# Lead Scoring Analysis:

## High-Potential Customers

A Logistic Regression Approach

Submitters: Usha V, Varshini R & Darshan V

# Table Of Contents

- Problem statement
- Project Goal
- Problem approach
- Data Processing
- EDA
- Model Building
- Model Evaluation
- Results
- Impact & Conclusion

# Problem Statement

- **X Education, which offers online courses for industry professionals**, is experiencing a significant challenge with its lead conversion process.

- Despite acquiring a substantial number of leads daily, **only about 30% of them are successfully converted** into customers.

- This inefficiency not only strains the sales team's resources but also **limits the company's growth potential**.

- To address this issue, X Education aims to **identify high-potential leads**, referred to as 'Hot Leads.' By focusing their sales efforts on these prioritized leads, **the company seeks to boost its lead conversion rate and optimize resource allocation effectively.**

# Project Goal

The project goal is to create a lead scoring system for Lead X that will:

- Assign a score from 0 to 100 to each lead.
- Facilitate the identification of high-potential leads.
- Support the company's objective of increasing their lead conversion rate to 80%.

# Problem Approach

To develop an effective lead scoring system, the approach focuses on:

1. Preparing data for analysis (cleaning, handling missing values, etc.).
2. Identifying key lead characteristics and their relationship to conversion.
3. Building a predictive model (logistic regression) to score leads.
4. Ensuring the model's accuracy and reliability through validation.
5. Delivering insights and recommendations to improve lead management and conversion rates.

# Data Processing – Current Status

## 1. Describe the data provided by X Education

- Dataset contains 9240 rows and 37 columns. It includes a mix of numerical (int64, float64) and categorical (object) data types.
- Data includes information on leads who have interacted with X Education's website.
- Variables cover lead demographics, source of lead, website activity, and conversion status.

## 2. Key variables and their relevance:

- Lead Source: Where the lead originated from (e.g., Google, organic search, referral).
- Total Time Spent on Website: How long the lead spent browsing courses.
- Page Views Per Visit: Number of pages viewed during a visit.
- Lead Origin: API, Landing Page
- Submission Converted: Whether the lead ultimately enrolled in a course (target variable)

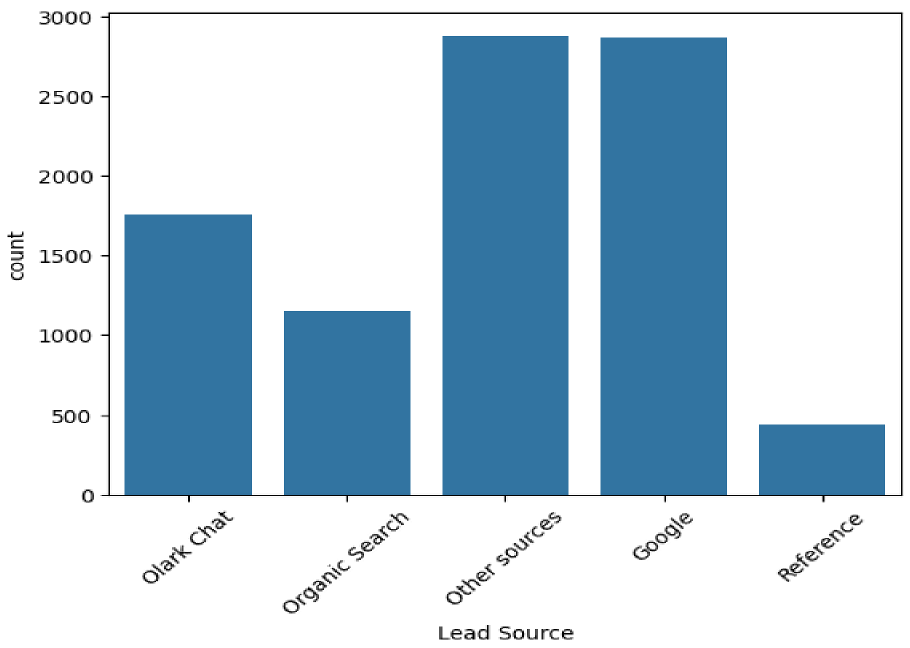
# Data Processing – Cleaning & Processing

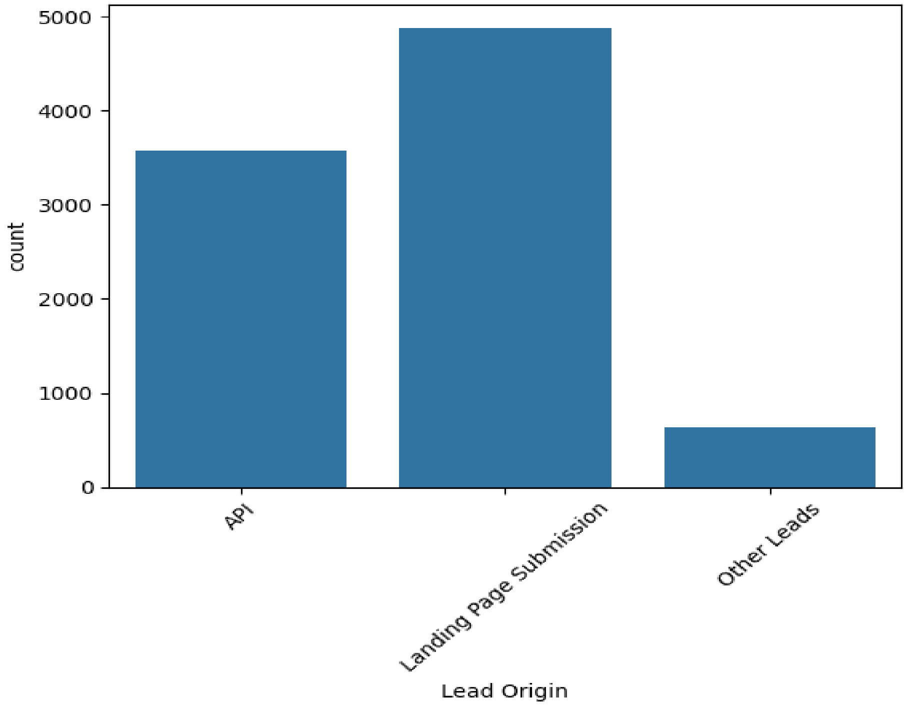
3. Steps taken to prepare X Education's data for analysis:

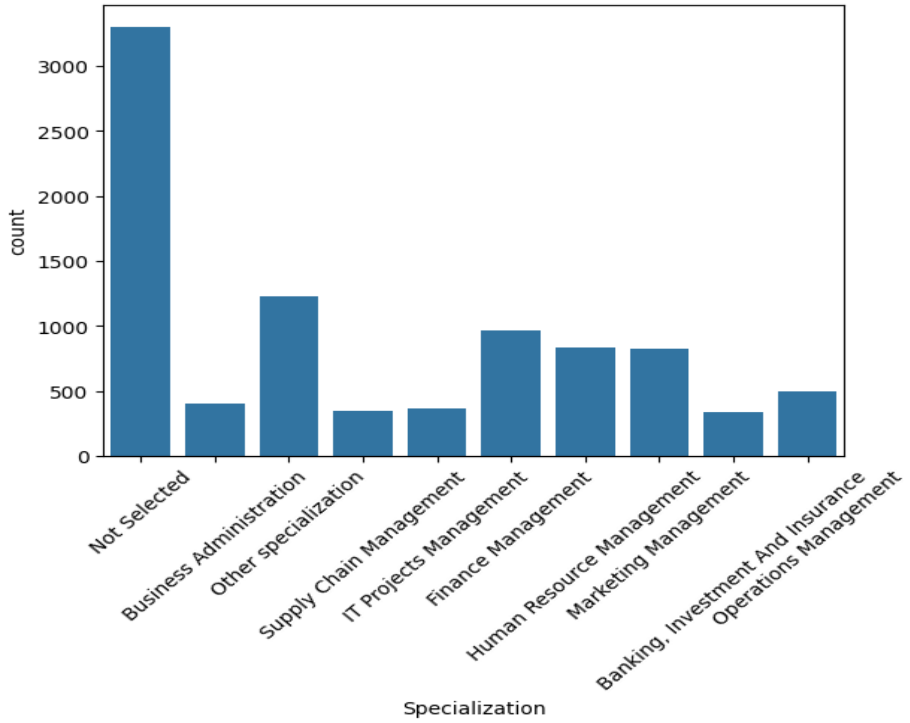
- Handling missing values: Missing values were addressed using appropriate imputation techniques (e.g., replacing with 'Unknown', mode) or removal where justified.
- Handling 'Select' values:  'Select' values were treated as 'Not Selected' to accurately reflect user input.
- Removing duplicates: Duplicate lead records were removed to ensure data integrity.
- Dropping columns like 'Magazine', 'Receive More Updates About Our Courses','I agree to pay the amount through cheque' ,'Update me on Supply Chain Content','Get updates on DM Content' as these fields have only one value and will not be useful for prediction.
- Missing values in City are imputed with mode value
- Dropping the rows where Total Visits is missing.
- **Categorical Column Analysis**: The value counts for categorical columns were calculated to understand the distribution of values. This analysis helped in identifying columns with low variance and potential grouping of categories.

