Lead Scoring Model for X Education

Case Study

"IMPROVING LEAD CONVERSION WITH LOGISTIC REGRESSION"

By:-
NEHA SHAHID
NEHA VERMA
NEETIKA JAMNAL

Problem Statement

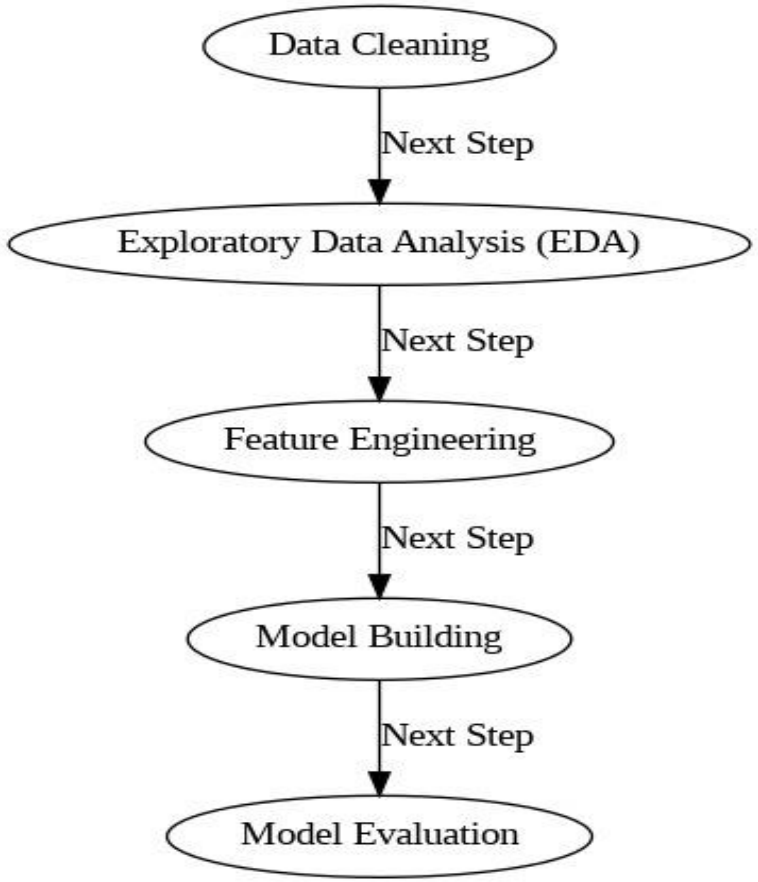
X Education is facing challenges with lead conversion, achieving only a 30% conversion rate . The goal is to improve this conversion rate to around 80% by prioritizing high-quality leads using a data-driven model . The objective of this analysis is to build a lead scoring model using logistic regression to predict the likelihood of a lead converting into a paying customer.

Goal: Improve X Education's lead conversion rate to around 80% by identifying "Hot Leads."

Solution: Build a Lead Scoring Model to prioritize leads most likely to convert.

Analysis Approach

- Data Cleaning: Handled missing values, removed irrelevant columns.
- Exploratory Data Analysis (EDA): Identified correlations between key features and conversion.
- Feature Engineering: Created dummy variables for categorical features like Lead Source, Last Activity.
- Model Building: Built a logistic regression model to predict lead conversion.
- Model Evaluation: Evaluated model accuracy, precision, recall, and ROC-AUC score.

