Lead Scoring Case Study Using Logistic Regression

By :
Mohammed Omer Ahmed
Paulami Sur Roy
Pratik P Punyawant

Contents

- Problem Statement
- Business Objectives
- Problem Approach
- Data Cleaning
- EDA
- Model Evaluation
- Observations
- Conclusion

Problem statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company classifies that individual as a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%. Now, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads.
- If they successfully identify this set of leads, 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

Business Objectives

- X Education requires the selection of the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Problem Approach

- Importing the data
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Scaling Features
- Splitting the data in Train and Test data
- Building the Logistic Regression model on train data
- Evaluating model using different measures and metrics
- Assigning Lead score for each lead
- Testing the model on the test set
- Measuring the accuracy and other model metrics

Data Cleaning - I

- Checking for duplicate values
- Dropping columns with more than 35% of missing values.
- Handling few of the columns that have 'select' value in them:
  - Specialization :

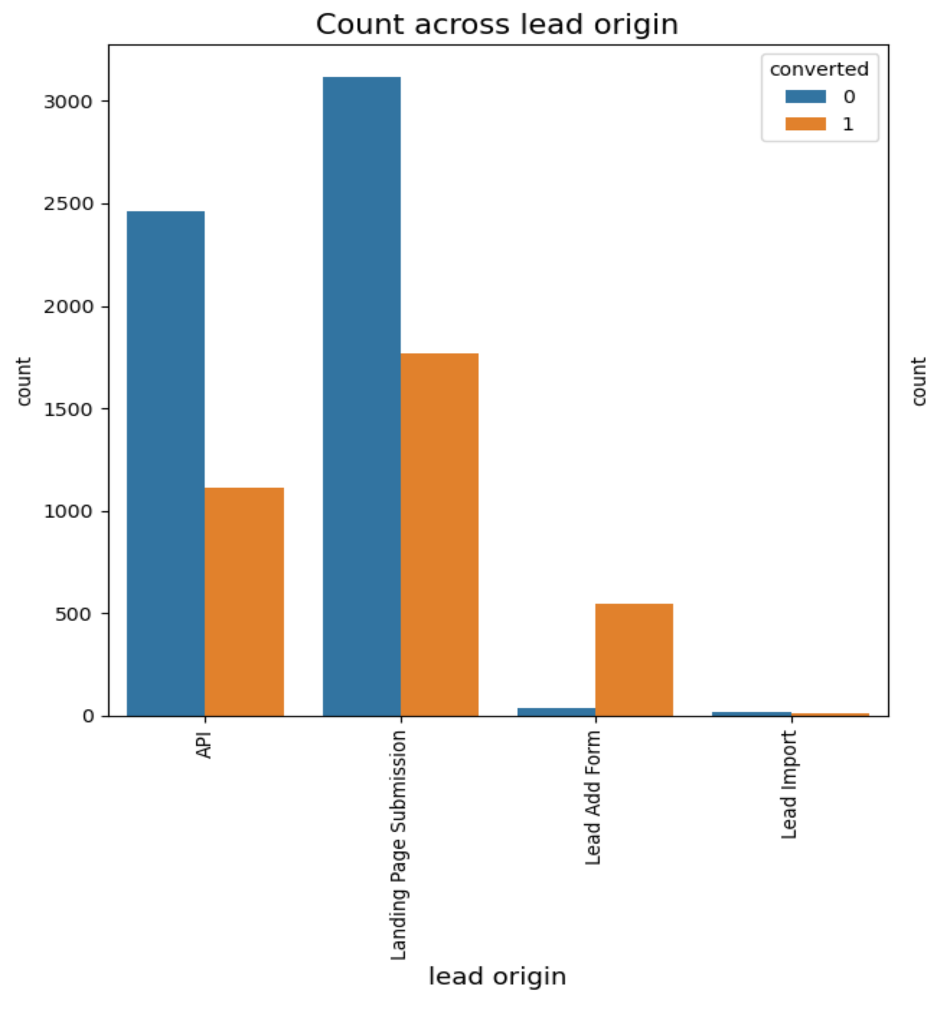
    Replacing 'select' values here with 'Not Specified'
  - How did you hear about x education & lead profile :

    These columns have high number of null values and also the value of the category 'select' is high compared to other categories, due to this we dropped these columns as they are highly kewed and might affect our analysis.

