Lead Score Analysis

X EDUCATION DATASET

Introduction

The analysis of leads data is crucial for businesses to understand customer behaviour, preferences, and conversion patterns.

This report summarizes the key findings from the analysis of leads data for X Education and implementation of Logistic Regression model to predict lead conversion that helps in highlighting the technical and business aspects of the analysis.

Problem Statement

An education company named X Education sells online courses to industry professionals.

The company markets its courses on several websites, search engines, social media, past referrals etc. 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 into successful sales, while most of the leads do not. The typical lead to successful sale conversion rate at X education is around 30%.

X Education aims to identify high-potential leads, also known as “Hot Leads.”

To achieve this, the company requires a predictive model that assigns a lead score, where higher scores indicate a greater

likelihood of conversion. This system will help prioritize leads with a higher probability of becoming customers.

The CEO has set a target lead conversion rate of 80%, emphasizing the need for an efficient lead qualification process.

Methodology

Data Cleaning & Preprocessing: removing duplicates, dropping irrelevant columns, handling missing values, detection and removal of outliers and checking for data imbalance

Exploratory Data Analysis (EDA): Identified key patterns and correlations using visualizations.

Logistic Regression Model: Built and optimized a model to predict conversion likelihood.

Evaluation Metrics: Used Accuracy, Precision, Recall, F1-score, and AUC-ROC forperformance assessment.

Observation, Recommendation and Conclusion

Data Cleaning and Preparation

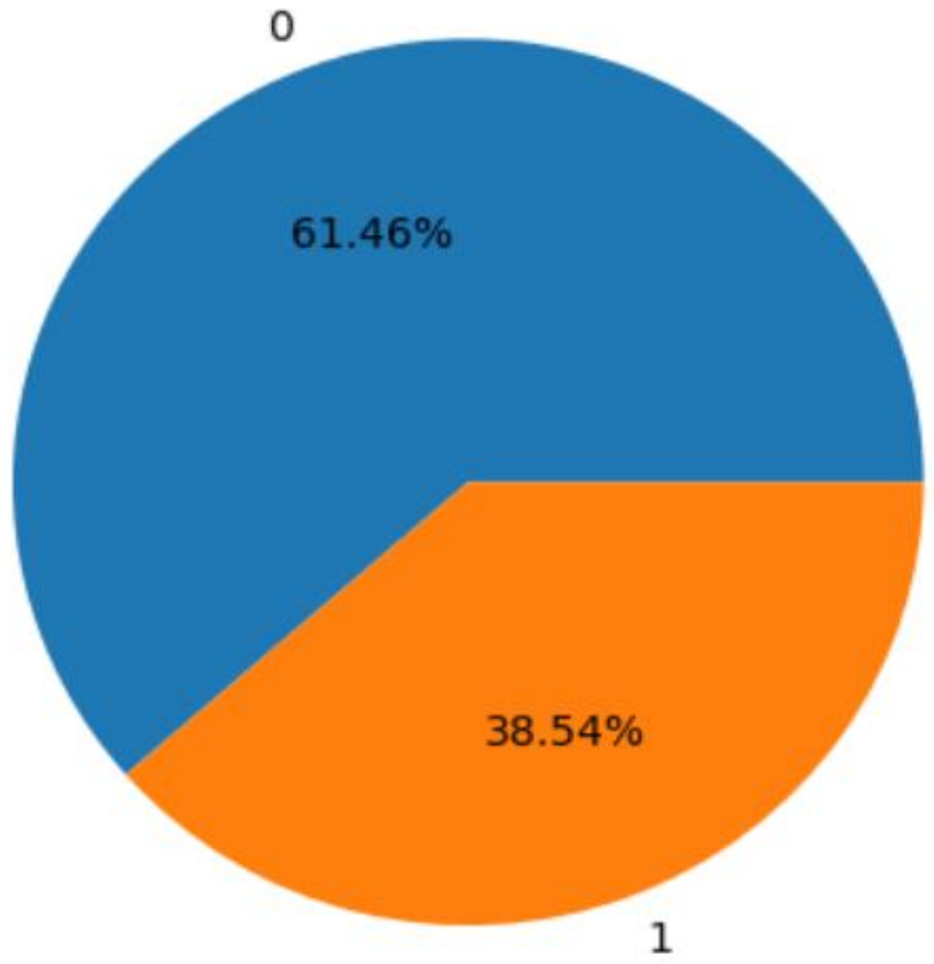
1. Removing the columns which are not contributing towards leads conversion and with null values above 40%
2. Checking for duplicates
3. Imputing the non relevant data points with relevant data points, e.g. Imputing ‘select’ with NAN and removing the columns with only one unique value
4. Handling missing values using appropriate imputation tehniques
5. Detecting and removing outliers using standardization techniques.
6. Checking data balance to ensure a fair modeling approach.

