# Lead Scoring Case Study

LEAD CONVERSION ANALYSIS

# Introduction

X Education, an online education company, provides courses for industry professionals through digital platforms. It attracts leads via websites, search engines like Google, and referrals. However, the company faces a low lead conversion rate of around 30%, despite acquiring numerous leads daily. To address this, X Education seeks a predictive model to identify high-potential leads or 'Hot Leads,' helping the sales team focus their efforts effectively. The goal is to improve the lead conversion rate to 80% by assigning a lead score between 0 and 100, indicating the likelihood of conversion.

# Business Understanding

X Education aims to enhance its lead conversion rate by identifying high-potential leads from its pool of prospects. The current conversion rate of 30% indicates inefficiencies in targeting the right leads. By building a predictive model to assign lead scores, the company seeks to optimize its sales efforts, prioritize promising leads, and achieve a target conversion rate of 80%, ultimately improving customer acquisition and overall business performance

# Problem Statement

The primary objective is to build a predictive model to assign lead scores between 0 and 100, enabling X Education to identify Hot Leads leads likely to convert into paying customers. This will help optimize sales efforts, improve lead conversion rates to the target of 80%, and enhance the efficiency of the sales funnel

If Hot Leads are successfully identified then the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

# Data Understanding and Preprocessing

## 1. Data Cleaning

- ❑ Removed columns with more than 30% null values
- ❑ Removed all the rows for the columns where the null values are less than 2%
- ❑ Replaced “Select” values with null. Also, if in case same is appearing while creating the dummy variables then dropping those specific columns
- ❑ Imputed missing values using median for numerical features and mode for categorical features
- ❑ Handling outliers: Removed extreme values which are above the **upper whisker** for numerical variables **'Total Visits'** and **'Page Views Per Visit'** to improve model accuracy

# Data Understanding and Preprocessing

## 2. Feature Engineering

- Grouped similar categories in **'Last Activity'** into broader categories: **'Email Activity'**: Combined all email-related actions (e.g., 'Email Opened,' 'Email Link Clicked')
- **'Website Activity'**: Merged website-related activities like 'Page Visited on Website
- **'Direct Communication'**: Unified direct interactions (e.g., 'Phone Conversation,' 'SMS Sent')
- **'Unreachable'**: Consolidated activities like 'Visited Booth in Tradeshow' and 'Unreachable

