Lead Score Analysis

Capstone Project Summary - Anusha N

Problem Statement

- X Education, an online education platform targeting industry professionals, attracts a significant volume of website traffic through digital marketing, search engines, referrals, and social media. Visitors who show interest by filling out a form with their contact details are classified as *leads*.
- Despite acquiring a large number of leads, the company's lead-to-conversion rate is only **30%**, indicating inefficiencies in the lead nurturing process. Sales representatives engage with all leads via emails, phone calls, and other channels, but many of these efforts result in no conversions.

Business & Analytical Objective

- Business: Increase conversion rate to ~80% by identifying 'Hot Leads'.
- Analytical: Build interpretable, scalable logistic regression model for lead scoring.

Dataset Overview

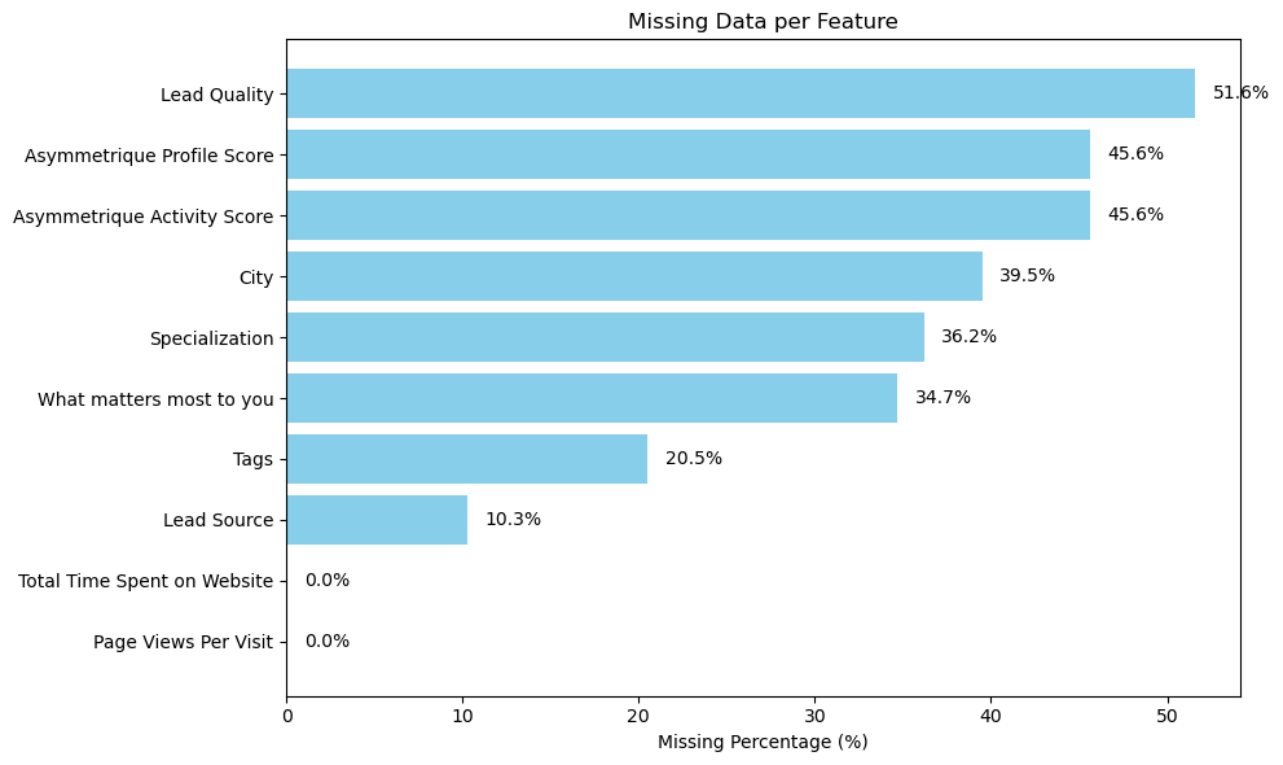
- ~9,000 past lead records
- Features include:
- **Target Variable**: Converted (1 = converted, 0 = not converted)

Category	Example Columns
Identifiers	Prospect ID, Lead Number (Dropped during cleaning)
Lead Origin	Lead Origin, Lead Source
Demographics	City, Country, Do Not Email, Do Not Call
Course Info	What is your current occupation?, Specialization
Engagement	Total Time Spent on Website, Page Views Per Visit
Marketing	Last Activity, Last Notable Activity, Tags
Communication	Email Opened, SMS Sent, Asymmetrique Activity Score
Target Variable	Converted (1 = converted to customer, 0 = not converted)

Missing Data Handling

- **Dropped High-Missing Columns**:
  - Columns like Lead Quality, Asymmetrique Activity Score, and Asymmetrique Profile Index had **>45% missing values** and were **dropped** due to low informational value.
- ◇ **Categorical Placeholders Treated as Missing**:
  - Placeholder entries like **"Select"** in columns such as Specialization, City, and What matters most... were considered missing and handled appropriately.
- ◇ **Imputation Strategy**:
  - For **categorical columns**: Missing values were filled using **mode** or set to **"Unknown"** (e.g., City, Specialization).
  - For **numerical columns**: Median imputation was applied to maintain robustness against outliers (e.g., Total Time Spent on Website).
- ◇ **Redundant Identifiers Removed**:
  - Prospect ID and Lead Number were dropped since they served only as unique keys and had **no predictive relevance**.