Exploratory Data Analysis

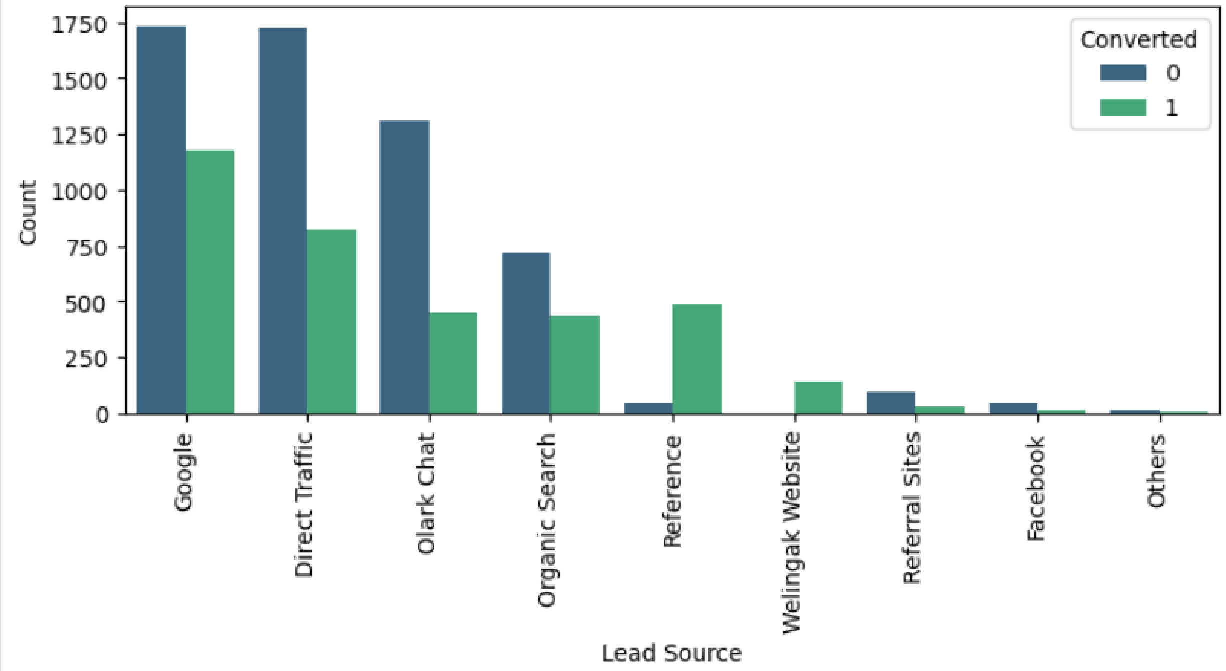
A. Target Column

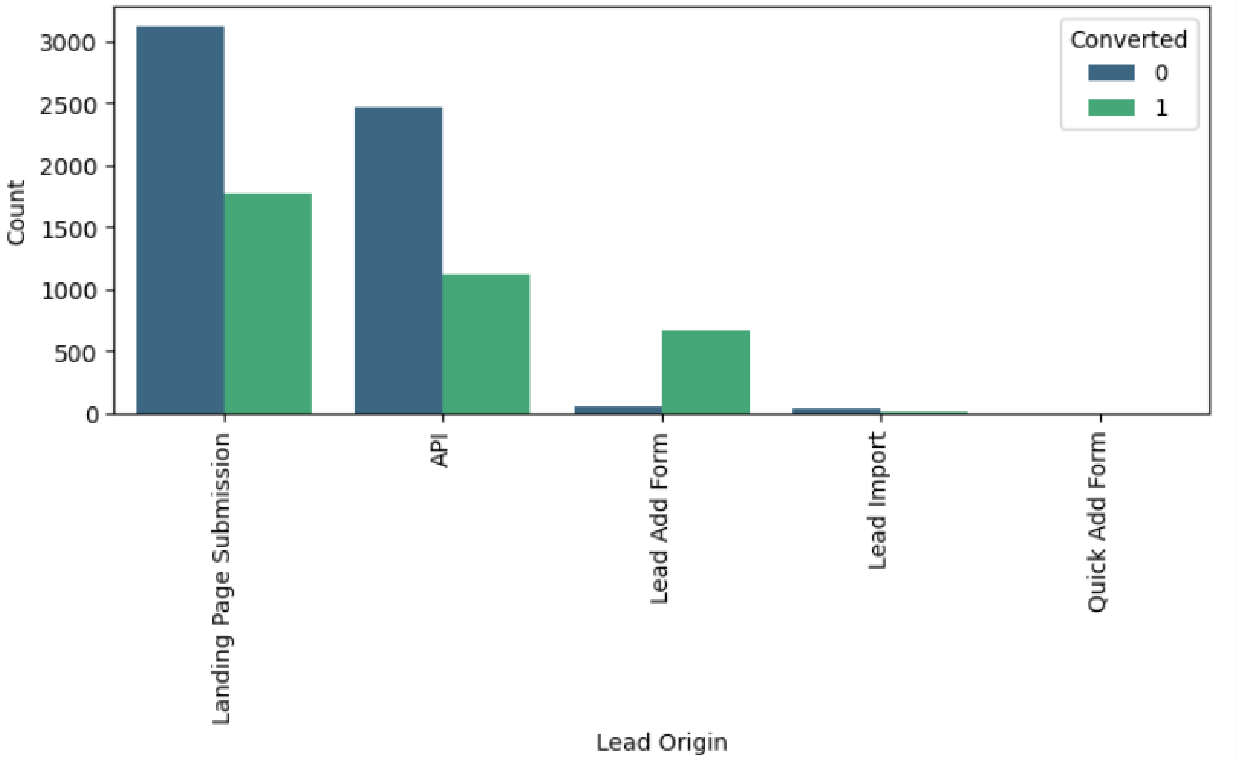
According to the pie chart,

there is 38.5% lead conversation rate



B. Categorical Features



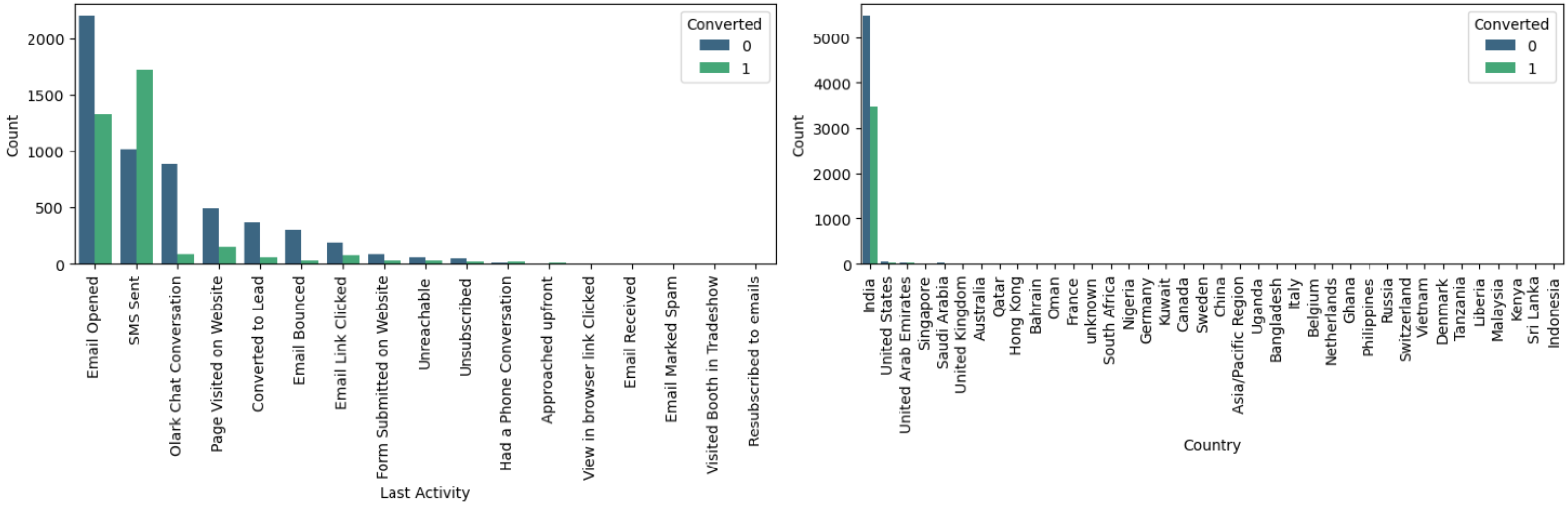


1. Top Lead Sources:

The primary sources of leads are Google, Direct Traffic, and Olark Chat, all of which show a strong conversion rate. Additionally, leads acquired through Referrals and the Welingak Website exhibit exceptionally high conversion rates.

2. Top Lead Origins:

The most common lead origins are Landing Page Submissions, API, and Lead Add Forms, with Lead Add Forms having the highest conversion rate.

