Lead Scoring Case Study

Submitted by:

- Kankani Venkata Hari
- Himanshu Shekhar
- Hrithik Kumar

Problem Statement

X Education, an online education company, has a low lead conversion rate despite acquiring many leads daily.

They aim to boost the conversion rate from 30% to 80% by identifying the leads most likely to become paying customers, called 'Hot Leads'.

Business Objectives

- ➔ Build a lead scoring model to prioritize leads based on their likelihood of conversion. The model should assign a lead score to each lead, with higher scores indicating a higher probability of conversion.
- ➔ This will enable the sales team to focus their efforts on communicating with potential leads more efficiently, thereby increasing the overall lead conversion rate.
- ➔ The target for conversion rate is 80%

Solution Methodology

- ❑ Importing Libraries and Data
    - ❑ Importing libraries, set warnings and set display
    - ❑ Reading the data
- ❑ Data Understanding
- ❑ Data Cleaning
    - ❑ Treatment for 'Select' values
    - ❑ Handling Missing Values
    - ❑ Columns with Categorical Data
    - ❑ Columns with Numerical Data
    - ❑ Removing Unwanted Columns
    - ❑ Checking & Dropping Category Columns that are Skewed
    - ❑ Outlier Analysis
    - ❑ Fixing Invalid values & Standardising Data in columns

- ❑ Data Analysis (EDA)
    - ❑ Checking Data Imbalance
    - ❑ Univariate Analysis
    - ❑ Bivariate Analysis
- ❑ Data Preparation
    - ❑ Creating Dummy Variables
- ❑ Test Train Split
- ❑ Feature Scaling
- ❑ Model Building
    - ❑ Feature Selection Using RFE
- ❑ Model Evaluation
- ❑ Making Predictions on Test Set
- ❑ Adding Lead Score Feature to Dataset
- ❑ Conclusion

Data Understanding

- Checking for unique values
- Checking data types of all columns
- Checking for duplicate values

Data Cleaning

➔ Treatment for 'Select' values

'Select' values are supposed to be treated as null values
All ‘Select’ values converted to ‘NaN’

➔ Handling Missing Values

Following columns were removed for having >40% missing values:

'How did you hear about X Education'
'Lead Profile', 'Lead Quality'
'Asymmetrique Profile Score'
'Asymmetrique Activity Score'
'Asymmetrique Activity Index'
'Asymmetrique Profile Index'

Data Cleaning

- ➔ Columns with Categorical Data

City: City has 39.71 % missing values. Imputing missing values with Mumbai will make the data more skewed. Skewness will later cause bias in the

Specialization: Specialization has 36.58 % missing values. The specialization selected is evenly distributed. Hence imputation or dropping is not a good choice. We need to create additional category called 'Others'.

Tags: Tags has 36.29 % missing values. Tags are assigned to customers indicating the current status of the lead. Since this is current status, this column will not be useful for modeling. Hence it can be dropped.

What matters most to you in choosing a course: This variable has 29.32 % missing values. 99.95% customers have selected 'better career prospects'. This is massively skewed and will not provide any insight.

What is your current occupation: We can impute the missing values with 'Unemployed' as it has the most values. This seems to be a important variable from business context, since X Education sells online courses and unemployed people might take this course to increase their chances of getting employed.

Country: X Education sells online courses and appx 96% of the customers are from India. Does not make business sense right now to impute missing values with India. Hence `Country column can be dropped.

Last Activity: "Email Opened" is having highest number of values and overall missing values in this column is just 1.11%, hence we will impute the missing values with label 'Email Opened'.

Lead Source: "Google" is having highest number of occurences and overall nulls in this column is just 0.39%, hence we will impute the missing values with label 'Google'

Dropping the following columns:

'City'
'Tags'
'Country'
'What matters most to you in choosing a course'

Imputing the following columns:

'Specialization',
'Lead Source',
'Last Activity',
'What is your current occupation'

➔ Columns with Numerical Data

'TotalVisits' missing values can be imputed with mode

'Page Views Per Visit' missing values can be imputed with mode

Data Cleaning

➔ Removing Unwanted Columns

Following columns have only 1 unique value:

'I agree to pay the amount through cheque'
'Get updates on DM Content'
'Update me on Supply Chain Content'
'Receive More Updates About Our Courses'
'Magazine'

'Prospect ID'
'Lead Number'
'Last Notable Activity'

Above columns do not add any value to the model. Dropping these columns will remove unnecessary data from the dataframe.

Data Cleaning

➔ Checking & Dropping Category Columns that are Skewed

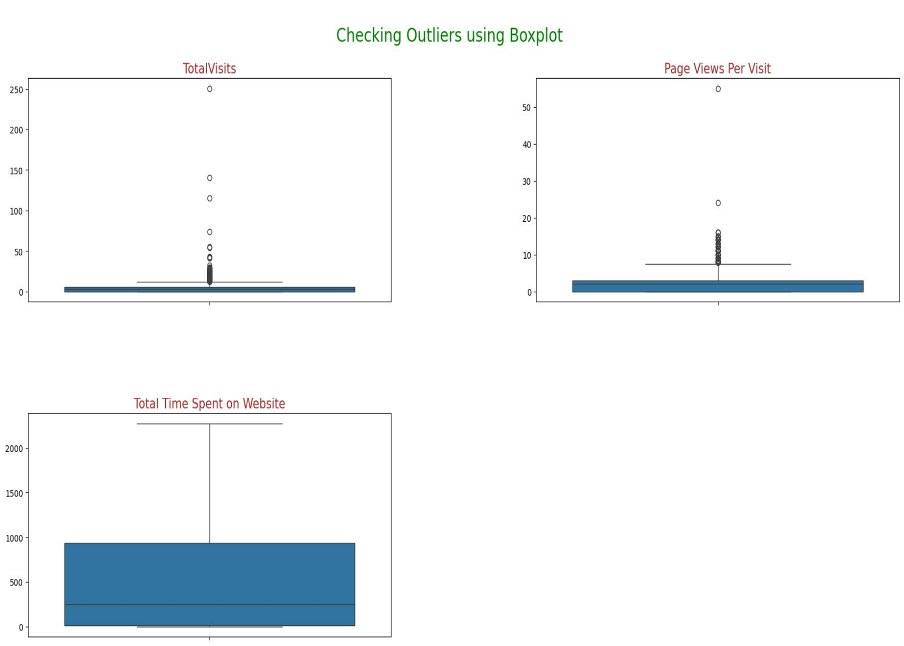
Following columns have data which is highly skewed :

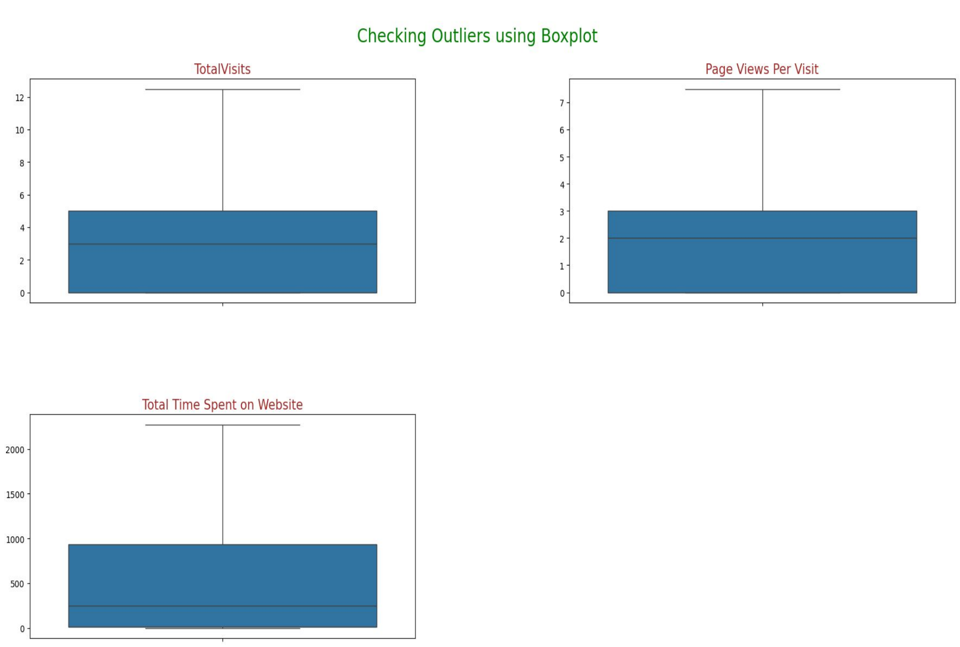
'Do Not Call'
'Search'
'Newspaper Article'
'X Education Forums'
'Newspaper'
'Digital Advertisement'
'Through Recommendations'

