Lead Score Case Study using Logical Regression

By Manashree Sawant, Manikanta Vuggina and Manisha Mathur

Content

Problem Statement Summary

An education company named X Education sells online courses to industry professionals. They want to improve the probability of converting potential leads into enrolled students.

X Education generates leads through website visits, form submissions, and referrals from past customers.

The sales team engages leads via calls and emails, but the current conversion rate is only around 30%.

Despite acquiring many leads daily, only about 30% convert into paying customers, leading to inefficiency.

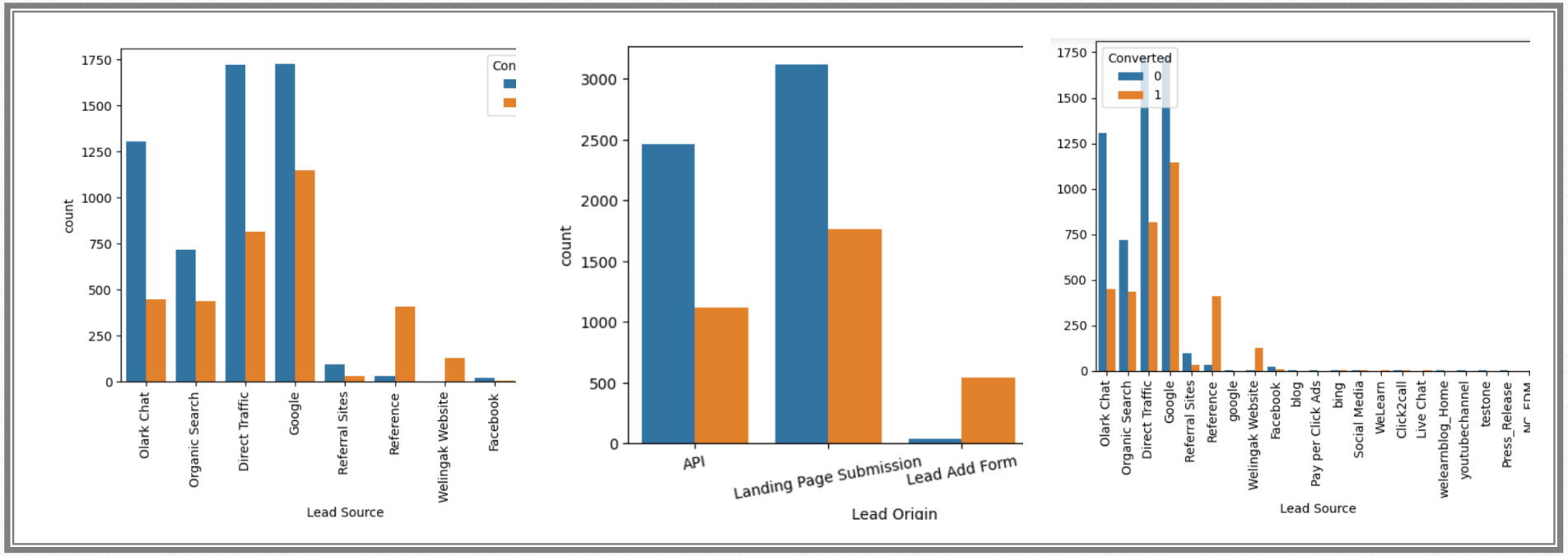
Business Goal

The company aims to improve conversion rates by identifying high-potential leads, allowing the sales team to prioritize efforts on the most promising prospects.

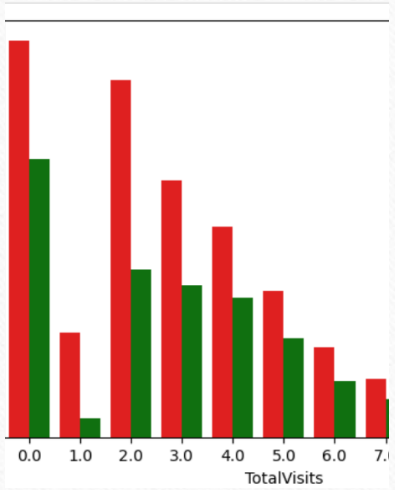
X Education wants to improve lead conversion by identifying and nurturing high-potential leads. The task is to build a model that assigns lead scores, helping prioritize leads to achieve a target conversion rate of 80%.

