


Lead Scoring Case Study (Course 2: Machine Learning)

Submission by:

- Rishav Raj
- Rajiv Kasera
- Kartik Narayana Ryali

Introduction

X Education provides online courses to industry professionals. The company generates the leads through various avenues like website, forms or videos. Basis the leads, the company would then reach out to them for converting into sales.


- **Problem Statement:** X Education's CEO wants to improve the sales conversion from current 30% to 80%. As a data expert, we are required to evaluate the leads and highlight hot leads (who have high chances of conversion) by assigning a lead score between 0 and 100 - high score indicating hot lead
- 

Goals

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads
- Ensure that the model can be adjusted to company's future requirements



Approach


- **Data Understanding:** Getting to know about the type of data available, columns etc
 - **Data Cleaning & Preparation:** Handling null values, identifying & handling redundant columns/values, Handling categorical variables by creating dummy columns and observing correlations between each columns
 - **Model Training Preparation:** Splitting of data into training & test data, scaling of columns, correlation analysis
 - **Model Building:** Feature selection using RFE, fitting logistic regression over iterations of adding/dropping features by evaluating p-value, VIF & r-square
 - **Model Evaluation:** Predicting values on trained model, comparing them with actual values by confusion matrix and evaluating the accuracy
 - **Making Predictions on Test Set:** Predict the data on Test Set and checking for accuracy
- 

Data Understanding

- The Dataset contains 37 columns/parameters and 9240 leads
- There are 7 numeric parameters & 30 non-numeric parameters
- The target variable is Converted (0 indicating not converted, 1 indicating converted)
- Summary of numeric parameters is as follows:

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Data Cleaning & Preparation

- Columns which contain more than 30% (3000) missing values, are dropped.
 - Columns, where the majority values are same (such as City and Country columns which have Mumbai & India as majority values) are dropped.
 - Post above two steps, the null rows are dropped for columns which have highest null value counts
 - After the cleanup of null values 69% of rows are left, which is a decent amount of information retained
 - High Cardinal columns (Prospect ID & Lead Number) are removed as it wouldn't help in analysis
 - Dummy variables are created for categorical columns/variables. Binary variables encoded with 1 for 'Yes' and 0 for 'No.'
- 

Model Preparation & Building

- Post split of data into test & training data (30-70% split), scaling of numeric values are performed for training data
- To select features for model building, we have followed automated approach of Recursive Feature Elimination(RFE), as there are lot of features
- Basis RFE, following features have been selected as a starting point:
 - *'TotalVisits', 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Reference', 'Lead Source_Welingak Website', 'Do Not Email_Yes', 'Last Activity_Had a Phone Conversation', 'Last Activity_SMS Sent', 'What is your current occupation_Housewife', 'What is your current occupation_Student', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Last Notable Activity_Had a Phone Conversation', 'Last Notable Activity_Unreachable'*


Model Preparation & Building

- In iterations, features which have high p-value (more than 0.05) & high VIF (more than 5) are dropped
- Post the eliminations, the final feature selection & model is as displayed at right:

Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4449
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2079.1
Date:	Tue, 21 Jan 2025	Deviance:	4158.1
Time:	11:36:13	Pearson chi2:	4.80e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3642
Covariance Type:	nonrobust		