Hence these columns will be dropped as they will not add any value to the model and affect it negatively

Data Cleaning

➔ Outlier Analysis





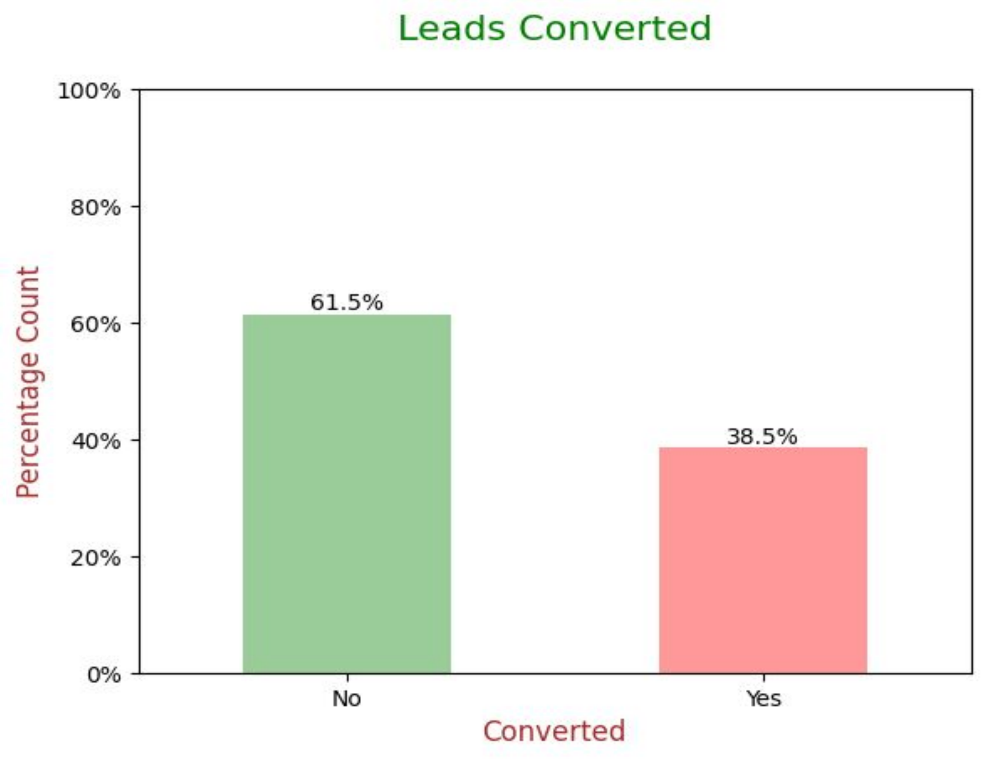
Data Analysis (EDA)

➔ Checking Data Imbalance

Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)

While 61.5% of the people did not convert to leads. (Majority)

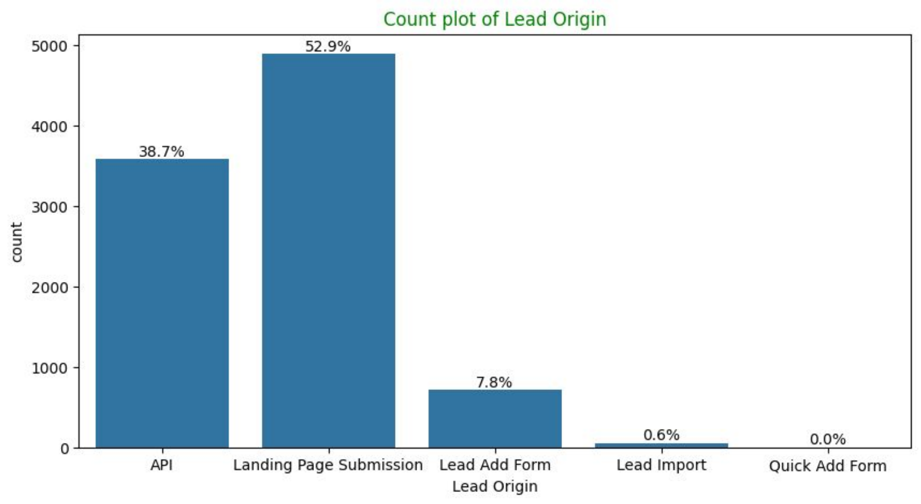
Data Imbalance Ratio: **1.59 : 1**

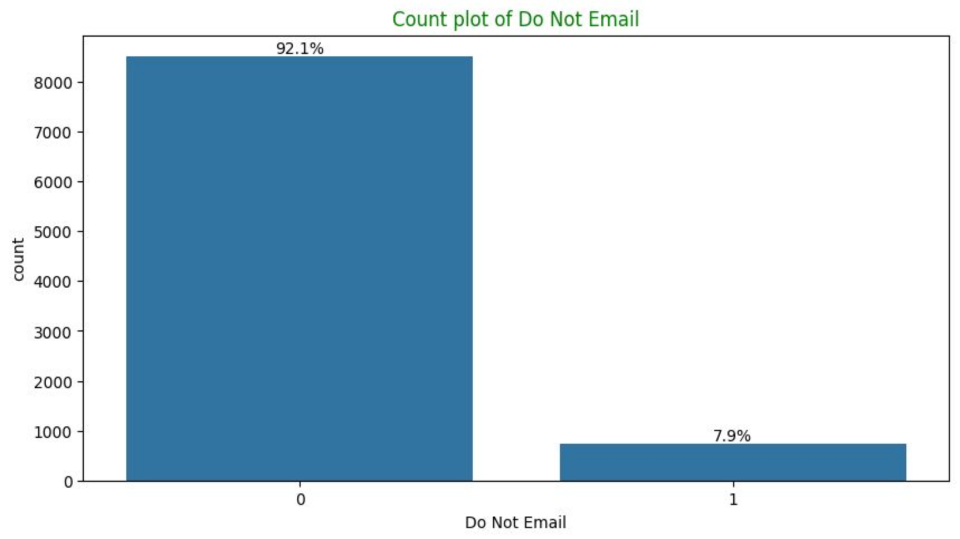


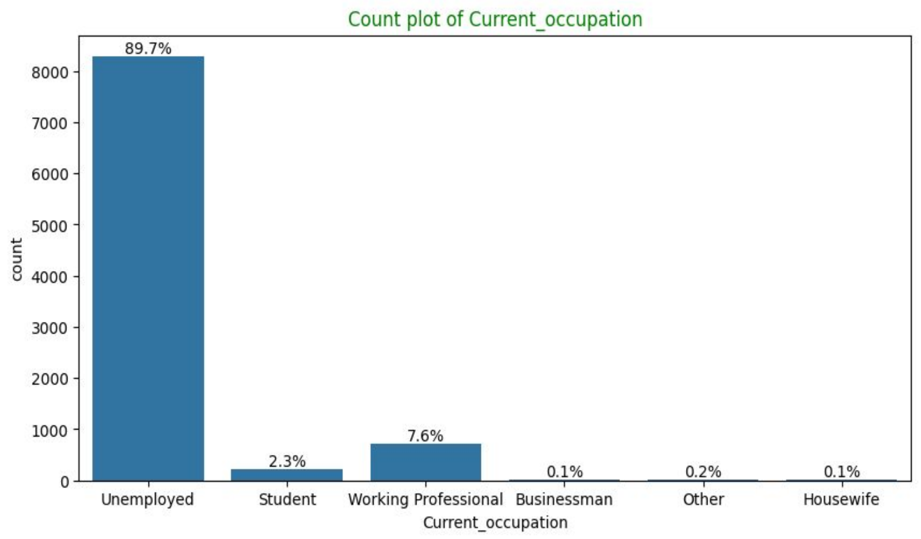
Data Analysis (EDA)

➔ Univariate Analysis

Countplot of columns with value_counts percentage as annotation







Data Analysis (EDA)

- ➔ Univariate Analysis

In Categorical Univariate Analysis we get to know the value counts percentage in each variable that how much is the distribution of values in each column.

Here is the list of features from variables which are present in majority (Converted and Not Converted included)

Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.