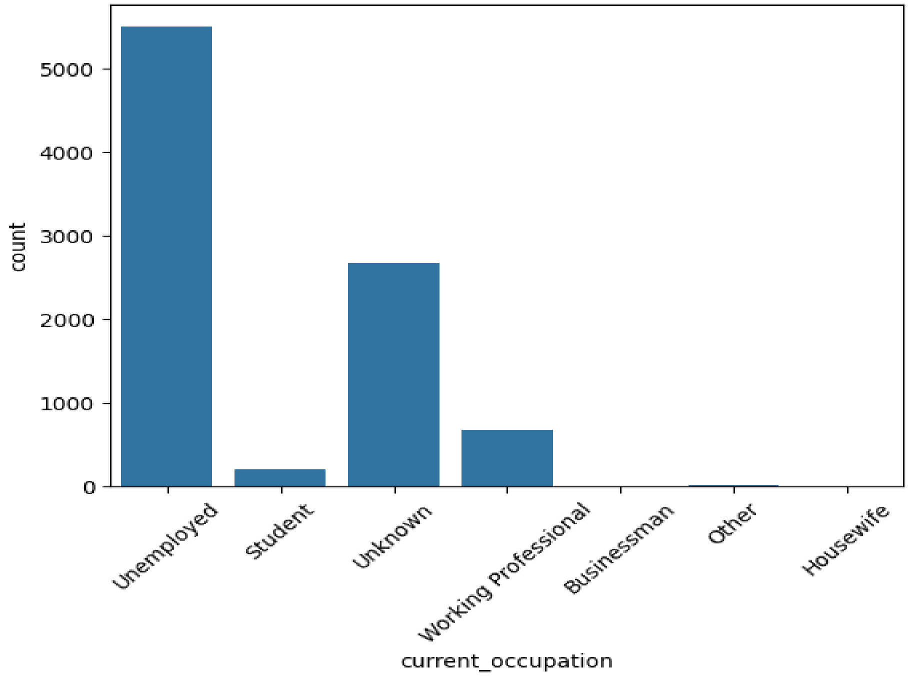
# EDA – Exploratory Data Analysis-Univariate

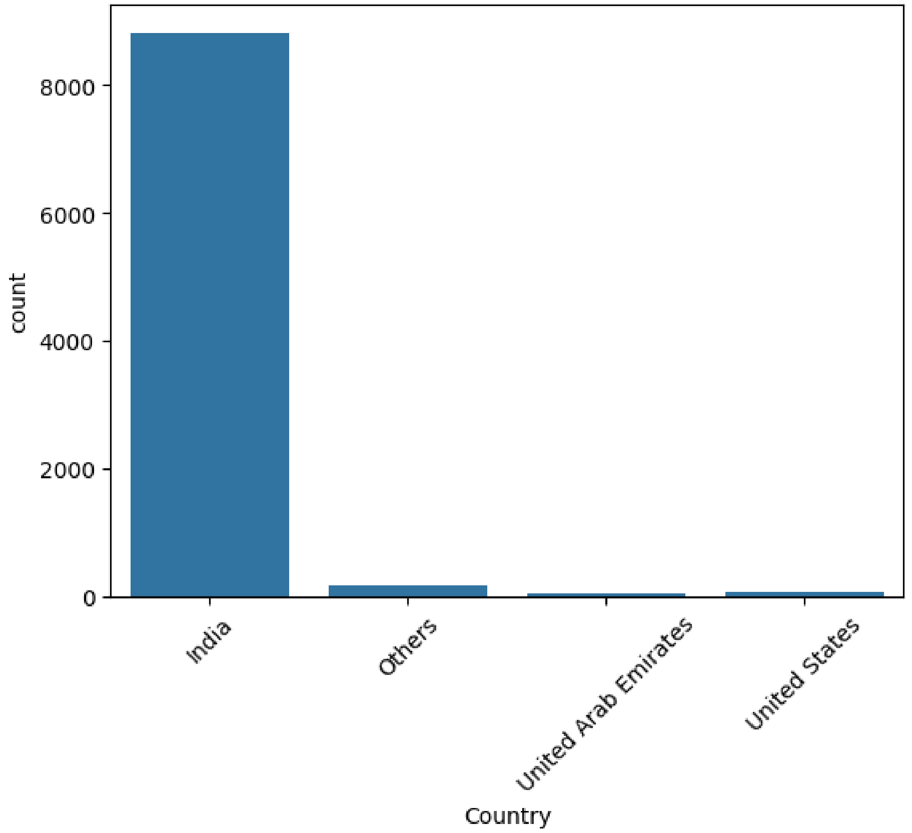
**Univariate Analysis**: Count plots and histograms were used to visualize the distribution of categorical and numerical variables, respectively.

