Lead Scoring Case Study

From:- Vedant Khairnar
Roshan Veervani
Shreyas Somani

Business Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Overall Approach

1. Reading and Understanding Data
2. Basic Data Cleaning
3. Visualizing Data & EDA
4. Data Preparation
5. Feature Engineering & Model Building
6. Model Evaluation on Train Data
7. Finding Optimal Probability Cutoff
8. Prediction on Test Data and Generating Lead Score (x100)
9. Model Evaluation on Test Data

Understanding Data Basic Data Cleaning

The dataset consists of 37 columns (30 categorical and 7 numeric) with 9,240 observations.

•The value *Select* appeared as a class in multiple columns, including *Specialization*, *How did you hear about X Education*, *Lead Profile*, and *City*. Since *Select* is likely a default placeholder in form dropdowns when no option is chosen, we replaced it with *NaN*.

•Columns such as *Magazine*, *Receive More Updates About Our Courses*, *Update me on Supply Chain Content*, *Get updates on DM Content*, and *I agree to pay the amount through cheque* contained only a single unique value and no missing data, making them redundant for EDA and model building. These were dropped due to lack of variance.

•Columns with more than 40% missing data, including *How did you hear about X Education*, *Lead Profile*, *Lead Quality*, *Asymmetries Activity Index*, *Asymmetries Profile Index*, *Asymmetries Activity Score*, and *Asymmetries Profile Score*, were also dropped from our analysis and modelling process.

•No rows in the dataset had more than 70% missing values.

•For categorical variables with a high number of classes but few data points, we created new bins, including columns like *Lead Origin*, *Lead Source*, *Last Activity*, *Last Notable Activity*, *Country*, *Specialization*, and *Occupation*.

•Missing values were treated based on business understanding. For instance, *NaN* values in *Specialization* and *Occupation* were replaced with a new category, *Not Disclosed*.

•Column names were modified for ease during EDA and model building: *What is your current occupation* was renamed to *Occupation*, and *What matters most to you in choosing a course* was renamed to *Reason choosing*.

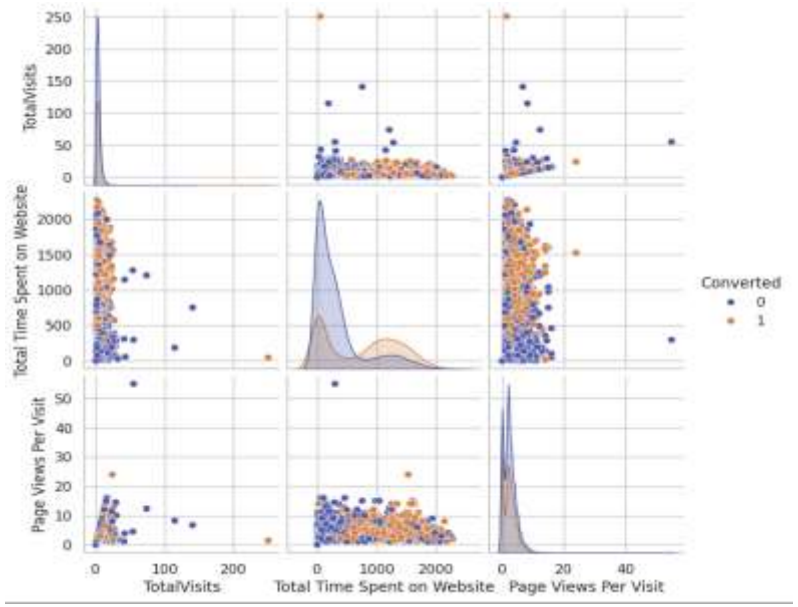
Visualizing Data & EDA: Numerical Variable

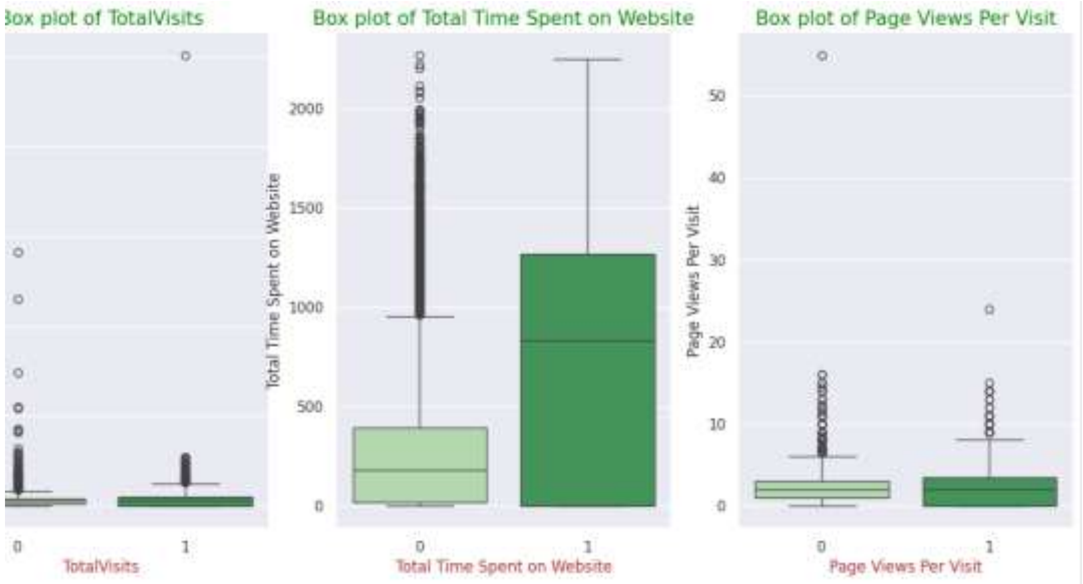
Inferences:

1.The median value of *Total Time Spent on Website* is significantly higher for converted leads compared to non-converted leads. This suggests that leads who spend more time on the website have a higher likelihood of conversion. The team should focus on targeting and engaging users who exhibit higher website engagement, as they have a greater potential for conversion.

2.A high number of outliers were observed in the *Total Visits* for non-converted leads (*Converted* = *0*). Despite multiple visits, many users are not opting for the course. The team should investigate potential reasons for this behaviour, such as financial constraints, lack of relevant course offerings, or the availability of better alternatives from competitors.

3.Numerous outliers were also identified in the *Total Time Spent on Website* for non-converted leads (*Converted* = *0*). This indicates that even some users who spent considerable time on the website did not convert, warranting further analysis to understand the barriers to conversion.

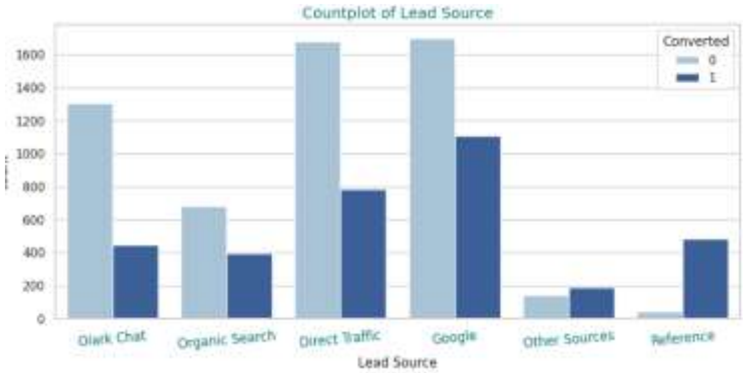


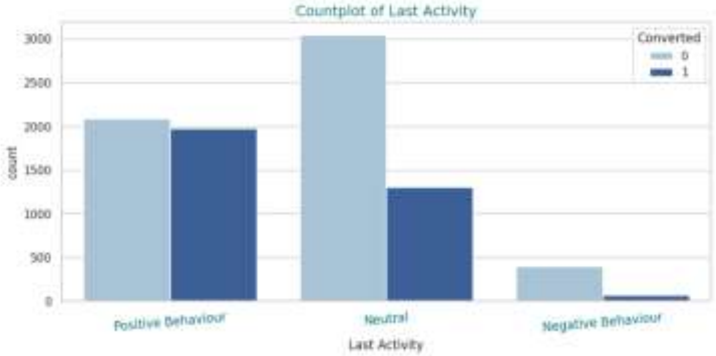


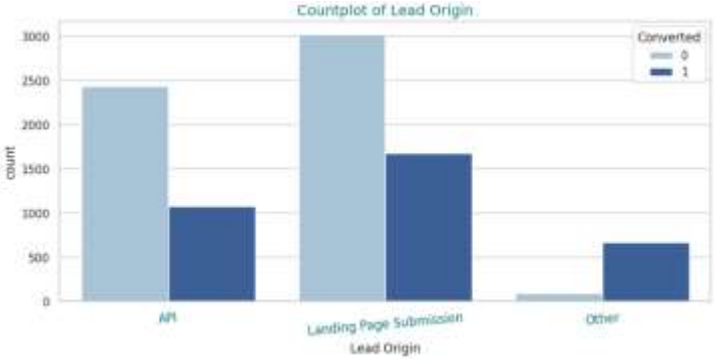
Visualizing Data & EDA: Categorical Variable

Leads with the "Other" type Lead Origin have a high likelihood of successful conversion. Reference-type Lead Sources show a strong success rate, making them a priority, while customers from Other Sources, though fewer in number, also have a high conversion rate.

Additionally, leads acquired through Organic Search have a significantly higher chance of conversion. Customers who have displayed positive behavior in their last activity are also more likely to convert successfully.



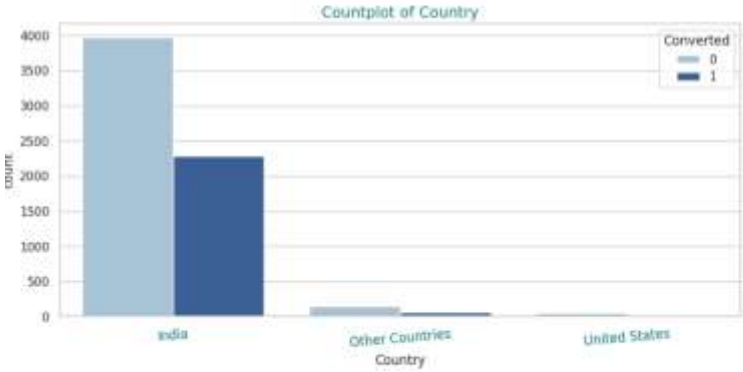


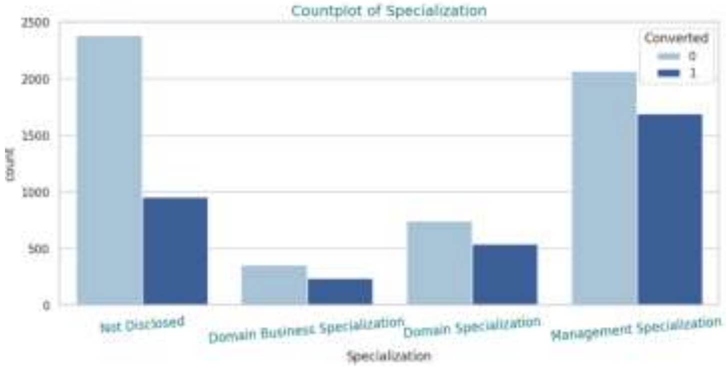


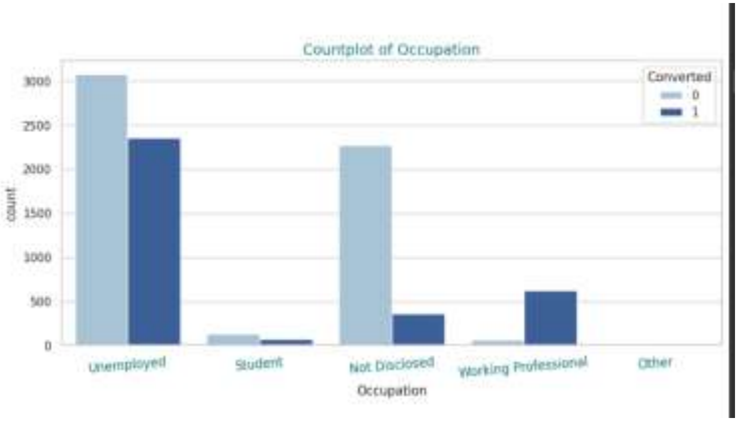
Visualizing Data & EDA: Categorical Variable

Most customers are from India and have Management Specializations. Those who have specified their specialization in the form are more likely to opt for the course. A significant number of interested customers are unemployed, while working professionals have a much higher chance of successful conversion.