Current_occupation: It has 90% of the customers as Unemployed

Do Not Email: 92% of the people has opted that they dont want to be emailed about the course

Lead Source: 58% Lead source is from Google & Direct Traffic combined

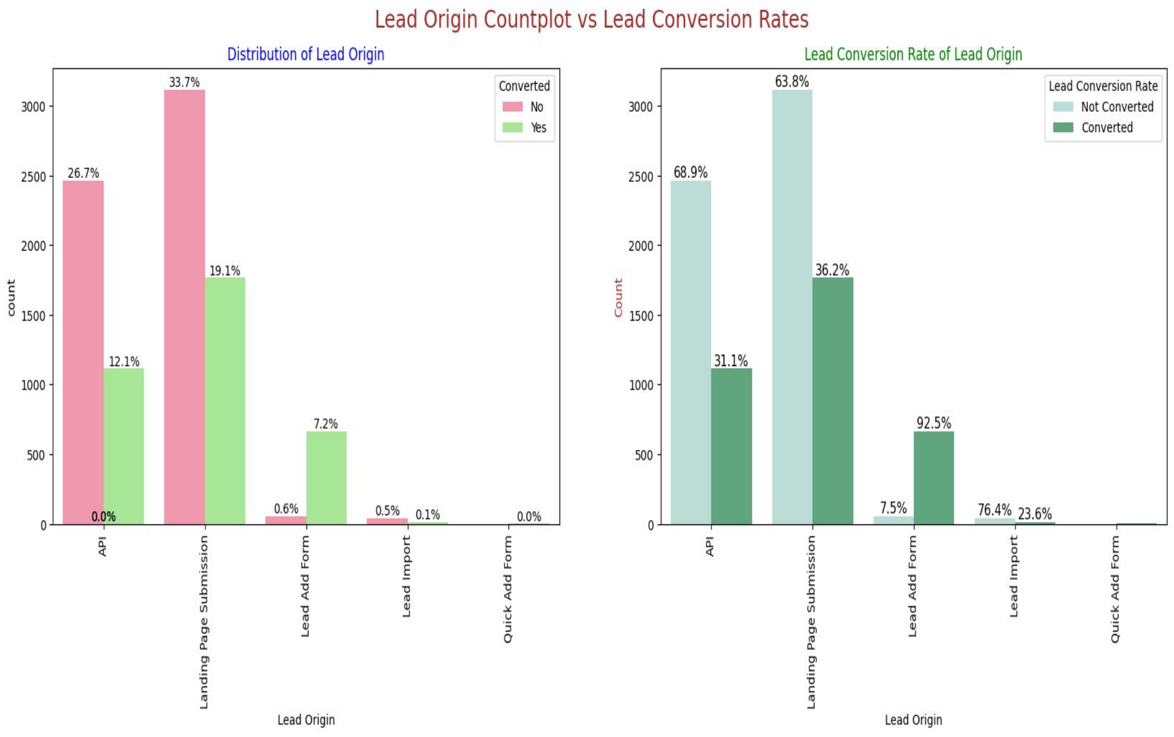
Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

Data Analysis (EDA)

➔ Bivariate Analysis

Lead Origin: Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%. The "API" identified approximately 39% of customers with a LCR of 31%.

Current_occupation: Around 90% of the customers are Unemployed with LCR of 34%. While Working Professional contribute only 7.6% of total customers with almost 92% lead conversion rate (LCR).

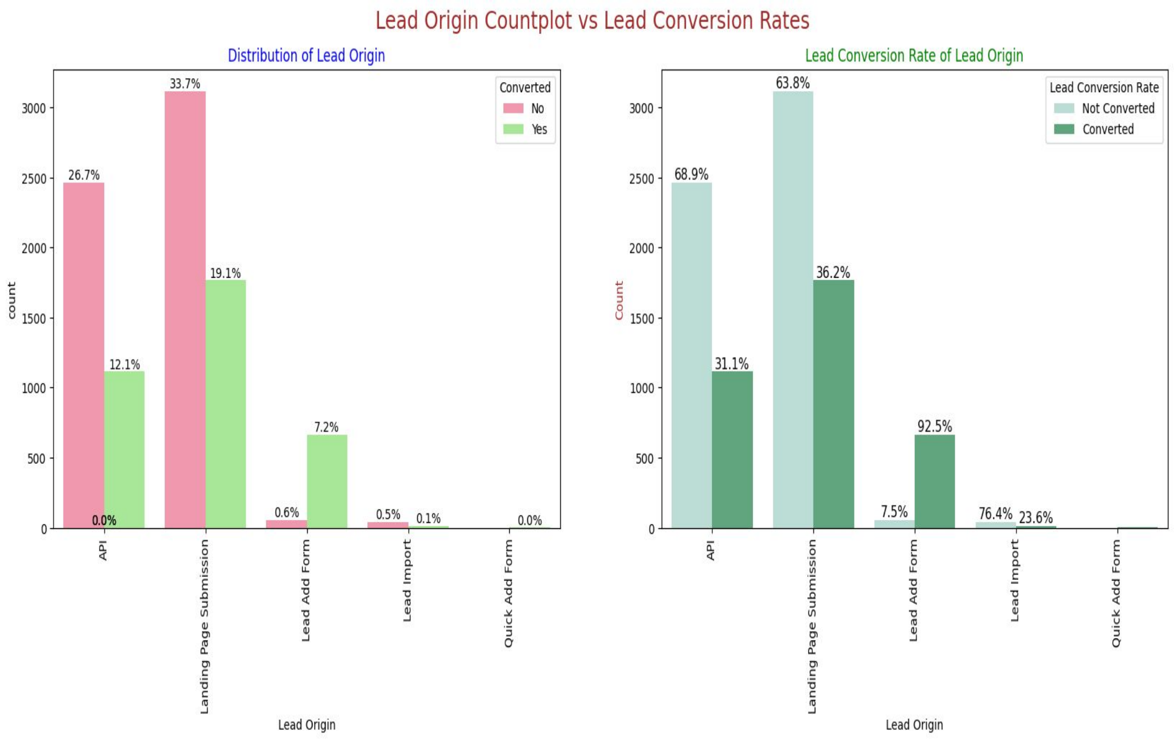


Data Analysis (EDA)

➔ Bivariate Analysis

Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.

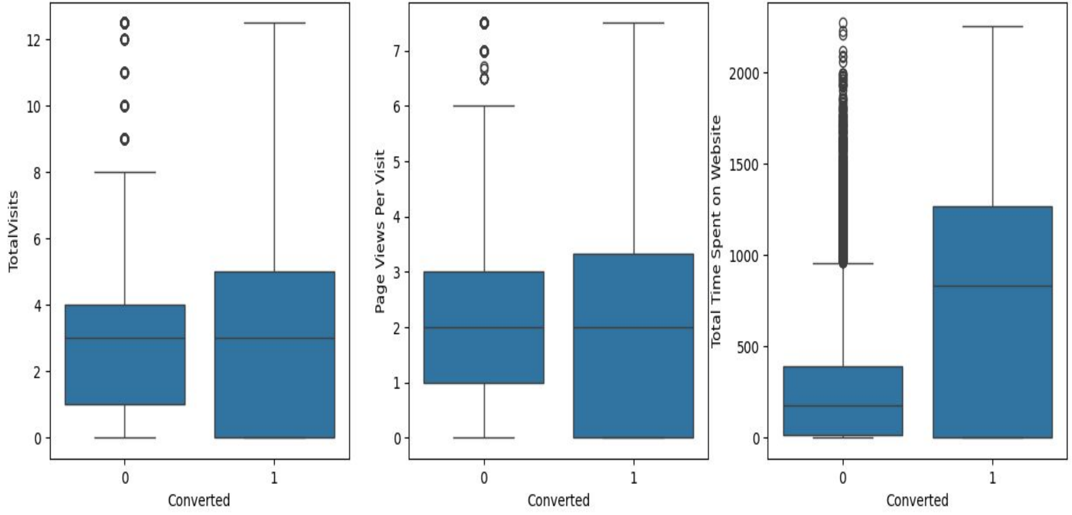
Lead Source: Google has LCR of 40% out of 31% customers, Direct Traffic contributes 32% LCR with 27% customers which is lower than Google, Organic Search also gives 37.8% of LCR but the contribution is by only 12.5% of customers, Reference has LCR of 91% but there are only around 6% of customers through this Lead Source.