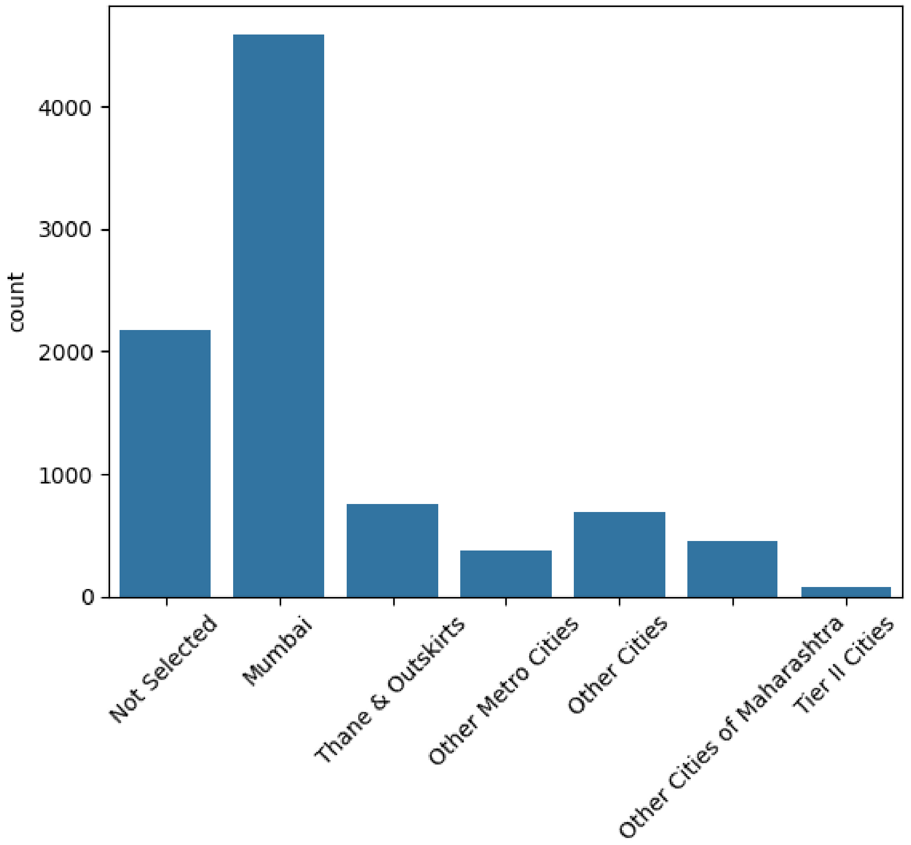


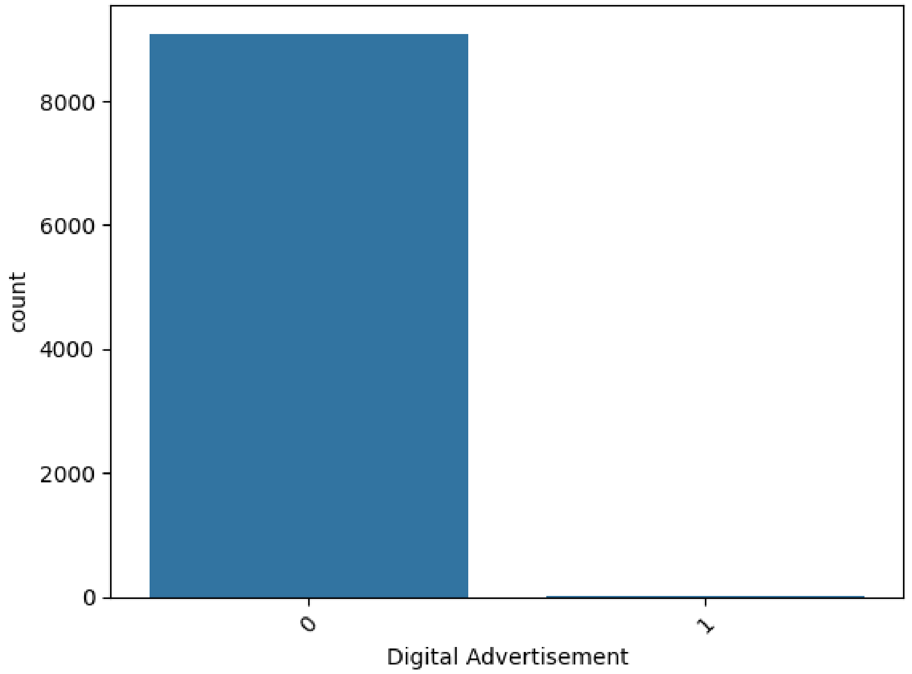


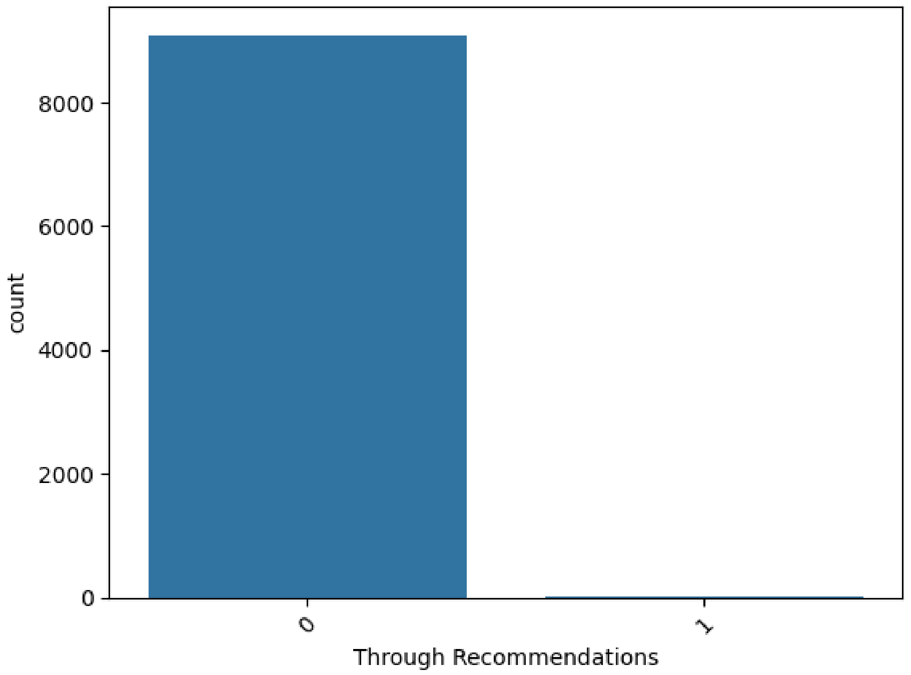






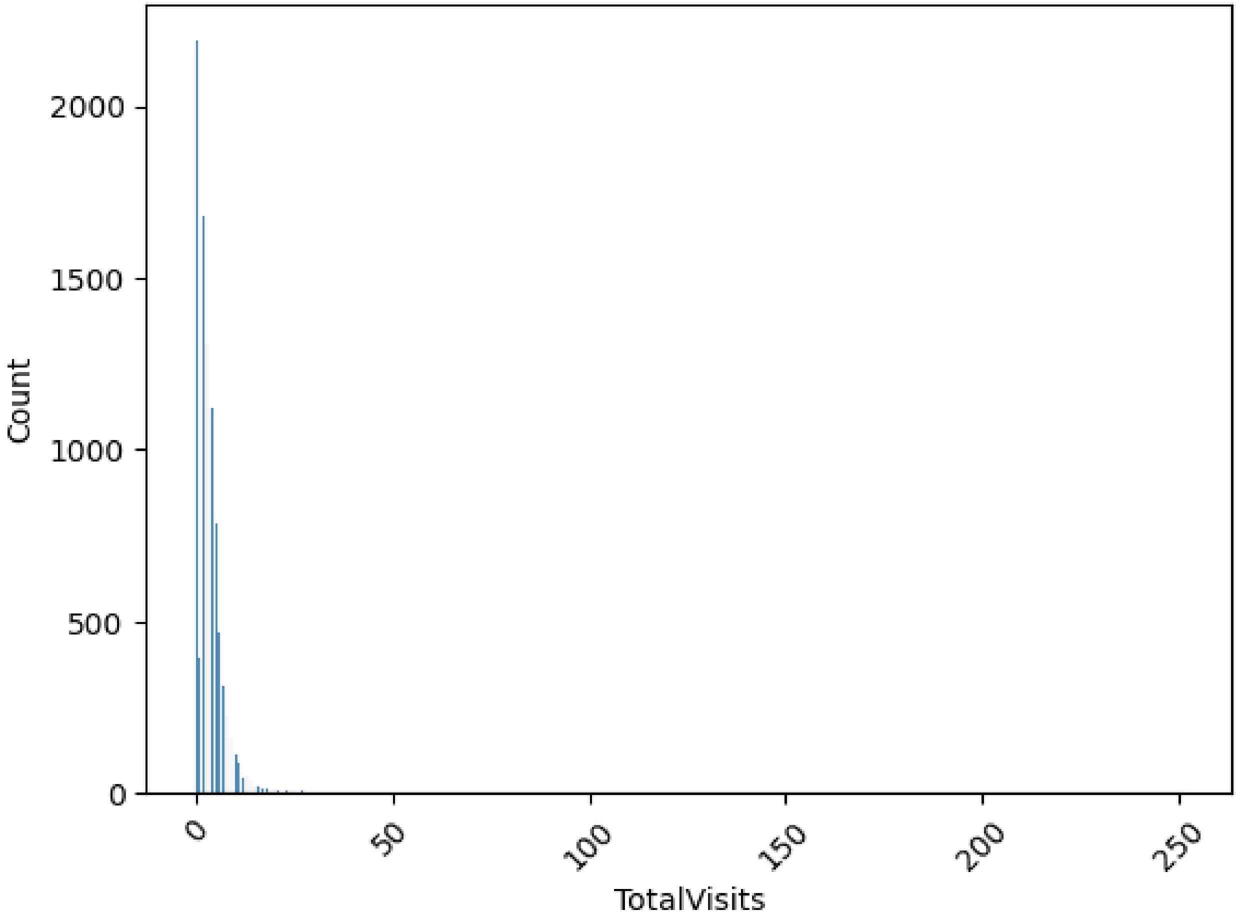


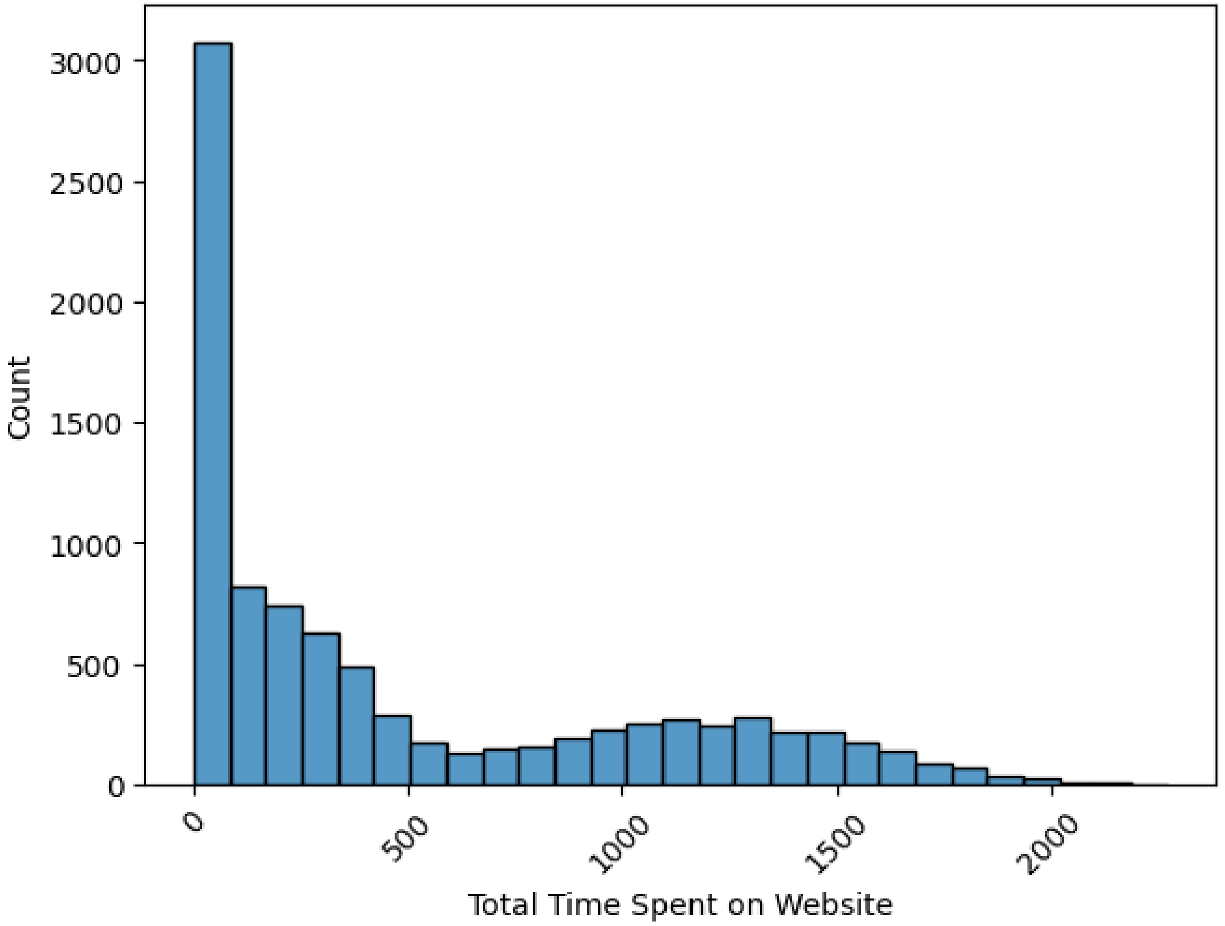


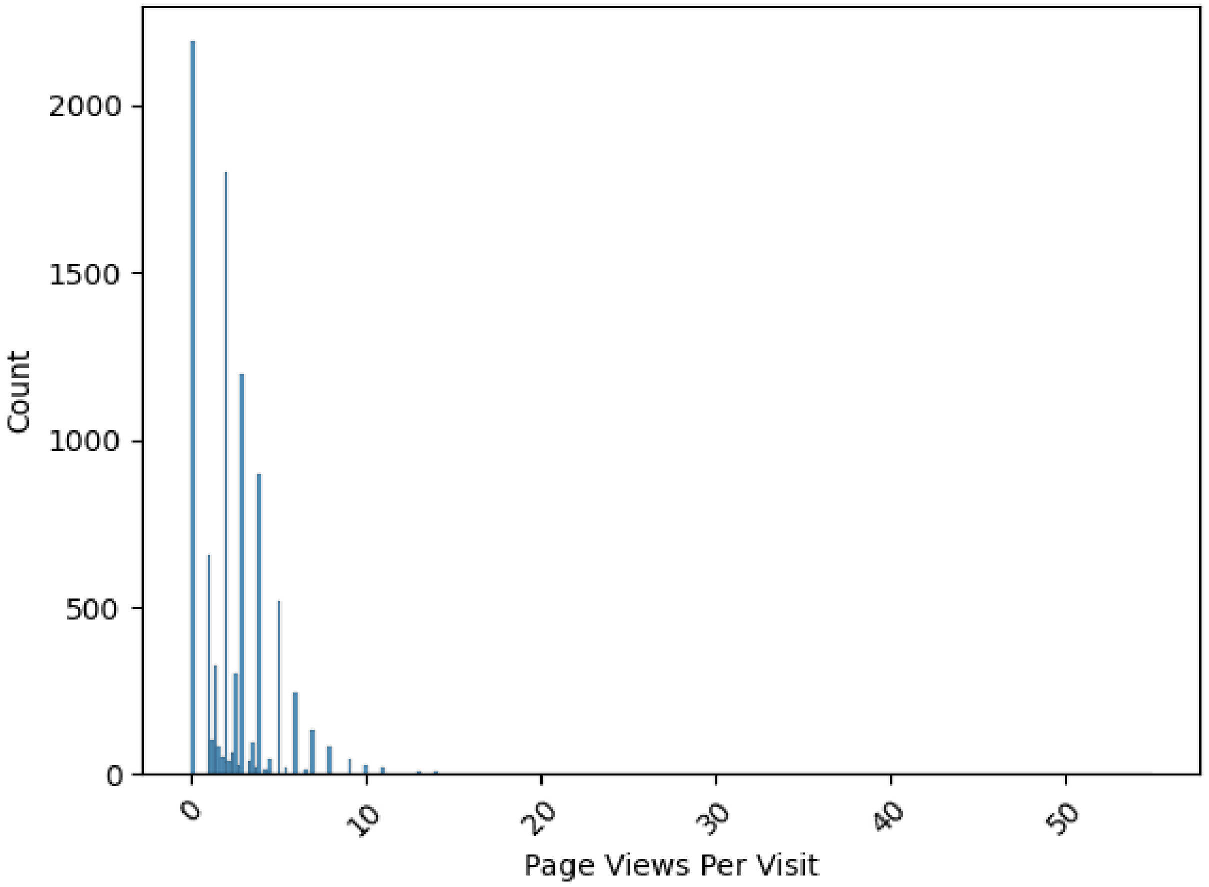


# EDA – Exploratory Data Analysis

**Univariate Analysis**: Count plots and