The sales team should consider launching campaigns to attract more working professionals. Additionally, leads who mentioned their employment status while filling out the form have a higher likelihood of conversion.







Data Preparation

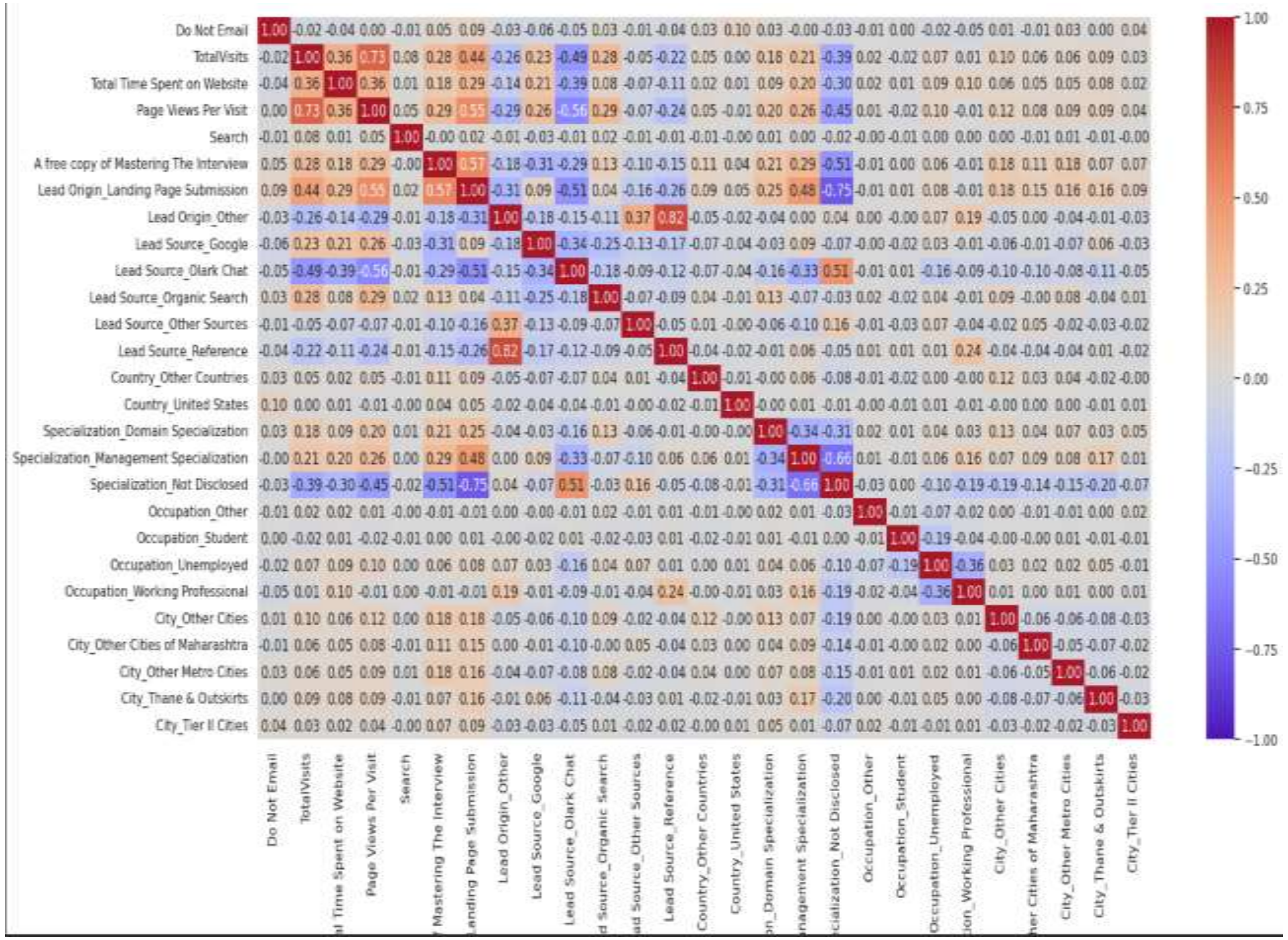
1. Outlier Treatment :- Identified 2.8% of total data (< 5%) as outliers and removed those rows
2. Train-Test Split :- The dataset was split into a 70:30 ratio, with the training dataset used to train the model and the test dataset used to evaluate its performance.
3. Missing Value Imputation :- The median and mode were calculated from the training dataset and used to impute missing values in both the training and test datasets. Statistical imputation was performed as follows: mode imputation for nominal categorical columns and median imputation for numeric columns
4. Categorical Variable Encoding :- Columns with *Yes* and *No* values (*Do Not Email, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations,* and *A free copy of Mastering The Interview*) were encoded by replacing *Yes* with 1 and *No* with 0. Additionally, dummy variables were created for categorical columns (*Lead Origin, Lead Source, Country, Specialization, Reason_choosing, Occupation,* and *City*), with the original columns and the first dummy variable for each category dropped from the dataset.
5. MinMax Scaling on Train Data :- MinMax Scaling (fit and transform) was performed on the training data for all numeric predictors, including *Total Visits*, *Total Time Spent on Website*, and *Page Views Per Visit*.
6. Variance Thresholding :- Variance thresholding was applied with a threshold of 0.001, and columns with lower variance were removed, including *Do Not Call, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations,* and *Reason_choosing_Flexibility & Convenience*.

Data Preparation : Pairwise Correlation

Observation:

Lead Origin_Other exhibited a high correlation (0.82) with *Lead Source_Reference*. To avoid multicollinearity, the *Lead Source_Reference* column was dropped.

Lead Origin_Landing Page Submission showed a high correlation (0.75) with *Specialization_Not Disclosed*. Consequently, the *Specialization_Not Disclosed* column was removed to maintain the integrity of the dataset during model building.



Model Building : Approach

- **Feature Selection:**

  Recursive Feature Elimination (RFE) was employed to select the top 16 features:

  - **Do Not Email:** Indicator variable selected by the customer to opt out of receiving course-related emails.
  - **Total Visits:** Total number of visits made by the customer on the website.
  - **Total Time Spent on Website:** Total time spent by the customer on the website.
  - **Page Views Per Visit:** Average number of pages viewed per visit.
  - **Lead Origin_Landing Page Submission:** Dummy variable for leads originating from landing page submissions.
  - **Lead Origin_Other:** Dummy variable for leads originating from other sources.
  - **Lead Source_Olark Chat:** Dummy variable for leads sourced from Olark Chat.
  - **Lead Source_Other Sources:** Dummy variable for leads from sources other than Google, Direct Traffic, Olark Chat, Organic Search, and Reference.
  - **Country_Other Countries:** Dummy variable for customers from countries other than India and the United States.
  - **Specialization_Domain Specialization:** Dummy variable for the domain specialization category.
  - **Specialization_Management Specialization:** Dummy variable for the management specialization category.
  - **Occupation_Other:** Dummy variable for the ‘Other’ occupation category.
  - **Occupation_Student:** Dummy variable for the ‘Student’ occupation category.
  - **Occupation_Unemployed:** Dummy variable for the ‘Unemployed’ occupation category.
  - **Occupation_Working Professional:** Dummy variable for the ‘Working Professional’ occupation category.
  - **City_Tier II Cities:** Dummy variable for the ‘Tier II Cities’ category.

Model Building: Approach

•**Model Development:**
The first logistic regression model was built using a Generalized Linear Model (GLM) in *statsmodels* with these 16 selected features.

•**Model Fine-Tuning:**
The model was manually fine-tuned by:

- Checking the statistical significance of features using p-values (with an acceptance threshold of $p < 0.05$).
- Removing multicollinearity using Variance Inflation Factors (VIF < 5).

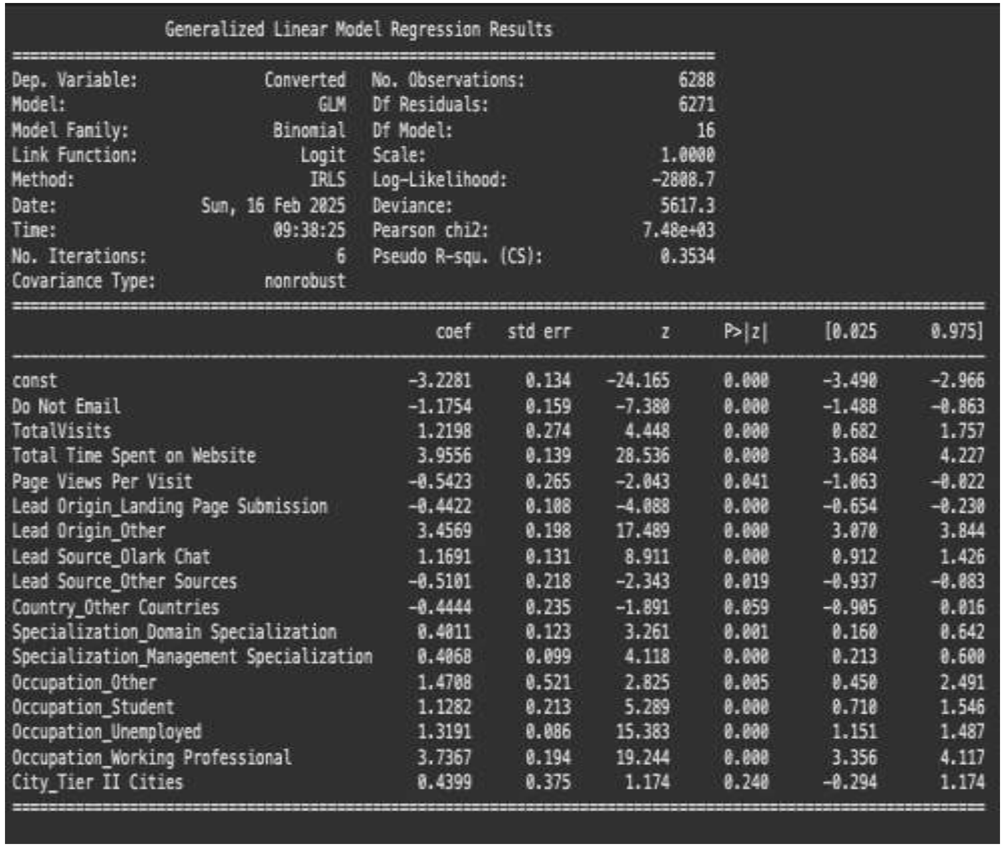
•**Iterative Process:**
A total of seven models were built. After each iteration:

  - p-values and VIFs were reviewed.
  - Features violating the thresholds were removed in subsequent models.
  - Model performance was evaluated using overall accuracy and confusion matrix comparisons to ensure improvements over previous iterations.