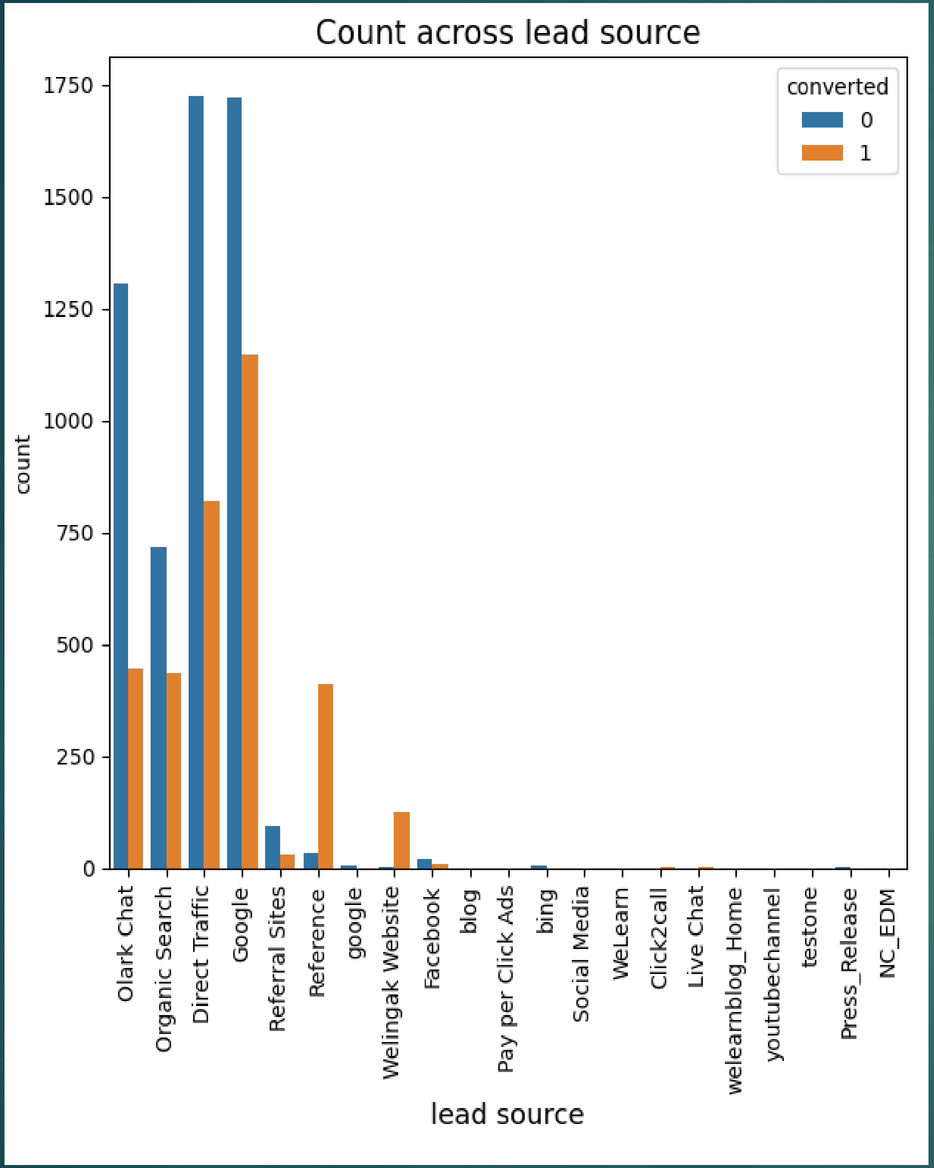
Data Cleaning - II

- Dropping columns that are heavily skewed:
  - Do not call
  - Search
  - Magazine
  - Newspaper article
  - X education forums
  - Newspaper
  - Digital advertisement
  - 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
  - Country
  - City
  - What matters most to you in choosing a course