histograms were used to visualize the distribution of categorical and numerical variables, respectively.







# EDA – Exploratory Data Analysis -BiVariate

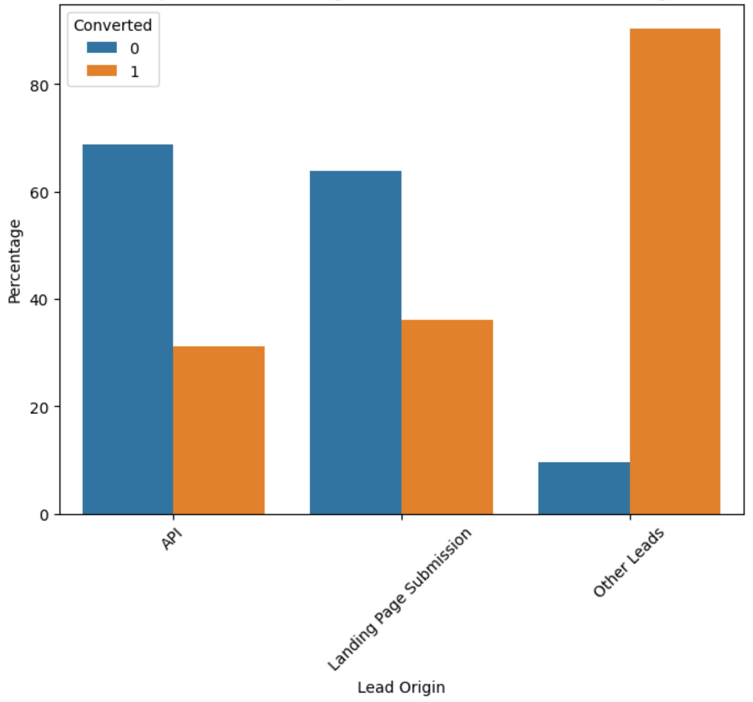
**Bivariate Analysis:**

Scatter plots were used to visualize the relationship between numerical variables and the target variable Converted.

