Improving Lead Conversion Rate at X Education

Using Logistic Regression to Identify Hot Leads

Submitted by :

Richa Chaturvedi

Manoj Kalbugi

Mayuri Pande

Problem Statements & Objectives of Case Study

Problem statement

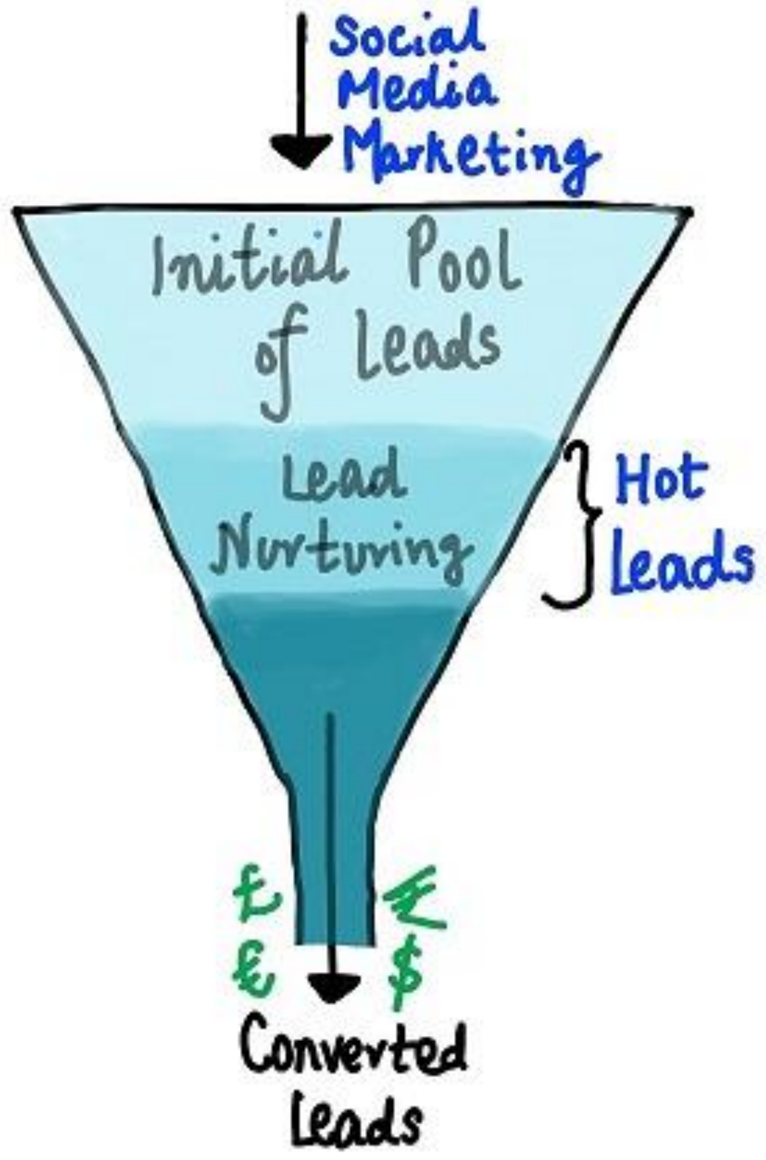
- X Education acquires huge number of leads everyday, however current **lead conversion rate is only ~30%**.
- The sales team spends significant time and effort reaching out to all leads, many of whom do not convert.
- Company wants to **identify high-potential leads (also called Hot Leads)** to improve efficiency and conversion rates.
- The goal is to increase the **lead conversion rate to around 80%** by prioritising the most promising leads.

Business Impacts:

- **Higher conversion rates** → More revenue.
- **Optimized sales efforts** → Focus on quality leads.
- **Data-driven decision-making** → Efficient resource allocation.

Objectives

- Develop a Lead Scoring Model using **Logistic Regression**.
- **Identify Hot Leads** with high probability of conversion.
- Improve lead **conversion rate to 80%**.
- Enhance sales team efficiency by focusing on quality leads.



Funnel Diagram: Typical lead conversion process with very low conversion rate.

Approach and Methodology

Major steps involved in case study:

- Data Source for Analysis
- Data Preprocessing
- Exploratory Data Analysis
- Data preparation.
- Model Selection & Training.
- Lead Scoring & Model Performance.
- Validation of model.

Data Source

- Checking provided data files.
- Basic inspection of the CSV file and data dictionary.

Data Preprocessing

- Checking for duplicates,
- Handling missing values and removing features with excessive missing data.,
- Data imputation.
- Identifying and resolving data issues.
- Detecting and removing outliers.

Exploratory Data Analysis

- Univariate Analysis. Bivariate Analysis (Split data analysis).
- Extracting insights from numerical and categorical features.

Data preparation

- Encoding categorical data and generating dummy features. Feature Scaling.
- Removing non-feature columns.
- Splitting the data into Test and Train data set.
- Scaling test and train datasets.

Model Selection & Training

- Using Logistic Regression to predict lead conversion probability.
- Feature selection using Recursive Feature Elimination (RFE).
- Refining the model by eliminating features based on p-values and Variance Inflation Factor (VIF).
- Generating the final model and predicting the target variable using the training data.

Lead Scoring, & Model performance

- Classifying leads using a probability threshold of 0.5 and evaluation of Model performance using different metrics- Accuracy, Sensitivity, specificity, precision, recall, etc (on train dataset predictions).
- Assign lead scores → Rank leads from 0 to 100
- Determining the optimal probability cutoff using Sensitivity-Specificity and Precision-Recall analysis.
- Re-evaluate model performance with optimal cutoffs and finalising the optimal cutoff.

Validation of model.

- Predicting the target variable using test data
- Evaluating model performance.

Data description

Data Source & Dataset Overview:

- There are two data files are provided: 1) Leads.csv having and 2) Leads Data Dictionary.xlsx.
- The dataset contains approximately 9240 leads with multiple attributes. There are 37 columns.
- Each lead has various features that may influence conversion.
- **Target Variable:** Column **‘Converted’** has entries 0 and 1.
    - **Converted (1):** Lead converted into a customer.
    - **Not Converted (0):** Lead did not convert.

Features in the Dataset: 36 features and 1 target variable. Some of these are listed below:

- **Lead Source:** Google, Facebook, Direct Traffic, Referral, etc.
- **Total Time Spent on Website:** Duration of user engagement.
- **Total Visits:** Number of times the lead visited the website.
- **Last Activity:** Last recorded interaction (Email Opened, Olark Chat Conversation, etc., etc.).
- **Other Attributes:** Lead Origin, Industry, City, Specialization, etc.

Data Cleaning Considerations

- Many categorical features have "Select" as a level, which is equivalent to a missing value and should be treated accordingly.
- Missing values need to be imputed or removed based on relevance.
- Categorical variables need to be converted into numerical form for modeling

Data preprocessing

Several pre-processing steps are followed as described in flowchart.

- No duplicate data found.
- Four columns such as ‘Lead Quality’, 'Asymmetrique Activity Index’, etc., have >30% missing data: We have dropped these features.
- Bar plot of each categorical variable shows 3 major data issues:

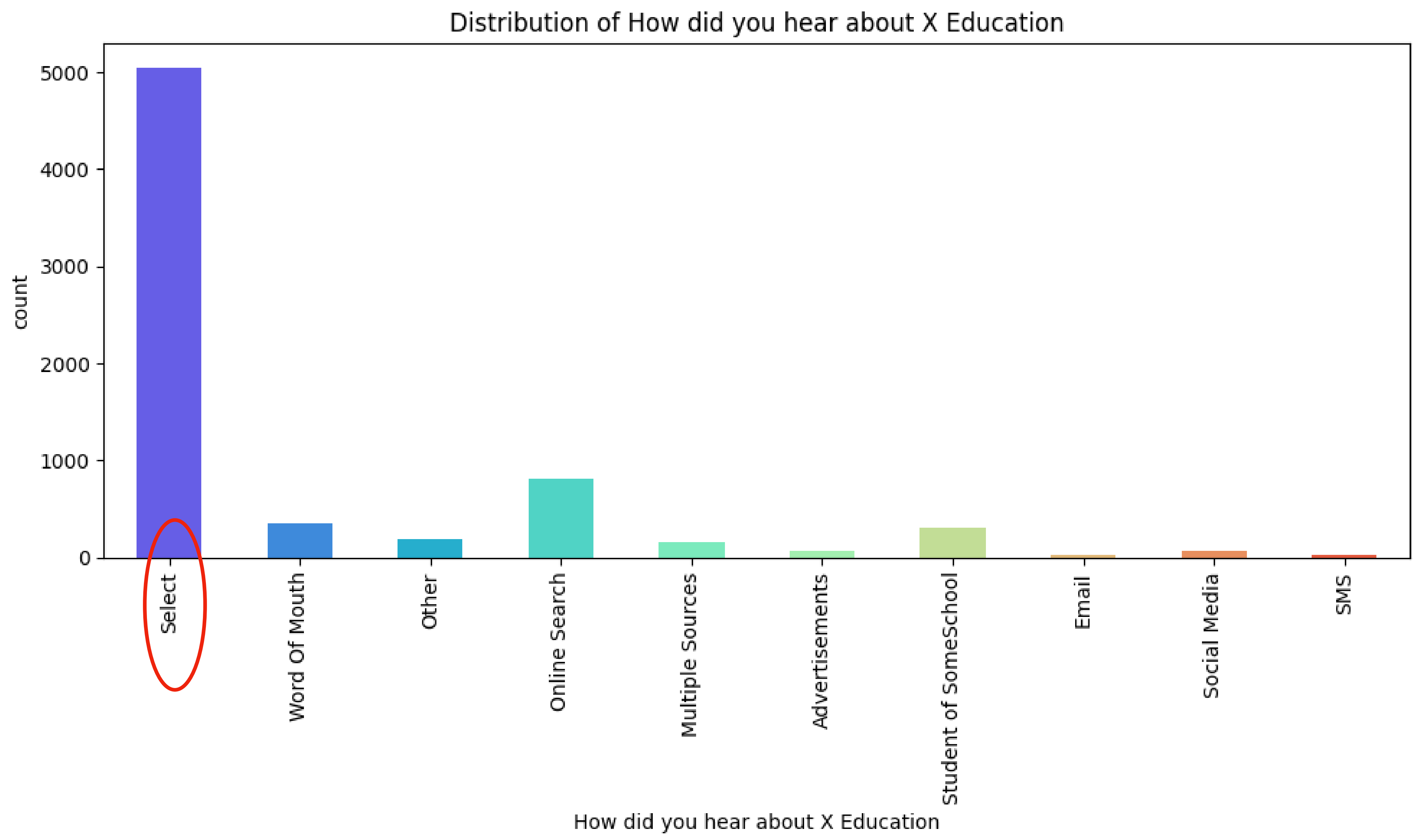
1) Four categorical variables have a value called “Select.” We consider these as missing data and Converted these values as NULL since the customer has not selected any options for these columns.

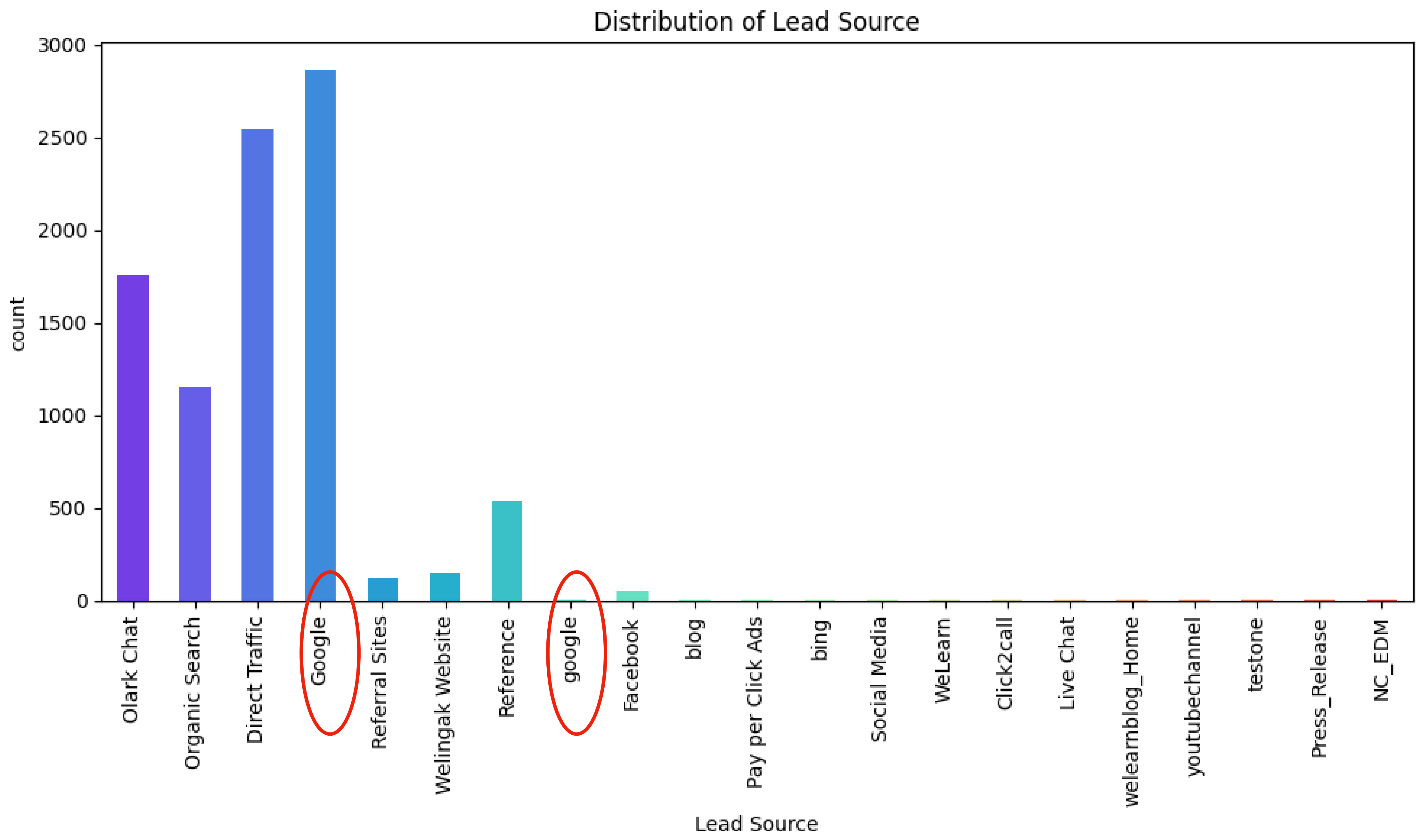
2) For Lead source values ‘Google’ is in both upper and lower cases , Updated lowercase google to uppercase.

