LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

Param Nagarsheth

CONTENTS

- Problem Statement
- Problem Approach
- Data Understanding and Cleaning
- EDA
- Model Approach
- Model Evaluation
- Observations
- Conclusion

PROBLEM STATEMENT

X Education, an online education company, wants to improve its **lead conversion rate,** which currently stands at around 30%. The goal is to identify **‘Hot Leads’**—potential customers who are most likely to convert into paying customers—so that the sales team can focus their efforts on these leads and achieve a target conversion rate of **80%.**

The company has provided a dataset of around 9,000 leads with various attributes like **Lead Source, Total Time Spent on Website, Last Activity**, etc. The task is to build a **logistic regression model** that assigns a **lead score (0–100)** to each lead. Higher scores indicate higher conversion probabilities.

The deliverables include:

1. **A predictive logistic regression model** to prioritize leads.
2. **Actionable insights and recommendations** for improving lead conversion efficiency.
3. A flexible approach to handle future business requirements and challenges.

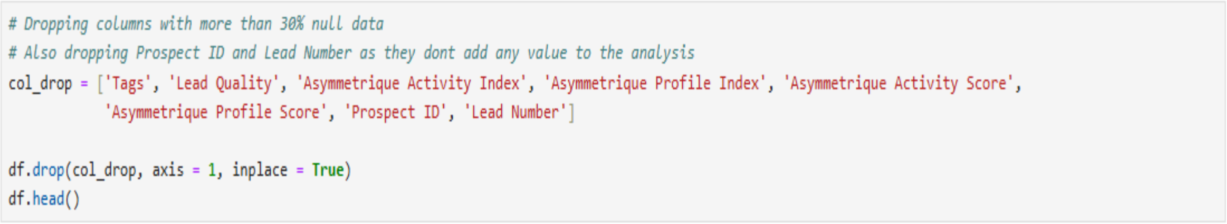
The ultimate objective is to optimize the sales funnel by focusing resources on the most promising leads, ensuring a higher ROI and achieving the company's desired growth targets.

PROBLEM APPROACH

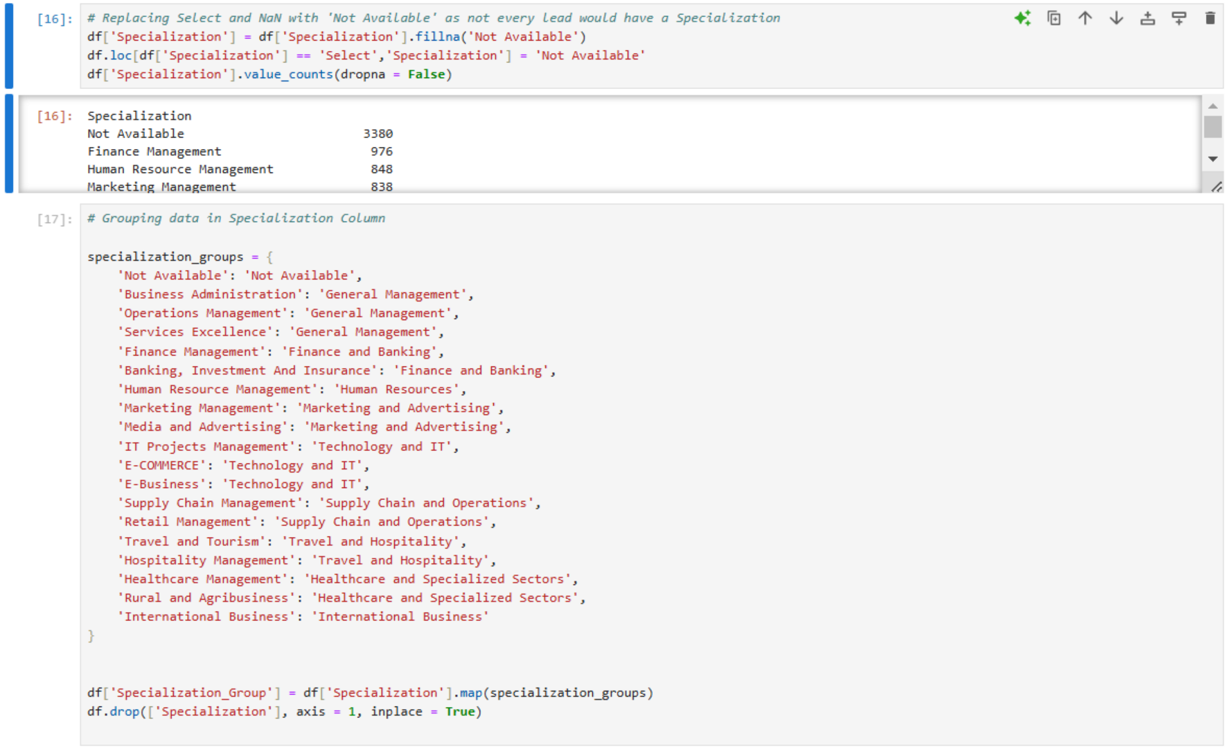
- Data Understanding: Analyzed the dataset of ~9,000 leads, identifying key variables impacting lead conversion.
- Data Cleaning: Handled missing values, removed irrelevant levels (e.g., "Select"), and treated categorical variables.
- Exploratory Data Analysis (EDA): Identified trends and key factors influencing lead conversion.
- Feature Engineering: Created dummy variables for categorical features and scaled numerical data.
- Model Building: Developed a logistic regression model to predict lead conversion probabilities.
- Model Evaluation: Evaluated performance using metrics like ROC-AUC, sensitivity, specificity, precision, and recall to ensure accuracy.