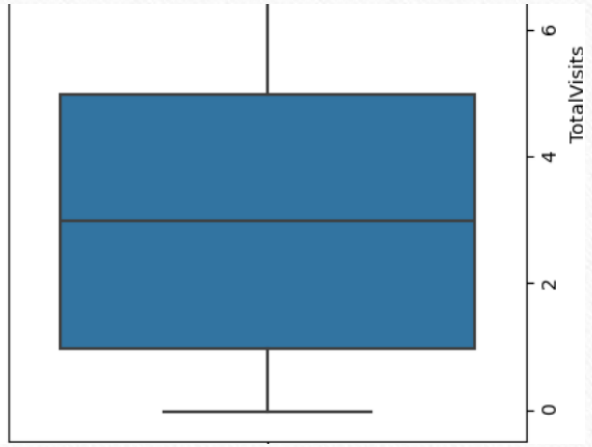
Problem Approach

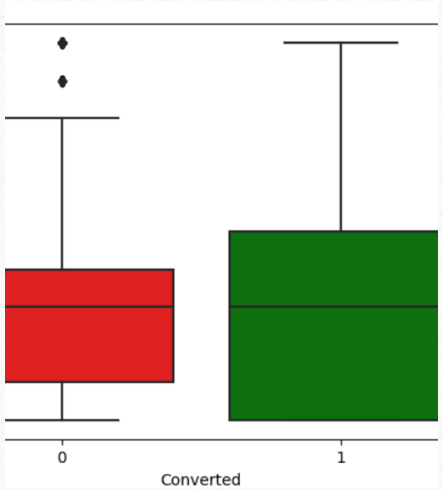
- Import Data
- Check data size, shape, information and data type
- Perform EDA
- Handle missing values
- Handling outliers
- Dropping rows with null values
- get list of object cols, convert them to numericals
- Train model
- Test split
- Running the model
- Refit the model with the new set of features
- Creating Predictions
- Model Evaluation
- Optimize ROC Curve
- Call the ROC function and evaluating the ROC area



Analysis of Lead Source: Olak and Google have the highest lead conversion count as compared to others