# Data Understanding and Preprocessing

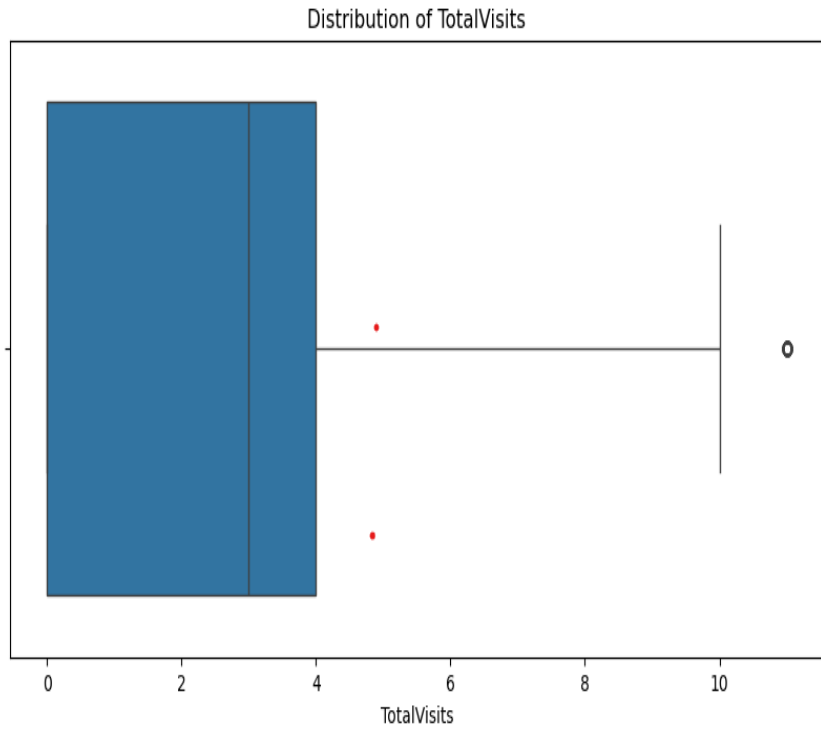
## 3. Final Data Validation

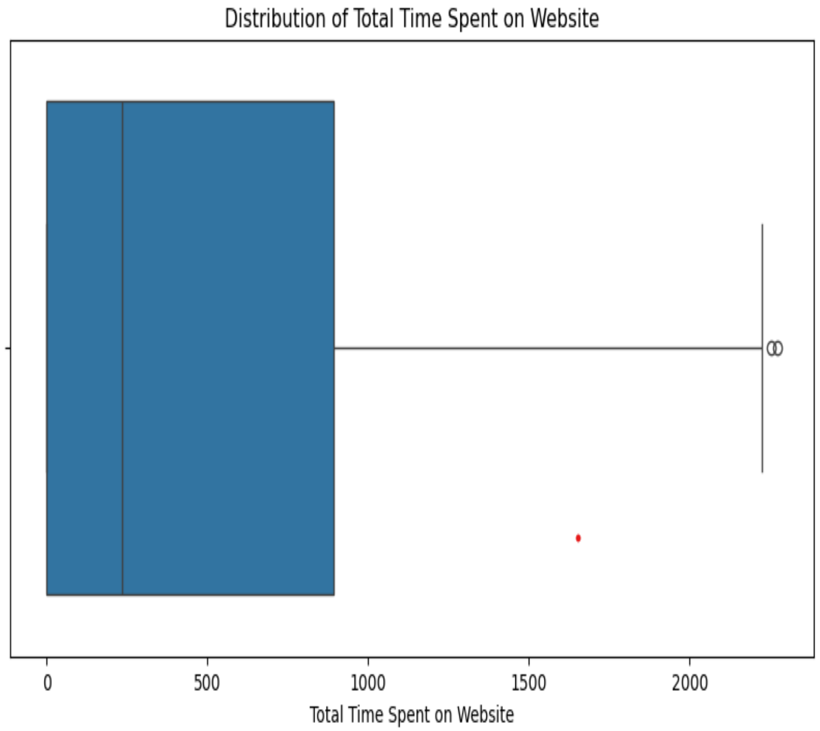
- Verified data transformations, category distributions and prepared the dataset for modeling
- Ensured no redundancy or inconsistencies in the final dataset

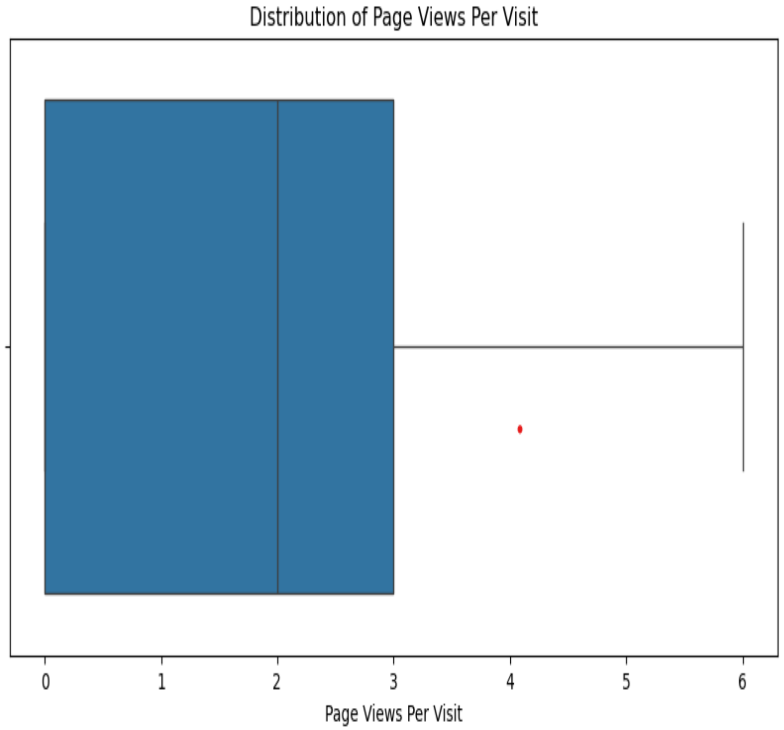
# Exploratory Data Analysis (EDA)