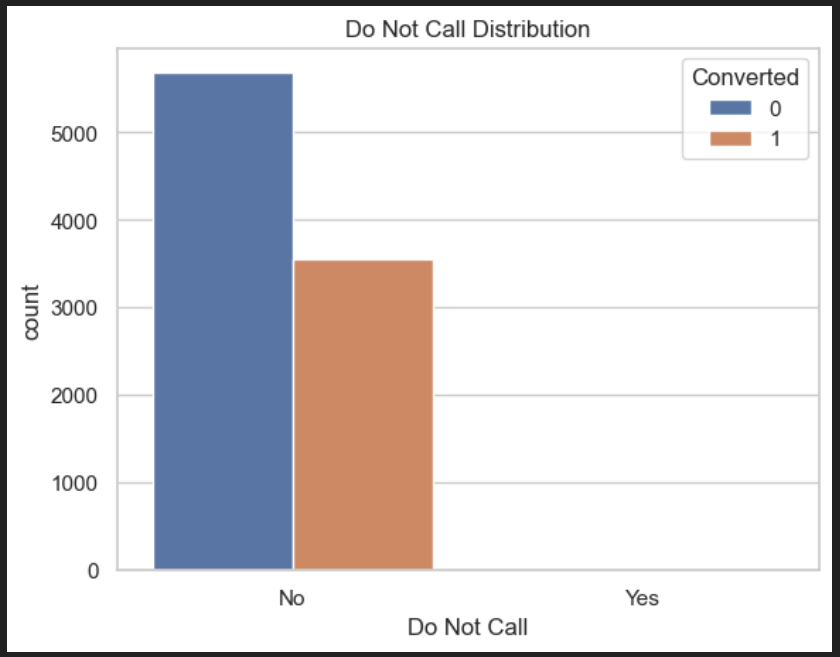
Visualising missing data

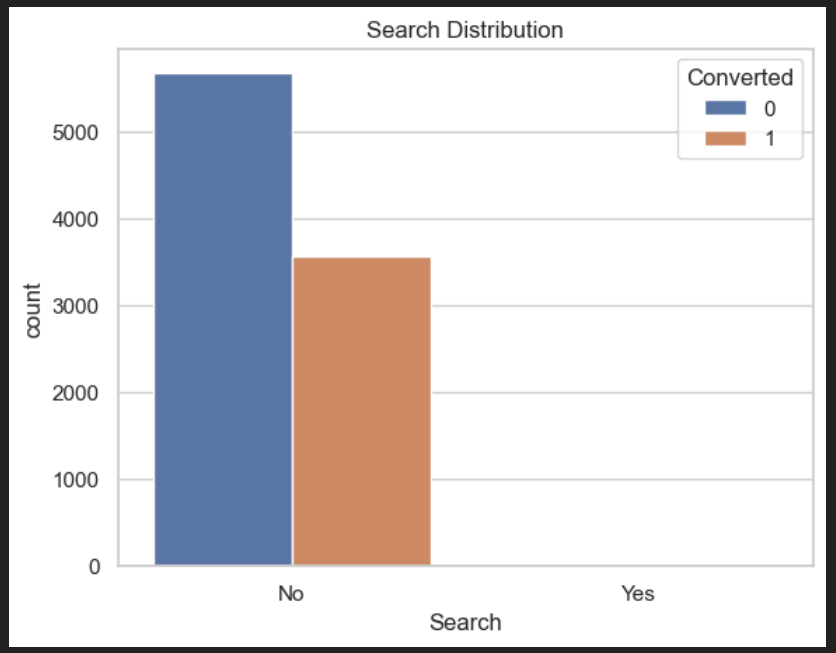


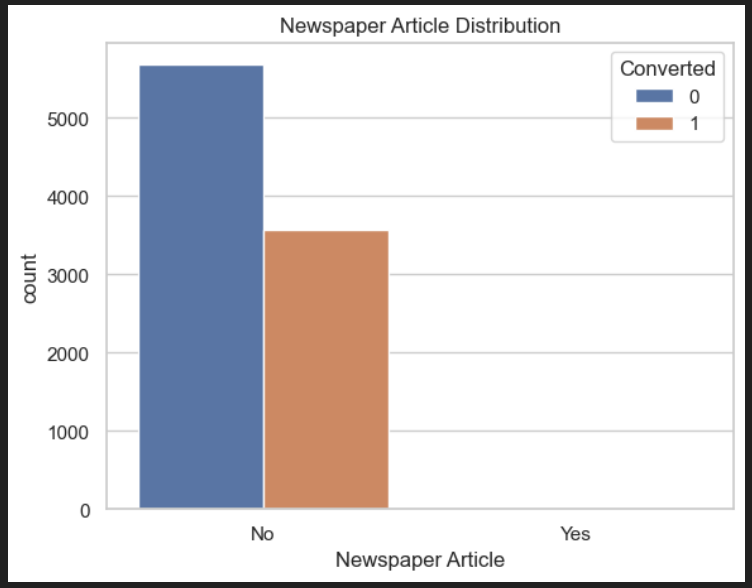
Data Cleaning Highlights

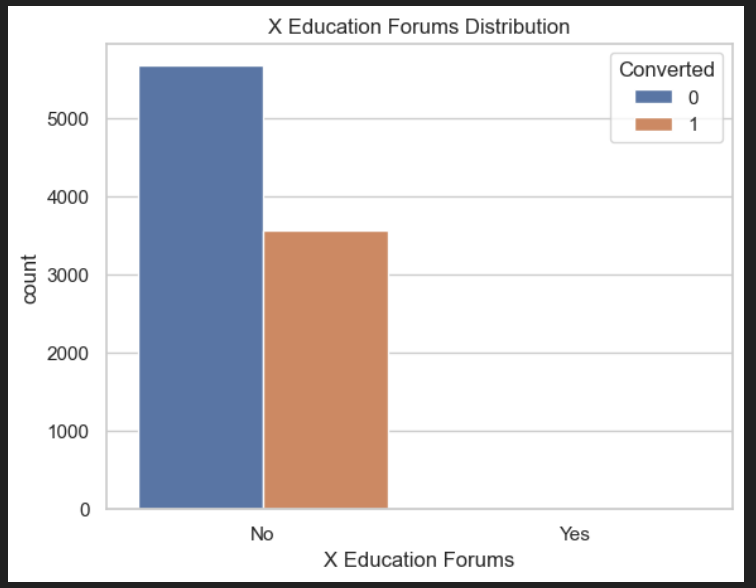
- **Dropped Low-Variance Features**:

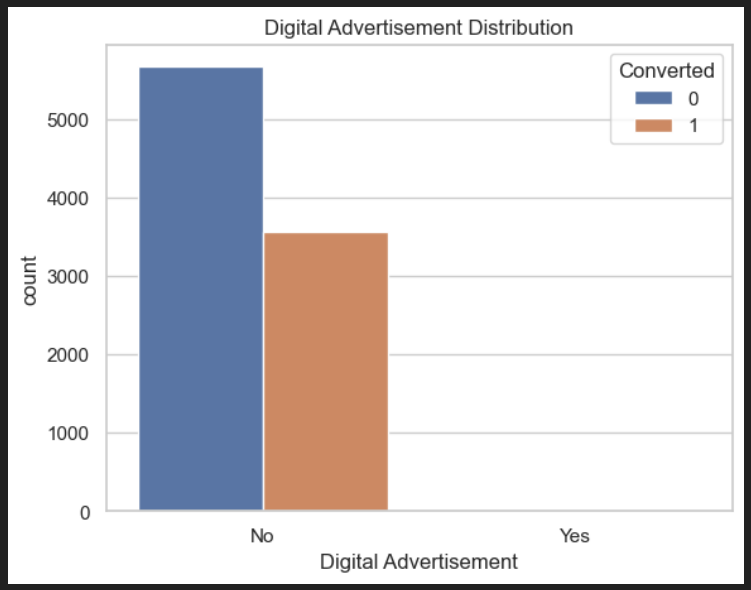
•Columns like Newspaper Article, A free copy of Master Program, Digital Advertisement