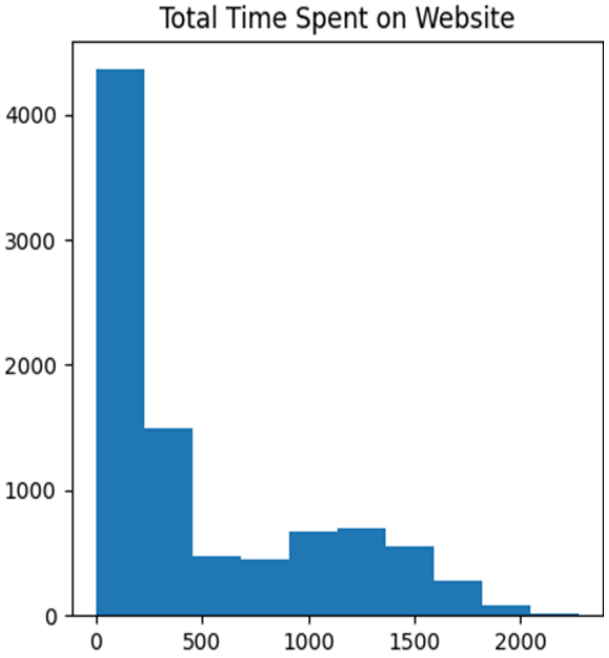


Data Overview

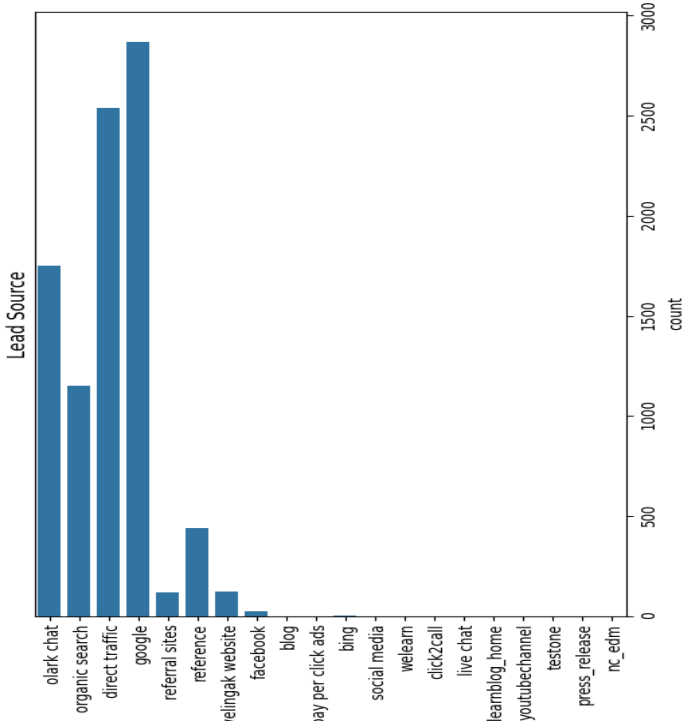
- Dataset contains around 9240 data points.
- Features include: Lead Source, Total Time Spent on Website, Last Activity, and more.
- Target variable: Converted (1 for converted, 0 for non-converted).
- Cleaned Data: Removed columns with >30% missing values, handled missing data (imputation/removal), and standardized text (e.g., lowercase).
- Filtered Irrelevant Categories: Replaced placeholders like "Select" with NaN for accurate processing.
- Ready for Modeling: Proceed with encoding categorical variables, feature scaling, and model evaluation.

Exploratory Data Analysis (EDA)

