

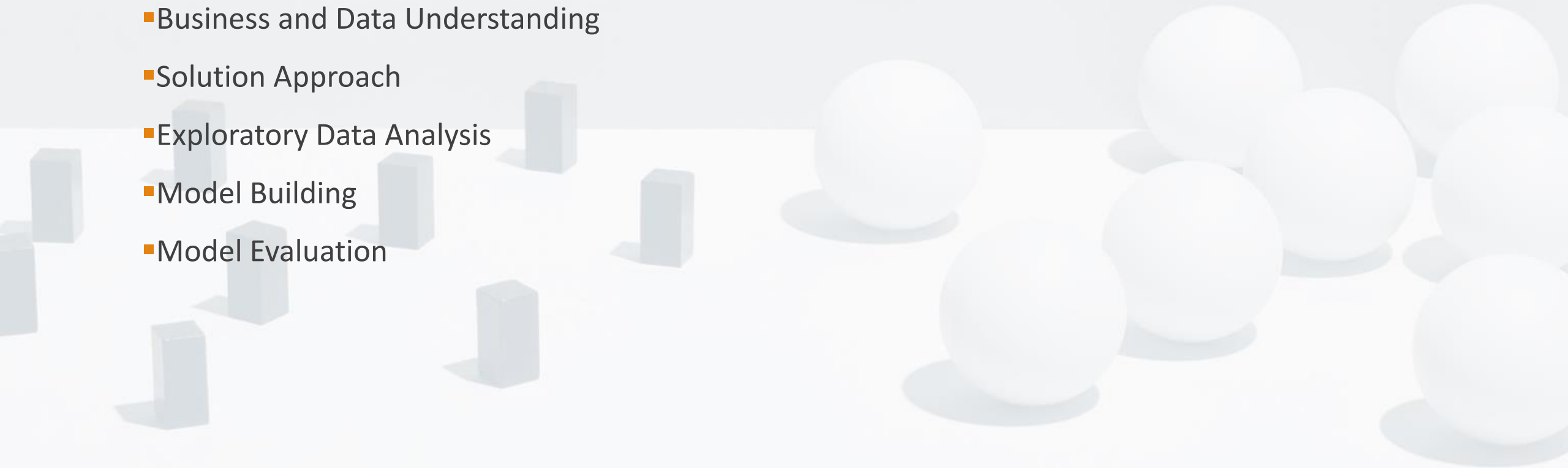
Lead Scoring Case Study C54

KRISHAN KUMAR

UMA DEVI K

SHREYA KANAKIA

Lead Scoring Case Study

- Problem Statement
 - Business and Data Understanding
 - Solution Approach
 - Exploratory Data Analysis
 - Model Building
 - Model Evaluation
- 

Problem Statement

An education company provides an online course to industry professional. The company markets their program through several websites and search engines. Based on the number visits, referrals from past students leads are acquired.

Problem here is lead conversion rate of the company is very low the is 30%.

Company wants to determine the factors that can help them to get more conversion rate.

Business and Data Understanding

The project started with understanding the dataset provided by X Education. It contained information on various attributes such as lead source, website activity, and last interaction.

We observed that several categorical variables had a 'Select' level, which was treated as a null value. Company's profit depends on number of enrollment acquired for different courses.

Company wants to target the individuals such that maximum conversion rate is achieved and hence more profit.

Prospect data is collected by number of users visiting the site through search engine or direct hit. Based on form filled by the prospect on the website data is collected.

Also, company monitors the various activities performed by the prospect on the website like time spent , pages viewed etc.

Business and Data Understanding

Dataset Size: The dataset contains 9,240 data points and 37 columns, with a mix of numerical and categorical features.

Target Variable: The target variable is '**Converted**,' which indicates whether a lead was converted (1) or not (0).

Missing Values: Several columns have missing values, with varying degrees of completeness. For example, 'Country,' 'Lead Quality,' and 'Tags' have a substantial number of missing values.

Data Types: Most of the columns are of type 'object,' indicating categorical variables, while a few columns are of type 'int64' and 'float64,' representing numerical variables.

Numeric Features: Numeric features like 'TotalVisits,' 'Total Time Spent on Website,' 'Page Views Per Visit,' 'Asymmetrique Activity Score,' and 'Asymmetrique Profile Score' provide insights into lead behavior.

Solution Approach

The goal here is to use logistic regression and assign a lead score between 0 to 100 to each prospect.

