LEADS SCORING CASE STUDY

Problem Statement

- **Business Context:** X Education, an online learning platform, acquires potential customers through multiple marketing channels.
- **Challenge:** The existing lead conversion rate is relatively low, averaging around 30%.
- **Goal:** Build a logistic regression model assign a lead score (ranging from 0 to 100) to each prospect, enabling the identification of high-potential leads. The objective is to enhance the conversion rate and move closer to the 80% target.

Approach to solution

Data Preprocessing & Cleaning:

1. Identify and remove duplicate records.
2. Detect and manage missing values.
3. Eliminate columns with excessive missing data that are not useful for analysis.
4. Impute missing values when necessary.
5. Detect and address outliers in the dataset.

Exploratory Data Analysis (EDA):

1. Univariate Analysis – Examining individual feature distributions and value counts.
2. Bivariate Analysis – Analyzing relationships between variables using correlation and patterns.

Feature Engineering: Apply feature scaling and encode categorical variables using dummy variables.

Model Development: Implement logistic regression for classification and lead scoring.

Model Evaluation & Validation: Assess model performance to ensure accuracy and reliability.

Presentation of Findings: Interpret model results and insights.

Key Takeaways & Recommendations: Provide data-driven suggestions to enhance lead conversion.

Exploratory Data Analysis

Dataset Overview

Data Characteristics

- Contains 9,240 records with 37 attributes, covering customer demographics, engagement metrics, lead quality, and other relevant factors.

Important Features

- Prospect ID, Lead Origin, Lead Source, Conversion Status (Converted), Total Visits, Time Spent on Website, Page Views Per Visit, Lead Quality, Tags, and more.

Challenges in Data

- Presence of missing values in certain attributes.
- Categorical variables such as *Lead Quality* and *Tags* contain dispersed categories, requiring preprocessing.

Data Imbalance

Data Distribution & Model Feasibility

Conversion Rate:

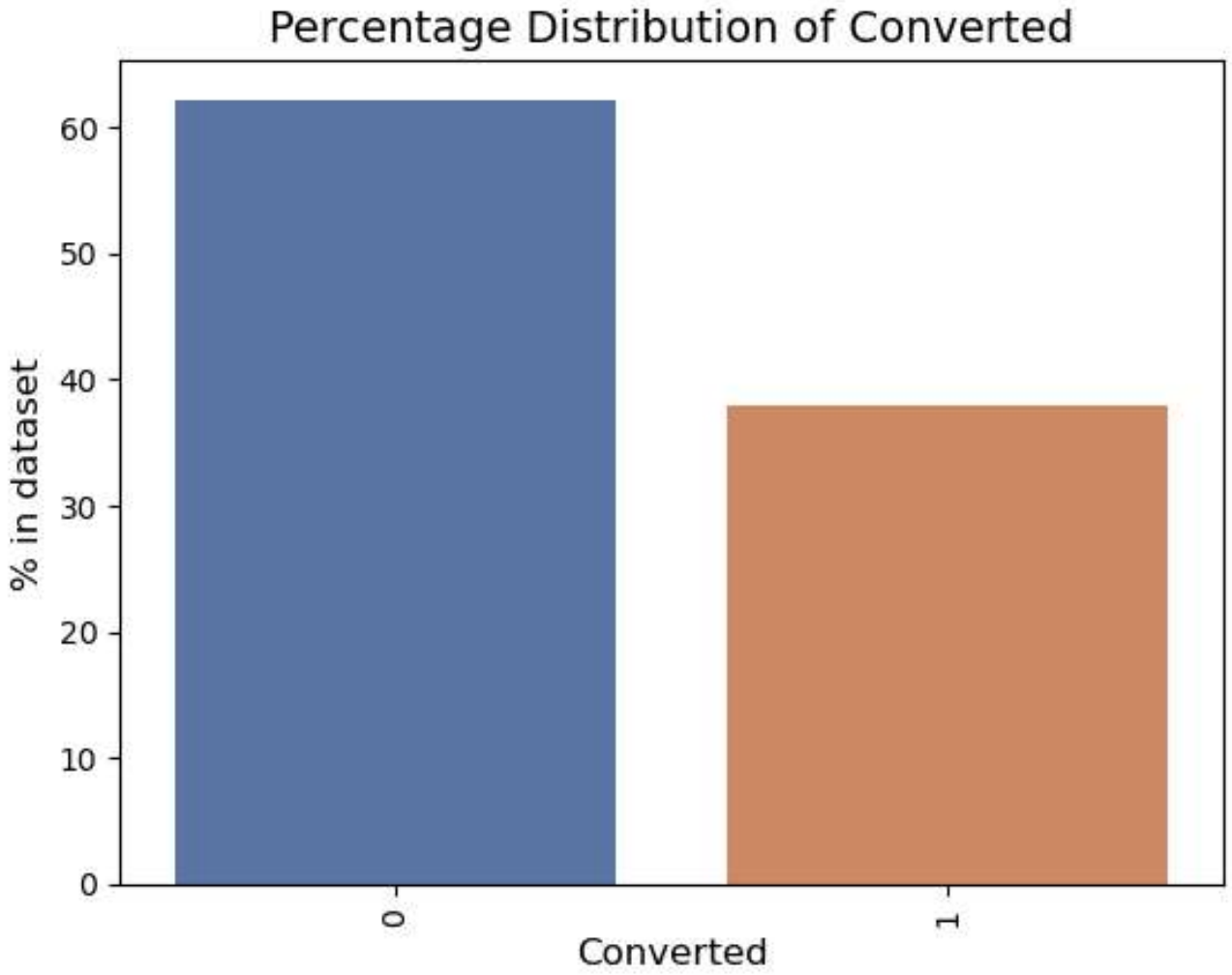
37.9% of the leads have successfully converted (i.e., 'Converted' = 1).