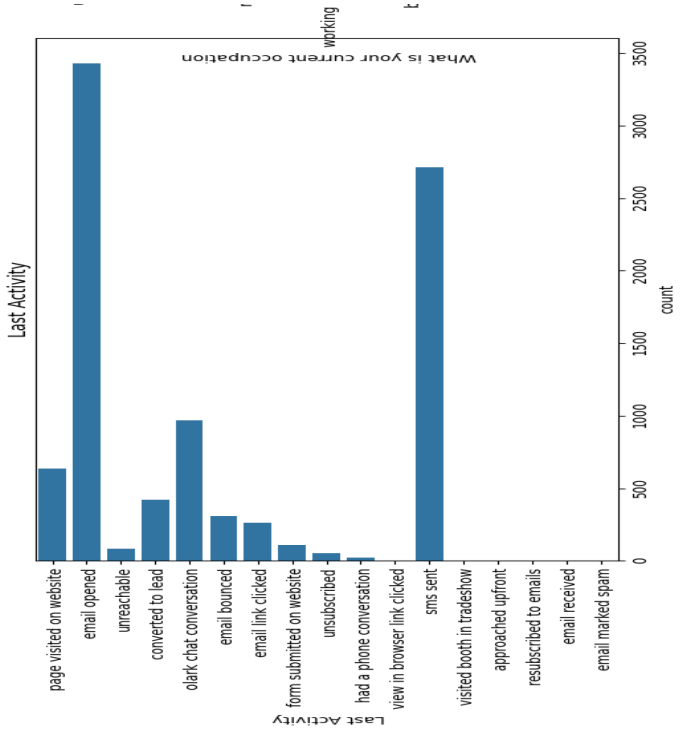
Univariate Analysis



Total time spent on website analysis :it was found that above 4000 minutes was spent by customers on website.