Conducted univariate analysis to understand key features like 'Total Visits' and 'Total Time Spent on Website' and their distributions. Bivariate analysis revealed significant predictors of conversion, such as higher time spent on the website correlating with increased conversion likelihood. A correlation matrix was used to identify relationships among numerical variables, highlighting key patterns and addressing potential multicollinearity issues.

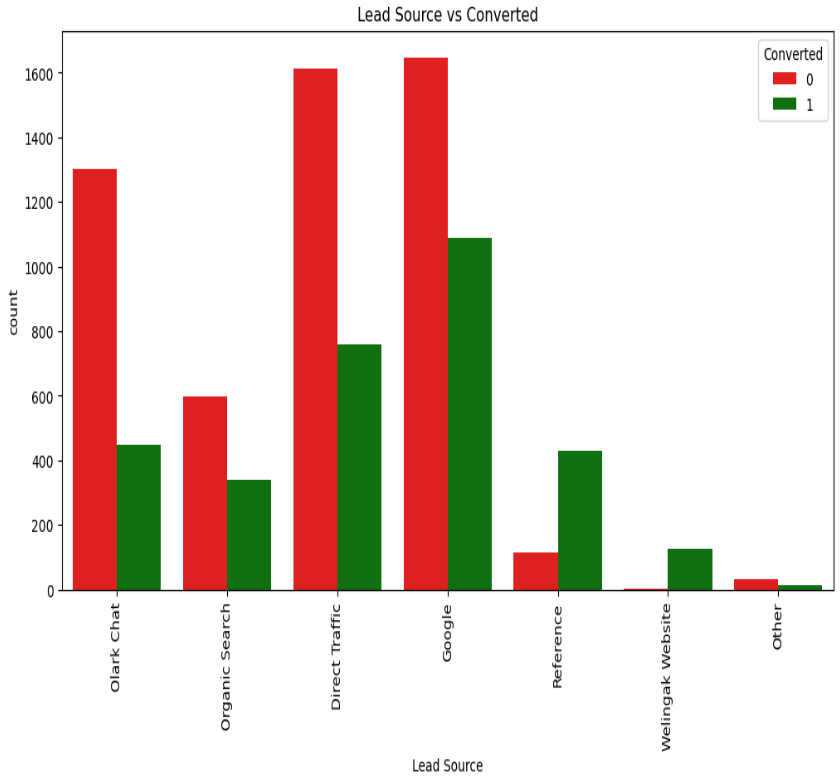
# Univariate Analysis


Distribution of TotalVisits