Dataset Balance:

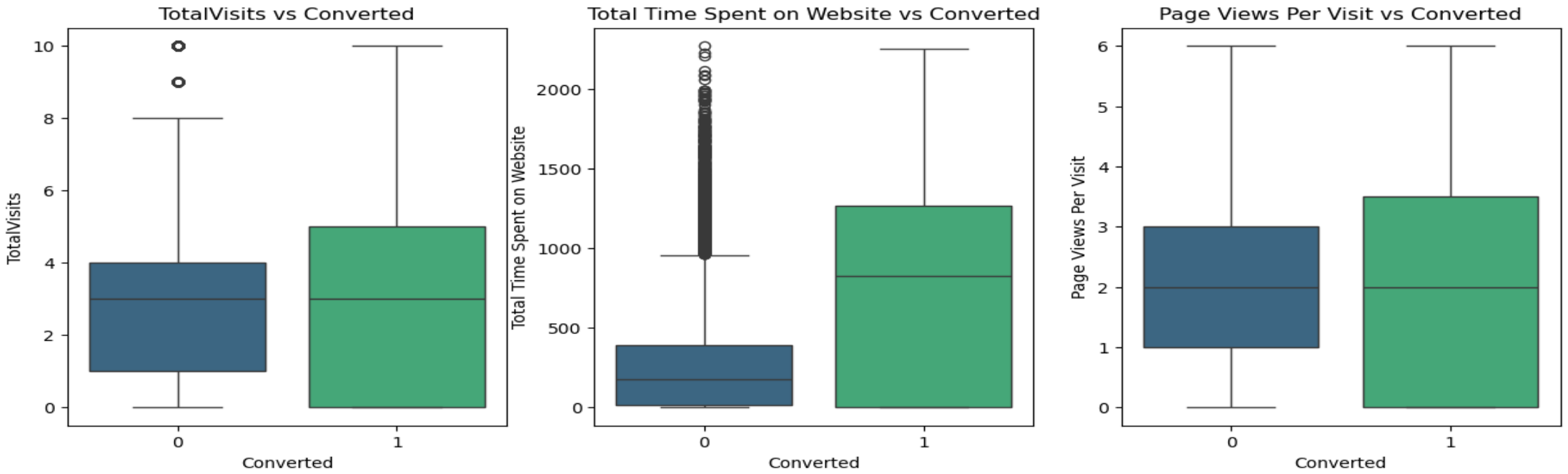
The dataset exhibits a moderate imbalance, which is manageable.