DATA UNDERSTANDING AND CLEANING

- Dropped columns with more than 30% empty rows.
- Dropped columns ‘Country’ and ‘City’ as there is a huge data imbalance.
- Imputing the Specialization column with ‘Not Available’ and grouping the data.
- Dropping empty rows for the remaining columns.



```
# Dropping columns with more than 30% null data
# Also dropping Prospect ID and Lead Number as they dont add any value to the analysis
col_drop = ['Tags', 'Lead Quality', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score',
            'Asymmetrique Profile Score', 'Prospect ID', 'Lead Number']

df.drop(col_drop, axis = 1, inplace = True)
df.head()
```



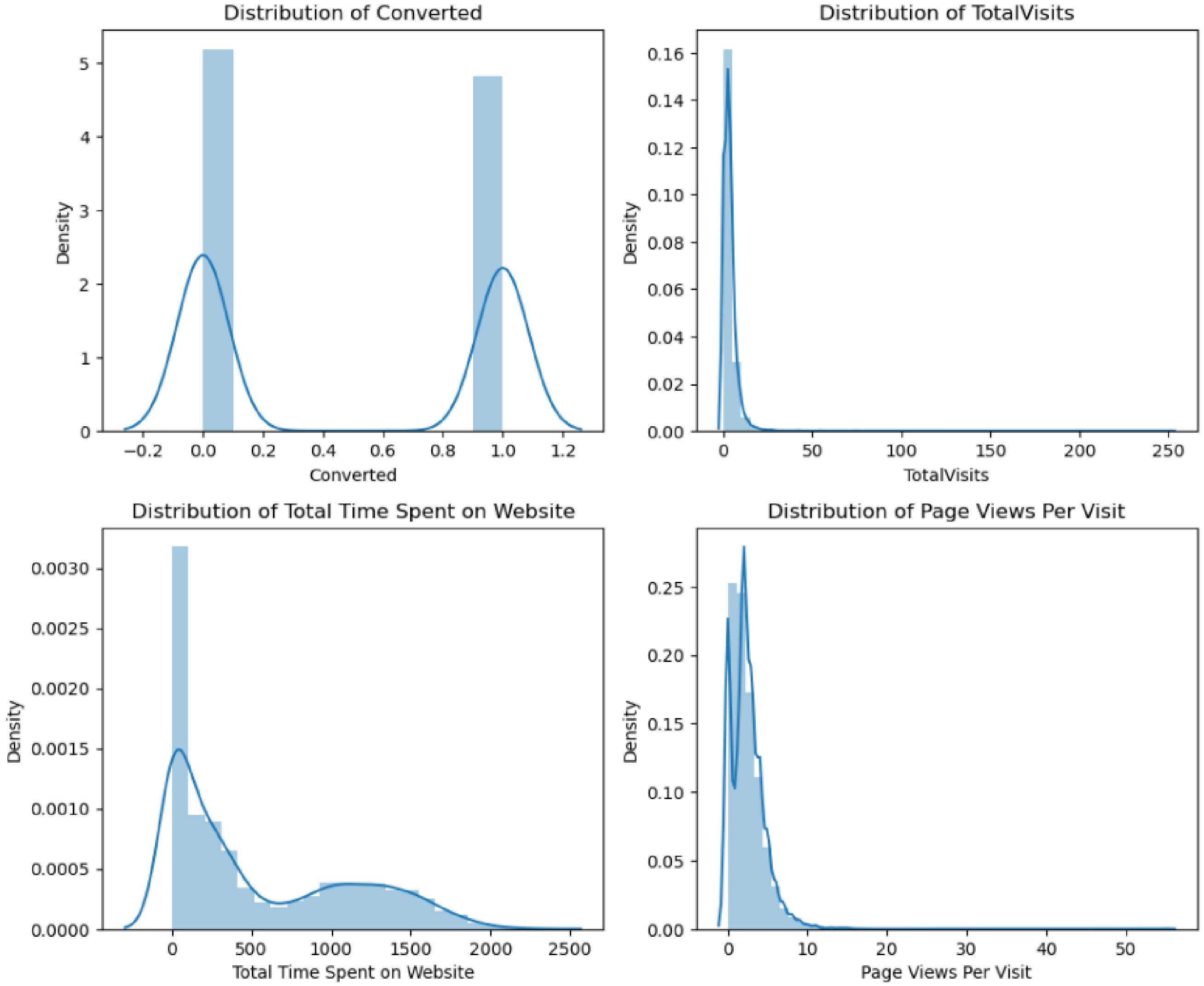
```
[16]: # Replacing Select and NaN with 'Not Available' as not every lead would have a Specialization
      df['Specialization'] = df['Specialization'].fillna('Not Available')
      df.loc[df['Specialization'] == 'Select','Specialization'] = 'Not Available'
      df['Specialization'].value_counts(dropna = False)
```

```
[16]: Specialization
      Not Available                 3380
      Finance Management             976
      Human Resource Management      848
      Marketing Management           838
```

```
[17]: # Grouping data in Specialization Column

      specialization_groups = {
          'Not Available': 'Not Available',
          'Business Administration': 'General Management',
          'Operations Management': 'General Management',
          'Services Excellence': 'General Management',
          'Finance Management': 'Finance and Banking',
          'Banking, Investment And Insurance': 'Finance and Banking',
          'Human Resource Management': 'Human Resources',
          'Marketing Management': 'Marketing and Advertising',
          'Media and Advertising': 'Marketing and Advertising',
          'IT Projects Management': 'Technology and IT',
          'E-COMMERCE': 'Technology and IT',
          'E-Business': 'Technology and IT',
          'Supply Chain Management': 'Supply Chain and Operations',
          'Retail Management': 'Supply Chain and Operations',
          'Travel and Tourism': 'Travel and Hospitality',
          'Hospitality Management': 'Travel and Hospitality',
          'Healthcare Management': 'Healthcare and Specialized Sectors',
          'Rural and Agribusiness': 'Healthcare and Specialized Sectors',
          'International Business': 'International Business'
      }


      df['Specialization_Group'] = df['Specialization'].map(specialization_groups)
      df.drop(['Specialization'], axis = 1, inplace = True)
```