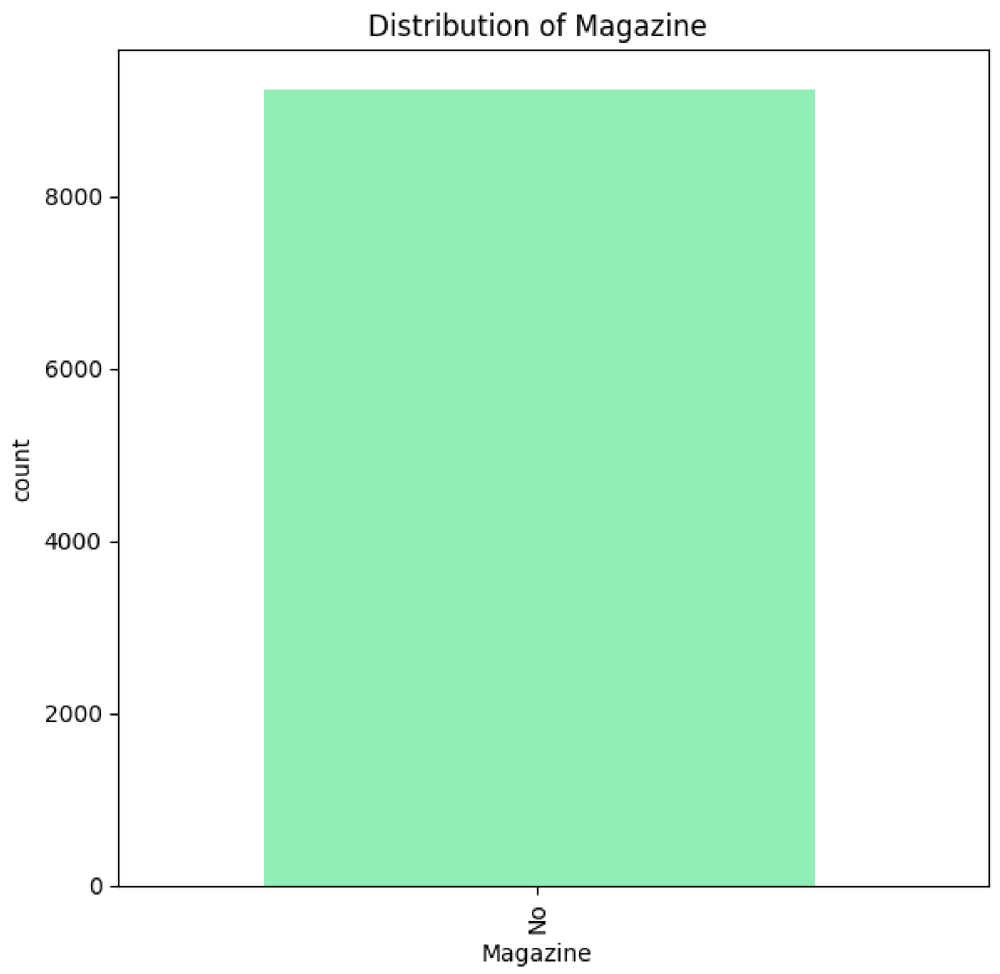
3) Five features such as ‘Magazine’ ,’Receive More Updates About Our Courses, etc., have only ‘NO’ entry. We dropped these columns as they have zero entropy I.e., no information.

- Seven features have <30% missing data are: Country , Lead Source , Total Visits , Page Views Per Visit , Last Activity, current occupation and What matters most to you in choosing a course.
- Imputation:

Categorical —> mode.
Numerical —> median value



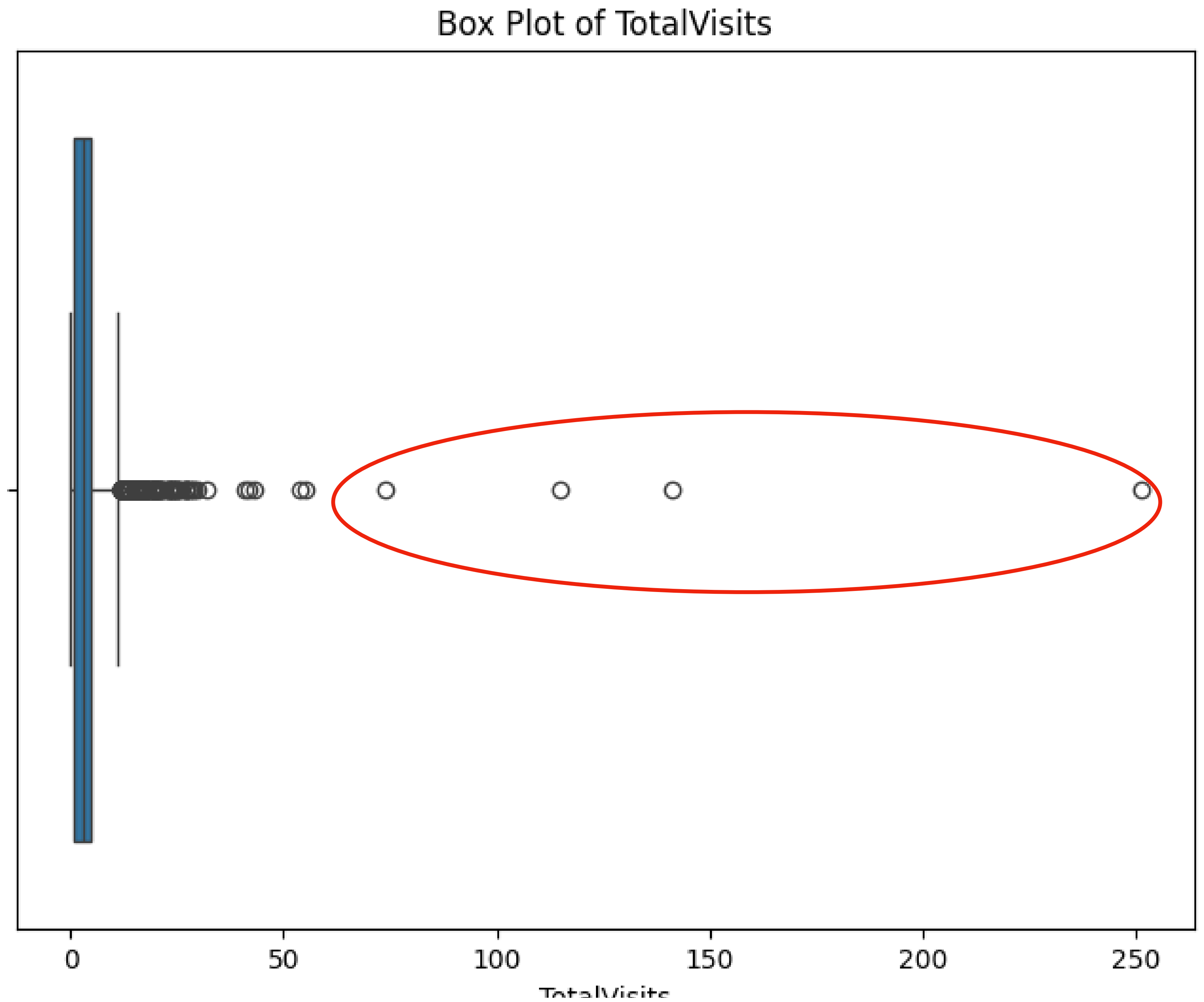


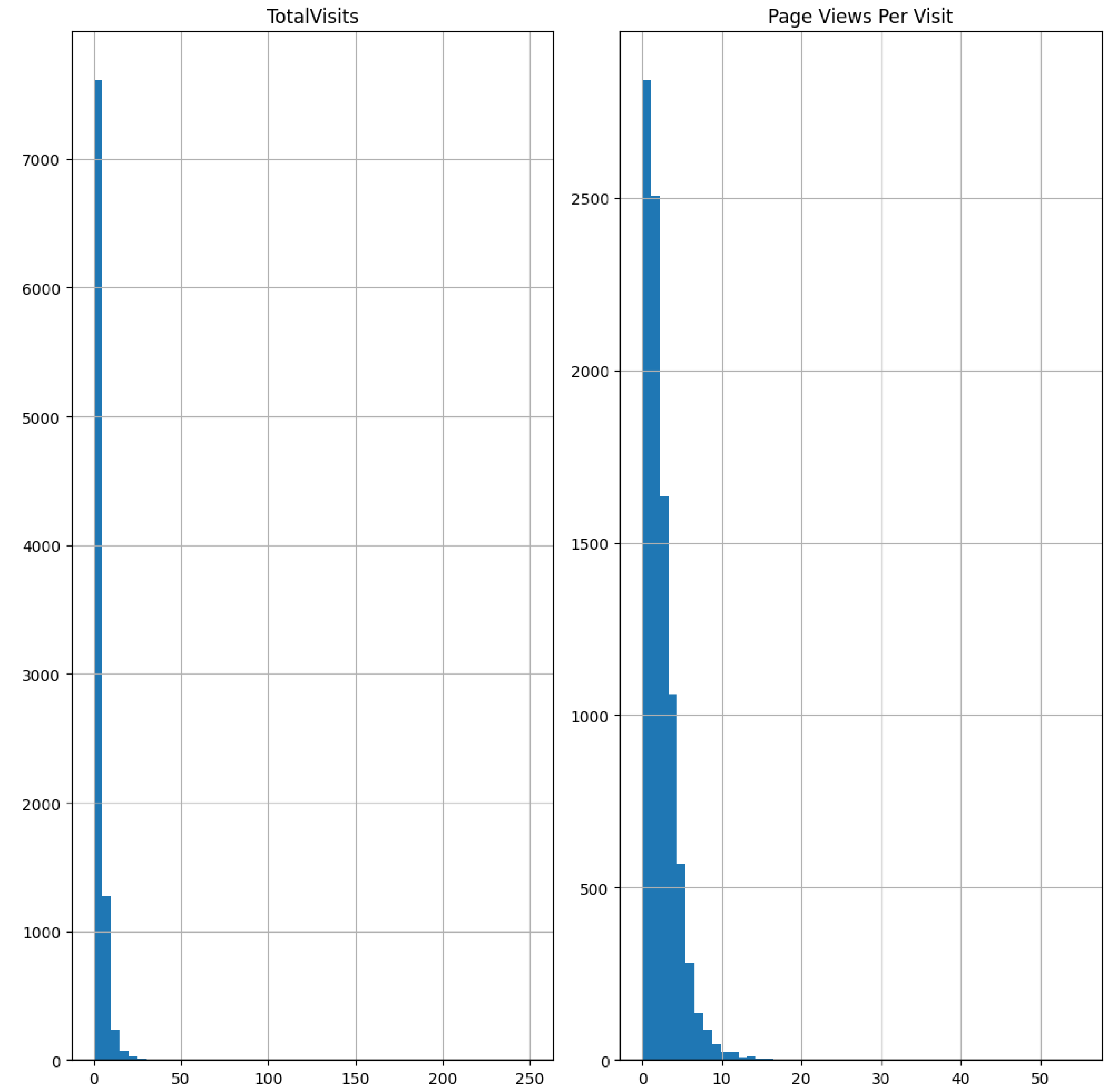


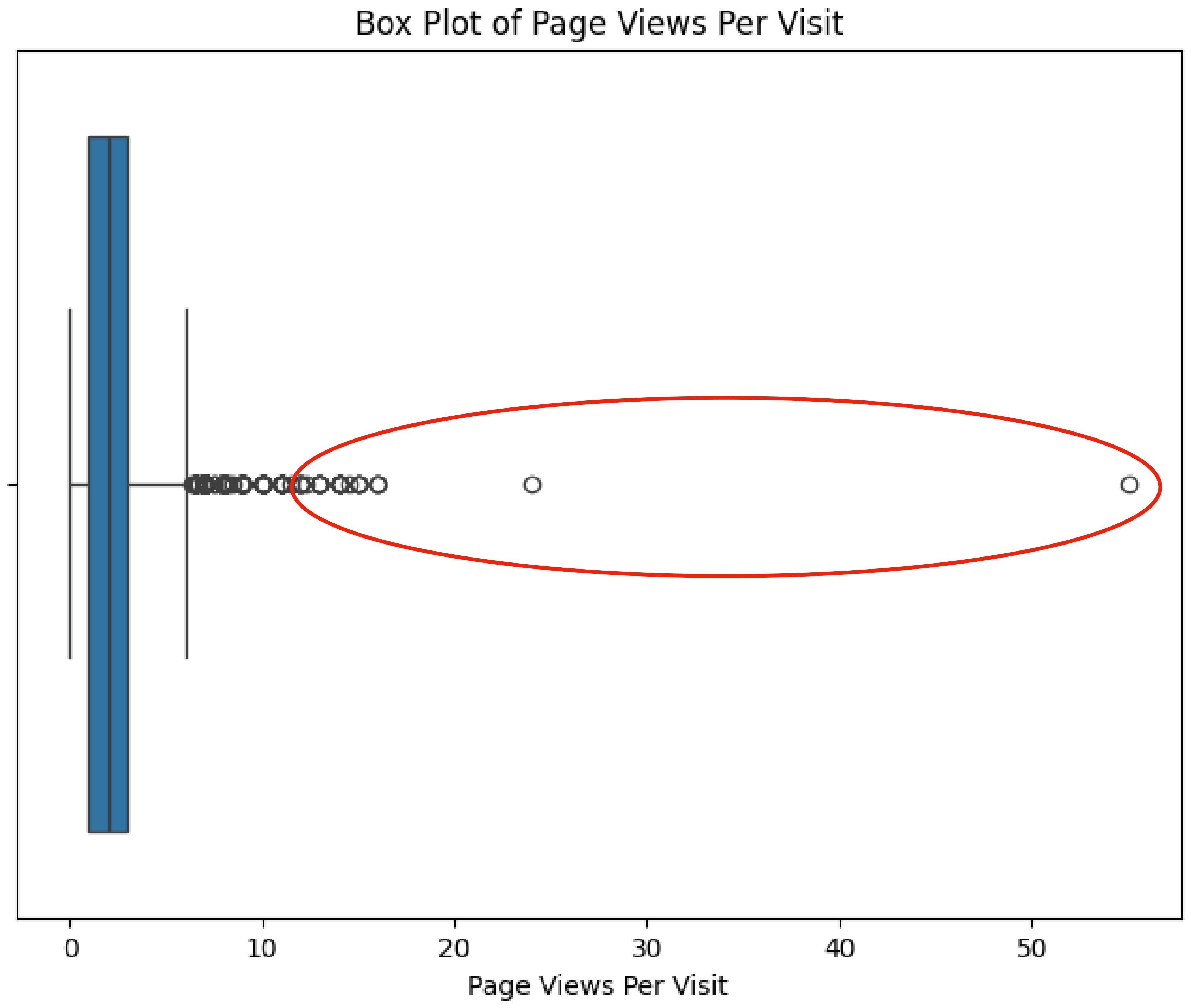
Data preprocessing

Outliers

- Box plot of 2 numerical features clearly showed outliers.
- Histogram of these shows highly left skewed data.
- Thus we used Q1-1.5xIQR and Q3+1.5xIQR as lower and upper bound, rest are assumed outlier and removed. Here IQR=Q3-Q1.



