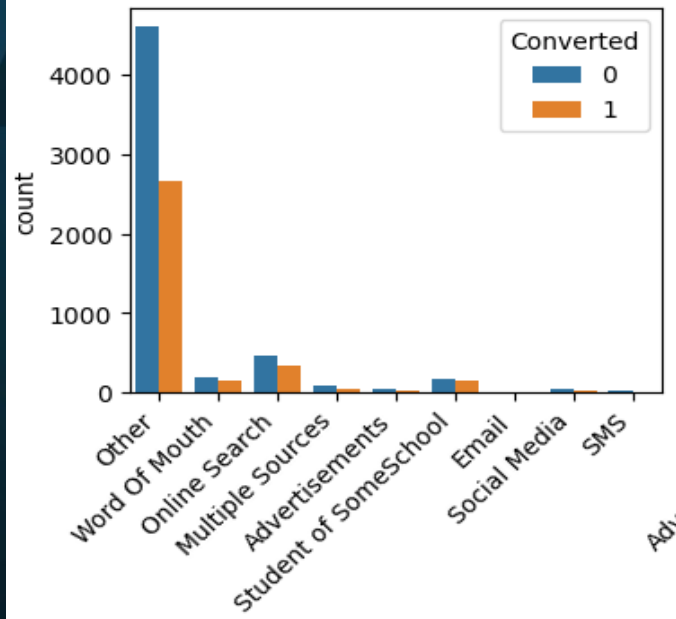
Visualize Categorical Data

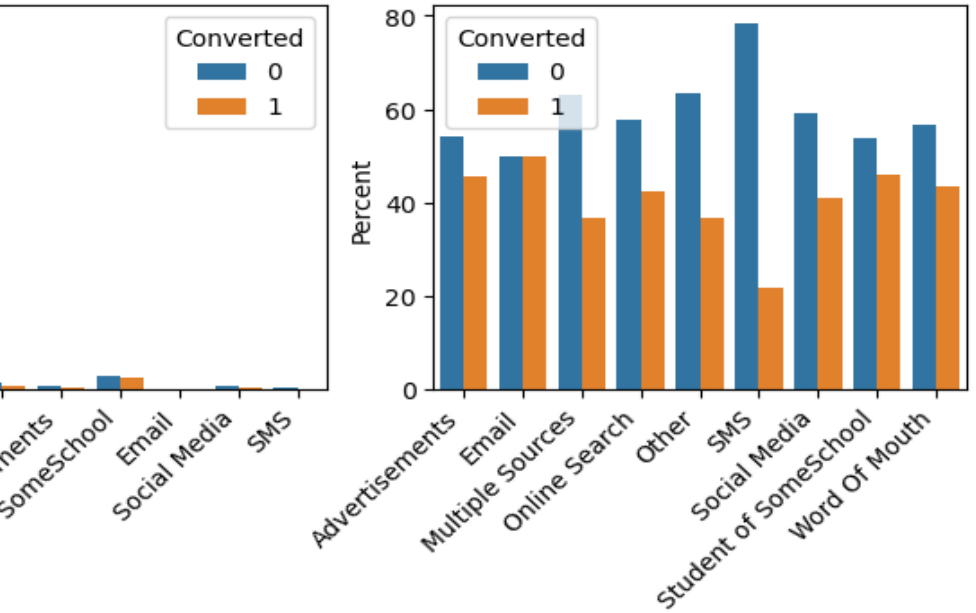
- Specialization
- How did you hear about X Education

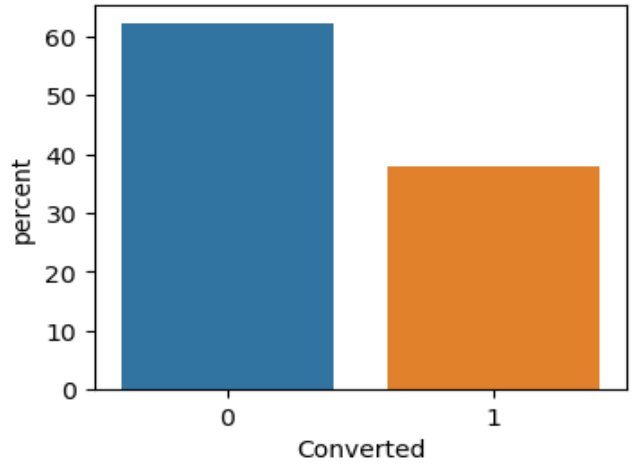






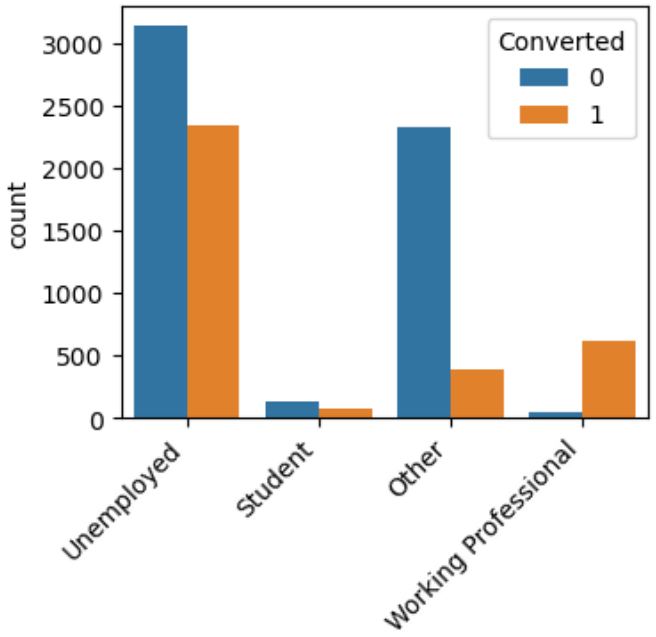


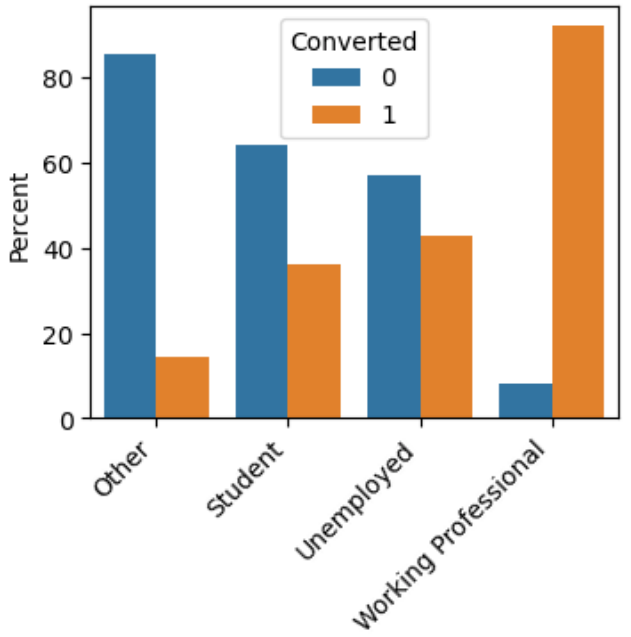


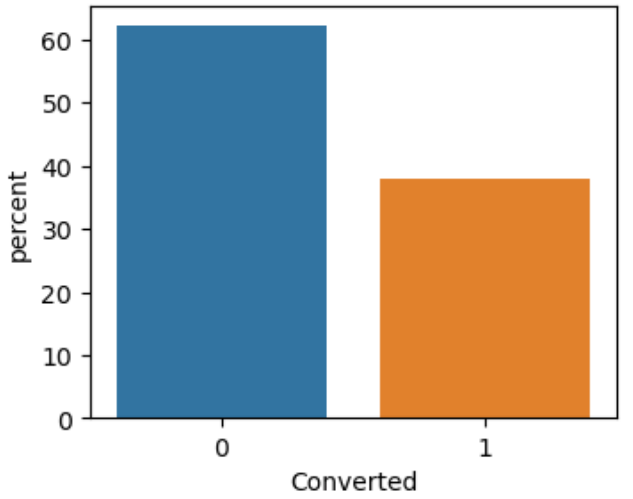


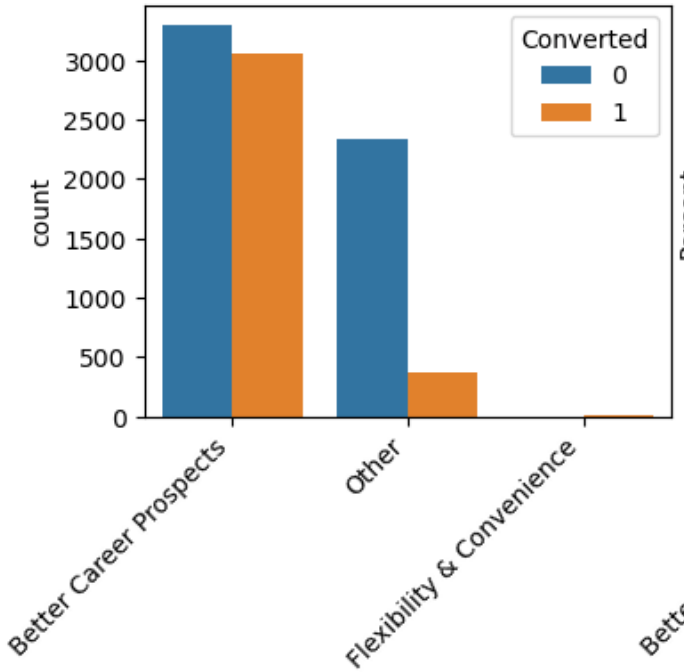
Visualize Categorical Data

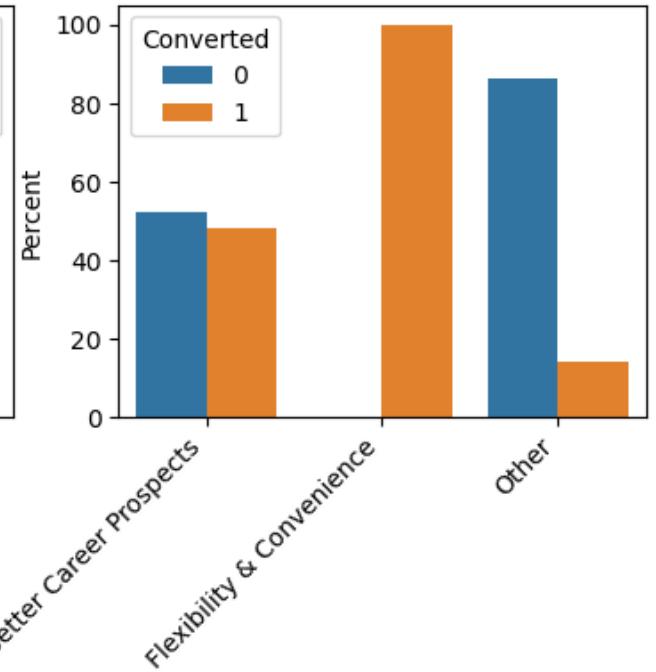
- What is your current occupation
- What matters most to you in choosing a course