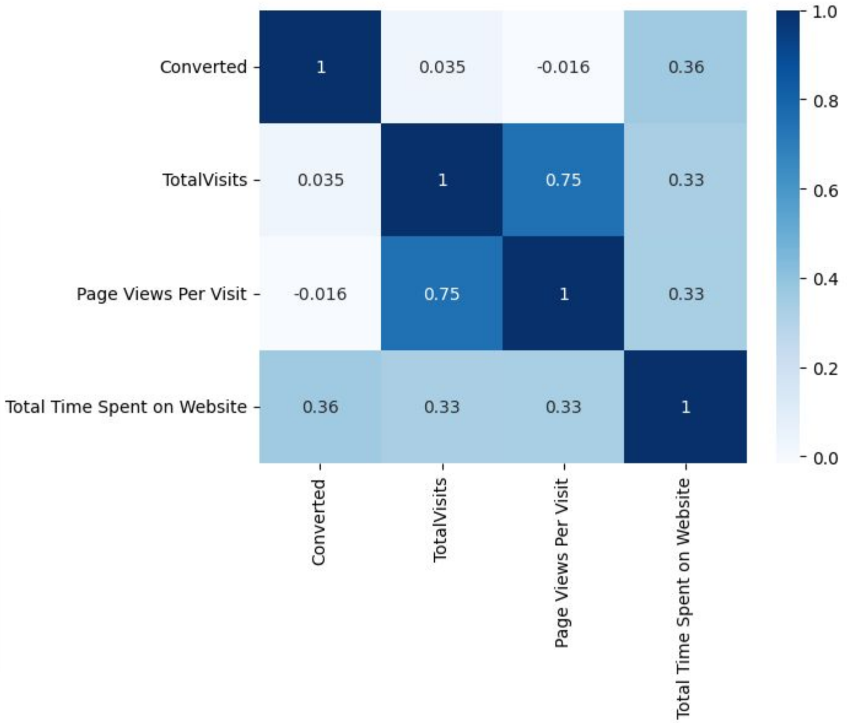

Lead Origin Countplot vs Lead Conversion Rates

Data Analysis (EDA)

➔ Bivariate Analysis

Past Leads who spend more time on website are successfully converted than those who spend less time





Test/Train Split

➔ Creating Dummy Variables

Setting Predictor variables to X

Setting Target variables to Y

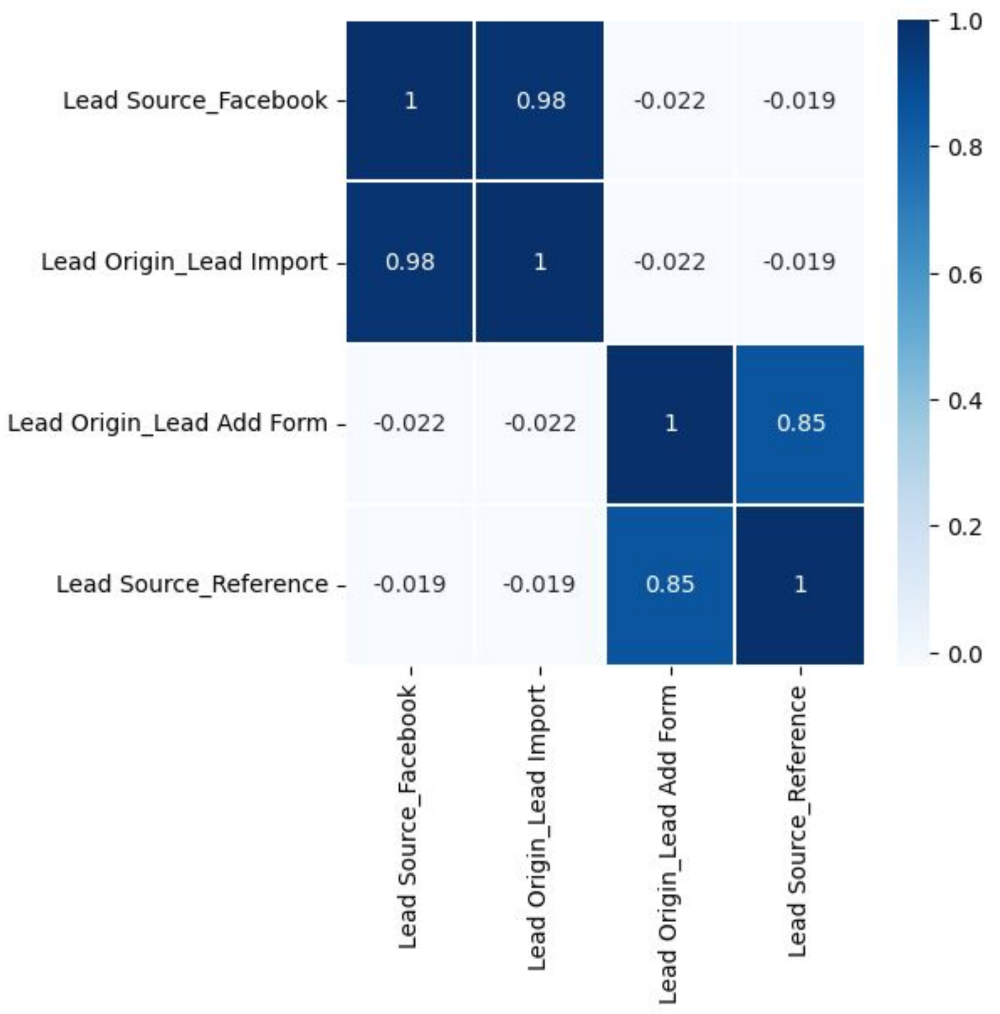
Splitting the data into Train and Test: 70:30 ratio

Feature Scaling

Using standard scaler for scaling the features.
Checking the Lead Conversion Rate (LCR) - "Converted" is our Target Variable:
38.5% conversion rate

➔ Looking at correlations

These predictor variables above are very highly correlated with each other near diagonal with (0.98 and 0.85), it is better that we drop one of these variables from each pair as they won’t add much value to the model. So , we can drop any of them, lets drop 'Lead Origin_Lead Import' and 'Lead Origin_Lead Add Form'.



Model Building

We will Build Logistic Regression Model for predicting categorical variable

Feature Selection Using RFE (Coarse tuning)

Manual fine-tuning using p-values and VIFs

Model Building

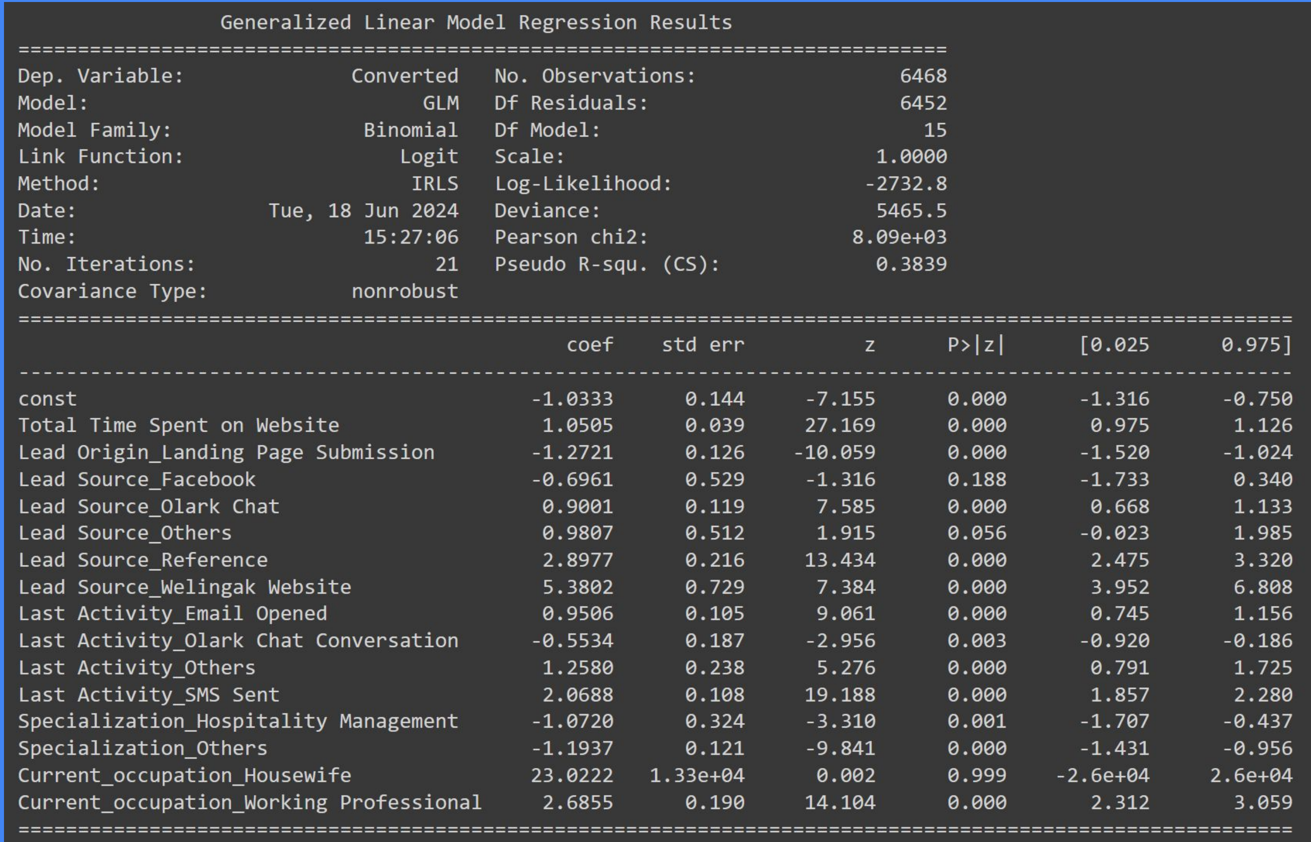
➔ Feature Selection Using RFE (Recursive Feature Elimination)