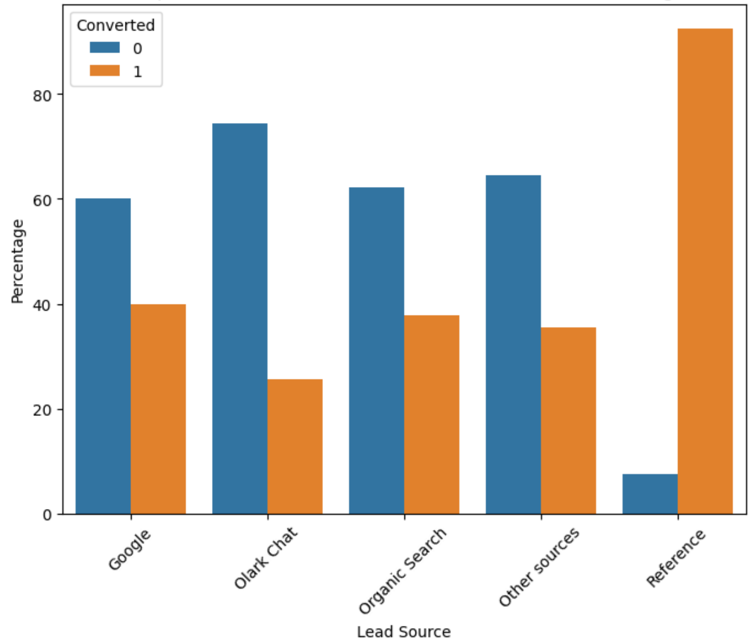
Bar plots were used to compare categorical variables with the Converted variable, showing the percentage of conversion for each category.

Box plots were used to visualize the distribution of numerical variables by Converted.

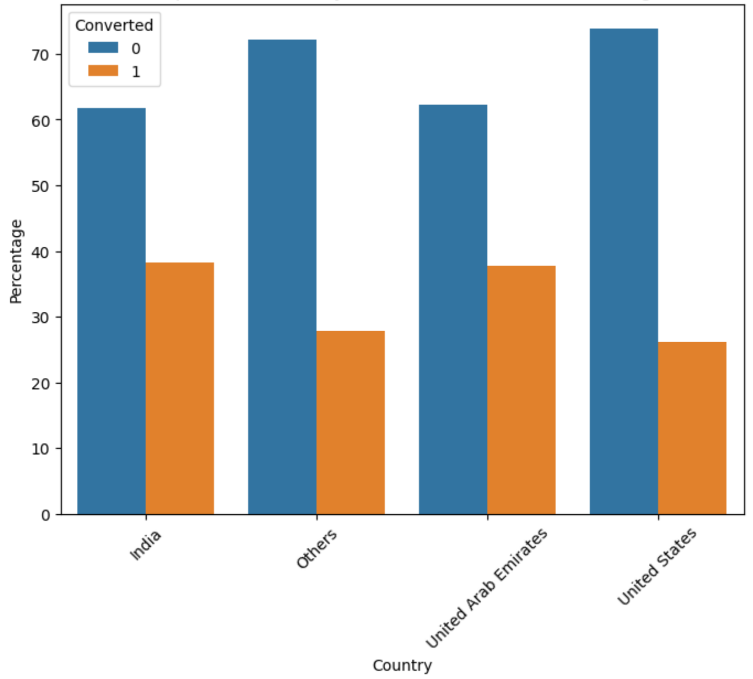
Comparison of Lead Origin with Lead Conversion (Percentage)



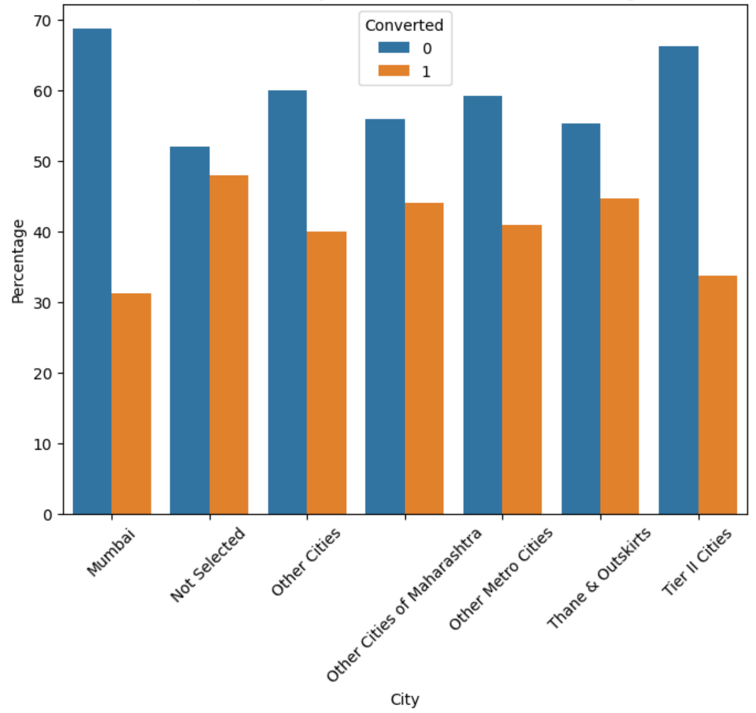
Comparison of Lead Source with Lead Conversion (Percentage)



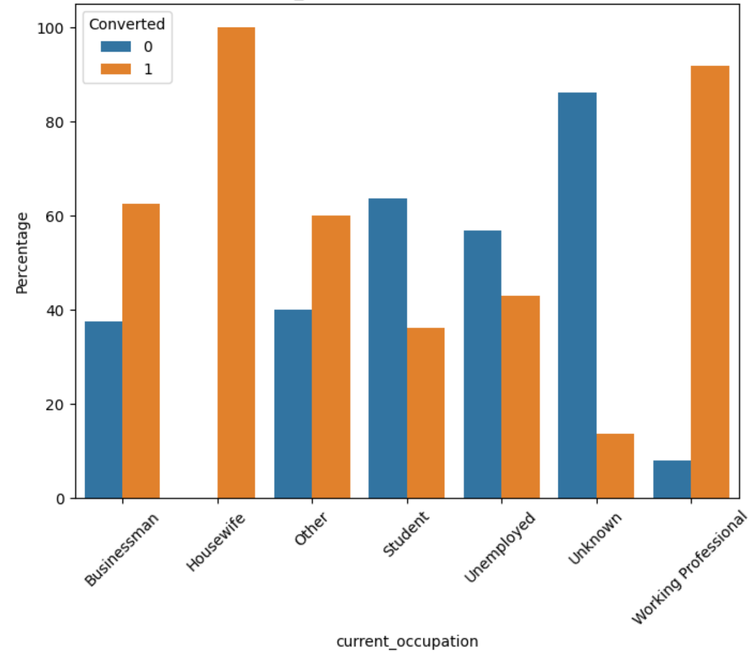
Comparison of Country with Lead Conversion (Percentage)