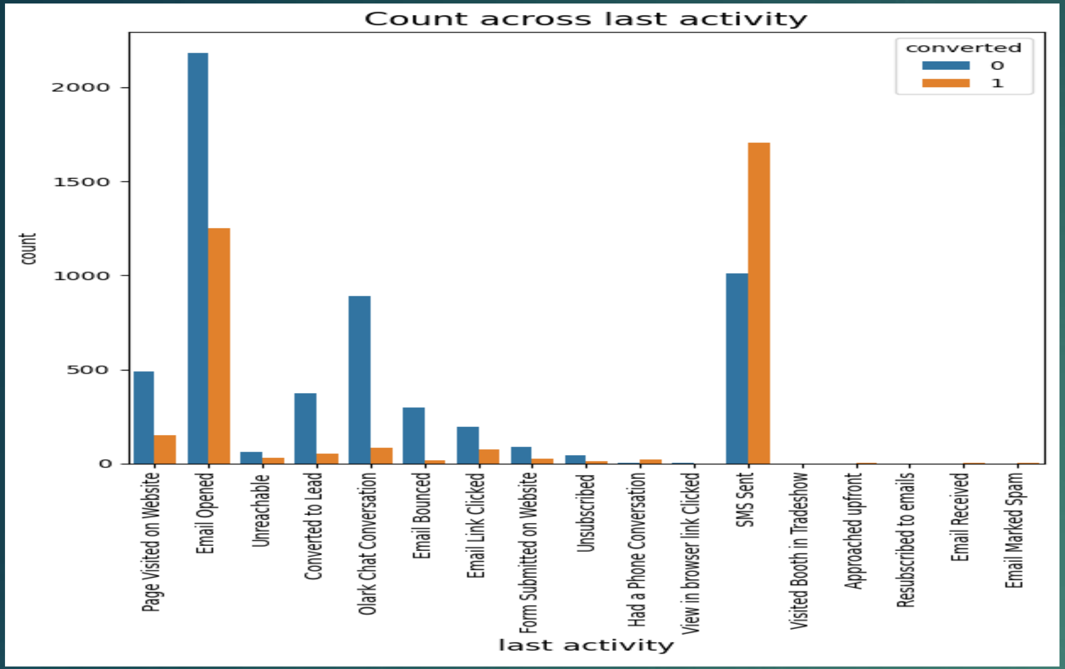
EDA



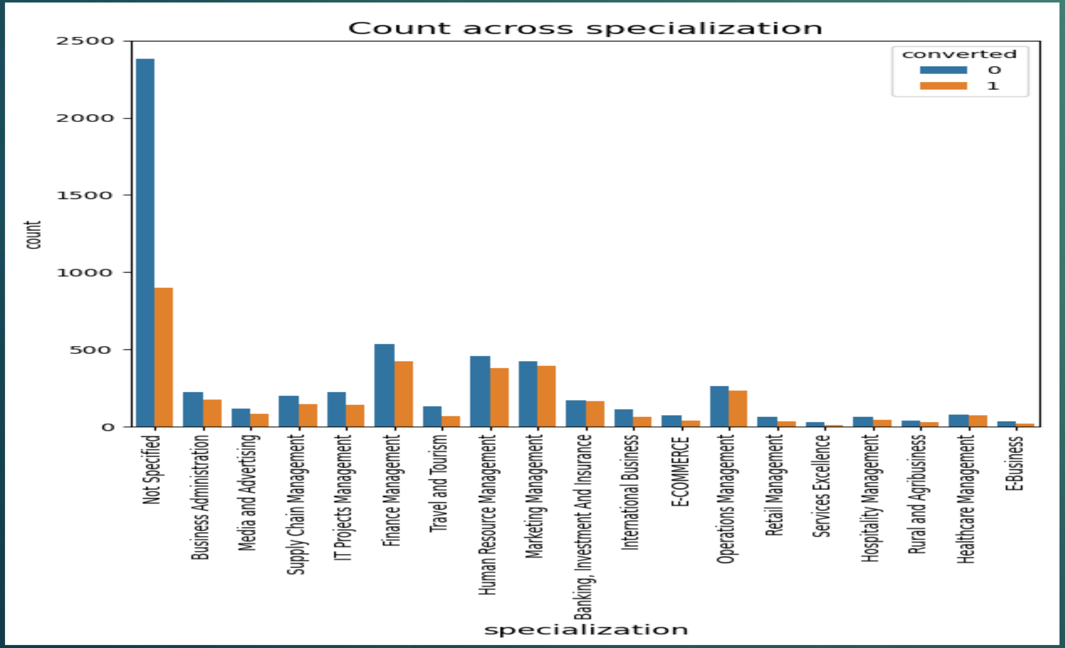
- Lead Origin:
  - For 'Lead Add Form' successful conversion is more than not converted
  - Count of 'Lead Import' is less.
  - To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission and generate more leads from Lead Add Form.