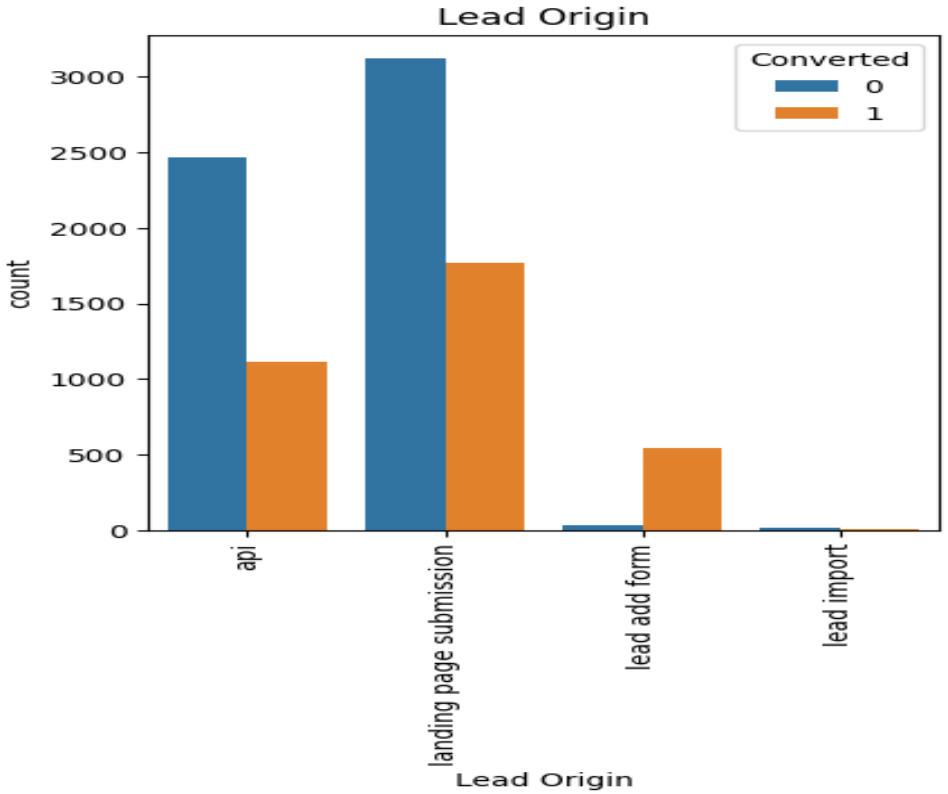


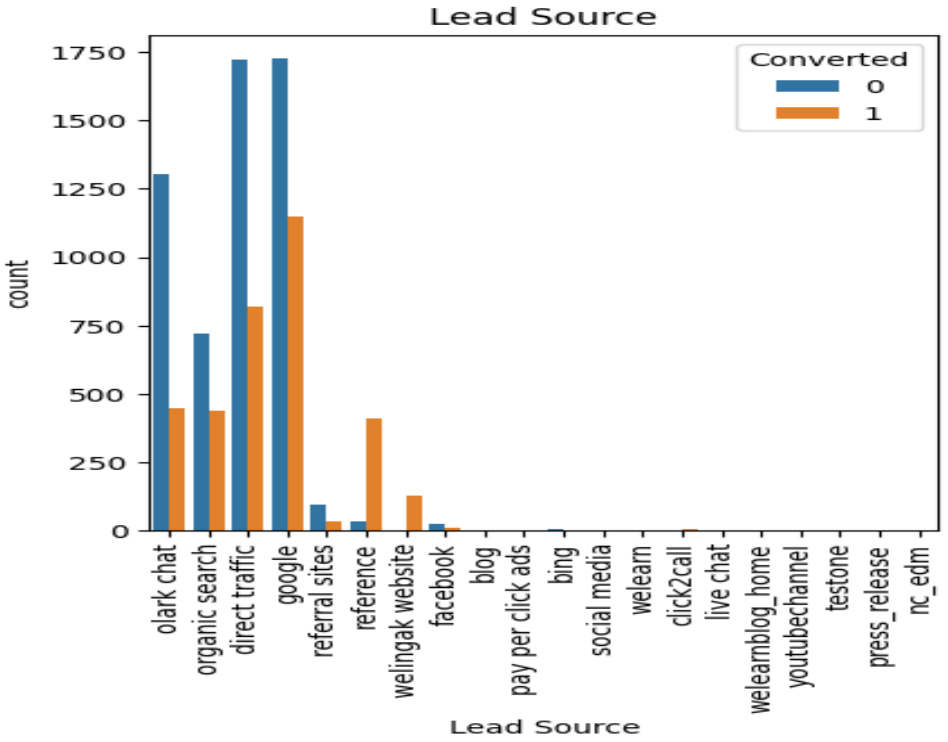
Lead Source Analysis: "Google" contributes the most among all lead sources, indicating its strong influence on conversions.



Last Activity Analysis: "Email Opened" has the highest contribution among all activities.