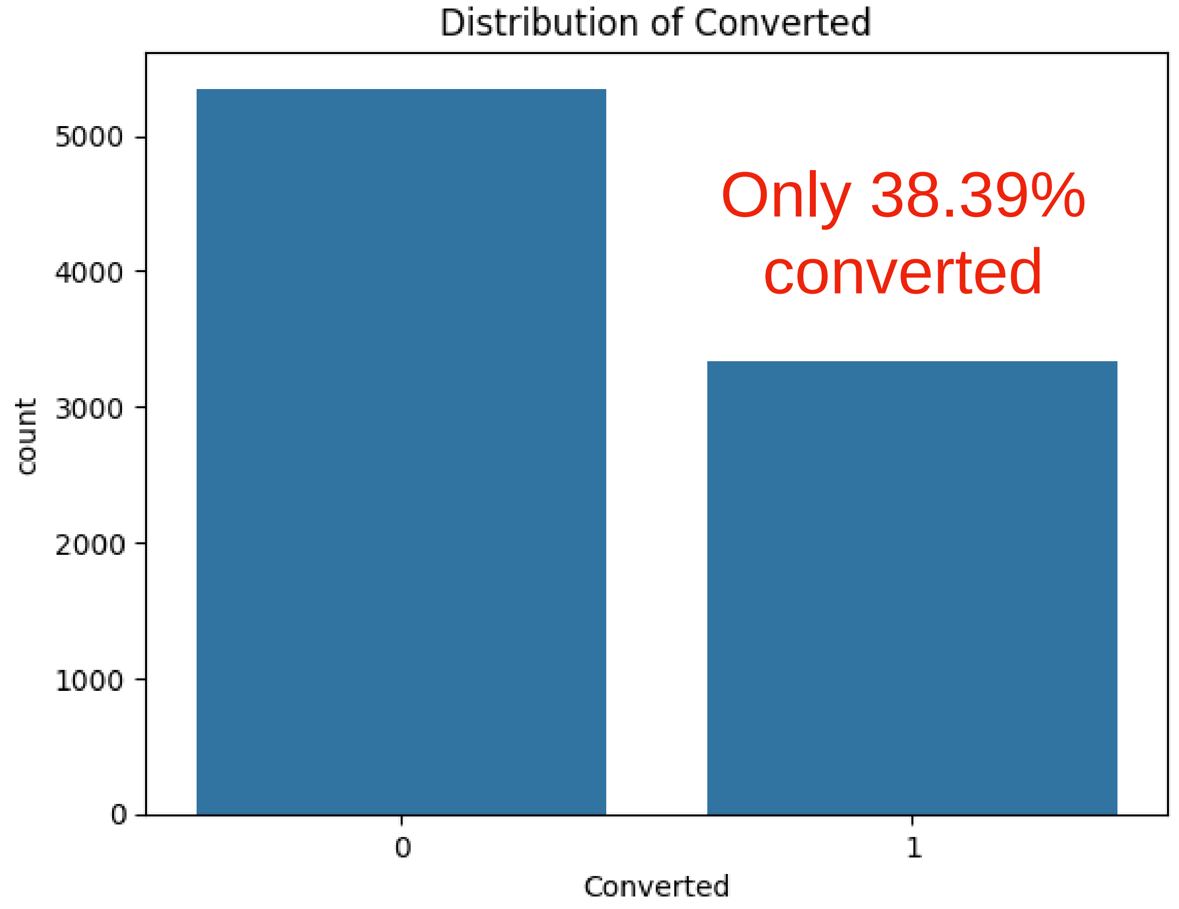




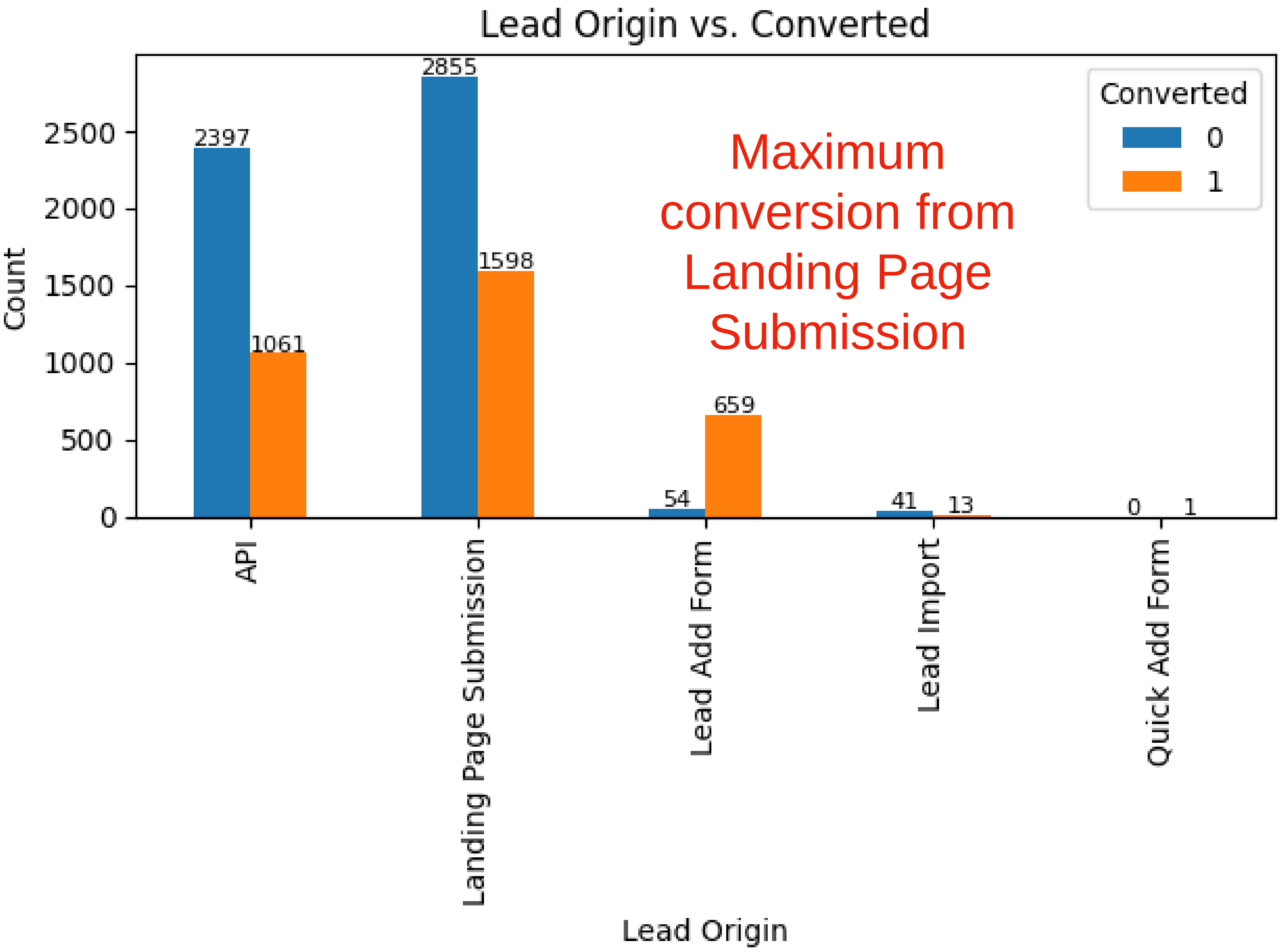
Exploratory Data Analysis

Bivariate Analysis with split bar plots.

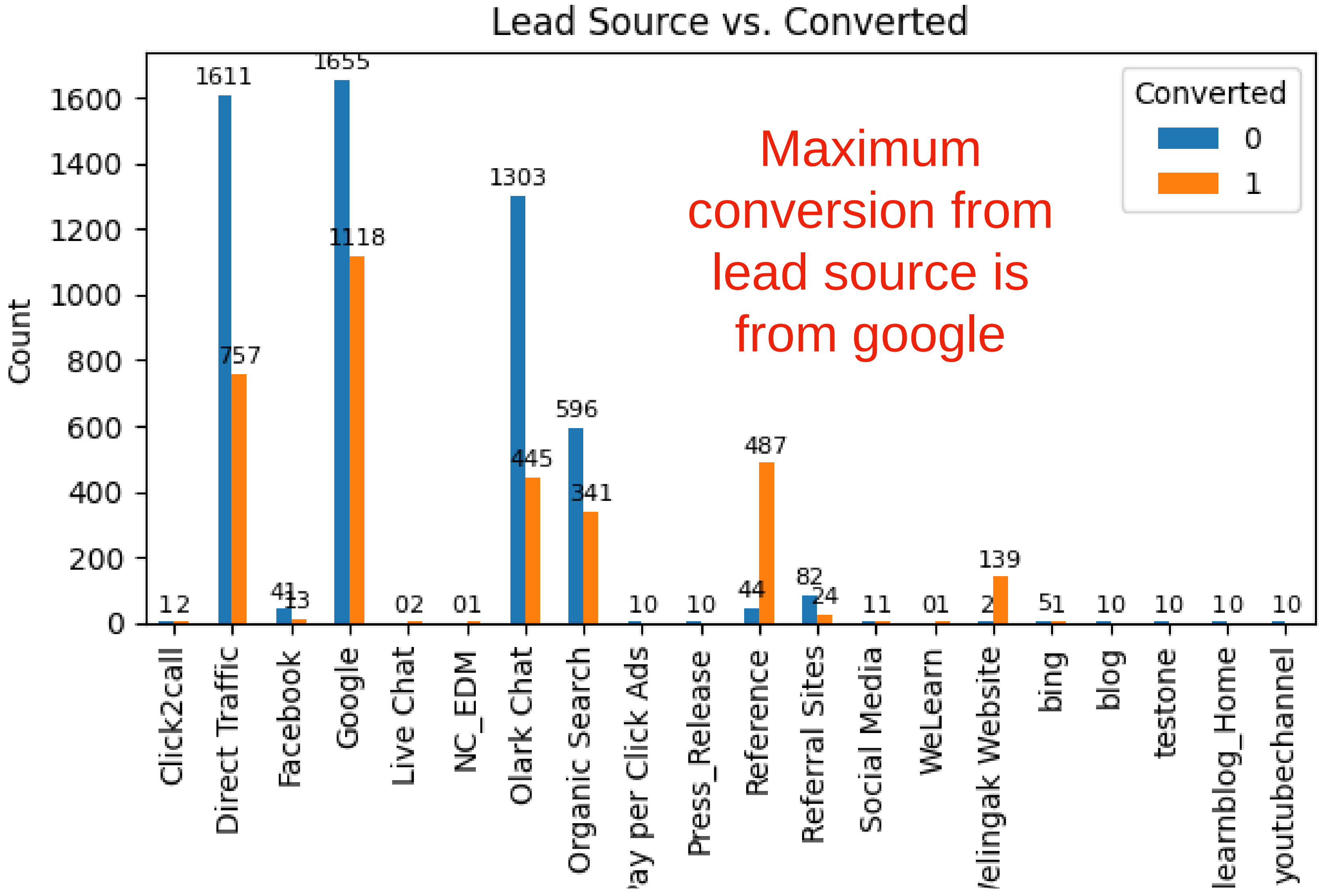
- Target ‘Converted’: Count of non-converted > converted case.
- For numerical and categorical, split box plot and split bar plot is analysed.