Model Suitability:

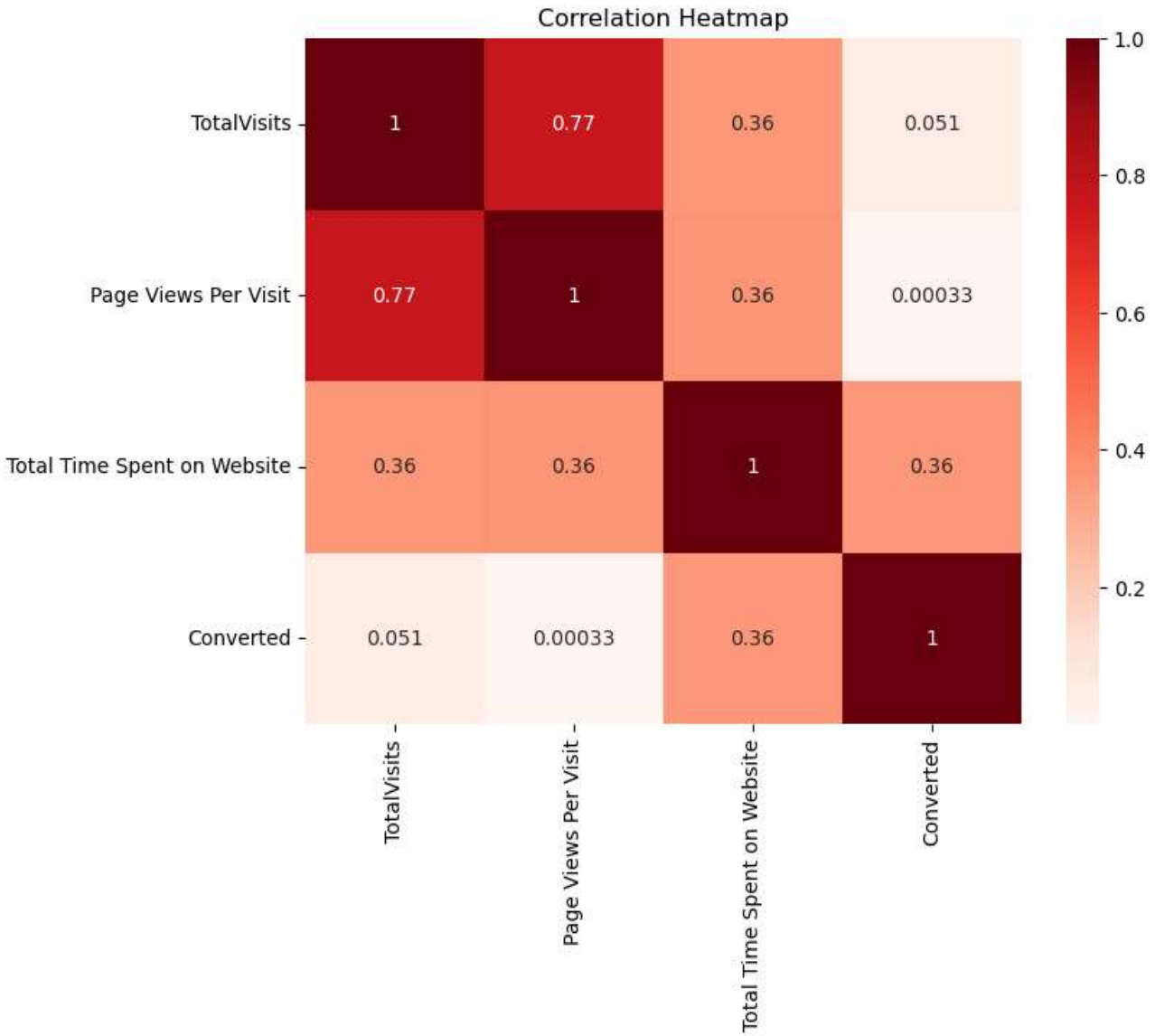
Logistic Regression is well-equipped to handle this level of class imbalance effectively.



