2023

Lead Scoring

Case Study

PROBLEM STATEMENT

- X Education is an education company selling online courses to industry professionals.
- They receive a significant number of leads through website visits, form submissions, and referrals.
- The company's current lead conversion rate is only 30%, indicating room for improvement.
- X Education wants to identify the most potential leads, referred to as 'Hot Leads,' to increase their conversion rate to approximately 80%.
- The objective is to build a lead scoring model that assigns scores to leads based on their likelihood of converting into paying customers.
- By focusing on leads with higher scores, the sales team can prioritize their efforts and increase the overall conversion rate.

APPROACH

- Domain variable knowledge
- Data and Metadata structure review
- Missing value & Outlier check
- Univariate analysis
- Bivariate analysis
- Multivariate Analysis
- Analysis of Merged Data

EDA

Overview of the dataset: 9000 data points with various attributes
Data cleaning and preparation:

- Handling null values and dropping irrelevant columns
- Handling categorical variables with the level 'Select'
- Dropping columns with a single dominant value
- Exploratory data analysis: Visualizing data distribution, correlations, and box plots

Feature engineering and encoding:

- Creating dummy variables for categorical features using one-hot encoding
- Handling the 'Specialization' variable separately due to the 'Select' level
- Splitting the dataset into training and testing sets using train_test_split()
- Scaling numerical features using MinMaxScaler()

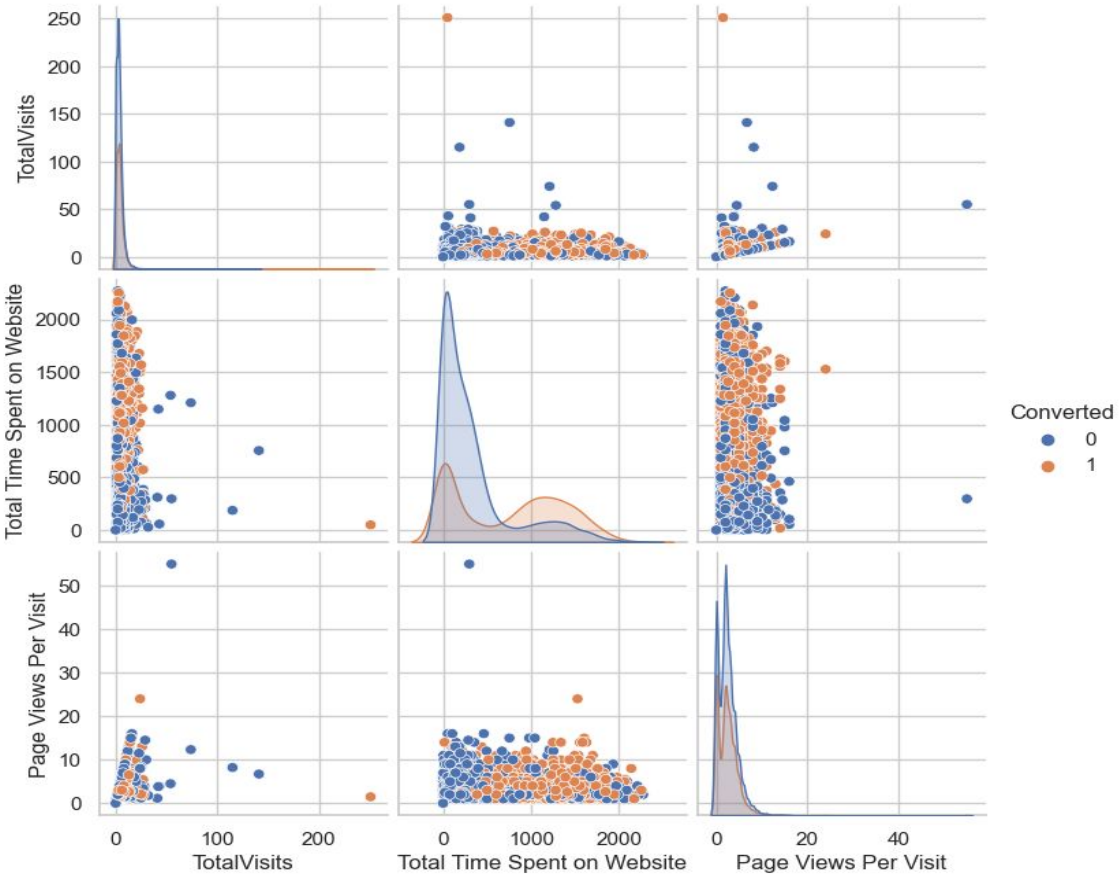
Building the logistic regression model:

- Importing necessary libraries
- Fitting the model on the training data
- Predicting on the test data
- Evaluating model performance using various metrics (classification report, accuracy, precision, recall, etc.)

Recommendations for the education company based on the logistic regression model:

- Assign lead scores between 0 and 100 to prioritize potential leads
- Focus on leads with higher scores for improved conversion rates
- Continuously update and refine the model as per changing requirements

EDA



The conversion rates were high for Total Visits, Total Time spent on website and page view per visit.

