Lead Conversion Analysis Using Logistic Regression

Predicting Lead Conversion Likelihood with Data Insights

Adarsh M Shetty B

Arnab Biswas

Bhavya Jain

Problem Statement

- **Objective**:
  To predict the likelihood of lead conversion based on historical data, enabling efficient lead prioritization and better resource allocation.

- **Expected Key Challenges After Initial Data Analysis**:
  - Handling missing data in key features like Lead Quality and Last Activity.
  - Managing class imbalance with a conversion rate of 38.5%.
  - Identifying meaningful predictors from numerous features, including categorical variables.

```python
# 2. Data Loading
data = pd.read_csv('leads.csv')  # Replace with actual file path
data.head()
```

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetrique Profile Index	As A
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02.Medium	
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02.Medium	
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead	Mumbai	02.Medium	01.High	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select	Mumbai	02.Medium	01.High	
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select	Mumbai	02.Medium	01.High	

5 rows × 37 columns

Approach Overview

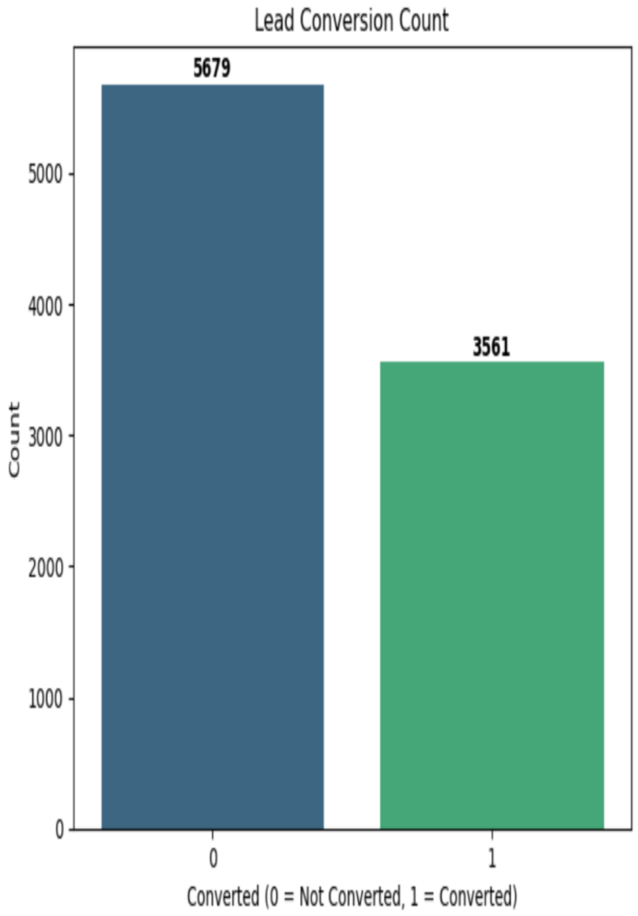
- **Data Understanding and Quality Checks**: Addressed missing values, ensured data consistency, and removed redundant columns.
- **Exploratory Data Analysis (EDA)**: Analyzed conversion trends and identified key predictors.
- **Data Preprocessing**: Imputed missing values, created dummy variables, and scaled numerical features.
- **Feature Engineering**: Selected features based on logistic regression coefficients and domain-specific insights.
- **Model Building and Evaluation**: Developed a logistic regression model and evaluated using metrics like accuracy, precision, recall, and ROC-AUC.
- **Lead Scoring**: Assigned a lead score (0-100) to each lead based on the predicted probability of conversion.