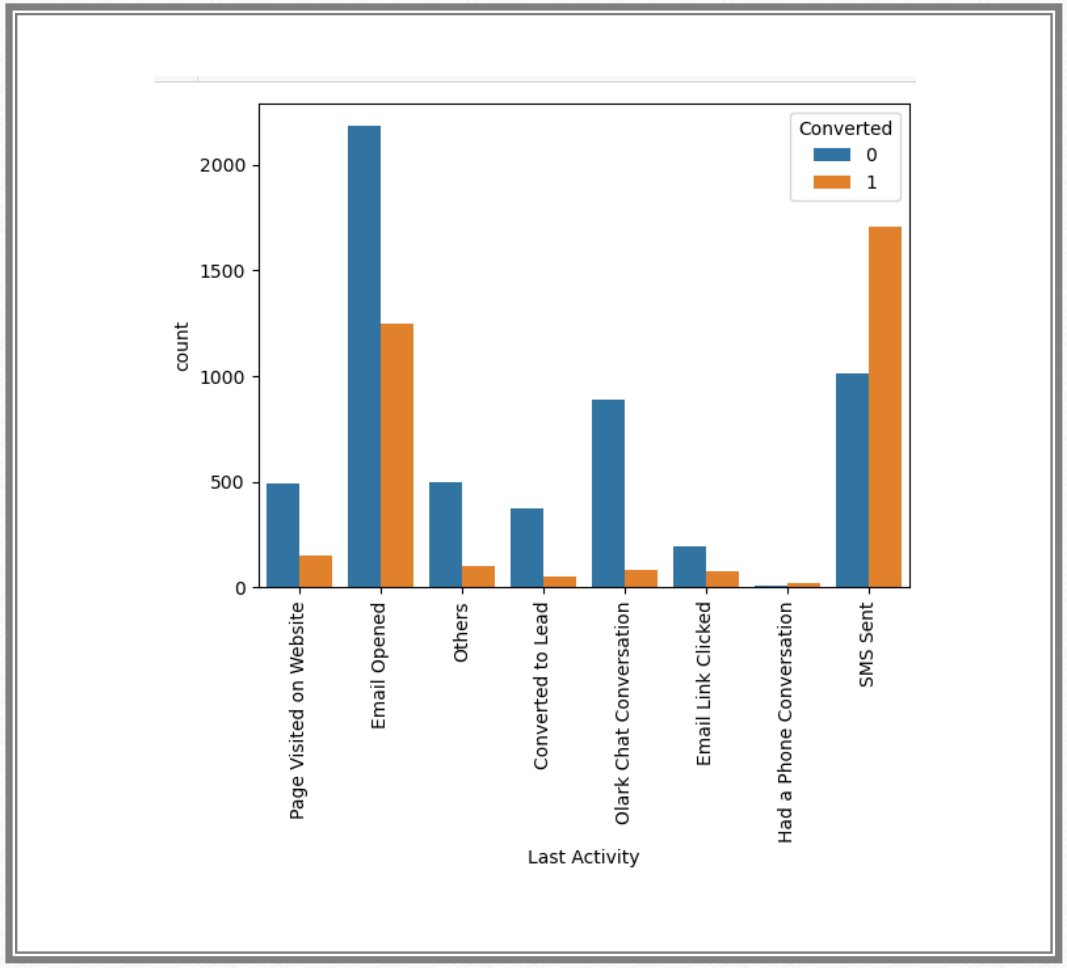






Analysis of Total Visit: People who have visited the website is more likely to get converted.

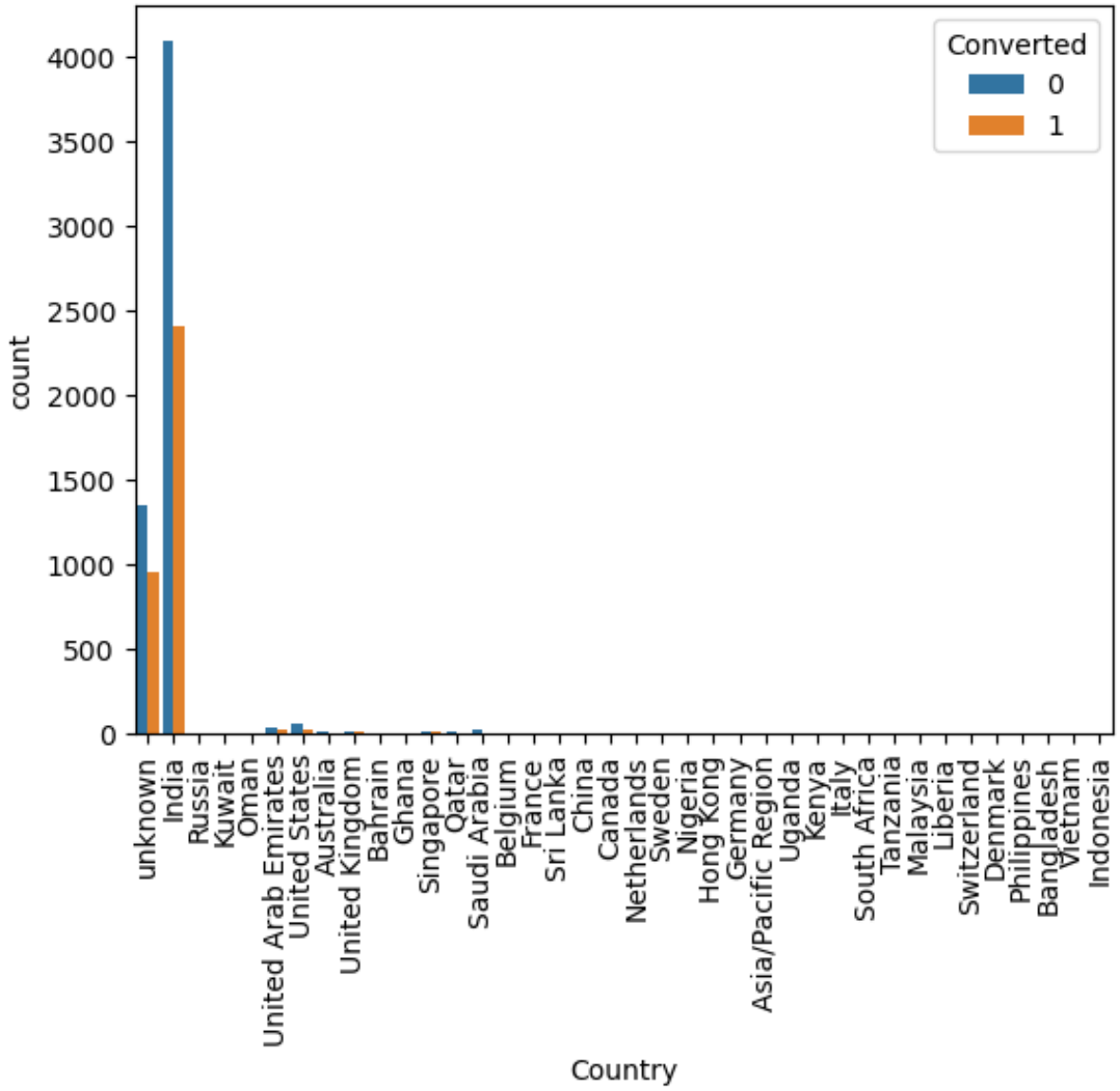
- **However, the total counts of customers visiting the site more often is less**



