LEADS SCORING CASE STUDY

SUBMITTED BY :

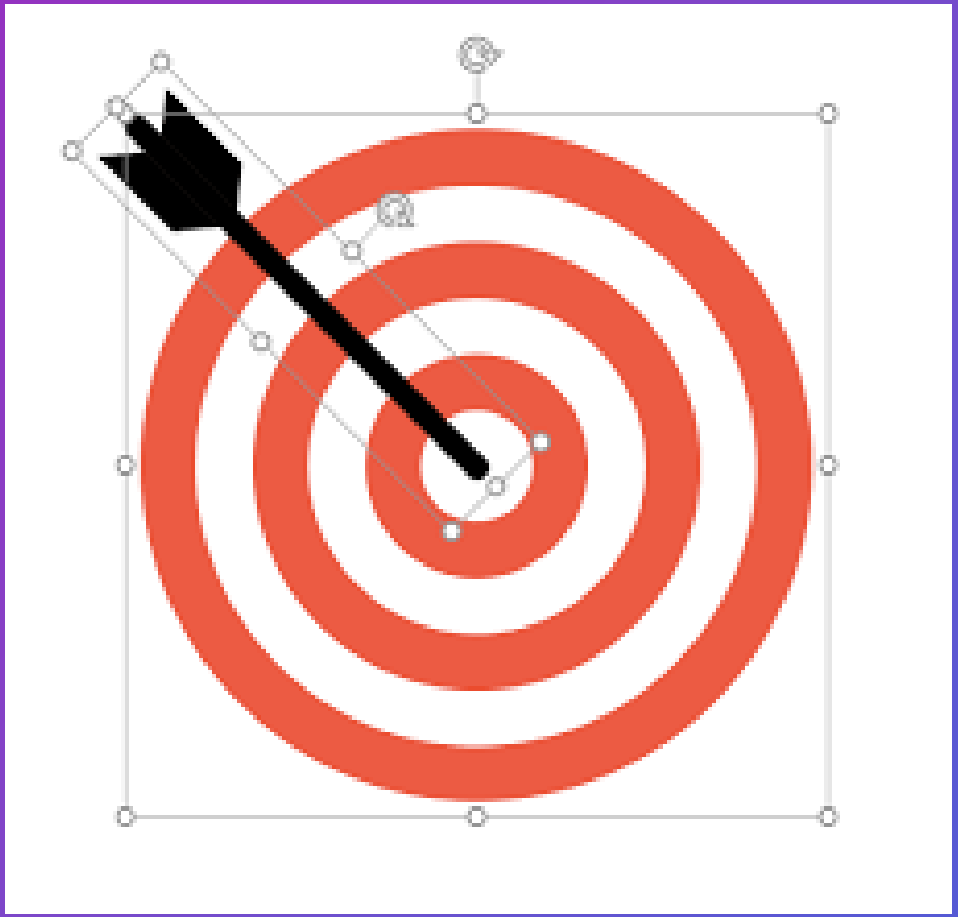
Shiva Chandra Kante, Krishnakumar V, Kamatchi M

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- Leads are acquired through website visits, from submissions, and referrals.
- Sales team follows up via calls and emails to convert leads.
- The typical lead conversion rate is around 30%.
- The company aims to identify high-potential leads (Hot Leads for efficiency.
- Focusing on Hot Leads can improve conversion rate and reduce effort on low-potential leads.

BUSINESS OBJECTIVE

- Develop a logistic Regression model to assign a Lead Score (0-100).
- Higher score – Hot Lead (High conversion probability).
- Lower score – Cold Lead (Low conversion probability).
- Ensure model flexibility to adapt to changing business requirements.
- Incorporate additional business problems as required.



APPROACH

1. Data cleaning and data manipulation.

1. Check and handle duplicate data.
2. Check and handle NA values and missing values and impute values, if necessary.
3. Drop columns, if it contains a large number of missing values and are not useful for the analysis. Check and handle outliers in data.

