LEAD SCORING CASE STUDY

Presented By:
•Nimisha Kashyap
•Sakshi Priyadarshi
•Darshith P K

PROBLEM STATEMENT

- X Education specializes in offering online courses targeted at professionals within various industries.
- Despite receiving numerous leads, X Education struggles with a notably low lead conversion rate. For instance, out of 100 leads obtained in a day, only around 30 are successfully converted.
- In a bid to streamline operations, the company aims to pinpoint the most promising leads, commonly referred to as 'Hot Leads'.
- Identifying this subset of leads is anticipated to boost the lead conversion rate, as the sales team can allocate more attention to communicating with potential leads rather than pursuing every lead indiscriminately.

Business Objective:

- X Education seeks to identify the most promising leads by constructing a model designed to recognize 'hot leads'.
- The development of this model is intended to facilitate the identification process.
- The company plans to deploy the model for future utilization, enhancing lead identification efficiency over time.

SOLUTION METHADOLOGY

- Conduct data cleaning and manipulation processes, including:
  - Identification and resolution of duplicate data.
  - Handling of NA values and missing data.
  - Elimination of columns containing a significant amount of missing values and deemed irrelevant for analysis.
  - Imputation of missing values if required.
  - Identification and handling of outliers in the dataset.
- Perform Exploratory Data Analysis (EDA) comprising:
  - Univariate analysis such as value counts and variable distributions.
  - Bivariate analysis including correlation coefficient calculations and examination of patterns between variables.
- Implement feature scaling, creation of dummy variables, and encoding of the data for further analysis.
- Utilize classification techniques, particularly logistic regression, for model development and prediction.
- Validate the model to ensure its reliability and accuracy.
- Present the finalized model along with relevant insights.
- Draw conclusions and provide recommendations based on the analysis conducted.

Data Manipulation

- The dataset comprises a total of 37 rows and 9240 columns.
- Features containing single values, such as "Magazine," "Receive More Updates About Our Courses," "Update me on Supply Chain Content," "Get updates on DM Content," and "I agree to pay the amount through cheque," have been excluded.
- Unnecessary columns like "Prospect ID" and "Lead Number" have been removed from the dataset.
- Upon examining the value counts for some object type variables, it was observed that certain features lacked sufficient variability and were thus dropped. These features include "Do Not Call," "What matters most to you in choosing course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," "Digital Advertisement," among others.
- Columns with more than 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile,' have been eliminated from the dataset.

Variable impacting the Conversion rate

- Total visits
- Total time spent on website
- Lead Source_Olark chat
- Lead origin_Lead Add form
- Lead Source_Wellingak Website
- Do not Email
- Lead source_Referral sites and etc.

Data Conversion

Normalization has been applied to numerical variables.