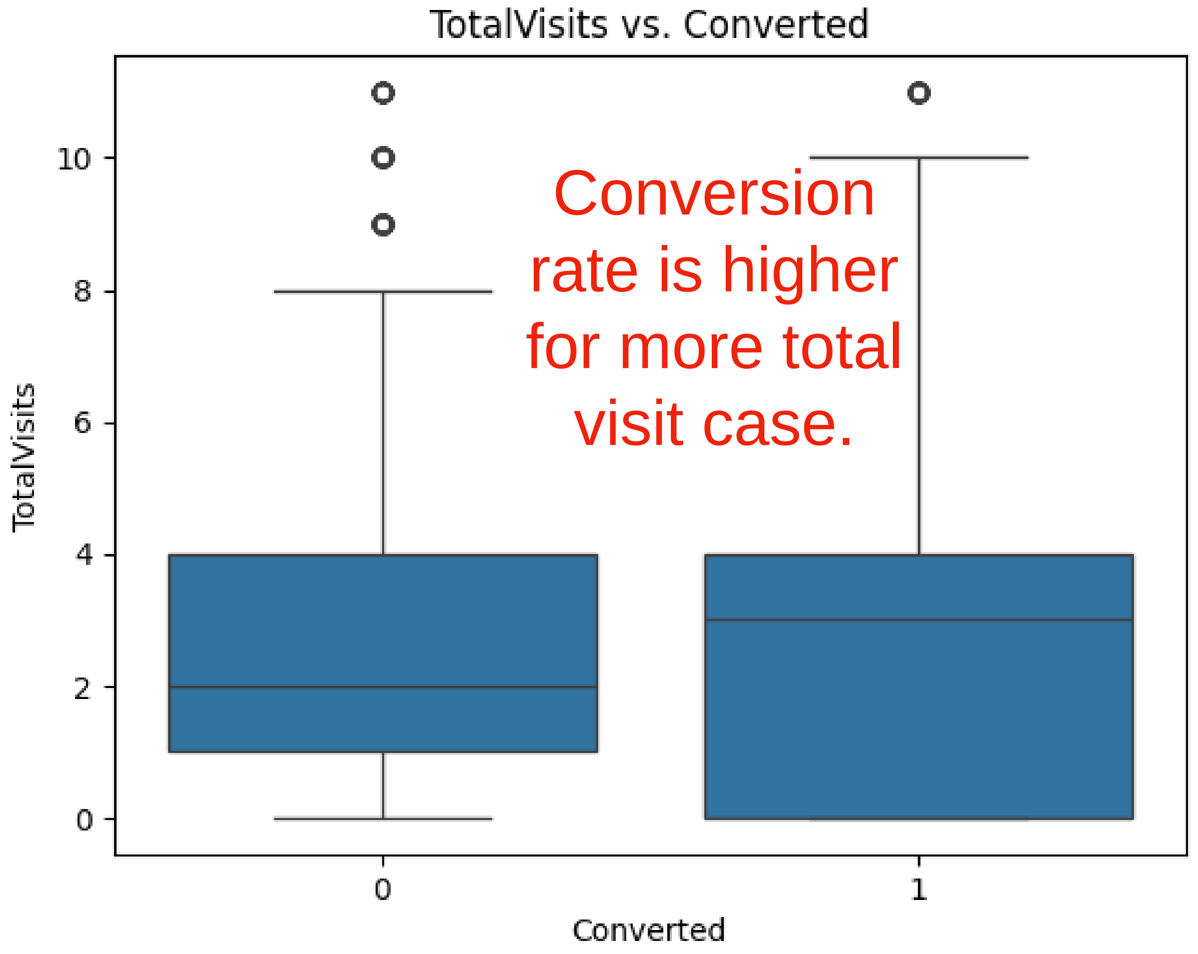
Only 38.39% converted



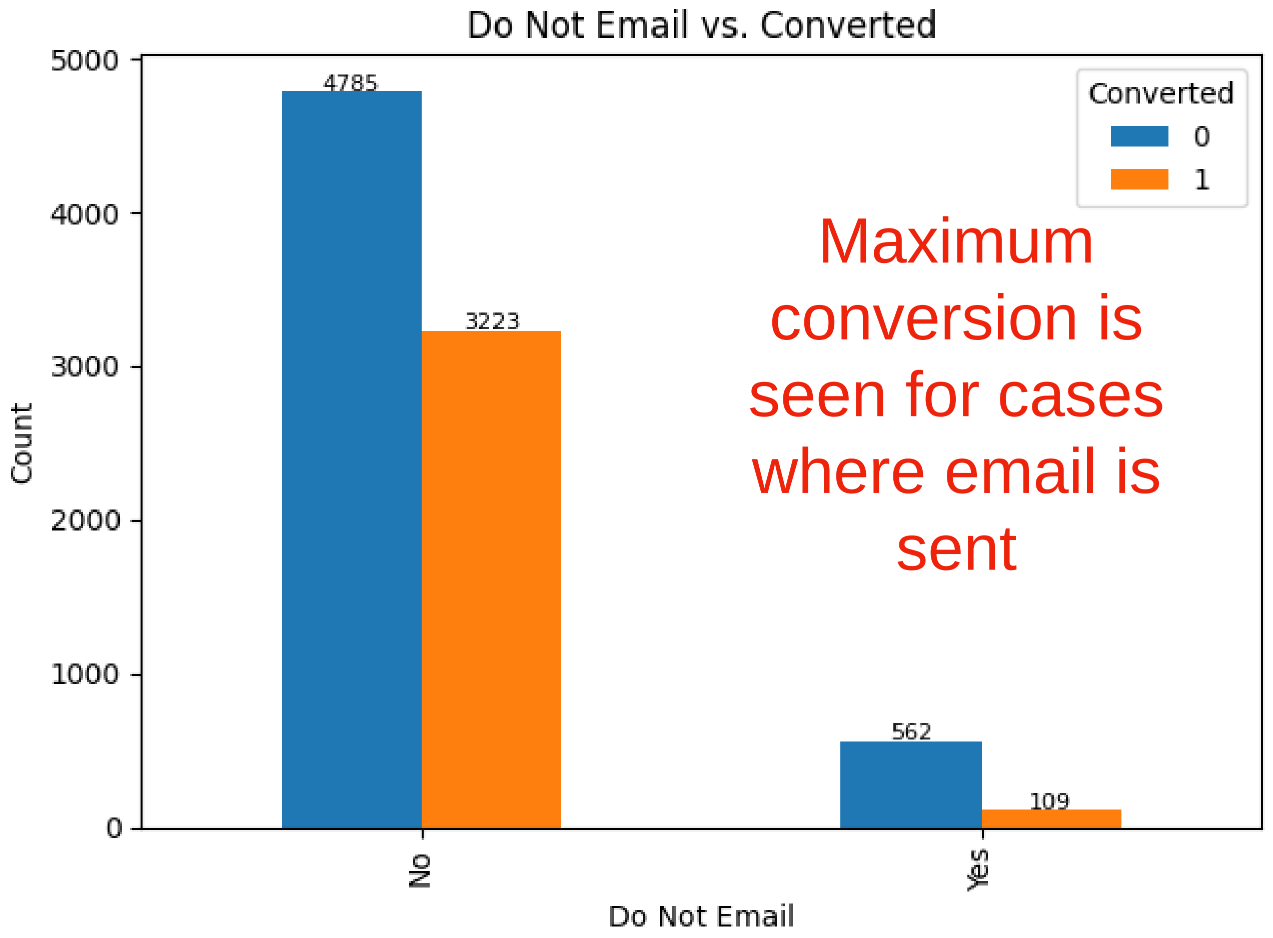
Maximum conversion from Landing Page Submission



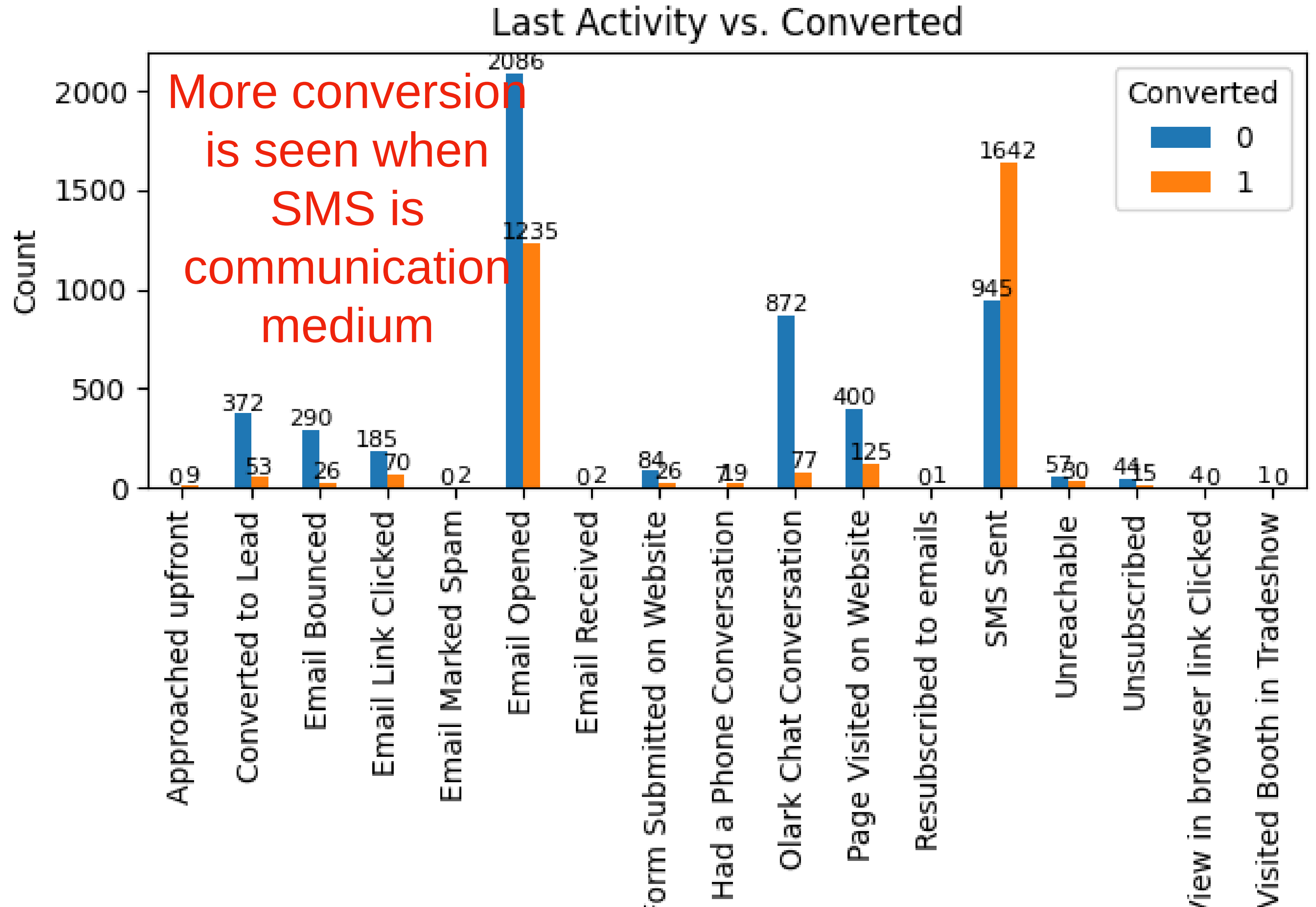
Maximum conversion from lead source is from google



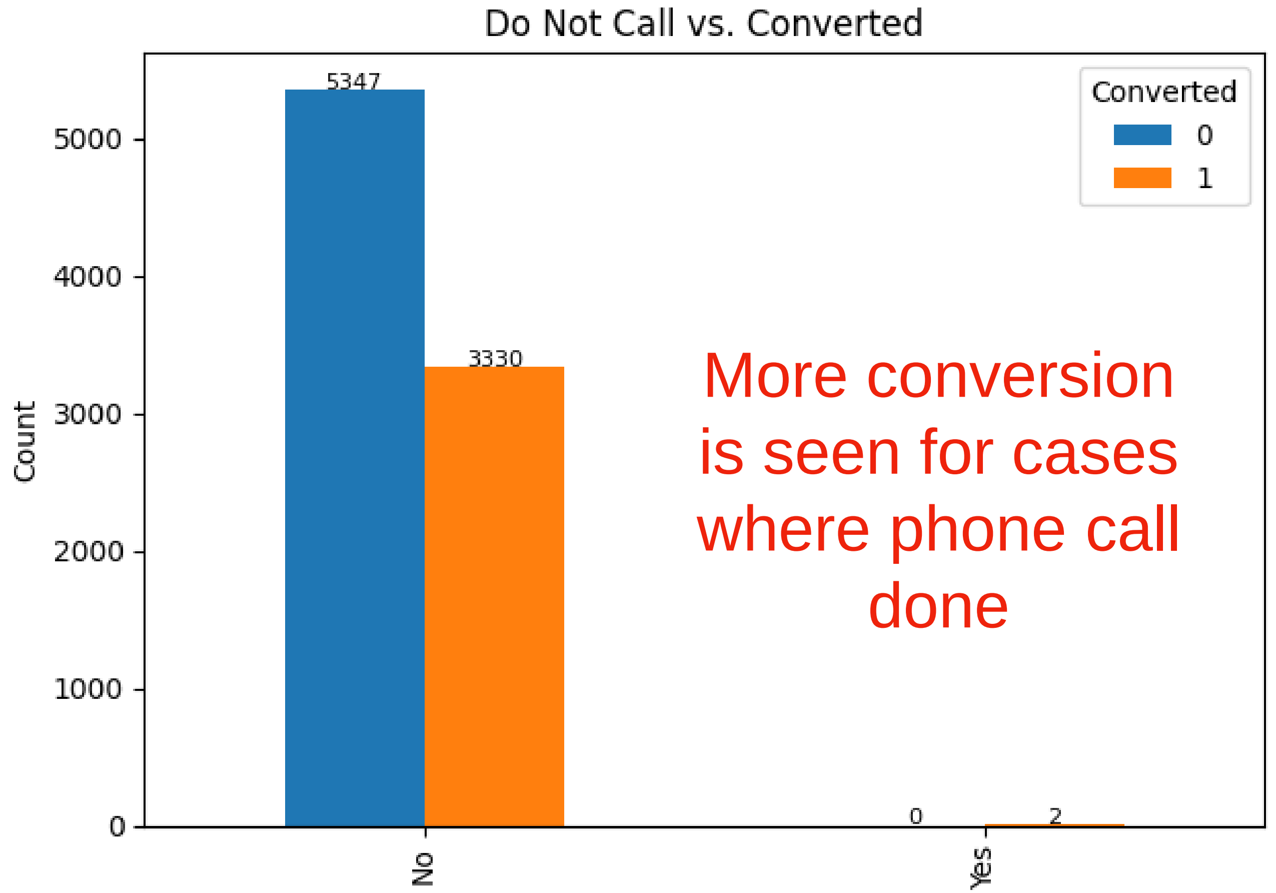
Conversion rate is higher for more total visit case.