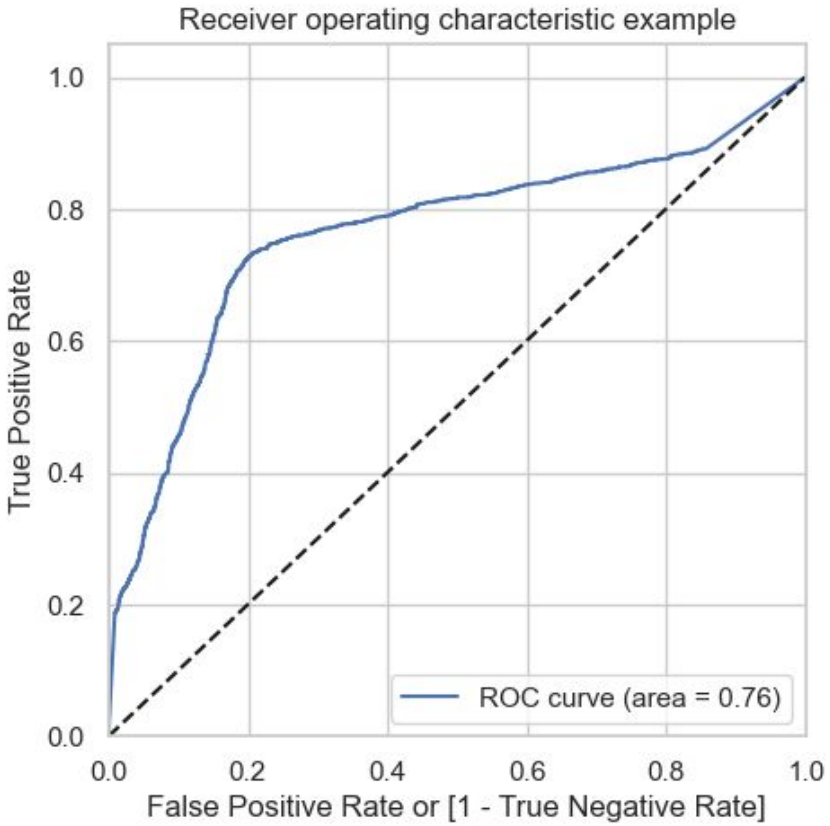
Observations

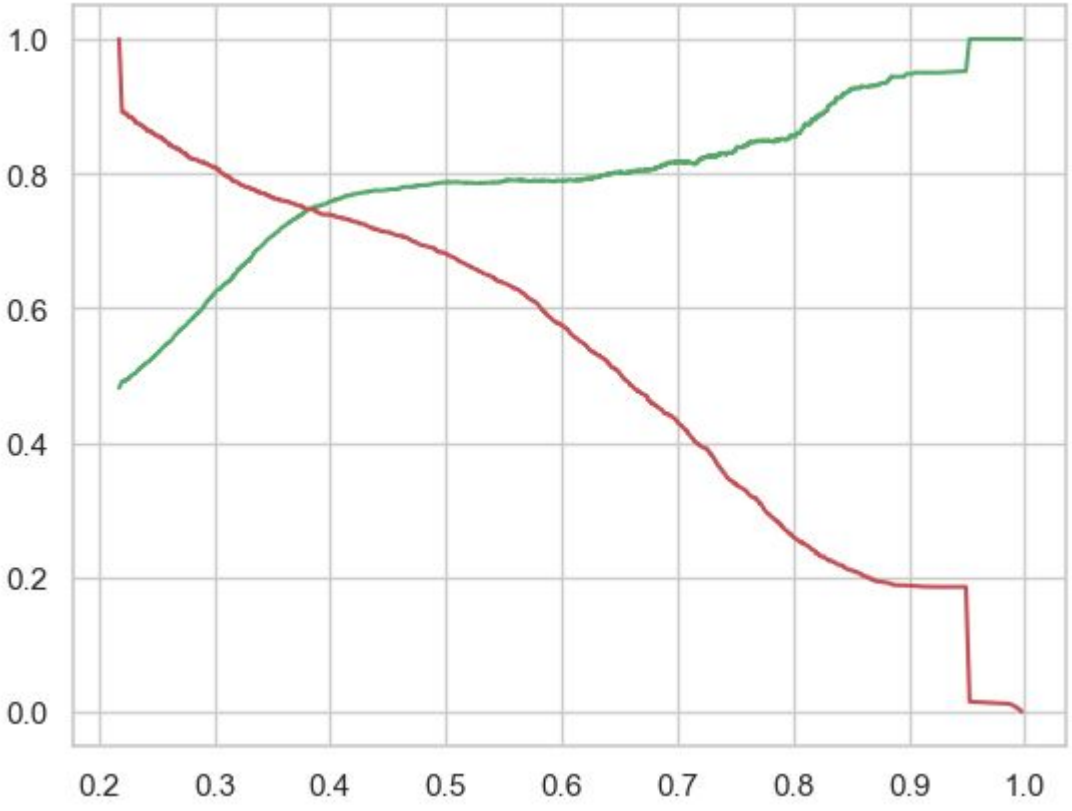
- Maximum Conversion happened from landing page through lead origin
- Maximum conversion happened from Email sent and calls made through Do not call
- Major conversions were from google in lead source.
- There was not much impact on conversion rates through Search, digital ads and recommendations
- Employed people were the ones that were converted more.

Variables impacting the conversion rate

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity - Unreachable

Model Evaluation





The Graph depicts an optimal cutoff of 0.42 based on precision and recall.

Precision : 79%
Recall : 71%

Conclusion

- Model performance:
  - Accuracy: Approximately 81%
  - Conversion rate: Around 80% in the train set
- Top contributing variables for lead conversion:
  - Total time spent on website
  - Lead add form from lead origin
  - Had a phone conversation from last notable activity
- Overall assessment: The model is deemed to be effective in identifying potential leads with a high likelihood of conversion.

THANK YOU