3. Model building

1. Feature scaling & Dummy variables and encoding of the data.
2. Classification technique: Used logistic regression for model making and prediction.
3. Calculated overall accuracy of the model

2. Exploratory Data Analysis (EDA)

1. Univariate data analysis: value count, distribution of variables, etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

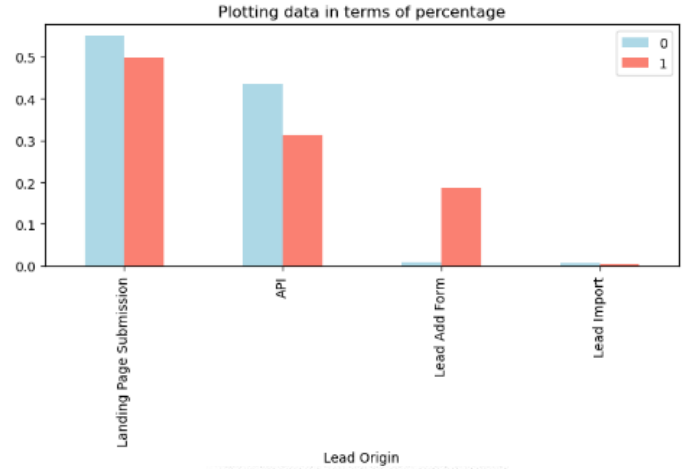
4. Model Evaluation

1. Used ROC curves between train and test data for evaluating the model
2. Assessed Accuracy, Sensitivity and specificity
3. Assessed Precision and Recall

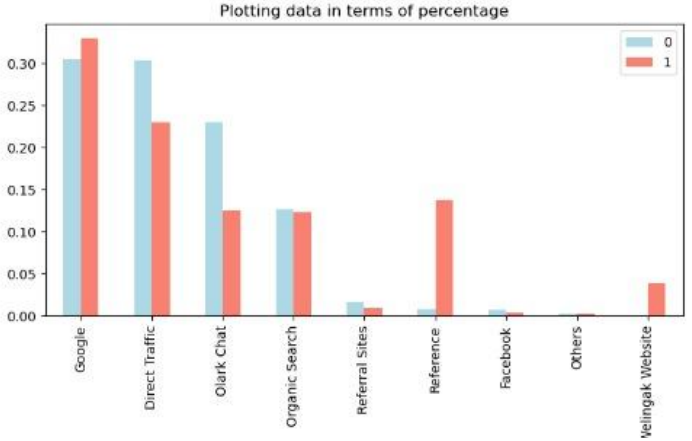
Exploratory Data Analysis

EDA

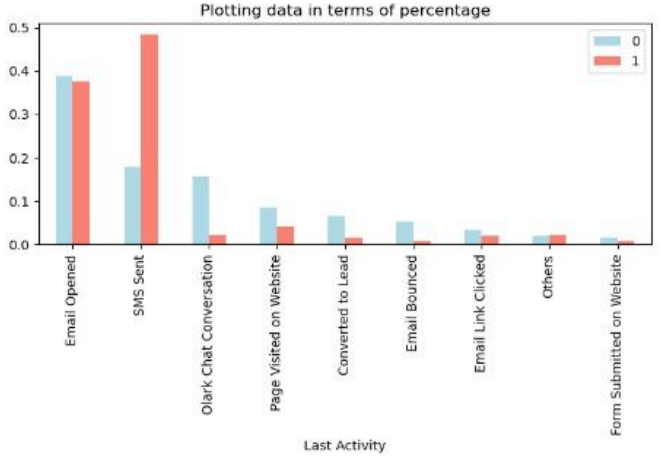
• Lead Origin: "Landing Page Submission" identified 53% of customers, while "API" identified 39%.



• Lead Source: 58% of leads come from Google & Direct Traffic combined.



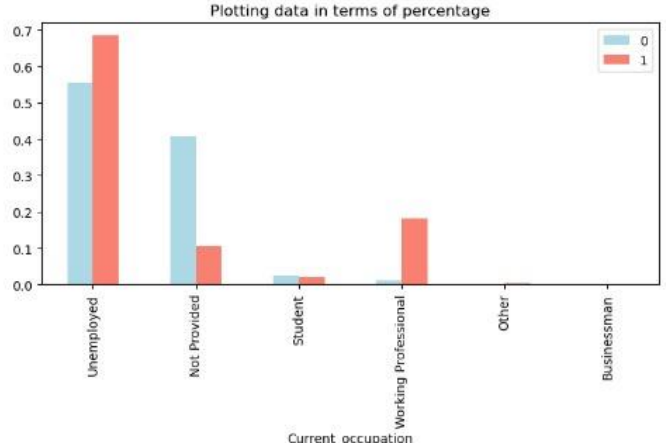
•Last Activity: 68% of customer interactions are from SMS Sent & Email Opened activities.