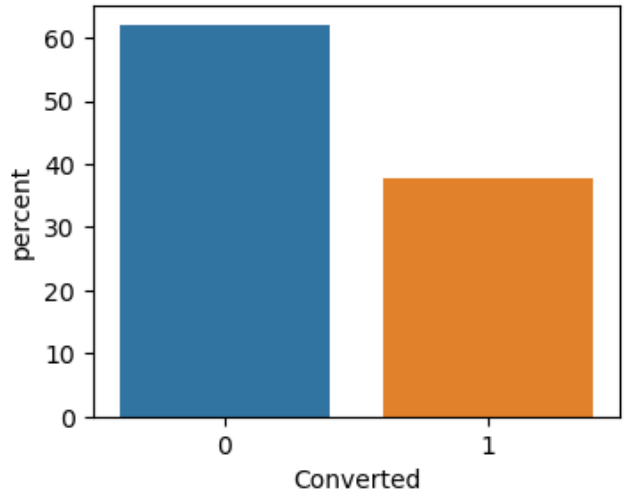






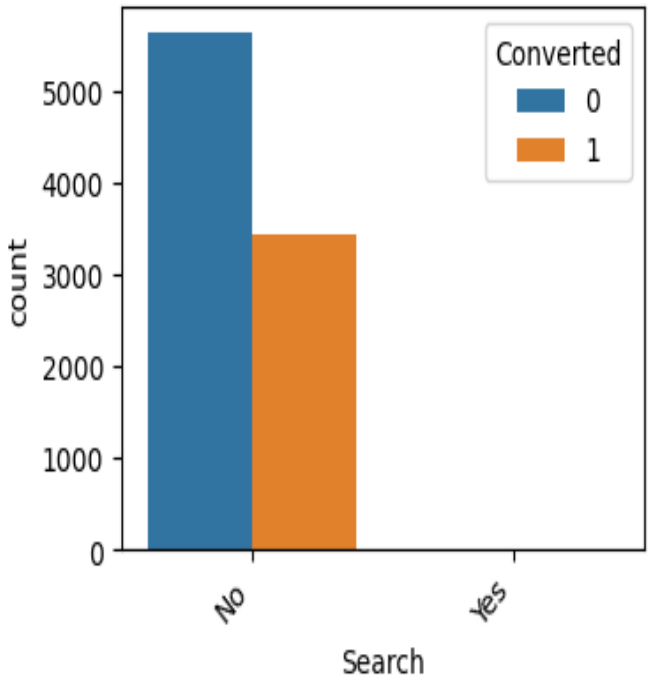


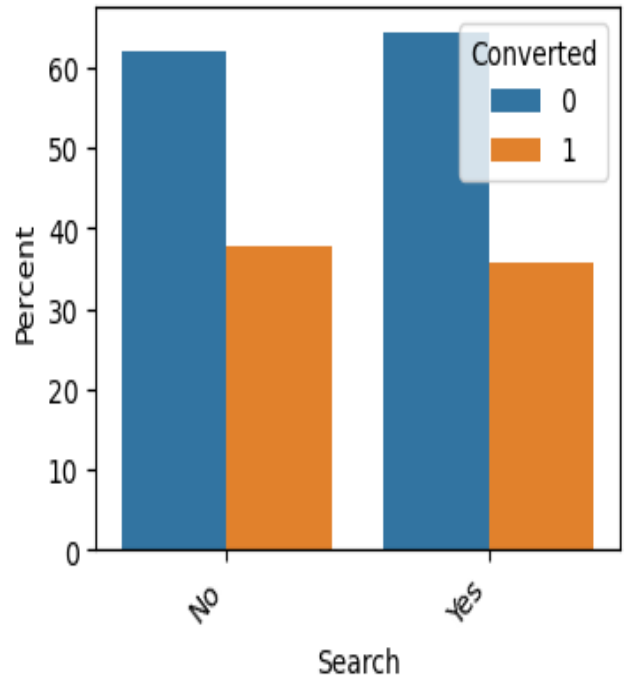


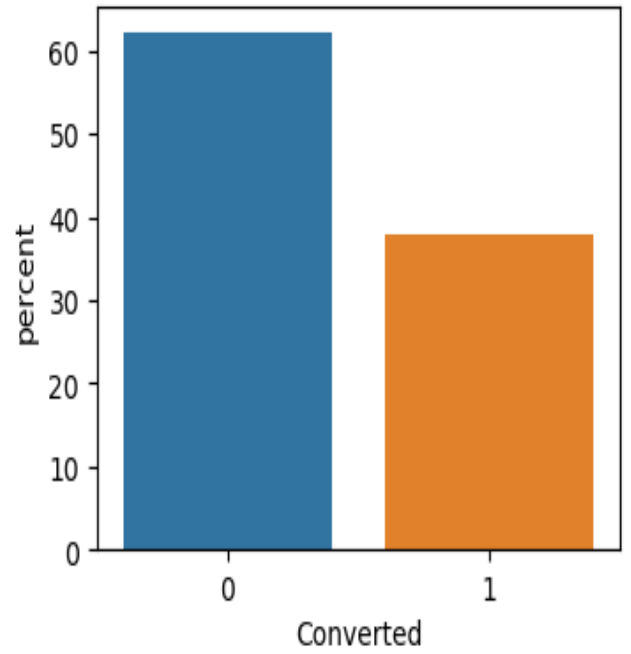


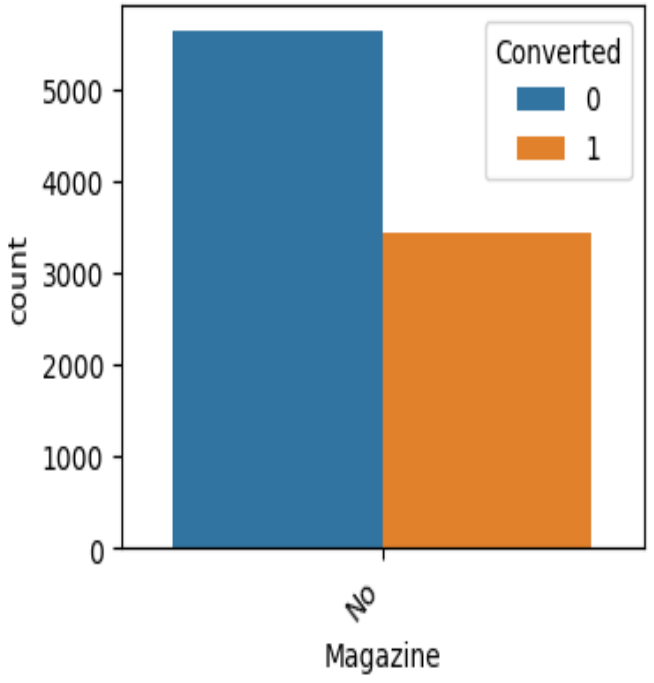
Visualize Categorical Data

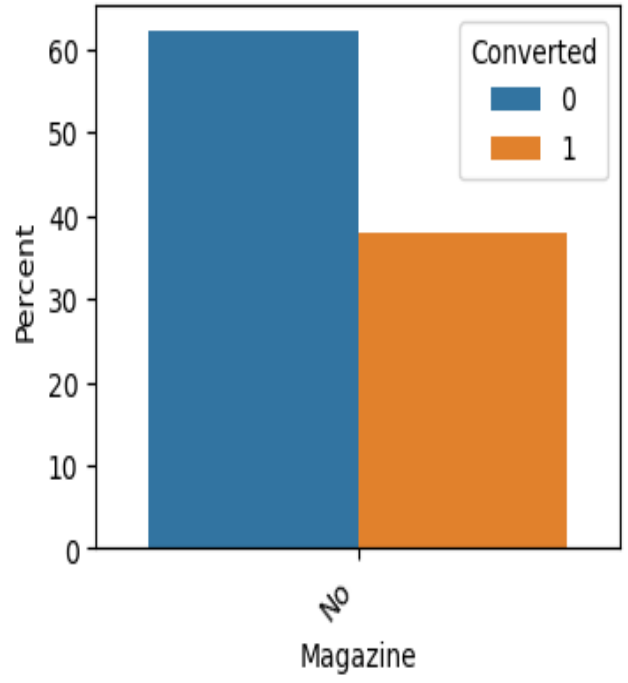
- Search
- Magazine

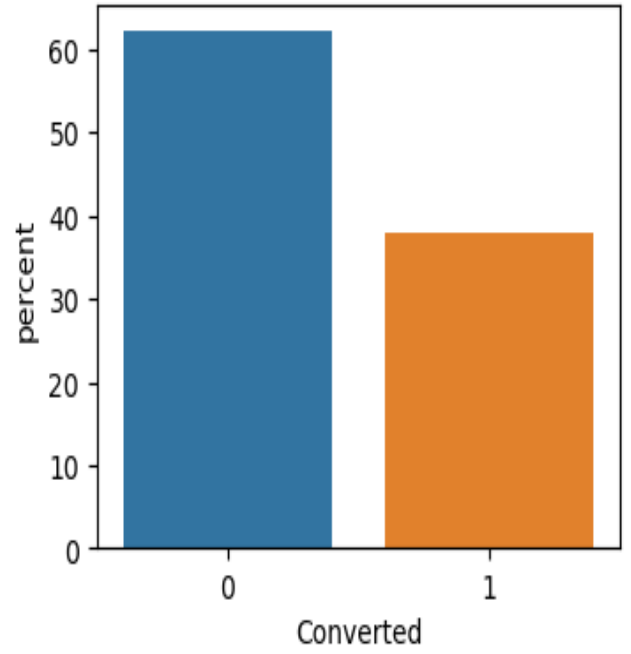






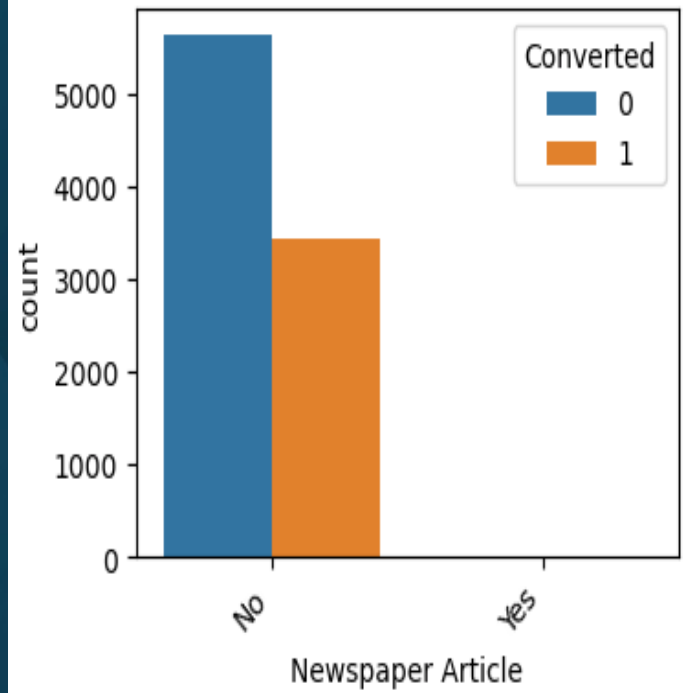


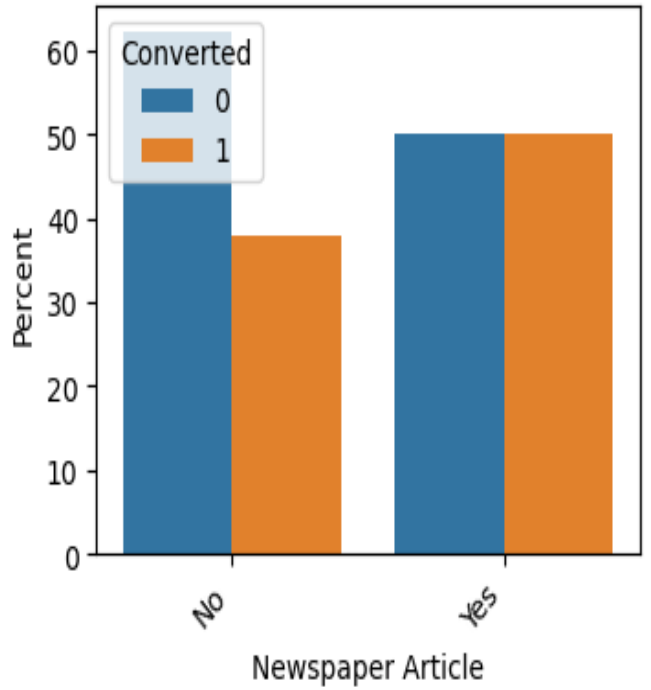


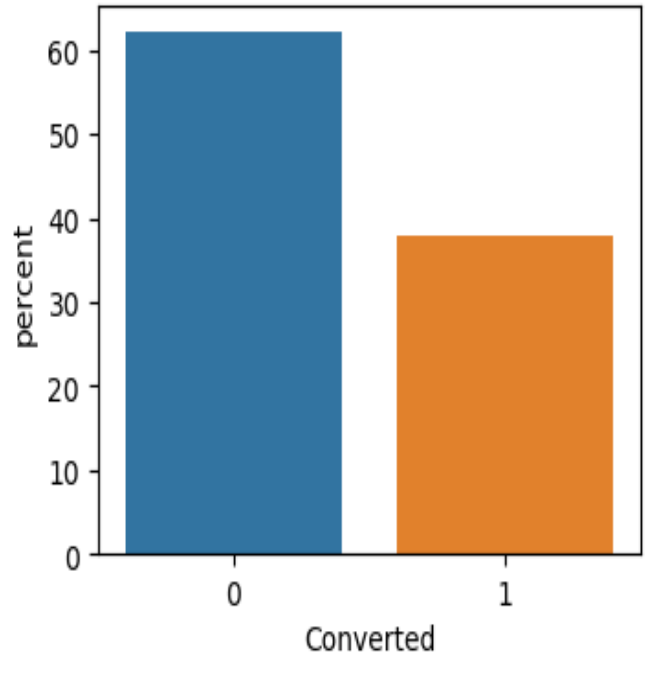


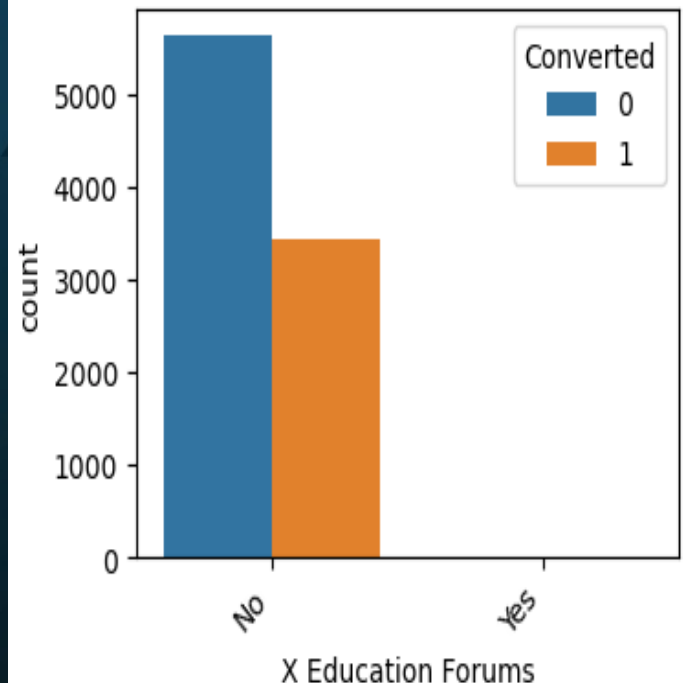
Visualize Categorical Data

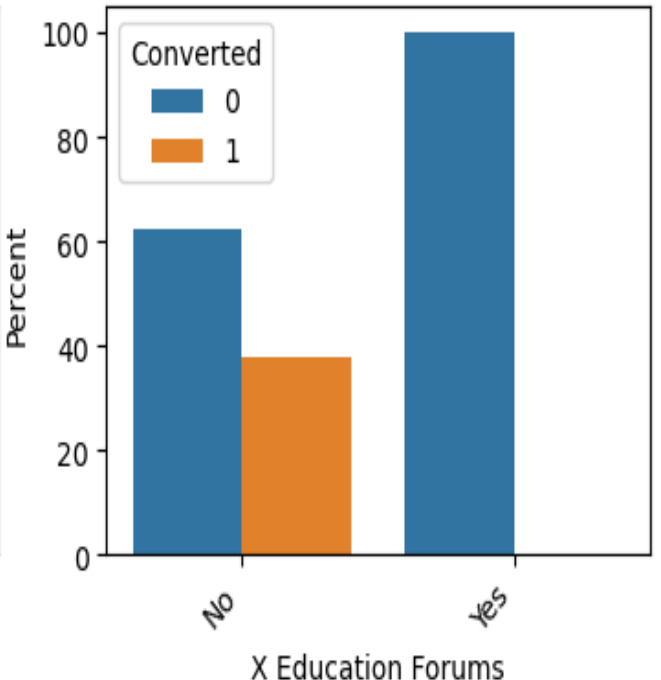
- Newspaper Article
- X Education Forums

