




Lead Scoring Model for X Education

by Utkarsh raj Verma
Sahil Madavi



Problem Statement:

X EDUCATION IS FACING A LOW LEAD CONVERSION RATE (30%). THE COMPANY WANTS TO IDENTIFY "HOT LEADS" WITH A HIGHER LIKELIHOOD OF CONVERTING TO PAYING CUSTOMERS. THE GOAL IS TO BUILD A MODEL THAT ASSIGNS A LEAD SCORE TO EACH LEAD, HELPING THE SALES TEAM FOCUS ON THE MOST PROMISING LEADS AND INCREASE THE LEAD CONVERSION RATE TO AROUND 80%.

Data Overview

- ▶ Data Source: The dataset consists of ~9,000 records with various features like Lead Source, Total Time Spent on Website, Total Visits, and Last Activity, which can potentially influence lead conversion.
- ▶ Target Variable: Converted (1 = converted, 0 = not converted).
- ▶ Preprocessing: Handling missing values (e.g., replacing "Select" entries), encoding categorical variables, and scaling numerical features.

Analysis Approach

- ▶ Data Cleaning: Addressed missing data and replaced "Select" values (treated as null). Removed redundant or irrelevant categories in categorical features.
- ▶ Feature Engineering: Encoded categorical variables using One-Hot Encoding. Standardized numerical features like Total Time Spent and Total Visits for model compatibility.
- ▶ Model Choice: Chose Logistic Regression as it outputs probabilities and is easy to interpret.
- ▶ Model Evaluation: Used performance metrics: Accuracy, Precision, Recall, F1 Score, and ROC AUC.