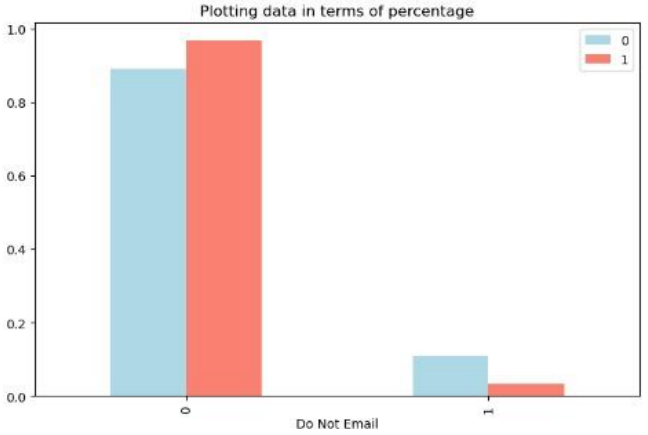
EXPLORATORY DATA ANALYSIS (EDA) UNIVARIATE ANALYSIS

EDA

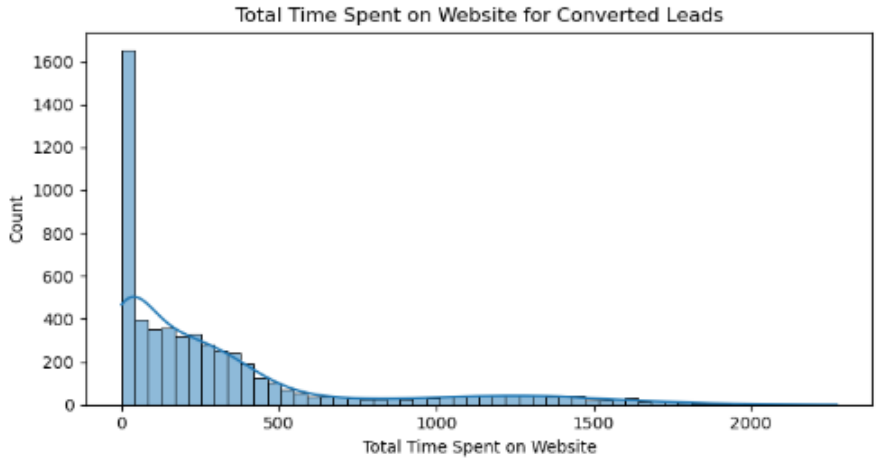
• Current Occupation: 90% of customers are Unemployed.



• Do Not Email: 92% of people opted not to receive emails about the course.



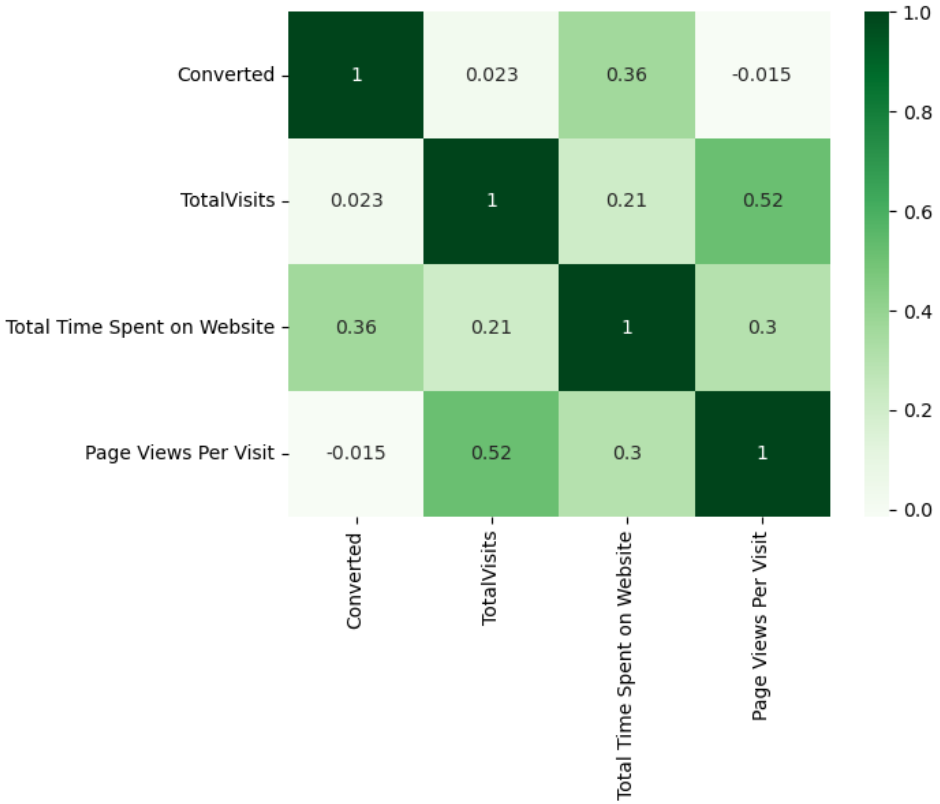
• Leads who spend more time on the website are more likely to convert into successful leads. This indicates that higher engagement correlates with a higher conversion rate.