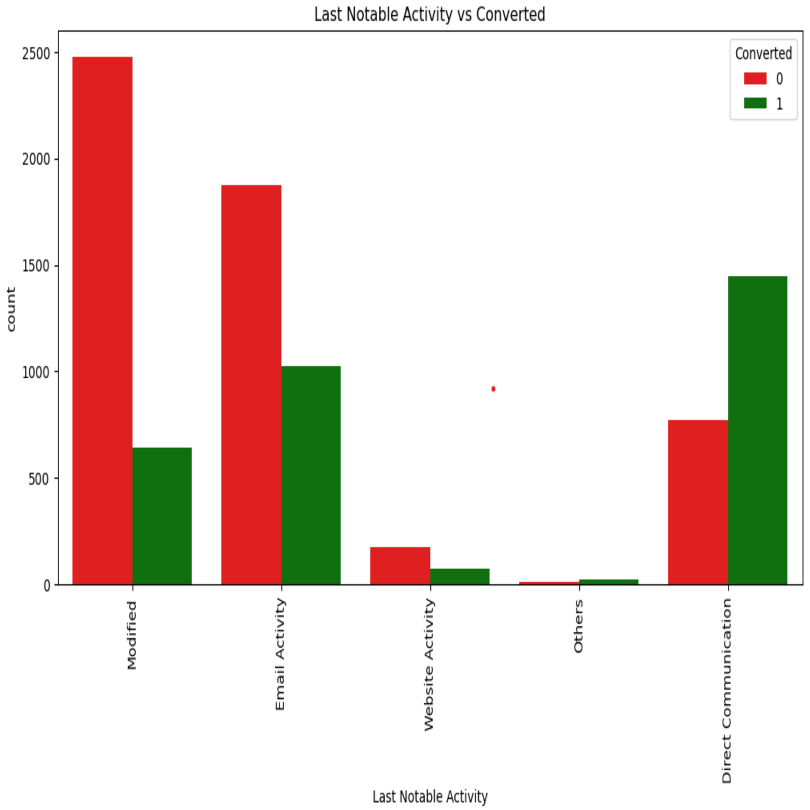

Distribution of Total Time Spent on Website


Distribution of Page Views Per Visit

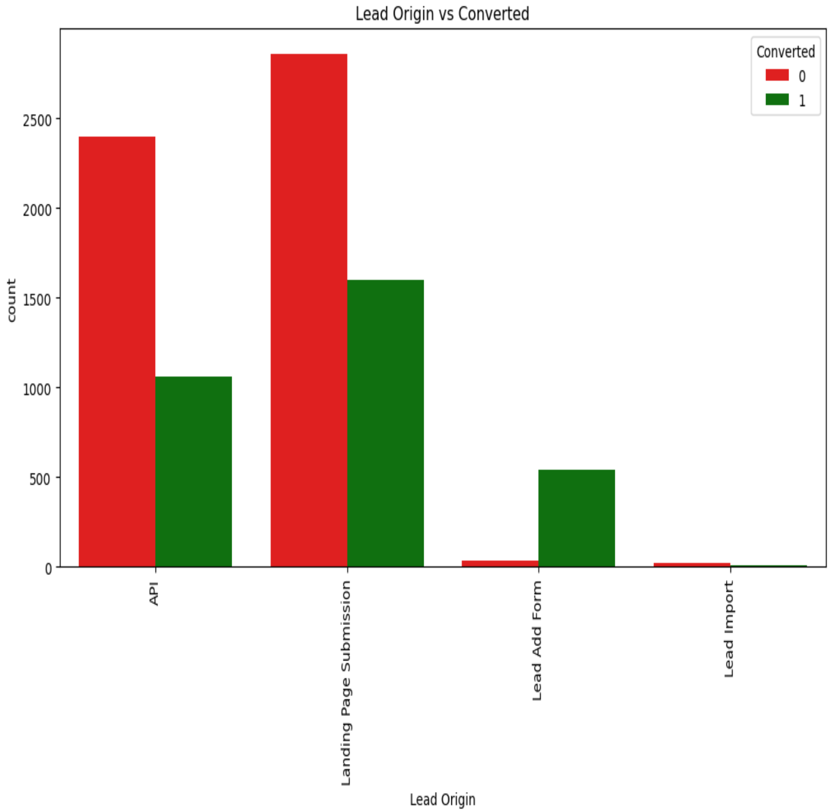
# Bivariate Analysis



Lead Source vs Converted

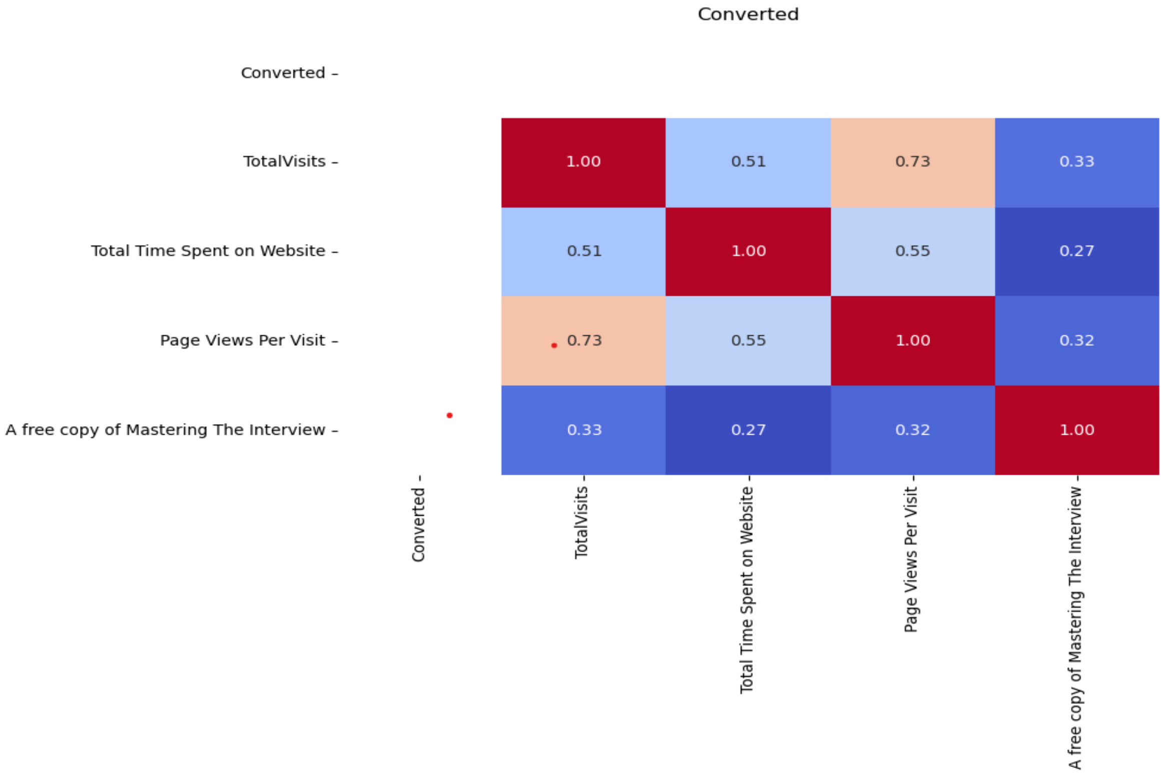


Last Notable Activity vs Converted



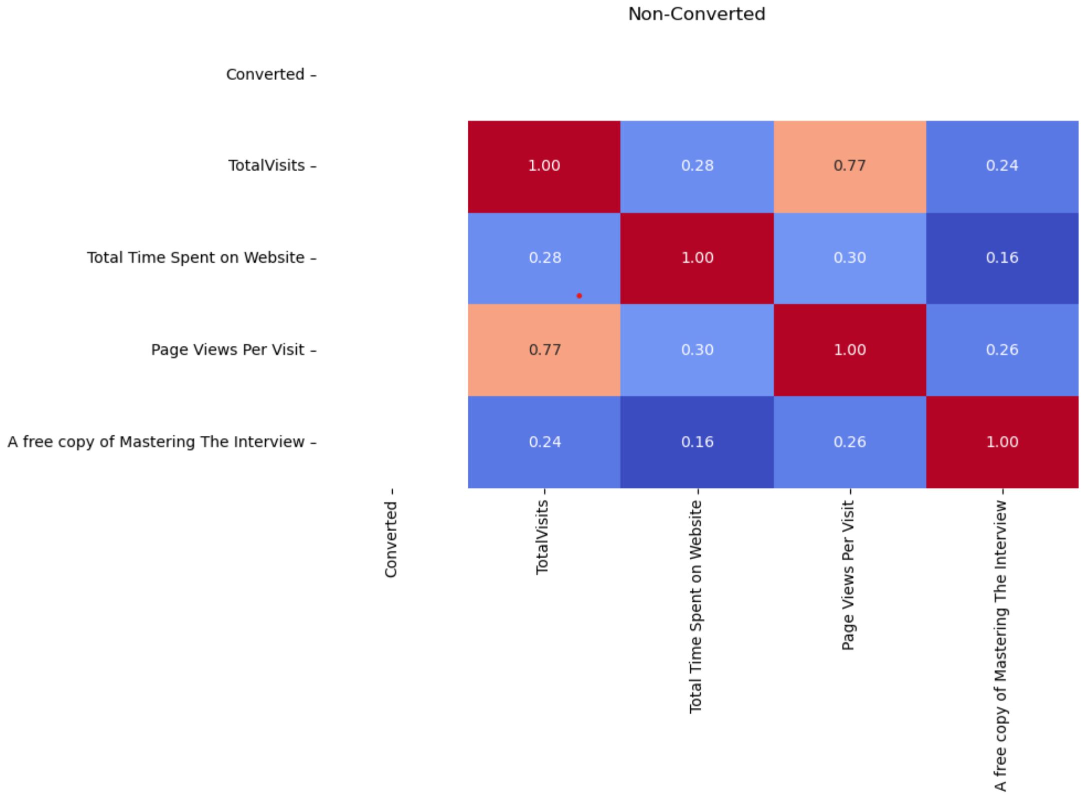
Lead Origin vs Converted

# Multivariate Analysis



Converted

| | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | A free copy of Mastering The Interview |
|---|---|---|---|---|---|
| Converted | | | | | |