were dropped as over **95% values were "No"**, offering **no useful variance**.





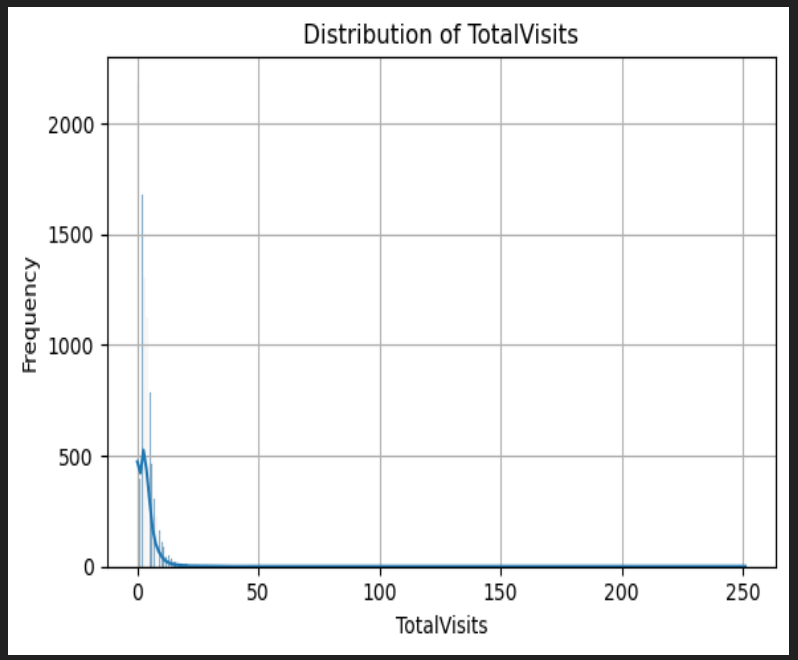

Newspaper Article Distribution
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Newspaper Article

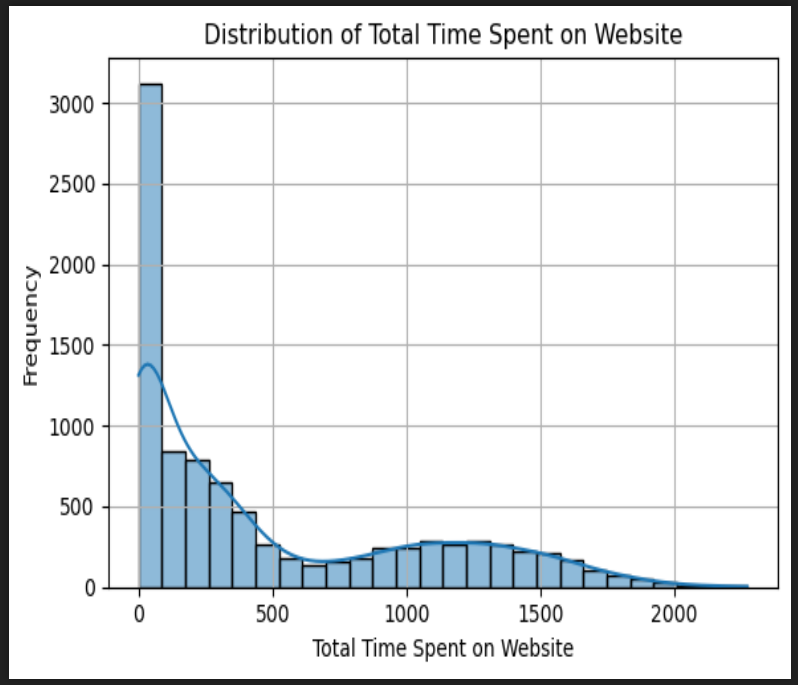

X Education Forums Distribution
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
X Education Forums