EXPLORATORY DATA ANALYSIS (EDA) UNIVARIATE ANALYSIS

EXPLORATORY DATA ANALYSIS (EDA) BIVARIATE ANALYSIS

HEATMAP TO SHOW CORRELATION BETWEEN NUMERICAL VALUES



- No multicollinearity found in numerical variables and not much inference can be made with this data.

Model building and evaluation

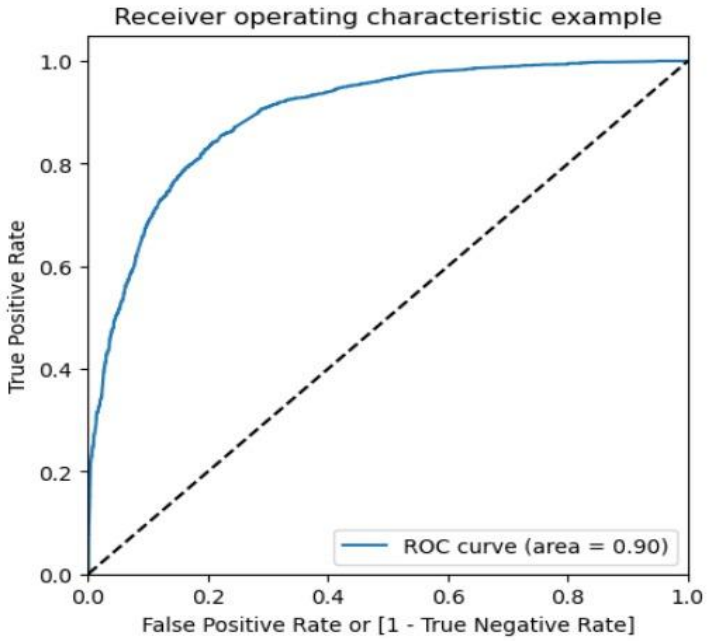
LOGISTIC MODEL BUILDING

Splitting the Data into Training and Testing Sets

- The first basic step for regression is train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature selection.
- Running RFE with 15 variables as output.
- Building Model by removing the variable whose p-value is greater than 0.05 and V_i value is greater than 5.