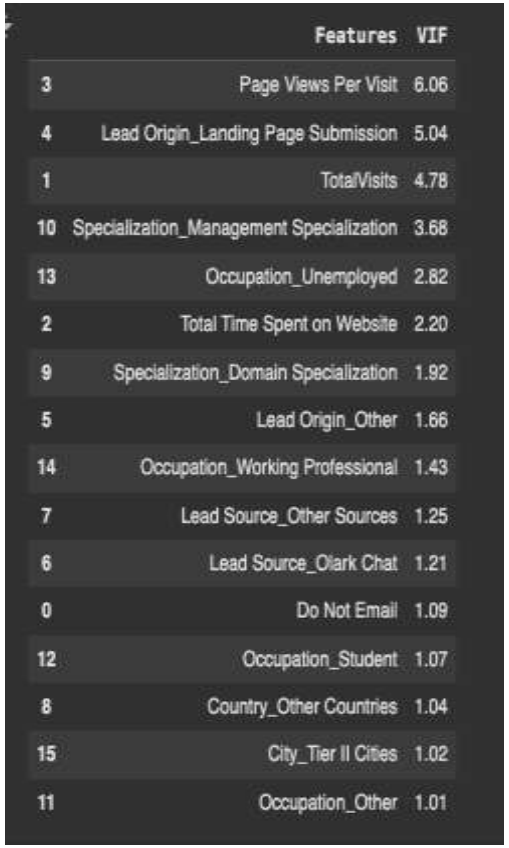
Model Building : Model 1

1.'City_Tier II Cities' and 'Country_Other Countries' have p value higher than .05. So their coefficient value is not statistically significant.

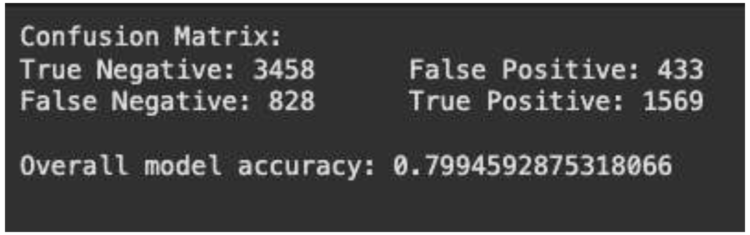
2.Let's see VIFs, to check if there is any multicollinearity present.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6288
Model:                            GLM   Df Residuals:                     6271
Model Family:                Binomial   Df Model:                           16
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2808.7
Date:                Sun, 16 Feb 2025   Deviance:                       5617.3
Time:                        09:38:25   Pearson chi2:                 7.48e+03
No. Iterations:                     6   Pseudo R-squ. (CS):             0.3534
Covariance Type:            nonrobust
==============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                      -3.2281      0.134    -24.165      0.000      -3.490      -2.966
Do Not Email                               -1.1754      0.159     -7.380      0.000      -1.488      -0.863
TotalVisits                                 1.2198      0.274      4.448      0.000       0.682       1.757
Total Time Spent on Website                 3.9556      0.139     28.536      0.000       3.684       4.227
Page Views Per Visit                       -0.5423      0.265     -2.043      0.041      -1.063      -0.022
Lead Origin_Landing Page Submission        -0.4422      0.108     -4.088      0.000      -0.654      -0.230
Lead Origin_Other                           3.4569      0.198     17.489      0.000       3.070       3.844
Lead Source_Olark Chat                      1.1691      0.131      8.911      0.000       0.912       1.426
Lead Source_Other Sources                  -0.5101      0.218     -2.343      0.019      -0.937      -0.083
Country_Other Countries                    -0.4444      0.235     -1.891      0.059      -0.905       0.016
Specialization_Domain Specialization        0.4011      0.123      3.261      0.001       0.160       0.642
Specialization_Management Specialization    0.4068      0.099      4.118      0.000       0.213       0.600
Occupation_Other                            1.4708      0.521      2.825      0.005       0.450       2.491
Occupation_Student                          1.1282      0.213      5.289      0.000       0.710       1.546
Occupation_Unemployed                       1.3191      0.086     15.383      0.000       1.151       1.487
Occupation_Working Professional             3.7367      0.194     19.244      0.000       3.356       4.117
City_Tier II Cities                         0.4399      0.375      1.174      0.240      -0.294       1.174
==============================================================================================================
```



	Features	VIF
3	Page Views Per Visit	6.06
4	Lead Origin_Landing Page Submission	5.04
1	TotalVisits	4.78
10	Specialization_Management Specialization	3.68
13	Occupation_Unemployed	2.82
2	Total Time Spent on Website	2.20
9	Specialization_Domain Specialization	1.92
5	Lead Origin_Other	1.66
14	Occupation_Working Professional	1.43
7	Lead Source_Other Sources	1.25
6	Lead Source_Olark Chat	1.21
0	Do Not Email	1.09
12	Occupation_Student	1.07
8	Country_Other Countries	1.04
15	City_Tier II Cities	1.02
11	Occupation_Other	1.01



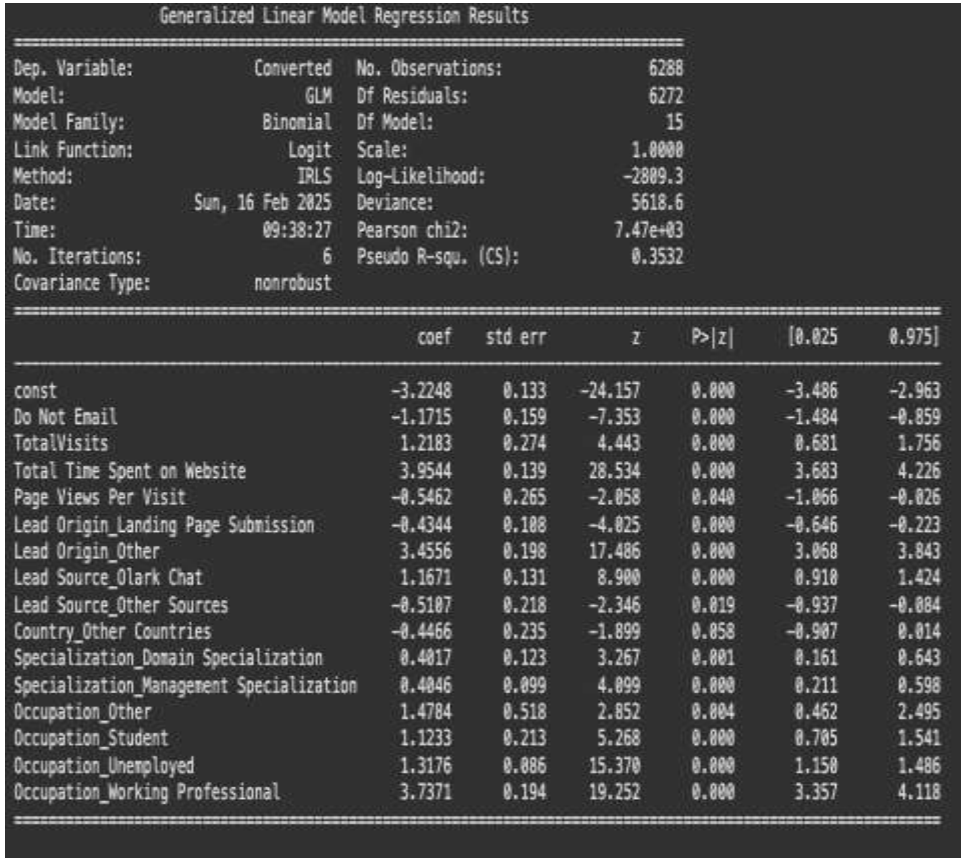
```
Confusion Matrix:
True Negative: 3458      False Positive: 433
False Negative: 828      True Positive: 1569

Overall model accuracy: 0.7994592875318066
```

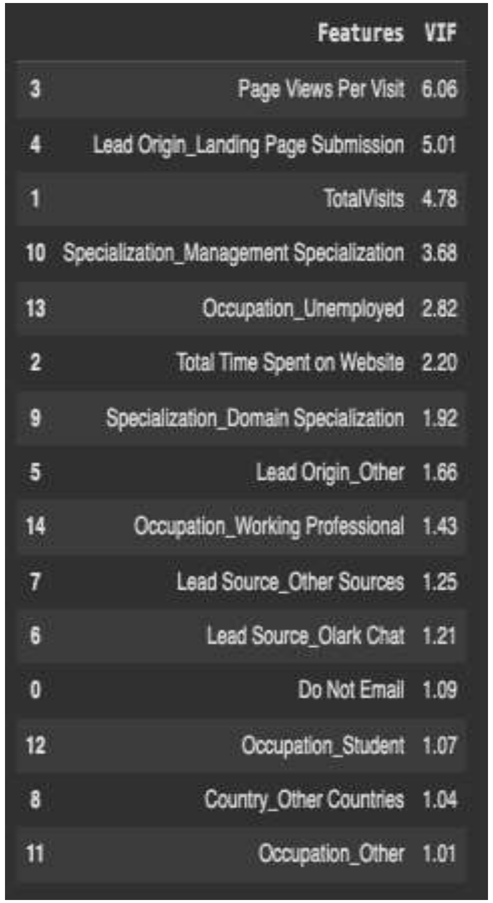
Model Building : Model 2

1.Country_Other Countries' has p value higher than .05. So its coefficient value is not statistically significant.

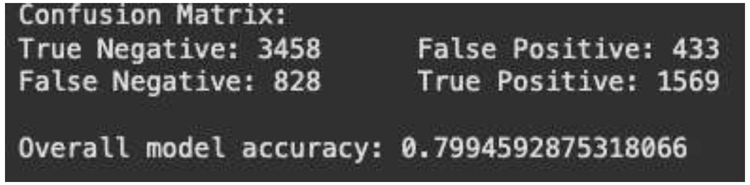
2.Let's see VIFs, to check if there is any multicollinearity present.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6288
Model:                            GLM   Df Residuals:                     6272
Model Family:                Binomial   Df Model:                           15
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2809.3
Date:                Sun, 16 Feb 2025   Deviance:                       5618.6
Time:                        09:38:27   Pearson chi2:                 7.47e+03
No. Iterations:                     6   Pseudo R-squ. (CS):             0.3532
Covariance Type:            nonrobust
============================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                      -3.2248      0.133    -24.157      0.000      -3.486      -2.963
Do Not Email                               -1.1715      0.159     -7.353      0.000      -1.484      -0.859
TotalVisits                                 1.2183      0.274      4.443      0.000       0.681       1.756
Total Time Spent on Website                 3.9544      0.139     28.534      0.000       3.683       4.226
Page Views Per Visit                       -0.5462      0.265     -2.058      0.040      -1.066      -0.026
Lead Origin_Landing Page Submission        -0.4344      0.108     -4.025      0.000      -0.646      -0.223
Lead Origin_Other                           3.4556      0.198     17.486      0.000       3.068       3.843
Lead Source_Olark Chat                      1.1671      0.131      8.900      0.000       0.910       1.424
Lead Source_Other Sources                  -0.5107      0.218     -2.346      0.019      -0.937      -0.084
Country_Other Countries                    -0.4466      0.235     -1.899      0.058      -0.907       0.014
Specialization_Domain Specialization        0.4017      0.123      3.267      0.001       0.161       0.643
Specialization_Management Specialization    0.4046      0.099      4.099      0.000       0.211       0.598
Occupation_Other                            1.4784      0.518      2.852      0.004       0.462       2.495
Occupation_Student                          1.1233      0.213      5.268      0.000       0.705       1.541
Occupation_Unemployed                       1.3176      0.086     15.370      0.000       1.150       1.486
Occupation_Working Professional             3.7371      0.194     19.252      0.000       3.357       4.118
============================================================================================================
```