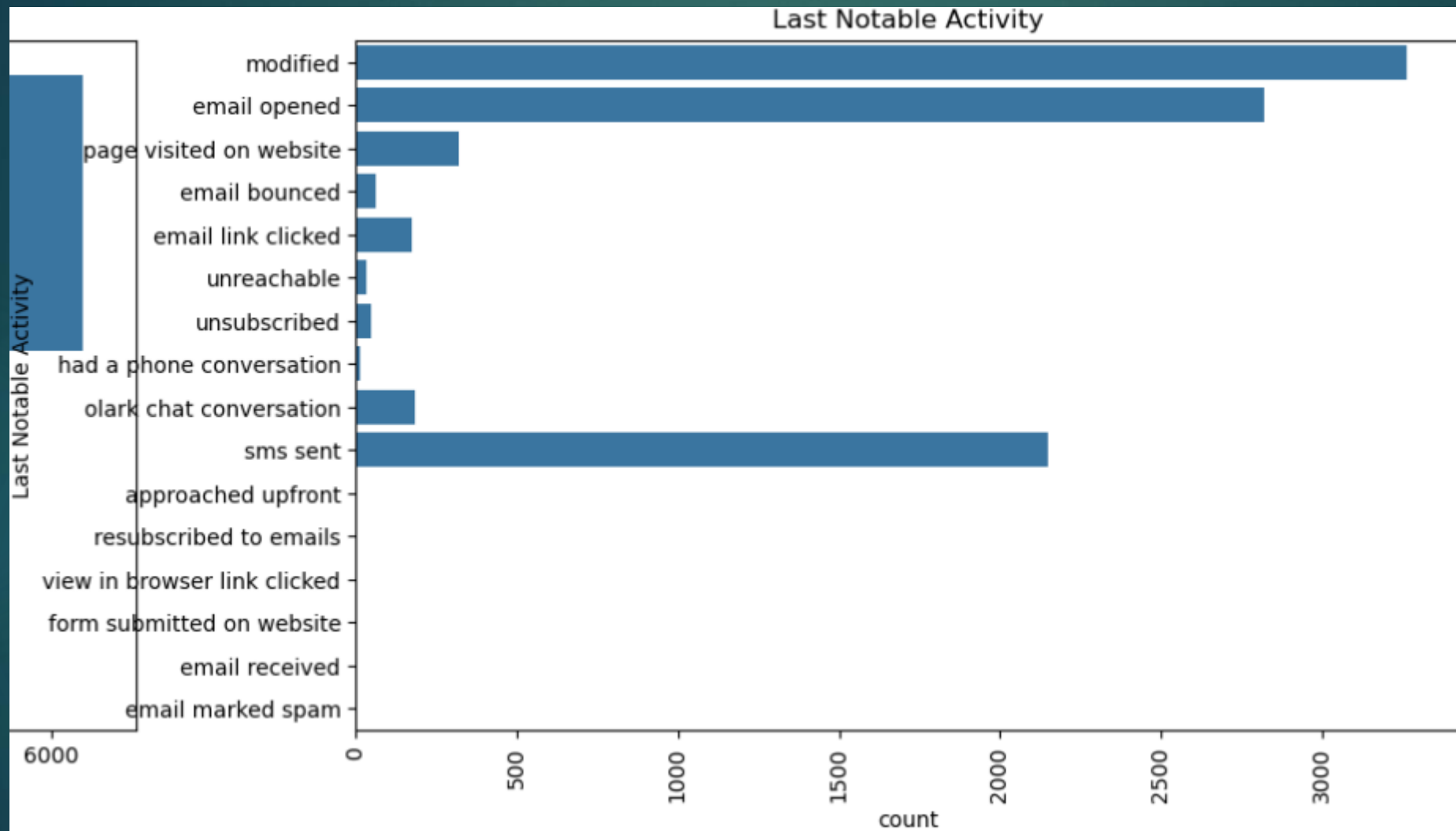
Data Manipulation

- ▶ Total Number of Rows =37, Total Number of Columns =9240.
- ▶ Single value features like “Magazine”, “Receive More Updates About Our Courses”,
- ▶ “Update me on Supply”
- ▶ Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through
- ▶ cheque” etc. have been dropped.
- ▶ Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- ▶ After checking for the value counts for some of the object type variables, we find some of
- ▶ the features which has no enough variance, which we have dropped, the features are:
- ▶ “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper
- ▶ Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- ▶ □ Dropping the columns having more than 35% as missing value such as ‘How did you hear
- ▶ about X Education’ and ‘Lead Profile’.