Numerical variable vs Target variable



The analysis shows that **Total Visits** has a similar median for both groups, but converted users exhibit greater variability. **Time Spent on Website** is significantly higher for converted leads, with a wider spread and fewer outliers, suggesting an **optimal engagement range** for conversions. In contrast, **Page Views Per Visit** has the same median for both groups, making it a weaker predictor. Overall, **time spent on the website is a key indicator of lead conversion**.



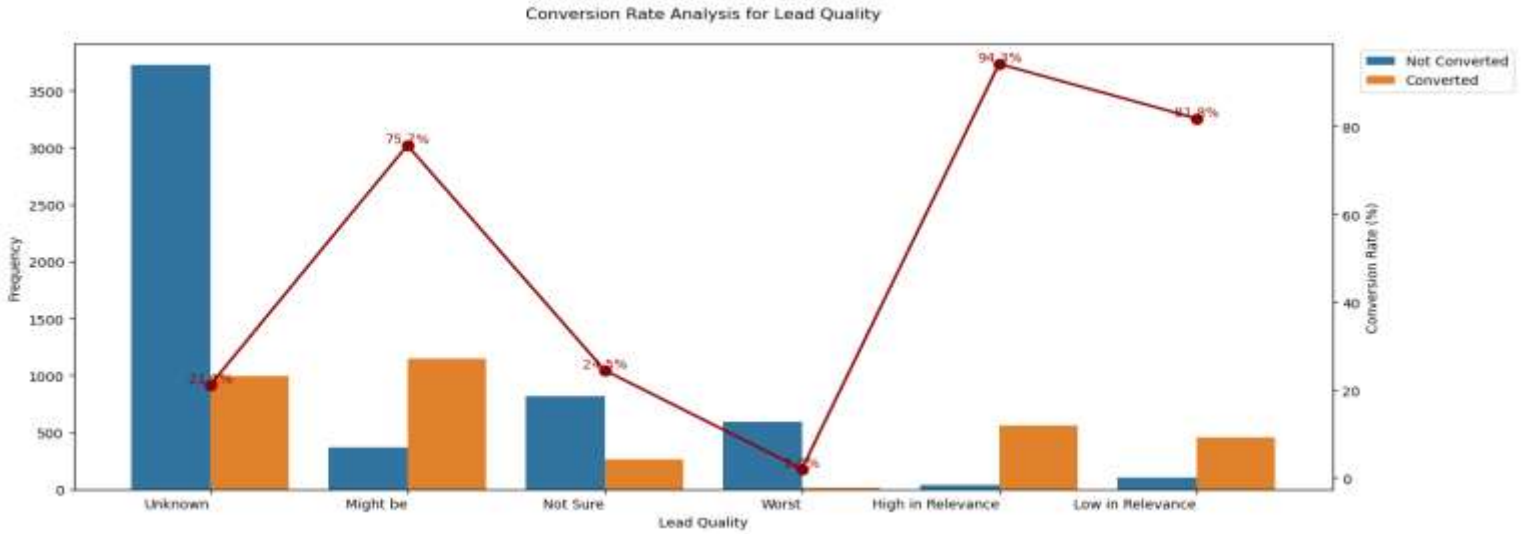
Correlation Heatmap

There is a **strong correlation** between **Total Visits** and **Page Views Per Visit**, indicating that users who visit more frequently also tend to explore more pages per session. Additionally, **Time Spent on Website** shows a **moderate correlation** with conversion, reinforcing its role as a **reliable predictor** of lead conversion.