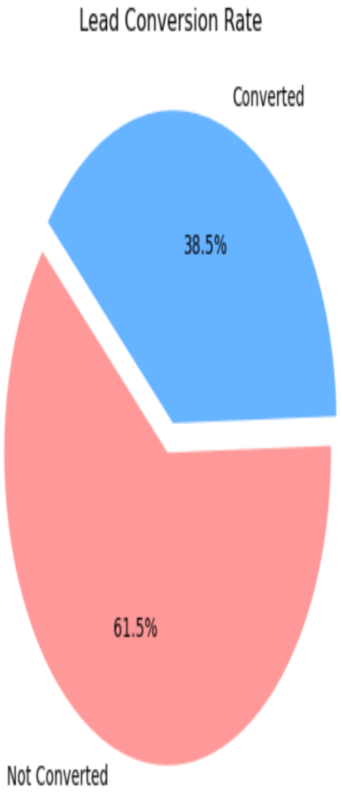
Data Preparation

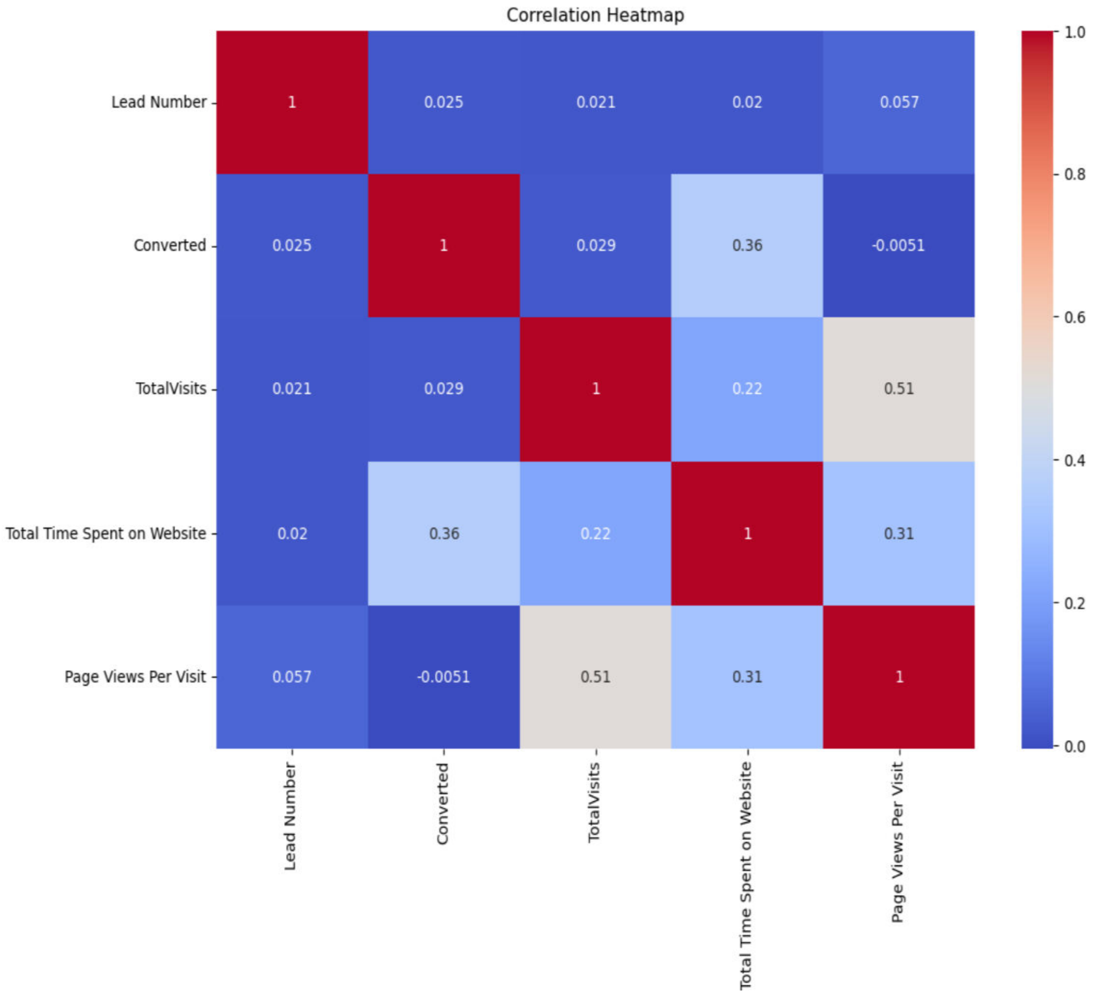
- **Missing Value Treatment**:
  - Lead Quality and Last Activity were imputed with mode and placeholder values, respectively.
  - Applied "Not Available" placeholders for categorical null values.