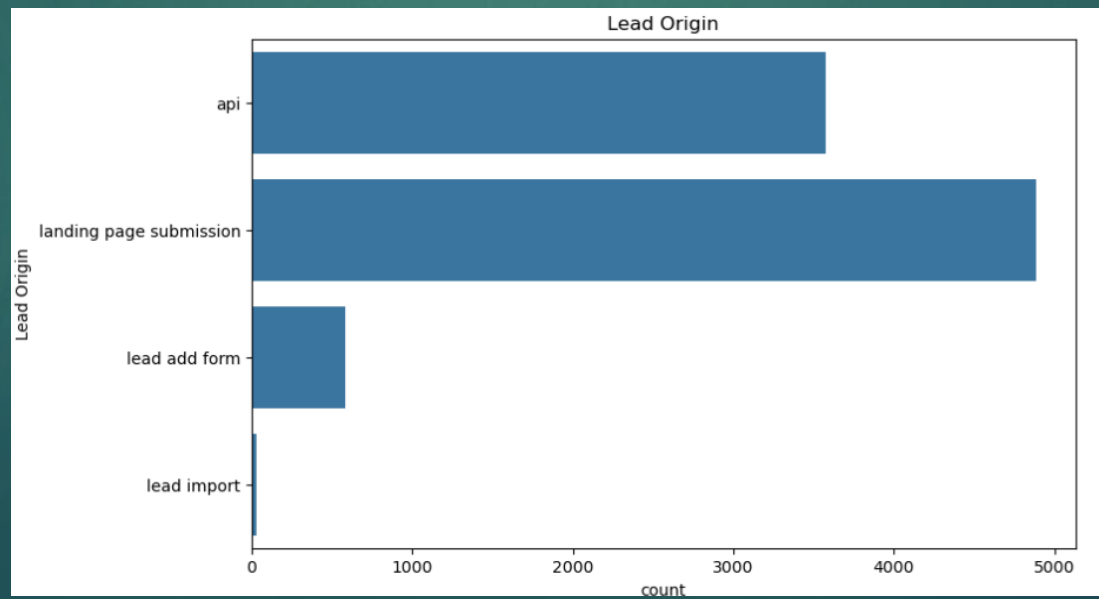
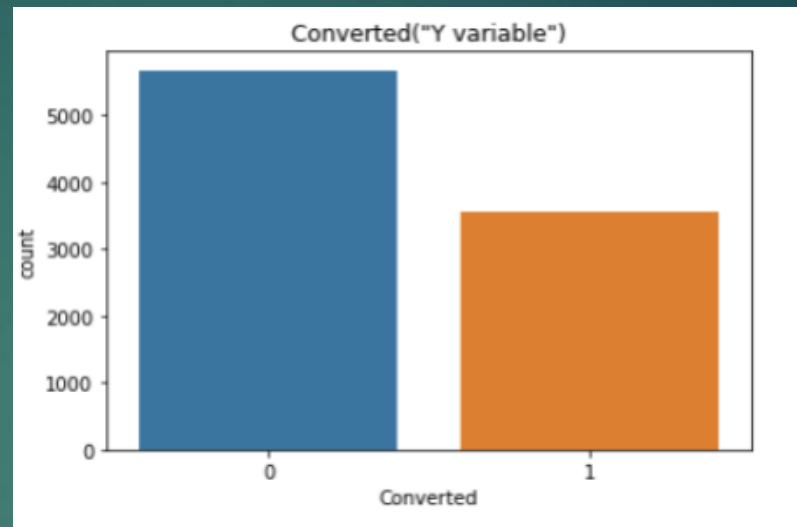
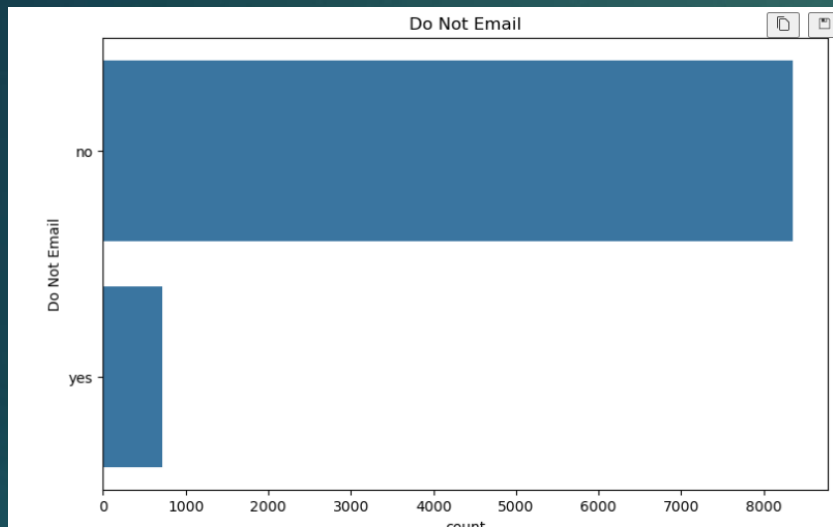
Exploratory Data Analysis (EDA)

- ▶ Visualization 1: Distribution of the Target Variable: A bar chart showing the imbalance in the dataset (more non-converted leads than converted).
- ▶ Visualization 2: Correlation Heatmap: Visualizing relationships between numerical features and the target variable. This highlights which features are more related to conversion.
- ▶ Visualization 3: Lead Source vs Conversion Rate: A bar chart showing conversion rates by lead source, helping us understand which lead sources are more effective.

