Improving Lead Conversion at X Education with Logistic Regression

Harsh Bhardwaj

Problem statement

- X Education has a low lead conversion rate (around 30%)
- High volume of leads suggests inefficiency in lead nurturing
- Goal: Increase conversion rate to 80% by identifying "hot leads"

Data and Approach

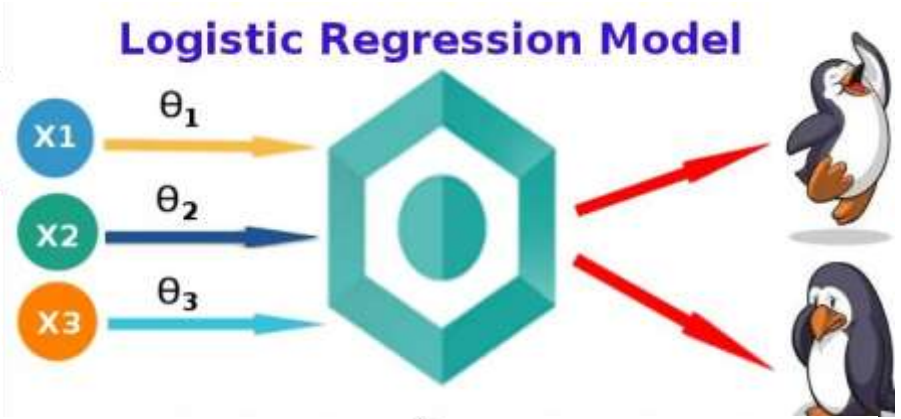
Dataset: 9,000 leads with various attributes (source, website activity, etc.)

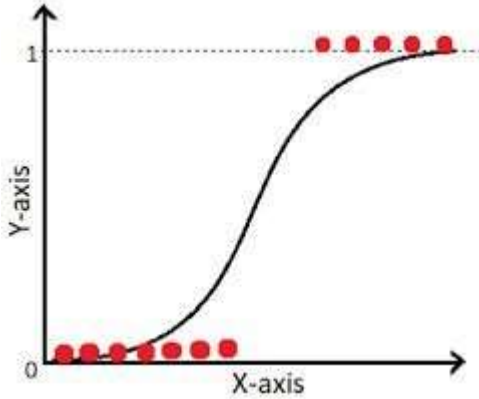
Target Variable: "Converted" (1=converted, 0=not converted)

We used a historical dataset of 9,000 leads provided by X Education. The data contains various attributes about each lead, such as their lead source, time spent on the website, and total visits. The target variable is a simple binary variable indicating whether the lead converted into a paying customer (1) or not (0).

Model Building Process

- Data Cleaning and Preprocessing
  - Address missing values
  - Remove unnecessary and duplicate data
  - Convert categorical features
- Exploratory Data Analysis (EDA)
  - Understand feature distribution
- Feature Engineering and Selection
  - Create dummy variables
  - Use RFE for feature selection
- Model Building and Evaluation
  - Train logistic regression model
  - Address multicollinearity
  - Evaluate model performance (test set)
  - Optimize probability cutoff

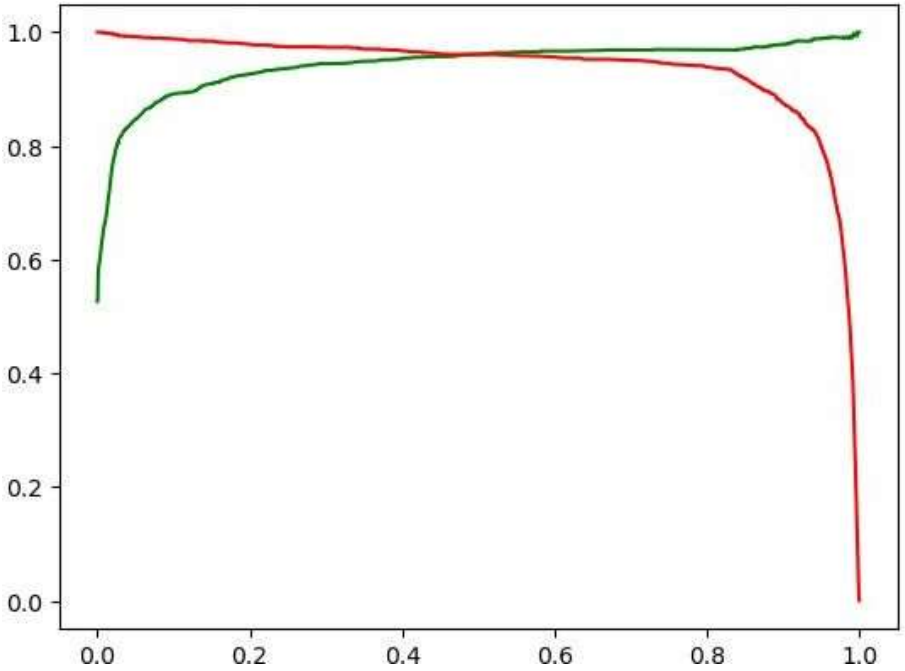




Model Performance

- On the training set:
  - Accuracy: 97.2%
  - True Positive (TP): 1321
  - True Negative (TN): 1101
  - False Positive (FP): 39
  - False Negative (FN): 38
  - Sensitivity (Recall): 96.1%
  - Specificity: 96.6%
  - Precision: 96.17%
- On the test set:
  - Accuracy: 96.1%
  - True Positive (TP): 623
  - True Negative (TN): 508
  - False Positive (FP): 24
  - False Negative (FN): 45
  - Sensitivity (Recall): 95.81%
  - Specificity: 95.45%
  - Precision: 96.67%

ROC Curve



- Precession
- Recall