Last Activity

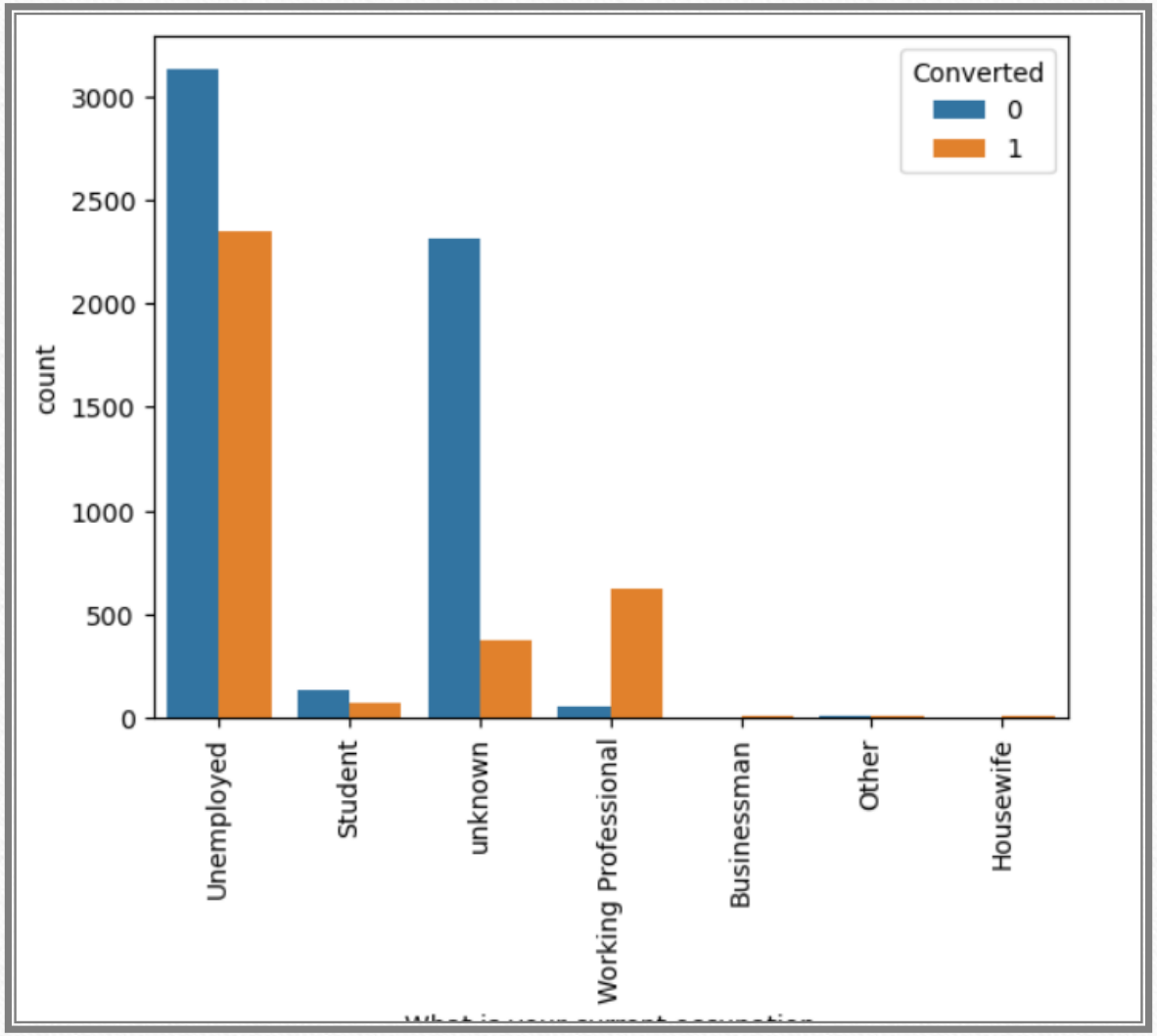
- Most of the lead have their Email opened as their last activity.
- Conversion rate for leads with last activity as SMS Sent the highest.
- Leads to whom SMSes are sent are most likely to get converted