Digital Advertisement Distribution
Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Digital Advertisement

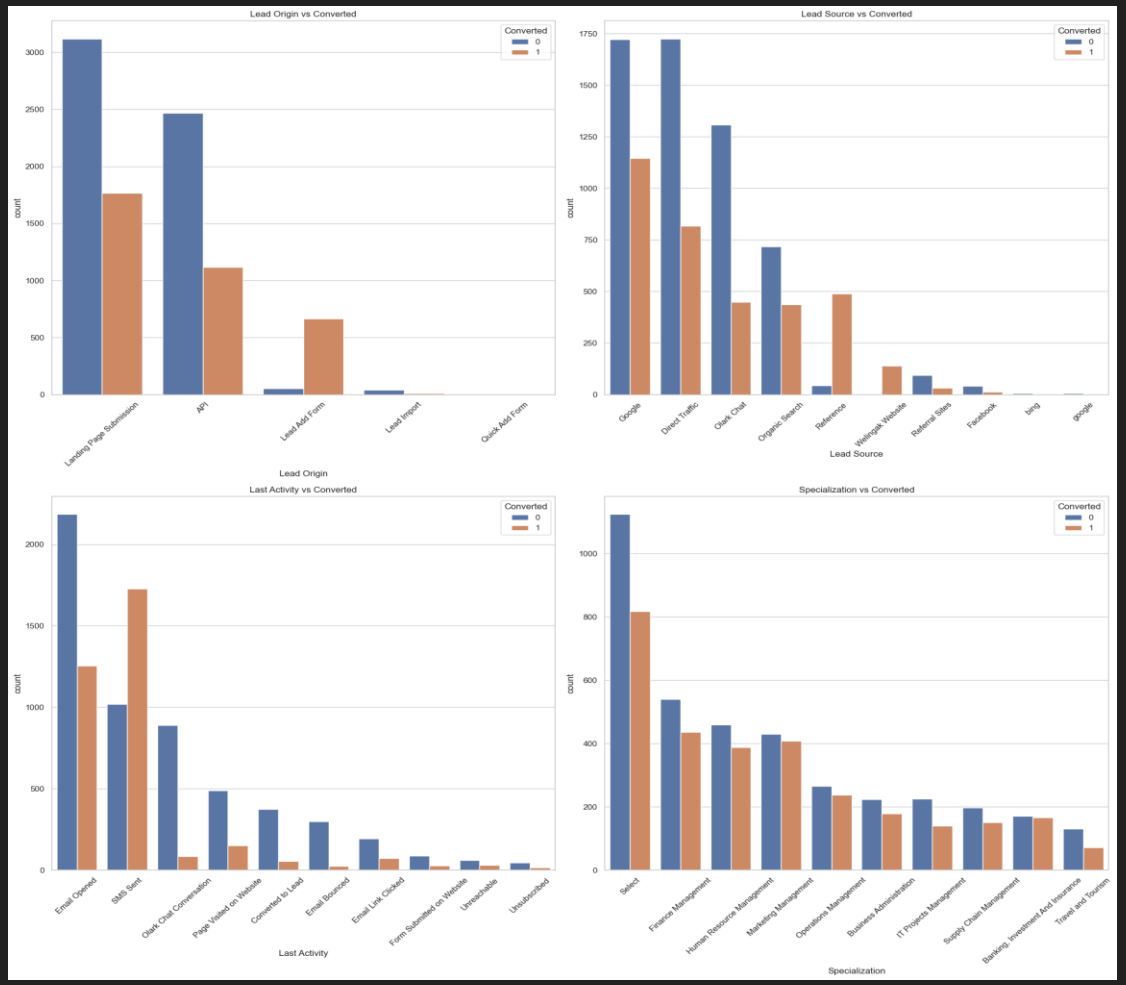
Univariate & Outlier Analysis

- Highly skewed numeric features capped using IQR method.
- Created new features: Engagement Score, Visit-Page Ratio.