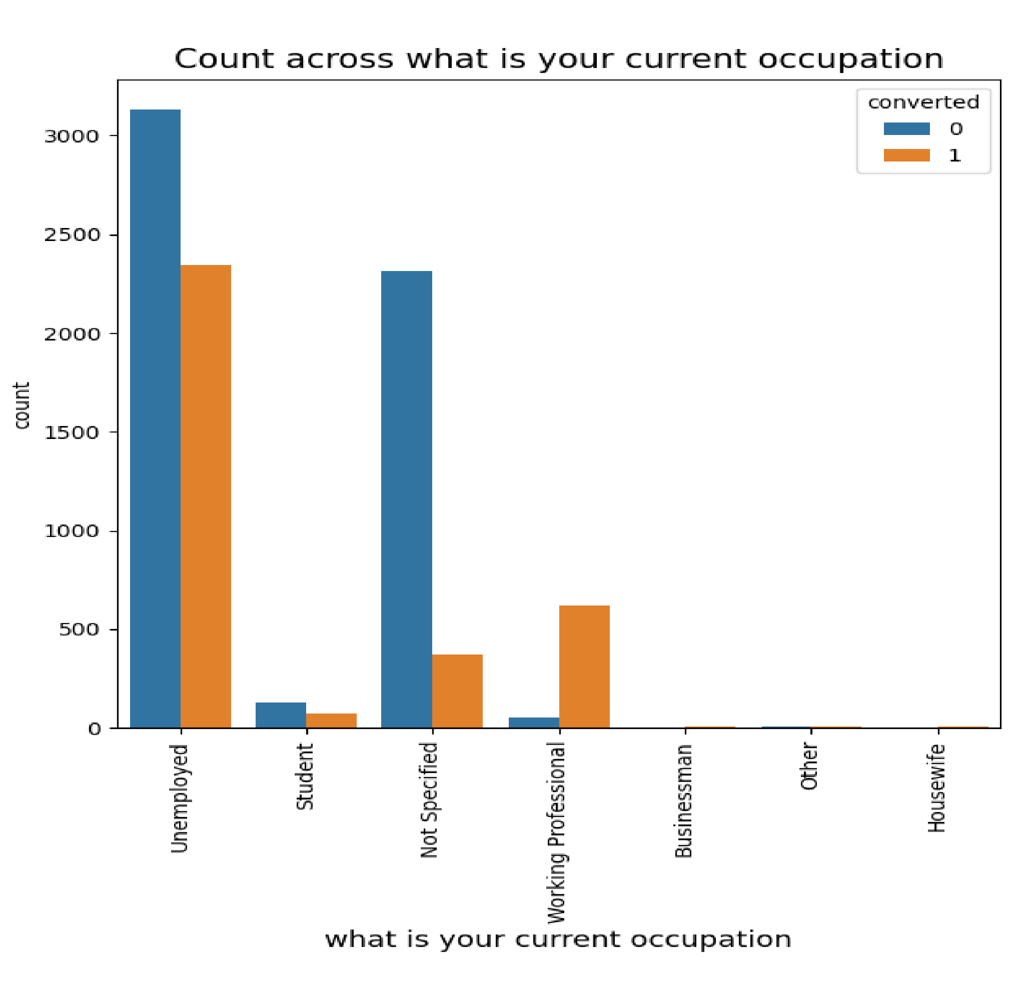
- Lead Source:
  - Google and Direct traffic generate maximum number of leads.
  - Conversion rate of 'Reference' and 'Welingak Website' leads is high.
  - To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.