Maximum conversion is seen for cases where email is sent



More conversion is seen when SMS is communication medium

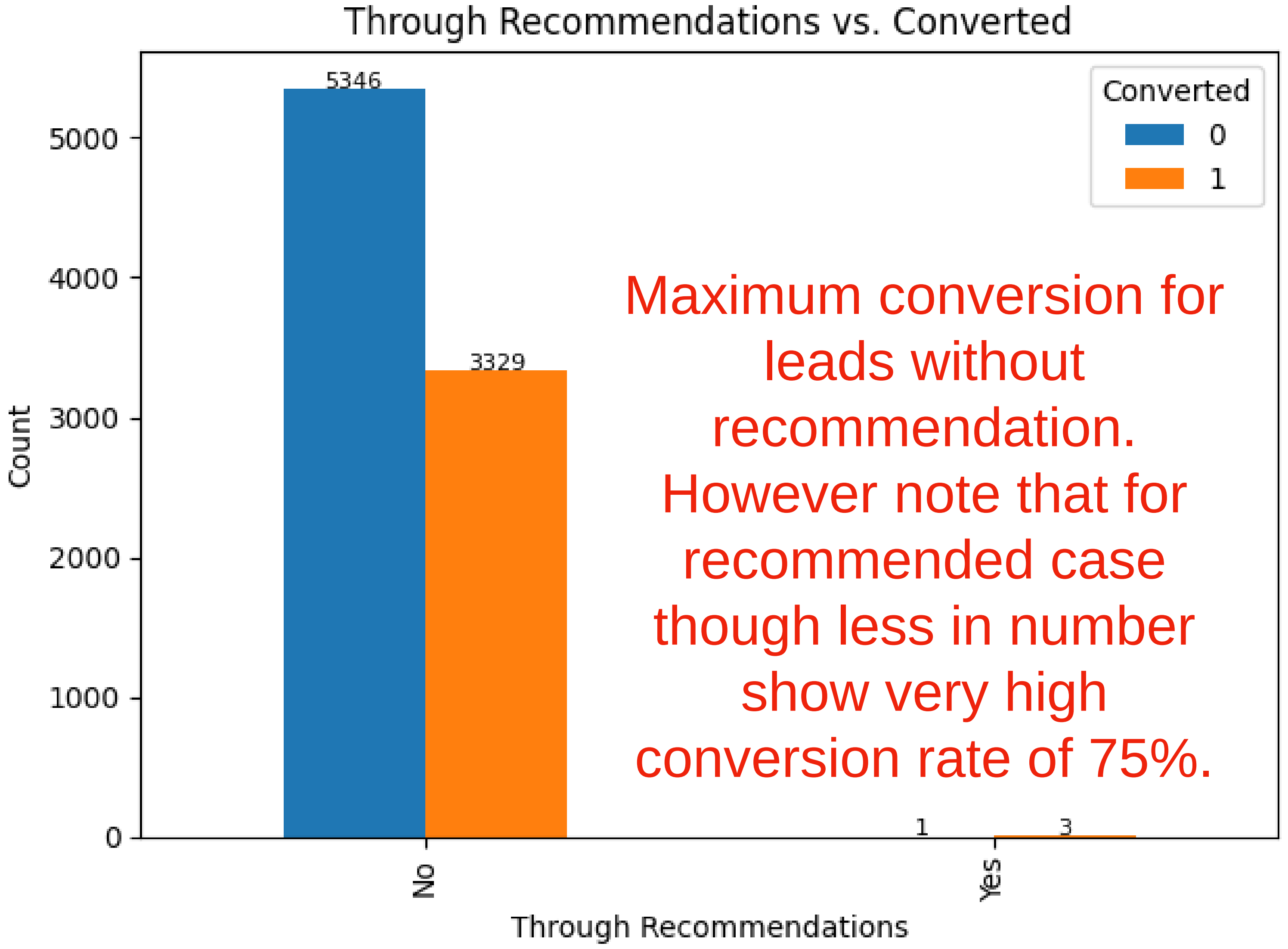


More conversion is seen for cases where phone call done

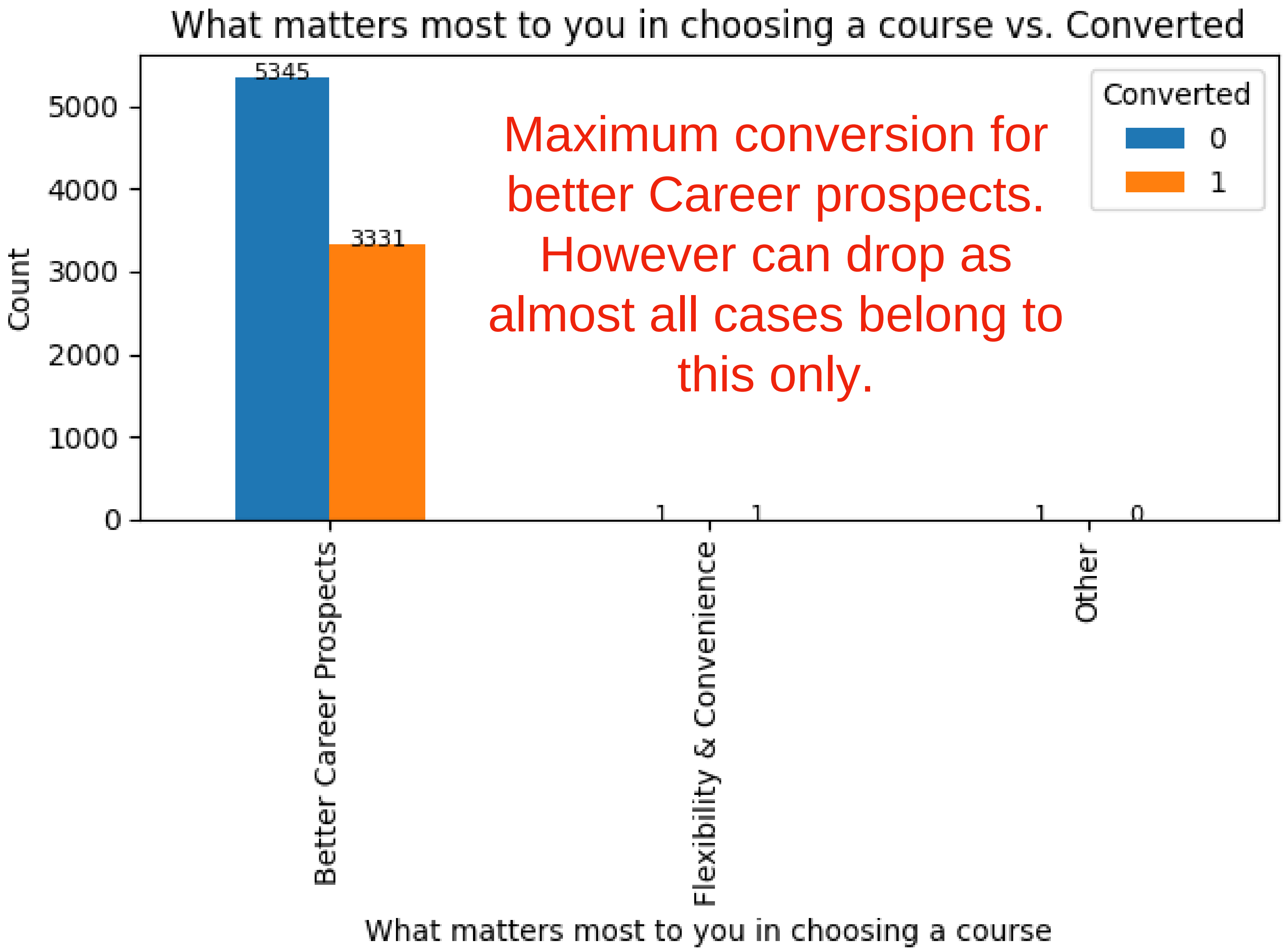
Exploratory Data Analysis

Bivariate Analysis with split bar plots.

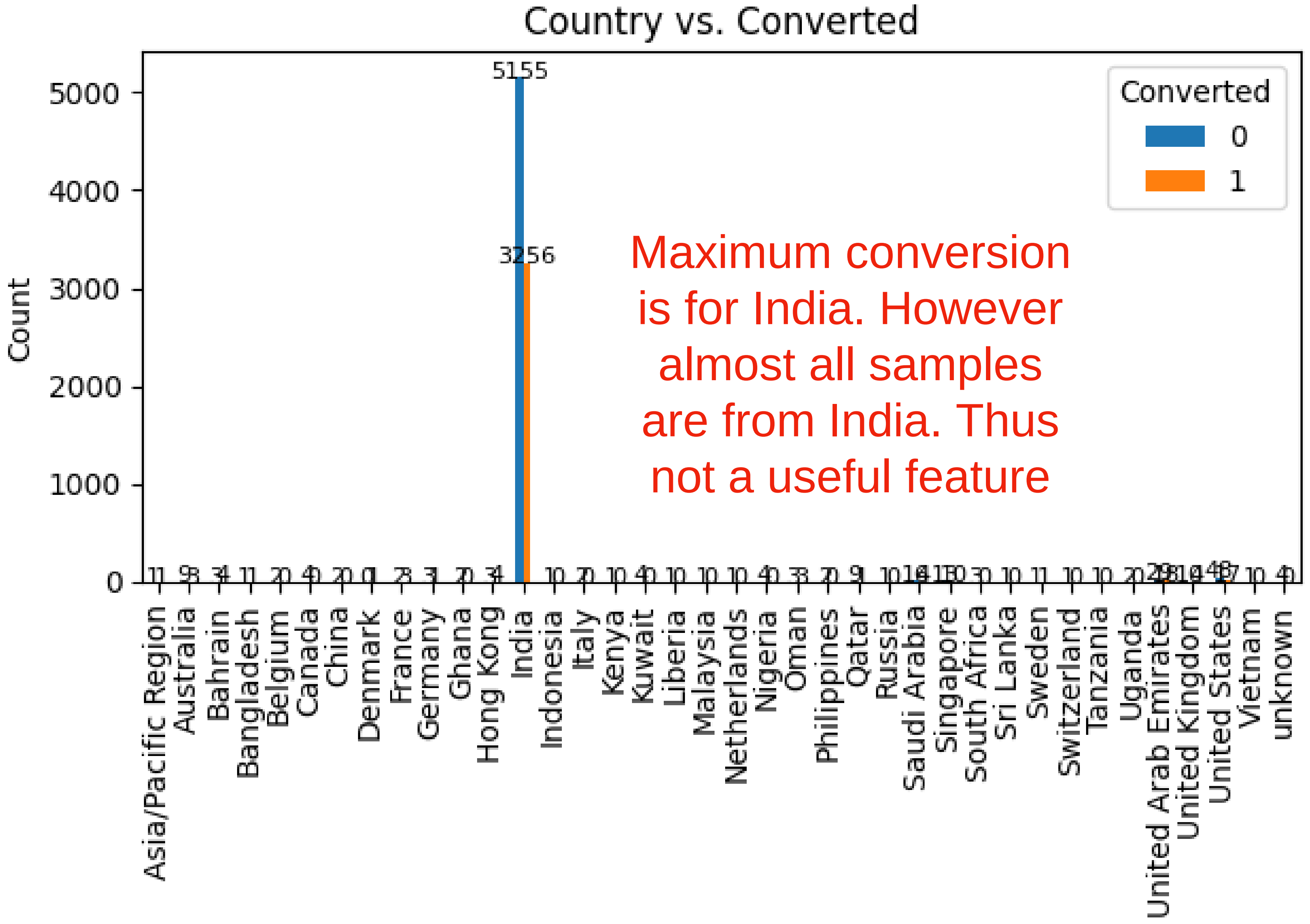
- Multiple features like ‘Newspaper Activity’ show only ’NO’ entry, thus are irrelevant, thus removed.
- Multiple features. Like ‘Country’ shows almost all samples from single category, such features also dropped.
- Iterestigly seen that Working professionals coverted at 75%.
- SMS is effective communication medium.



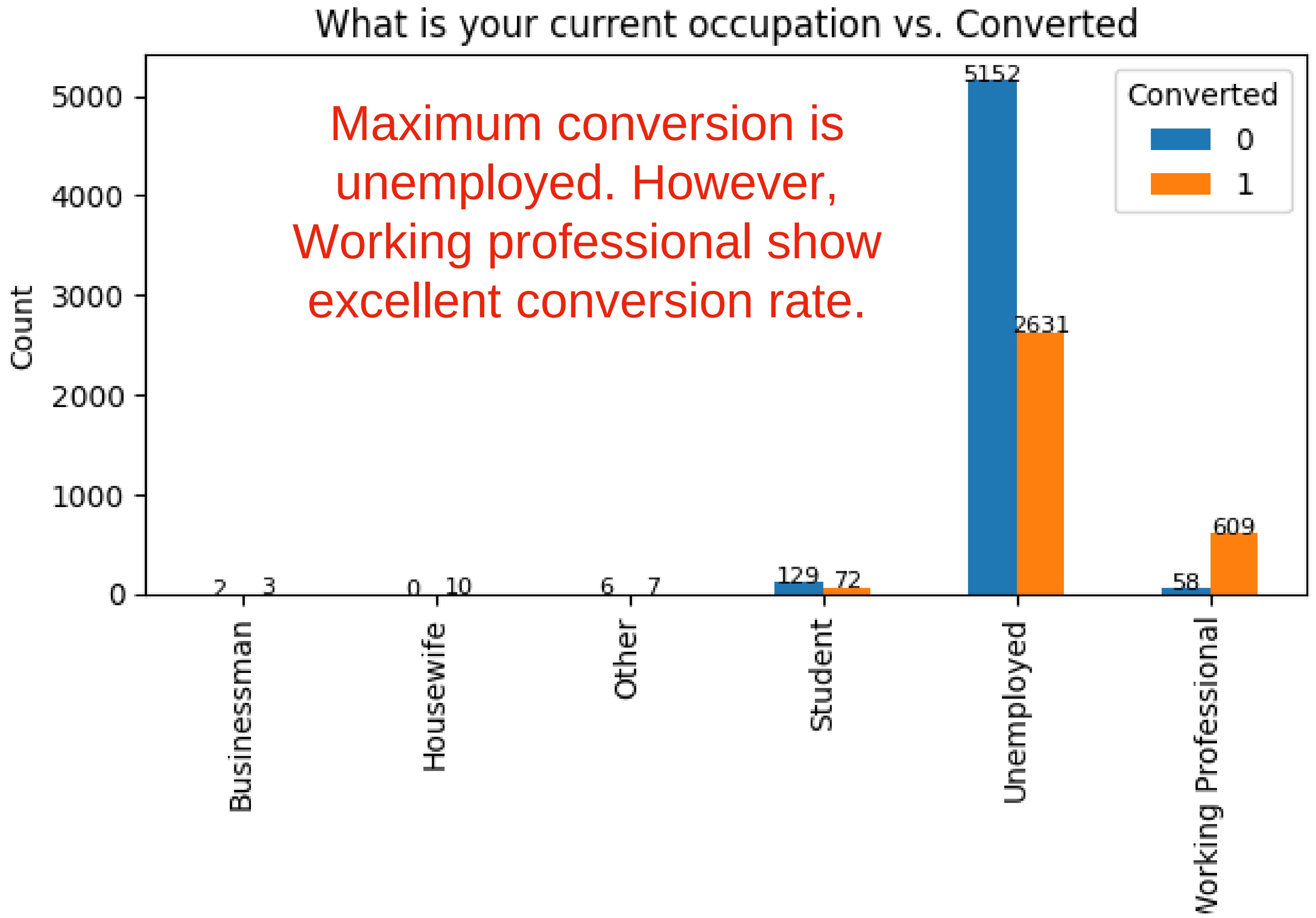
Maximum conversion for leads without recommendation. However note that for recommended case though less in number show very high conversion rate of 75%.



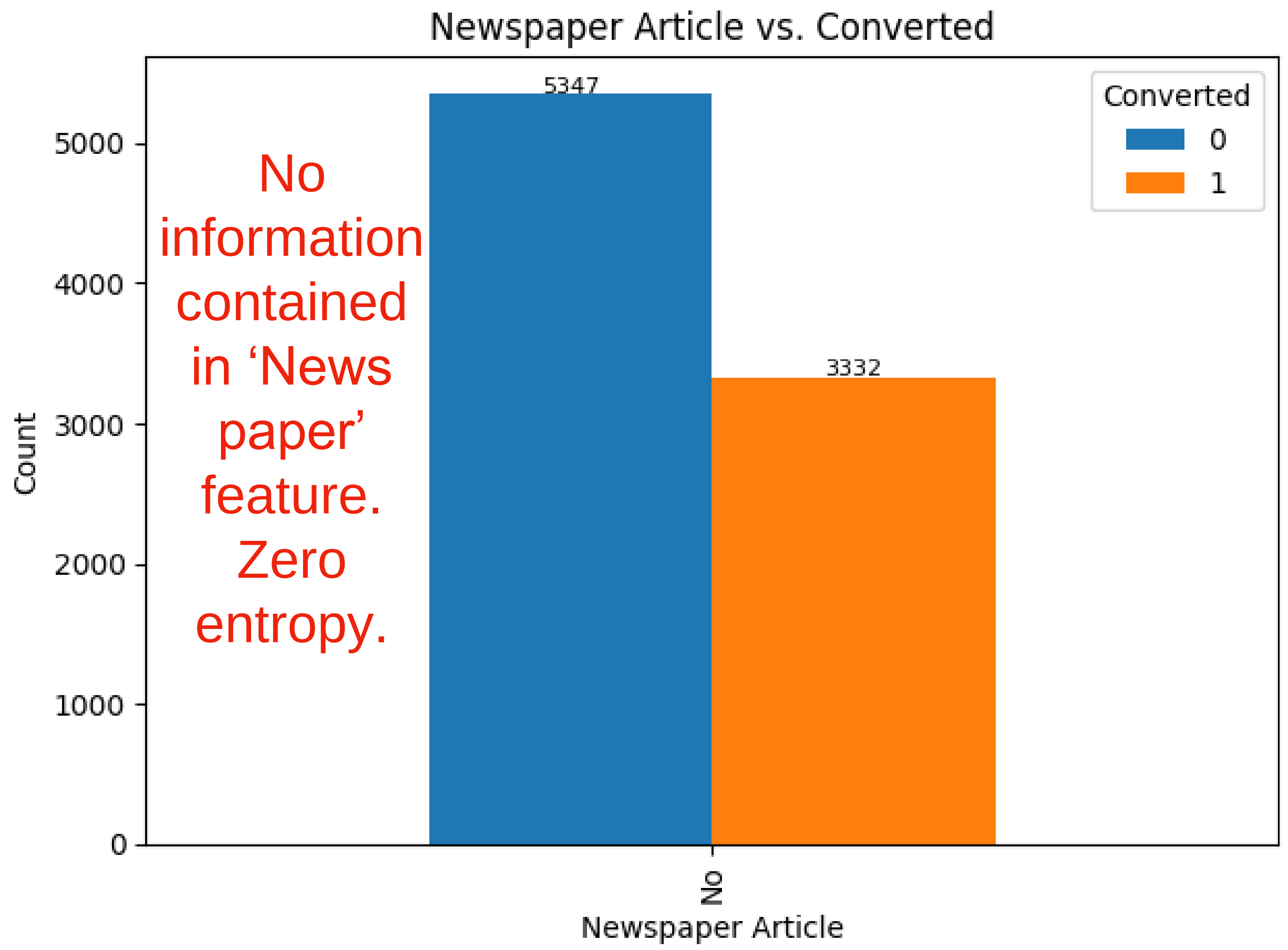
Maximum conversion for better Career prospects. However can drop as almost all cases belong to this only.