Conversion Analysis

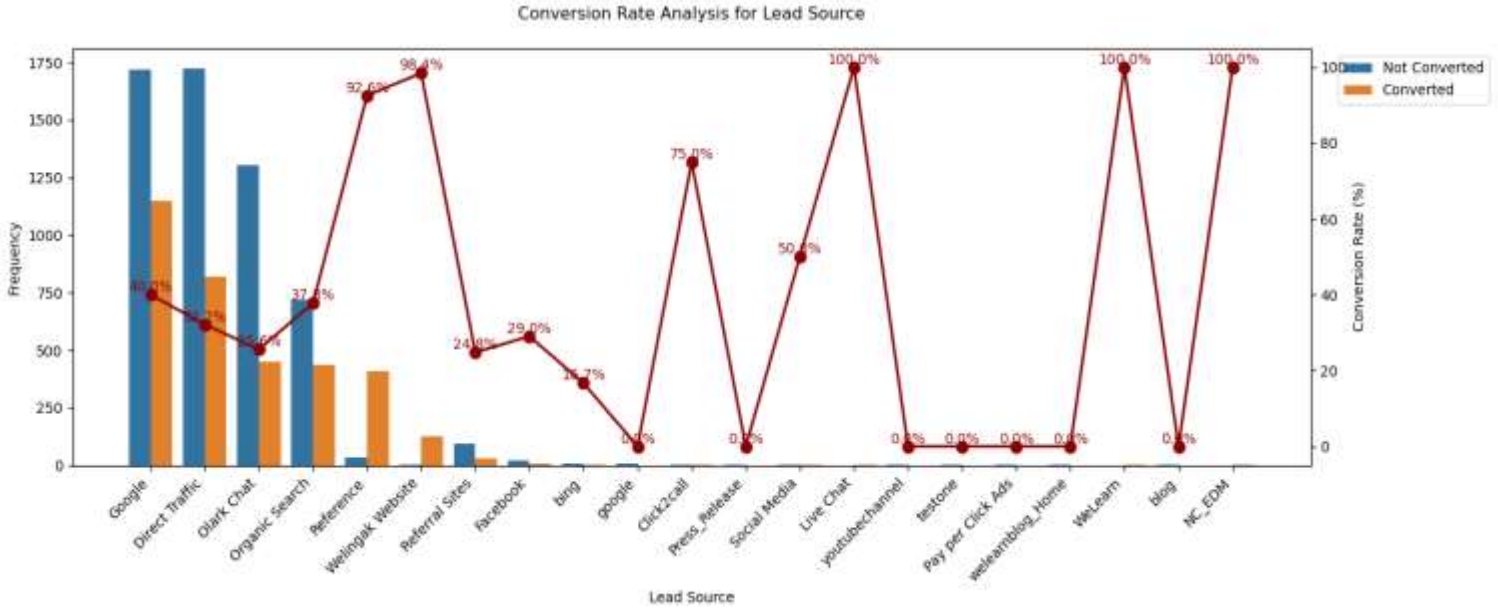
Lead Quality:

"High in Relevance" and "Low in Relevance" leads have the highest conversion rates (~94%) but low volume, while "Unknown" leads have high volume but low conversions, making them a key target for optimization. "Might be" leads offer a good balance of volume and conversion.



Lead Source:

Search, direct, and chat sources drive high traffic but low conversions, whereas referral marketing shows strong conversion rates due to higher trust. Digital marketing channels underperform, except for Social Media and Facebook, highlighting the need for strategic improvements.



Data Scaling & Splitting

The numerical features 'TotalVisits,' 'Page Views Per Visit,' and 'Total Time Spent on Website’ were scaled using MinMaxScaler, which normalizes each feature within a range of 0 to 1. This process reduces the impact of features with larger inherent scales and ensures a balanced contribution to the model. After scaling, the dataset was divided into training (80%) and testing (20%) subsets to train the model and evaluate its performance on the test data.