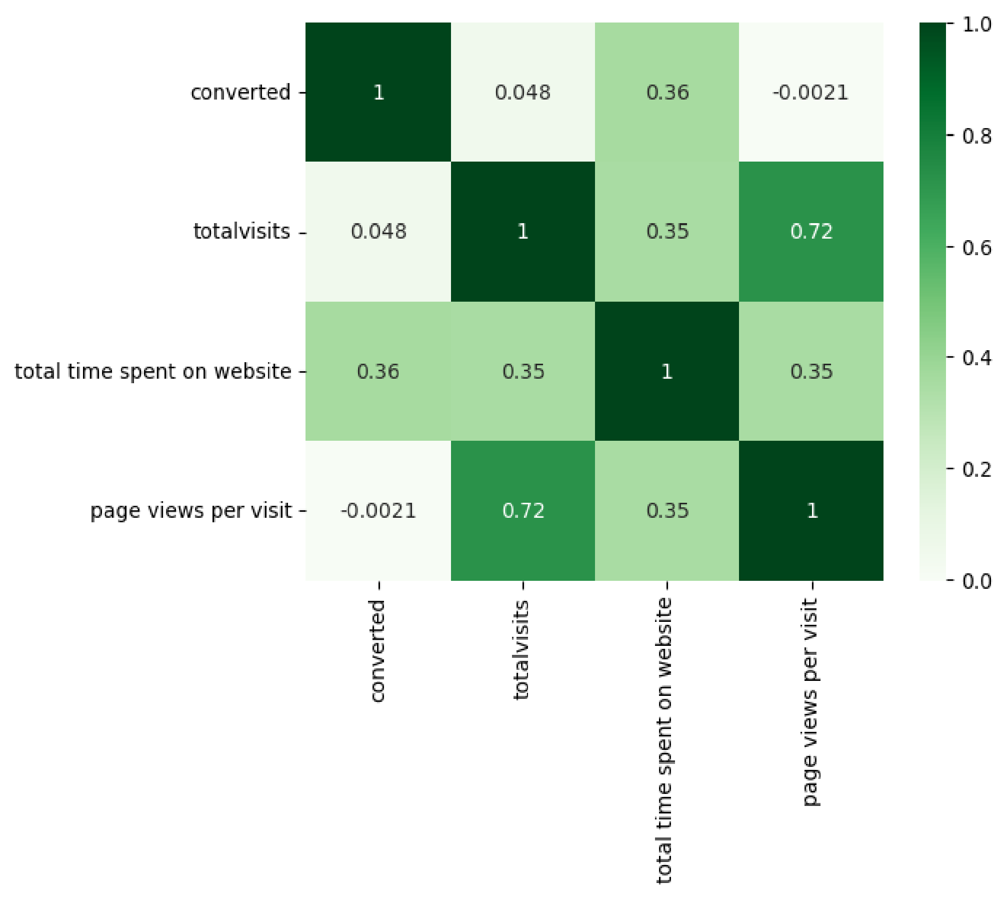
- Last Activity:
  - Highest count of last activity of leads is 'Email Opened'
  - Leads to whom the last activity was 'sms sent' have a higher conversion rate



- Specialization:
  - ❑ Count of various 'Management' specialization is high.
  - ❑ 'Marketing management' has the highest conversion rate