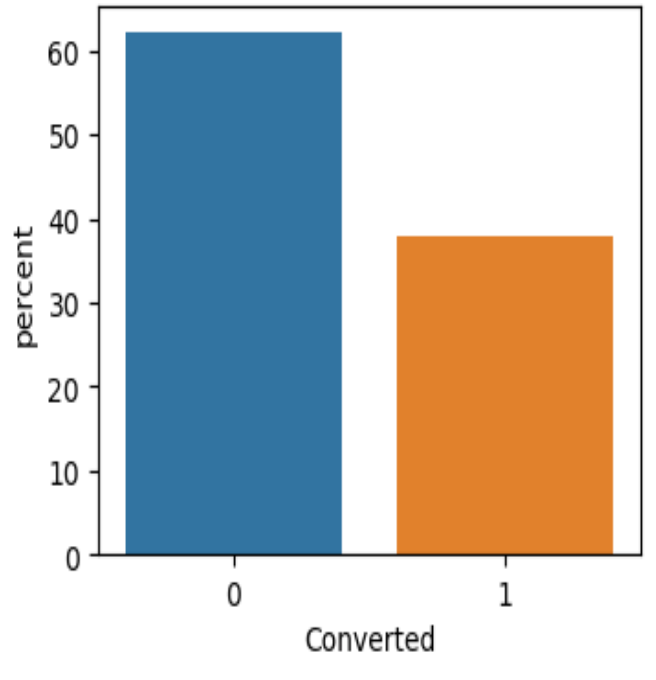






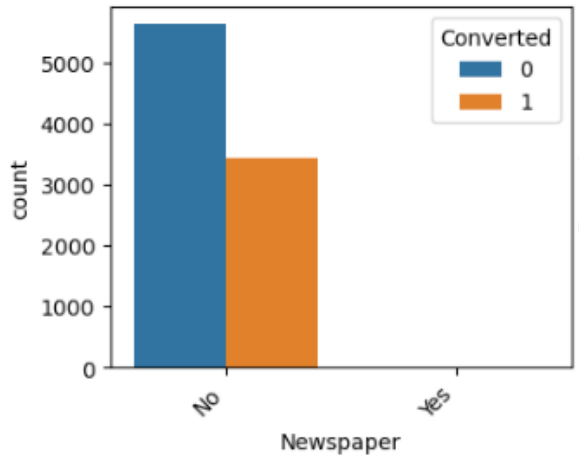


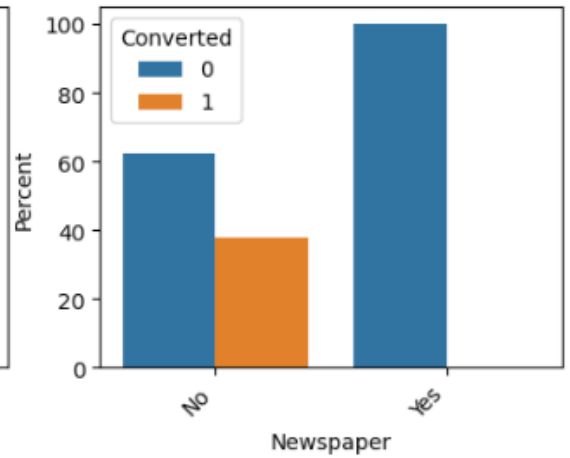


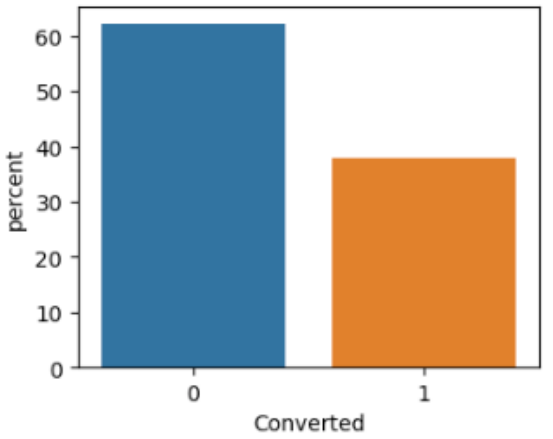


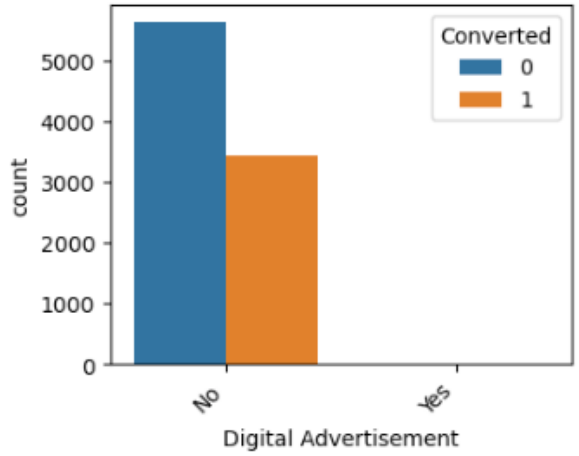
Visualize Categorical Data

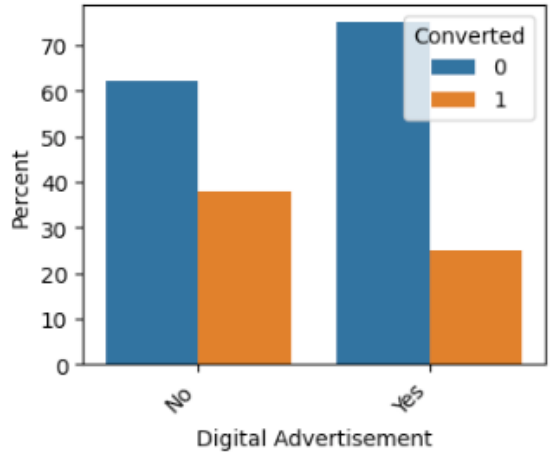
- Newspaper
- Digital Advertisement
- Through Recommendations