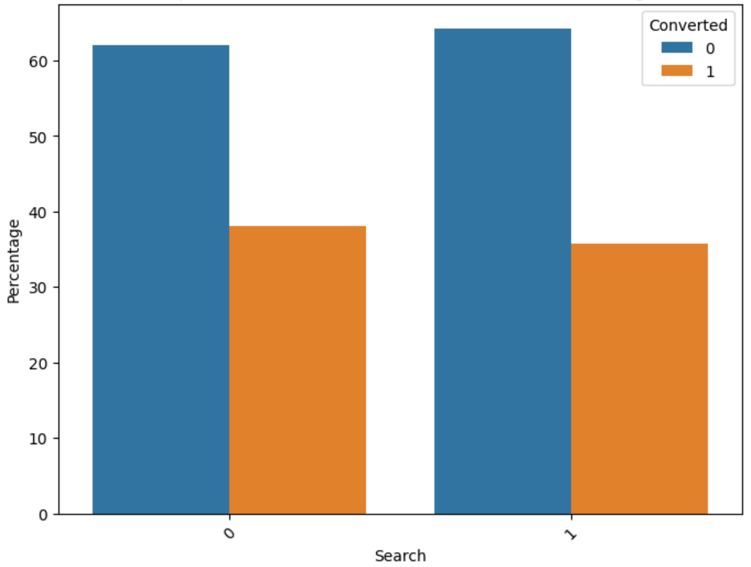
Comparison of City with Lead Conversion (Percentage)



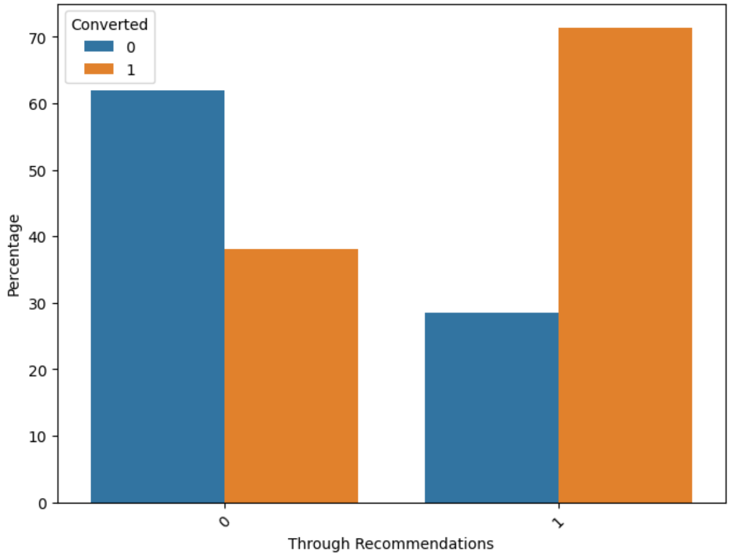
Comparison of current_occupation with Lead Conversion (Percentage)



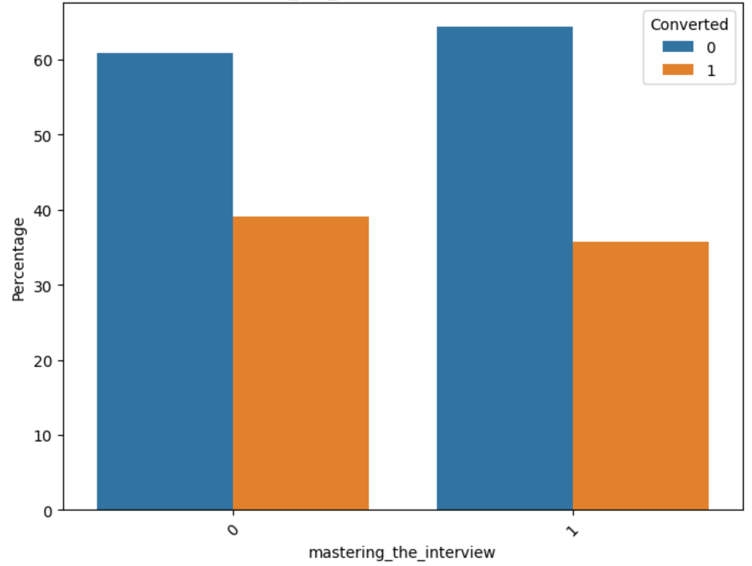
Comparison of Search with Lead Conversion (Percentage)



Comparison of Through Recommendations with Lead Conversion (Percentage)

