Lead Scoring Case Study

Analysis

Problem Statement

1. X Education is struggling with the conversion of customers though there are some potential leads

2. Model needs to be created for locating hot leads which can be targeted for conversion

3. Feedback on what should be focused on for future prospects

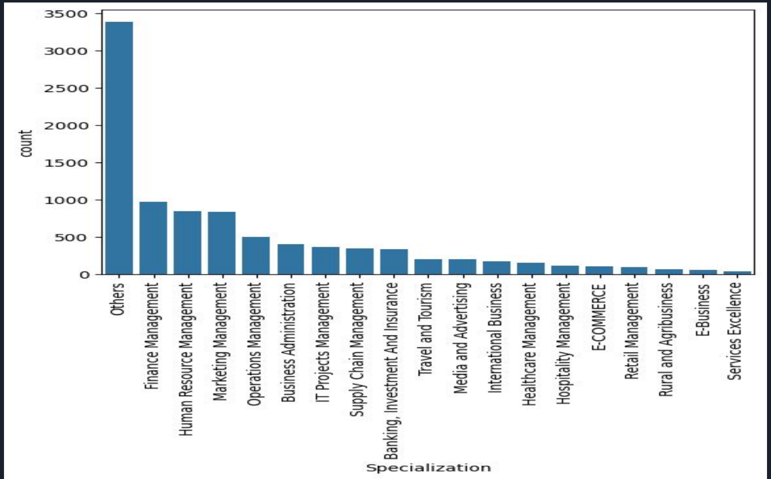
Solution Approach

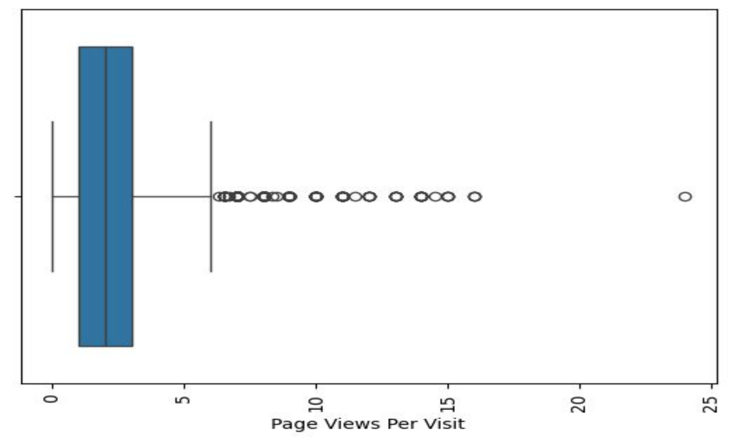
- Data Description and Sanity
  - Data Checking
  - Discovering Null values and outliers
  - Checking feature variance
- Exploratory Data Analysis
  - Univariate Analysis
  - Bivariate Analysis
  - Feature Extraction
- Data Modelling
  - Handling Categorical data
  - Scaling
- Modelling
  - RFE
  - Regression Modelling
- Model Evaluation
  - Checking cutoff
  - ROC-AUC

Data Handling

1. Data Provided: 36 features, 9240 entries
2. 7 features having extensive null values
3. Features about Asymmetrique indexes and lead profile dropped due to null values
4. Features based on search result: “Search”, “Magazine” etc. have too less variance and so combined under 1 feature
5. Features like Country binned to create 2 major classes due to imbalance.
6. “Prospect Id” and “lead number ” dropped as they are primary keys