Feature Engineering & Transformation

Categorical features such as 'Tags,' 'Occupation,' and 'Asymmetrique Activity Index' were grouped to streamline categories and enhance interpretability. Columns with minimal informational value, including 'Country,' 'How did you hear about X Education,' 'What matters most to you in choosing a course,' and 'Lead Profile,' were removed. Binary categorical variables ('Yes/No') were transformed into numerical values (1/0), while the remaining categorical features were encoded using Label Encoding.

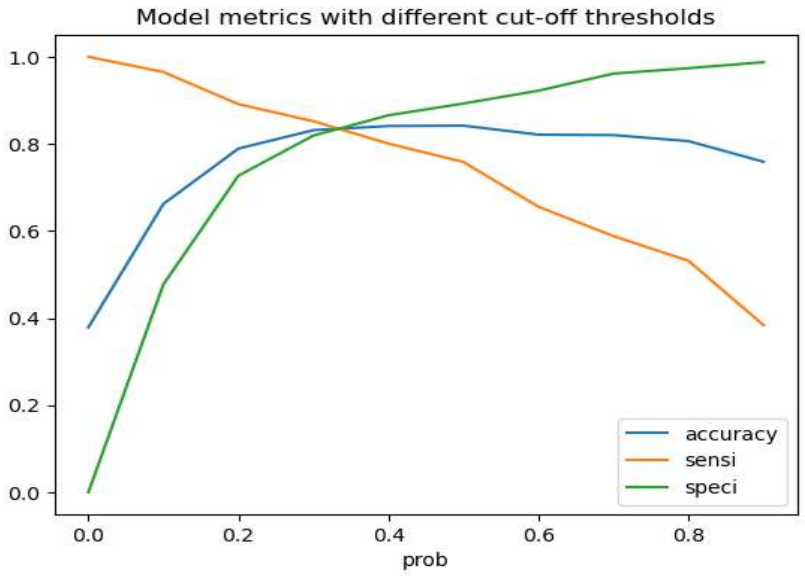
Feature Selection using Recursive Feature Elimination(RFE)

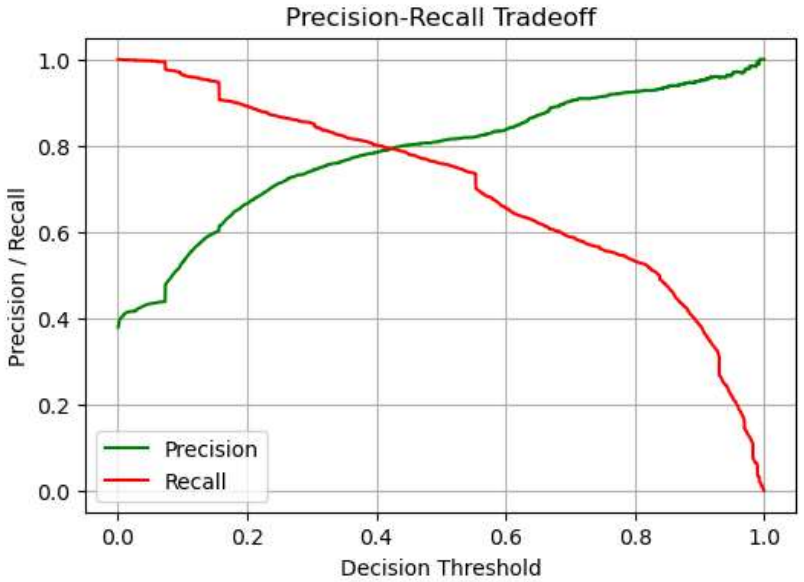
We utilized Recursive Feature Elimination (RFE) to systematically identify and rank features, ultimately selecting the top 15. This process highlighted key features such as Lead Origin, Lead Quality, Tags, and Total Time Spent on Website, which play a significant role in predicting lead conversion.