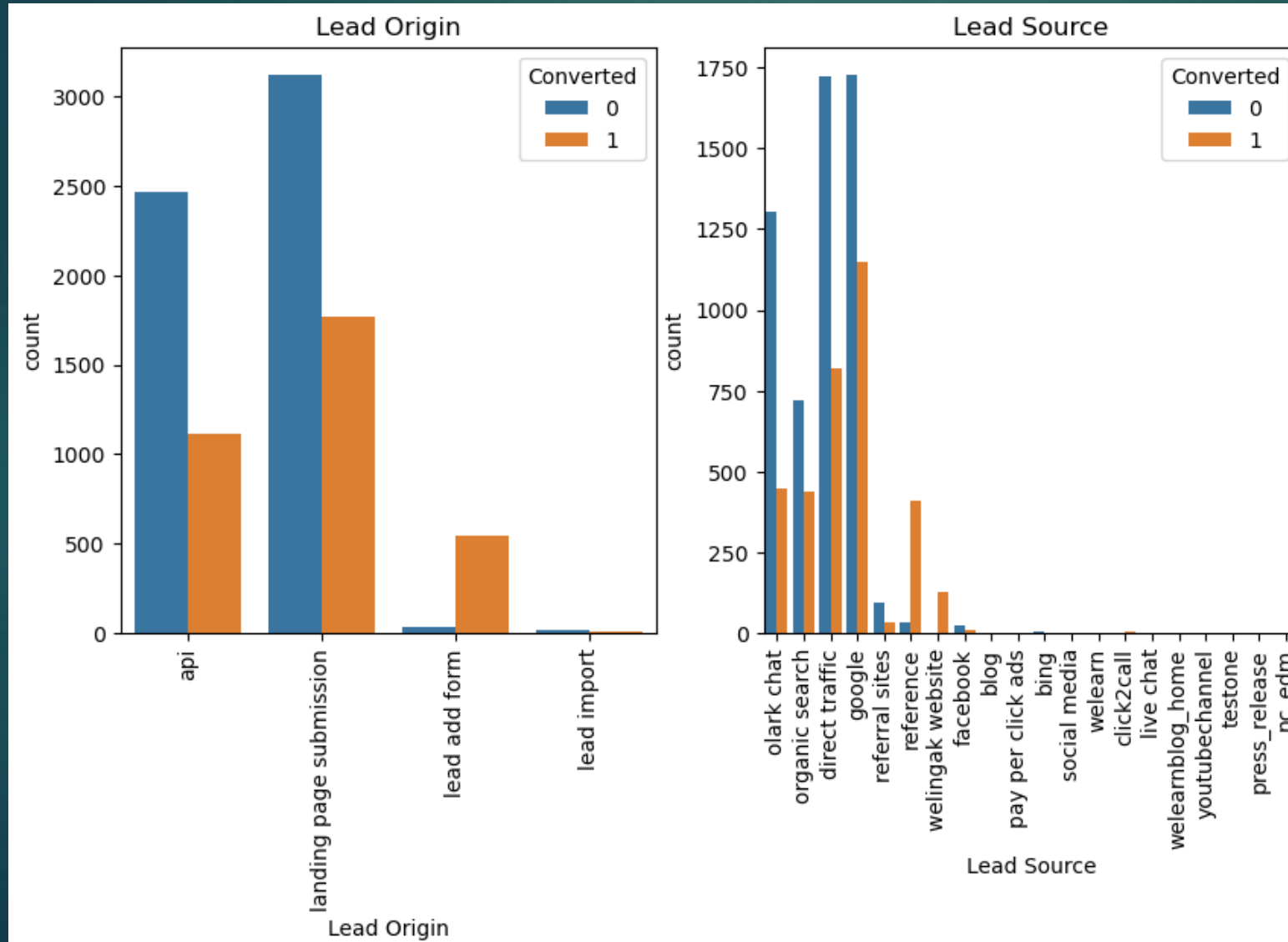
Visualization 1