Bivariate Analysis



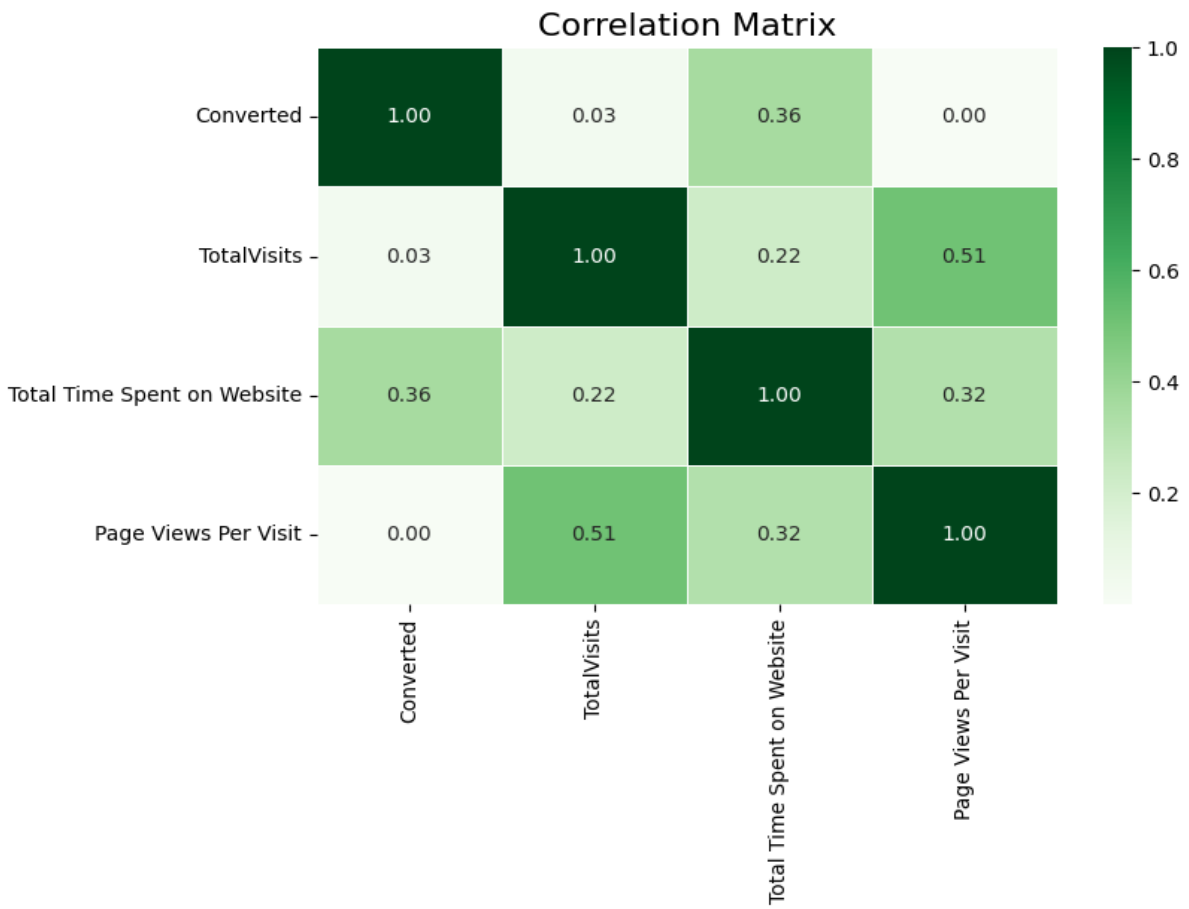


The graph shows that **"Landing Page Submission“, "Direct Traffic“ and “Google** contribute the most to lead conversions.

Correlation Analysis

Key Insights:

- Found strong positive correlations between Total Time Spent on Website and lead conversion.
- Lead Source, Last Activity, and Lead Quality significantly impact conversion.
- Visualized relationships between features and target variable.
- **Insights regarding "Total Visits":**
- Converted: The correlation between Total Visits and Converted is 0.03, which is very weak.
- Total Time Spent on Website: The correlation is 0.22, indicating a mild positive relationship.
- Page Views Per Visit: The correlation is 0.51, showing a moderately strong positive relationship.



Model Building

- Feature Selection using RFE (Recursive Feature Elimination)
- RFE identifies the most important features for the model by recursively fitting a model and eliminating the least significant features.
- .Building the Logistic Regression Model
- RFE Selected 15 Variables : Features were refined using Recursive Feature Elimination.
- Dropped High VIF Columns : Removed variables with VIF > 5 to address multicollinearity.
- Dropped Insignificant Variables : Excluded features with p-value > 0.05 for statistical significance.