	Features	VIF
3	Page Views Per Visit	6.06
4	Lead Origin_Landing Page Submission	5.01
1	TotalVisits	4.78
10	Specialization_Management Specialization	3.68
13	Occupation_Unemployed	2.82
2	Total Time Spent on Website	2.20
9	Specialization_Domain Specialization	1.92
5	Lead Origin_Other	1.66
14	Occupation_Working Professional	1.43
7	Lead Source_Other Sources	1.25
6	Lead Source_Olark Chat	1.21
0	Do Not Email	1.09
12	Occupation_Student	1.07
8	Country_Other Countries	1.04
11	Occupation_Other	1.01



```
Confusion Matrix:
True Negative: 3458      False Positive: 433
False Negative: 828      True Positive: 1569

Overall model accuracy: 0.7994592875318066
```

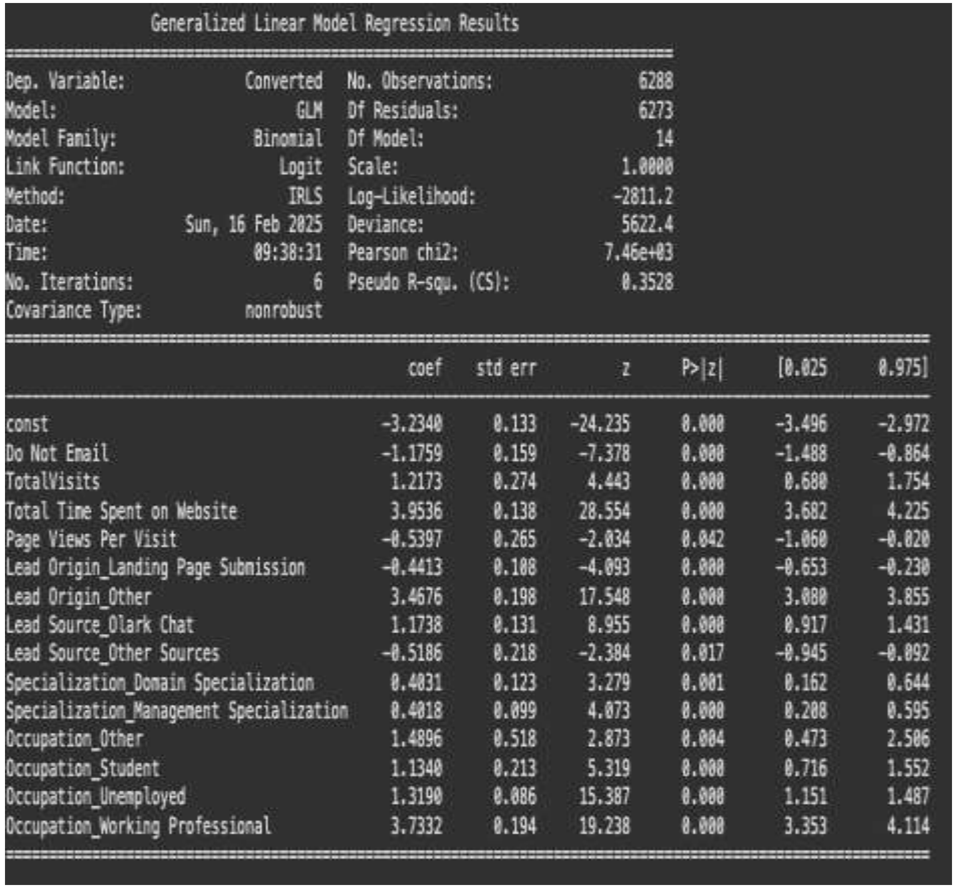
Model Building : Model 3

Observations:

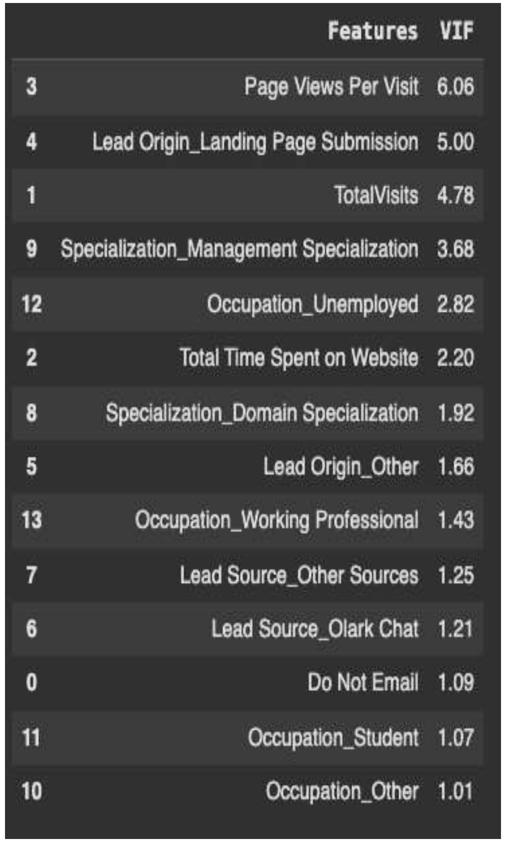
•Comparing Model 7 with Model 2 reveals a minimal change in the confusion matrix. The number of True Positives (TP) slightly increased, while the number of True Negatives (TN) decreased by the same margin.

•This change did not lead to any significant reduction in the overall model accuracy.

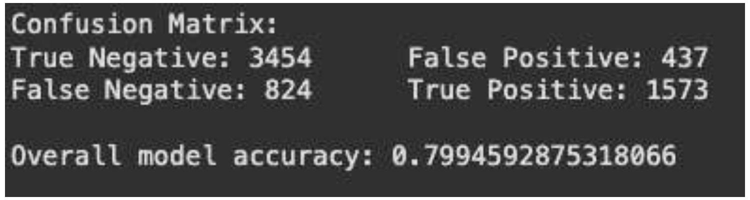
•The feature *Page Views Per Visit* still exhibited a slightly high Variance Inflation Factor (VIF) value, prompting its exclusion in the next model iteration to address multicollinearity concerns.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6288
Model:                            GLM   Df Residuals:                     6273
Model Family:                Binomial   Df Model:                           14
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2811.2
Date:                Sun, 16 Feb 2025   Deviance:                       5622.4
Time:                        09:38:31   Pearson chi2:                 7.46e+03
No. Iterations:                     6   Pseudo R-squ. (CS):             0.3528
Covariance Type:            nonrobust
==========================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                      -3.2340      0.133    -24.235      0.000      -3.496      -2.972
Do Not Email                               -1.1759      0.159     -7.378      0.000      -1.488      -0.864
TotalVisits                                 1.2173      0.274      4.443      0.000       0.680       1.754
Total Time Spent on Website                 3.9536      0.138     28.554      0.000       3.682       4.225
Page Views Per Visit                       -0.5397      0.265     -2.034      0.042      -1.060      -0.020
Lead Origin_Landing Page Submission        -0.4413      0.108     -4.093      0.000      -0.653      -0.230
Lead Origin_Other                           3.4676      0.198     17.548      0.000       3.080       3.855
Lead Source_Olark Chat                      1.1738      0.131      8.955      0.000       0.917       1.431
Lead Source_Other Sources                  -0.5186      0.218     -2.384      0.017      -0.945      -0.092
Specialization_Domain Specialization        0.4031      0.123      3.279      0.001       0.162       0.644
Specialization_Management Specialization    0.4018      0.099      4.073      0.000       0.208       0.595
Occupation_Other                            1.4896      0.518      2.873      0.004       0.473       2.506
Occupation_Student                          1.1340      0.213      5.319      0.000       0.716       1.552
Occupation_Unemployed                       1.3190      0.086     15.387      0.000       1.151       1.487
Occupation_Working Professional             3.7332      0.194     19.238      0.000       3.353       4.114
==========================================================================================================
```



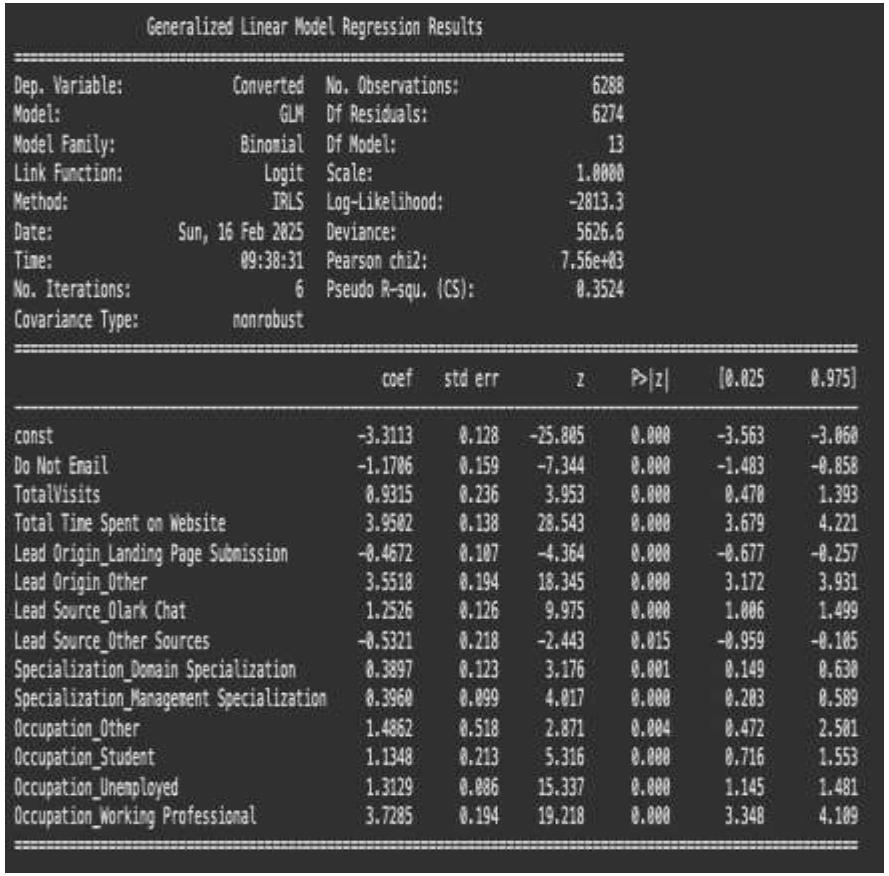
	Features	VIF
3	Page Views Per Visit	6.06
4	Lead Origin_Landing Page Submission	5.00
1	TotalVisits	4.78
9	Specialization_Management Specialization	3.68
12	Occupation_Unemployed	2.82
2	Total Time Spent on Website	2.20
8	Specialization_Domain Specialization	1.92
5	Lead Origin_Other	1.66
13	Occupation_Working Professional	1.43
7	Lead Source_Other Sources	1.25
6	Lead Source_Olark Chat	1.21
0	Do Not Email	1.09
11	Occupation_Student	1.07
10	Occupation_Other	1.01



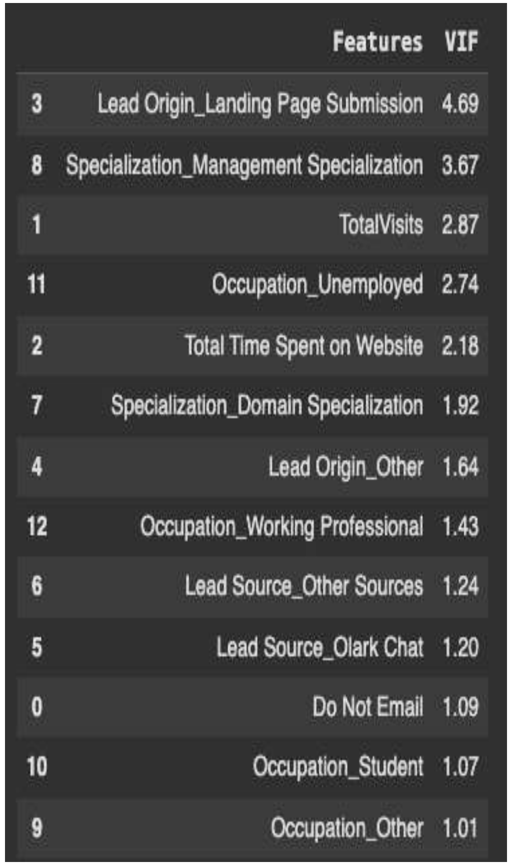
```
Confusion Matrix:
True Negative: 3454      False Positive: 437
False Negative: 824      True Positive: 1573