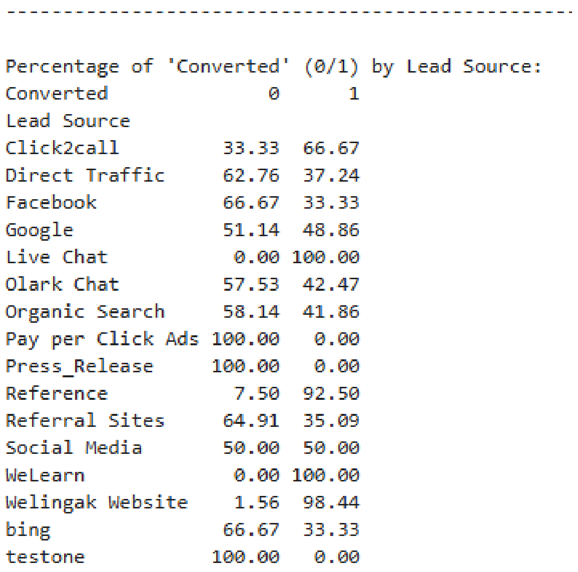
EDA – NUMERICAL DATA

- There is no data imbalance. The amount of converted and non-converted leads is similar.
- TotalVisits and Page Views Per Visit have upper limit outliers which was fixed by dropping rows with values greater than the 95th percentiles.

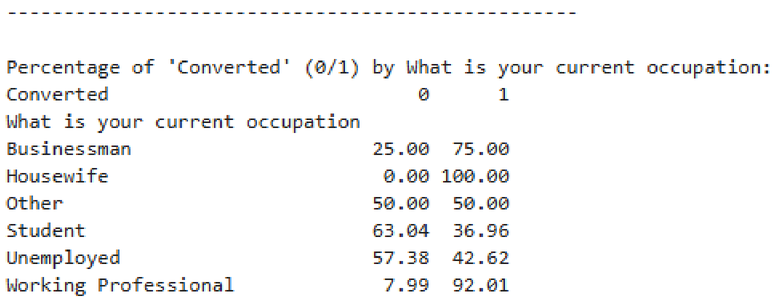


EDA – CATEGORICAL DATA

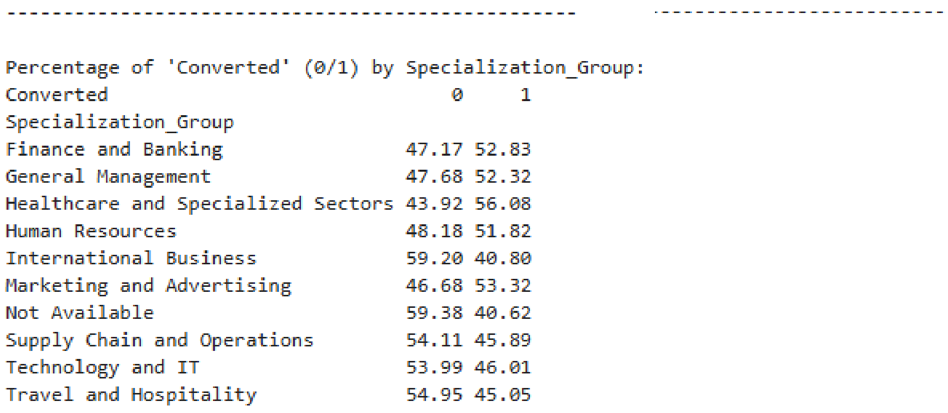
- Google, Reference and Welingak Website leads have a higher conversion rate.
- Leads of Working Professionals and Unemployed have a higher conversion rate.
- Leads with (Finance and Banking, General Management, Healthcare and Specialized Sectors, Human Resources, Marketing and Advertising) specialisation groups have a higher conversion rate.