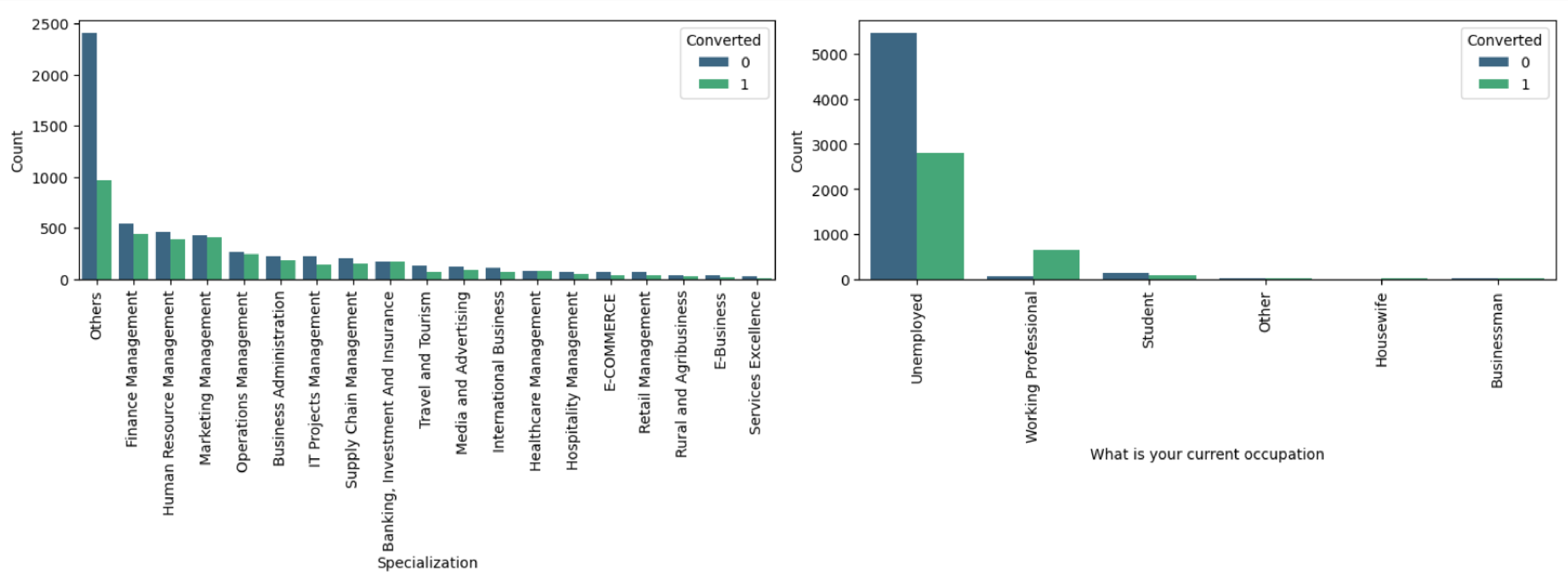


1. Top User Activities:

The most frequent last activities of leads include Email Opened, SMS Sent, and Olark Chat Conversations, with SMS Sent leading to the highest conversion rate.

2. Geographic Insights:

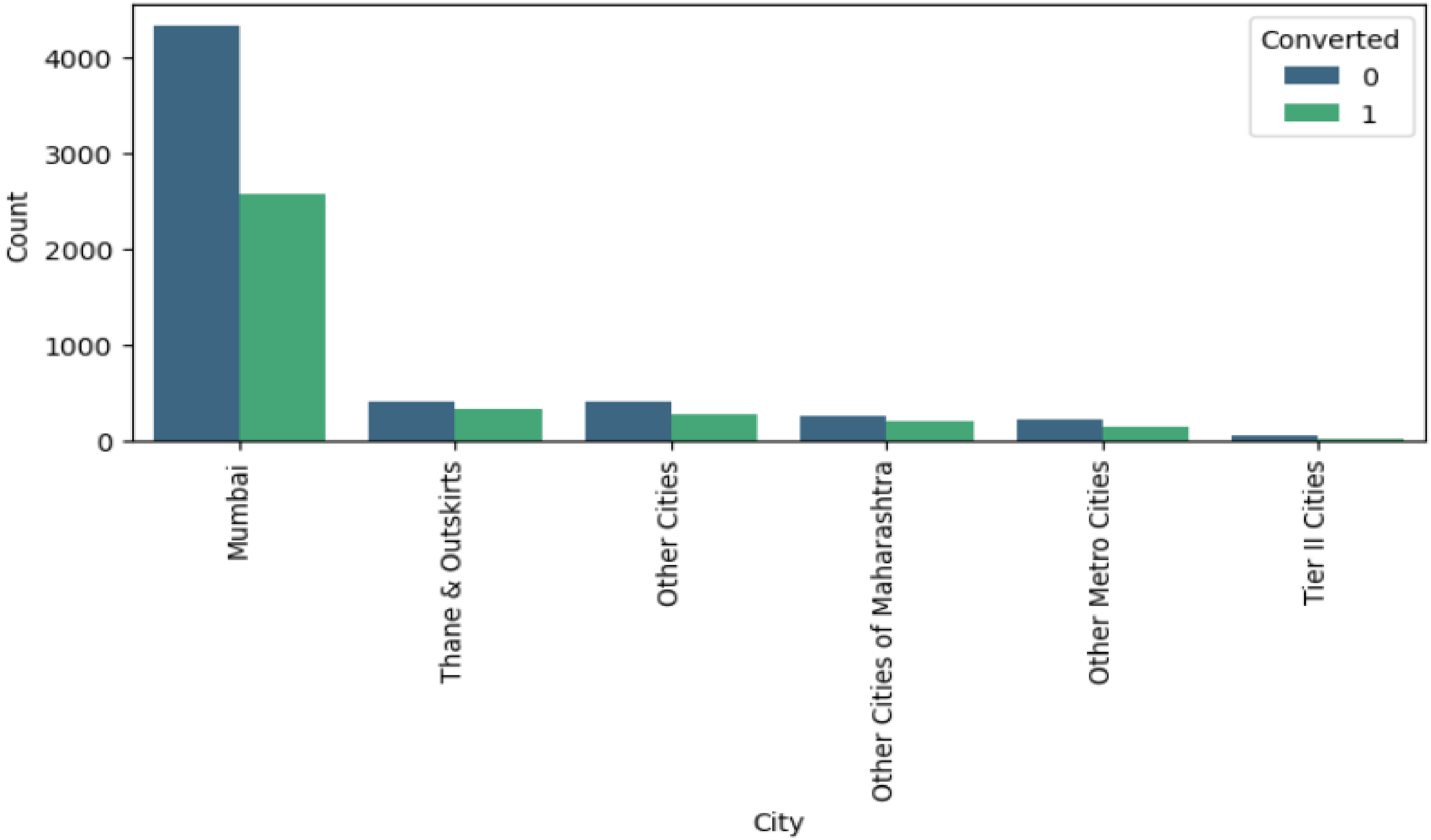
A significant majority of leads originate from India.