- **Feature Transformation:**
  - Dummy variables were created for categorical features like Lead Source and Tags.
  - Numerical columns like Total Time Spent on Website were scaled for model compatibility.

EDA Insights



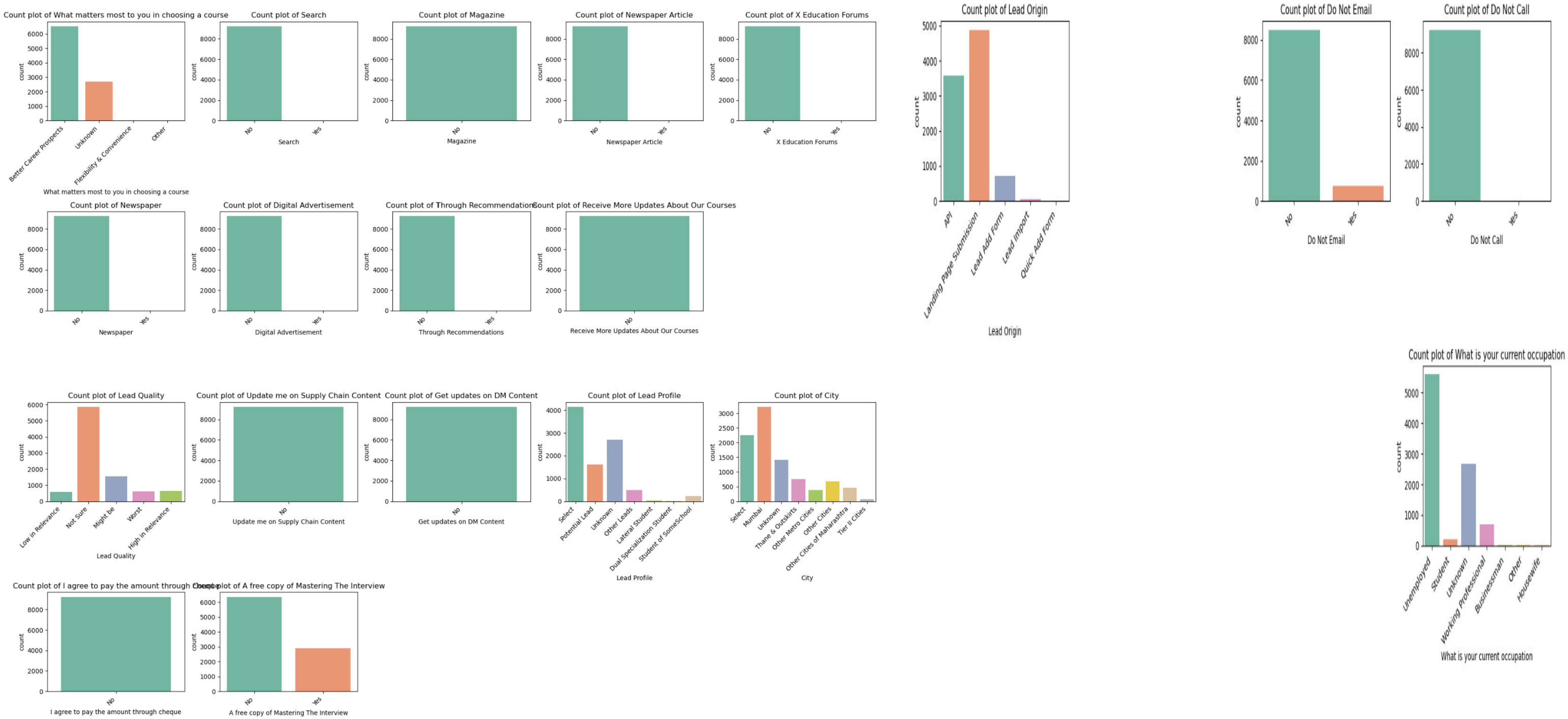




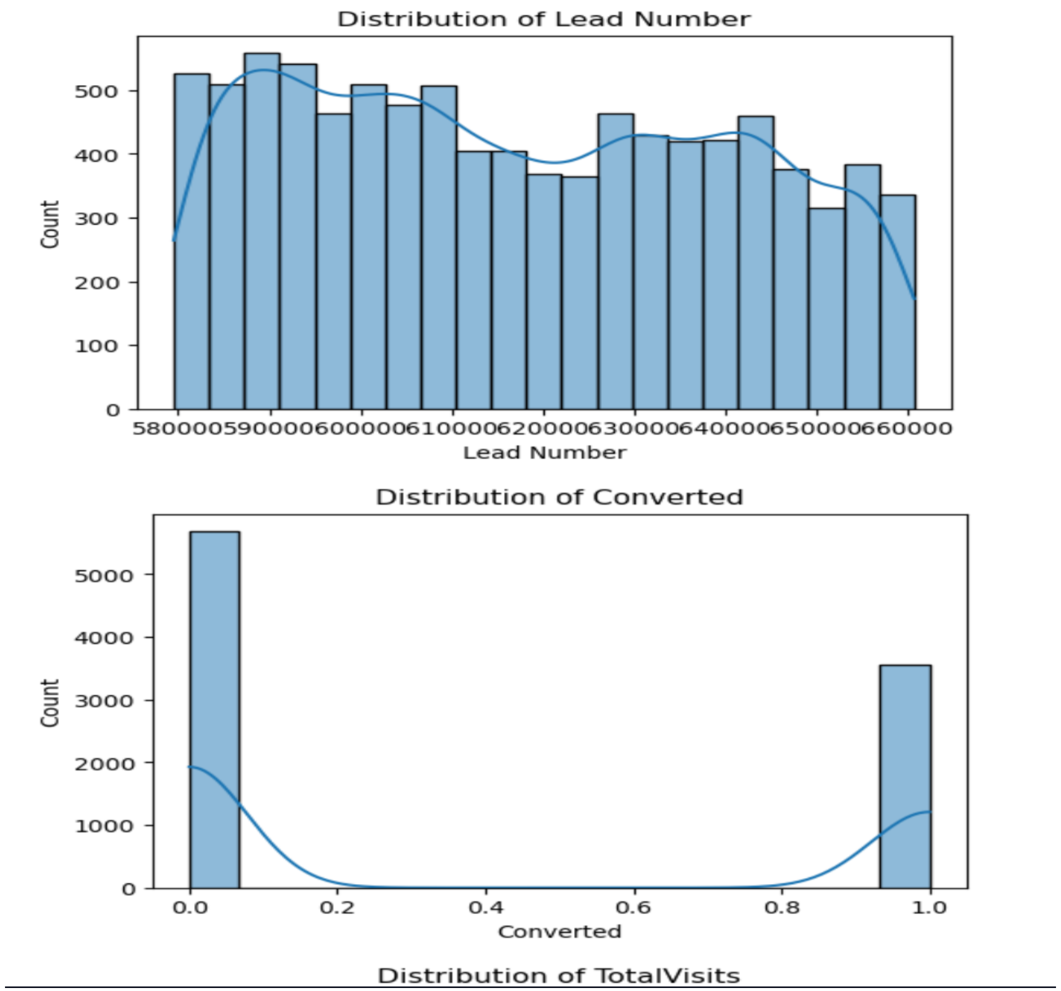
Conversion rate = 38.5% (3561 out of 9240 leads).