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

Percentage of 'Converted' (0/1) by Lead Source:
Converted              0      1
Lead Source
Click2call         33.33  66.67
Direct Traffic     62.76  37.24
Facebook           66.67  33.33
Google             51.14  48.86
Live Chat           0.00 100.00
Olark Chat         57.53  42.47
Organic Search     58.14  41.86
Pay per Click Ads 100.00   0.00
Press_Release     100.00   0.00
Reference           7.50  92.50
Referral Sites     64.91  35.09
Social Media       50.00  50.00
WeLearn             0.00 100.00
Welingak Website    1.56  98.44
bing               66.67  33.33
testone           100.00   0.00
```

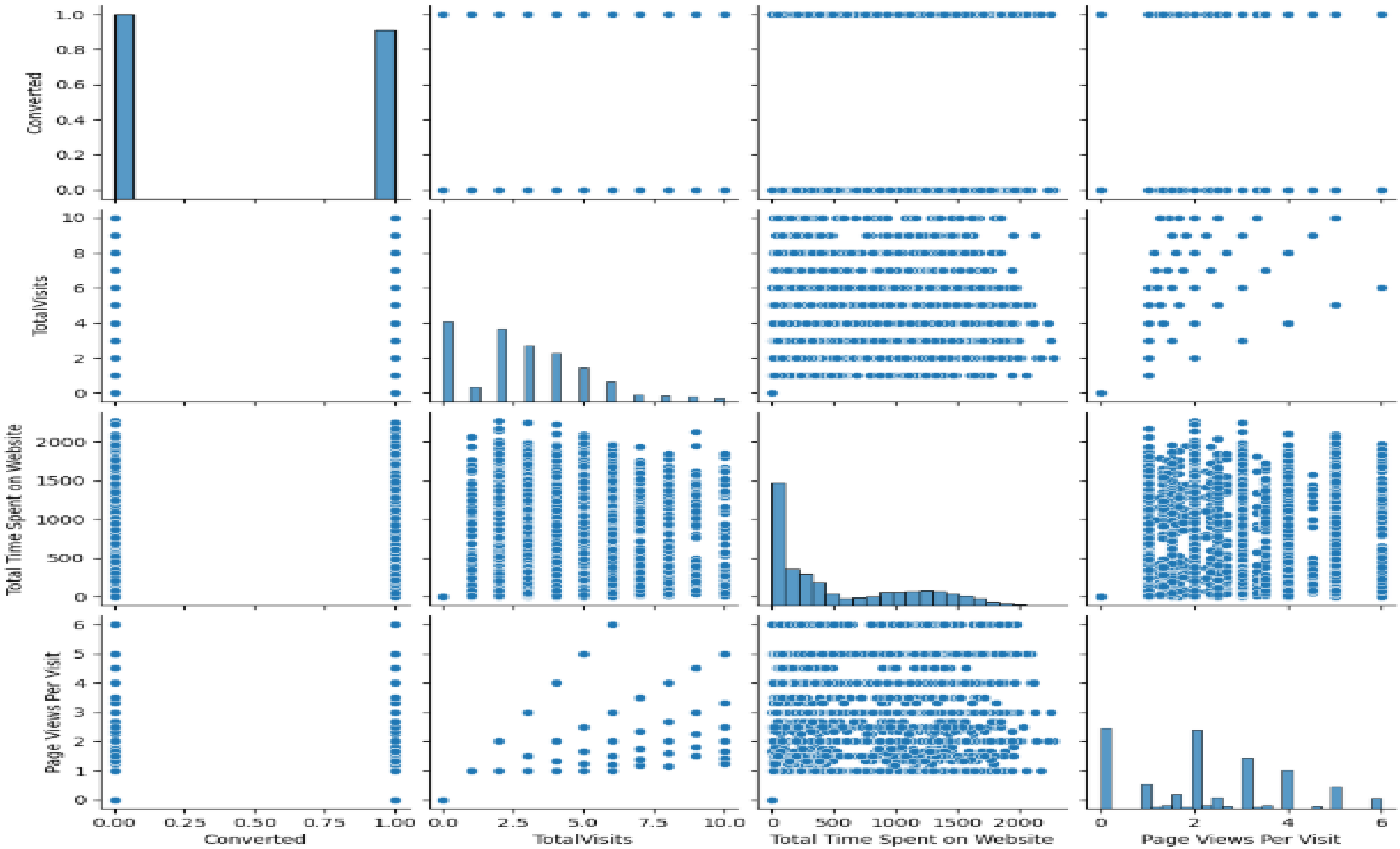


```
--------------------------------------------------

Percentage of 'Converted' (0/1) by What is your current occupation:
Converted                                   0      1
What is your current occupation
Businessman                             25.00  75.00
Housewife                                0.00 100.00
Other                                   50.00  50.00
Student                                 63.04  36.96
Unemployed                              57.38  42.62
Working Professional                     7.99  92.01
```



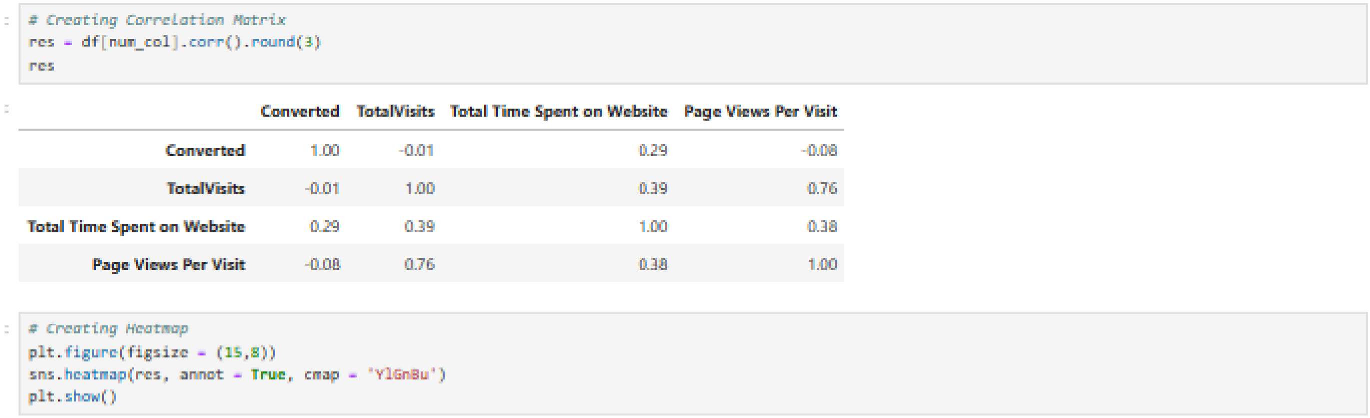
```
--------------------------------------------------     -------------------------