1. Specialization Category: The "Others" category in specialization accounts for the highest number of leads and conversions.

2. Employment Status:

While most leads come from unemployed individuals, working professionals demonstrate the highest conversion rates.



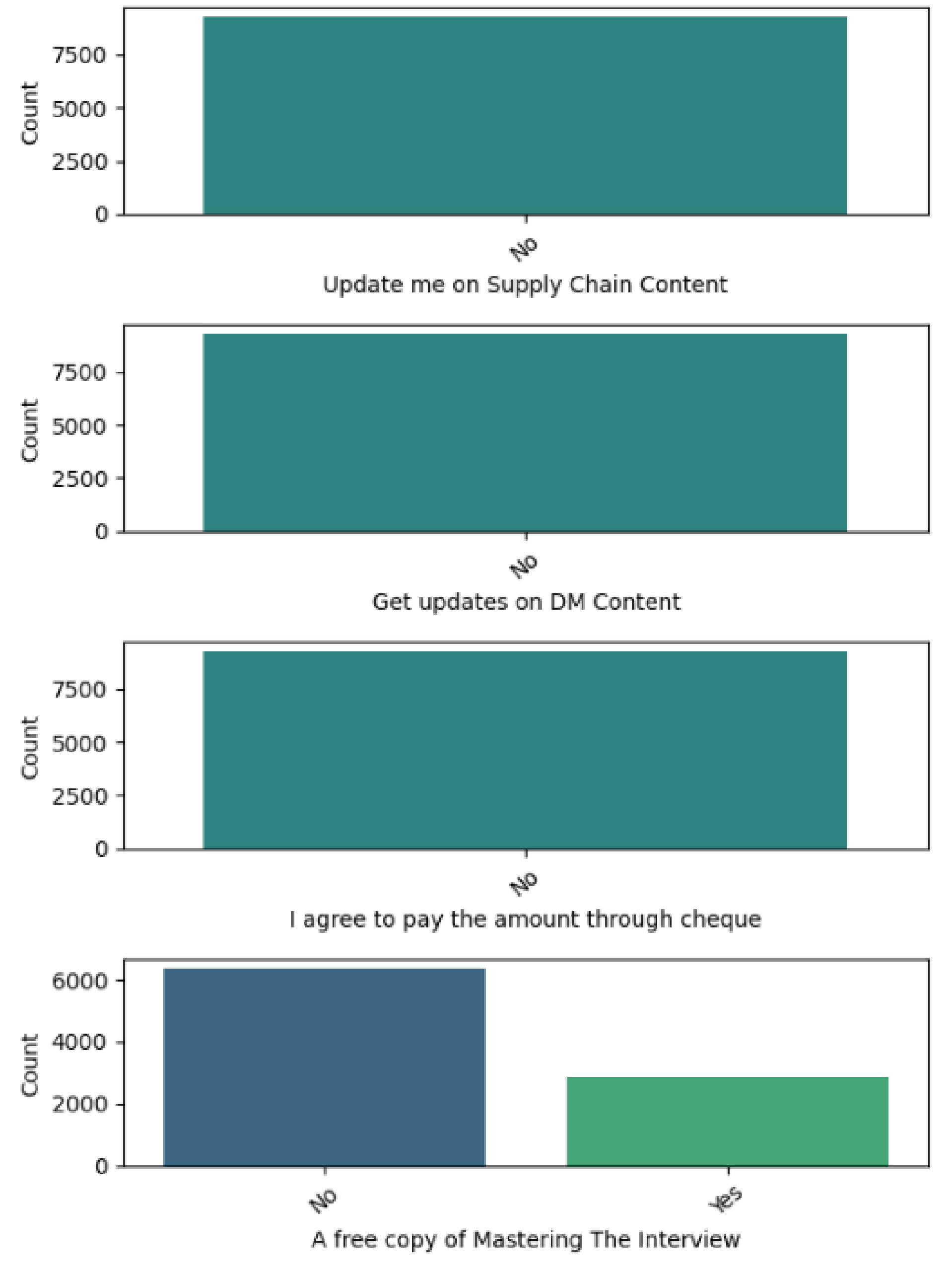
1. Regional Insights: The majority of leads are generated from Mumbaiwhich also records a good conversion rate.