Following columns were chosen as per RFE
'Total Time Spent on Website'
'Lead Origin_Landing Page Submission'
'Lead Source_Facebook'
'Lead Source_Olark Chat'
'Lead Source_Others'
'Lead Source_Reference'
'Lead Source_Welingak Website'
'Last Activity_Email Opened'
'Last Activity_Olark Chat Conversation'
'Last Activity_Others'
'Last Activity_SMS Sent'
'Specialization_Hospitality Management'
'Specialization_Others'
'Current_occupation_Housewife'
'Current_occupation_Working Professional'

Model Evaluation

➔ Model 1

"Current_occupation_Housewife" column will be removed from model due to high p-value of 0.999, which is above the accepted threshold of 0.05 for statistical significance.



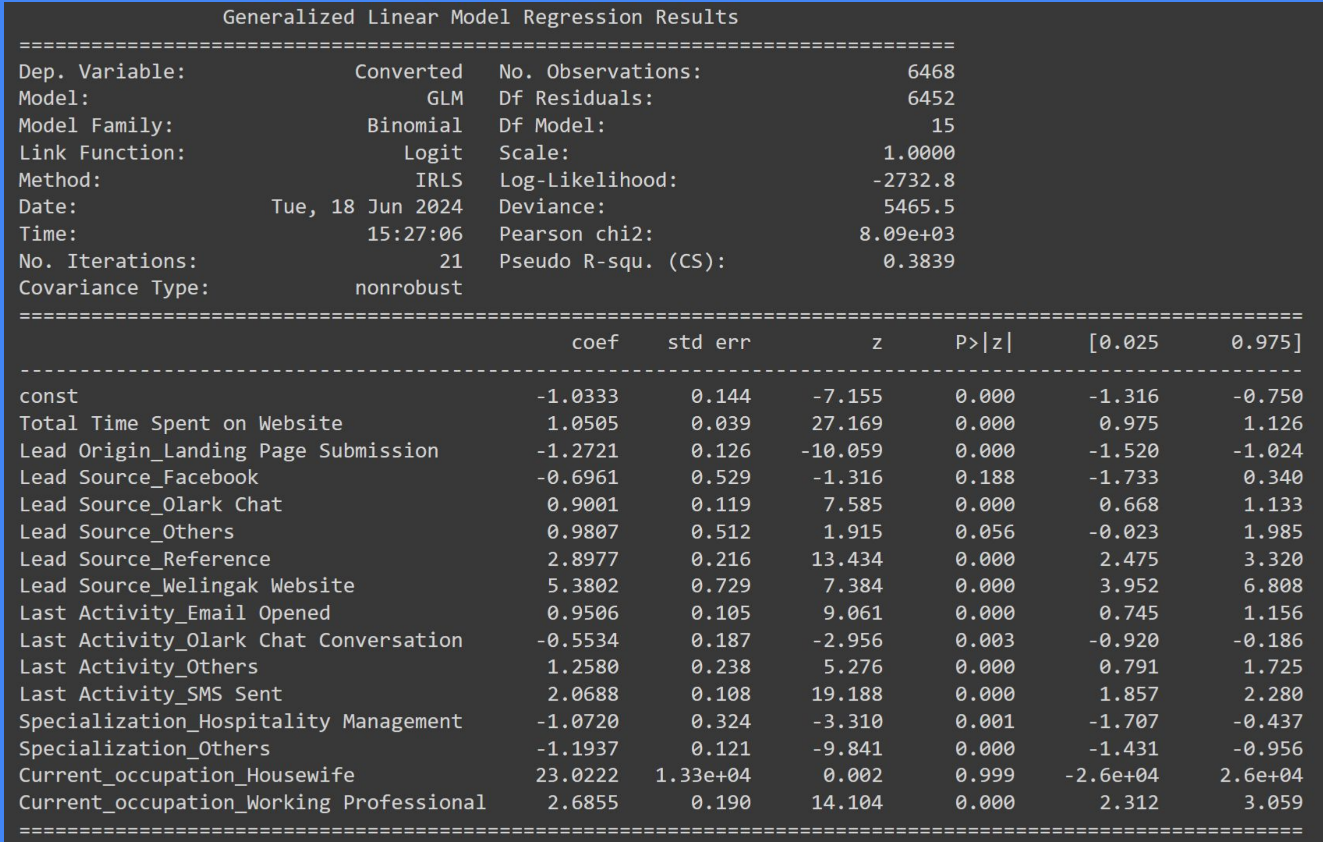
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6452
Model Family:                Binomial   Df Model:                           15
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2732.8
Date:                Tue, 18 Jun 2024   Deviance:                       5465.5
Time:                        15:27:06   Pearson chi2:                 8.09e+03
No. Iterations:                    21   Pseudo R-squ. (CS):             0.3839
Covariance Type:            nonrobust
============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                      -1.0333      0.144     -7.155      0.000      -1.316      -0.750
Total Time Spent on Website                 1.0505      0.039     27.169      0.000       0.975       1.126
Lead Origin_Landing Page Submission        -1.2721      0.126    -10.059      0.000      -1.520      -1.024
Lead Source_Facebook                       -0.6961      0.529     -1.316      0.188      -1.733       0.340
Lead Source_Olark Chat                      0.9001      0.119      7.585      0.000       0.668       1.133
Lead Source_Others                          0.9807      0.512      1.915      0.056      -0.023       1.985
Lead Source_Reference                       2.8977      0.216     13.434      0.000       2.475       3.320
Lead Source_Welingak Website                5.3802      0.729      7.384      0.000       3.952       6.808
Last Activity_Email Opened                  0.9506      0.105      9.061      0.000       0.745       1.156
Last Activity_Olark Chat Conversation      -0.5534      0.187     -2.956      0.003      -0.920      -0.186
Last Activity_Others                        1.2580      0.238      5.276      0.000       0.791       1.725
Last Activity_SMS Sent                      2.0688      0.108     19.188      0.000       1.857       2.280
Specialization_Hospitality Management      -1.0720      0.324     -3.310      0.001      -1.707      -0.437
Specialization_Others                      -1.1937      0.121     -9.841      0.000      -1.431      -0.956
Current_occupation_Housewife               23.0222   1.33e+04      0.002      0.999    -2.6e+04     2.6e+04
Current_occupation_Working Professional     2.6855      0.190     14.104      0.000       2.312       3.059
============================================================================================================
```

Model Evaluation

➔ Model 2

Dropping 'Current_occupation_Housewife' column

"Lead Source_Facebook" column will be removed from model due to high p-value of 0.187, which is above the accepted threshold of 0.05 for statistical significance.



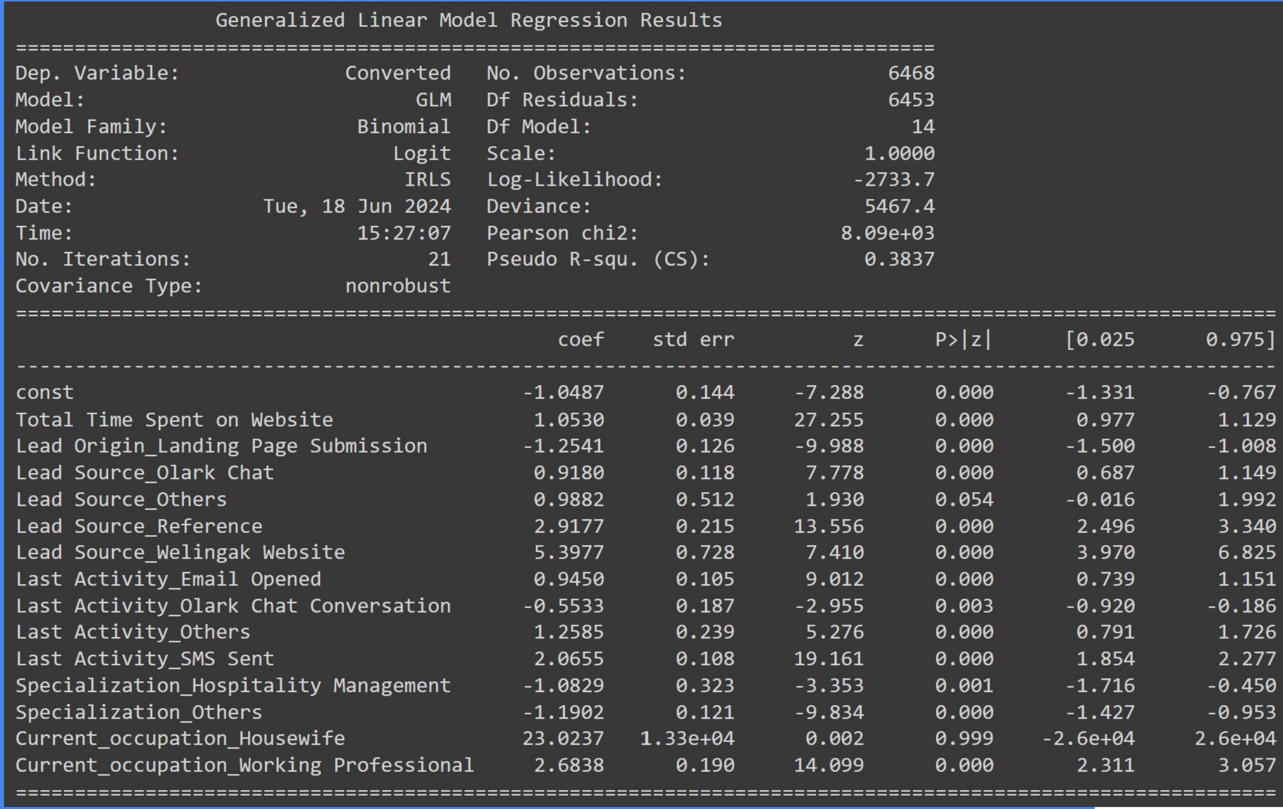
```
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6452
Model Family:                Binomial   Df Model:                           15
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2732.8
Date:                Tue, 18 Jun 2024   Deviance:                       5465.5
Time:                        15:27:06   Pearson chi2:                 8.09e+03
No. Iterations:                    21   Pseudo R-squ. (CS):             0.3839
Covariance Type:            nonrobust                                         
===========================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
const                                      -1.0333      0.144     -7.155      0.000      -1.316      -0.750
Total Time Spent on Website                 1.0505      0.039     27.169      0.000       0.975       1.126
Lead Origin_Landing Page Submission        -1.2721      0.126    -10.059      0.000      -1.520      -1.024