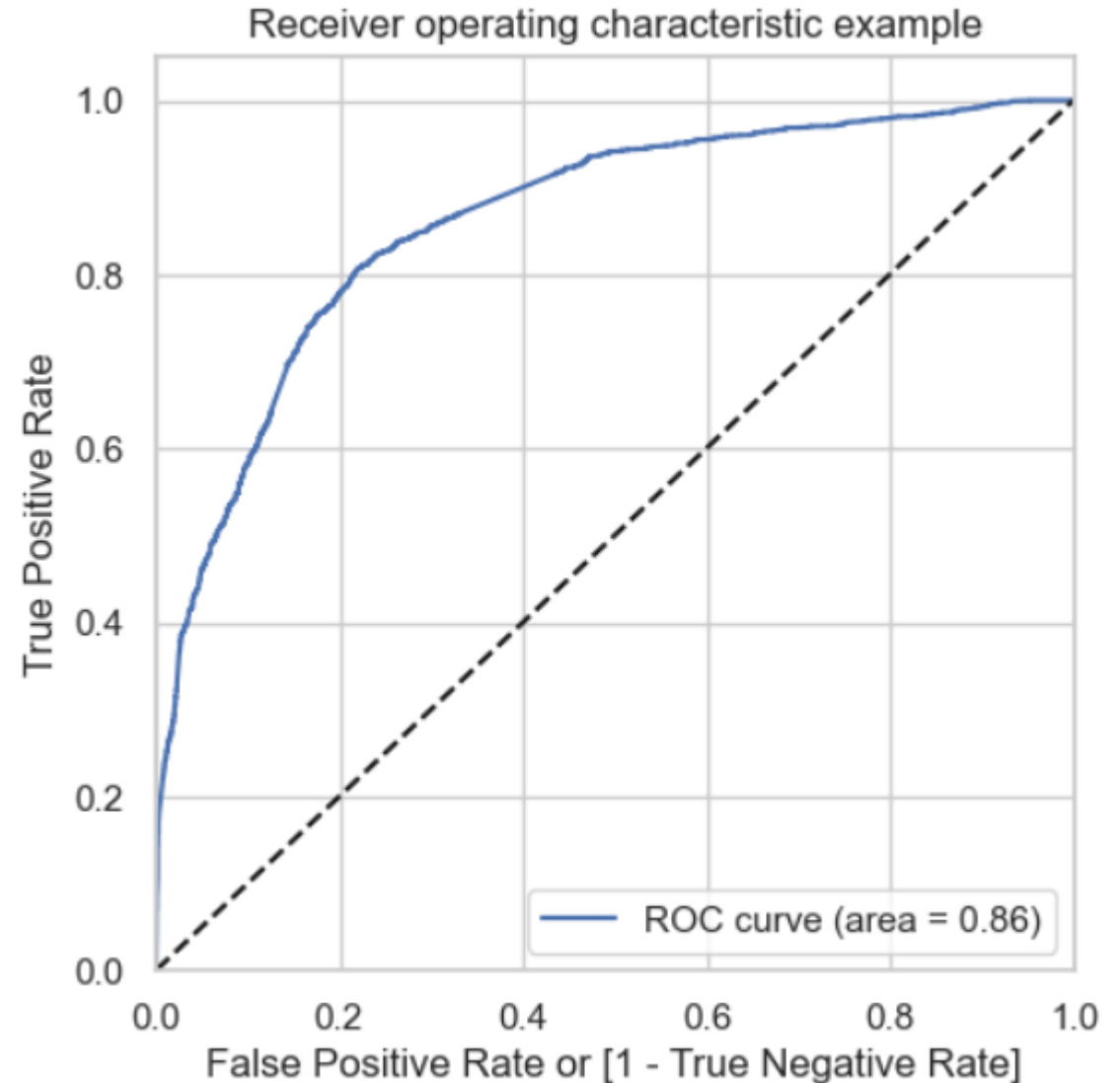
	coef	std err	z	P> z	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

Model Evaluation

- Based on the model & trained data, the converted probability is predicted and compared alongside the actual converted values.
 - 0.5 probability has been taken as a cutoff to converting probability to predicted values
 - Basis the predicted & actual value, confusion matrix is created.
 - Accuracy = 0.79
 - Sensitivity = 0.74
 - Specificity = 0.83
- 