Percentage of 'Converted' (0/1) by Specialization_Group:
Converted                              0     1
Specialization_Group
Finance and Banking                47.17 52.83
General Management                 47.68 52.32
Healthcare and Specialized Sectors 43.92 56.08
Human Resources                    48.18 51.82
International Business             59.20 40.80
Marketing and Advertising          46.68 53.32
Not Available                      59.38 40.62
Supply Chain and Operations        54.11 45.89
Technology and IT                  53.99 46.01
Travel and Hospitality             54.95 45.05
```

EDA – SCATTER PLOT



EDA – CORRELATION

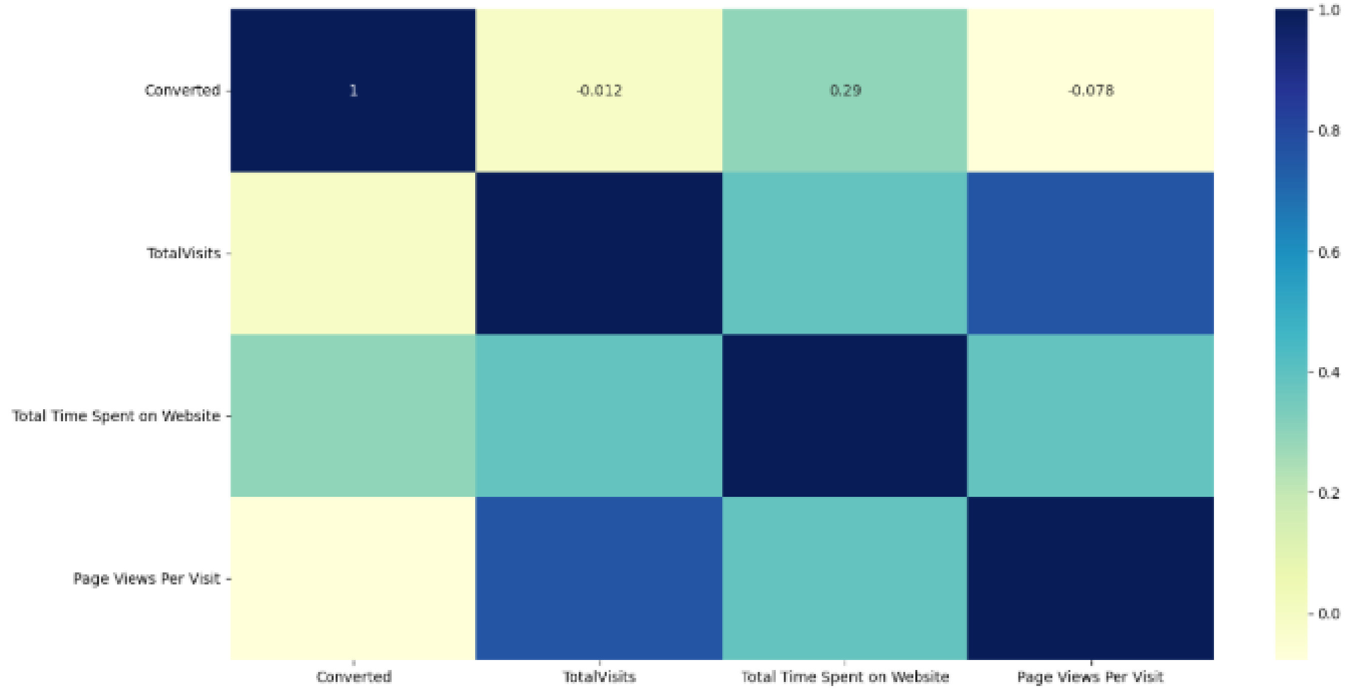
- There is no high correlation with the target variable – 'Converted'
- 'TotalVisits' and 'Page Views Per Visit' are highly correlated



```
# Creating Correlation Matrix
res = df[num_col].corr().round(3)
res
```

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1.00	-0.01	0.29	-0.08
TotalVisits	-0.01	1.00	0.39	0.76
Total Time Spent on Website	0.29	0.39	1.00	0.38
Page Views Per Visit	-0.08	0.76	0.38	1.00