Country wise leads generated: India has the maximum leads



Current Occupation

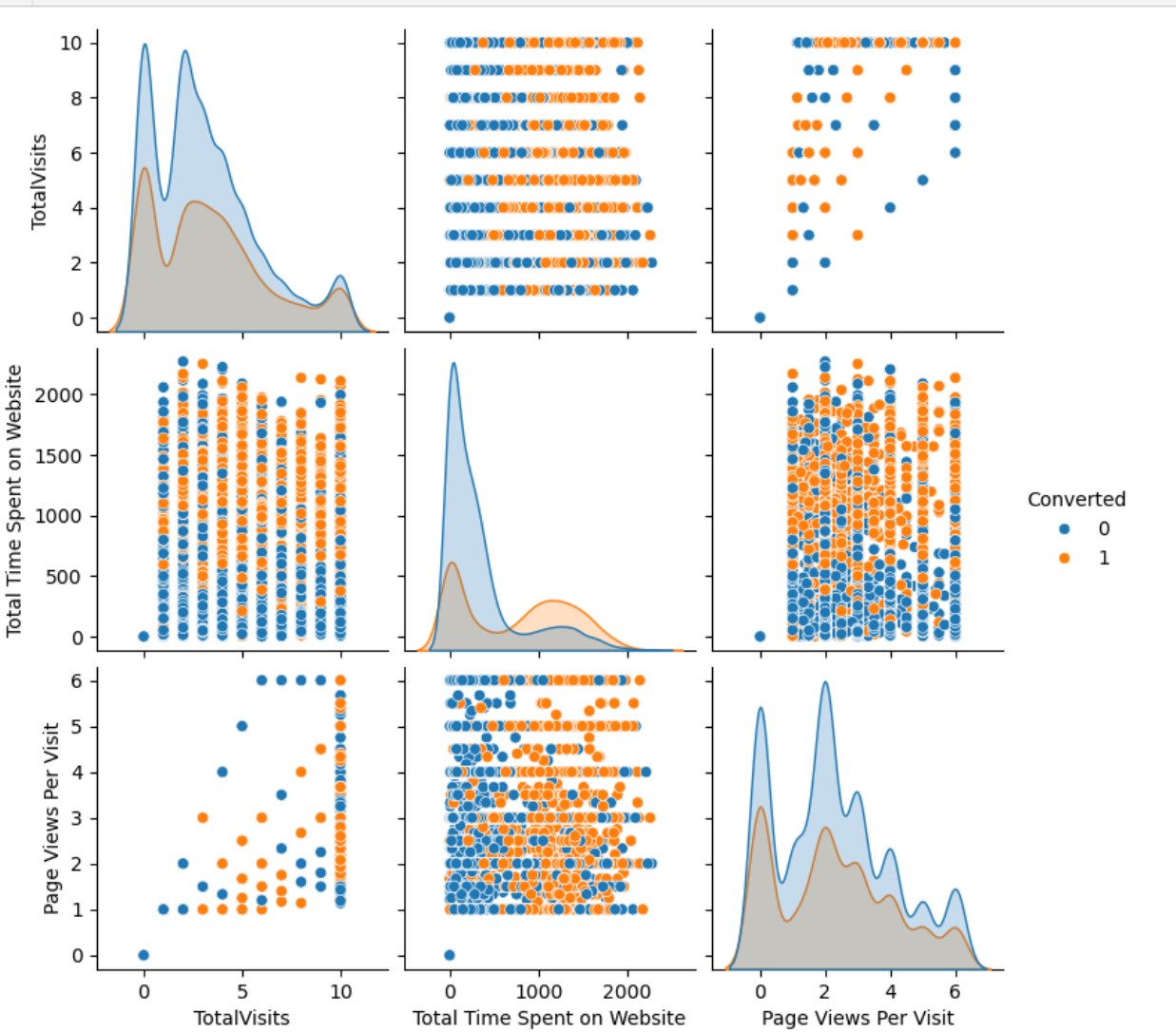
- Likely to get converted:
  - Unemployed
  - Working professionals