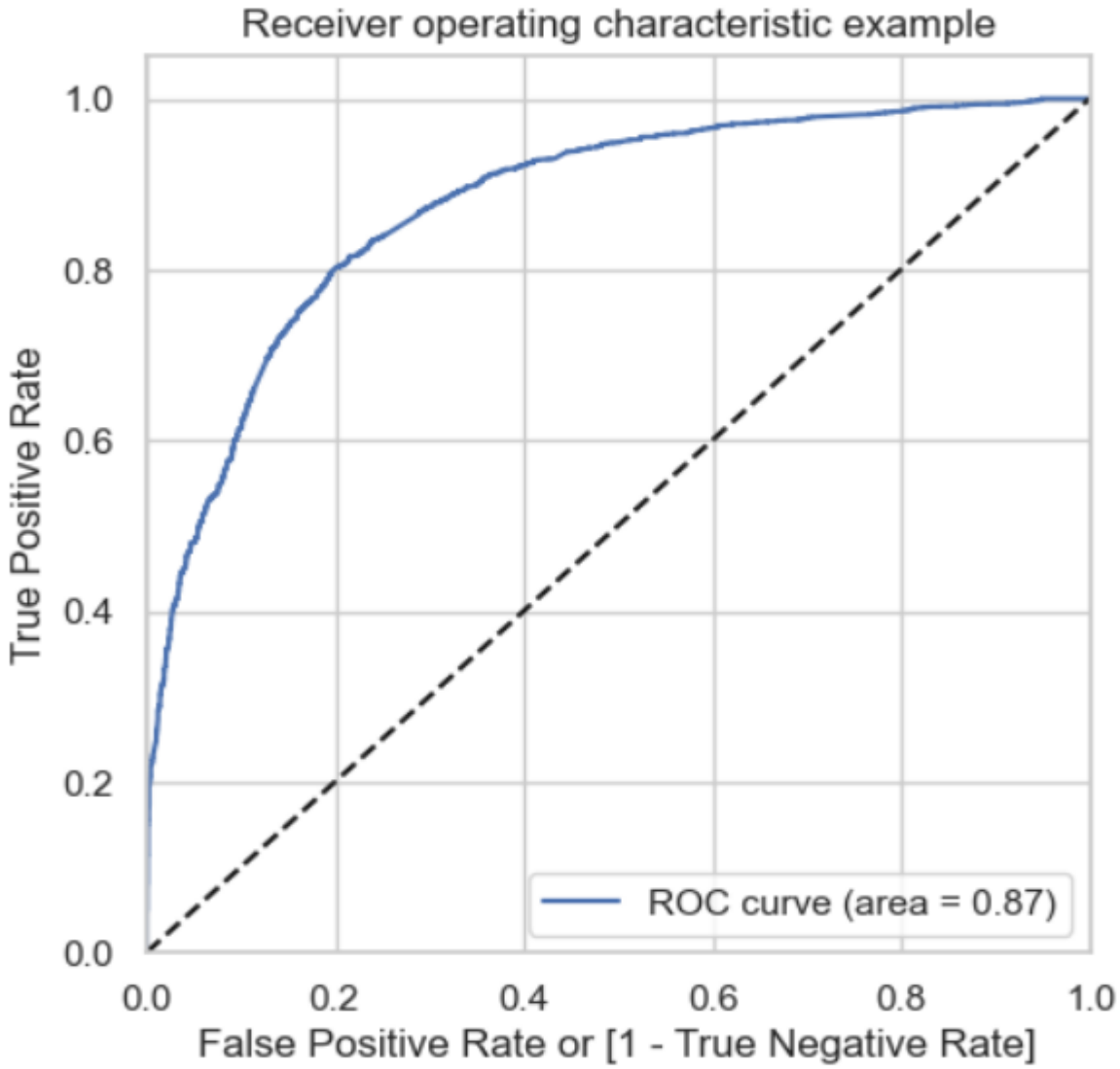
Object type variables have been encoded into dummy variables.