| TotalVisits | | 1.00 | 0.51 | 0.73 | 0.33 |
| Total Time Spent on Website | | 0.51 | 1.00 | 0.55 | 0.27 |
| Page Views Per Visit | | 0.73 | 0.55 | 1.00 | 0.32 |
| A free copy of Mastering The Interview | | 0.33 | 0.27 | 0.32 | 1.00 |



Non-Converted

| | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | A free copy of Mastering The Interview |
|---|---|---|---|---|---|
| Converted | | | | | |
| TotalVisits | | 1.00 | 0.28 | 0.77 | 0.24 |
| Total Time Spent on Website | | 0.28 | 1.00 | 0.30 | 0.16 |
| Page Views Per Visit | | 0.77 | 0.30 | 1.00 | 0.26 |
| A free copy of Mastering The Interview | | 0.24 | 0.16 | 0.26 | 1.00 |

# Model Building

Built a logistic regression model with the goal of minimizing false negatives (0s = not converted leads) and maximizing true positives (1s = converted leads).

Key predictors used in the model include:

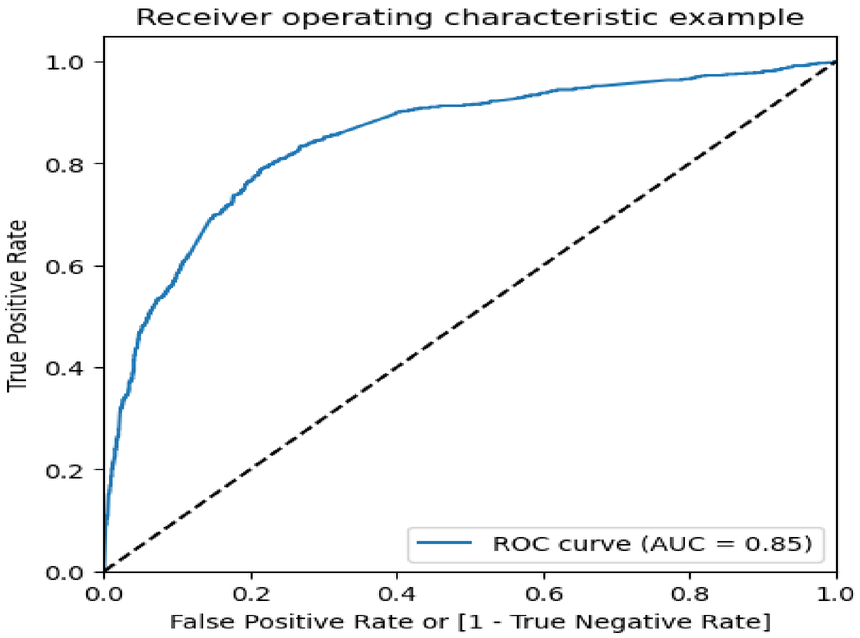
- Last Activity: Email Activity
- Last Notable Activity: Email Activity, Modified and Website Activity
- Lead Origin: Lead Add Form
- Lead Source: Direct Traffic and Olark Chat
- Total Time Spent on Website
- TotalVisits