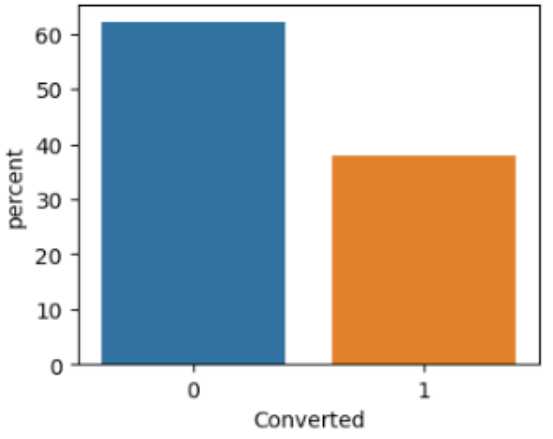


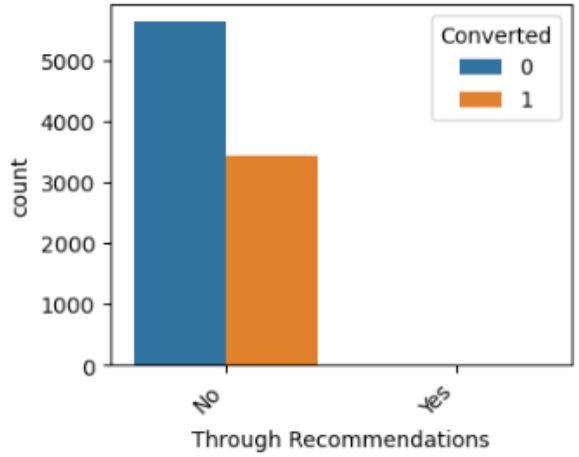


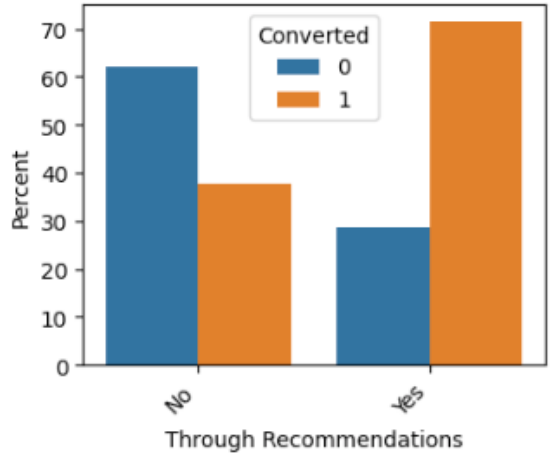


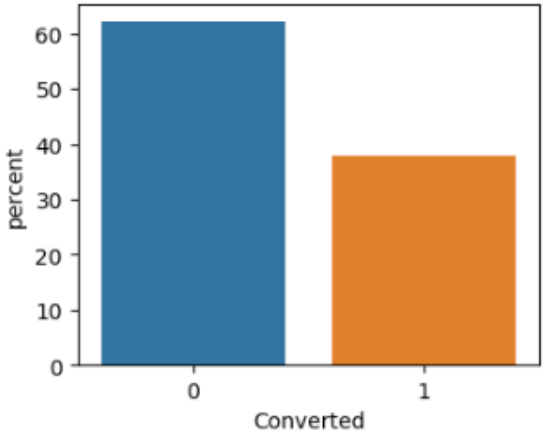






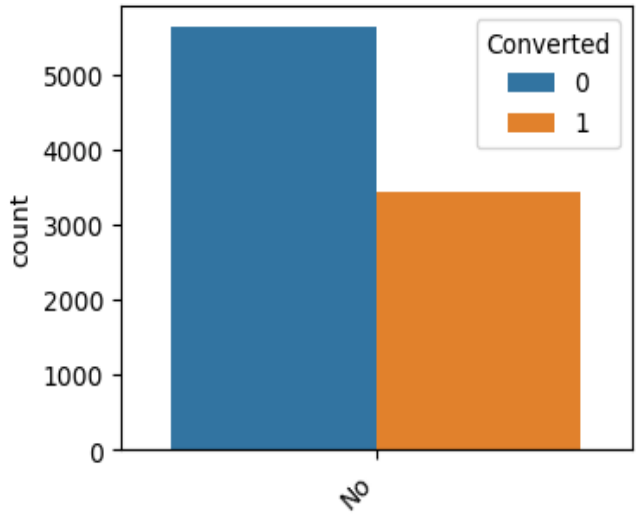


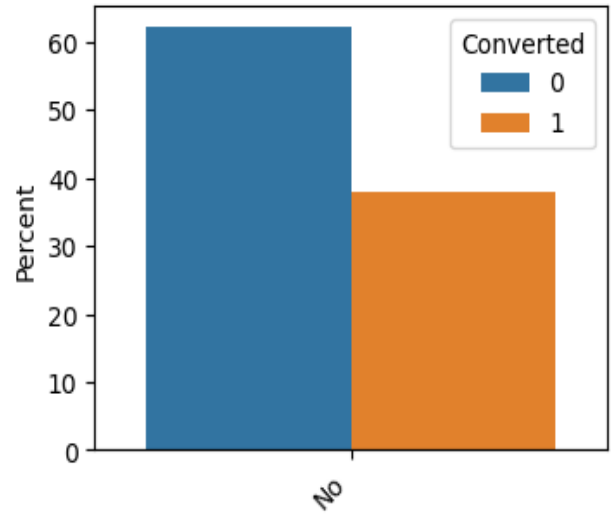


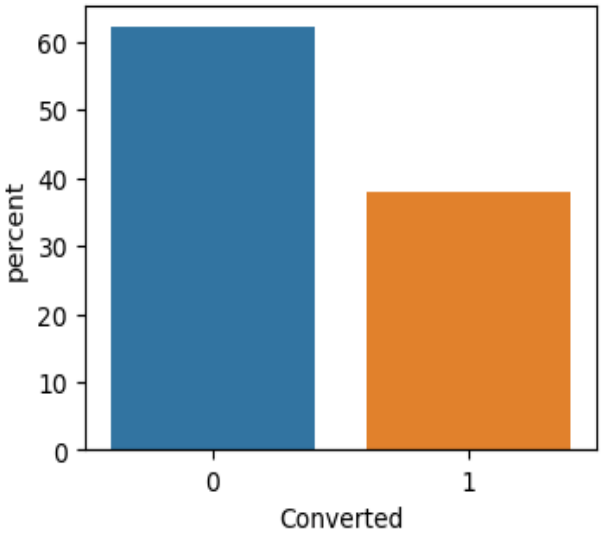


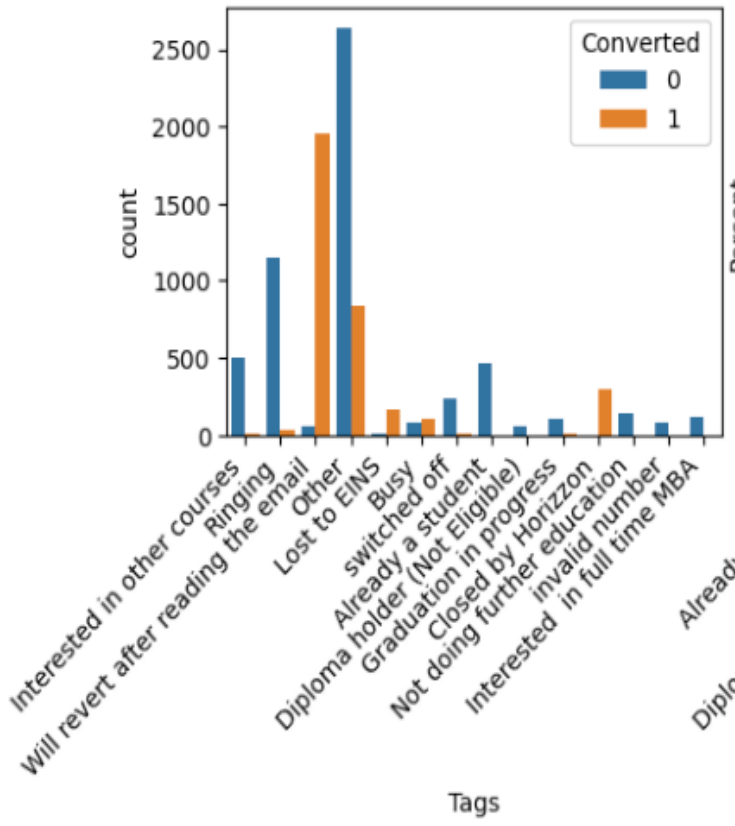
Visualize Categorical Data

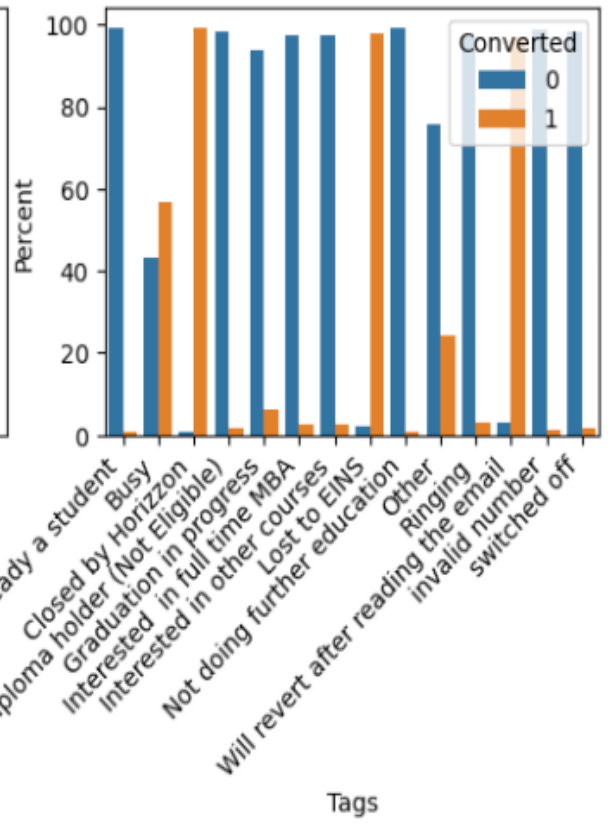
- Receive More Updates About Our Courses
- Tags

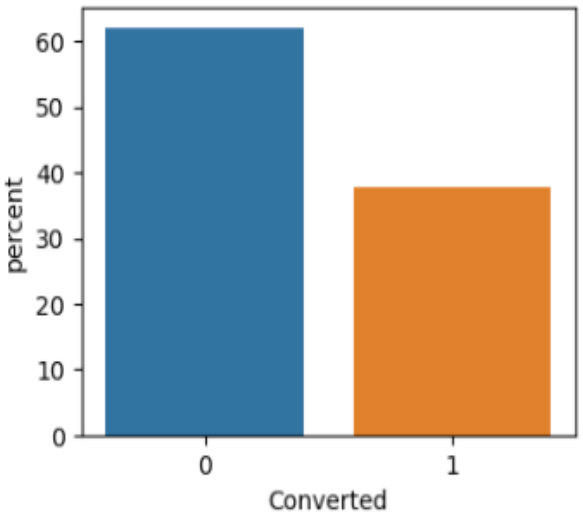






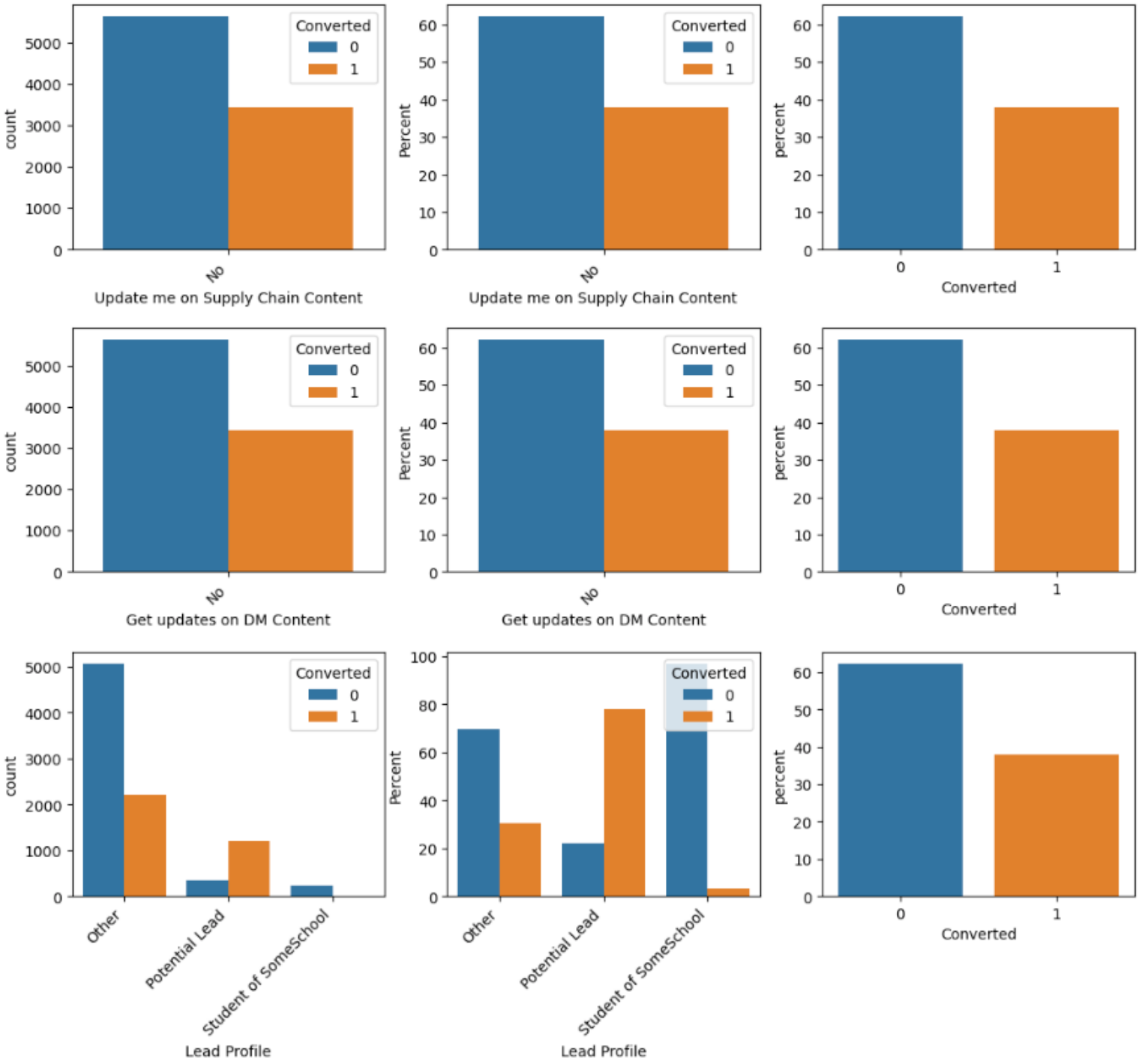






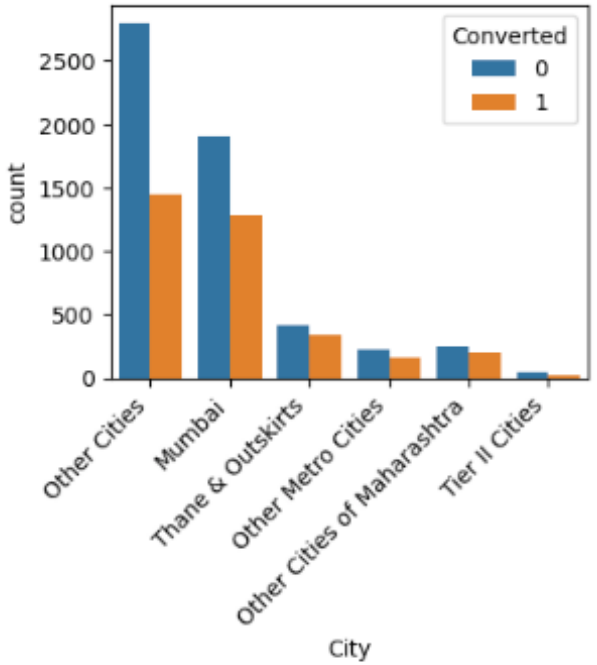
Visualize Categorical Data

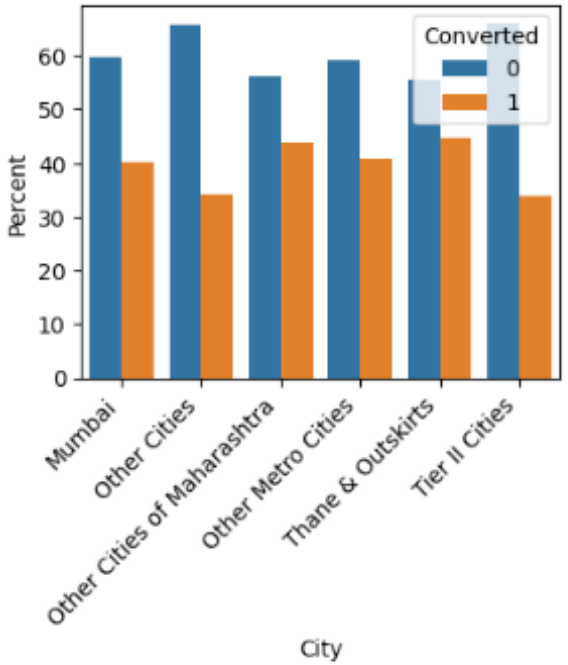
- Update me on Supply Chain Content
- Get updates on DM content
- Lead Profile

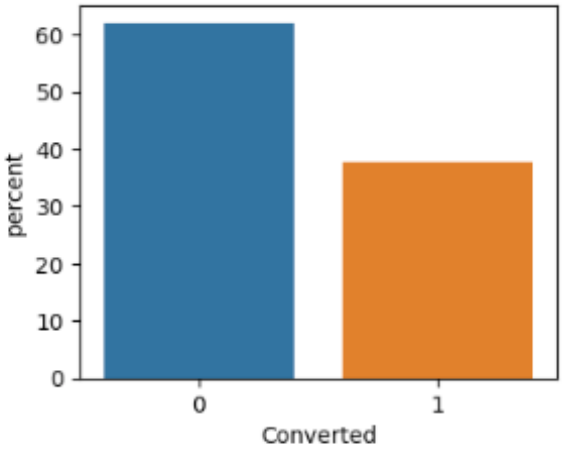


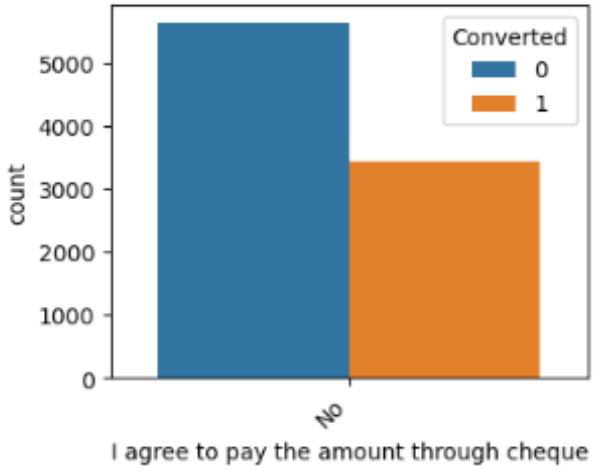
Visualize Categorical Data

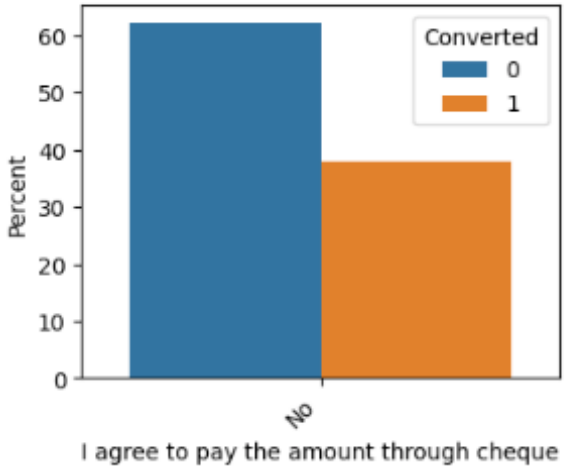
- City
- I agree to pay the amount through cheque