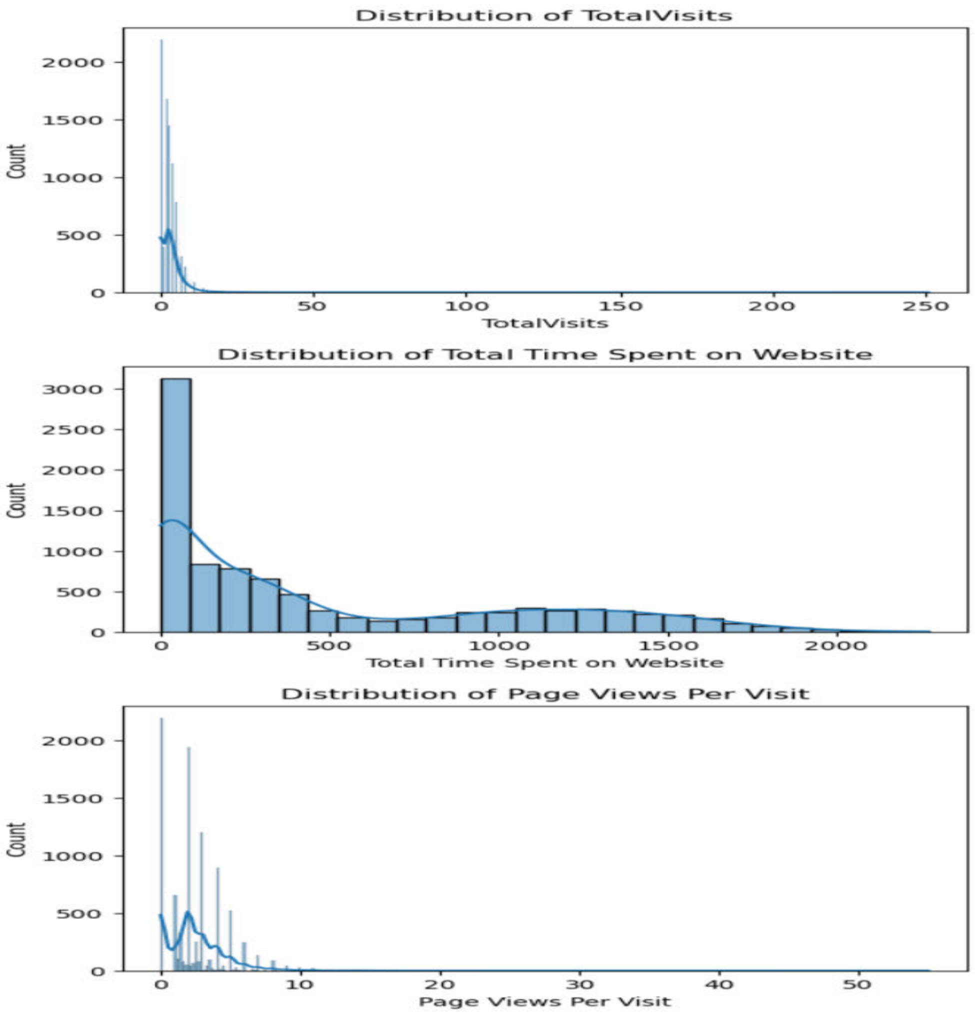
No significant linear correlation between selected features

Visualizations [Categorical Features]



Visualizations[Numerical Features]