Lead Source_Facebook                       -0.6961      0.529     -1.316      0.188      -1.733       0.340
Lead Source_Olark Chat                      0.9001      0.119      7.585      0.000       0.668       1.133
Lead Source_Others                          0.9807      0.512      1.915      0.056      -0.023       1.985
Lead Source_Reference                       2.8977      0.216     13.434      0.000       2.475       3.320
Lead Source_Welingak Website                5.3802      0.729      7.384      0.000       3.952       6.808
Last Activity_Email Opened                  0.9506      0.105      9.061      0.000       0.745       1.156
Last Activity_Olark Chat Conversation      -0.5534      0.187     -2.956      0.003      -0.920      -0.186
Last Activity_Others                        1.2580      0.238      5.276      0.000       0.791       1.725
Last Activity_SMS Sent                      2.0688      0.108     19.188      0.000       1.857       2.280
Specialization_Hospitality Management      -1.0720      0.324     -3.310      0.001      -1.707      -0.437
Specialization_Others                      -1.1937      0.121     -9.841      0.000      -1.431      -0.956
Current_occupation_Housewife               23.0222   1.33e+04      0.002      0.999    -2.6e+04     2.6e+04
Current_occupation_Working Professional     2.6855      0.190     14.104      0.000       2.312       3.059
===========================================================================================================
```

Model Evaluation

- Model 3

# Dropping 'Lead Source_Facebook' column

"Lead Source_Others" column will be removed from model due to high p-value of 0.055, which is above the accepted threshold of 0.05 for statistical significance.



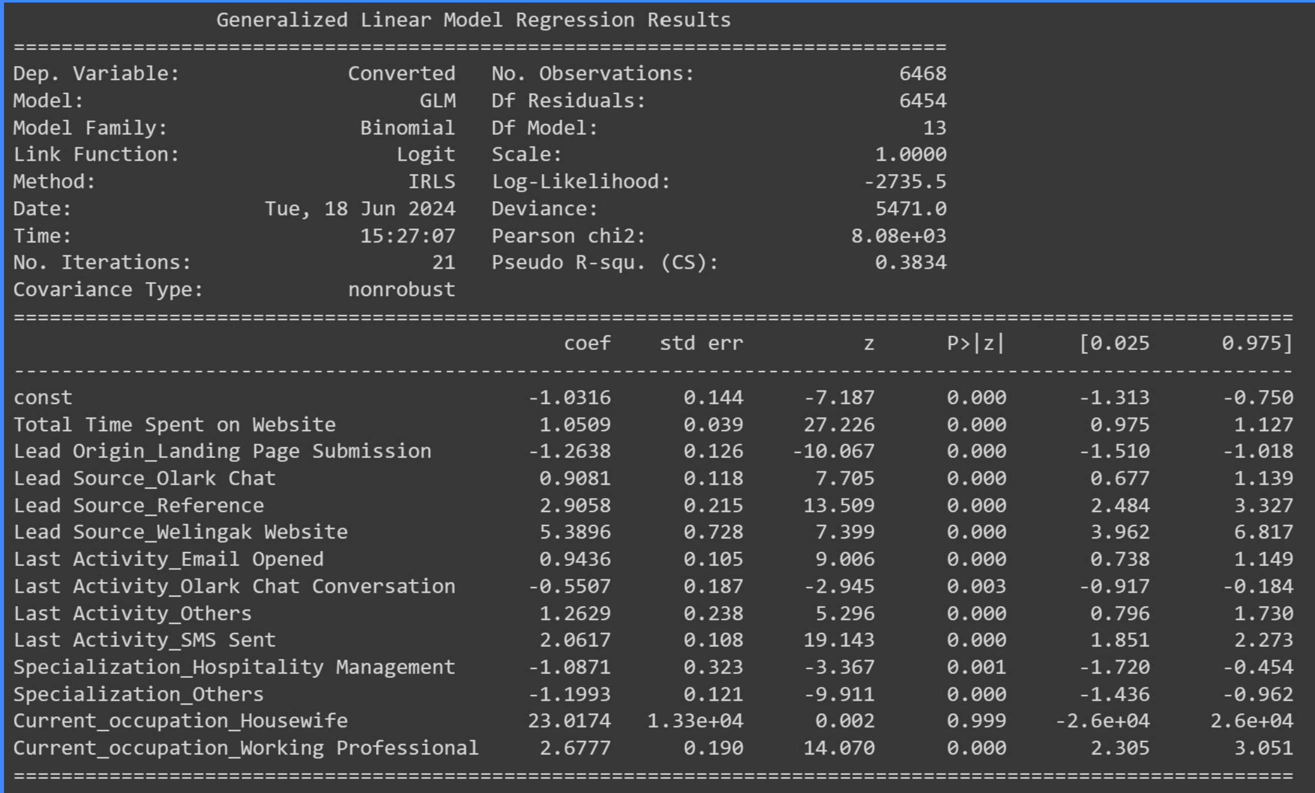
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6453
Model Family:                Binomial   Df Model:                           14
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2733.7
Date:                Tue, 18 Jun 2024   Deviance:                       5467.4
Time:                        15:27:07   Pearson chi2:                 8.09e+03
No. Iterations:                    21   Pseudo R-squ. (CS):             0.3837
Covariance Type:            nonrobust
=============================================================================================================
                                             coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                                     -1.0487      0.144     -7.288      0.000      -1.331      -0.767
Total Time Spent on Website                1.0530      0.039     27.255      0.000       0.977       1.129
Lead Origin_Landing Page Submission       -1.2541      0.126     -9.988      0.000      -1.500      -1.008
Lead Source_Olark Chat                     0.9180      0.118      7.778      0.000       0.687       1.149
Lead Source_Others                         0.9882      0.512      1.930      0.054      -0.016       1.992
Lead Source_Reference                      2.9177      0.215     13.556      0.000       2.496       3.340
Lead Source_Welingak Website               5.3977      0.728      7.410      0.000       3.970       6.825
Last Activity_Email Opened                 0.9450      0.105      9.012      0.000       0.739       1.151
Last Activity_Olark Chat Conversation     -0.5533      0.187     -2.955      0.003      -0.920      -0.186
Last Activity_Others                       1.2585      0.239      5.276      0.000       0.791       1.726
Last Activity_SMS Sent                     2.0655      0.108     19.161      0.000       1.854       2.277
Specialization_Hospitality Management     -1.0829      0.323     -3.353      0.001      -1.716      -0.450
Specialization_Others                     -1.1902      0.121     -9.834      0.000      -1.427      -0.953
Current_occupation_Housewife              23.0237   1.33e+04      0.002      0.999    -2.6e+04     2.6e+04
Current_occupation_Working Professional    2.6838      0.190     14.099      0.000       2.311       3.057
=============================================================================================================
```