Overall model accuracy: 0.7994592875318066
```

Model Building : Model 4



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6288
Model:                            GLM   Df Residuals:                     6274
Model Family:                Binomial   Df Model:                           13
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2813.3
Date:                Sun, 16 Feb 2025   Deviance:                       5626.6
Time:                        09:38:31   Pearson chi2:                 7.56e+03
No. Iterations:                     6   Pseudo R-squ. (CS):             0.3524
Covariance Type:            nonrobust
=========================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                      -3.3113      0.128    -25.805      0.000      -3.563      -3.060
Do Not Email                               -1.1706      0.159     -7.344      0.000      -1.483      -0.858
TotalVisits                                 0.9315      0.236      3.953      0.000       0.470       1.393
Total Time Spent on Website                 3.9502      0.138     28.543      0.000       3.679       4.221
Lead Origin_Landing Page Submission        -0.4672      0.107     -4.364      0.000      -0.677      -0.257
Lead Origin_Other                           3.5518      0.194     18.345      0.000       3.172       3.931
Lead Source_Olark Chat                      1.2526      0.126      9.975      0.000       1.006       1.499
Lead Source_Other Sources                  -0.5321      0.218     -2.443      0.015      -0.959      -0.105
Specialization_Domain Specialization        0.3897      0.123      3.176      0.001       0.149       0.630
Specialization_Management Specialization    0.3960      0.099      4.017      0.000       0.203       0.589
Occupation_Other                            1.4862      0.518      2.871      0.004       0.472       2.501
Occupation_Student                          1.1348      0.213      5.316      0.000       0.716       1.553
Occupation_Unemployed                       1.3129      0.086     15.337      0.000       1.145       1.481
Occupation_Working Professional             3.7285      0.194     19.218      0.000       3.348       4.109
=========================================================================================================
```



	Features	VIF
3	Lead Origin_Landing Page Submission	4.69
8	Specialization_Management Specialization	3.67
1	TotalVisits	2.87
11	Occupation_Unemployed	2.74
2	Total Time Spent on Website	2.18
7	Specialization_Domain Specialization	1.92
4	Lead Origin_Other	1.64
12	Occupation_Working Professional	1.43
6	Lead Source_Other Sources	1.24
5	Lead Source_Olark Chat	1.20
0	Do Not Email	1.09
10	Occupation_Student	1.07
9	Occupation_Other	1.01

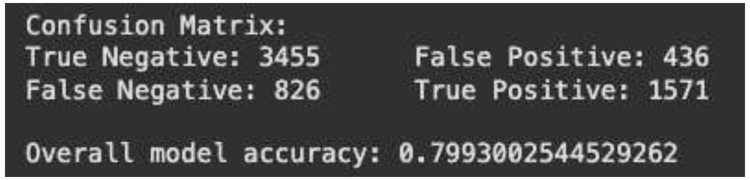
Observations:

•The model accuracy in Model 4 remains nearly the same as in Model 3.

•In Model 4, the p-values of all predictor coefficients are within the acceptable range, indicating that all predictors are statistically significant.

•However, *Lead Origin_Landing Page Submission* exhibited a slightly higher VIF value, though still below the acceptable threshold of 5.

•To address potential multicollinearity, *Lead Origin_Landing Page Submission* was dropped in the next model iteration. This step aimed to assess whether its removal leads to any significant change in the overall model accuracy.



```
Confusion Matrix:
True Negative: 3455      False Positive: 436
False Negative: 826      True Positive: 1571

Overall model accuracy: 0.7993002544529262
```

Model Building : Model 5

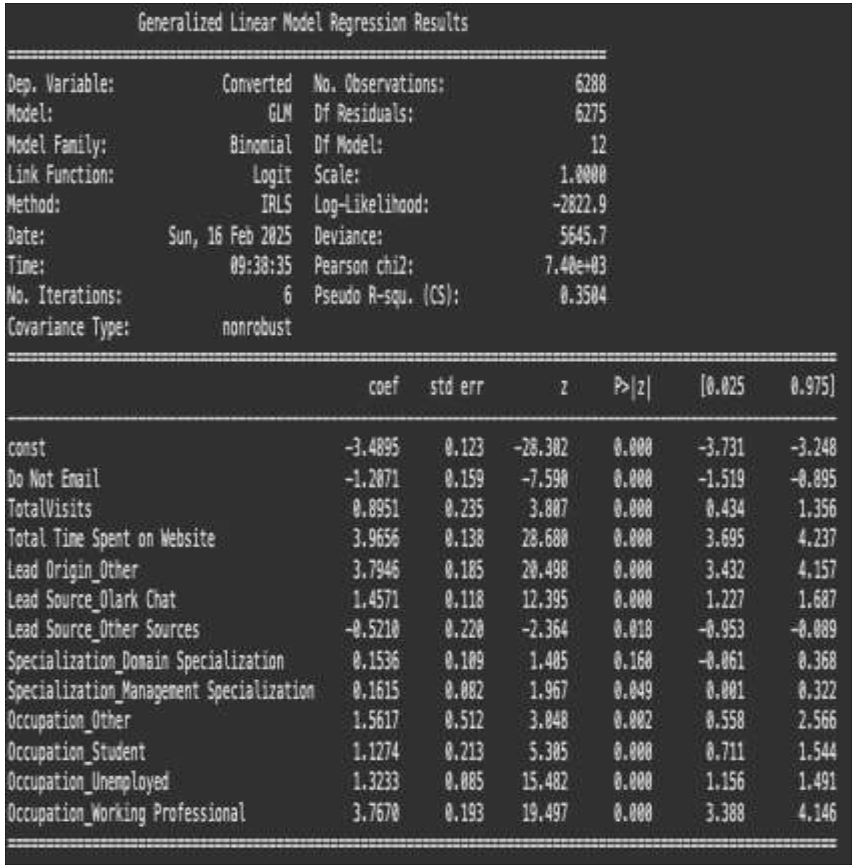
Observations:

•Model 5 demonstrates nearly the same overall accuracy as the previous model.

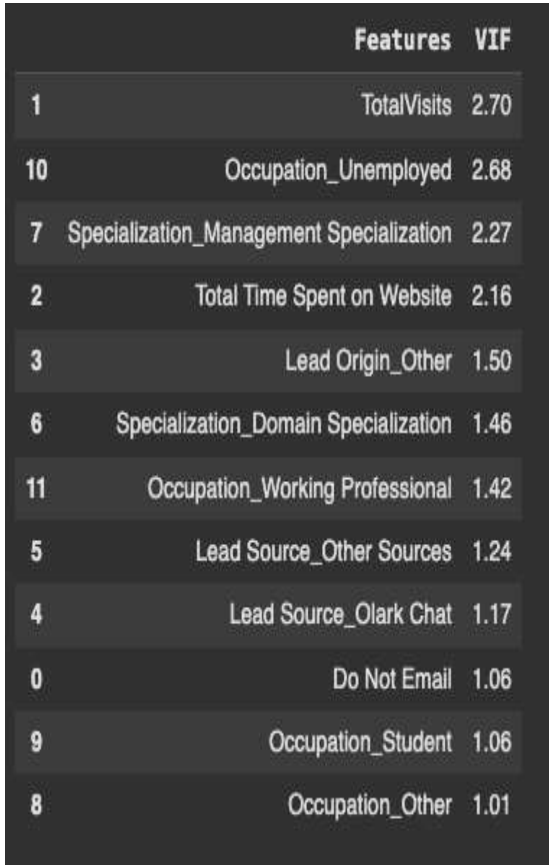
•After removing *Lead Origin_Landing Page Submission*, the VIF value of *Specialization_Management Specialization* significantly decreased, indicating that multicollinearity is no longer a concern in the model.

•However, the p-value of the *Specialization_Domain Specialization* coefficient is now higher than the acceptable threshold of 0.05, suggesting that this predictor is no longer statistically significant.

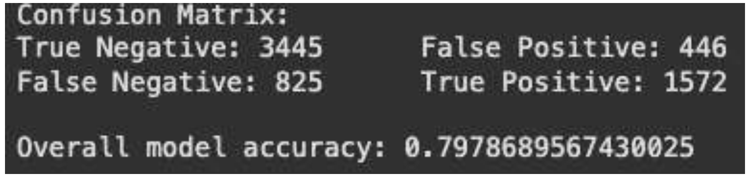
•Consequently, *Specialization_Domain Specialization* was excluded in the next model iteration to improve the model’s statistical robustness.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6288
Model:                            GLM   Df Residuals:                     6275
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2822.9
Date:                Sun, 16 Feb 2025   Deviance:                       5645.7
Time:                        09:38:35   Pearson chi2:                 7.40e+03
No. Iterations:                     6   Pseudo R-squ. (CS):             0.3504
Covariance Type:            nonrobust
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                       -3.4895      0.123    -28.302      0.000      -3.731      -3.248
Do Not Email                                -1.2071      0.159     -7.590      0.000      -1.519      -0.895
TotalVisits                                  0.8951      0.235      3.807      0.000       0.434       1.356
Total Time Spent on Website                  3.9656      0.138     28.680      0.000       3.695       4.237
Lead Origin_Other                            3.7946      0.185     20.498      0.000       3.432       4.157
Lead Source_Olark Chat                       1.4571      0.118     12.395      0.000       1.227       1.687
Lead Source_Other Sources                   -0.5210      0.220     -2.364      0.018      -0.953      -0.089
Specialization_Domain Specialization         0.1536      0.109      1.405      0.160      -0.061       0.368
Specialization_Management Specialization     0.1615      0.082      1.967      0.049       0.001       0.322
Occupation_Other                             1.5617      0.512      3.048      0.002       0.558       2.566
Occupation_Student                           1.1274      0.213      5.305      0.000       0.711       1.544
Occupation_Unemployed                        1.3233      0.085     15.482      0.000       1.156       1.491
Occupation_Working Professional              3.7670      0.193     19.497      0.000       3.388       4.146
============================================================================================================
```



	Features	VIF
1	TotalVisits	2.70
10	Occupation_Unemployed	2.68
7	Specialization_Management Specialization	2.27
2	Total Time Spent on Website	2.16
3	Lead Origin_Other	1.50
6	Specialization_Domain Specialization	1.46
11	Occupation_Working Professional	1.42
5	Lead Source_Other Sources	1.24
4	Lead Source_Olark Chat	1.17
0	Do Not Email	1.06
9	Occupation_Student	1.06
8	Occupation_Other	1.01



```
Confusion Matrix:
True Negative: 3445      False Positive: 446
False Negative: 825      True Positive: 1572