Maximum conversion is for India. However almost all samples are from India. Thus not a useful feature



Maximum conversion is unemployed. However, Working professional show excellent conversion rate.



No information contained in ‘News paper’ feature. Zero entropy.

Many more interesting fact from EDA detailed in notebook.

Data Prepration

- **Mapped features with binary entry (yes, no) to (1,0)**

```
Binary_features_list
['Do Not Email', 'Do Not Call',
'Search', 'Newspaper', 'Digital
Advertisement', 'Through
Recommendations', 'A free copy of
Mastering The Interview']
```

- **Dummy variables generated for five categorical features .**

```
categorical_features for which dummy generated
['Lead Origin', 'Lead Source','Last
Activity','What is your current
occupation', 'Last Notable Activity']
```

- **Dropped Prospect ID and lead Number as they are not valid features for modelling.**
- **Converted the datatype from ‘bool’ to numerical as need in logistic regression modeming.**
- **After completing data preparation there are 65 features with 8679 entries ready for modelling.**

Train-Test data split

- **Split final data into train (70%) and test (30%) for model development.**

Split data into train and test.

Shape of training data	(6075,	64)
Shape of test data	(2604,	64)

- **Scaled all numerical features using StandardScaler function from sklearn.**

Many more interesting fact from EDA detailed in notebook.

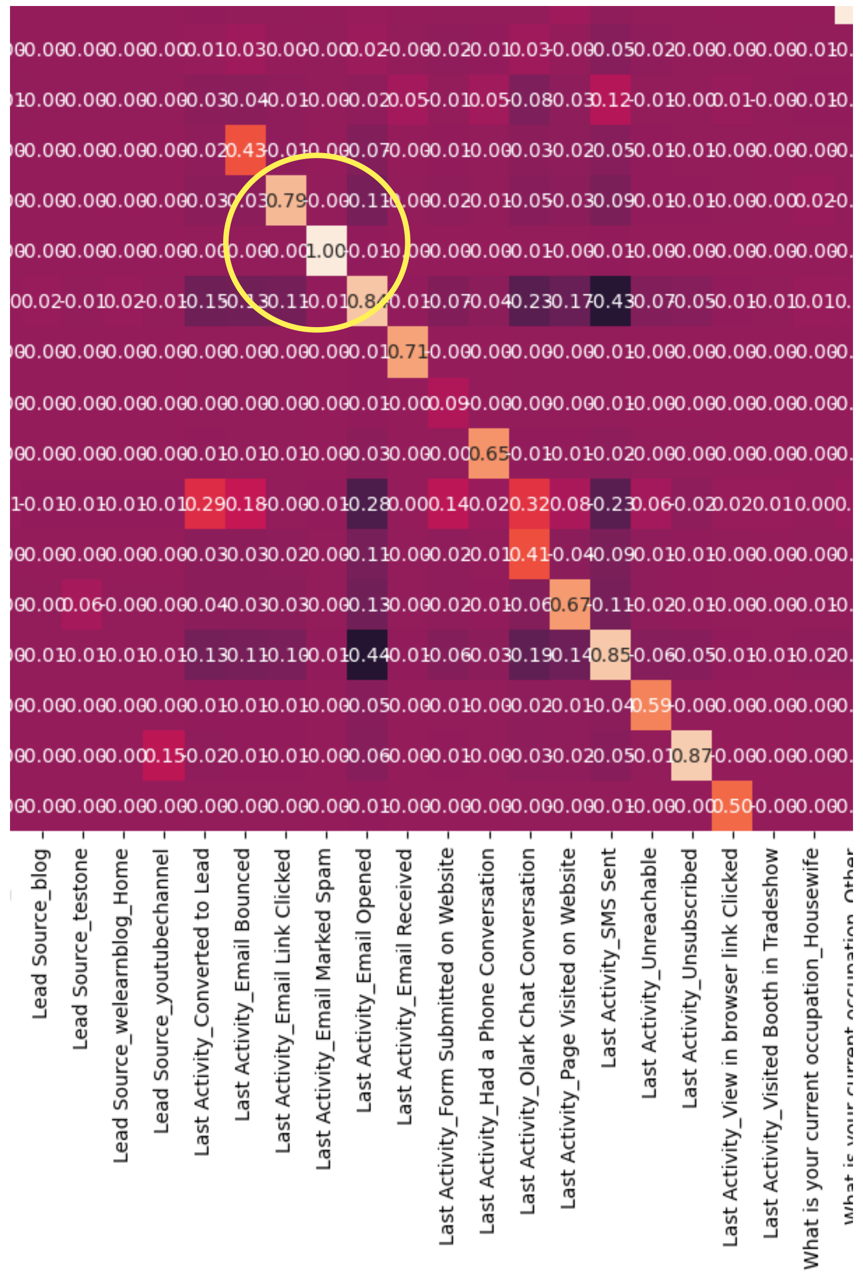
Data Prepration

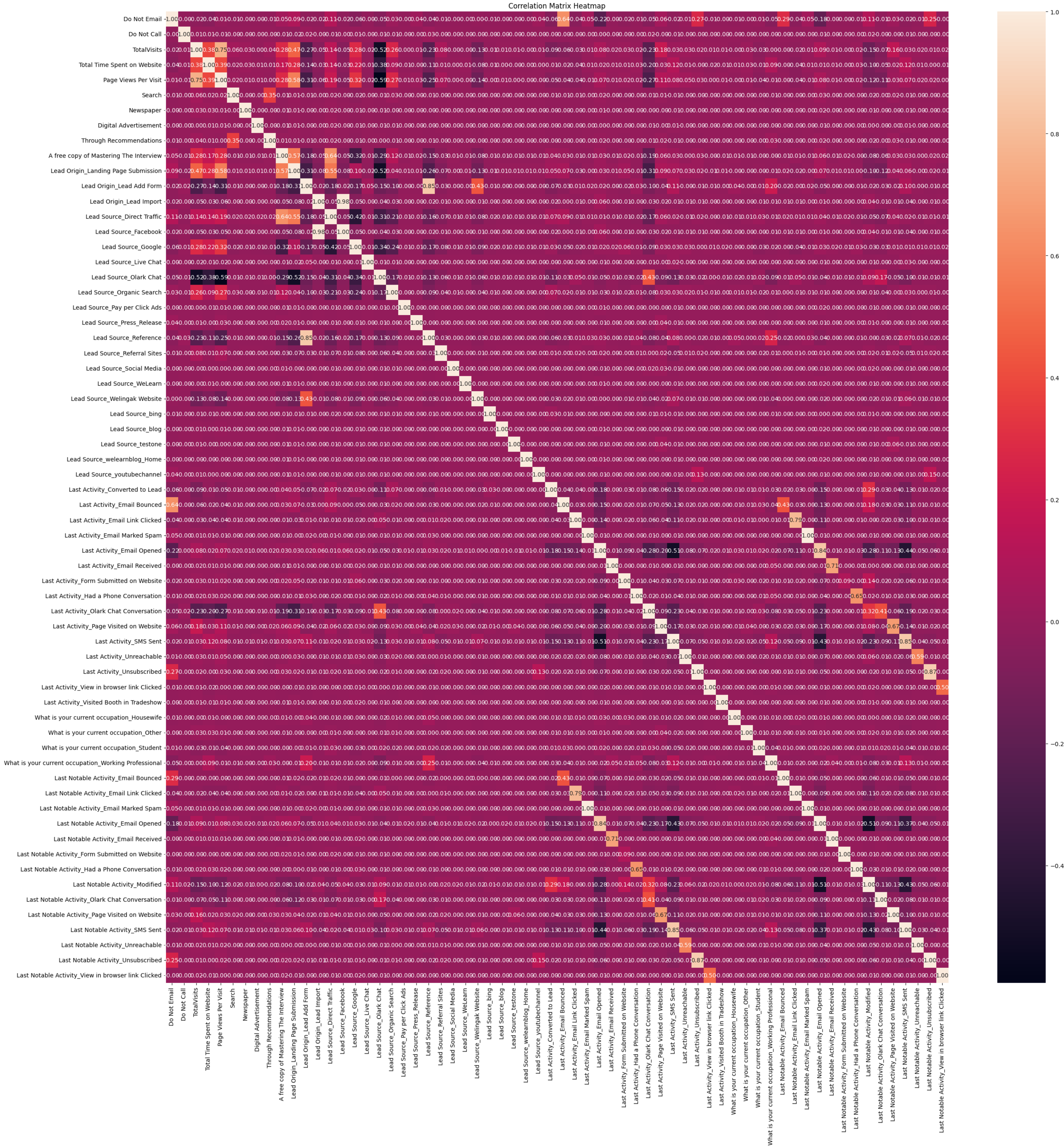
These combinations of features shows correlation between 0.9 to 1.0. Thus removed one from each pair.

(Last Activity_Unsubscribed,
Last Notable Activity_Unsubscibed)

(Last Activity_SMS Sent,
Last Notable Activity_SMS Sent)

(Last Notable Activity_Email Marked Spam,
Last Activity_Email Marked Spam)





Model development: Feature selection using RFE.

- The goal is to develop a model with **≤15 parameters**.
- Our approach is to perform **Recursive feature elimination (RFE) with 20 features** and then **fine-tuning to <=15 features using p-value and VIF criteria.**
- RFE with15 features and no fine-tuning based on p-values and VIF may lead to unnecessary or redundant features.
- By performing RFE with 20 features initially, we ensure that model should capture a comprehensive set of relevant predictors and select <15 features using p-values and VIF analysis to remove statistically insignificant or multicollinear variables.

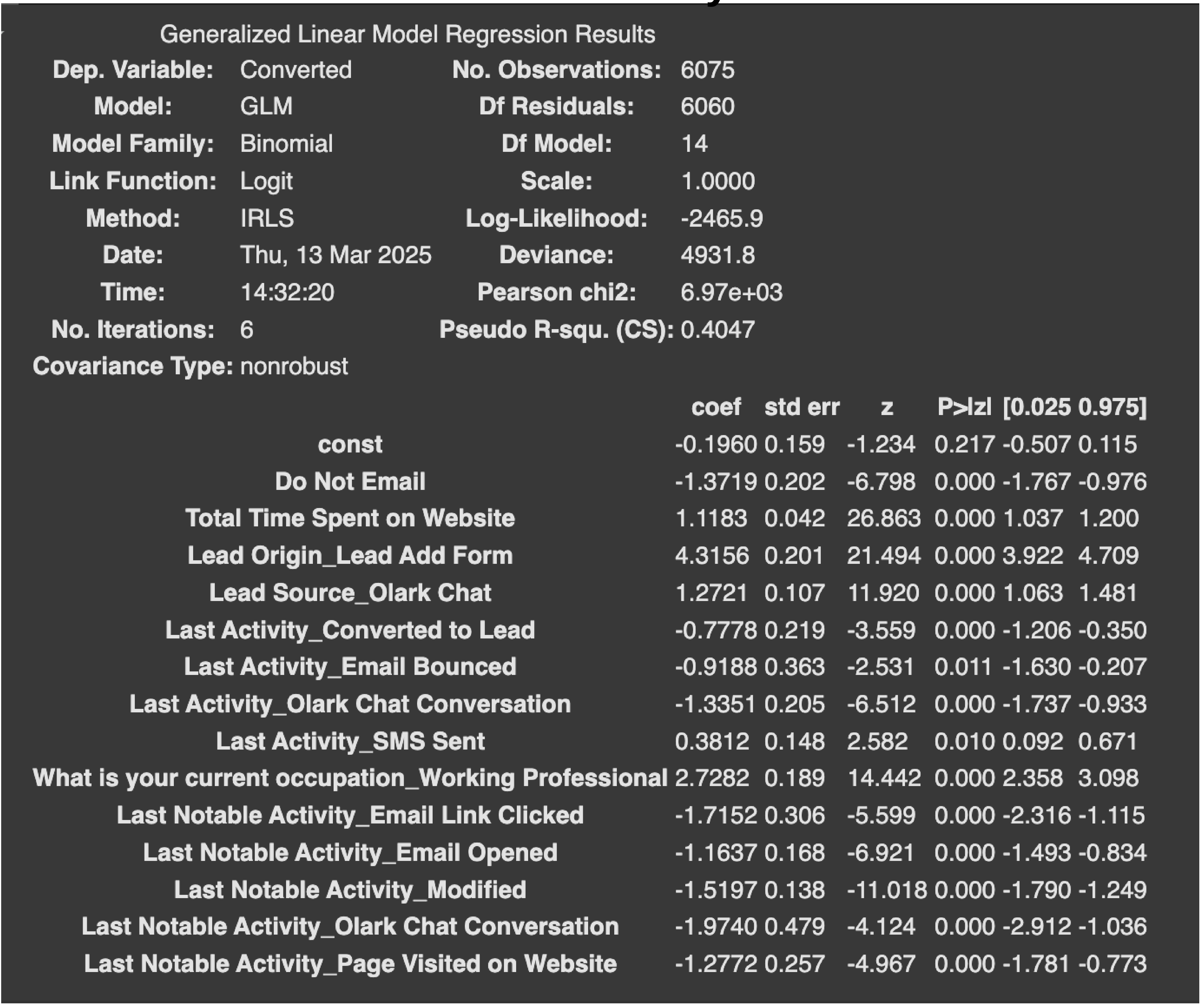
```
Features Selected by RFE:
------------------------------
- Do Not Email
- Do Not Call
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Last Activity_Converted to Lead
- Last Activity_Email Bounced
- Last Activity_Had a Phone Conversation
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- What is your current occupation_Housewife
- What is your current occupation_Working Professional
- Last Notable Activity_Email Link Clicked
- Last Notable Activity_Email Opened
- Last Notable Activity_Had a Phone Conversation
- Last Notable Activity_Modified
- Last Notable Activity_Olark Chat Conversation
- Last Notable Activity_Page Visited on Website
- Last Notable Activity_Unreachable
```

```
Features Not Selected by RFE:
------------------------------
- TotalVisits
- Page Views Per Visit
- Search
- Newspaper
- Digital Advertisement
- Through Recommendations
- A free copy of Mastering The Interview
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Import
- Lead Source_Direct Traffic
- Lead Source_Facebook
- Lead Source_Google
- Lead Source_Live Chat
- Lead Source_Organic Search
- Lead Source_Pay per Click Ads
- Lead Source_Press_Release
- Lead Source_Reference
- Lead Source_Referral Sites
- Lead Source_Social Media
- Lead Source_WeLearn
- Lead Source_bing
- Lead Source_blog
- Lead Source_testone
- Lead Source_welearnblog_Home
- Lead Source_youtubechannel
- Last Activity_Email Link Clicked
- Last Activity_Email Marked Spam
- Last Activity_Email Opened
- Last Activity_Email Received
- Last Activity_Form Submitted on Website
- Last Activity_Pae Visited on Website
- Last Activity_Unreachable
- Last Activity_View in browser link Clicked
- Last Activity_Visited Booth in Tradeshow
- What is your current occupation_Other
- What is your current occupation_Student
- Last Notable Activity_Email Bounced
- Last Notable Activity_Email Received
- Last Notable Activity_Form Submitted on Website
- Last Notable Activity_Unsubscribed
- Last Notable Activity_View in browser link Clicked
```

Model refinement: Feature selection using p-values and VIF analysis.