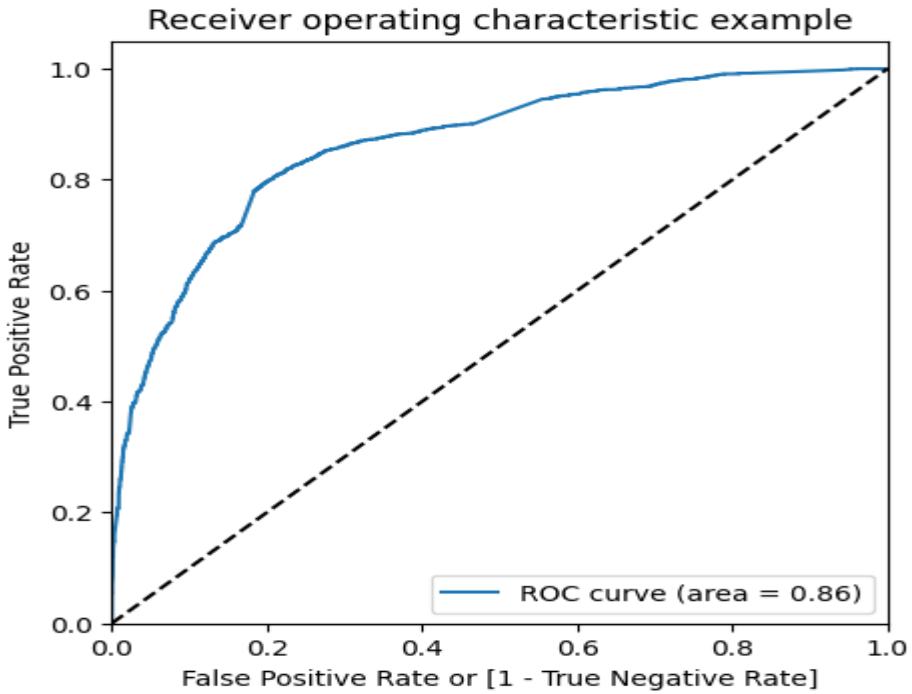
- .

Model Evaluation

- Precision: High precision indicating the model correctly identifies the leads most likely to convert.
- Recall: High recall ensuring the model captures most converted leads.
- ROC-AUC: The model performed well with an ROC-AUC score above 0.8.
- Confusion Matrix: Visual showing the true positives, false positives, etc.
- Performance Metrics Table (Accuracy, Precision, Recall, F1 Score).