Below are the steps that can be followed to achieve it:

- High level Observation of dataset
- Data Cleaning
- Categorical Attribute Analysis
- Drop unnecessary rows and columns based on null value analysis
- Outlier analysis and approach to remove the outliers
- Splitting dataset and Building Logistic Regression model
- Evaluate the model on the test data

Exploratory Data Analysis

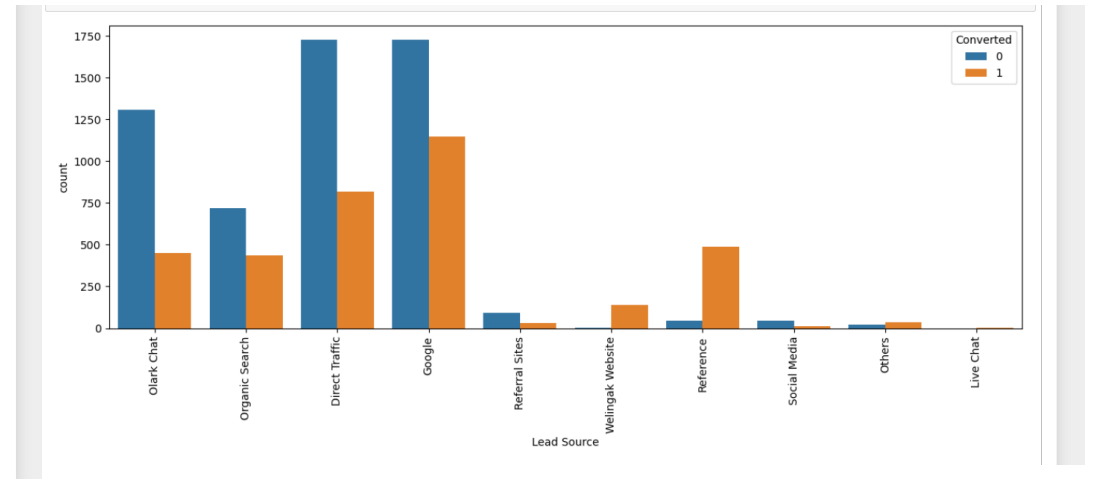
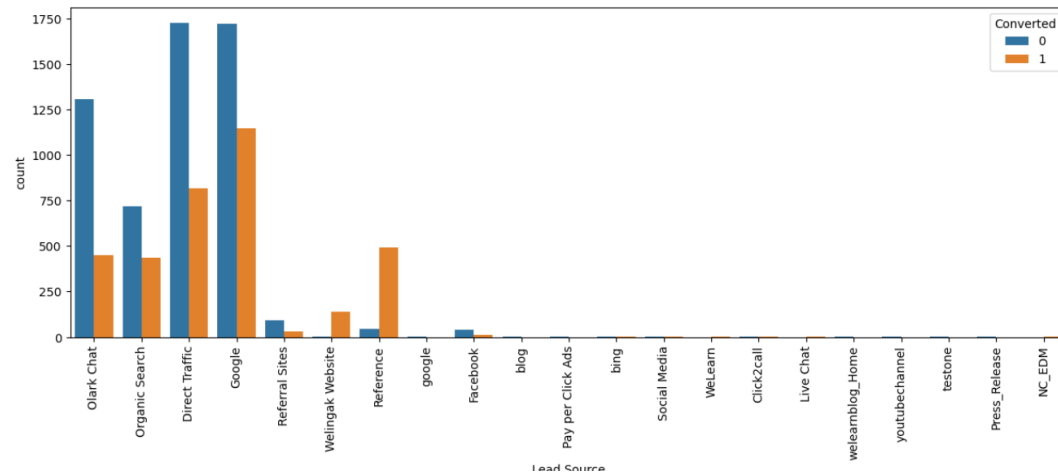
- Checked the number of columns and rows in the dataset.
- Understand the terms used in the column description.
- Checked the datatypes of each variables.
- Analyzed missing values in the datasets and removed the columns having more than 45% values missing. Removed unnecessary columns which does not have significance for model building like prospect ID, Lead number, Tags etc.
- Analyzed outliers in the datasets and remove the outliers with .
- Key findings included the significance of the total time spent on the website, the number of total visits, and the impact of certain lead sources on conversion rates.

Exploratory Data Analysis

- We examined the 'Country' column, found that 'India' was overwhelmingly dominant (97%), and decided to drop this column due to its lack of variability.
- For the 'Specialization' column, we replaced missing values with 'Not Specified' to account for leads who did not provide this information, retaining the column for analysis.
- Remove columns like country where most of the columns have same value as "India". Replaced "Select" value with NaN.
- Drop the variables which has most of the values as "No" as it does not have any impact on the model.
- These actions were taken to ensure the dataset's quality and relevance for further analysis and model development. These steps helped clean and prepare the dataset for further analysis and modeling by addressing missing values and eliminating less informative columns.