```
# Creating Heatmap
plt.figure(figsize = (15,8))
sns.heatmap(res, annot = True, cmap = 'YlGnBu')
plt.show()
```

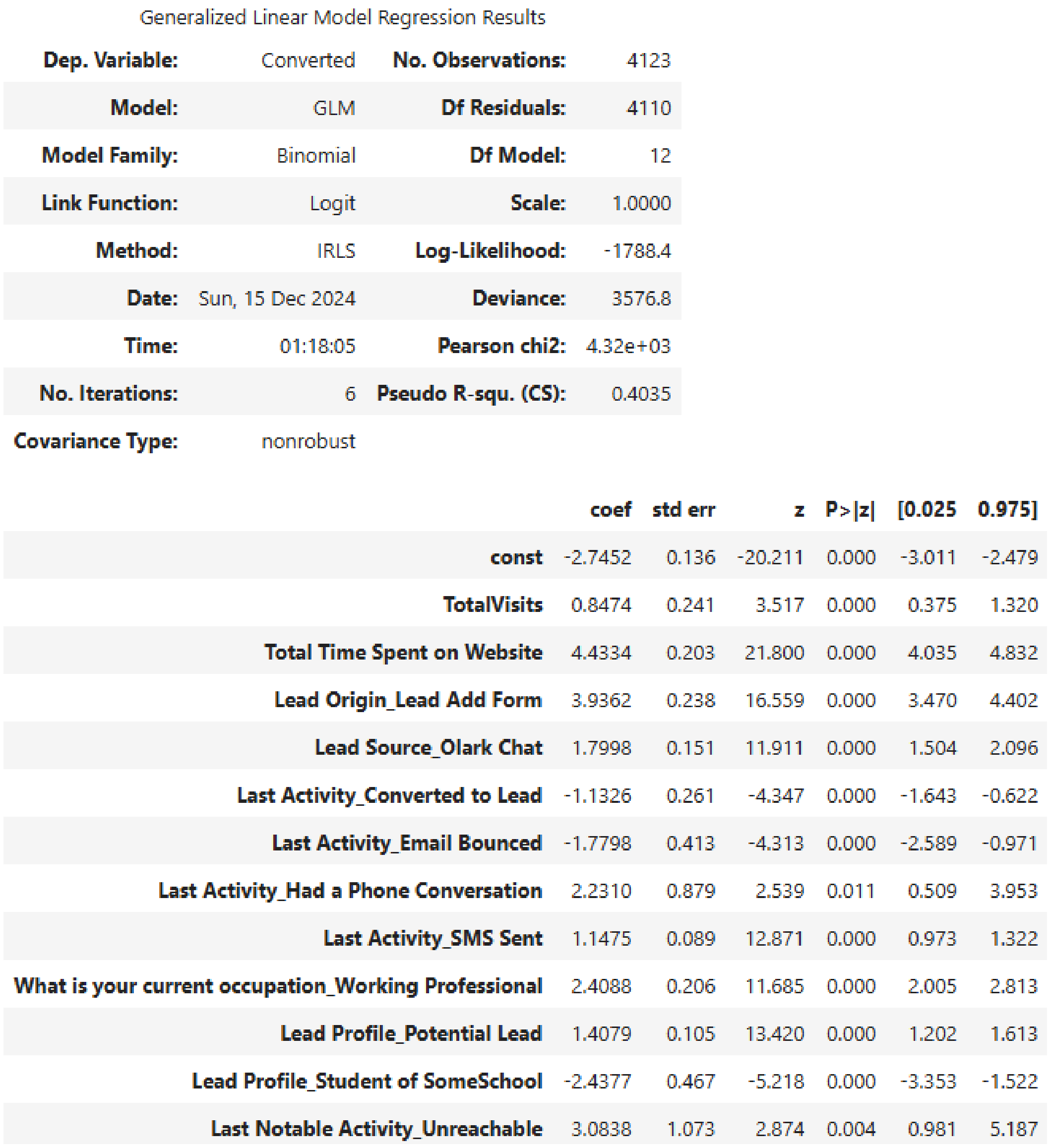


MODEL APPROACH

- The first step involved preparing the data by converting categorical variables into dummy variables.
Next, the data was split into training and testing sets using a 70-30 split ratio.
Numerical variables were then scaled using the **MinMaxScaler.**

- The model-building process began with Recursive Feature Elimination (RFE), where the top 15 features were initially selected. Features were subsequently eliminated based on their p-values and VIF values, leading to **Model-4**, the final model.

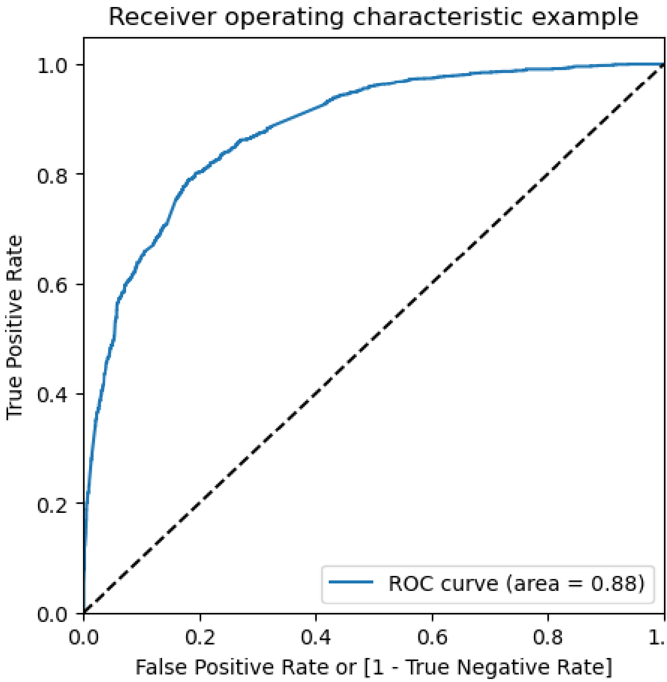
- Predictions were made on the training set using an initial cut-off of 0.5, and the model's performance was evaluated in terms of accuracy, sensitivity, and specificity:
  - **Sensitivity**: 0.78
  - **Specificity**: 0.82
  - **Accuracy**: 0.80

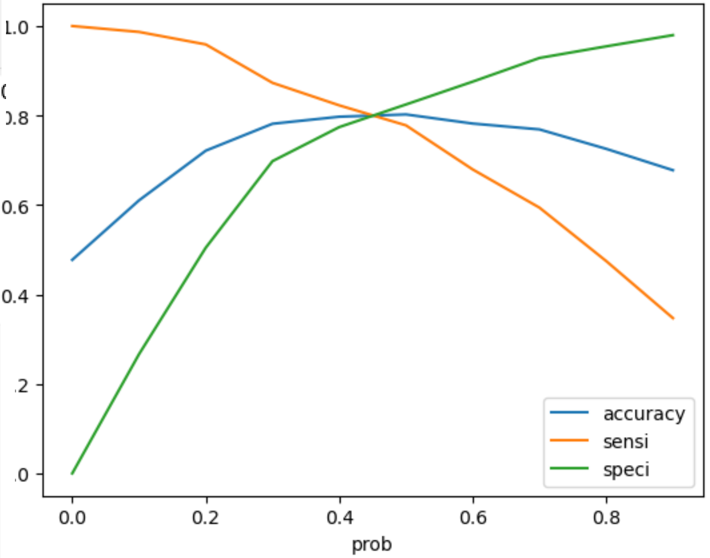


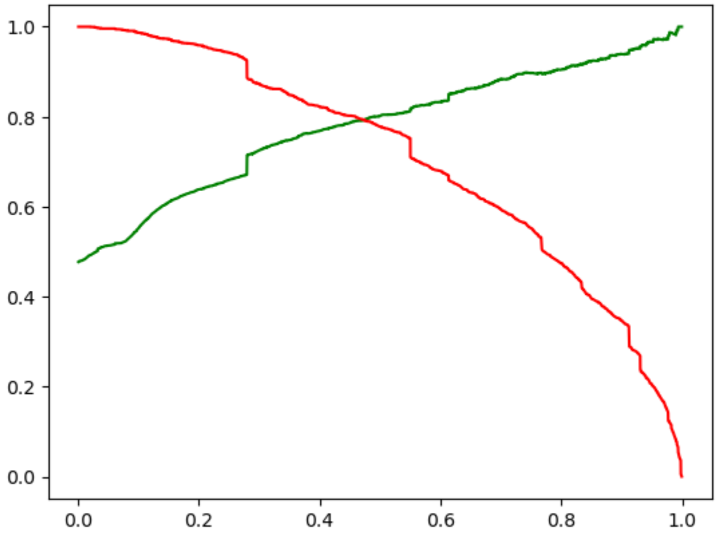
```
                 Generalized Linear Model Regression Results
   Dep. Variable:          Converted      No. Observations:      4123
           Model:                GLM          Df Residuals:      4110
    Model Family:           Binomial              Df Model:        12
   Link Function:              Logit                 Scale:    1.0000
          Method:               IRLS        Log-Likelihood:   -1788.4
            Date:   Sun, 15 Dec 2024              Deviance:    3576.8
            Time:           01:18:05         Pearson chi2:   4.32e+03
  No. Iterations:                  6  Pseudo R-squ. (CS):      0.4035
 Covariance Type:          nonrobust