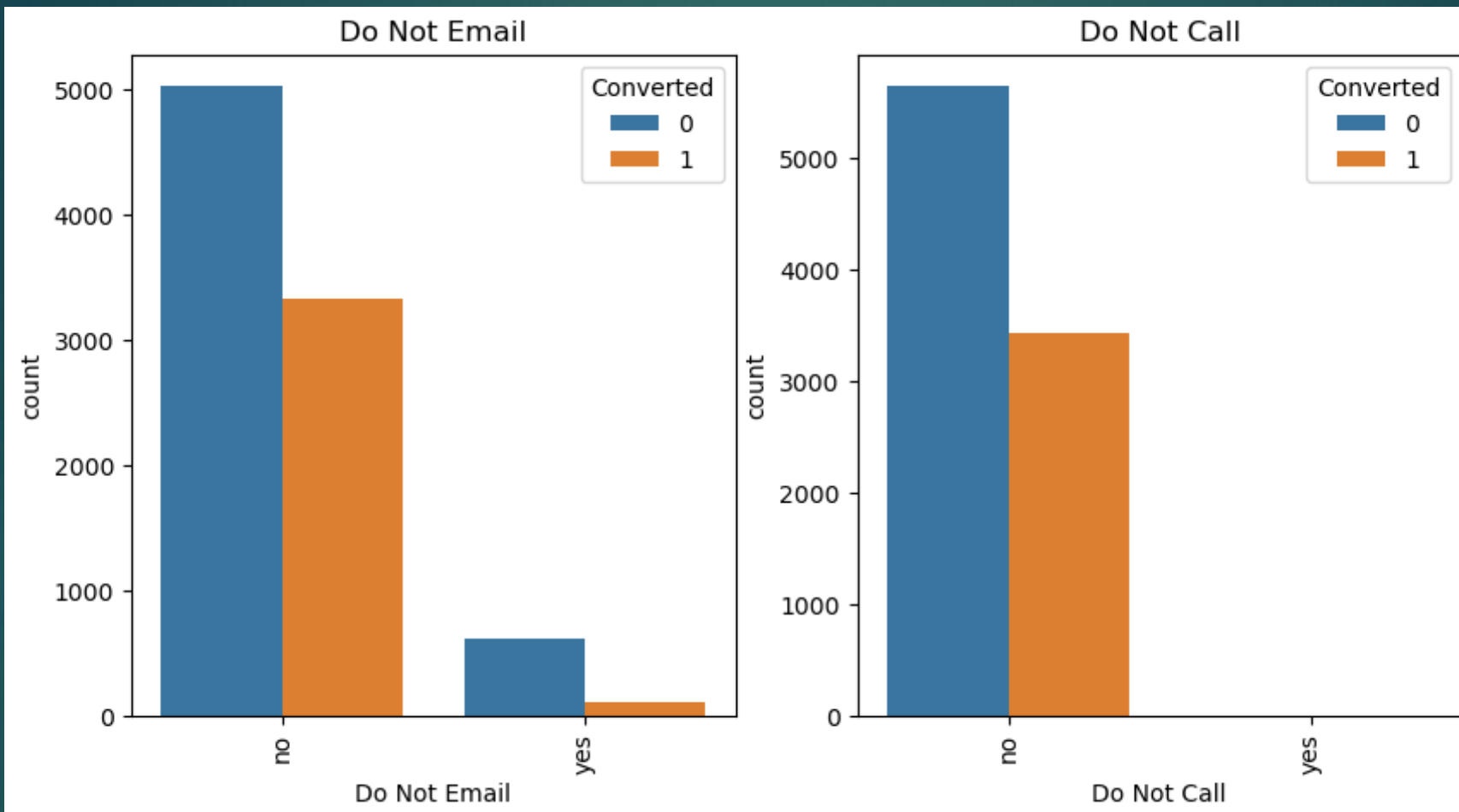


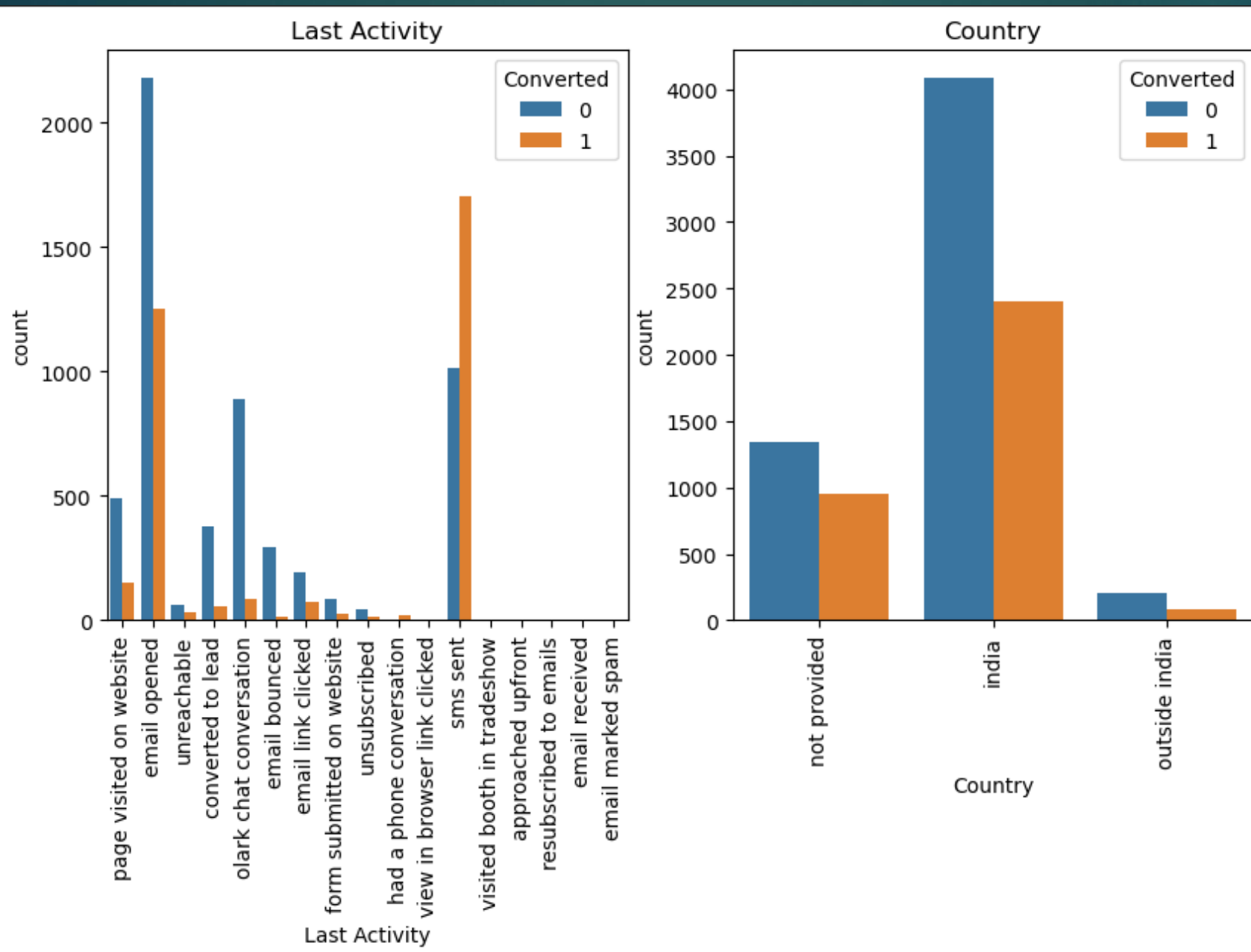