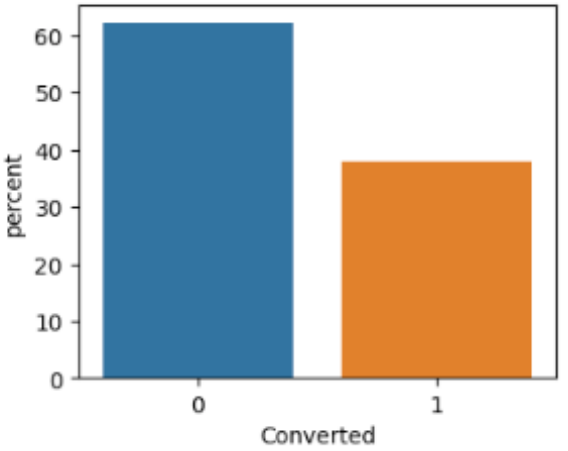






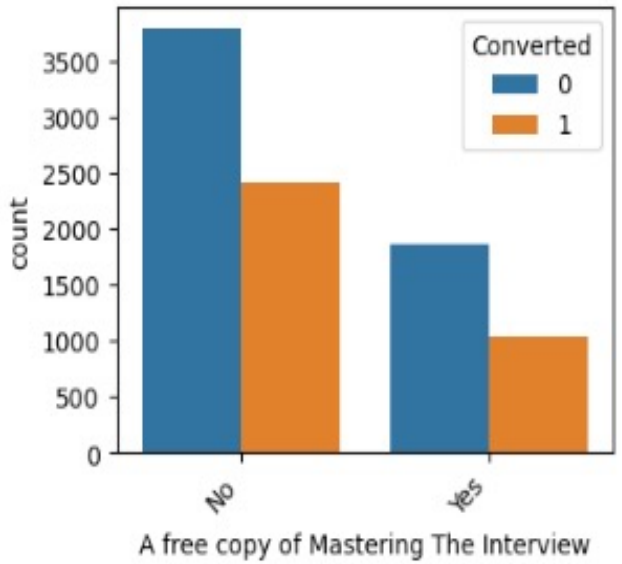


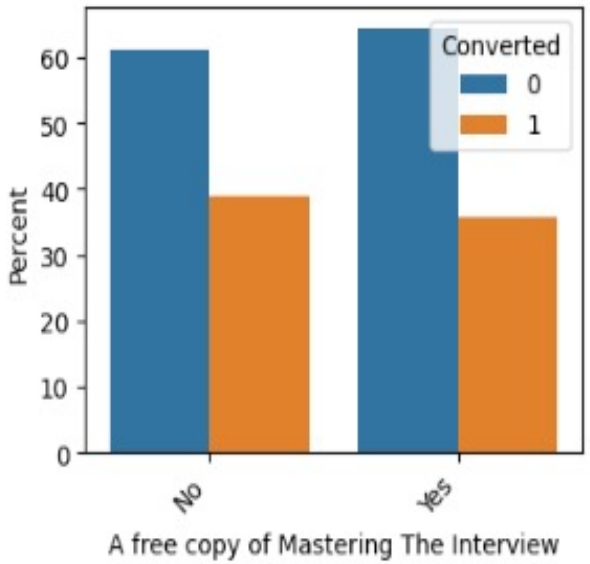


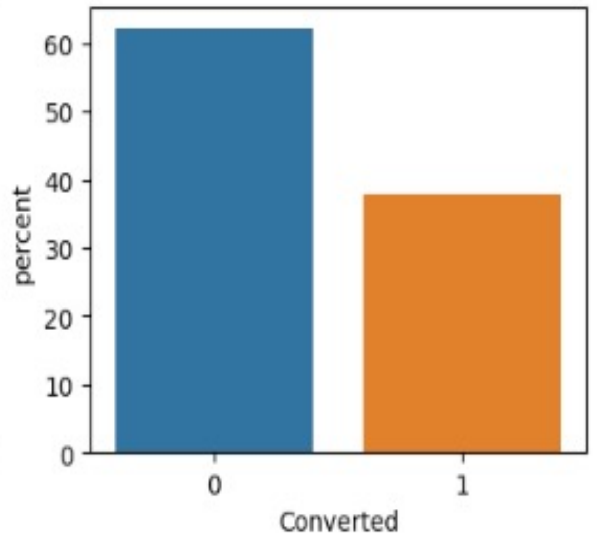


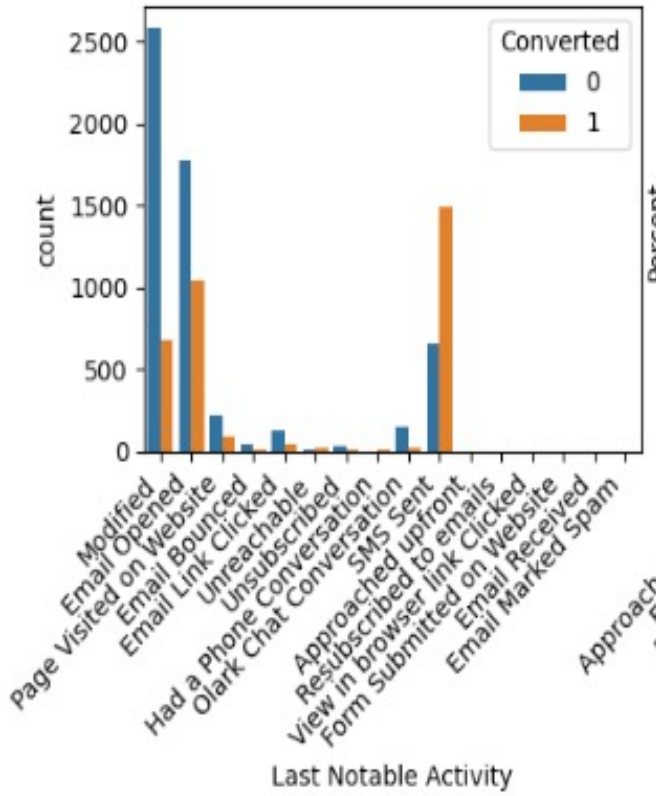
Visualize Categorical Data

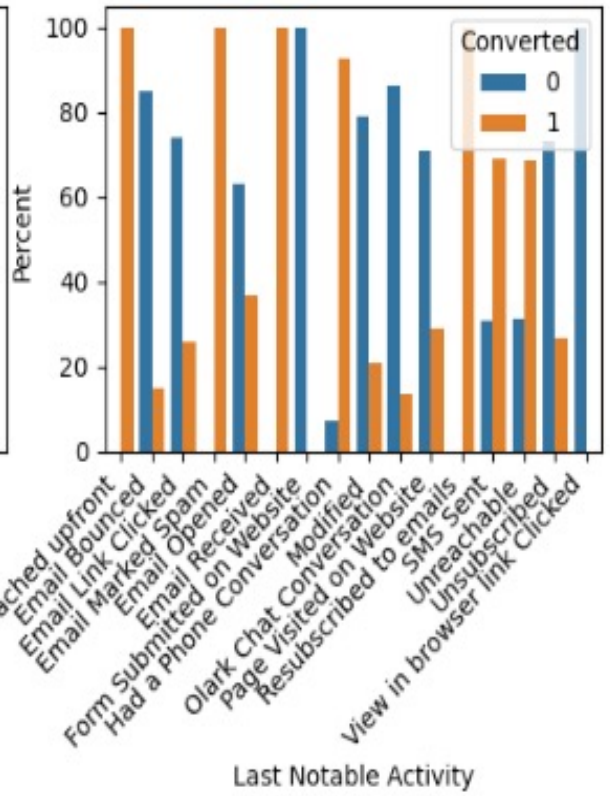
- A free copy of Mastering The Interview
- Last Notable Activity

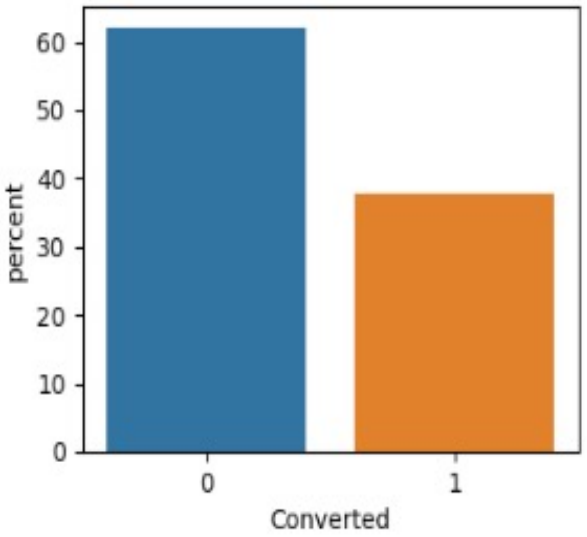












Numerical Correlation Heatmap

II. Analysis Approach

- EDA
- Data Visualization
- Build the logistic regression model
- Model evaluation

Build Logistic Regression Model - Approach

A. Data Transformations for modelling

1. Encoding binary categorical (Yes/No --> 1/0)
2. Encoding categorical with dummy variables
   a. Auto Dropping 1st dummy var
   b. Manual Dropping a selected dummy var

B. Model Preparation

1. Check dataset imbalance
2. Create X and y
3. Split TRAIN-SET
4. Scale numerical data for TRAIN set

C. Auto-Eliminate Features with RFE

1. Start running RFE with number of selected features = 50% to eliminate features
2. Build logistic regression with statsmodel for interpreting easily
3. Verify number of insignificant features <= 12 (p-Value>0.05)
4. If number of insignificant>12, repeat step 1 with new smaller number of selected feature

D. Manual Eliminate Features

1. Prepare list of features to build model, initially with selected features from RFE
2. Build logistic regression model with statsmodel with selected features and fit the model
3. [Checkpoint 1] If there's insignificant feature (pValue>0.05), remove the the feature with highest pValue from selected feature, and back to step 2
4. [Checkpoint 2 – No insignificant feature] Calculate VIF score to detect the multicollinearity issue (VIF>5). If there is VIF>5, we remove the feature with highest VIF from selected features and back to step 2

Final model Evaluation

1. All features are significant (pValue<=0.05)
2. All features do not have multicollinearity problem (VIF<=5)
3. Build the confusion matrix
4. Calculate model metrics:
   a. Accuracy score
   b. Sensitivity vs. specificity
   c. TPR vs. FPR
   d. Precision vs. Recall
   e. Positive Prediction Ratio vs. Negative Prediction Ratio

II. Build Logistic Regression Model

A. Data Transformations – Encoding categorical data

[Categorical] Encoding Yes/No to 1/0

1.Do Not Email

2.Do Not Call

3.Search

4.Magazine (No only)

5.Newspaper Article

6.X Education Forums

7.Newspaper

8.Digital Advertisement

9.Through Recommendations

10.Receive More Updates About Our Courses

11.Update me on Supply Chain Content

12.Get updates on DM Content

13.I agree to pay the amount through cheque

14.A free copy of Mastering The Interview

[Categorical] Drop 1st dummy var

1.Lead Origin

2.Lead Source

3.Last Activity

4.Last Notable Activity

All values in each categorical have the same level of information value. So we drop the 1st dummy variable (randomly)

[Categorical] Drop "Other" dummy var

1.Country

2.How did you hear about X Education

3.What is your current occupation

4.What matters most to you in choosing a course

5.Tags

6.Lead Profile

7.City

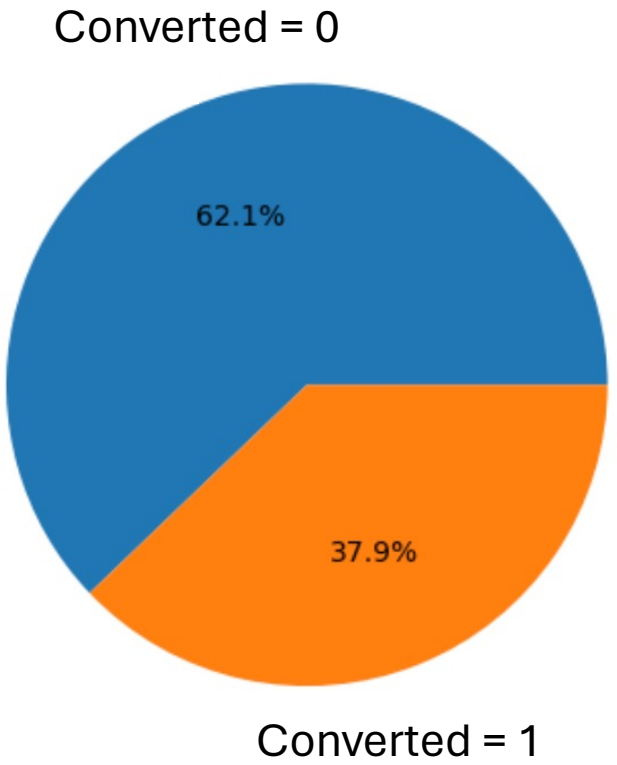
8.Specialization

The "Other" value doesn't contain the information value much vs. remained values in categorical. So we drop the "Other" dummy variable