                                                    coef  std err        z  P>|z|  [0.025  0.975]
                                        const    -2.7452    0.136  -20.211  0.000  -3.011  -2.479
                                  TotalVisits     0.8474    0.241    3.517  0.000   0.375   1.320
                    Total Time Spent on Website   4.4334    0.203   21.800  0.000   4.035   4.832
                      Lead Origin_Lead Add Form   3.9362    0.238   16.559  0.000   3.470   4.402
                       Lead Source_Olark Chat     1.7998    0.151   11.911  0.000   1.504   2.096
                 Last Activity_Converted to Lead  -1.1326   0.261   -4.347  0.000  -1.643  -0.622
                    Last Activity_Email Bounced   -1.7798   0.413   -4.313  0.000  -2.589  -0.971
         Last Activity_Had a Phone Conversation   2.2310   0.879    2.539  0.011   0.509   3.953
                         Last Activity_SMS Sent   1.1475   0.089   12.871  0.000   0.973   1.322
What is your current occupation_Working Professional  2.4088  0.206  11.685  0.000  2.005  2.813
                    Lead Profile_Potential Lead   1.4079    0.105   13.420  0.000   1.202   1.613
            Lead Profile_Student of SomeSchool   -2.4377    0.467   -5.218  0.000  -3.353  -1.522
             Last Notable Activity_Unreachable    3.0838    1.073    2.874  0.004   0.981   5.187
```

MODEL APPROACH

- The **ROC curve** and the plot of accuracy, sensitivity, and specificity were used to determine the optimal cut-off value, which was found to be **0.4**.
- Using this new cut-off value, the model's performance was re-evaluated, yielding the following results:
  - **Sensitivity**: 0.82
  - **Specificity**: 0.77
  - **Accuracy**: 0.797
- The model was also double-checked using Precision and Recall
  - **Recall**: 0.82
  - **Precision**: 0.77

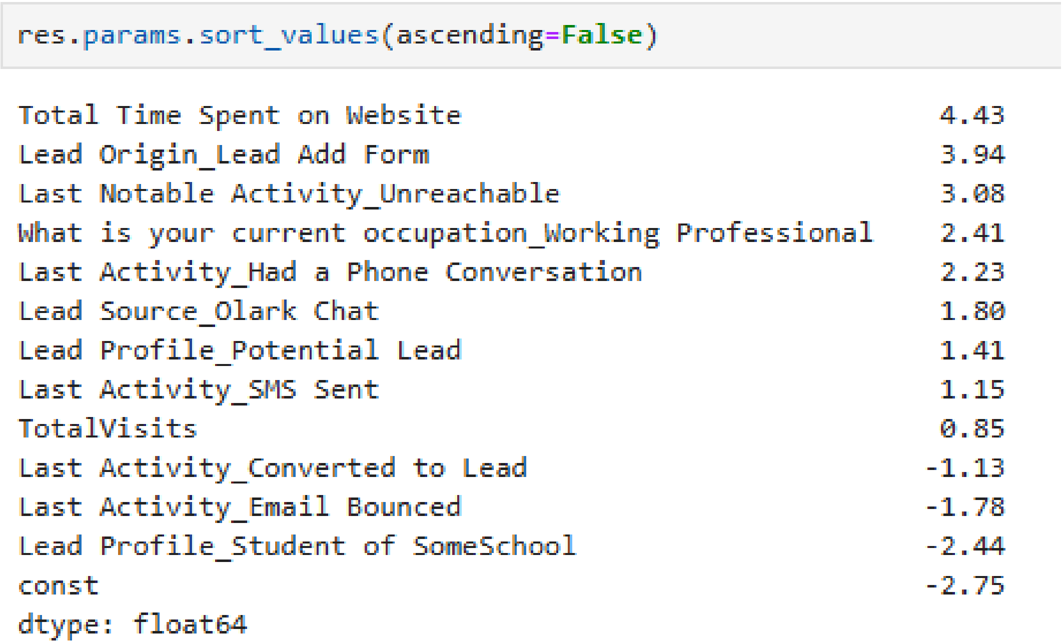






MODEL EVALUATION

- To evaluate the model, predictions were made on the test set using **Model-4** and the optimal cut-off value of **0.4**. The model's performance was assessed based on accuracy, sensitivity, specificity, precision, and recall:
  - **Accuracy**: 0.79
  - **Sensitivity**: 0.80
  - **Specificity**: 0.78
  - **Precision**: 0.77
  - **Recall**: 0.80