- Rebuilding model multiple time to elimination features with high p-values (p-values>0.05)
- After 8 iterations Model-8 gives p-values<0.05 and VIF <5 which shows the stability of the model.

Final Model Summary



```
                 Generalized Linear Model Regression Results
   Dep. Variable:  Converted          No. Observations:  6075
          Model:   GLM                   Df Residuals:   6060
   Model Family:   Binomial                 Df Model:    14
  Link Function:   Logit                       Scale:    1.0000
         Method:   IRLS                Log-Likelihood:   -2465.9
           Date:   Thu, 13 Mar 2025          Deviance:   4931.8
           Time:   14:32:20               Pearson chi2:  6.97e+03
  No. Iterations:  6             Pseudo R-squ. (CS):  0.4047
Covariance Type:   nonrobust
                                               coef    std err  z       P>|z|  [0.025  0.975]
                    const                      -0.1960 0.159   -1.234   0.217  -0.507  0.115
                 Do Not Email                  -1.3719 0.202   -6.798   0.000  -1.767  -0.976
           Total Time Spent on Website         1.1183  0.042   26.863   0.000  1.037   1.200
           Lead Origin_Lead Add Form           4.3156  0.201   21.494   0.000  3.922   4.709
             Lead Source_Olark Chat            1.2721  0.107   11.920   0.000  1.063   1.481
         Last Activity_Converted to Lead       -0.7778 0.219   -3.559   0.000  -1.206  -0.350
           Last Activity_Email Bounced         -0.9188 0.363   -2.531   0.011  -1.630  -0.207
      Last Activity_Olark Chat Conversation    -1.3351 0.205   -6.512   0.000  -1.737  -0.933
             Last Activity_SMS Sent            0.3812  0.148   2.582    0.010  0.092   0.671
What is your current occupation_Working Professional 2.7282 0.189 14.442 0.000  2.358   3.098
      Last Notable Activity_Email Link Clicked -1.7152 0.306   -5.599   0.000  -2.316  -1.115
        Last Notable Activity_Email Opened     -1.1637 0.168   -6.921   0.000  -1.493  -0.834
          Last Notable Activity_Modified       -1.5197 0.138   -11.018  0.000  -1.790  -1.249
  Last Notable Activity_Olark Chat Conversation -1.9740 0.479  -4.124   0.000  -2.912  -1.036
 Last Notable Activity_Page Visited on Website -1.2772 0.257   -4.967   0.000  -1.781  -0.773
```

VIF analysis

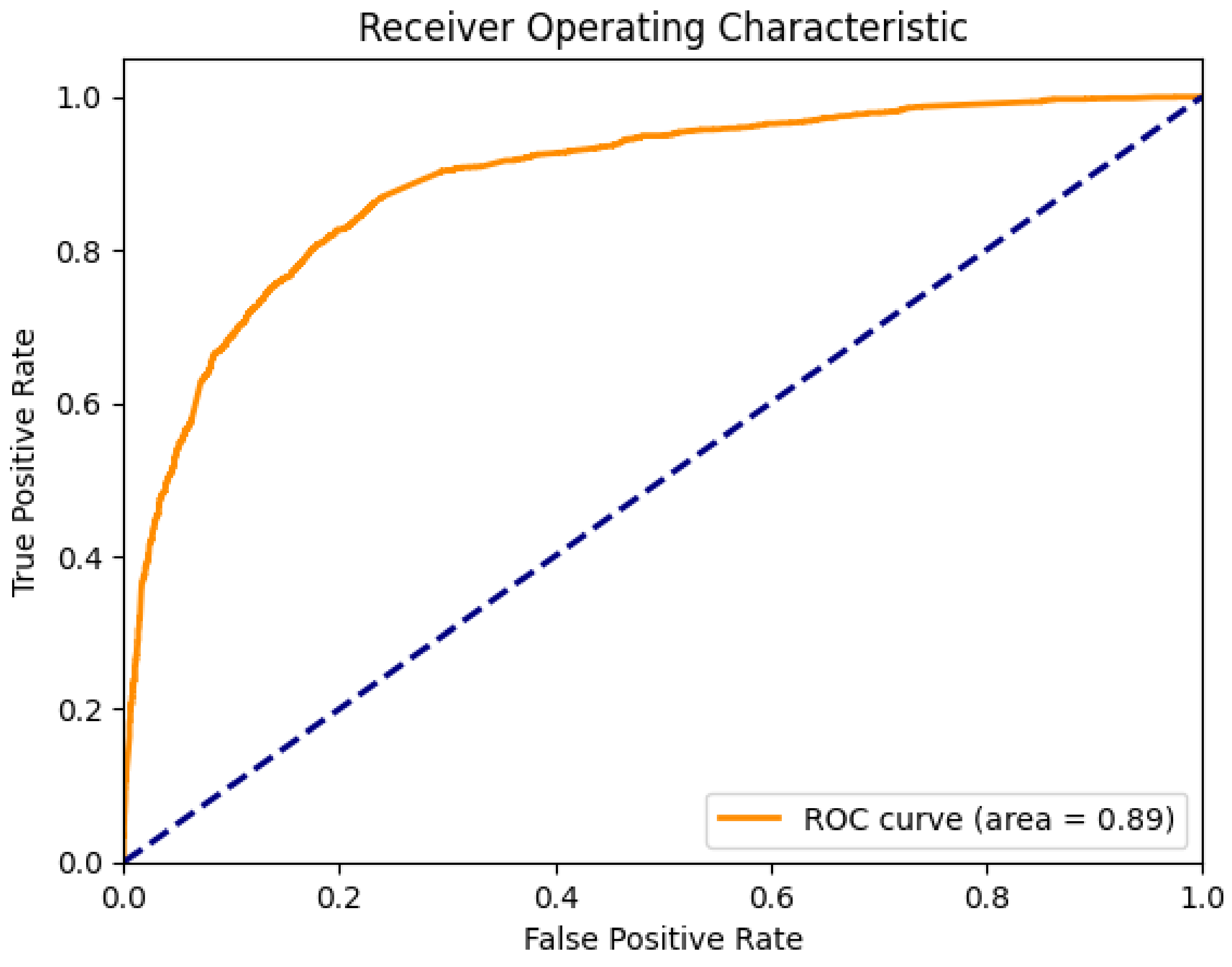
	Feature	VIF
0	Do Not Email	1.829968
1	Total Time Spent on Website	1.246661
2	Lead Origin_Lead Add Form	1.223749
3	Lead Source_Olark Chat	1.809388
4	Last Activity_Converted to Lead	1.267344
5	Last Activity_Email Bounced	1.837162
6	Last Activity_Olark Chat Conversation	2.065129
7	Last Activity_SMS Sent	1.268290
8	What is your current occupation_Working Profes...	1.148010
9	Last Notable Activity_Email Link Clicked	1.020274
10	Last Notable Activity_Email Opened	1.120737
11	Last Notable Activity_Modified	1.988950
12	Last Notable Activity_Olark Chat Conversation	1.322326
13	Last Notable Activity_Page Visited on Website	1.023511

Post Model development Analysis: Training dataset

- **Model performance on training data set:** Assumed probability threshold of 0.5.

```
Accuracy: 0.8187
Sensitivity: 0.7043
Specificity: 0.8897
False Positive Rate: 0.1103
Positive Predictive Value: 0.7980
Negative Predictive Value: 0.8293
```

Predicted Actual	Non converted	Converted
Non Converted	3338	414
Converted	687	1636



- **ROC curve:** To understand how good the finalised Model-8 is at separating the two classes (converted and non-converted) and how this varies if we change the model's confidence level (i.e., cutoff point used to convert the probabilities to binary outcome).

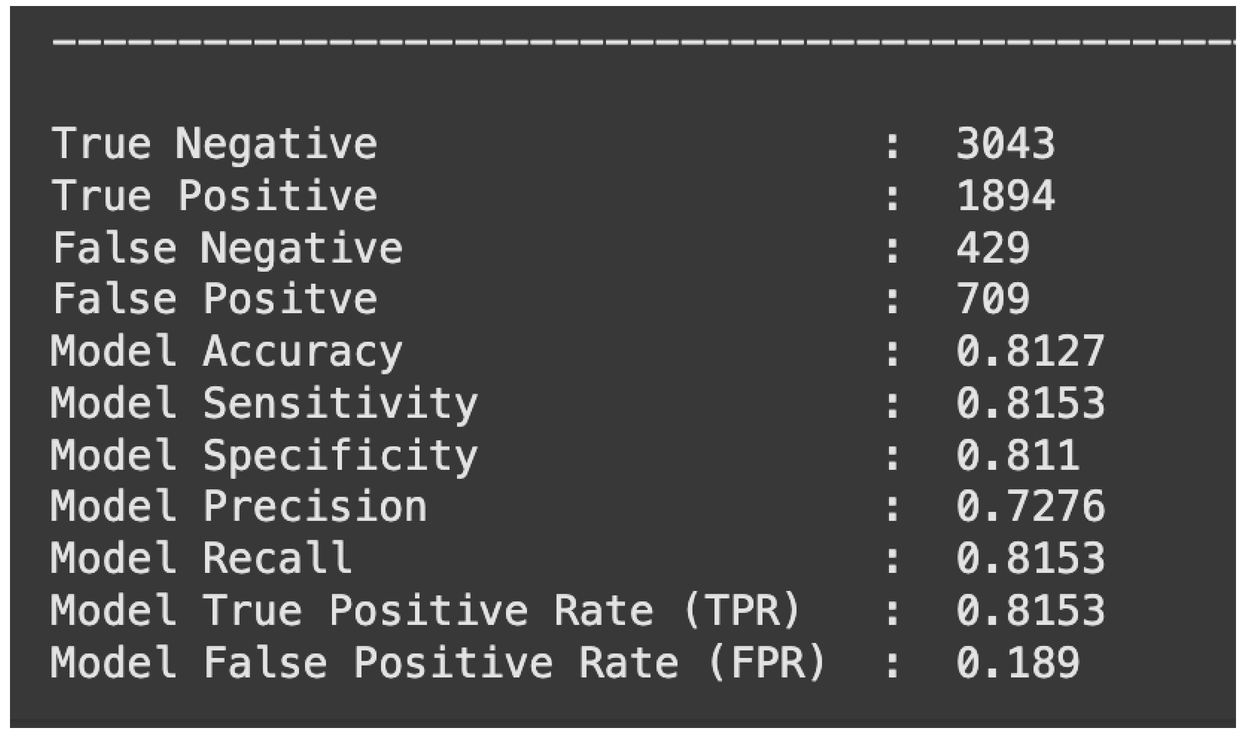
  A higher curve (closer to the top-left corner) and a larger area under the curve (AUC) indicate better performance.

- **Optimal cutoff point (Sensitivity-Specificity view)** : A cut off of 0.35 improves the performance and gives balanced accuracy, sensitivity & specificity all almost equal to 81%.

Confusion matrix with cutoff of 0.35

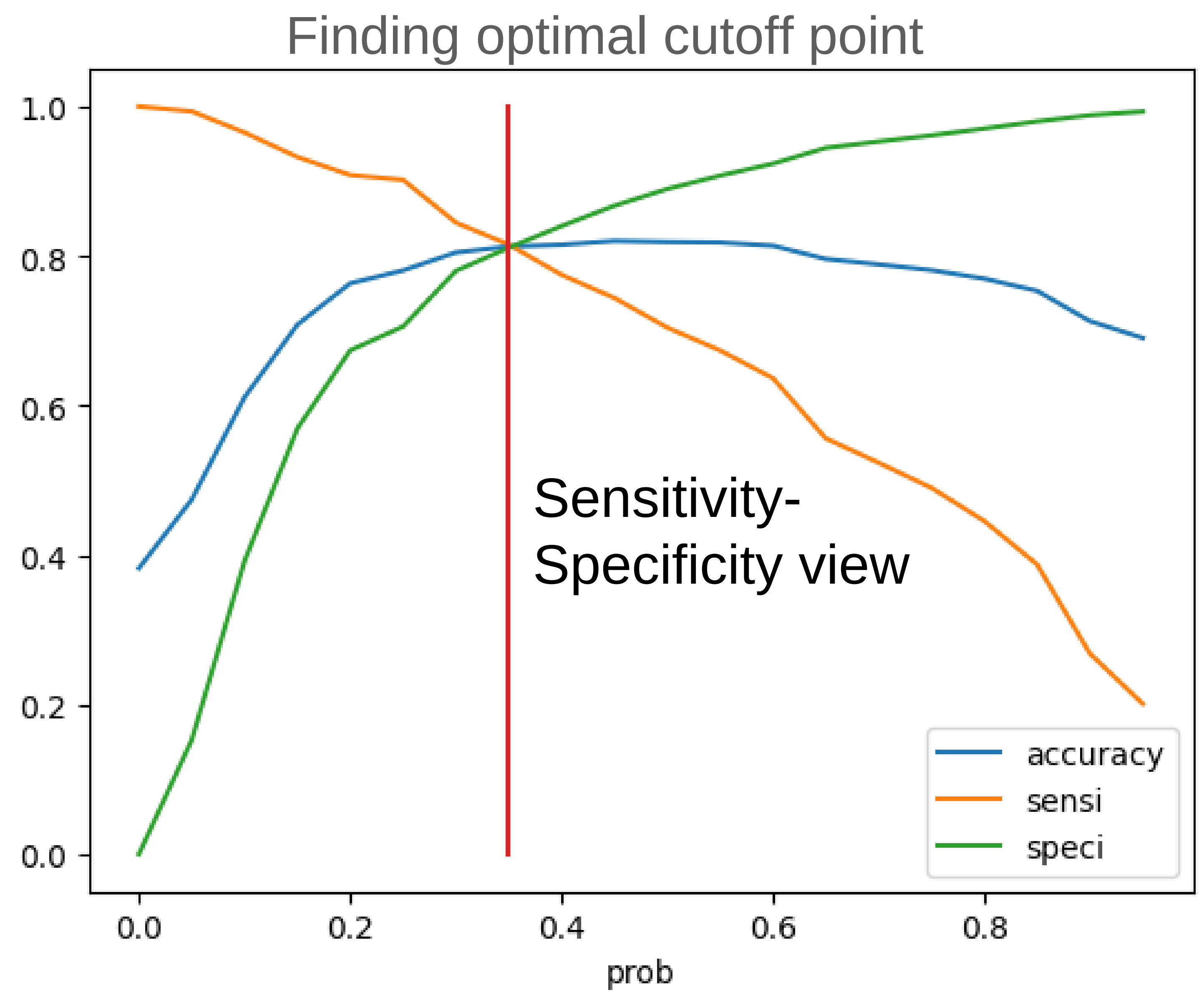
Predicted Actual	Non converted	Converted
Non Converted	3043	789
Converted	429	1894