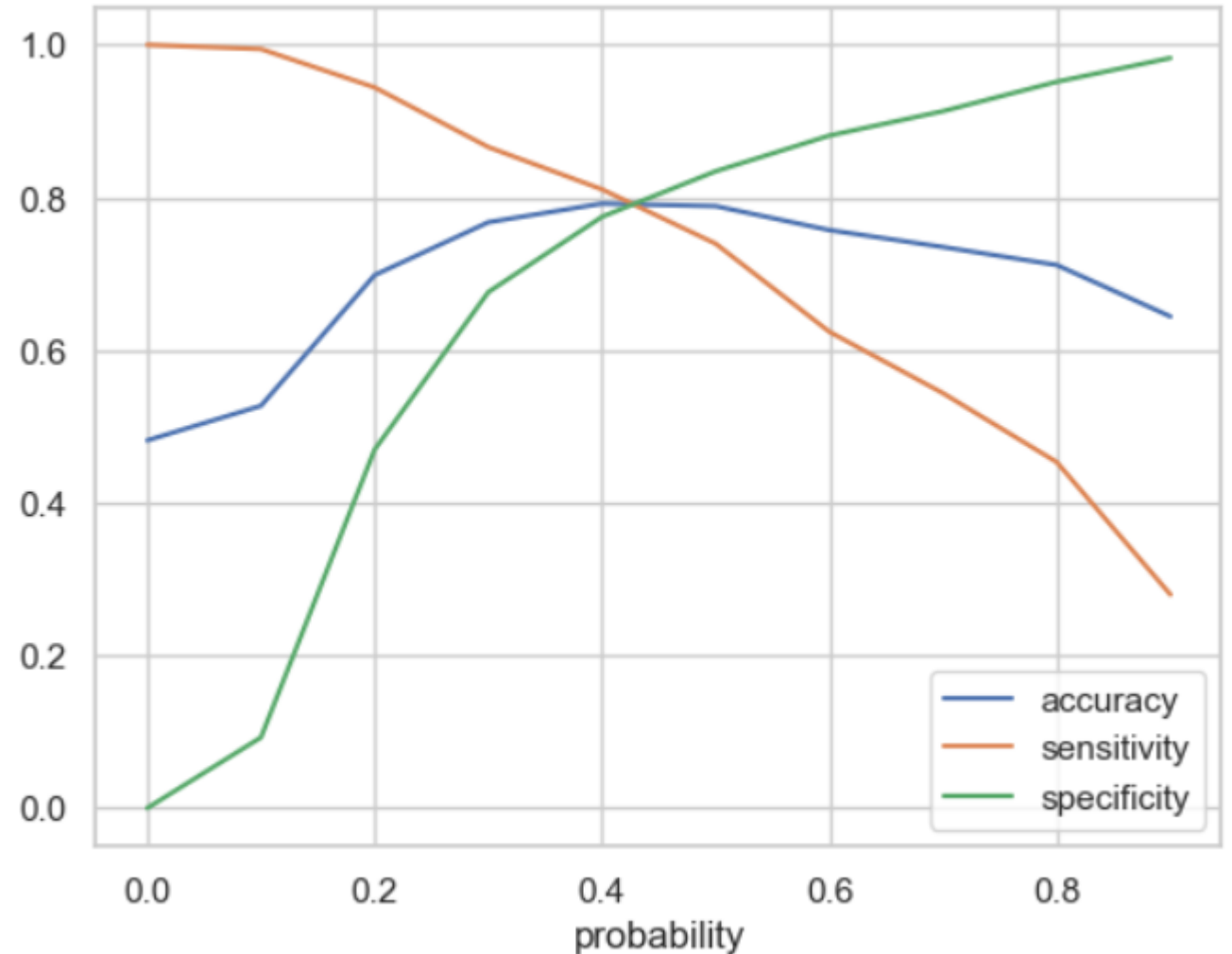
Model Evaluation

- For 0.5 probability cutoff, the Area Under the Curve of the ROC curve is 0.86 which shows that the model can accurately distinguish between positive and negative instances.