Predictions on test data set

- Overall accuracy 80%

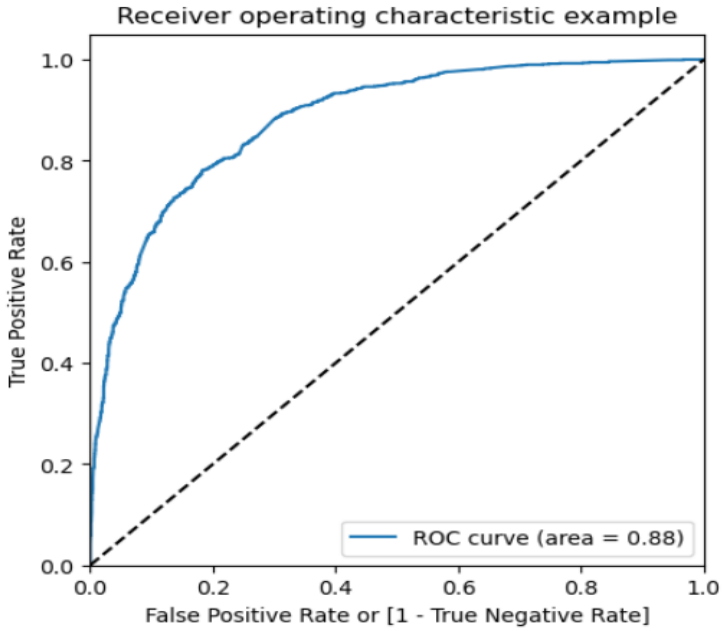


ROC curve for train data

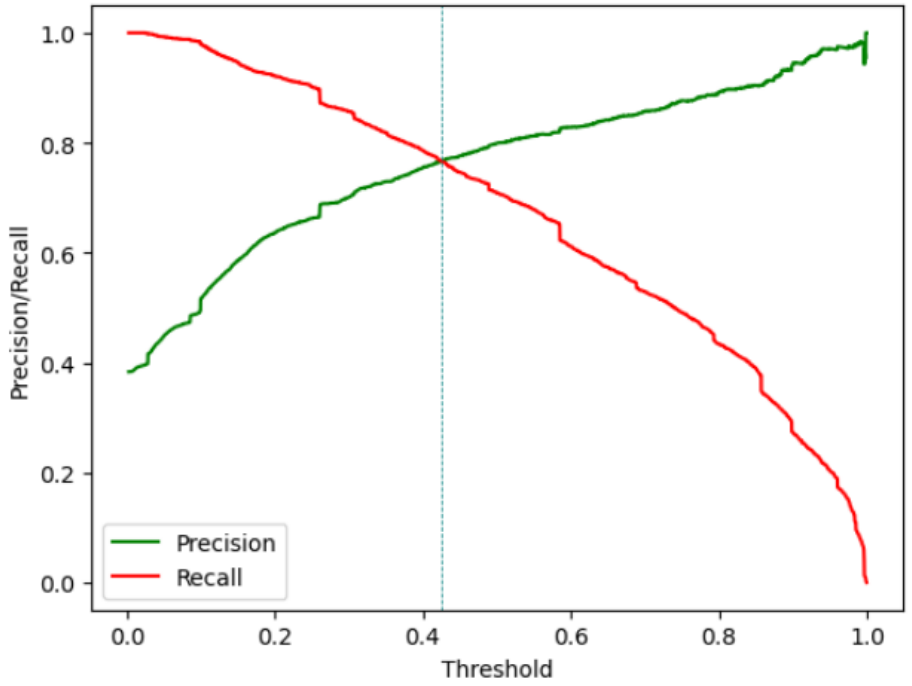
• For ROC in train data, Area under ROC curve is 0.90 out of 1 which indicates a good predictive model

ROC curve for test data

• For ROC in test data, Area under ROC curve is 0.88 out of 1 which indicates a good predictive model



<u>There is not much difference between ROC curves from test and train data. It seems like a good model.</u>

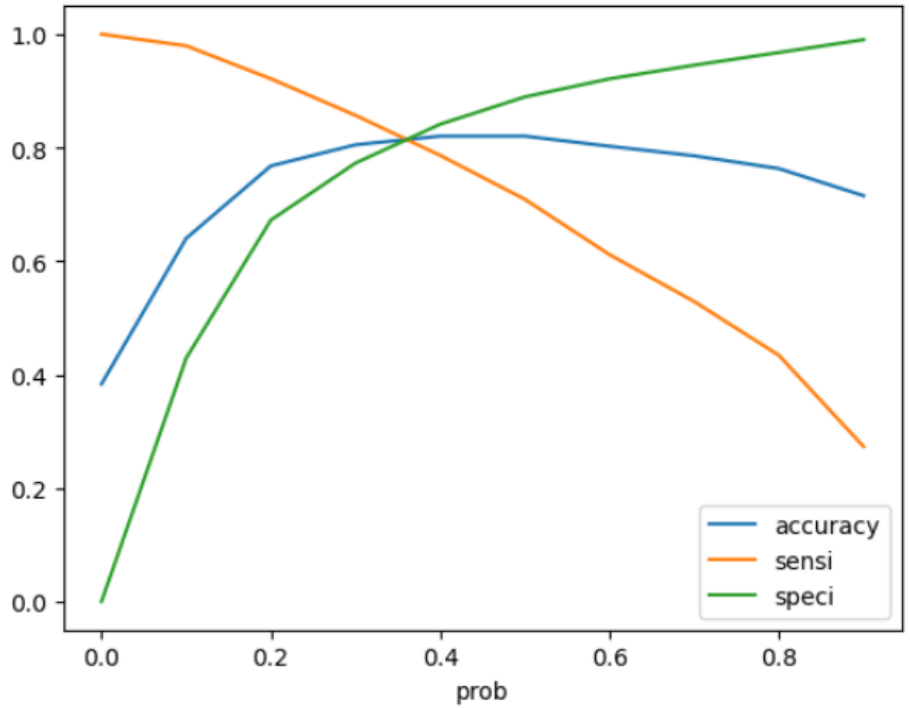


Precision and Recall

•The intersection point of the curve represents the threshold where the model balances precision and recall. This value helps optimize model performance based on business requirements. From the curve above, the optimal probability threshold is approximately 0.42.

Accuracy, Sensitivity and specificity

•0.36 is the approx. point where all the curves meet, so 0.36 seems to be our Optimal cutoff point for probability threshold .



Thank you