Exploratory Data Analysis

Bucket variables having low frequency values into “Others” for variables like Lead Source, Last Activity etc.



Model Building

We have used Stats Model and RFE to build a model. Number of feature selected for RFE is 20.

After removing the values with high P value and VIF on the 9th iteration we produced the final model

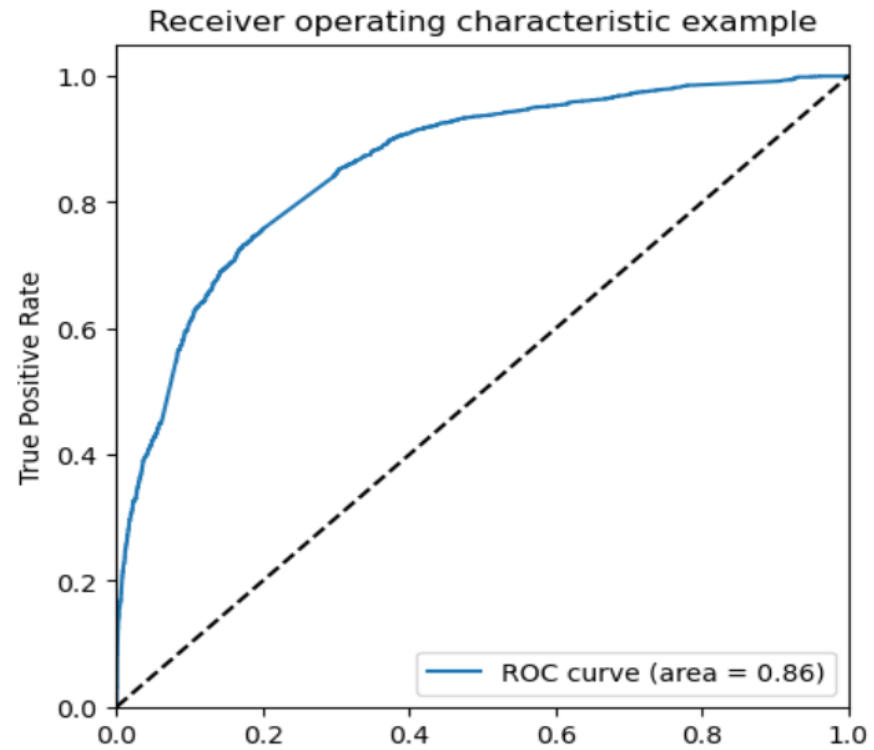
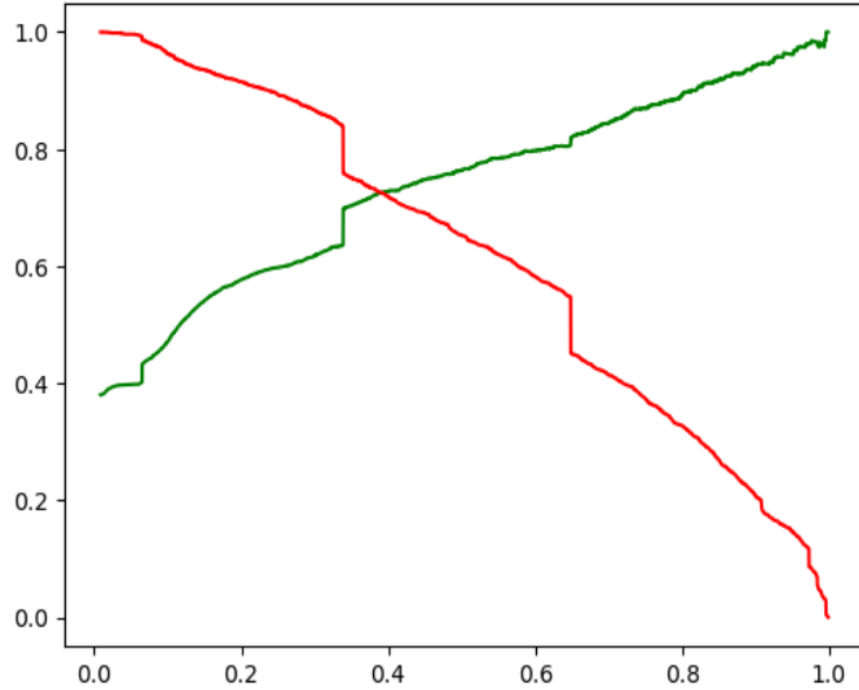
Below are the matrices for the final model :