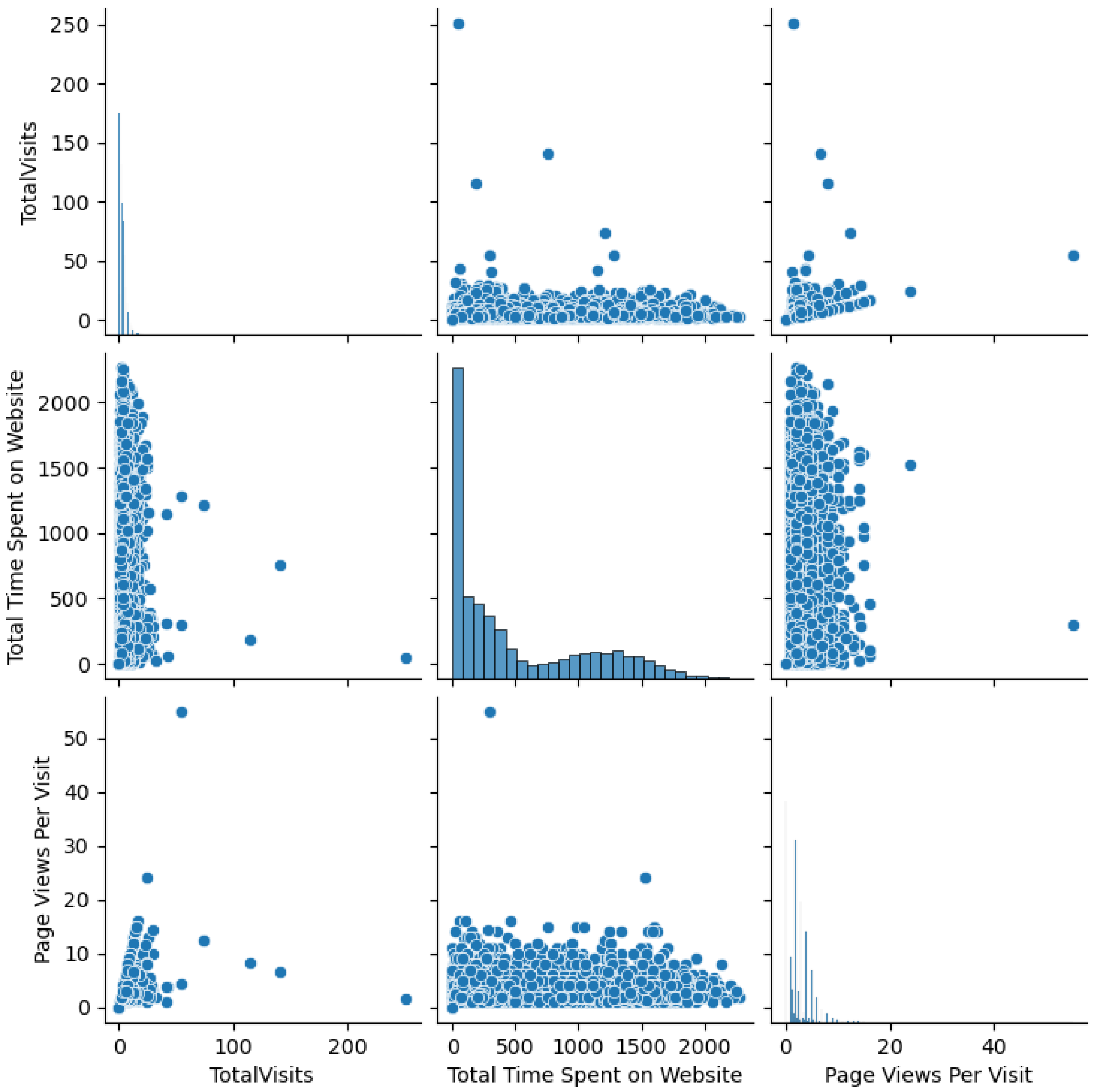


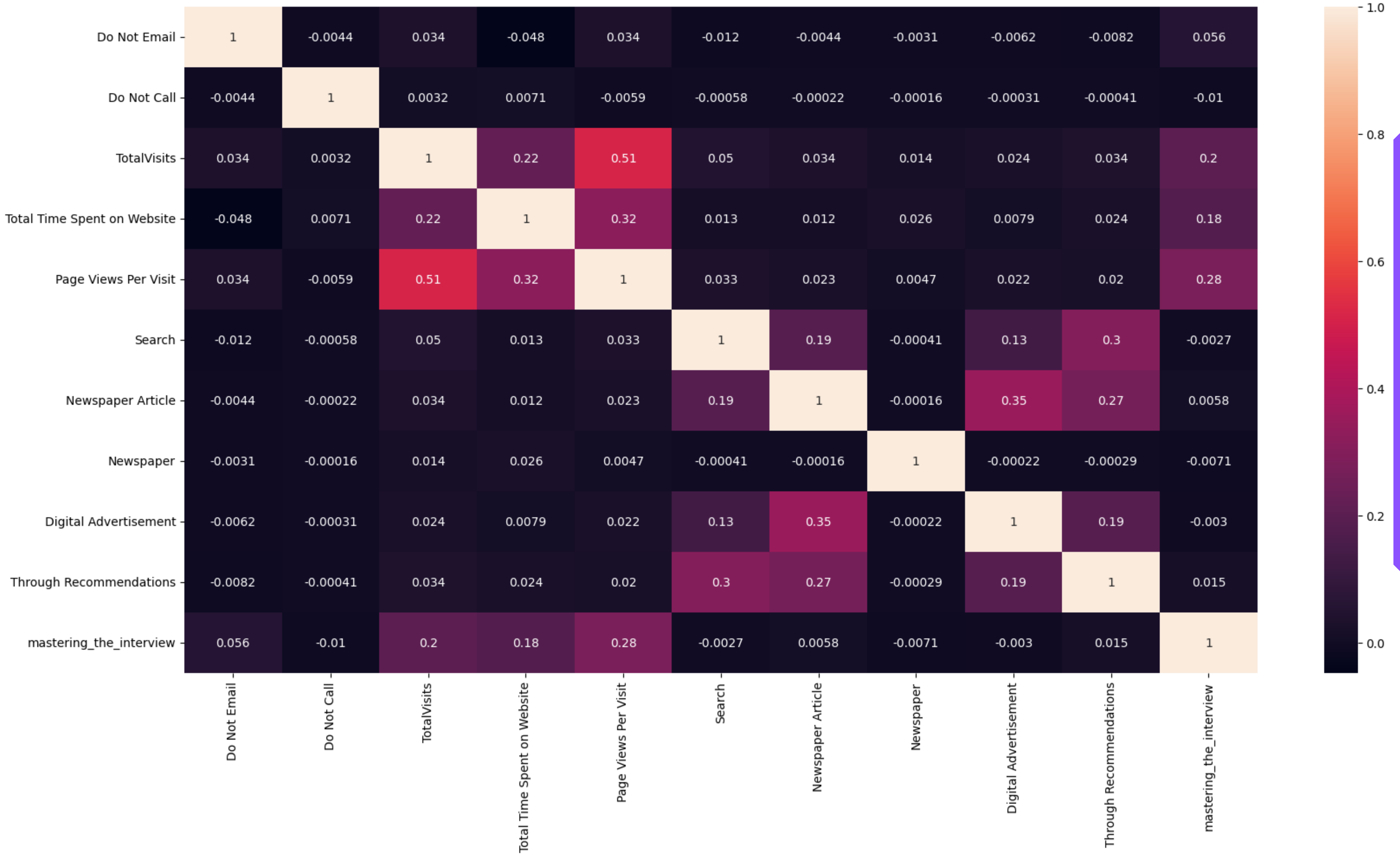
Comparison of mastering_the_interview with Lead Conversion (Percentage)



# EDA – Exploratory Data Analysis - Overview

1. Leads from sources like 'Reference' have the highest conversion rates, making them a valuable focus.
2. Although small in volume, leads from 'Other Leads' sources (e.g., Lead Add Form) are of high quality.
3. Leads from outside India, especially the United States, show lower conversion rates compared to Indian leads.
4. Indicating specialization (like Finance or Marketing) correlates with a higher likelihood of conversion.
5. Professionals like working individuals and businessmen have significantly higher conversion rates.
6. Metrics such as 'Total Visits' and 'Page Views Per Visit' are strongly correlated; keeping one is sufficient to simplify analysis.





# Model Building – Logistic Regression

## 1. Data Preparation:

- Categorical variables were converted into numerical variables using dummy variable creation (one-hot encoding).
- The data was split into training(70%)  and testing (30%) sets to evaluate model performance on unseen data.
- Before applying RFE, the numerical features are scaled using StandardScaler.
- Final list of Feature after multiple steps of RFE and VIF analysis

| | **Feature** | **VIF** |
|---|---|---|
| **0** | const | **10.265020** |
| **1** | Do Not Email | 1.101994 |
| **2** | Total Time Spent on Website | 1.325185 |
| **3** | Lead Origin_Landing Page Submission | 1.822494 |
| **4** | Lead Origin_Other Leads | 1.477892 |
| **5** | Lead Source_Olark Chat | 1.862756 |
| **6** | Last Activity_Email Opened | 1.760867 |
| **7** | Last Activity_Page Visited on Website | 1.235516 |
| **8** | Last Activity_SMS Sent | 1.752312 |
| **9** | current_occupation_Student | 1.056153 |
| **10** | current_occupation_Unemployed | 1.306268 |
| **11** | current_occupation_Working Professional | 1.291249 |
| **12** | Specialization_Finance Management | 1.053077 |

## 2. Model Building:

- A logistic regression model was used for predicting lead conversion.
- **Lead score** for each lead was assigned using the predicted probability from the model

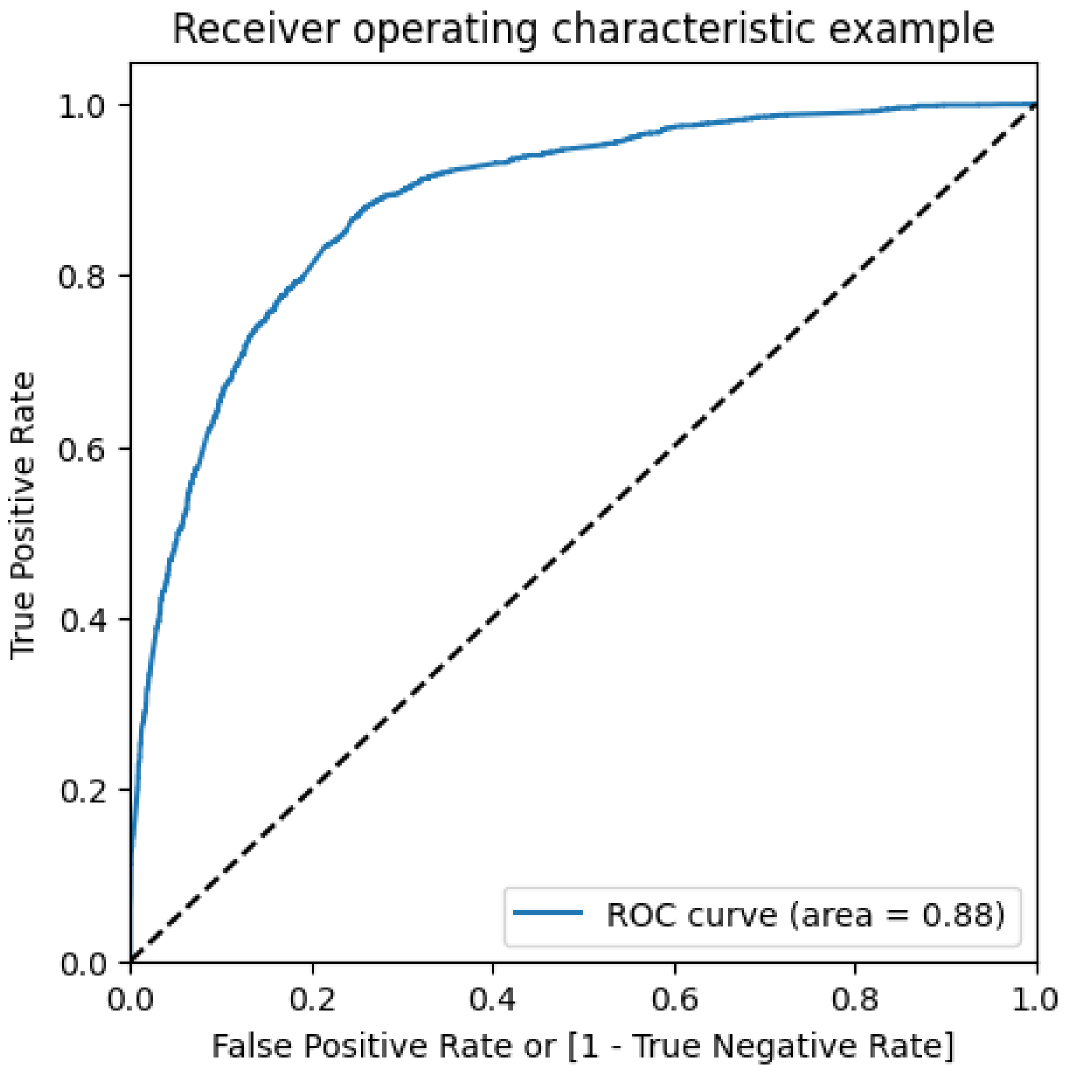
## 3. Model Evaluation:

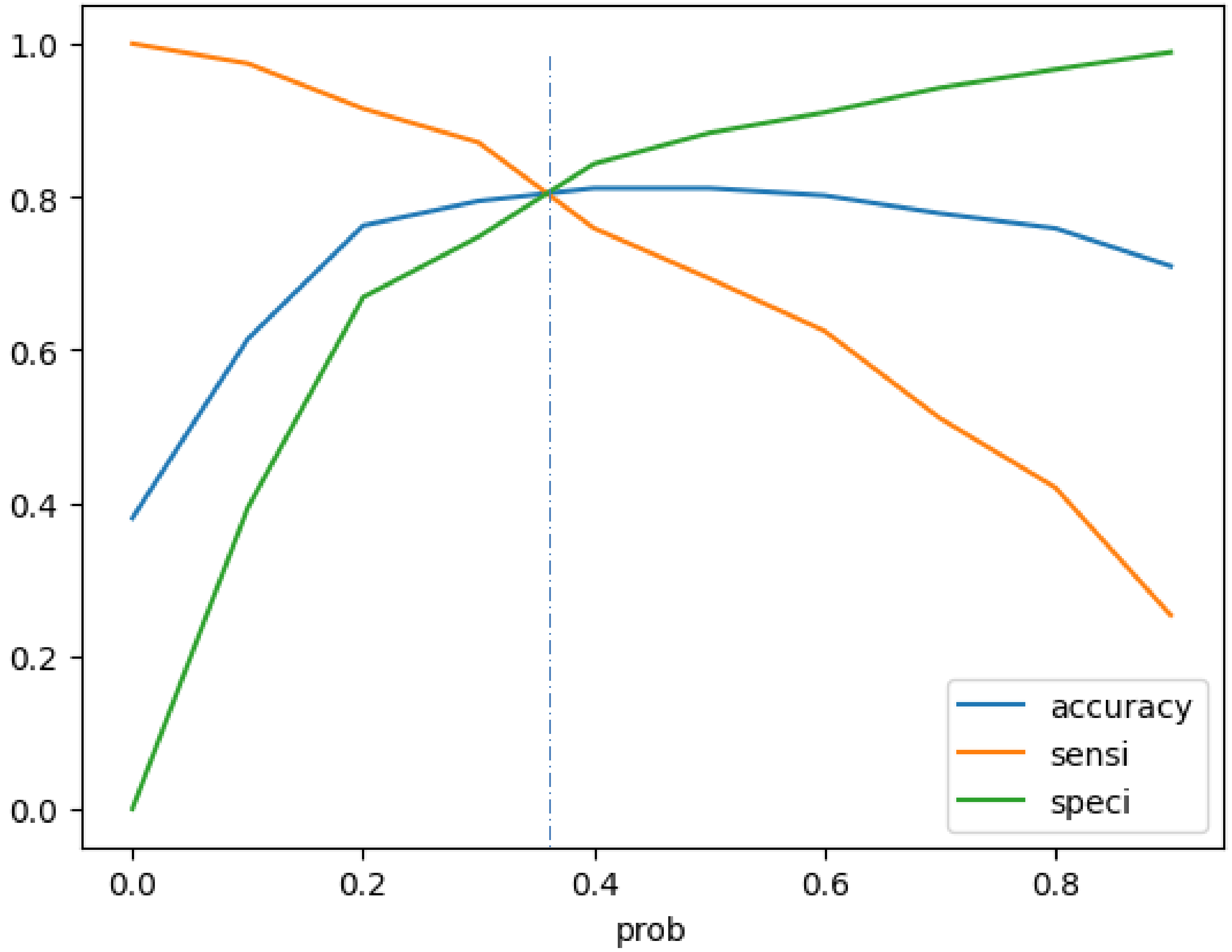
- The model's performance was evaluated on the test dataset.
- Evaluation metrics included accuracy, sensitivity, specificity, False Positive Rate, positive predictive value, and negative predictive value

# Model Building – ROC Curve

**Finding Optimal Cut off Point**

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the **optimal cut off is at >0.33**.





**Lead Score Creation**

- The optimal cut-off 0.33 was used on the predicted probability to assign 0 and 1 (1 indicates that the lead will convert) to each leads.
- Lead score was calculated for each lead on test and train dataset for the leads using the probabilities.

# Model Evaluation – Logistic Regression

The logistic regression model's performance was evaluated on the test dataset.

Here's a summary of the key metrics:

| Key Metrics | Test | Train |
|---|---|---|
| Accuracy | 80.3% | 81.1% |
| Sensitivity | 83.40% | 84.1% |
| Specificity | 78.5% | 77.2% |
| False Positive Rate | 21.4% | 22.8% |
| Positive Prediction Value | 69.4% | 69.3% |
| Negative Prediction Value | 88.9% | 88.8% |

## Interpretation

1. The model demonstrates a good balance between accuracy, sensitivity, and specificity.
2. The sensitivity of 83.4% indicates that the model is reasonably effective at identifying potential conversions.
3. The specificity of 78.5% suggests that the model is also reasonably good at identifying leads that are unlikely to convert.
4. The positive predictive value suggests that when the model predicts a lead will convert, it is correct about 69.5% of the time.
5. The negative predictive value suggests that when the model predicts a lead will not convert, it is correct about 89% of the time.

# Results– Insights

## Key Findings

1. The model identifies key factors that predict **lead conversion for X Education, including lead source, total time spent on the website, and lead origin**.
2. Leads originating from **Reference sources are highly likely to convert, indicating the effectiveness of referral programs**. **X Education should invest in strengthening referral channels.**
3. **Working professionals and businessmen show a higher conversion rate**. X Education can tailor marketing and sales approaches to target these segments effectively.
4. Leads with a specialization are more likely to convert which indicates leads having specialization are high interest leads. **X education can target the leads having specialization.**
5. Total Time Spent on Website is a strong predictor of conversion. **X Education should focus on optimizing website content and user experience to encourage longer engagement.**

## Actionable recommendations tailored to X Education:

1. Prioritize sales outreach to leads with **high lead scores**, focusing on those from high-converting sources.
2. **Develop targeted marketing campaigns for specific lead sources** and professional segments.
3. Optimize the website to increase engagement and encourage leads to spend more time browsing courses.

# Impact & Conclusion

1. **Increased lead conversion rate:** By focusing on hot leads, X Education can expect to see a significant improvement in their conversion rate above the current 30% baseline.

2. **Improved sales efficiency:** The sales team will be able to prioritize their efforts, leading to more efficient use of resources and increased productivity.

3. **Reduced customer acquisition costs:** By converting more leads with the same effort, X Education can lower the cost of acquiring each new customer.

4. **Better lead management:** The lead scoring system will provide X Education with a more structured and data-driven approach to lead management."

# Thank you