```
res.params.sort_values(ascending=False)
```

```
Total Time Spent on Website                              4.43
Lead Origin_Lead Add Form                                3.94
Last Notable Activity_Unreachable                        3.08
What is your current occupation_Working Professional     2.41
Last Activity_Had a Phone Conversation                   2.23
Lead Source_Olark Chat                                   1.80
Lead Profile_Potential Lead                              1.41
Last Activity_SMS Sent                                   1.15
TotalVisits                                              0.85
Last Activity_Converted to Lead                         -1.13
Last Activity_Email Bounced                             -1.78
Lead Profile_Student of SomeSchool                      -2.44
const                                                   -2.75
dtype: float64
```

OBSERVATION

1. **Total Time Spent on Website (4.43) - Leads spending more time on the website are significantly more likely to convert. This variable has the highest positive coefficient, indicating it strongly influences lead conversion.**

2. **Lead Origin: Lead Add Form (3.94) - Leads generated through the *Lead Add Form* are highly likely to convert, suggesting this lead origin is a critical channel for attracting high-quality leads.**

3. **Last Notable Activity: Unreachable (3.08) - Leads marked as *Unreachable* in their last activity still show strong conversion potential. It indicates that these leads may not respond immediately but are still valuable prospects.**

4. **Current Occupation: Working Professional (2.41) - Leads categorized as *Working Professionals* are more likely to convert compared to other occupations. This demographic should be a primary target group.**

5. **Last Activity: Had a Phone Conversation (2.23) - Leads who had a *Phone Conversation* in their last recorded activity are strong candidates for conversion, indicating that personal interaction boosts conversion likelihood.**

6. **Lead Source: Olark Chat (1.80) - Leads acquired via *Olark Chat* show higher conversion probabilities, highlighting this as a productive communication channel for engaging leads.**

OBSERVATION

1. **Lead Profile: Potential Lead (1.41) - Leads categorized as *Potential Leads* have a positive likelihood of conversion, supporting the reliability of the lead profiling process.**

2. **Last Activity: SMS Sent (1.15) - Sending an *SMS* as the last activity positively impacts lead conversion, demonstrating that timely follow-ups via SMS can yield favourable results.**

3. **Total Visits (0.85) - A higher number of visits to the website correlates positively with conversions, though the impact is relatively lower compared to *Time Spent on the Website*.**

4. **Last Activity: Converted to Lead (-1.13) - Leads marked as *Converted to Lead* in their last activity are less likely to convert. This may indicate these leads are further down the funnel and have stagnated.**

5. **Last Activity: Email Bounced (-1.78) - Leads whose *Email Bounced* are unlikely to convert. This indicates poor data quality or lack of engagement from these leads.**

6. **Lead Profile: Student of Some School (-2.44) - Leads categorized as *Students of Some Schools* have a significantly lower likelihood of conversion, suggesting this demographic is not a strong focus for sales efforts.**

7. **Intercept (Constant: -2.75)- The negative intercept indicates that without significant positive features, leads are unlikely to convert.**

BUSINESS RECOMMENDATIONS

Focus on High-Engagement Leads

- Prioritize leads spending **more time on the website** and those with **higher total visits.**
- Optimize website content and user experience to increase engagement.

Leverage Top-Performing Channels

- Invest in **Lead Add Form** and **Olark Chat** as key lead generation sources.
- Ensure chat availability and streamline form processes.

Target Working Professionals

- Tailor campaigns for **Working Professionals** with personalized messaging and career-focused content.

Enhance Follow-Up Strategies

- Prioritize leads with Phone Conversations and SMS interactions.
- Implement structured follow-ups via calls and SMS for high-scoring leads.

Reduce Effort on Low-Converting Segments

- Limit focus on Students of Some Schools and leads with Email Bounces.
- Improve data quality with email verification tools.

Adopt Lead Scoring

- Use the logistic regression model to assign lead scores and prioritize "hot leads" (cut-off $\geq$ 0.4).
- Automate lead routing for faster sales team engagement.

Monitor and Optimize

- Continuously track performance metrics and refine strategies based on lead behaviour insights.

CONCLUSION

By focusing on the top-performing variables such as *Total Time Spent on Website*, *Lead Add Form*, and *Working Professionals*, and streamlining resources toward high-scoring leads, X Education can significantly improve its lead conversion rate. Implementing these recommendations will help achieve the target conversion rate of 80% while optimizing the sales team's efforts for higher ROI.