Visualization 2



Categorical Variable Relation







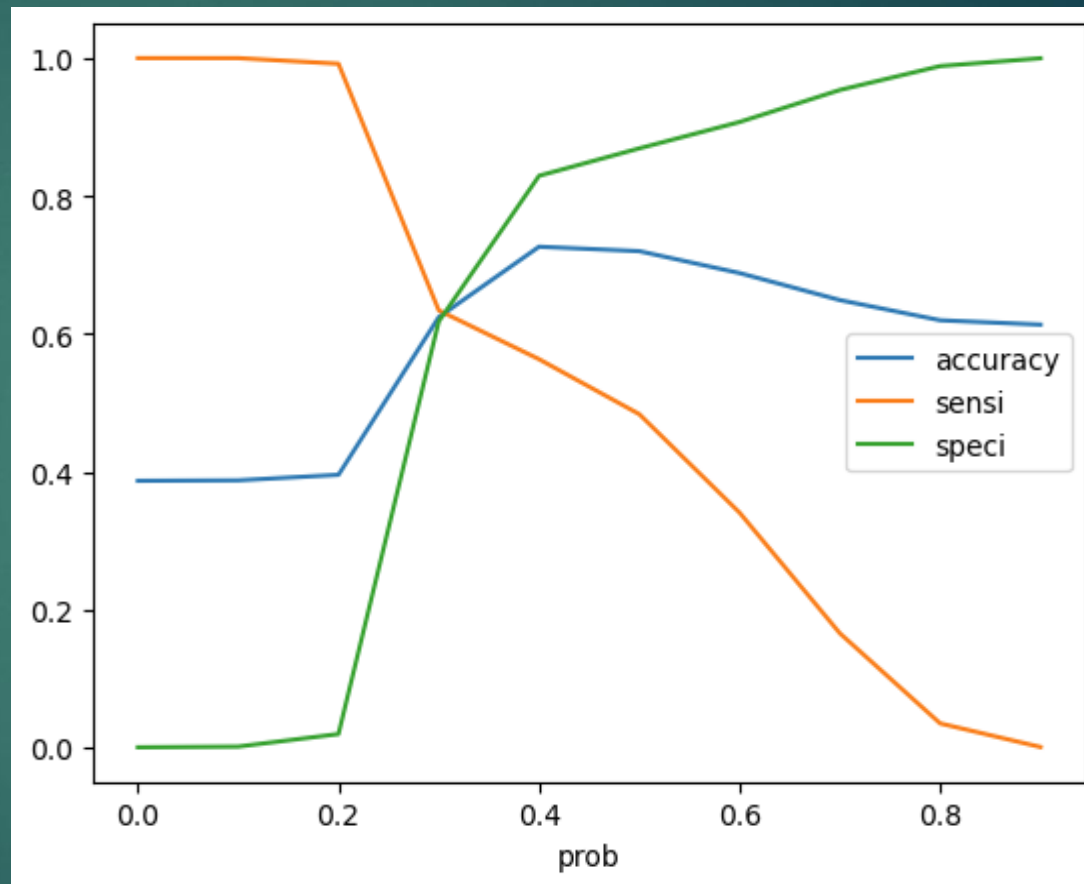