Overall model accuracy: 0.7978689567430025
```

Model Building : Model 6

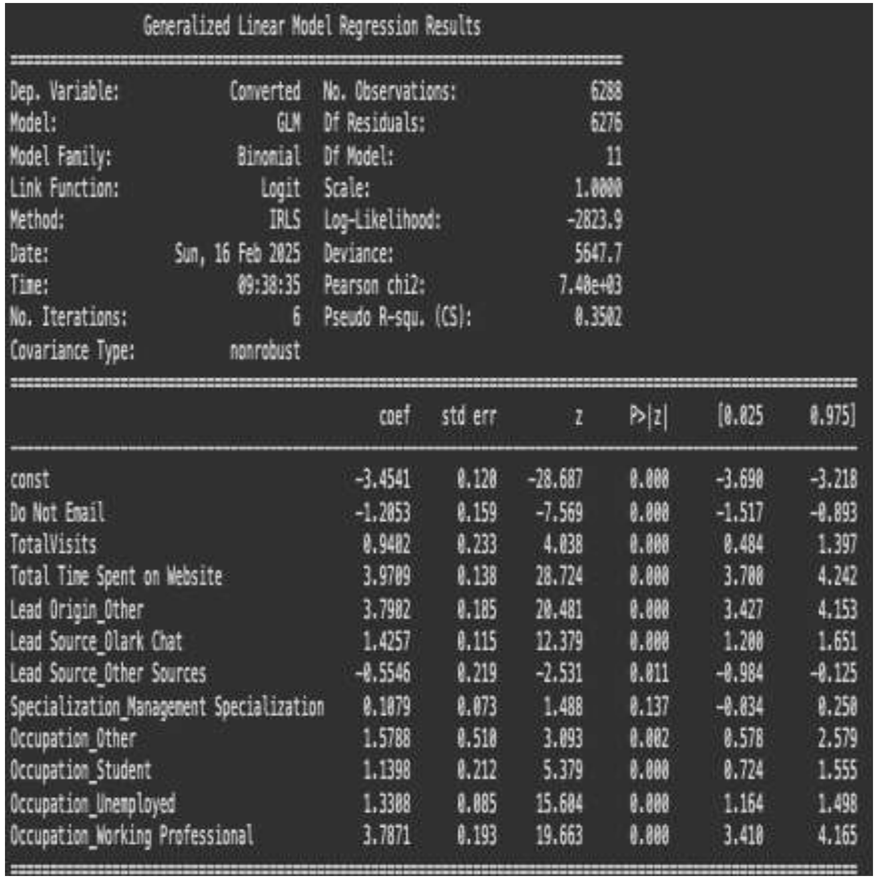
Observations:

•Even after dropping *Specialization_Management Specialization*, there is no significant change in the overall model accuracy.

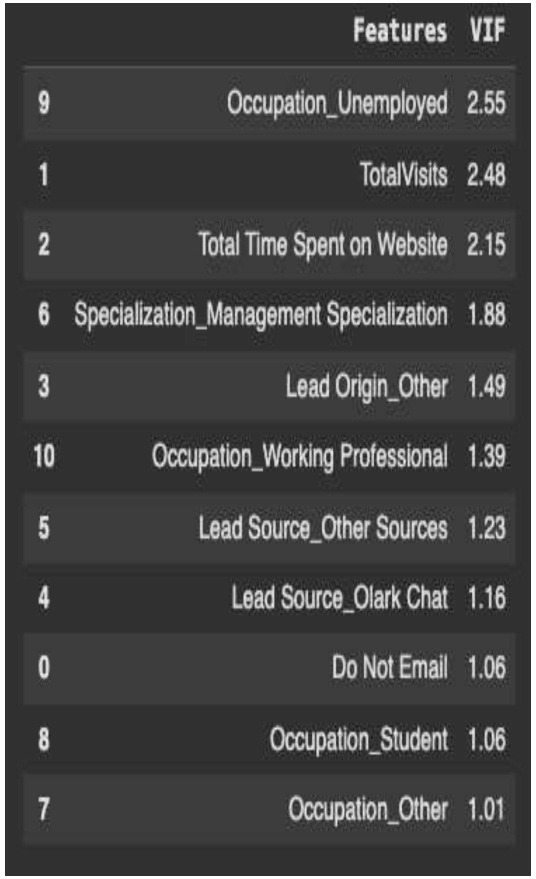
•However, the beta coefficient of *Specialization_Management Specialization* now exhibits a higher p-value, indicating it is not statistically significant.

•Multicollinearity is no longer present in Model 6.

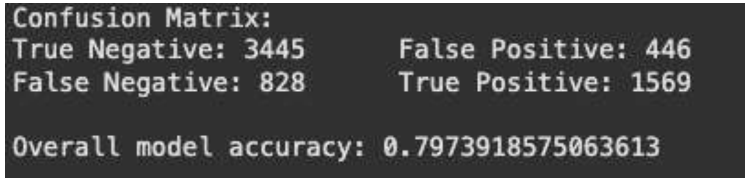
•As a result, *Specialization_Management Specialization* was dropped in the next model iteration due to its lack of statistical significance.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6288
Model:                            GLM   Df Residuals:                     6276
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2823.9
Date:                Sun, 16 Feb 2025   Deviance:                       5647.7
Time:                        09:38:35   Pearson chi2:                 7.40e+03
No. Iterations:                     6   Pseudo R-squ. (CS):             0.3502
Covariance Type:            nonrobust
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                       -3.4541      0.120    -28.687      0.000      -3.690      -3.218
Do Not Email                                -1.2053      0.159     -7.569      0.000      -1.517      -0.893
TotalVisits                                  0.9402      0.233      4.038      0.000       0.484       1.397
Total Time Spent on Website                  3.9709      0.138     28.724      0.000       3.700       4.242
Lead Origin_Other                            3.7902      0.185     20.481      0.000       3.427       4.153
Lead Source_Olark Chat                       1.4257      0.115     12.379      0.000       1.200       1.651
Lead Source_Other Sources                   -0.5546      0.219     -2.531      0.011      -0.984      -0.125
Specialization_Management Specialization     0.1079      0.073      1.488      0.137      -0.034       0.250
Occupation_Other                             1.5788      0.510      3.093      0.002       0.578       2.579
Occupation_Student                           1.1398      0.212      5.379      0.000       0.724       1.555
Occupation_Unemployed                        1.3308      0.085     15.604      0.000       1.164       1.498
Occupation_Working Professional              3.7871      0.193     19.663      0.000       3.410       4.165
============================================================================================================
```



	Features	VIF
9	Occupation_Unemployed	2.55
1	TotalVisits	2.48
2	Total Time Spent on Website	2.15
6	Specialization_Management Specialization	1.88
3	Lead Origin_Other	1.49
10	Occupation_Working Professional	1.39
5	Lead Source_Other Sources	1.23
4	Lead Source_Olark Chat	1.16
0	Do Not Email	1.06
8	Occupation_Student	1.06
7	Occupation_Other	1.01



```
Confusion Matrix:
True Negative: 3445      False Positive: 446
False Negative: 828      True Positive: 1569

Overall model accuracy: 0.7973918575063613
```

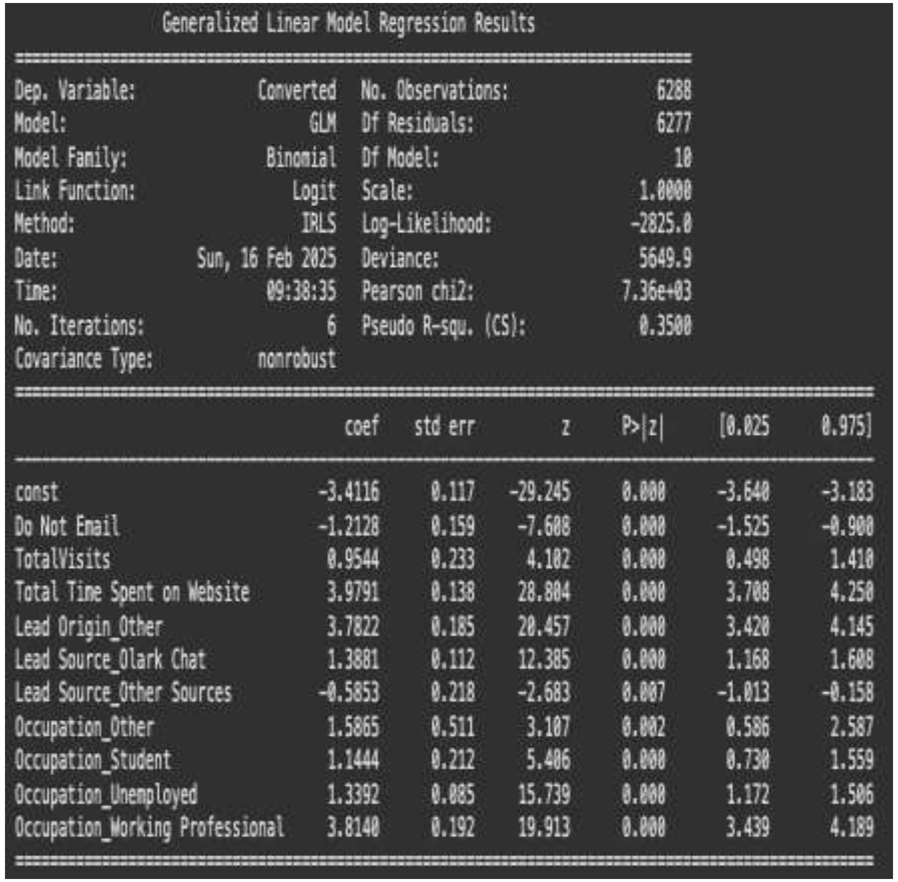
Model Building : Model 7

Observations:

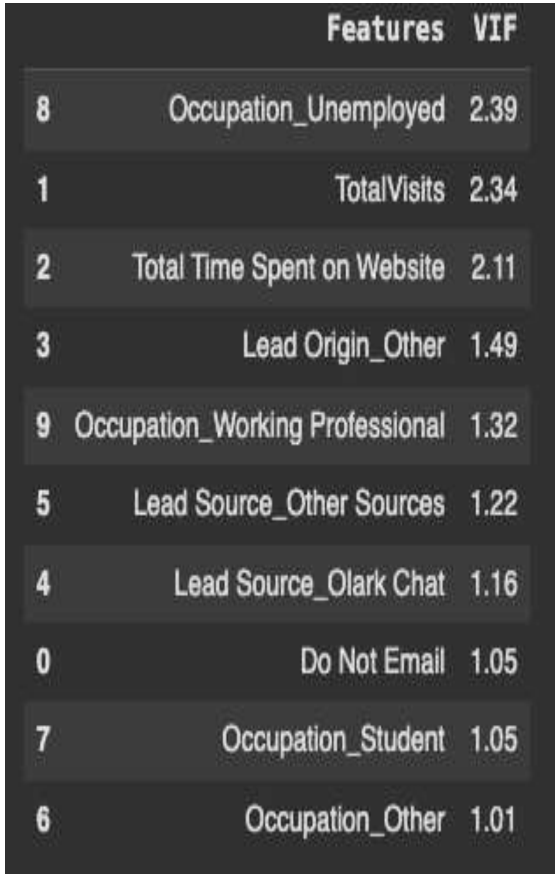
After droping 'Specialization_Management Specialization' our model accuracy has been improved a bit.