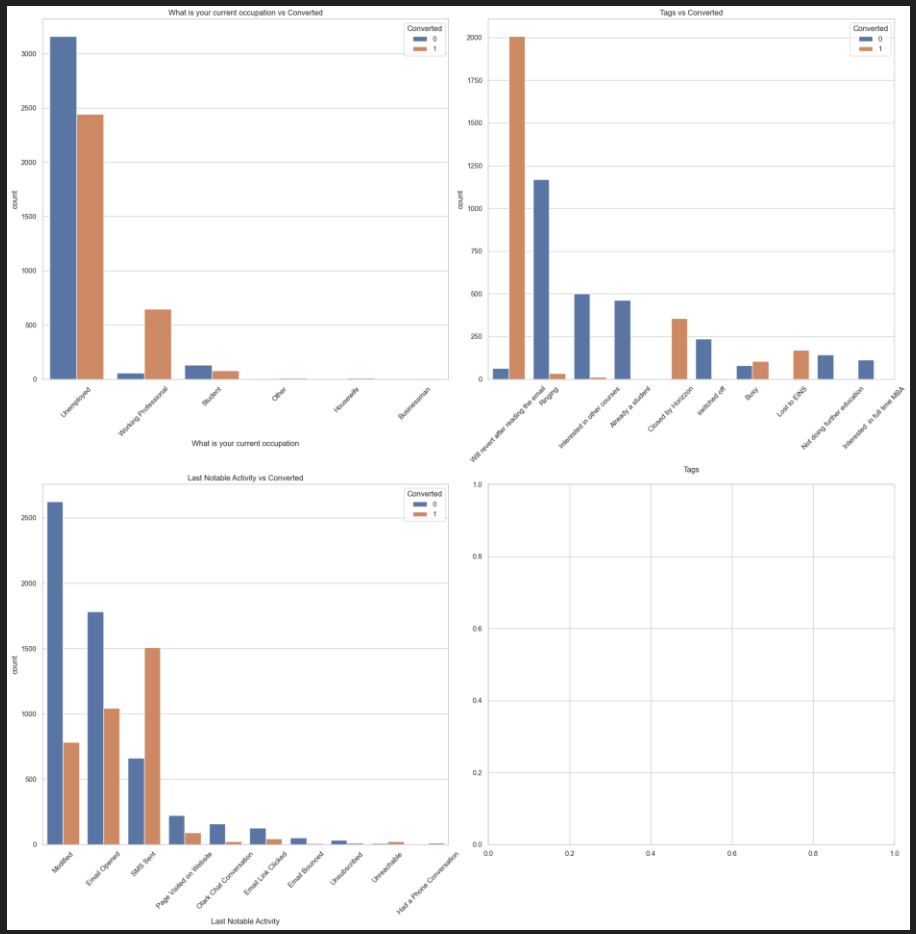


Categorical column Analysis



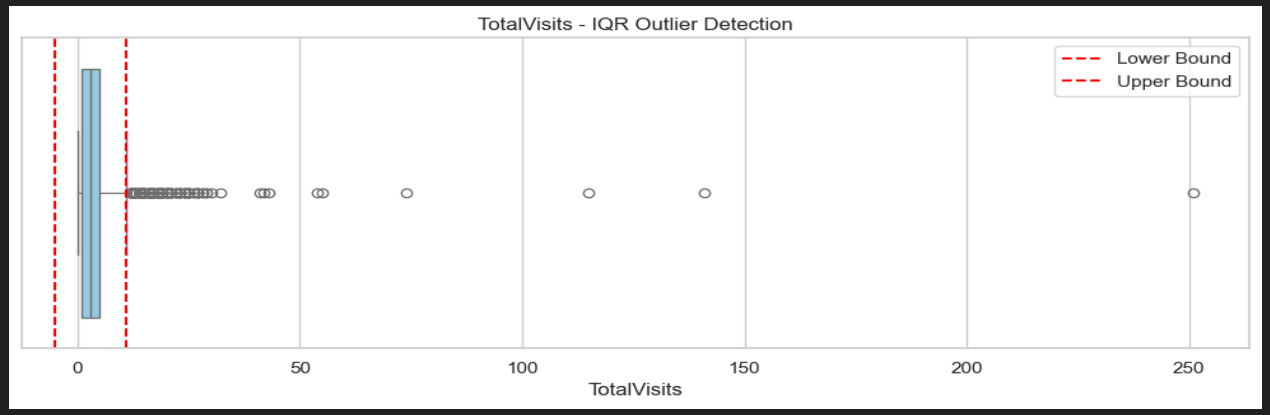
- **Lead Origin**:
  - *Landing Page Submission* and *API* bring in the highest number of leads.
  - These also show **higher conversion rates**, indicating effective acquisition channels.
- **Lead Source**:
  - *Google*, *Direct Traffic*, *Olark Chat*, *Organic Search*, and *Referral Sites* are the top traffic sources.
  - Among them, **Google**, **Referral Sites**, and **Direct Traffic** contribute to **higher conversions**.
- **Last Activity**:
  - Leads who **opened emails** or **received SMS** are more likely to convert.
  - These activities indicate higher engagement and intent.
- **Specialization**:
  - *Finance Management* has the highest number of leads.
  - Could be due to broader interest or better campaign targeting