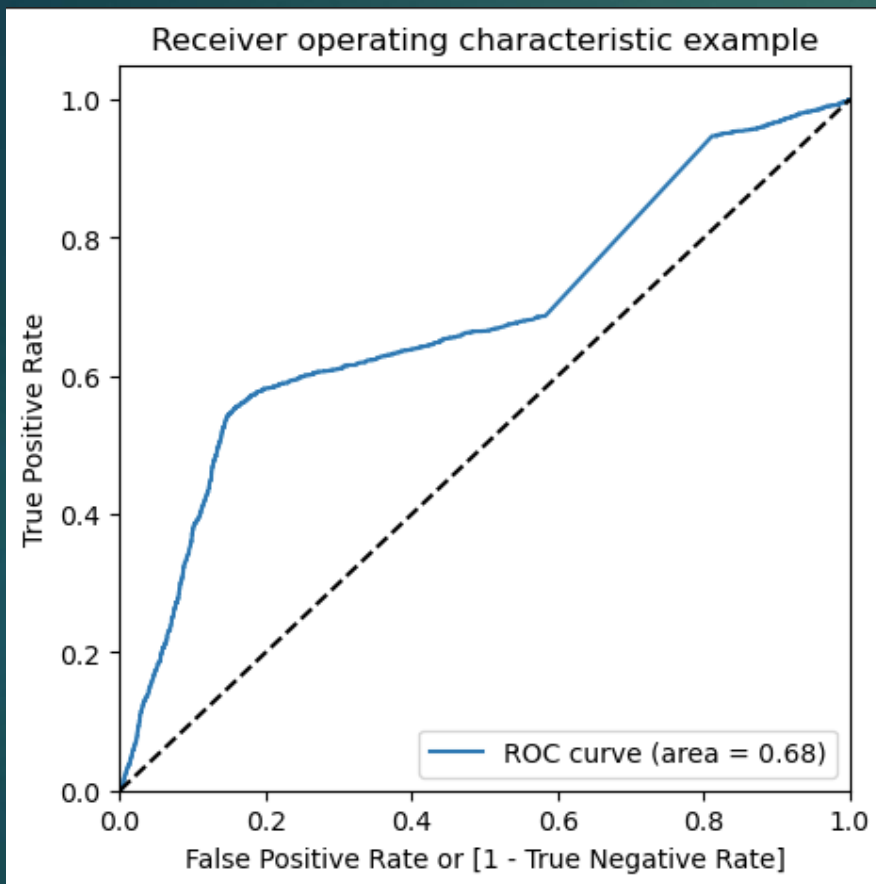
Data Conversion