Total Visits

Total Time Spent on Website

Page Views Per Visit

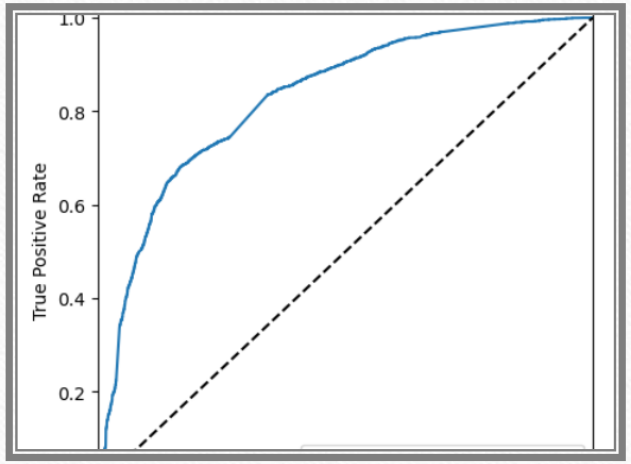


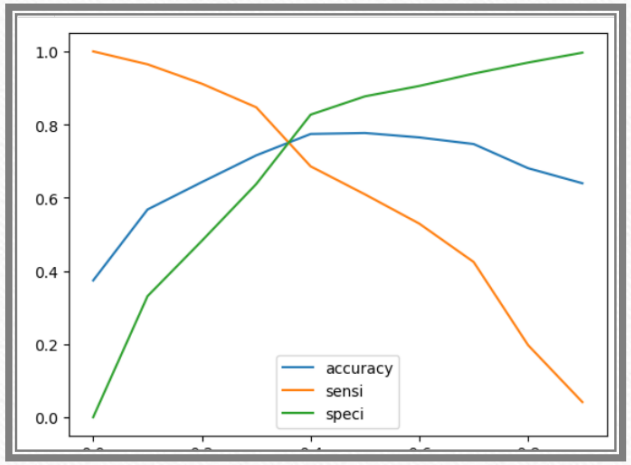
Approach

- **Data Preprocessing:**
- Used **Label Encoder** to convert categorical variables into numerical form.
- Handled missing values and ensured data consistency.
- **Train-Test Split:**
- Split the dataset into **30% train** and **70% test** sets for model training and evaluation.
- **Feature Selection:**
- Applied **Recursive Feature Elimination (RFE)** to select the top **15 important variables**.
- **Model Building:**
- Iteratively removed variables with **p-value > 0.05** (statistical insignificance).
- Dropped variables with **VIF > 5** to reduce multicollinearity.
- **Model Evaluation:**
- Made predictions on **test data**.
- Calculated **accuracy** and other relevant metrics.

Observations

- The model has 78% accuracy
- With a cut off as 0.5 we have accuracy of 78%, sensitivity of 61% and Specificity of 88%
- Graph 1: The area under ROC curve is 0.83 which is a good value
- Graph 2: It is visible that optimal cut off is at 0.35





Conclusion (1/2)

- The top three variables that contribute towards the result are:
  - Total Time spent on website
  - Lead Source with elements Google
  - Page Views Per website
- The top 3 Categorical variables which increase probability are:
  - Lead Source with elements Google
  - Lead Source with elements Direct traffic
  - Lead Source with Olark chat

Conclusion (2/2)

- Last Activity:
  - Most of the lead have their Email opened as their last activity.
  - Conversion rate for leads with last activity as SMS Sent the highest.
  - Leads to whom SMS are sent are most likely to get converted
- Maximum leads are from India
- Likely to get converted are working professionals and unemployed