Accuracy : 78.37%

Sensitivity : 75.76%

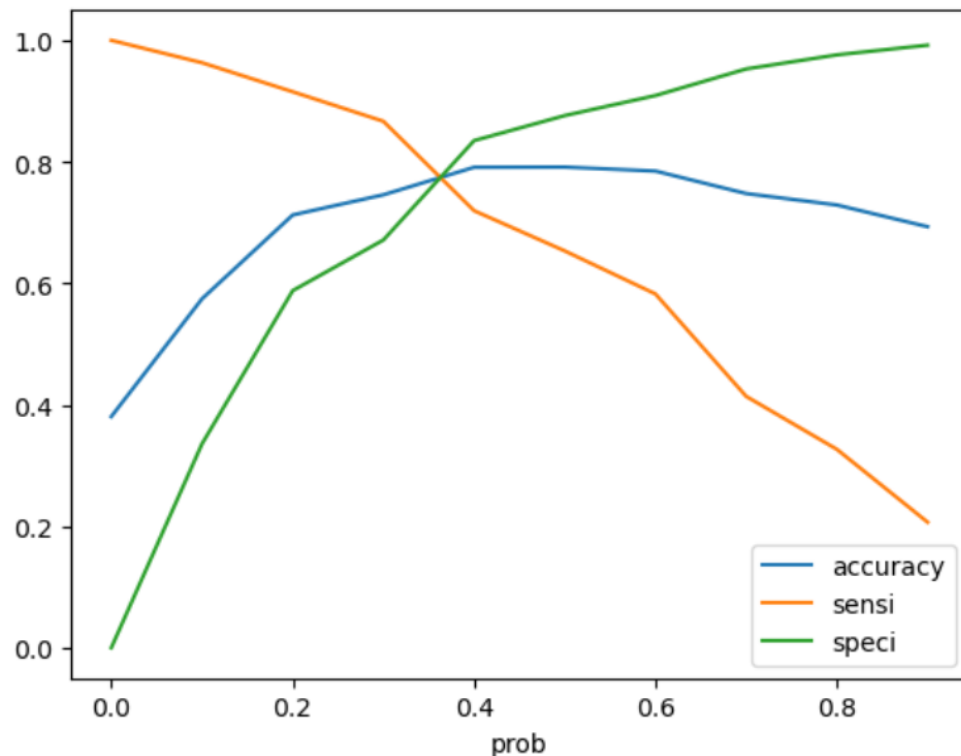
Specificity : 79.98%

Precision Recall Curve



Model Building

Accuracy Sensitivity and Specificity plot for various probabilities



We see 0.34 is the optimum point to take it as a cutoff probability.

Model Building

Below are the features and their coefficient :

	coef	std err	z	P> z	[0.025	0.975]
const	0.2842	0.076	3.734	0.000	0.135	0.433
Total Time Spent on Website	1.0939	0.039	27.876	0.000	1.017	1.171
What is your current occupation_Working Professional	2.9515	0.190	15.511	0.000	2.579	3.324
Specialization_Banking, Investment And Insurance	0.5998	0.172	3.488	0.000	0.263	0.937
Lead Source_Direct Traffic	-2.0679	0.106	-19.546	0.000	-2.275	-1.861
Lead Source_Google	-1.6301	0.100	-16.328	0.000	-1.826	-1.434
Lead Source_Organic Search	-1.7344	0.124	-13.957	0.000	-1.978	-1.491
Lead Source_Referral Sites	-1.9676	0.333	-5.908	0.000	-2.620	-1.315
Lead Source_Welingak Website	4.7787	1.017	4.698	0.000	2.785	6.772
Last Activity_Email Bounced	-1.7817	0.315	-5.653	0.000	-2.399	-1.164
Last Activity_Olark Chat Conversation	-1.9946	0.166	-12.045	0.000	-2.319	-1.670
Last Activity_SMS Sent	1.2811	0.071	18.024	0.000	1.142	1.420

Based on coefficient company should target the prospect whose Source is from the website, who spent more time on website and whose occupation is working professional.

Model Evaluation

Train data observations vs Test data observations

	Accuracy	Specificity	Sensitivity
Train Data	78.37%	79.98%	75.76%
Test Data	78.44%	76.23%	79.78%

Accuracy , specificity and Sensitivity numbers are closed for test and train data indicates that model is good.