- ▶ Numerical Variables are Normalised
- ▶ Dummy Variables are created for object type variables
- ▶ Total Rows for Analysis: 8792
- ▶ Total Columns for Analysis: 43

Model Building

- ▶ Splitting the Data into Training and Testing Sets
- ▶ The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- ▶ Use RFE for Feature Selection
- ▶ Running RFE with 15 variables as output
- ▶ Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- ▶ Predictions on test data set
- ▶ Overall accuracy 72%

ROC Curve



Business Impact

- ▶ Improved Lead Prioritization: The model allows X Education to prioritize leads with a higher likelihood of conversion, ensuring that the sales team focuses on the most promising candidates.
- ▶ Higher Conversion Rate: By using the model to score leads, the company can expect a higher conversion rate as they engage with the right prospects.
- ▶ Efficient Resource Allocation: Resources spent on cold leads (leads with low conversion probability) will be reduced, leading to more efficient use of time and effort.

Next Steps & Recommendations

- ▶ **Threshold Adjustment:** Fine-tuning the decision threshold could increase the recall (identify more potential conversions), at the cost of precision.
- ▶ **Cross-validation:** Perform cross-validation for a more robust performance assessment across multiple data splits.
- ▶ **Exploring Advanced Models:** Test other models (e.g., Random Forest, Gradient Boosting) to compare performance and further improve lead scoring accuracy.
- ▶ **Feature Engineering:** Further investigation into additional features (e.g., interaction terms) could improve model predictions.

Conclusion

- ▶ Summary: The logistic regression model developed successfully predicts lead conversion, assigning a lead score between 0 and 100. This helps X Education focus on high-conversion leads, improving efficiency and increasing the chances of higher lead conversion.
- ▶ Key Takeaway: Using data-driven insights, X Education can significantly improve its sales strategy by focusing on the most promising leads, ultimately increasing revenue and reducing unnecessary effort on cold leads.