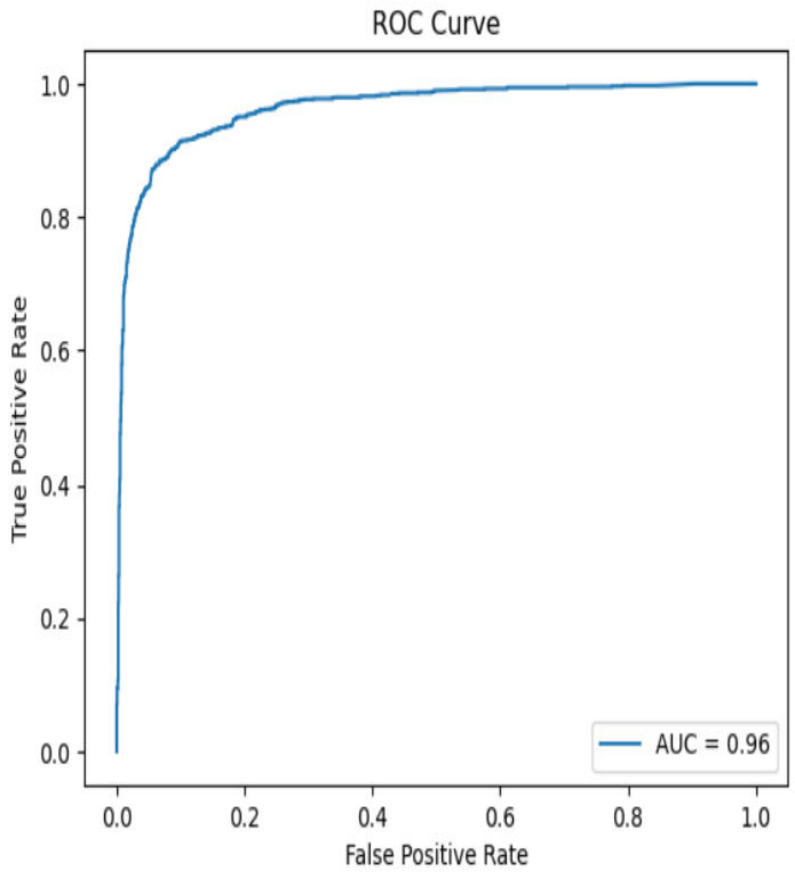


Model Development

- Logistic regression was chosen for interpretability and efficiency.
- Class imbalance addressed using weighted loss functions.
- Lead scores were computed as:
  - lead_scores = y_pred_prob * 100
  - data_test['Lead Score'] = lead_scores
- Thresholds adjusted for specific strategies:Aggressive conversion: Threshold = 0.3 (Targeted leads = 1123).
- Minimized calls: Threshold = 0.7 (Targeted leads = 967).

Model Evaluation

- **Confusion Matrix**:
  [[1571 133]
  [118 950]]
- **Classification Report**:
  - Accuracy: **91%**
  - Precision (Converted Leads): **88%**
  - Recall (Converted Leads): **89%**
  - F1 Score (Converted Leads): **88%**
- **ROC-AUC Score**: **0.958**



- **Threshold Impact (Customize on Need Basis)**:
  - Threshold = 0.3: Higher sensitivity for aggressive conversion.
  - Threshold = 0.7: Focused targeting with minimized waste.

Business Implications

1. **Lead Scoring**:
   1. Assigning lead scores (0-100) for prioritization.
   2. Example: Leads with scores >70 are "hot" and should be prioritized.

2. **Actionable Insights**:
   1. **Tags_Ringing**: Indicates less promising leads; adjust follow-up strategies.
   2. **Total Time Spent on Website**: Key metric for identifying engaged leads.

3. **Campaign Optimization**:
   1. Reallocate resources to effective lead sources and Tags categories.

Business Key Insights after using prediction Model

- **Top Predictors (Numerical)**:
  - ✓ **Total Time Spent on Website**: Positively correlated with conversion (Coefficient: 1.35).

- **Top Predictors (Categorical)**:
  - ✓ **Tags_Ringing**: Strong negative impact on conversion (Coefficient: -1.71).
  - ✓ **Tags_Will revert after reading the email**: Positive correlation with conversion (Coefficient: 1.59).
  - ✓ **Tags_Lost to EINS**: Positive impact (Coefficient: 0.93).

Conclusion and Recommendations

Summary of Findings:

- Logistic regression model achieved **91% accuracy** and **0.958 ROC-AUC**, providing actionable insights.
- Top predictors (numerical and categorical) were identified to guide lead prioritization.

Recommendations:

1. Focus on leads with high scores for conversion campaigns.
2. Adjust follow-up intensity based on thresholds (e.g., aggressive vs. focused strategy).
3. Regularly retrain the model with new data to adapt to changing trends.

Thank You