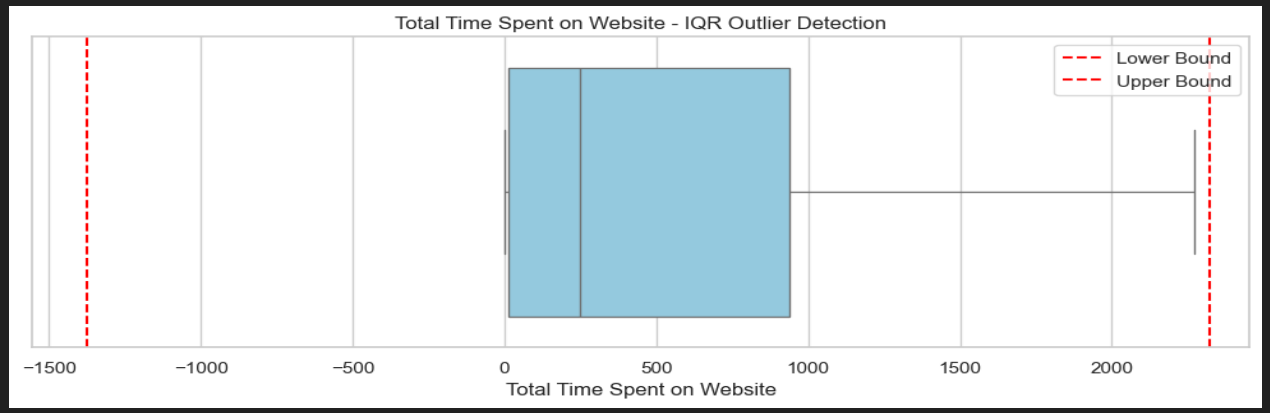
Categorical column Analysis

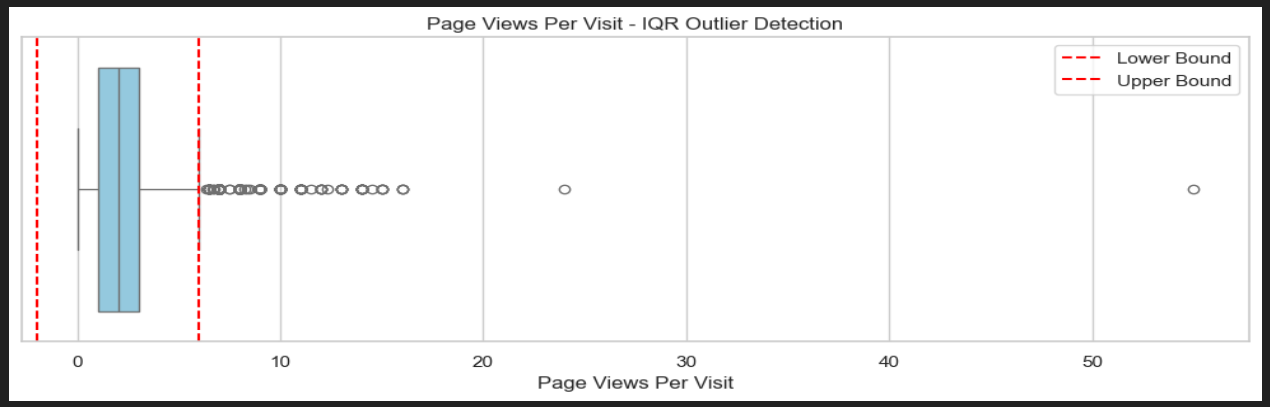


- **What is your current occupation**:
  - Most leads are **working professionals**, and they show **higher interest** in courses.
- **Tags**:
  - The tag **"Will revert after reading the email"** is strongly associated with conversions.
  - Indicates positive intent and email-driven engagement.
- **Last Notable Activity**:
  - *SMS Sent* appears most frequently as the last touchpoint before conversion.
  - May act as a final nudge to convert leads.

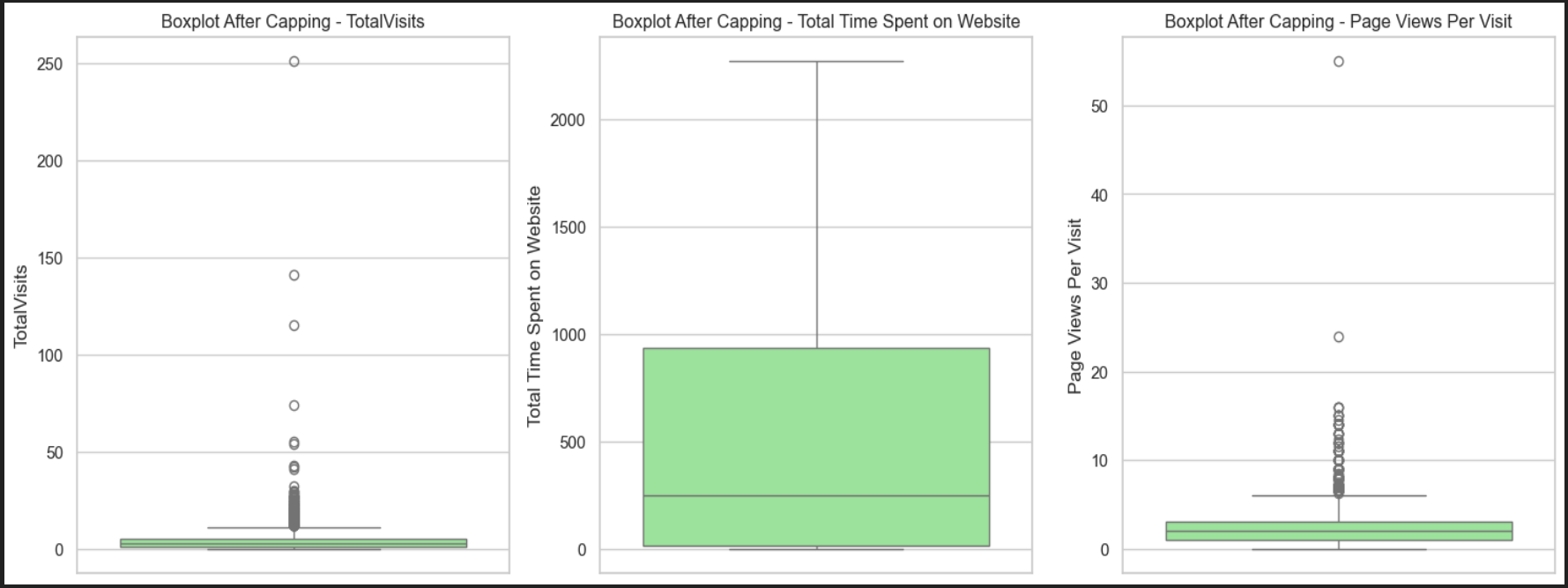
Outlier Analysis







After handling outliers with IQR capping



Feature Engineering

- **Engagement Score**:
  - Combined three related metrics:

    Engagement Score = Total Time Spent on Website (in minutes) + TotalVisits + Page Views Per Visit
  - **Why?**: Individually these features give partial insight. Together, they form a comprehensive picture of how engaged a lead is.
- **Visit/Page Ratio**:
  - Visit_Page_Ratio = TotalVisits / Page Views Per Visit
  - **Why?**: This reveals user navigation depth—whether they browse deeply in fewer visits or visit often but skim.