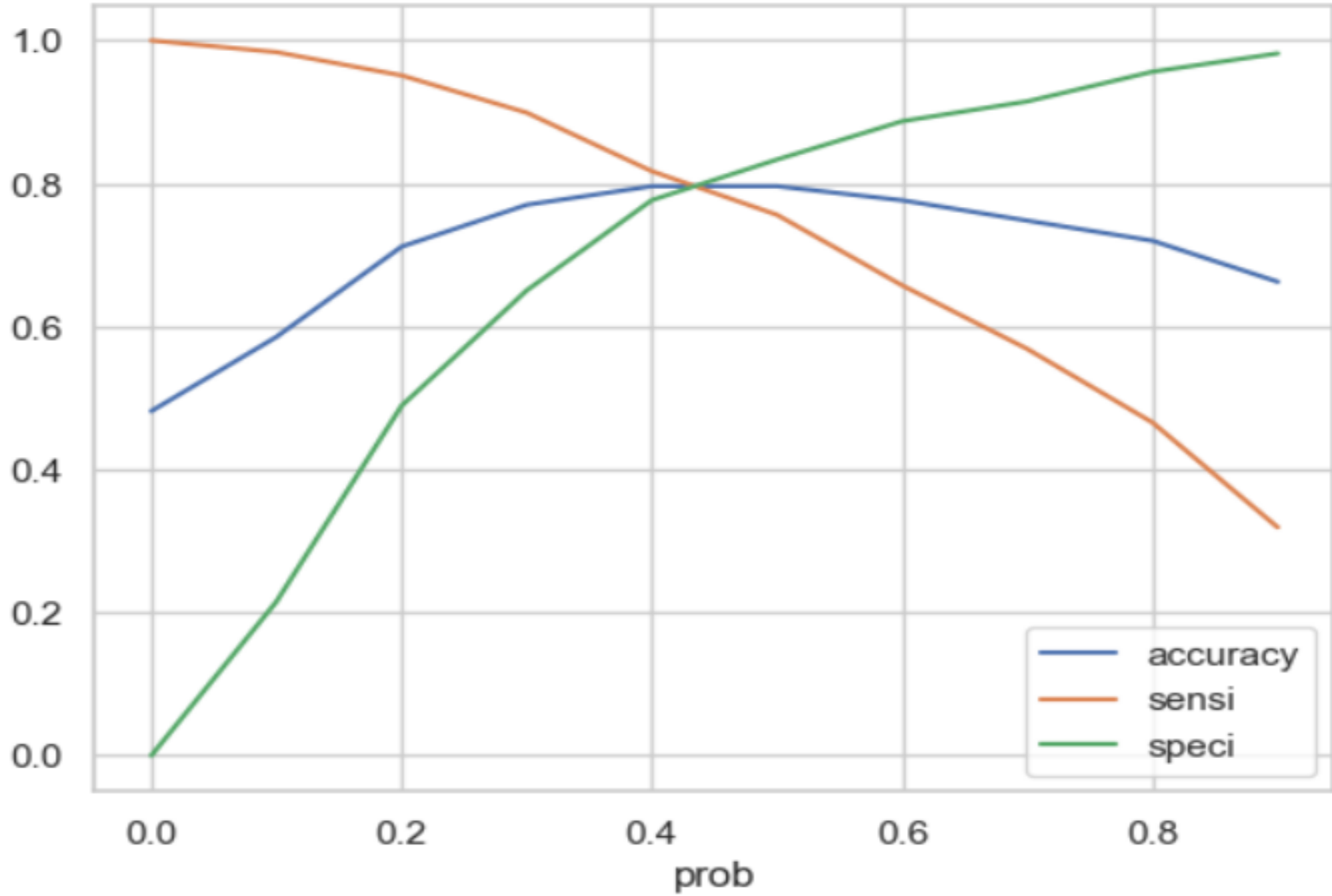
The dataset for analysis comprises 8792 rows and 43 columns.

Model Building

- Data is divided into training and testing sets.
- The initial step in regression involves a traintest split, selected at a ratio of 70:30.
- Recursive Feature Elimination (RFE) is employed for feature selection.
- RFE is executed with an output of 15 variables.
- The model is constructed by excluding variables with a pvalue exceeding 0.05 and a Variance Inflation Factor (VIF) greater than 5.
- Predictions are made on the test dataset.
- The overall accuracy achieved is 81%.