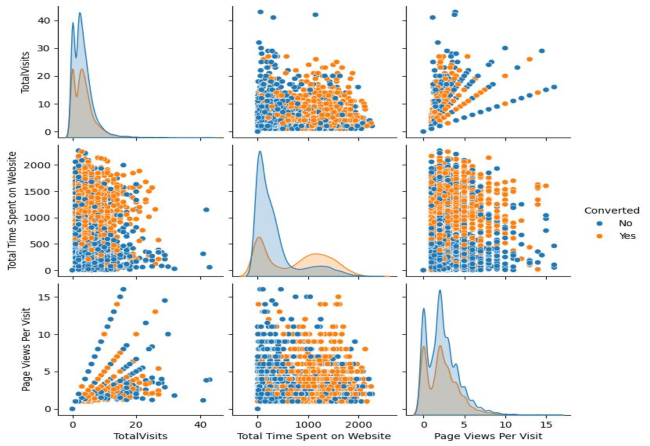
Univariate Analysis

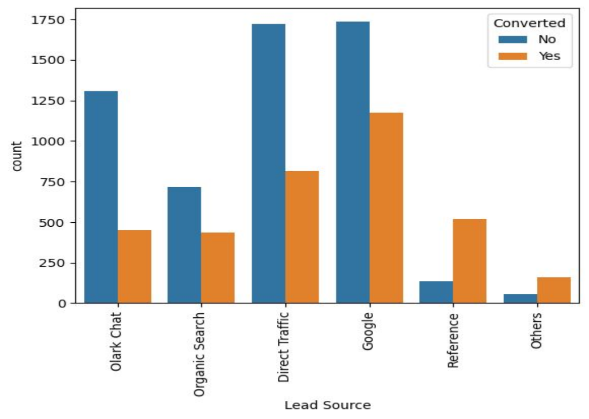




Distributions of categorical data and numerical data checked through count plots and box plots respectively and cleaning steps taken.

Bivariate Analysis





Bar charts, scatter plots and heat maps used to study internal correlations and impact of the features on the target variable.

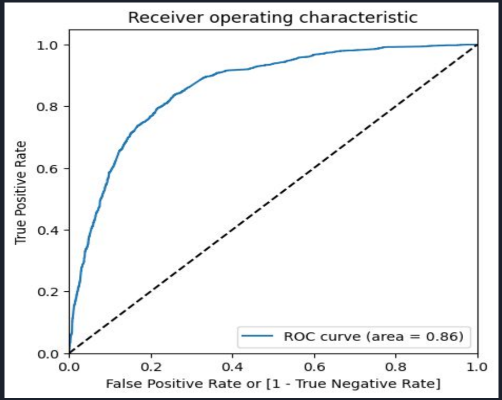
Data Modelling

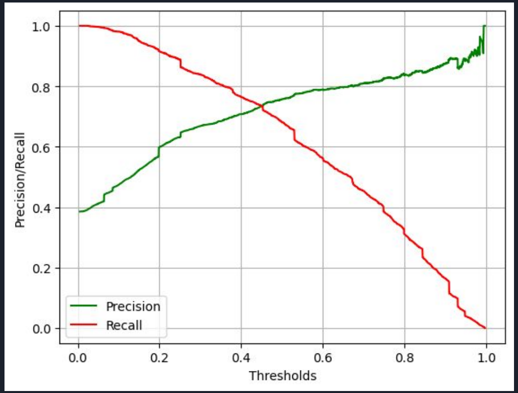
1. Finally reached to a data set of 15 features using Recursive feature extraction
2. Numerical data like Total Visits and Total time spent on website min-max scaled
3. Categorical features are one hot encoded.
4. Total 50 features after encoding. Reduced to 40 based on low correlation with target

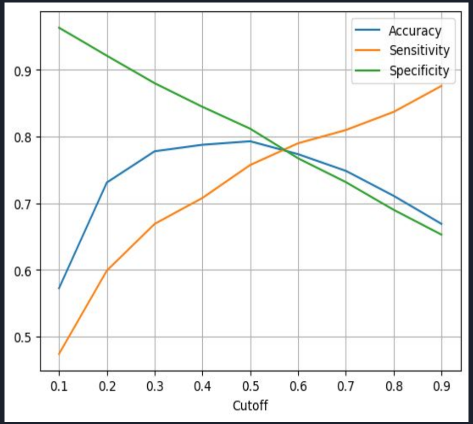
Predictive Modelling

1. Logistic Regression Model established
2. 70:30 - train: test data established
3. Final model input size : 9 features
4. Features eliminated based on high p-values and VIF
5. Established cutoff at 56% probability for prediction

Learning Curves







Prediction Results

1. Train Data

    a. 79% accuracy on Train data set

    b. 78% precision and 75% recall on training data

2. Test data

    a. 78% accuracy on Train data set

    b. 79% precision and 75% recall on training data

Feedback

X education should focus on:

1. Unemployed people
2. People coming from forms
3. Olark Chats, SMS and Emails
4. Maximizing people's visit time on website
5. Maximize total visits for people

THANK YOU