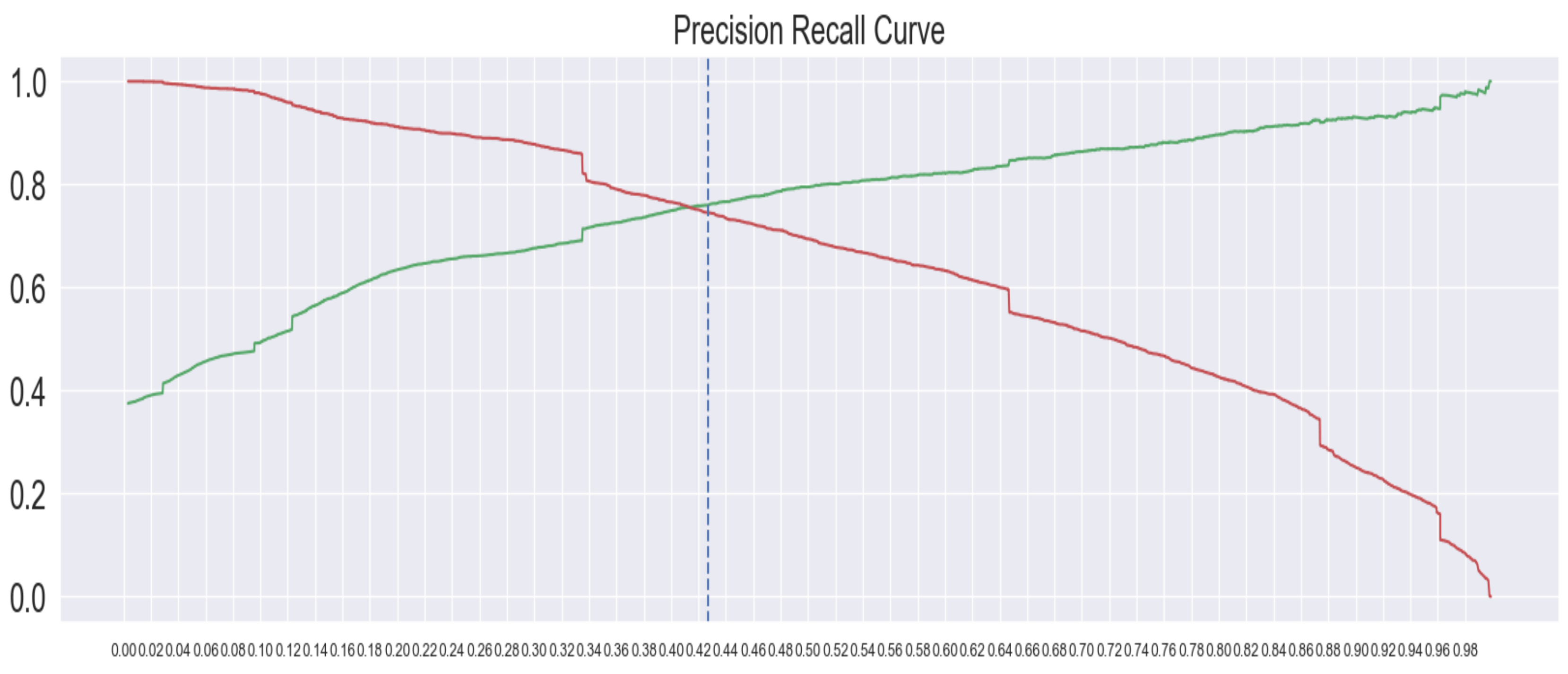
- What is your current Occupation:
  - The highest count is of unemployed people as they could be students
  - Conversion rate is higher for 'Working Professionals'.
  - To improve overall lead conversion rate, we need to focus more on improving lead conversion of 'unemployed people' such as students and generate more leads from 'Working Professionals'.