After droping 'Specialization_Management Specialization', we can see that all the beta coefficients are now

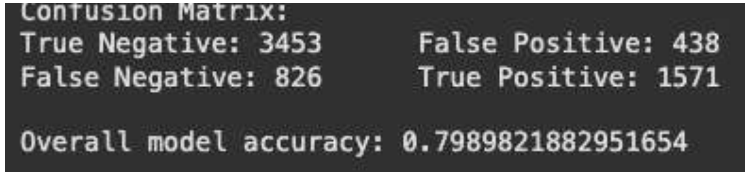
statistically significant also there is no multicolinearity present in Model 7.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            Converted   No. Observations:                 6288
Model:                          GLM   Df Residuals:                     6277
Model Family:              Binomial   Df Model:                           10
Link Function:                Logit   Scale:                          1.0000
Method:                        IRLS   Log-Likelihood:                -2825.0
Date:              Sun, 16 Feb 2025   Deviance:                       5649.9
Time:                      09:38:35   Pearson chi2:                 7.36e+03
No. Iterations:                   6   Pseudo R-squ. (CS):             0.3500
Covariance Type:          nonrobust
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                   -3.4116      0.117    -29.245      0.000      -3.640      -3.183
Do Not Email                            -1.2128      0.159     -7.608      0.000      -1.525      -0.900
TotalVisits                              0.9544      0.233      4.102      0.000       0.498       1.410
Total Time Spent on Website              3.9791      0.138     28.804      0.000       3.708       4.250
Lead Origin_Other                        3.7822      0.185     20.457      0.000       3.420       4.145
Lead Source_Olark Chat                   1.3881      0.112     12.385      0.000       1.168       1.608
Lead Source_Other Sources               -0.5853      0.218     -2.683      0.007      -1.013      -0.158
Occupation_Other                         1.5865      0.511      3.107      0.002       0.586       2.587
Occupation_Student                       1.1444      0.212      5.406      0.000       0.730       1.559
Occupation_Unemployed                    1.3392      0.085     15.739      0.000       1.172       1.506
Occupation_Working Professional          3.8140      0.192     19.913      0.000       3.439       4.189
========================================================================================================
```



	Features	VIF
8	Occupation_Unemployed	2.39
1	TotalVisits	2.34
2	Total Time Spent on Website	2.11
3	Lead Origin_Other	1.49
9	Occupation_Working Professional	1.32
5	Lead Source_Other Sources	1.22
4	Lead Source_Olark Chat	1.16
0	Do Not Email	1.05
7	Occupation_Student	1.05
6	Occupation_Other	1.01



```
Confusion Matrix:
True Negative: 3453      False Positive: 438
False Negative: 826      True Positive: 1571

Overall model accuracy: 0.7989821882951654
```

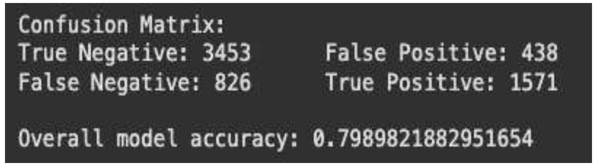
Prediction & Model Evaluation : Training Data Cutoff 0.5

Model Prediction and Evaluation:

•Prediction Approach:
Model 7 was used to predict the probability of conversion for all observations in the training dataset. A probability cut-off of 0.5 was applied, where:

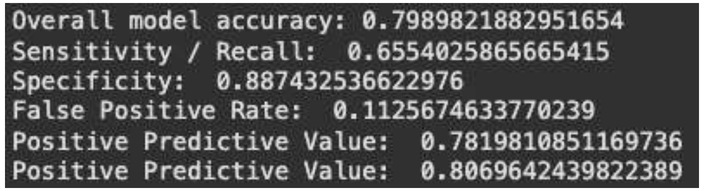
- *Probability > 0.5* was classified as *Converted = 1 (Yes)*.
- *Probability ≤ 0.5* was classified as *Converted = 0 (No)*.

•Model Evaluation**:**
After predicting the target variable on the training dataset, various evaluation metrics were calculated to assess model performance.



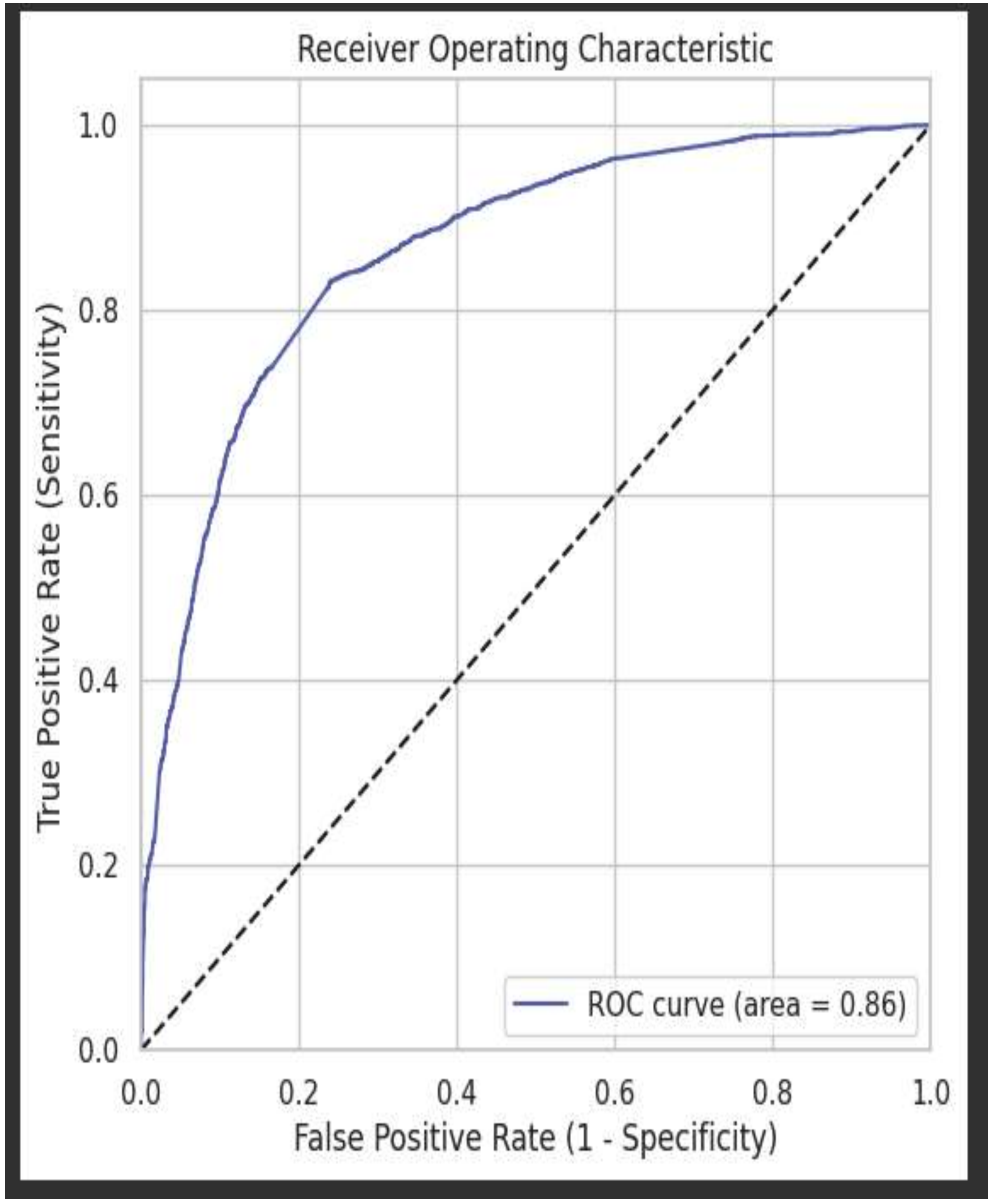
```
Confusion Matrix:
True Negative: 3453      False Positive: 438
False Negative: 826      True Positive: 1571

Overall model accuracy: 0.7989821882951654
```

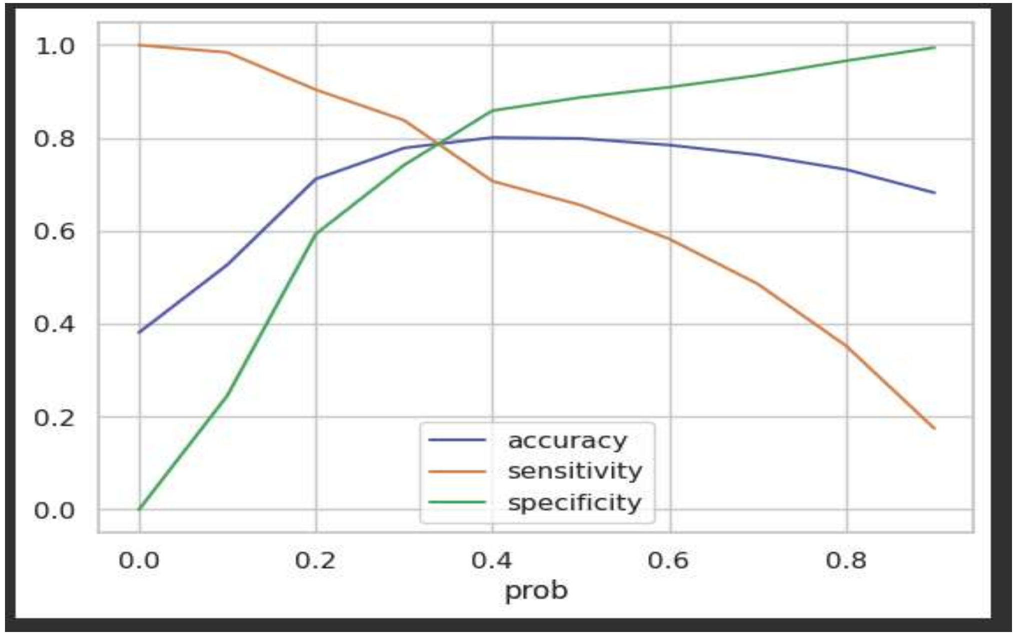


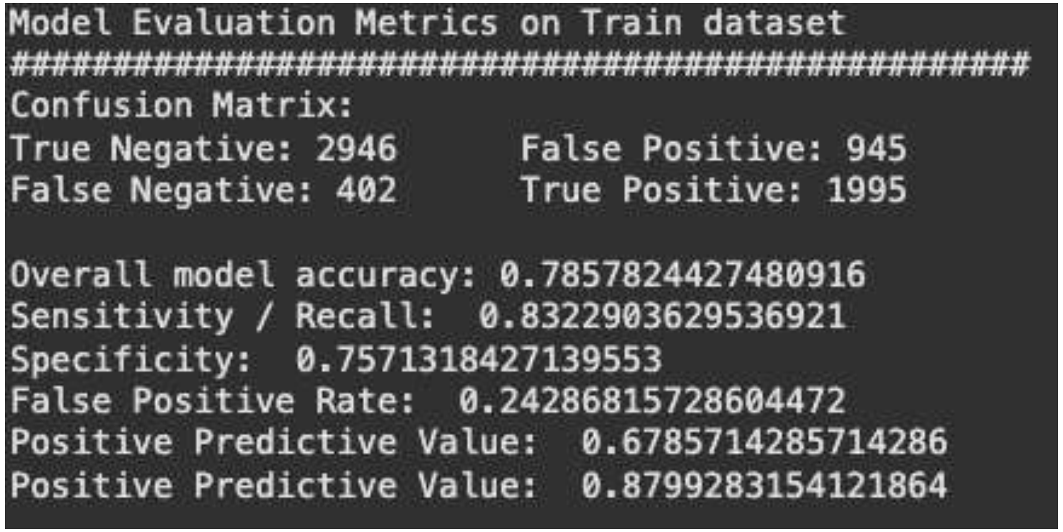
```
Overall model accuracy: 0.7989821882951654
Sensitivity / Recall:  0.6554025865665415
Specificity:  0.887432536622976
False Positive Rate:  0.1125674633770239
Positive Predictive Value:  0.7819810851169736
Positive Predictive Value:  0.8069642439822389
```

•Sensitivity-Specificity Trade-off:
The initial model demonstrated low sensitivity at the default 0.5 probability cut-off. To improve the model's performance, an optimal probability cut-off value was determined by performing a trade-off analysis between sensitivity and specificity.



Finding Optimal Probability Cutoff & Evaluating on Train Data