C. Binary Categorical Features

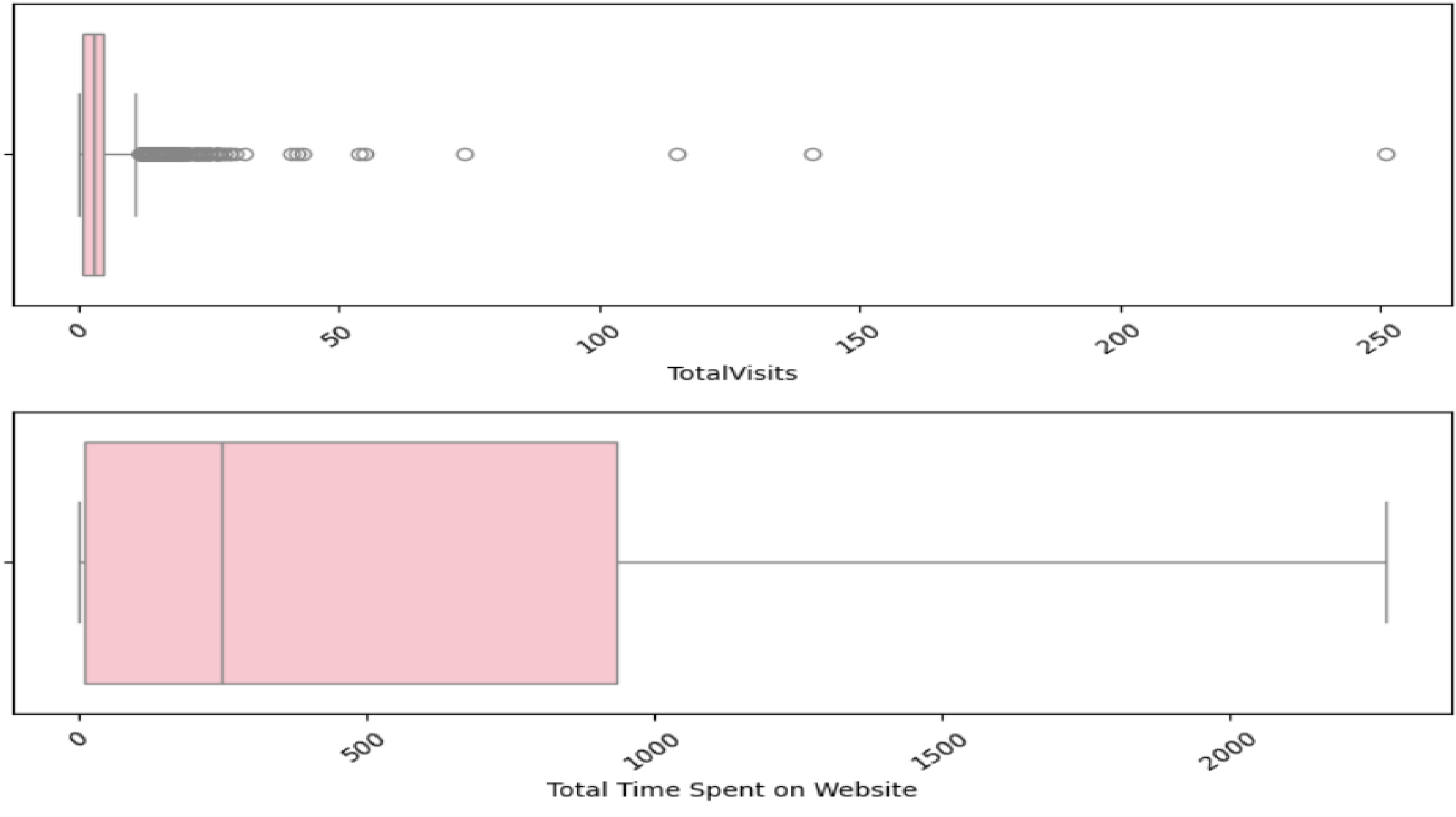
1.Most of the Binary categories are having only one unique value, i.e. either NO or YES