II. Build Logistic Regression Model

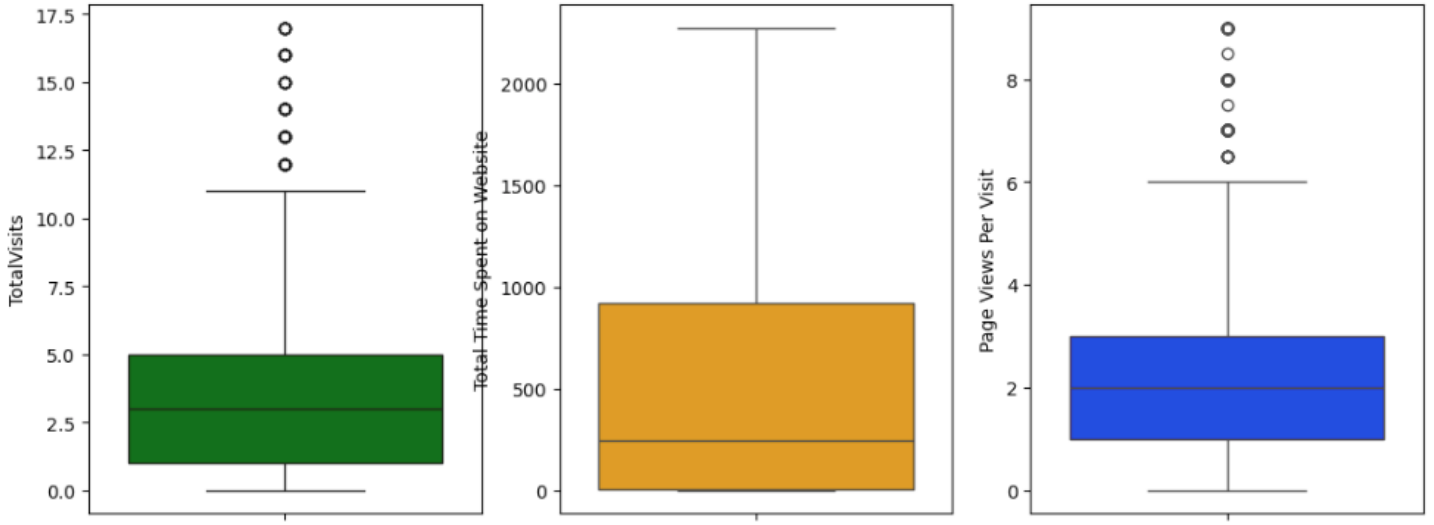
B1. Model Preparation – Check Imbalance Dataset

Dataset is not imbalance! It’s good for ML

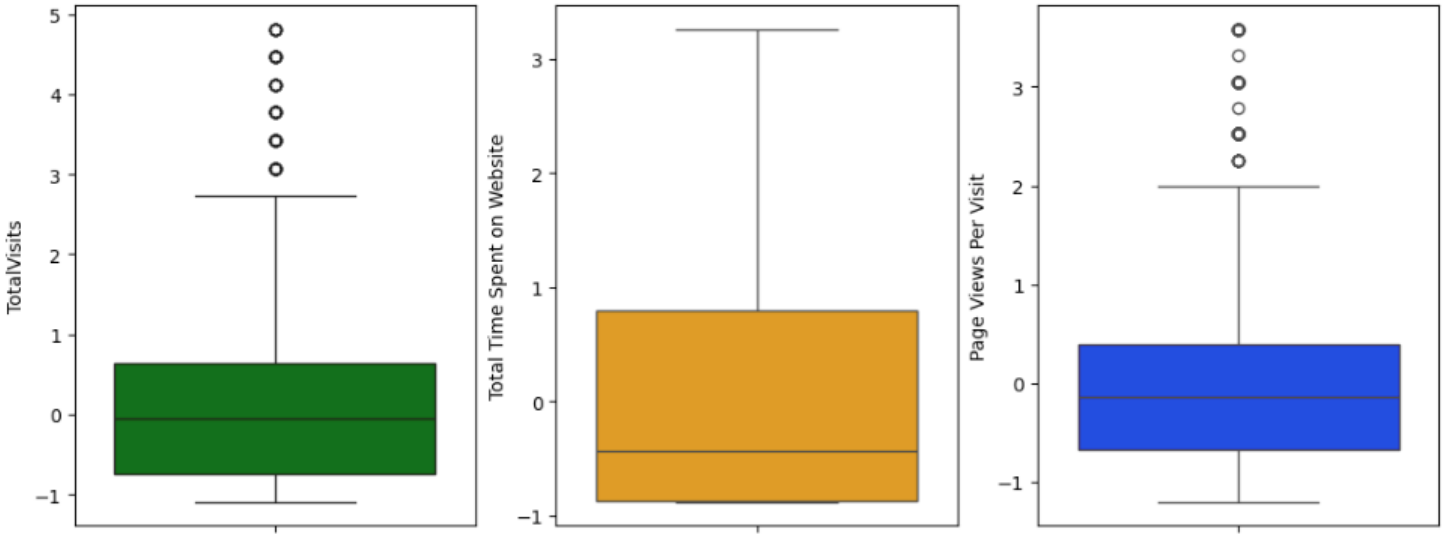


B4. Model Preparation– Scale the numerical data

Before scaling...

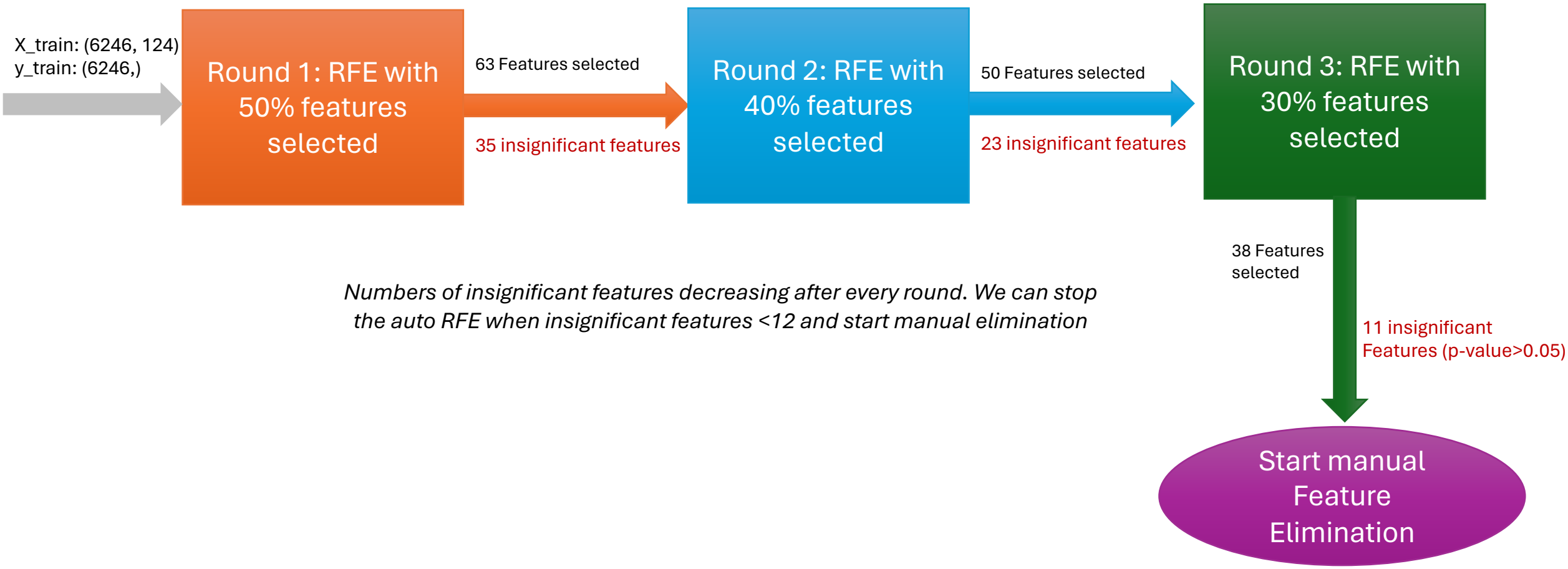


After scaling...



II. Build Logistic Regression Model

C. Auto Eliminate Features with RFE



Numbers of insignificant features decreasing after every round. We can stop the auto RFE when insignificant features <12 and start manual elimination

II. Build Logistic Regression Model

D. Manual Eliminate Features

Round	Insignificant/ Total Features	Insignificant feature (Highest p-Value >5%)	Multicollinear feature (Highest VIF >5)	Decision
4	10/37	"Last Activity_Email Bounced" (42%)		Drop insignificant feature
5	8/36	"Specialization_Hospitality Management" (21.3%)		Drop insignificant feature
6	7/35	"What is your current occupation_Working Professional" (11%)		Drop insignificant feature
7	6/34	"City_Other Metro Cities" (10.3%)		Drop insignificant feature
8	5/33	"Lead Profile_Student of SomeSchool" (10%)		Drop insignificant feature
9	3/32	"Last Activity_Had a Phone Conversation" (8.2%)		Drop insignificant feature
10	3/31	"Lead Origin_Lead Add Form" (6.3%)		Drop insignificant feature
11	2/30	"Last Notable Activity_Unsubscribed" (5.5%)		Drop insignificant feature
12	0/29		"Last Notable Activity_SMS Sent" (7.4)	Drop multicollinear feature
13	3/28	"Last Activity_Olark Chat Conversation" (9.1%)		Drop insignificant feature
14	3/27	Last Activity_Converted to Lead (11.4%)		Drop insignificant feature

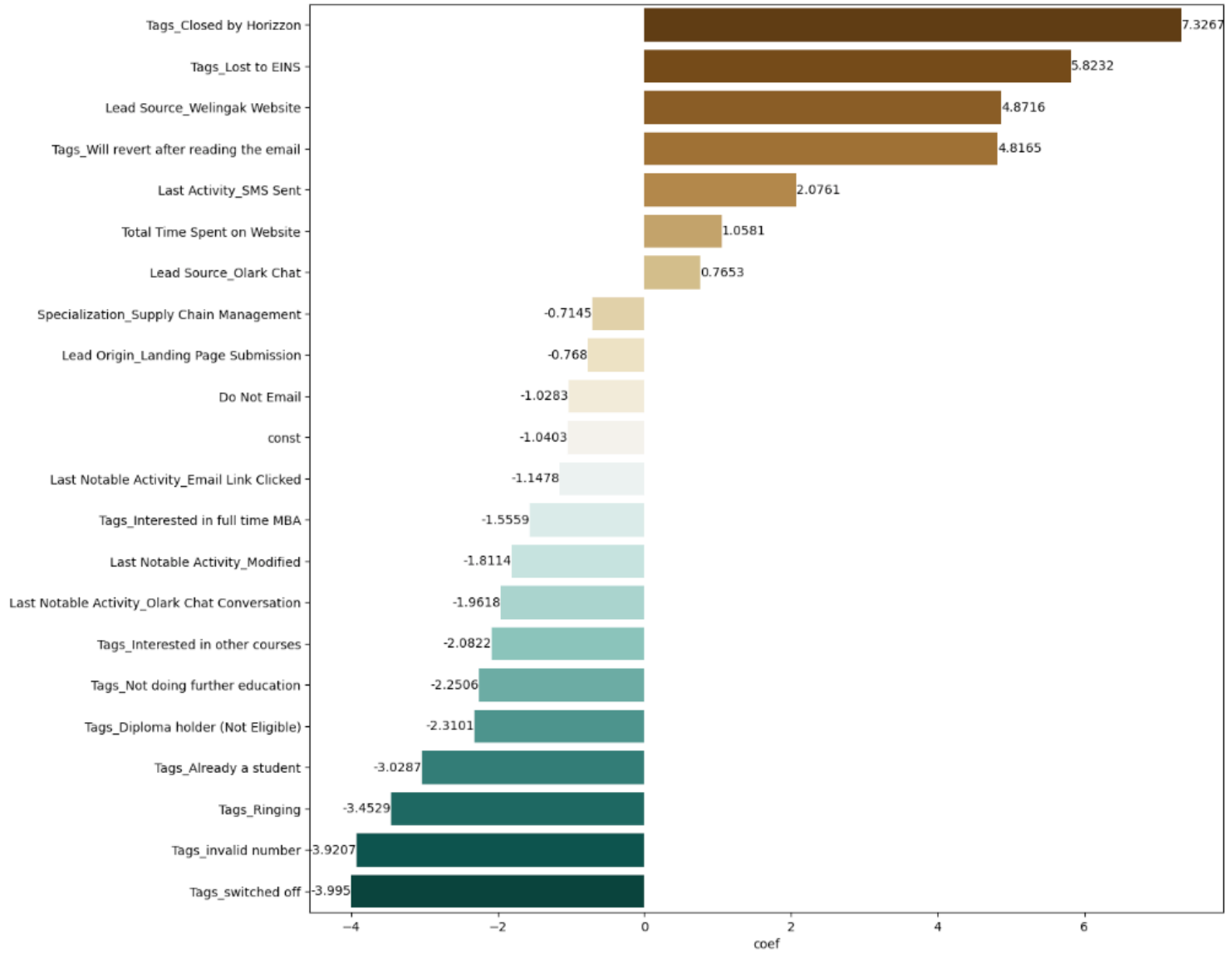
II. Build Logistic Regression Model

D. Manual Eliminate Features

Round	Insignificant/ Total Features	Insignificant feature (Highest p-Value >5%)	Multicollinear feature (Highest VIF >5)	Decision
15	2/26	Last Activity_Page Visited on Website (9.3%)		Drop insignificant feature
16	1/25	Specialization_Travel and Tourism (6.1%)		Drop insignificant feature
17	0/24		"What matters most to you in choosing a course" (7.2)	Drop insignificant feature
18	1/23	Tags_Graduation in progress (40%)		Drop insignificant feature
19	0/22			All features are significant and there is no multicollinear problem. This is final model