```
Model Evaluation Metrics on Train dataset
##################################################
Confusion Matrix:
True Negative: 2946      False Positive: 945
False Negative: 402      True Positive: 1995

Overall model accuracy: 0.7857824427480916
Sensitivity / Recall:  0.8322903629536921
Specificity:  0.7571318427139553
False Positive Rate:  0.24286815728604472
Positive Predictive Value:  0.6785714285714286
Positive Predictive Value:  0.8799283154121864
```

Observations:

•The sensitivity of the model has improved without any significant reduction in overall accuracy.

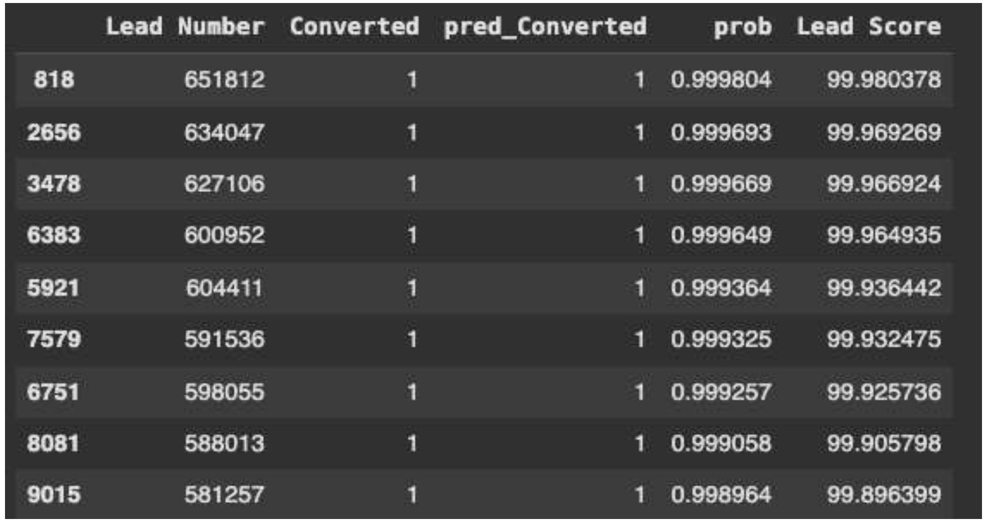
•The updated specificity also falls within an acceptable range, ensuring a well-balanced model performance.

Prediction & Generating Lead Score (Business Requirement)

Using Model 7, we calculated the probability on the test dataset and applied a cutoff value of 0.32 to predict the *pred_Converted* (0 or 1).

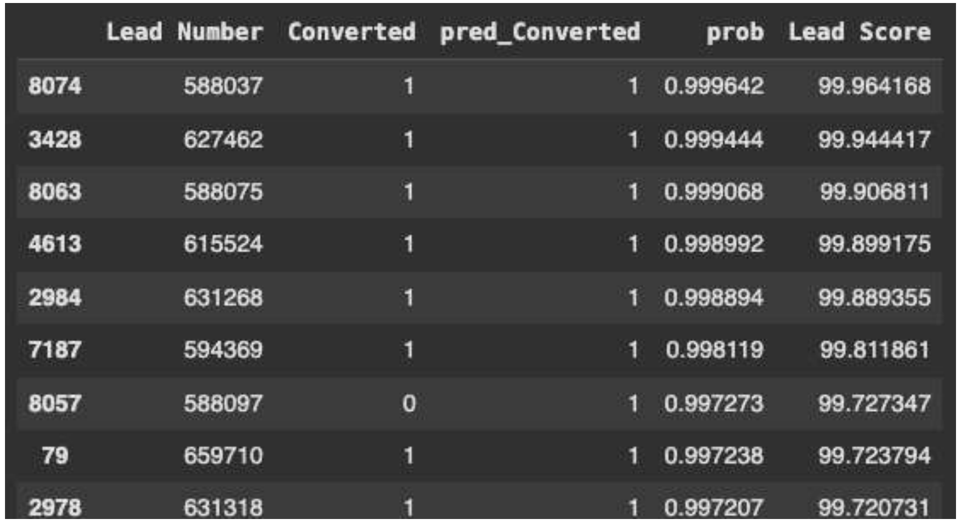
In line with business requirements, we created a *Lead Score* column (ranging from 0 to 100) to represent the likelihood of lead conversion. A higher score indicates a *hot lead* (most likely to convert), while a lower score signifies a *cold lead* (less likely to convert). The *Lead Score* was generated by multiplying the predicted probability (*pred_Converted*) by 100.

Train Dataset



	Lead Number	Converted	pred_Converted	prob	Lead Score
818	651812	1	1	0.999804	99.980378
2656	634047	1	1	0.999693	99.969269
3478	627106	1	1	0.999669	99.966924
6383	600952	1	1	0.999649	99.964935
5921	604411	1	1	0.999364	99.936442
7579	591536	1	1	0.999325	99.932475
6751	598055	1	1	0.999257	99.925736
8081	588013	1	1	0.999058	99.905798
9015	581257	1	1	0.998964	99.896399

Test Dataset



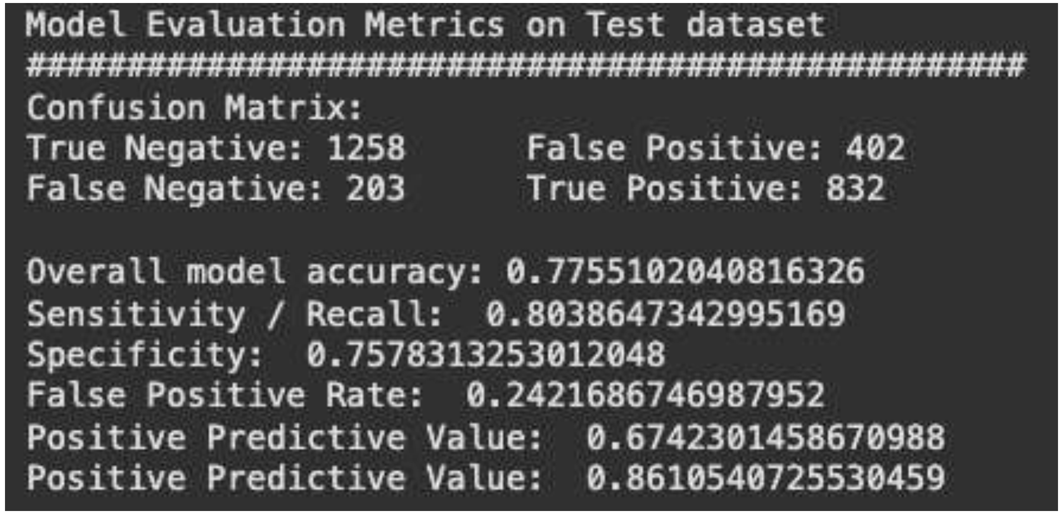
	Lead Number	Converted	pred_Converted	prob	Lead Score
8074	588037	1	1	0.999642	99.964168
3428	627462	1	1	0.999444	99.944417
8063	588075	1	1	0.999068	99.906811
4613	615524	1	1	0.998992	99.899175
2984	631268	1	1	0.998894	99.889355
7187	594369	1	1	0.998119	99.811861
8057	588097	0	1	0.997273	99.727347
79	659710	1	1	0.997238	99.723794
2978	631318	1	1	0.997207	99.720731

Model Evaluation : Test data

The model performed well on the test data, achieving a sensitivity of 80%, specificity of 76%, and an overall accuracy of 78%.

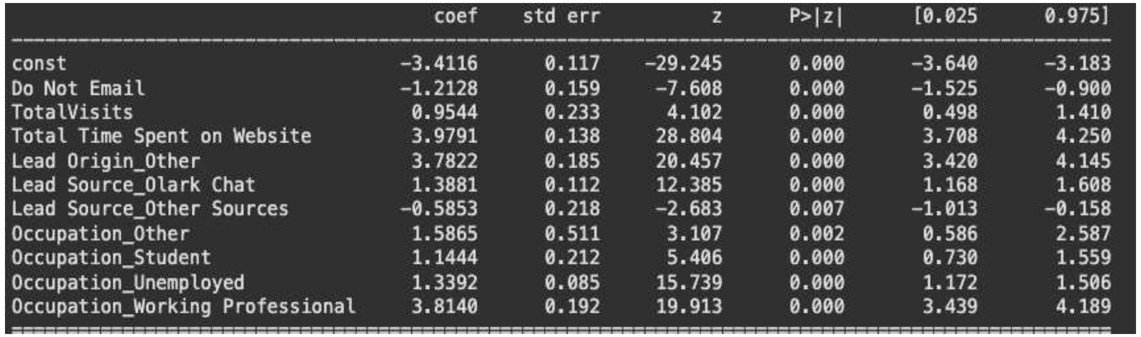
The top three variables contributing most to the probability of a lead getting converted are:

1. Total Time Spent on Website,
2. Current Occupation (Working Professional),
3. Lead Origin (Other).



```
Model Evaluation Metrics on Test dataset
##################################################
Confusion Matrix:
True Negative: 1258      False Positive: 402
False Negative: 203      True Positive: 832

Overall model accuracy: 0.7755102040816326
Sensitivity / Recall:  0.8038647342995169
Specificity:  0.7578313253012048
False Positive Rate:  0.2421686746987952
Positive Predictive Value:  0.6742301458670988
Positive Predictive Value:  0.8610540725530459
```



```
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                         -3.4116      0.117    -29.245      0.000      -3.640      -3.183
Do Not Email                  -1.2128      0.159     -7.608      0.000      -1.525      -0.900
TotalVisits                    0.9544      0.233      4.102      0.000       0.498       1.410
Total Time Spent on Website    3.9791      0.138     28.804      0.000       3.708       4.250
Lead Origin_Other              3.7822      0.185     20.457      0.000       3.420       4.145
Lead Source_Olark Chat         1.3881      0.112     12.385      0.000       1.168       1.608
Lead Source_Other Sources     -0.5853      0.218     -2.683      0.007      -1.013      -0.158
Occupation_Other               1.5865      0.511      3.107      0.002       0.586       2.587
Occupation_Student             1.1444      0.212      5.406      0.000       0.730       1.559
Occupation_Unemployed          1.3392      0.085     15.739      0.000       1.172       1.506
Occupation_Working Professional 3.8140     0.192     19.913      0.000       3.439       4.189
==============================================================================================
```

Conclusion and Recommendations

- A Logistic Regression model was used to calculate the Lead Score (ranging from 0 to 100) for each lead, where a higher score indicates a hot lead with a high likelihood of conversion, and a lower score indicates a cold lead with a lower probability of conversion.
- Sorting leads in descending order based on their Lead Scores enables faster and more efficient identification of hot leads, reducing conversion time and increasing conversion rates.
- Priority should be given to contacting leads with higher scores first, with special attention such as assigning a dedicated support SPOC for a small batch of high-scoring leads to enhance conversion chances.
- Medium-scoring leads also have good potential for conversion and should be contacted to understand their needs and address any concerns, such as modifying existing courses, introducing new courses, adjusting class schedules, or providing flexible financial options.
- Cold leads, with lower conversion chances, can be targeted later as part of an aggressive marketing strategy once high and medium-scoring leads are successfully converted.