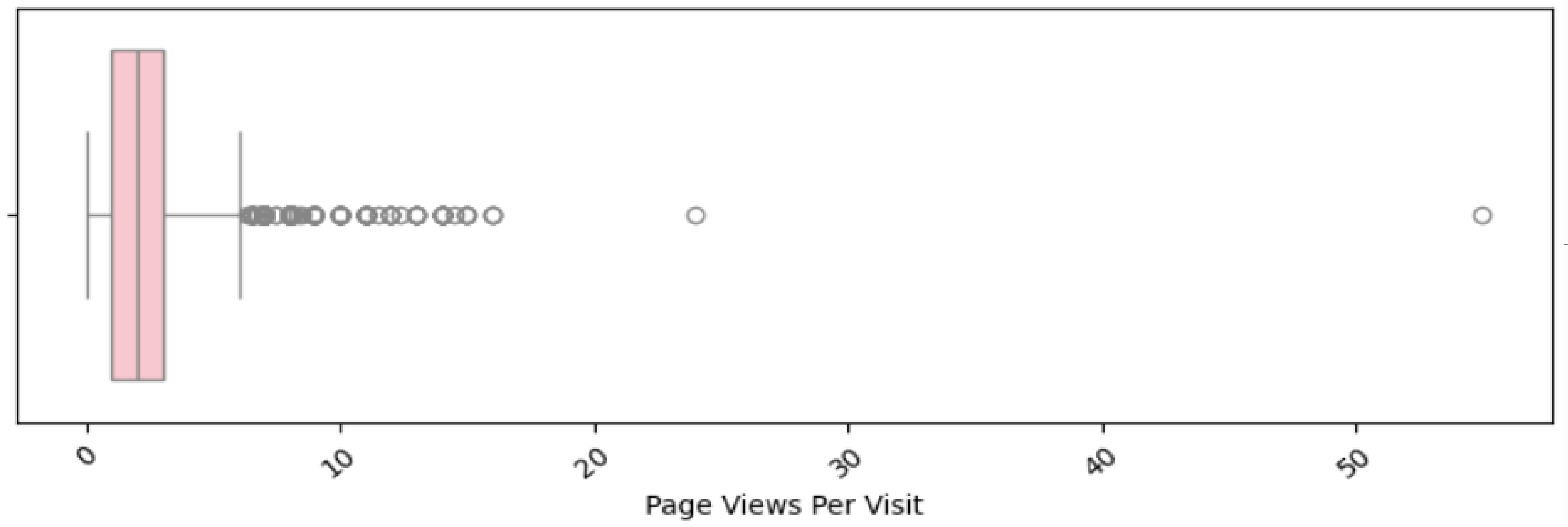
2.Only one feature, "A free copy of mastering the interview" has relevant no. of yes

3.Hence, all the unnecessary features can be removed before building our model¶



D. Numerical Features





From the boxplots, it can be observed that

There are no outliers present in the feature "total time spent on Website", whereas "Page Views Per Visit" and "TotalVisits" has multiple outliers

TotalVisits : the majority of values lie between 0-10

Total time spent on Website : The values 0-2500 (minutes)

Page Views Per Visit : The majority of value lie between

Model Training

The model training process involved several key steps:

• Multicollinearity Check: Ensured that independent variables were not highly correlated.

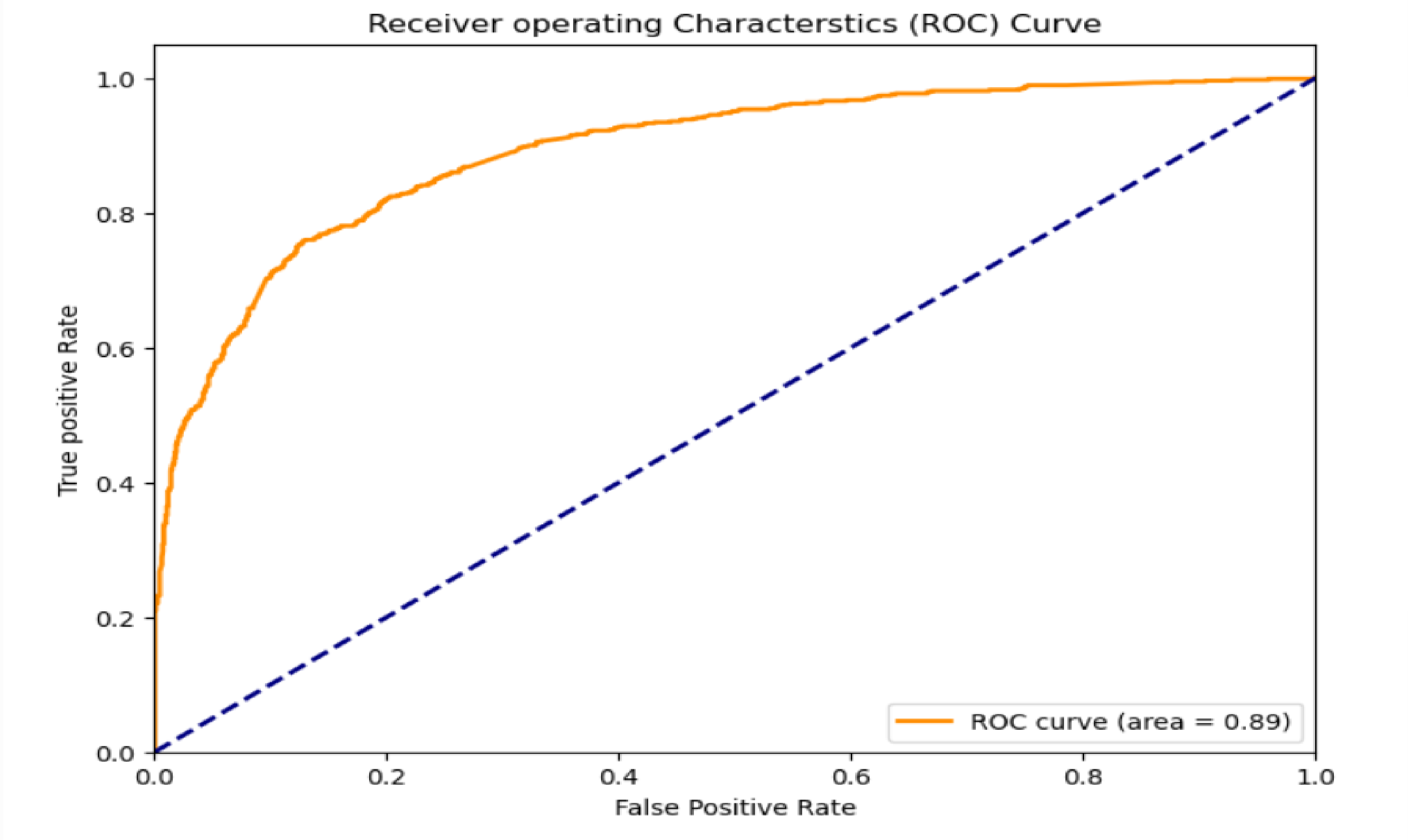
• Data Splitting: The dataset was split into 80% training and 20% testing using the train_test_split method.