The model was evaluated using a precision-recall curve to account for the imbalanced distribution of the target variable ('Converted').

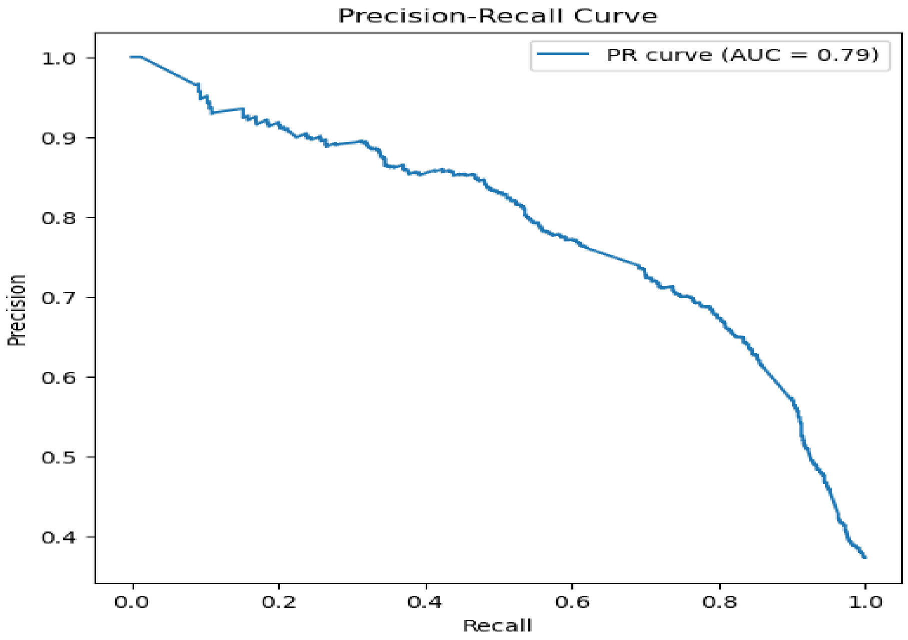
This approach ensures the identification of high-potential leads while minimizing resource wastage on unlikely conversions

# Evaluation

## ROC-AUC CURVE



## PRECISION-RECALL CURVE



# Recommendations

- Prioritize Leads with High Lead Score: Focus on leads with higher lead scores (e.g., 80+), as they have a higher probability of converting. Example: Lead Score of 89.68 shows strong conversion potential.

- Target Leads with Higher Predicted Conversion Probabilities: Leads with higher predicted probabilities of conversion (e.g., >0.5) should be prioritized, as they show higher interest in the product.

- Focus on High Engagement Leads: Leads with more website visits and time spent are more likely to convert. For example, leads who spend significant time on the website (Total Time Spent on Website) are more engaged

- Avoid leads with low scores (e.g., a Lead Score below 50) or other negative engagement indicators, as they are less likely to convert and should be deprioritized.

# Conclusions

- Improved Conversion Rates: By focusing on leads with higher engagement and positive signals, the likelihood of conversion increases
- Optimized Sales Efforts: Sales teams can save time by focusing on leads that are more likely to convert, resulting in better efficiency
- Better Resource Allocation: By targeting the right leads based on behavior, sales resources are used effectively, increasing the overall success rate