Scaling

. Standardization using StandardScaler

- Applied to:
  - TotalVisits, Total Time Spent on Website, Page Views Per Visit, and derived features like Engagement Score, Visit_Page_Ratio.
- **StandardScaler** Where μ = mean and σ = standard deviation.
- ☑ **Why?**
  - Logistic regression is sensitive to feature scale.
  - Prevents features with large magnitude from dominating the model.
  - Ensures faster and more stable convergence during training.

Encoding

- **Binary Encoding**

Converted Yes/No fields (e.g., Do Not Email, SMS Sent) to **1/0** for model compatibility.

- **One-Hot Encoding**

Applied to categorical variables like Lead Source, Specialization, Tags to avoid ordinal bias.

- **Handled Unknown Values**

Replaced missing or "Select" entries with **“Unknown”** or most frequent category (**mode**).

- **Reduced Noise from Rare Categories**

Grouped low-frequency values into "Other" to prevent overfitting.

Modeling Approach

- ☑ **Train-Test Split**

Used an **80:20 stratified split** to preserve the class distribution of the Converted target variable.

- ☑ **Feature Selection with RFE**

Applied **Recursive Feature Elimination (RFE)** with Logistic Regression to select the **top 15 predictive features**.

- ☑ **Model Refinement**

Used statsmodels to examine **p-values** and removed statistically insignificant features iteratively.

- ☑ **Final Model: Logistic Regression**

Chosen for its **interpretability**, **scalability**, and alignment with business requirements.

- ☑ **Model is Business-Ready**

Easy to explain using feature coefficients and odds ratios; suitable for non-technical stakeholders.

Dropping Features Based on p-values

•After selecting top features using **RFE**, you fitted the model with statsmodels.GLM and checked the **p-values and standard errors**.
•The following features had:

- •Extremely **high p-values (≈0.999)**
- •**Very large standard errors**, indicating instability

Dropped Features:
•Tags_number not provided
•Tags_wrong number given
•Lead Origin_Lead Add Form (in later refinement)

These were removed because:
•Their **coefficients were unreliable** (e.g., -22.49 with standard error ~18,500)
•**p-values ≈ 0.999** indicate no predictive power
•Could **distort the model** or introduce noise

Result After Dropping:

- Final model retained **12 meaningful features**
- All retained features had **$p < 0.05$**, making them statistically significant
- Improved **pseudo $R^2 \approx 0.591$** and stable performance on both training and test sets

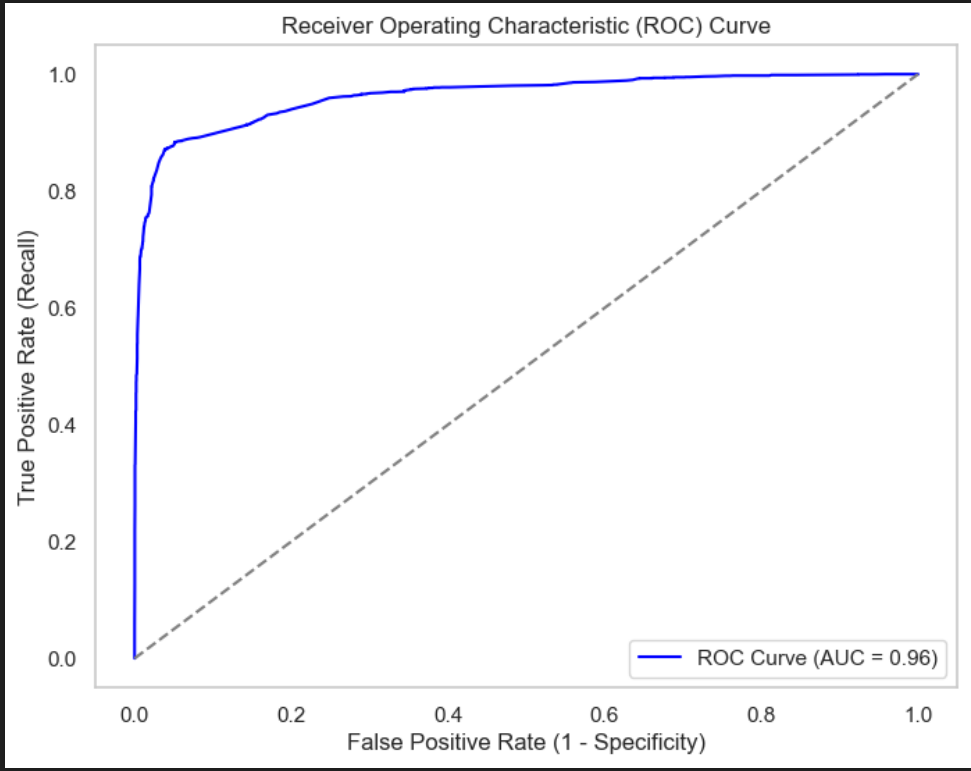
Multicollinearity Check – VIF Analysis

Feature	VIF Score
const	10.41 ☑ (Intercept, high but expected)
Tags_Will revert after reading the email	2.11 ☑
Tags_Unknown	2.09 ☑
Tags_Ringing	1.65 ☑
Tags_Closed by Horizzon	1.20 ☑
Last Activity_SMS Sent	1.16 ☑
Last Notable Activity_Modified	1.15 ☑
Tags_switched off	1.14 ☑
Tags_Busy	1.13 ☑
Tags_Lost to EINS	1.09 ☑
Visit_Page_Ratio	1.07 ☑
Lead Source_Welingak Website	1.05 ☑
Tags_invalid number	1.05 ☑