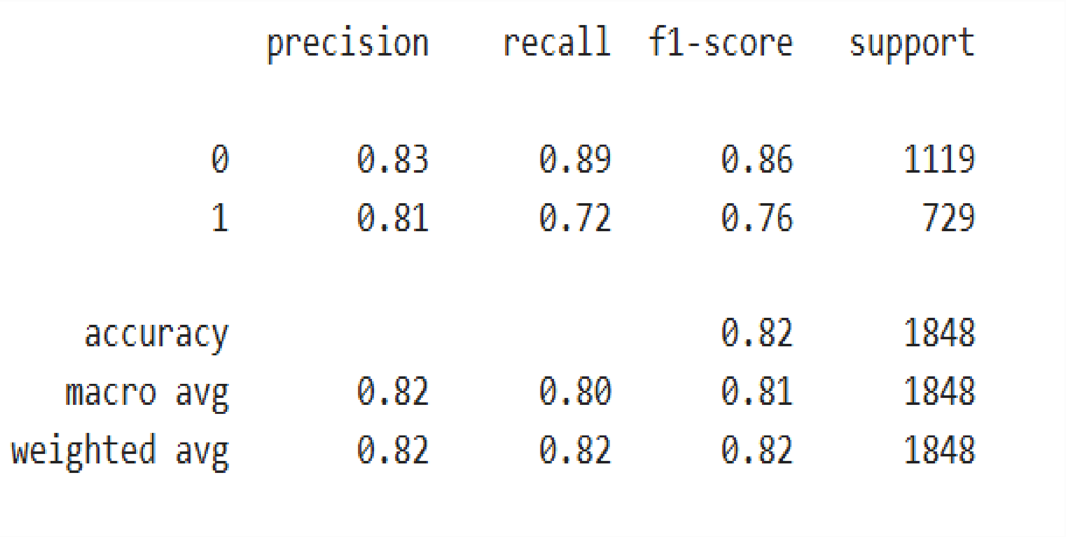
• Feature Selection: Recursive Feature Elimination (RFE) was applied to identify the top 10 most important features for the model.

• Model Optimization: Hyperparameter tuning was performed to improve performance and prevent overfitting.

• Model Performance: The final optimized Logistic Regression model achieved an overall accuracy of 82.25% on the test data.

Model Evaluation





```
              precision    recall  f1-score   support

           0       0.83      0.89      0.86      1119
           1       0.81      0.72      0.76       729

    accuracy                           0.82      1848
   macro avg       0.82      0.80      0.81      1848
weighted avg       0.82      0.82      0.82      1848
```

1. Overall Model Performance Accuracy = 82% → The model correctly classifies 82% of total samples, which is fairly good. Weighted Avg F1-score = 0.82 → Since this considers class imbalance, it confirms that the model has balanced performance across both classes.

2. Class-Wise Performance

Class 0 (Non-converted leads) Performance:

High Recall (0.89): The model captures most of the actual non-converted leads correctly. High Precision (0.83): When the model predicts a lead as "non-converted," it's usually correct. Conclusion: The model is very good at identifying non-converted leads, with few false negatives.

Class 1 (Converted leads) Performance:

Lower Recall (0.72): Some converted leads are misclassified as non-converted (False Negatives). Decent Precision (0.81): When the model predicts a lead as converted, it's right 81% of the time.

Conclusion: The model misses some actual conversions but performs reasonably well overall.

Recommendations

1. Implement a data-driven approach to refine marketing strategies.

2. Regularly update and analyze lead data to adapt to market changes.

3. Establish KPIs to monitor lead performance and conversion rates.

4. Focus on high-performing lead sources for marketing efforts.

5. Tailor campaigns based on customer behaviour insights.

6. Allocate resources to optimize lead generation from effective channels.

Conclusion

The logistic regression model we developed proved to be a superior lead scoring model. In nearly 82% of cases, it correctly assigns a higher lead score to leads that will convert compared to a lead who will not convert. By using this lead scoring model, the sales team can increase their conversion rate to 82% by focusing on the quality features that we get from the model.

The analysis of leads data provides valuable insights into customer behaviour and conversion strategies. The key findings and recommendations outlined in this report can help X Education refine its marketing strategies, optimize resource allocation, and improve conversion rates. By implementing these recommendations, X Education can improve its business outcomes and stay competitive in the market.