Model Evaluation

- ➔ Model 4

# Dropping 'Lead Source_Facebook' column

Model 4 is stable and has significant p-values within the threshold (p-values < 0.05), so we will use it for further analysis.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6454
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2735.5
Date:                Tue, 18 Jun 2024   Deviance:                       5471.0
Time:                        15:27:07   Pearson chi2:                 8.08e+03
No. Iterations:                    21   Pseudo R-squ. (CS):             0.3834
Covariance Type:            nonrobust
===========================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
const                                      -1.0316      0.144     -7.187      0.000      -1.313      -0.750
Total Time Spent on Website                 1.0509      0.039     27.226      0.000       0.975       1.127
Lead Origin_Landing Page Submission        -1.2638      0.126    -10.067      0.000      -1.510      -1.018
Lead Source_Olark Chat                      0.9081      0.118      7.705      0.000       0.677       1.139
Lead Source_Reference                       2.9058      0.215     13.509      0.000       2.484       3.327
Lead Source_Welingak Website                5.3896      0.728      7.399      0.000       3.962       6.817
Last Activity_Email Opened                  0.9436      0.105      9.006      0.000       0.738       1.149
Last Activity_Olark Chat Conversation      -0.5507      0.187     -2.945      0.003      -0.917      -0.184
Last Activity_Others                        1.2629      0.238      5.296      0.000       0.796       1.730
Last Activity_SMS Sent                      2.0617      0.108     19.143      0.000       1.851       2.273
Specialization_Hospitality Management      -1.0871      0.323     -3.367      0.001      -1.720      -0.454
Specialization_Others                      -1.1993      0.121     -9.911      0.000      -1.436      -0.962
Current_occupation_Housewife               23.0174   1.33e+04      0.002      0.999    -2.6e+04     2.6e+04
Current_occupation_Working Professional     2.6777      0.190     14.070      0.000       2.305       3.051
===========================================================================================================
```

Model Evaluation

- ➔ Confusion Matrix

```
Predicted      not_converted  |  converted
Actual                                   |
------------------------------------------------------------
Not_converted        3588     |        414
converted            846      |       1620
```

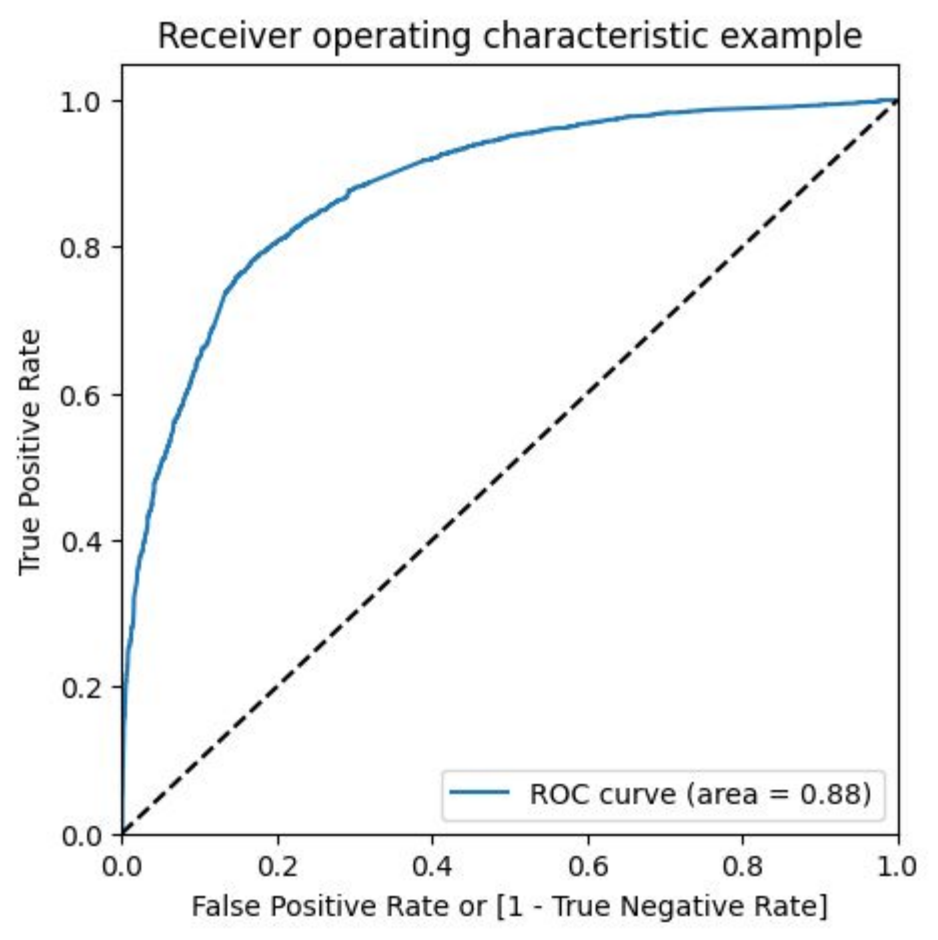
Above is the confusion matrix when we use threshold of probability as 0.5

- ➔ Accuracy: 80.55%
- ➔ Sensitivity : 65.77%
- ➔ Specificity : 89.65%

Model Evaluation

➔ Plotting the ROC Curve

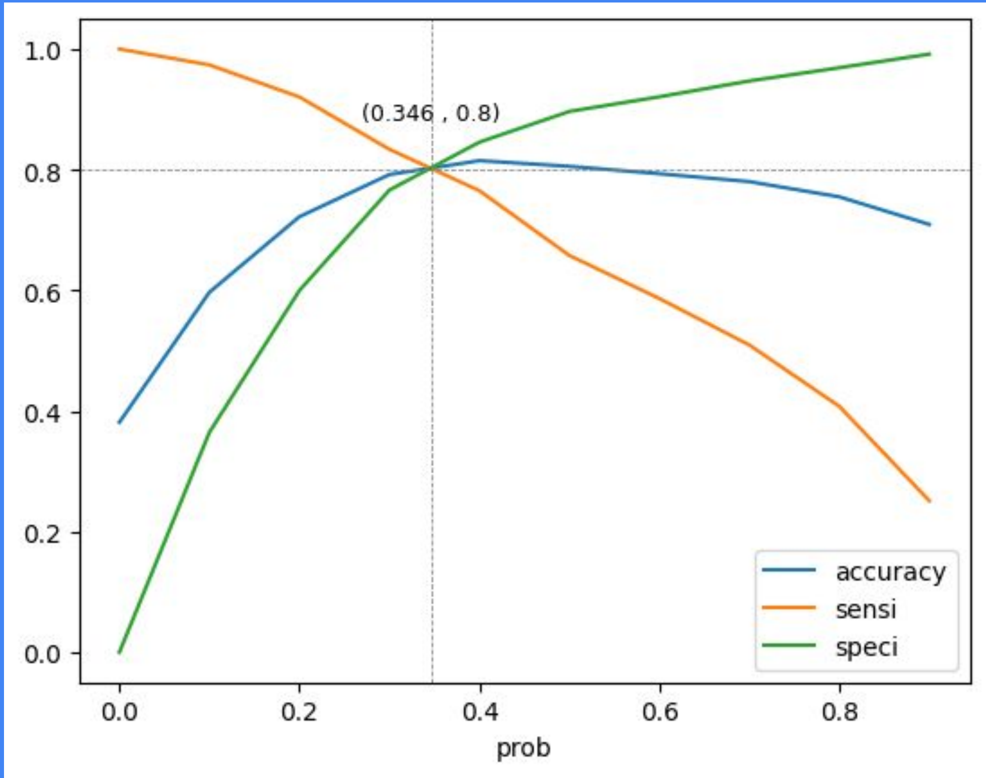
Area under ROC curve is 0.88 out of 1
which indicates a good predictive model



Model Evaluation

➔ Finding Optimal Cutoff Point/ Probability

0.345 is the approx. point where all the curves meet, so 0.345 seems to be our Optimal cutoff point for probability threshold.



Model Evaluation

➔ Calculating all metrics using confusion matrix for Train

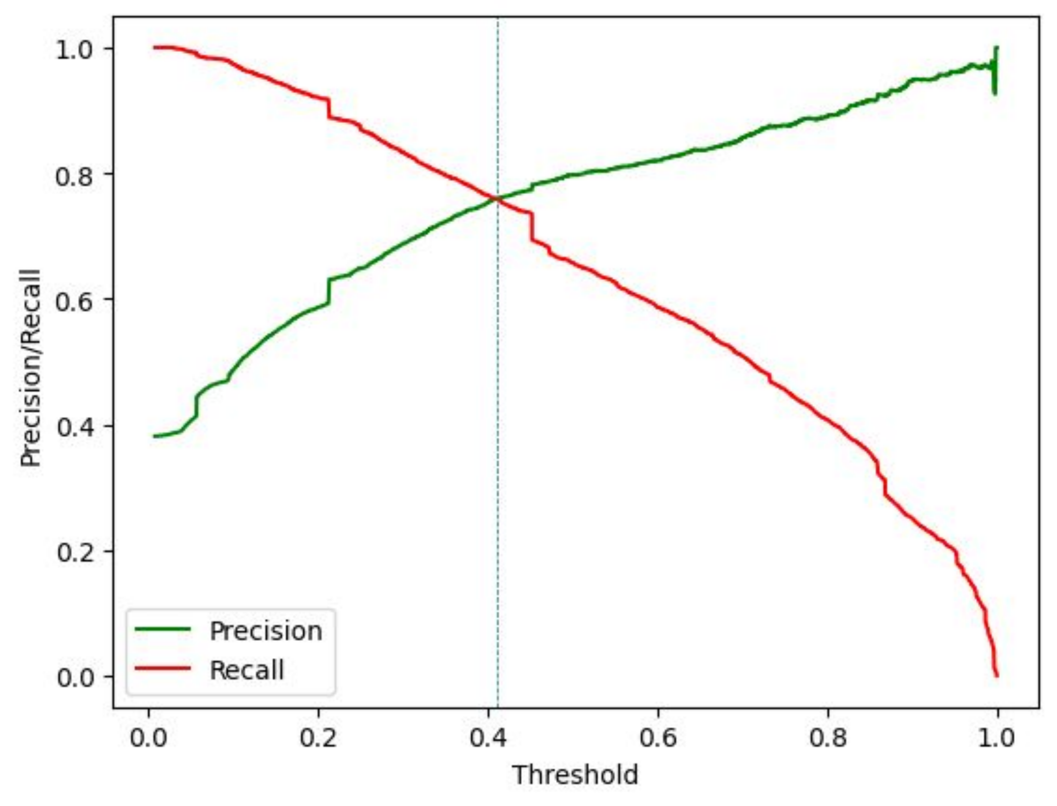
```
Confusion Matrix
[[3233  769] [ 492 1974]]