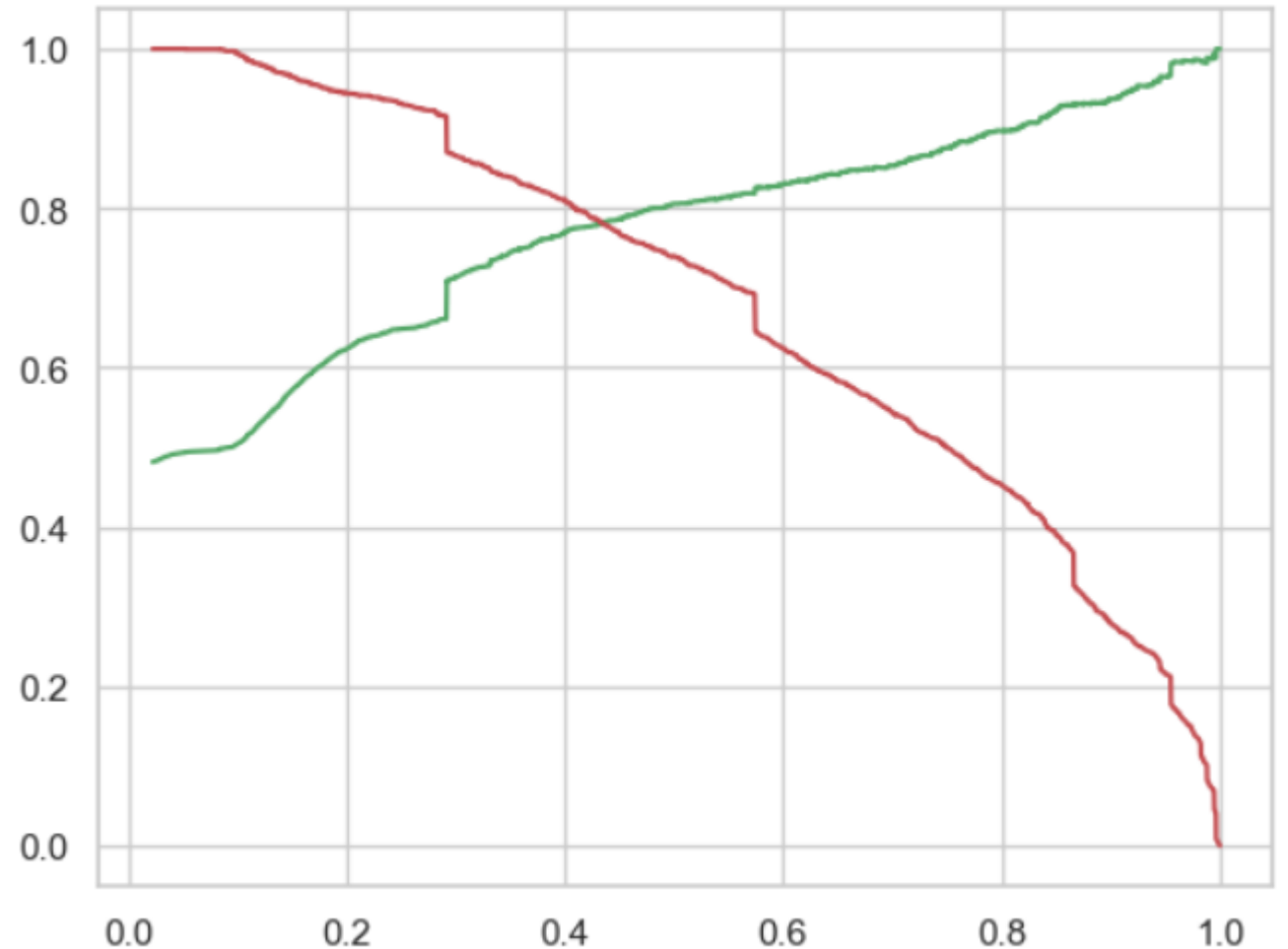
Model Evaluation

- Upon plotting accuracy, sensitivity & specificity, it is observed that 0.42 is coming to be an optimal cutoff



Making predictions on test set

- 0.44 is coming to be an optimal cutoff by plotting a tradeoff graph between precision & recall
- Basis 0.44 cutoff model, the precision is 0.78 and recall is 0.77



Summary

- 11 Features have been picked up for preparing model for predicting lead conversion
 - *TotalVisits*
 - *Total Time Spent on Website*
 - *Lead Origin_Lead Add Form*
 - *Lead Source_Olark Chat*
 - *Lead Source_Welingak Website*
 - *Do Not Email_ Yes*
 - *Last Activity_Had a Phone Conversation*
 - *Last Activity_SMS Sent*
 - *What is your current occupation_Student*
 - *What is your current occupation_Unemployed*
 - *Last Notable Activity_Unreachable*
- Total Visits is a major indicator to predict lead conversions. If a person has high number of visits to the website, probability to convert is high
- Lead being is student or unemployed and probability to conversion is inversely proportional
- Lead opting out of emails and probability to conversion is inversely proportional