Performance metrics with cutoff of 0.35



```
----------------------------------------------------

True Negative                     :  3043
True Positive                     :  1894
False Negative                    :  429
False Positve                     :  709
Model Accuracy                    :  0.8127
Model Sensitivity                 :  0.8153
Model Specificity                 :  0.811
Model Precision                   :  0.7276
Model Recall                      :  0.8153
Model True Positive Rate (TPR)    :  0.8153
Model False Positive Rate (FPR)   :  0.189
```



Over all sensitivity (positive conversion rate) of the model-8 on training dataset is found to be around 82%. Its a good accuracy, and more than target 80% desired by X Education CEO

Post Model development Analysis: Training dataset

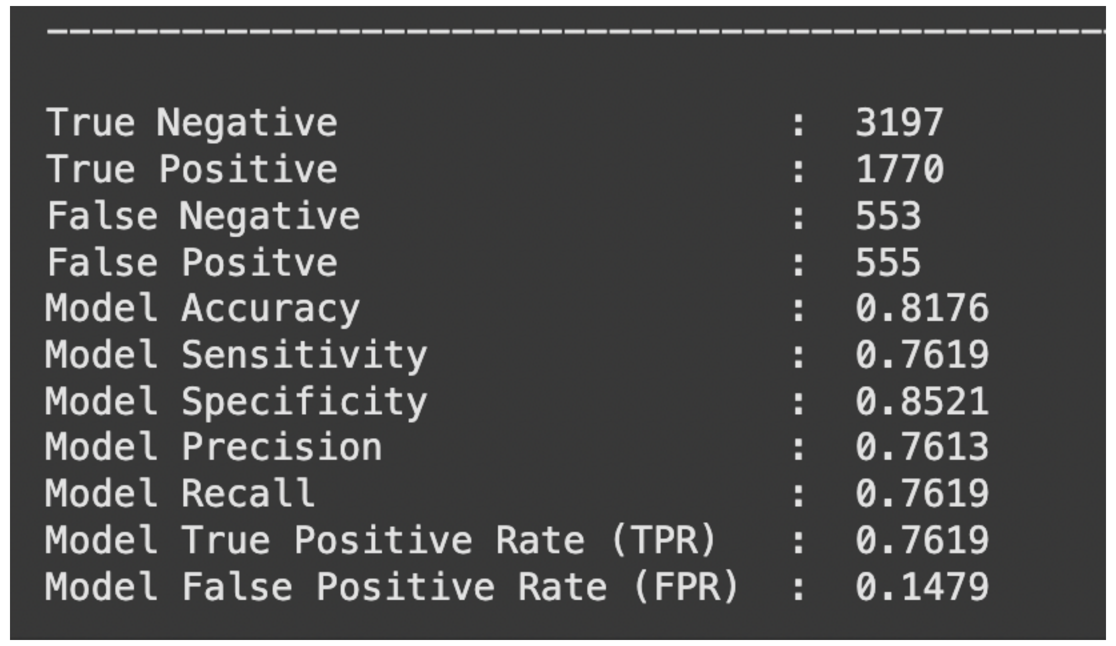
- **Precision and Recall Tradeoff (Cutoff using precision recall view):**

  A cutoff value of 0.42 is found as tradeoff between precision and Recall.

Confusion matrix with cutoff of 0.42

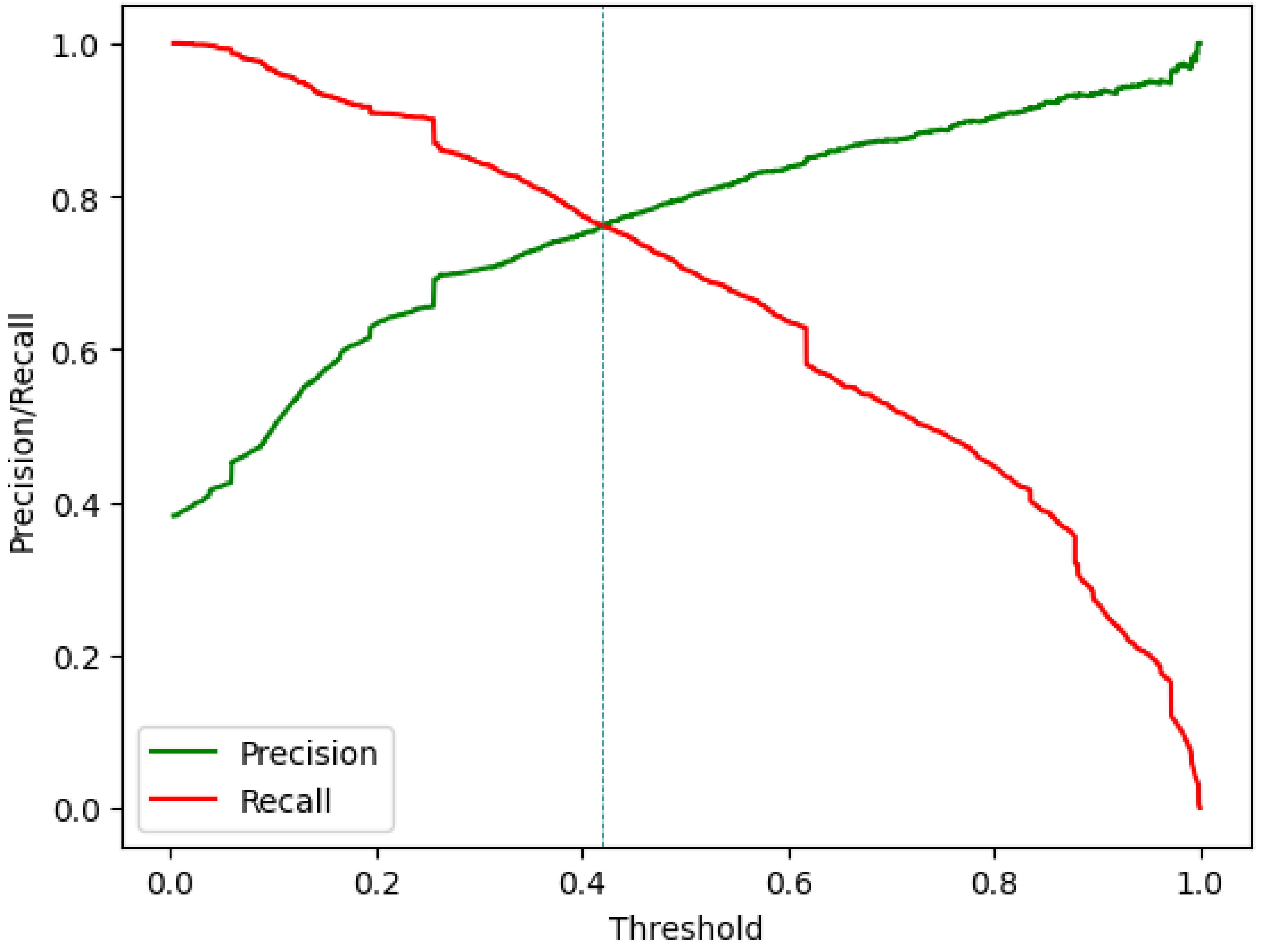
Predicted **Actual**	**Non converted**	**Converted**
Non Converted	338	414
Converted	687	1636

Performance metrics with cutoff of 0.42



```
--------------------------------------------------

True Negative                     :  3197
True Positive                     :  1770
False Negative                    :  553
False Positve                     :  555
Model Accuracy                    :  0.8176
Model Sensitivity                 :  0.7619
Model Specificity                 :  0.8521
Model Precision                   :  0.7613
Model Recall                      :  0.7619
Model True Positive Rate (TPR)    :  0.7619
Model False Positive Rate (FPR)   :  0.1479
```



- A cutoff of 0.42 decreases the conversion rate to 76% which is significantly lesser than the target of 80% as desired by CEO.

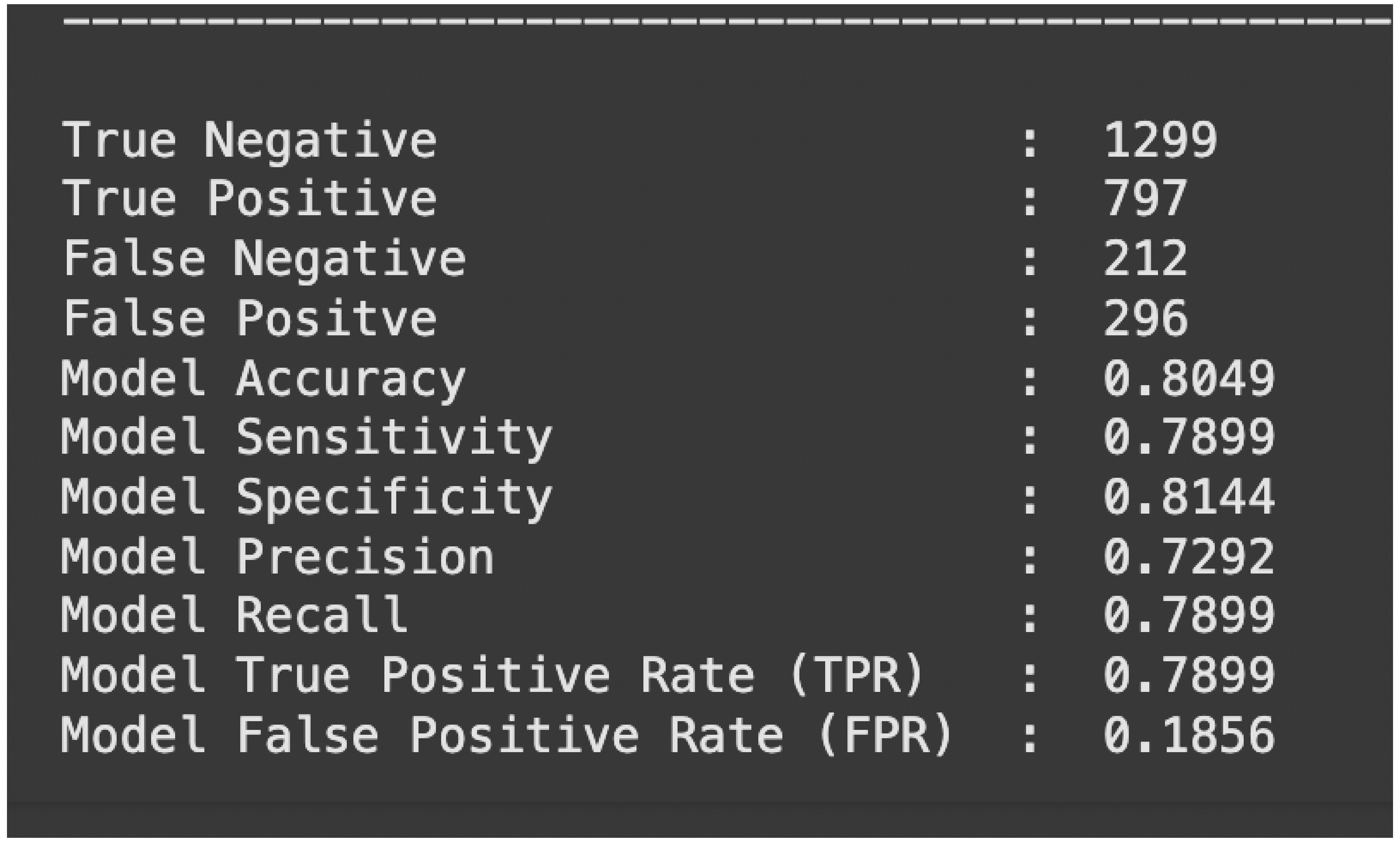
- **Thus final cutoff of 0.35 obtained by sensitivity-specificity view will be used that achieves the desired target of 80% conversion rate.**

Validation of Model: Prediction on test data.

Confusion matrix of prediction on test dataset. Cutoff of 0.35 is used.

Predicted Actual	Non converted	Converted
Non Converted	1299	296
Converted	212	797

Performance metrics of Model on test dataset.



```
---------------------------------------------

True Negative                       :  1299
True Positive                       :  797
False Negative                      :  212
False Positve                       :  296
Model Accuracy                      :  0.8049
Model Sensitivity                   :  0.7899
Model Specificity                   :  0.8144
Model Precision                     :  0.7292
Model Recall                        :  0.7899
Model True Positive Rate (TPR)      :  0.7899
Model False Positive Rate (FPR)     :  0.1856
```

The final prediction of conversions (sensitivity) on test data have a target rate of 79% (78.99%) (Around 1 % short of the predictions made on training data set).

The evaluation metrics are close to each other so it indicates that the model is stable across different evaluation metrics both test and train dataset.

These metrics are very close to train set, so out final model-8 is performing with good consistency on both Train & Test s and is also giving final prediction of conversions 79% on test which is very close to 80% desired by X education CEO.

- **Thus Model-8 shows good performance and achieves conversion rate of around 80% as desired by CEO of X Education on both test and train data.**

About Final Model-8

- Table on right shows model parameters.
- Final model have 14 features which is lesser than 15 as desired.
- A high positive coefficient for 'Lead Origin_Lead Add Form'and 'What is your current occupation_Working Professional', indicates these variable has stronger influence on predicting the probability of leads converting to take up X-Education's course.

Features	Coefficents
Lead Origin_Lead Add Form	4.315643
What is your current occupation_Working Professional	2.728219
Lead Source_Olark Chat	1.272080
Total Time Spent on Website	1.118258
Last Activity_SMS Sent	0.381176
const	-0.196029
Last Activity_Converted to Lead	-0.777787
Last Activity_Email Bounced	-0.918849
Last Notable Activity_Email Opened	-1.163690
Last Notable Activity_Page Visited on Website	-1.277239
Last Activity_Olark Chat Conversation	-1.335138
Do Not Email	-1.371936
Last Notable Activity_Modified	-1.519683
Last Notable Activity_Email Link Clicked	-1.715187
Last Notable Activity_Olark Chat Conversation	-1.973961

Conclusions

- We successfully developed a logistic regression model using lead data provided.
- Using RFE, p-value and VIF analysis, we have selected most important features giving stable model with consistent results on train as well as test dataset.
- The finalised model have 14 features, which is less than 15 as desired by the given instructions.
- We did both Sensitivity-Specificity as well as Precision-Recall tradeoff analysis giving probability optimal cutoff of 0.35 and 0.42, respectively.
- The Model performance metrics shows that with threshold of 0.42 decreases the sensitivity to 76%, while with a cutoff of 0.35 the sensitivity is around 80% (81% and 79%, respectively, on train and test data.)
- Since X-Education CEO has asked a sensitivity of around 80%, therefore cutoff of 0.35 based on sensitivity-specificity metrics is finalised.
- The model can **identify high-potential leads (also called Hot Leads)** twith improved conversion rate of around 80% as desired by the business goal.
- The top 3 variables that contribute for lead getting converted in the model are:

  - Lead Origin_Lead Add Form,

  - What is your current occupation_Working professionals, and

  - Lead Source_Olark Chat

Over all final model gives good and stable results.

Thanks