True Negative                    :  3233
True Positive                   :  1974
False Negative                   :  492
False Positive                   :  769
Model Accuracy                    :  0.805
Model Sensitivity                :  0.8005
Model Specificity                :  0.8078
Model Precision                  :  0.7197
Model Recall                    :  0.8005
Model True Positive Rate (TPR)   :  0.8005
Model False Positive Rate (FPR)  :  0.1922
```

Model Evaluation

➔ Precision and recall tradeoff

The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimise the performance of the model based on business requirement

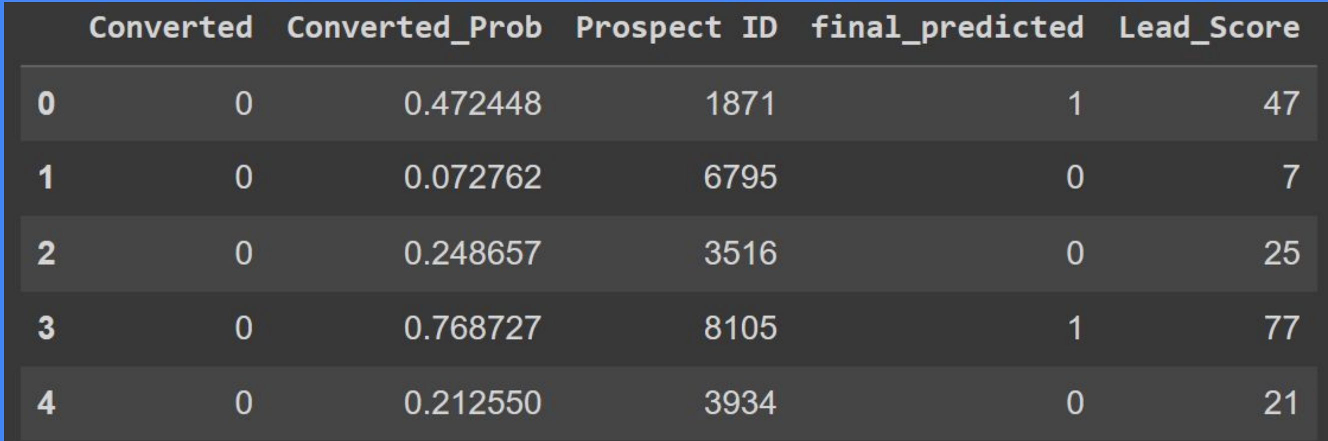
Here our probability threshold is 0.41 approximately from above curve.



Adding Lead Score Feature to Training dataframe

A higher score would mean that the lead is hot, i.e. is most likely to convert.

Lead Score is assigned to the customers
- The customers with a higher lead score have a higher conversion chance
- The customers with a lower lead score have a lower conversion chance.



	Converted	Converted_Prob	Prospect ID	final_predicted	Lead_Score
0	0	0.472448	1871	1	47
1	0	0.072762	6795	0	7
2	0	0.248657	3516	0	25
3	0	0.768727	8105	1	77
4	0	0.212550	3934	0	21

Making Predictions on Test Set

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

For Test set

- ➔ - Accuracy : 80.34%
- ➔ - Sensitivity : 79.82% ≈ 80%
- ➔ - Specificity : 80.68%

These metrics are very close to train set, so out final model logm4 is performing with good consistency on both Train & Test set

Making Predictions on Test Set

Features and their coefficient for final model:

Current_occupation_Housewife: 23.017356
Lead Source_Welingak Website: 5.389647
Lead Source_Reference: 2.905848
Current_occupation_Working Professional: 2.677711
Last Activity_SMS Sent: 2.061737
Last Activity_Others: 1.262880
Total Time Spent on Website: 1.050944
Last Activity_Email Opened: 0.943646
Lead Source_Olark Chat: 0.908112
Last Activity_Olark Chat Conversation: -0.550690
const: -1.031602
Specialization_Hospitality Management: -1.087051
Specialization_Others: -1.199302
Lead Origin_Landing Page Submission: -1.263769

Conclusion

Train Data Set:
Accuracy: 80.46%
Sensitivity: 80.05%
Specificity: 80.71%

Test Data Set:
Accuracy: 80.34%
Sensitivity: 79.82%
Specificity: 80.68%

NOTE: The evaluation metrics are close, indicating consistent performance across both datasets.

- The model achieved a sensitivity of 80.05% in the train and 79.82% in the test set using a cut-off value of 0.345.
- Sensitivity here reflects the proportion of correctly identified converting leads out of all potential converting leads.
- The CEO's target sensitivity of around 80% was achieved.
- The model also achieved an accuracy of 80.46%, aligning with the study's objectives.

Conclusion

Model parameters

The final Logistic Regression Model has 12 features

Top 3 features that contribute positively to predicting hot leads in the model are:

- ➔ Lead Source_Welingak Website
- ➔ Lead Source_Reference
- ➔ Current_occupation_Working Professional

NOTE: The optimal cutoff probability point is 0.345.
Converted probability greater than 0.345 will be predicted as Converted lead (Hot lead) & probability smaller than 0.345 will be predicted as not Converted lead (Cold lead).

Conclusion

Recommendations:

To increase our Lead Conversion Rates:

➔ Focus on features with positive coefficients for targeted marketing strategies.
➔ Develop strategies to attract high-quality leads from top-performing lead sources.
➔ Engage working professionals with tailored messaging.
➔ Optimize communication channels based on lead engagement impact.
➔ More budget/spend can be done on Welingak Website in terms of advertising, etc.
➔ Incentives/discounts for providing reference that convert to lead, encourage providing more references.
➔ Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement:

➔ Analyze negative coefficients in specialization offerings.
➔ Review landing page submission process for areas of improvement.