E. Final Model

Coefficients of final model



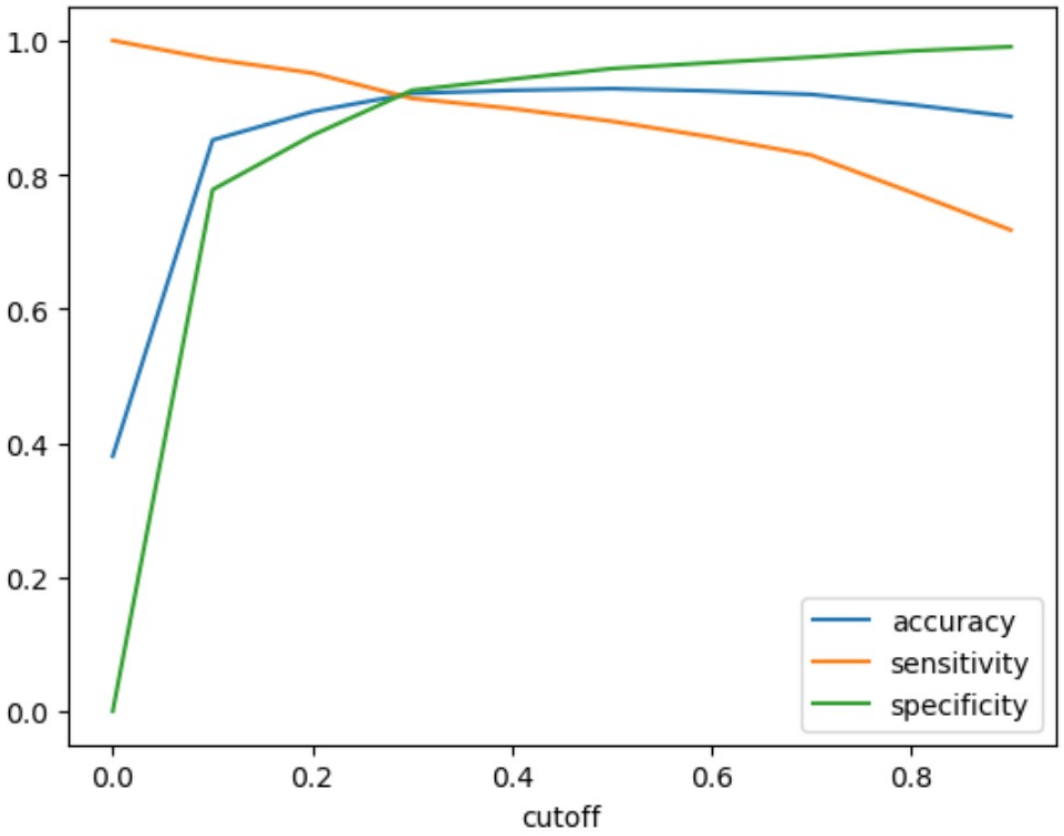
II. Analysis Approach

- EDA
- Data Visualization
- Build the logistic regression model
- Model evaluation

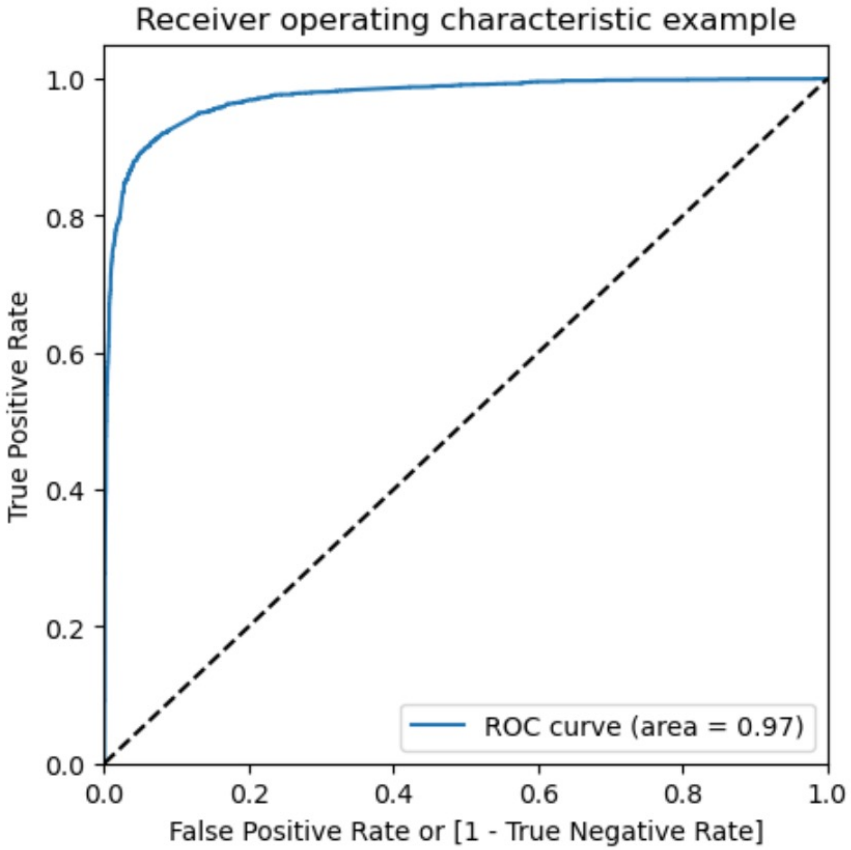
E. Model Evaluation - Metrics for all cut-off

cutoff	accuracy	sensitivity	specificity	TPR	FPR	precision	recall	PP_ratio	NP_ratio
0.0	0.380243	1.000000	0.000000	1.000000	1.000000	0.380243	1.000000	0.380243	NaN
0.1	0.851585	0.972211	0.777577	0.972211	0.222423	0.728391	0.972211	0.728391	0.978544
0.2	0.894012	0.951579	0.858693	0.951579	0.141307	0.805130	0.951579	0.805130	0.966560
0.3	0.921230	0.913684	0.925859	0.913684	0.074141	0.883191	0.913684	0.883191	0.945896
0.4	0.925712	0.898526	0.942392	0.898526	0.057608	0.905388	0.898526	0.905388	0.938030
0.5	0.928274	0.879579	0.958150	0.879579	0.041850	0.928032	0.879579	0.928032	0.928411
0.6	0.924752	0.856000	0.966934	0.856000	0.033066	0.940768	0.856000	0.940768	0.916279
0.7	0.919629	0.829053	0.975200	0.829053	0.024800	0.953511	0.829053	0.953511	0.902894
0.8	0.904419	0.773895	0.984500	0.773895	0.015500	0.968388	0.773895	0.968388	0.876495
0.9	0.886808	0.717474	0.990700	0.717474	0.009300	0.979310	0.717474	0.979310	0.851087

E. Model Evaluation - Find the optimum cutoff



The optimum cutoff is 0.3
(intersect of accuracy, sensitivity and specificity)



The area under the curve of the ROC is 0.97
which is quite good.

II. Build Logistic Regression Model

E. Final Model Evaluation (optimum cutoff=0.3)

Evaluate the TRAIN dataset

```
============Confusion Matrix:==========
[[3584  287]
 [ 205 2170]]
============Metrics:==========
  1. Accuracy score:  0.9212295869356388

  2.1 sensitivity (TPR):  0.9136842105263158
  2.2 specificity:  0.9258589511754068

  3.1 TPR:  0.9136842105263158
  3.2 FPR:  0.07414104882459313

  4.1 Precision:  0.8831908831908832
  4.2 recall:  0.9136842105263158

  5.1 Possitive_Prediction_ratio:  0.8831908831908832
  5.2 Negative_Prediction_ratio:  0.9458960147796253
```

Evaluate the TEST dataset

```
============Confusion Matrix:==========
[[1555  129]
 [  79  915]]
============Metrics:==========
  1. Accuracy score:  0.9223300970873787

  2.1 sensitivity (TPR):  0.920523138832998
  2.2 specificity:  0.9233966745843231

  3.1 TPR:  0.920523138832998
  3.2 FPR:  0.07660332541567696

  4.1 Precision:  0.8764367816091954
  4.2 recall:  0.920523138832998

  5.1 Possitive_Prediction_ratio:  0.8764367816091954
  5.2 Negative_Prediction_ratio:  0.9516523867809058
```

All metrics are good for both TRAIN and TEST dataset, equivalent between TRAIN and TEST. Especially the accuracy score, precision and recall are important metrics for this business:

- Precision: number of predicted YES correctly over Total predicted YES (converted=1)
- Recall: number of predicted YES correctly over Total actual YES (converted=1)

III. Interpret The Business Results

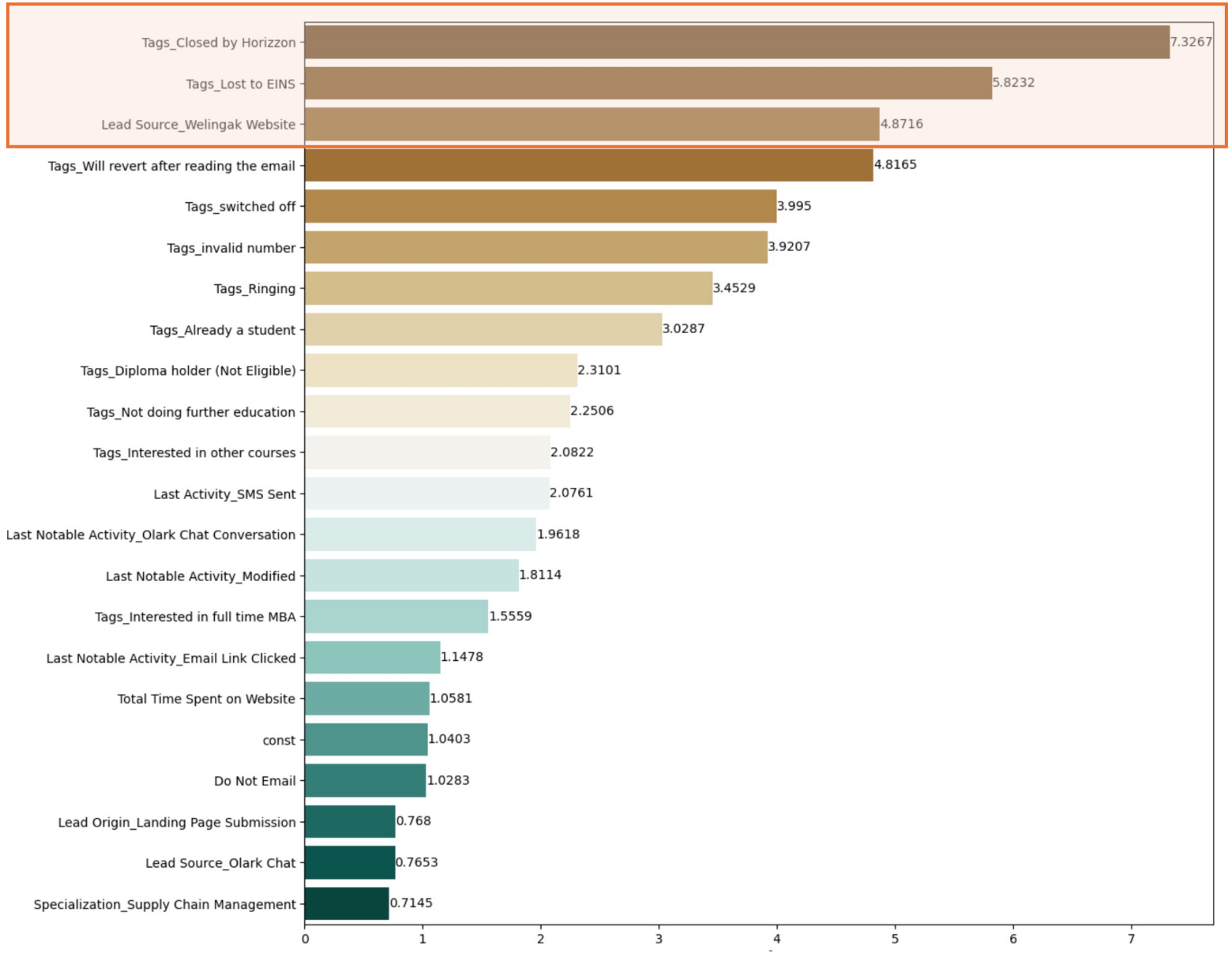
III. Interpret the results

Top 3 Variables Contribute Most

Top 3 Most Contributed Variables

These variables have highest absolute coefficients, so they are most contributed variables:

- Tags_Closed by Horizzon (coef = 7.3267)
  - Positive contribution
  - If the lead has tag “Closed by Horizzon”, there is a high probability to convert to customer. So company should focus on such leads
- Tags_Lost to EINS (coef = 5.8232)
  - Positive contribution
  - If the lead has tag “Lost to EINS”, there is a high probability to convert to customer. So company should focus on such leads
- Lead Source_Welingak Website (coef = 4.8716)
  - Positive contribution
  - If the source of leads is “Welingak Website”, there is a high probability to convert to customer. So company should focus on such leads



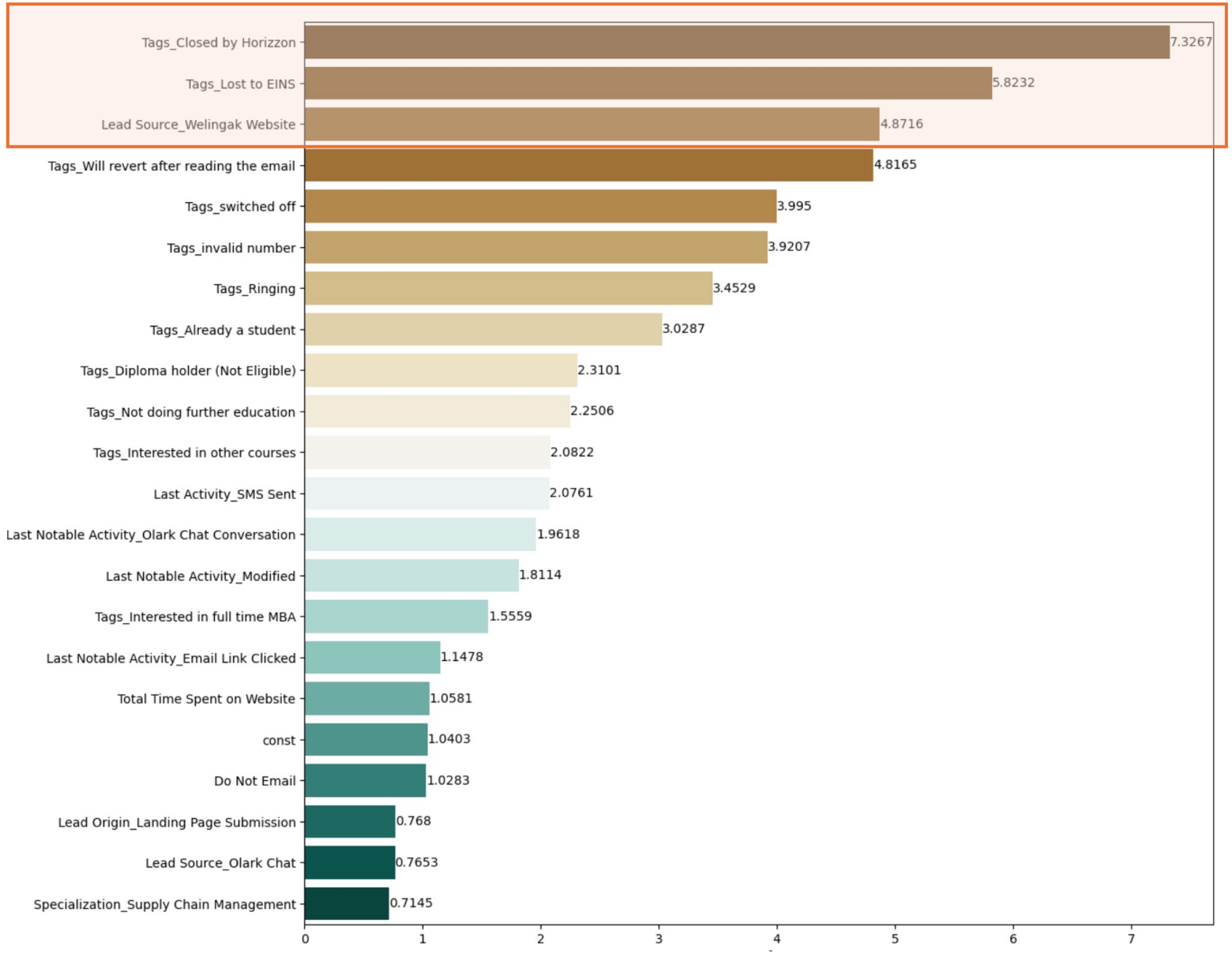
III. Interpret the results

Top 3 Categorical/Dummy Variables

Top 3 Most Contributed Variables

The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are :

- Tags_Closed by Horizzon (coef = 7.3267)
  - Positive contribution
  - If the lead has tag “Closed by Horizzon”, there is a high probability to convert to customer. So company should focus on such leads
- Tags_Lost to EINS (coef = 5.8232)
  - Positive contribution
  - If the lead has tag “Lost to EINS”, there is a high probability to convert to customer. So company should focus on such leads
- Lead Source_Welingak Website (coef = 4.8716)
  - Positive contribution
  - If the source of leads is “Welingak Website”, there is a high probability to convert to customer. So company should focus on such leads



III. Interpret the results

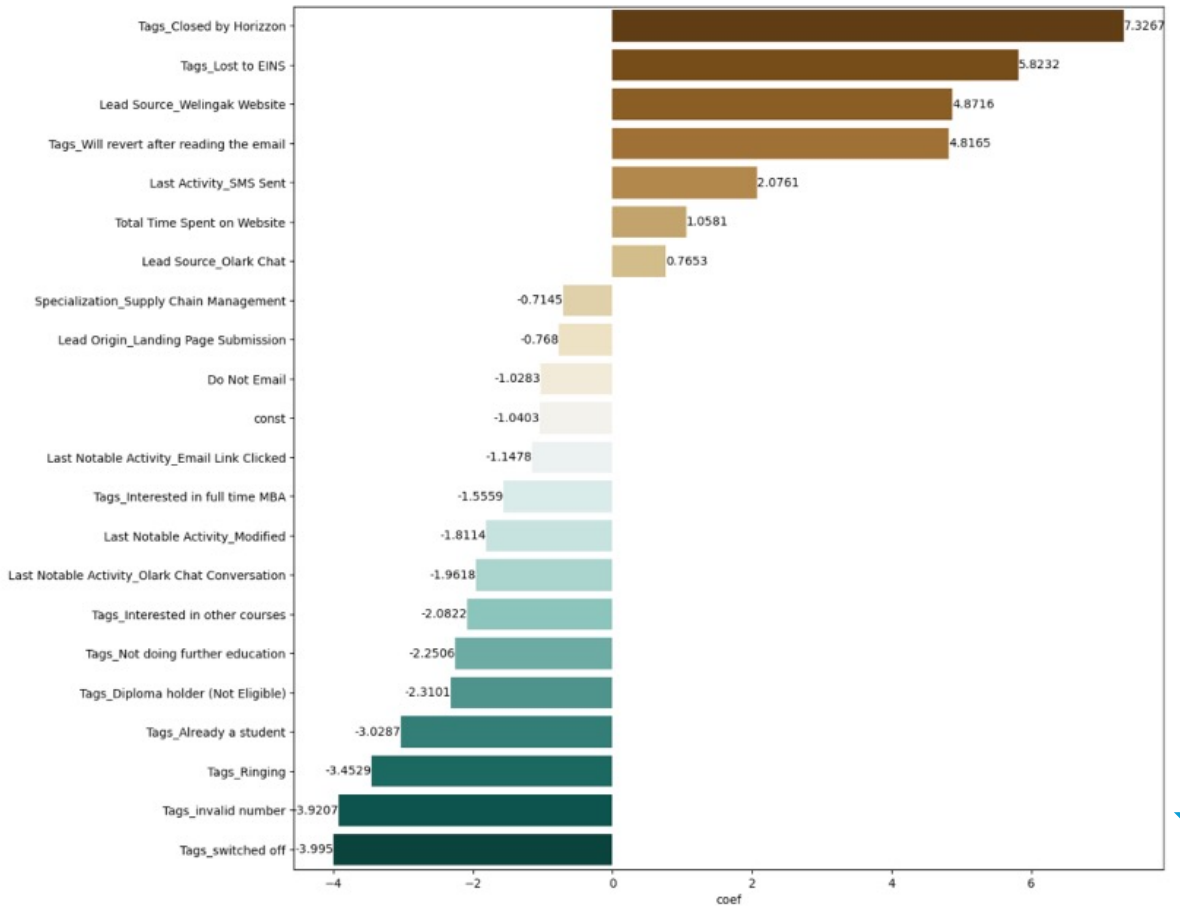
Strategy to prioritize the Leads to increase the conversions

Priority 1: Highest Lead Score

Priority 2 (Same Lead Score): Leads who have most contributed feature

- Priority 1: Target leads who have the higher predict lead_score. The lead_score is in range 0..100. The higher lead_score, the higher priority to make the phone calls.

- Priority 2: Among the leads who have the same lead_score, we will priority for the leads who have most contribution features.
  - The higher coef, the higher contribution to conversion probability. The below chart sorting descending on the contribution.
  - For the categorical/dummy variables (eg: “Tags_Closed by Horizzon”), if a lead have possitive on this feature, this lead will have higher chances to convert than a lead don’t have this feature
  - For the numerical variables (such as Total Time Spent on Website…), we will prioritize for the lead have higher coef * numerical_value



III. Interpret the results

Strategy to prioritize the Leads to reduce useless calls

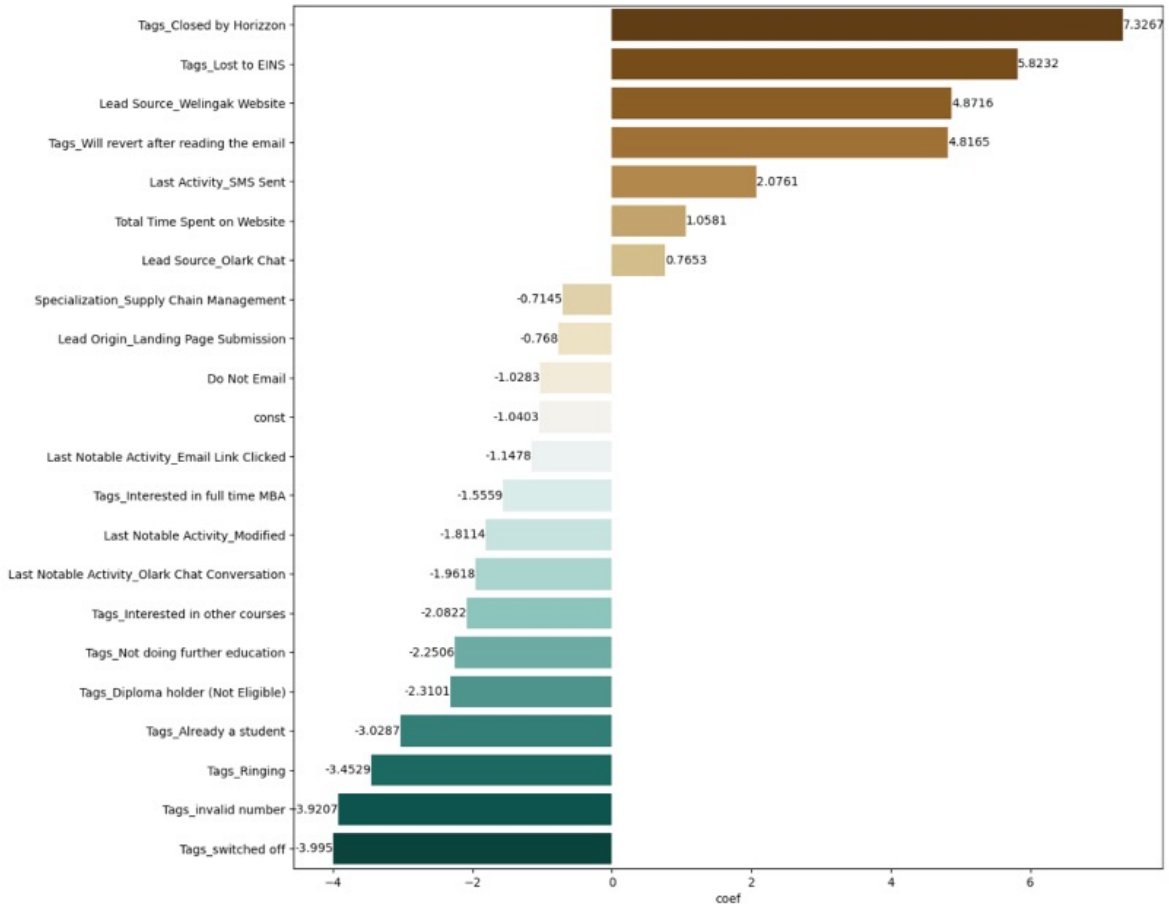
To minimize the useless phone calls, we will prevent to call the leads who have negative coef features (eg: Tags_switched off, Tags_invalid number…).

- These features have negative coef, means that it will have negative impact on the conversions.
- For categorical variables with negative coef: the higher negative coef, the higher higher prevention to make the call these leads.
- For the numerical variables with negative coef: we consider

magnitude of a feature = coef * numeric_feature_value

The higher “magnitude of a feature”, the higher prevention to make the phone call.

- For the feature priority, from above chart, we will go from bottom to top (lowest coef to highest coef)



Prevent to call the Leads who have the Bottom features

The End