Iterative Model Building, RFE & VIF Analysis

- **Feature Selection:** Used Recursive Model Building (RFE) to choose the most important features.
- **Handling Multicollinearity:** Applied Variance Inflation Factor (VIF) analysis to remove highly correlated features like 'What is your current occupation' and 'Do Not Call.'
- **Key Features Identified:** Lead Origin, Lead Source, Do Not Email, Total Time Spent on Website, Tags, Lead Quality, Last Notable Activity, Page Views per visit.
- **Goal:** Improved model accuracy and stability.

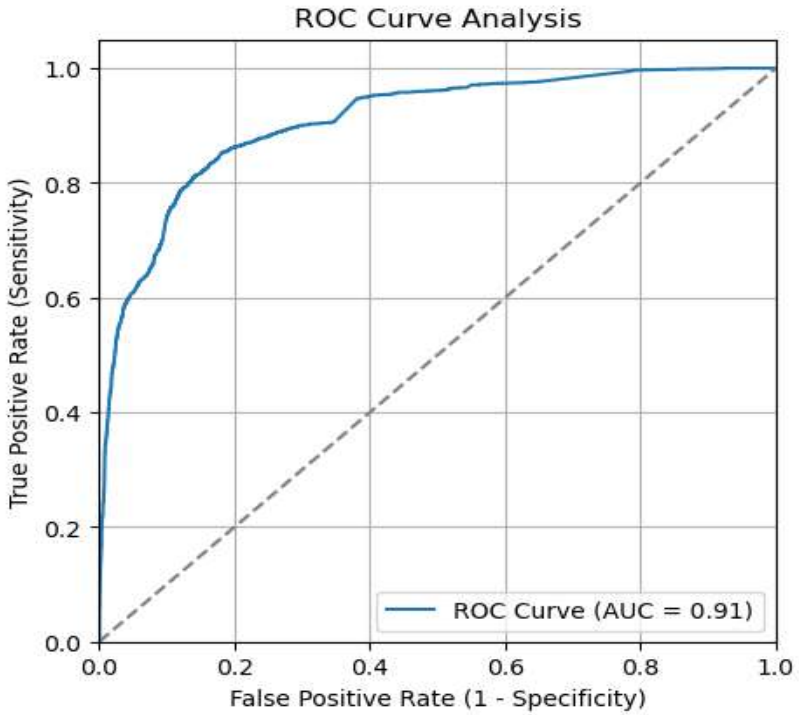
Optimal Cut-off & Confusion Matrix





With a 0.4 cut-off, the model effectively balances accurately identifying potential leads while reducing false positives.

ROC Curve



- **AUC = 0.91**, demonstrating strong discriminatory power.
- ROC Curve **positioned well above the diagonal**, indicating the model performs significantly better than random classification.