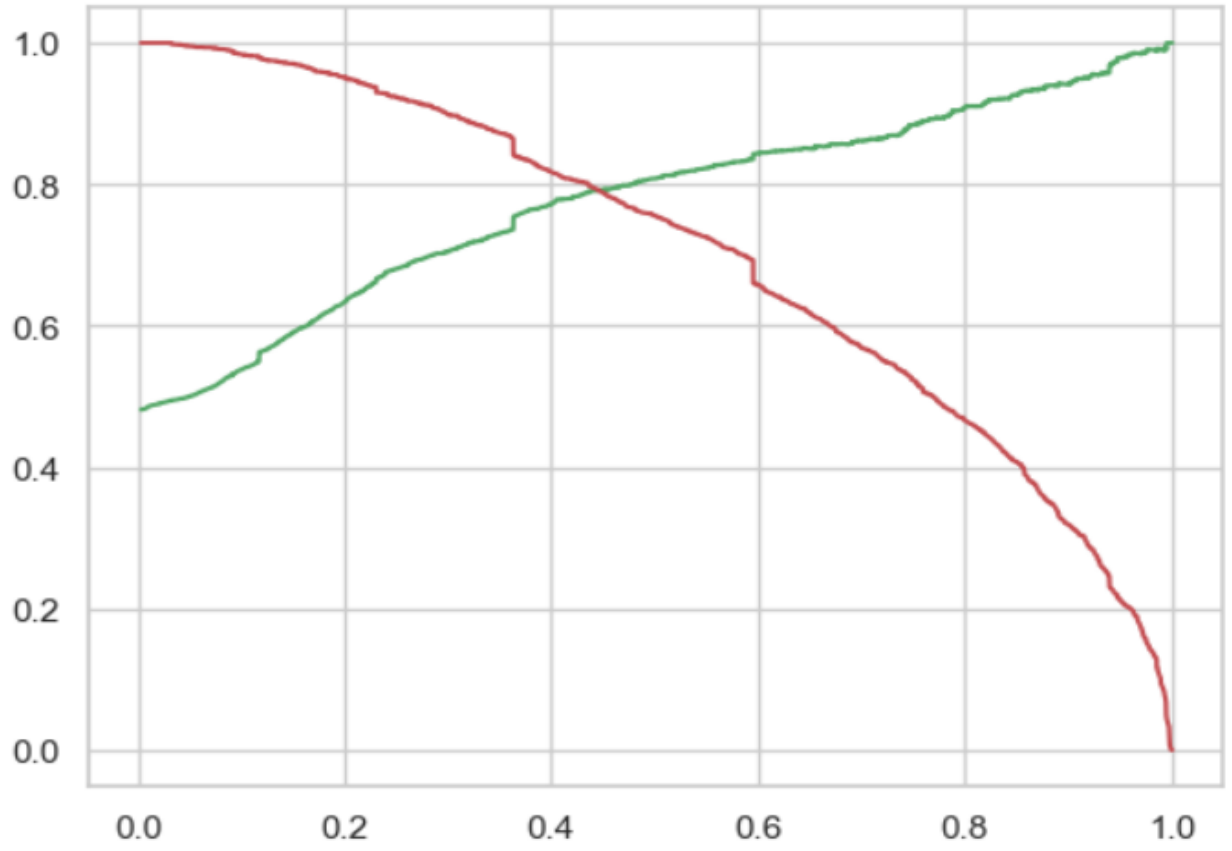
ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.87)


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
prob
accuracy
sensi
speci

- Upon examination of the ROC graphs provided above...
- Identifying the Optimal Cutoff Threshold
- The optimal cutoff probability represents the point where a balance between sensitivity and specificity is achieved.
- Noting that around 0.42, we achieve optimal values for all three metrics, we opt for 0.42 as our cutoff threshold.
- The accuracy stands at 0.79.
- Sensitivity is recorded at 0.80.
- Specificity is measured at 0.789.

Precision and Recall tradeoff



Precision :0.79
Recall: 0.795

Summary

- Initially, many leads were generated but only a few converted to paying customers.
- To increase conversion rates, nurture leads with education and communication.
- Identify high-potential prospects based on website engagement metrics.
- Maintain a list of leads for personalized updates.
- Tailor communication to address specific needs and interests.
- Focus on converted leads and gauge their interest through inquiries.
- Key factors influencing potential buyers include website engagement, lead source, last activity, lead origin, and occupation.
- Leveraging these insights can significantly improve conversion rates for X Education.