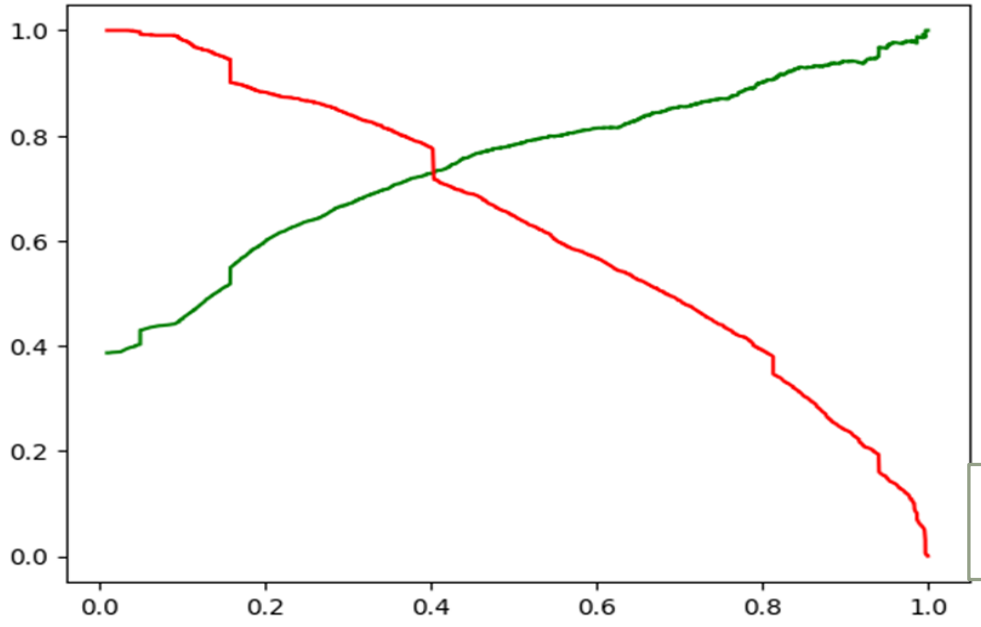
Model Evaluation

ROC CURVE



The area under ROC curve is 0.86

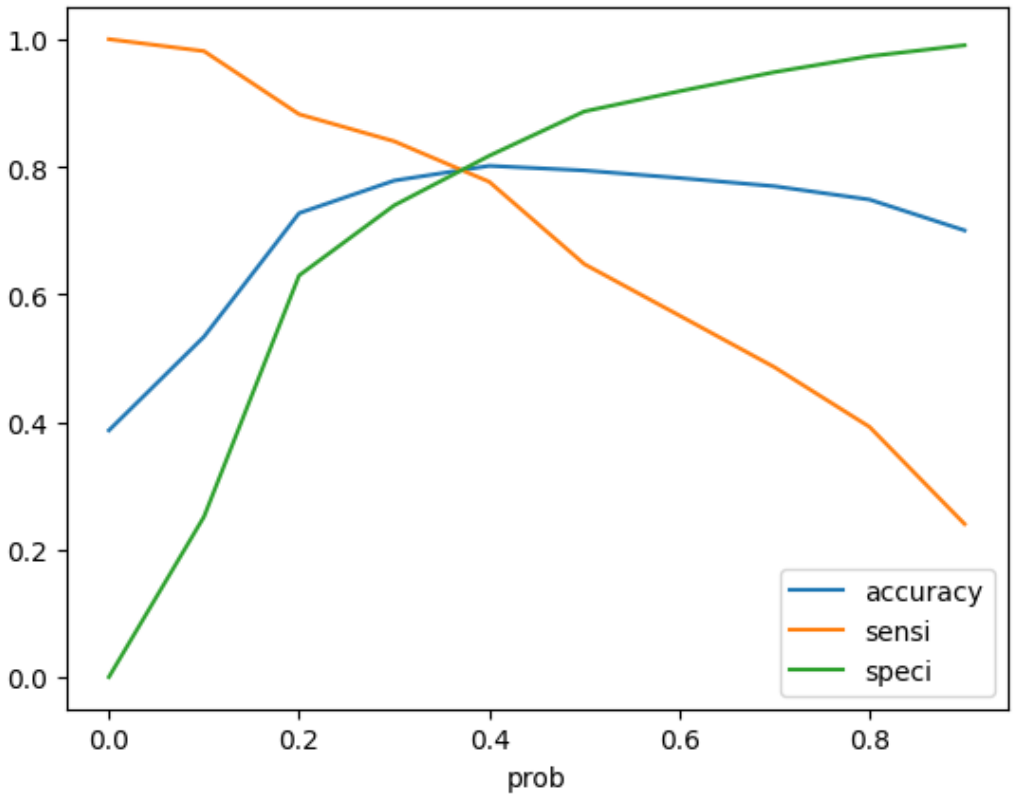
.



Green:- Precision
Red :- Recall

Precision is 78% , Recall is 64%

Model Evaluation



- Accuracy :- 79%
- Precision:- 78%
- Recall:- 64%
- ROC-AUC:- 0.86
- Sensitivity:- 82%
- Specificity:-78%
- Conclusion:- Model performs well in identifying high-potential leads
- Graph depict an optimal cut-off of 0.35 basis on accuracy, sensitivity, specificity

Business Insights

- Lead Source: Strengthen marketing on channels with high-quality leads (e.g., Google Search, Referrals).
- Last Activity: Focus on leads who have recently engaged (opened emails, visited key pages).
- Lead Quality: Prioritize "Hot" leads that are more likely to convert.

Sales Strategy Suggestion

1)**Aggressive Conversion Phase**

- Objective: Maximize lead conversion during intern phase (2 months).
- Strategy : Focus on High-Priority Leads (Hot Leads).

Interns handle Medium-Priority Leads with automated follow-ups.

Use multichannel outreach (phone, email, SMS) for Hot Leads.

2)Minimizing Useless Phone Calls

- Objective: Minimize outreach during downtime to focus on high-potential leads.
- Strategy : Raise the lead score threshold to only target the highest potential leads.

Use activity-based filtering to prioritize leads with recent engagement.

Automate follow-ups for low-priority leads using CRM.