Top Features

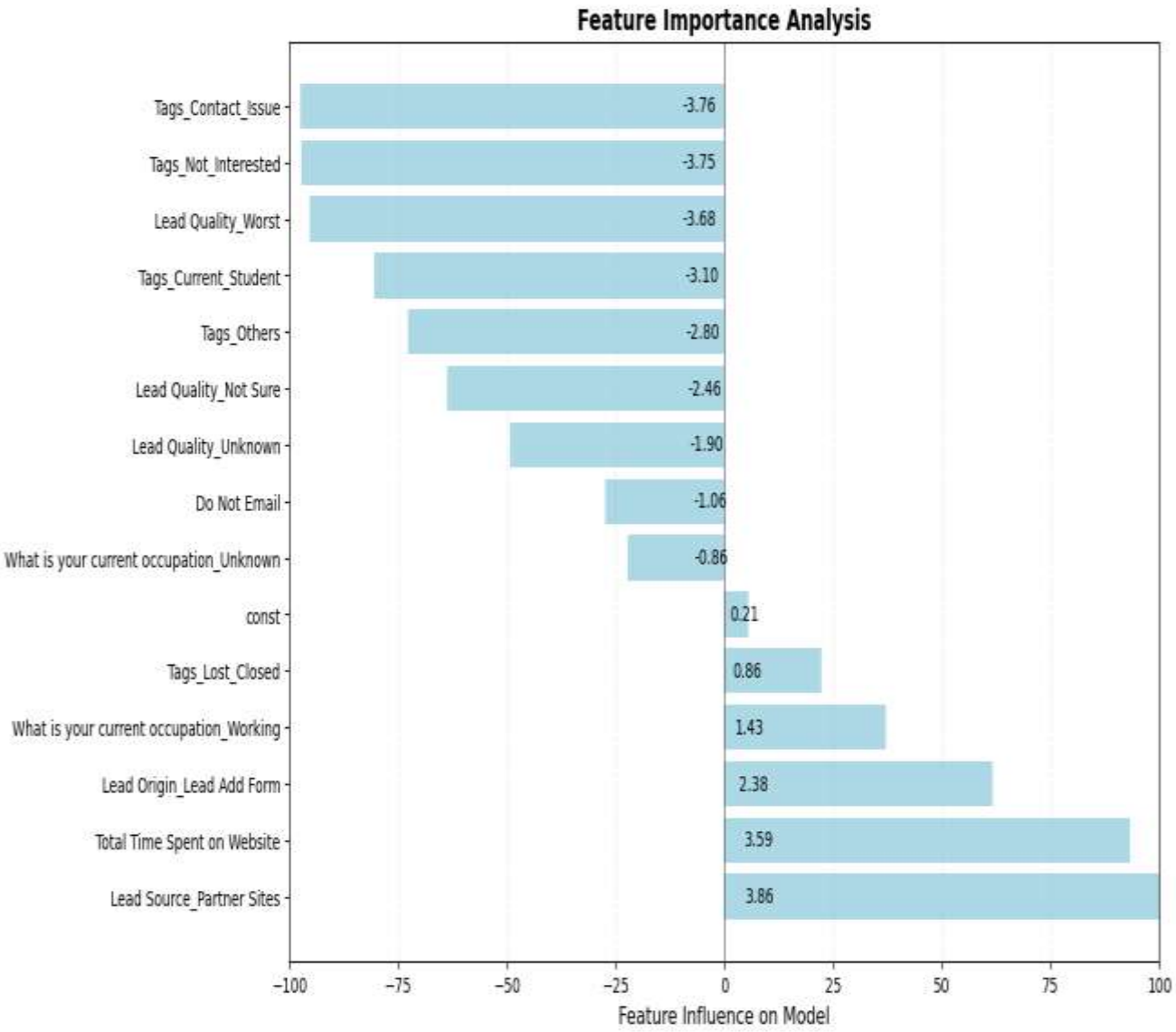
Key Predictors Impacting Conversion:

✅ **Positive Predictors (Increase Conversion Probability)**

- Lead Source_Partner Sites (3.86): Builds trust, boosting enrollments.
- Total Time Spent on Website (3.59): More time spent correlates with higher conversions.
- Lead Origin_Lead Add Form (2.38): Form submissions indicate strong interest.

❌**Negative Predictors (Decrease Conversion Probability)**

- Tags_Contact_Issue (-3.76): Leads with contact issues are unlikely to convert.
- Tags_Not_Interested (-3.75): Explicit disinterest lowers conversion chances.
- Lead Quality_Worst (-3.68): Poor lead quality reduces conversion.



Recommendations

- **Enhance Website Engagement:** Focus on improving the website experience to encourage longer visitor sessions and deeper content interaction, effectively capturing high-intent leads.
- **Improve Lead Quality Assessment:** Establish a standardized and refined process for evaluating lead quality to accurately identify and prioritize high-potential prospects for targeted outreach.
- **Maximize Key Lead Channels:** Direct marketing efforts and resources toward high-performing channels like Google and Direct Traffic to optimize lead acquisition.
- **Customize Lead Communication:** Adapt communication strategies and content based on lead behavior and specific interests to increase relevance and boost engagement.


THANK
YOU !