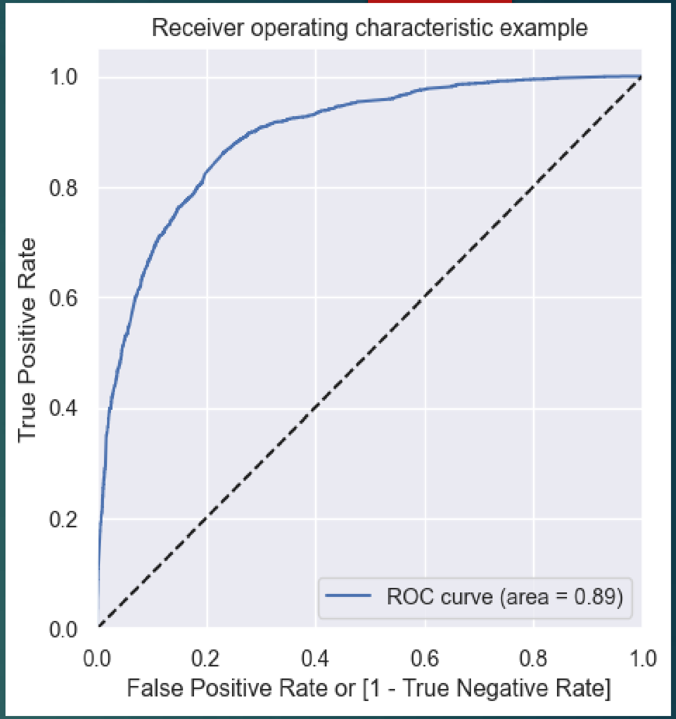


- Correlation:
  - ❑ 'Total Visits' and 'Page Views per Visit' are highly correlated with correlation of .72
  - ❑ 'Total Time Spent on Website' has correlation of 0.36 with target variable 'Converted'.

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 of Train : Test Data.
- Use RFE for Feature Selection and running it 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 ~80%.





- ➢ We are getting a good value of 0.89 as the area under the ROC Curve, indicating a good predictive model as ROC Curve should be as close to 1 as possible.

- ➢ From above 'Precision Recall Tradeoff Curve ' we can see that cutoff point is 0.427. Using this threshold value for Data Evaluation

Observations

- Positive Coefficients: Predictors with positive coefficients increase the likelihood of conversion.
  - Lead origin_Lead Add Form: Strong positive effect (Coef: 3.8049).
  - Lead source_Olark Chat: Positive effect (Coef: 1.1843).
  - Last activity_Had a Phone Conversation: Positive effect (Coef: 2.2467).
  - Last activity_SMS Sent: Positive effect (Coef: 1.2895).
  - Last activity_Unsubscribed: Positive effect (Coef: 1.0815).
  - Occupation_Working Professional: Positive effect (Coef: 2.5511).
  - Total time spent on website: Strong positive effect (Coef: 1.108).

- Negative Coefficients: Predictors with negative coefficients decrease the likelihood of conversion.
  - Lead origin_Landing Page Submission: Negative effect (Coef: -0.3402).
  - Last activity_Converted to Lead: Negative effect (Coef: -1.1218).
  - Last activity_Email Bounced: Negative effect (Coef: -1.0678).
  - Last activity_Olark Chat Conversation: Negative effect (Coef: -1.5616).
  - Occupation_Not Specified: Negative effect (Coef: -1.2776).
  - Do not email: Negative effect (Coef: -1.1981)

- Evaluation Metrics for the train Dataset:
  - Accuracy : 80.94%
  - Sensitivity : 79.25%
  - Specificity : 81.96%
  - Precision : 72.51%
  - Recall : 79.25%

- Evaluation Metrics for the test Dataset:
  - Accuracy : 79.91%
  - Sensitivity : 77.87%
  - Specificity : 81.17%
  - Precision : 71.98%
  - Recall : 77.87%

Recommendations

- To improve the potential lead conversion rate X-Education will have to mainly focus on the important features responsible for good conversion rate which are:
  - **Lead Source_Welingak Website :** As conversion rate is higher for those leads who got to know about course from 'Welingak Website',so company can focus on this website to get more number of potential leads.
  - **Lead Origin_Lead Add Form**: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it to get more number of leads cause have a higher chances of getting converted.
  - **What is your current occupation_Working Professional :** The lead whose occupation is 'Working Professional' having higher lead conversion rate ,company should focus on working professionals and try to get more number of leads.
  - **Last Activity_SMS Sent:** Lead whose last activity is sms sent can be potential lead for company.
  - **Total Time Spent on website:** Leads spending more time on website can be our potential lead.

THANK YOU!