☑ Interpretation & Conclusion

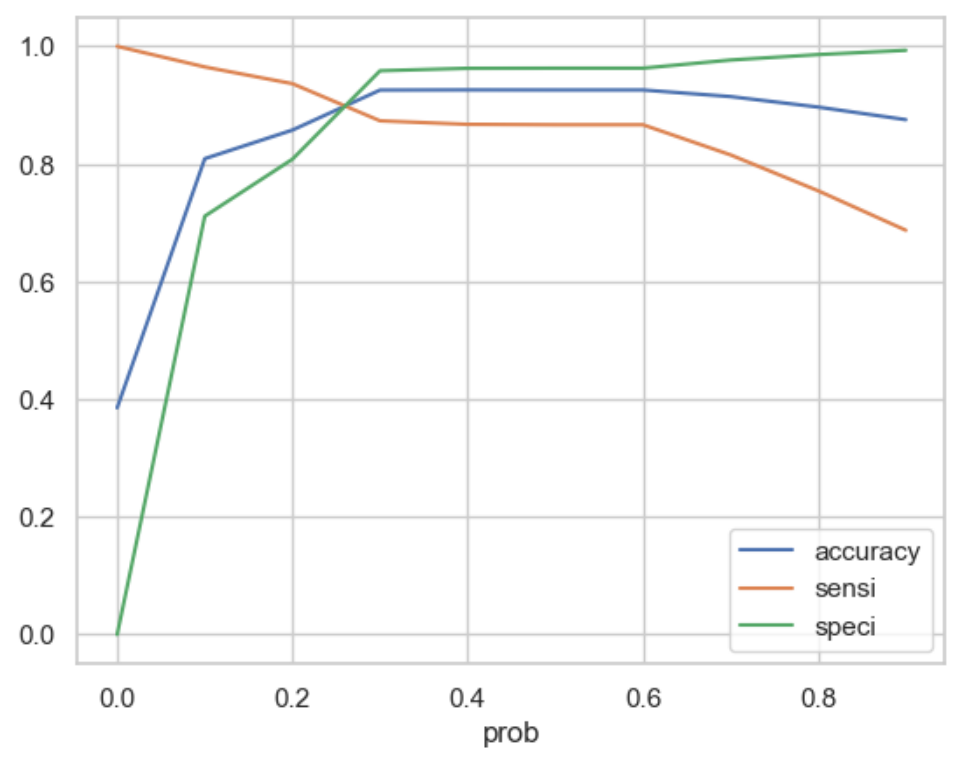
- **All feature VIFs are < 2.2**, indicating:
  - **No significant multicollinearity** exists among the variables.
  - Model coefficients are **stable** and **independently meaningful**.
- **Intercept (const)** has a high VIF by design — it is **not a concern**.

ROC CURVE



ROC =0.96 is close to 1. which is a very good value

Probability cutoff curve



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Final Logistic Regression Model

Feature Name	Type	Interpretation
Tags_Will revert after reading the email	Categorical (1/0)	Strong positive indicator of conversion
Tags_Unknown	Categorical (1/0)	Moderate positive contribution
Tags_Ringing	Categorical (1/0)	Indicates potential interest
Tags_Closed by Horizzon	Categorical (1/0)	Positive correlation
Tags_switched off	Categorical (1/0)	Slight positive relation
Tags_Busy	Categorical (1/0)	Neutral/slightly positive effect
Tags_Lost to EINS	Categorical (1/0)	Negatively correlated with conversion
Tags_invalid number	Categorical (1/0)	Negative contribution
Last Activity_SMS Sent	Categorical (1/0)	Strong positive impact
Last Notable Activity_Modified	Categorical (1/0)	Positive signal of engagement
Lead Source_Welingak Website	Categorical (1/0)	High intent lead source
Visit_Page_Ratio	Numerical (Scaled)	Indicates user's browsing depth

Model Evaluation

- **Model Performance**
- **Technique**: Logistic Regression (with RFE and p-value refinement)
- **Pseudo R^2**: ~**0.591** – good explanatory power for classification
- **VIF Scores**: All **< 2.2** – no multicollinearity detected
- **Business Interpretability**: Model outputs are clear and actionable
- Training Set Performance (at 0.3 cutoff) Accuracy : 0.9257
- Sensitivity : 0.8666 (High Recall)
- Specificity : 0.9628
- Precision : 0.9359
- NPV : 0.9200
- Test Set Performance (at 0.3 cutoff) Accuracy : 0.9210
- Sensitivity : 0.8722
- Specificity : 0.9516
- Precision : 0.9186
- NPV : 0.9224

Business Recommendations

- Focus marketing on Google, Direct Traffic, Referral Sites.
- Use SMS and Email campaigns more actively.
- Prioritize working professionals and Finance specialization leads.

Key Takeaways

- Data cleaning and proper encoding significantly impact model performance.
- Tags and activity-based features are highly predictive.
- Simple logistic regression can deliver high business value when engineered well.