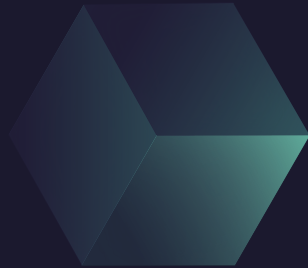


# Lead Scoring Case study

Sowndarya Venkateswaran

# Agenda

- Background
- Problem Statement
- Analysis Approach
- Inferences
- Recommendations
- Summary





# Background

---

On any given day, many professionals who are interested in the courses offered by X - Education land on their website and browse for courses.

---

Marketing is done on several websites and search engines like Google.

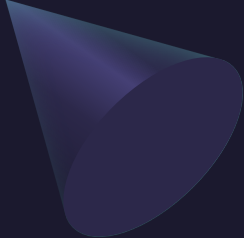
---

People who land on the website may fill up a form and watch related videos.

---

People who fill up the form providing their email address or phone number, are considered as leads. Leads can also be via past referrals.

---



For the acquired leads, sales team start making calls, writing emails ,etc., to convert them as potential customers. Typical Lead conversion rate is 30%

# Problem Statement

---

X- Education has lots of leads but a poor conversion rate.

---

Company wants to increase the lead conversion rate and to target the potential leads .

---

X-Education wishes to identify the most potential leads, also known as 'Hot Leads' who are most likely to convert into paying customers.

Expectation - target lead conversion rate to be around 80%.





# Analysis Approach



## Preparing the data

- Data Cleaning
- Handling categorical and text attributes

## Model Selection and training

- Train-test split
- Feature Scaling
- Evaluating Training set
- Feature Selection using RFE
- Fine-tuning the Model

## Model Evaluation

- Predicting probabilities and Converted
- Confusion Matrix
- Performance Measures
- Sensitivity Vs Specificity
- Precision and Recall
- ROC Curve

## Model Performance

- Prediction on test set

## Lead Score

# Inferences

---

Correlations

---

Model with essential features

---

ROC Curve

---

Optimal cut-off (Sensitivity vs Specificity vs Accuracy)

---

Precision vs Recall trade-off

---

Model Performance

---

Lead Score

# Correlations of features with over 60%

lead_origin_Lead Import	lead_source_Facebook	0.981903
lead_source_Facebook	lead_origin_Lead Import	0.981903
last_activity_SMS Sent	last_notable_activity_SMS Sent	0.890591
last_notable_activity_SMS Sent	last_activity_SMS Sent	0.890591
last_activity_Unsubscribed	last_notable_activity_Unsubscribed	0.879716
last_notable_activity_Unsubscribed	last_activity_Unsubscribed	0.879716
last_activity_Email Opened	last_notable_activity_Email Opened	0.866192
last_notable_activity_Email Opened	last_activity_Email Opened	0.866192
lead_source_Reference	lead_origin_Lead Add Form	0.862980
lead_origin_Lead Add Form	lead_source_Reference	0.862980
last_activity_Email Link Clicked	last_notable_activity_Email Link Clicked	0.781836
last_notable_activity_Email Link Clicked	last_activity_Email Link Clicked	0.781836
last_activity_Had a Phone Conversation	last_notable_activity_Had a Phone Conversation	0.751218
last_notable_activity_Had a Phone Conversation	last_activity_Had a Phone Conversation	0.751218
last_activity_Email Received	last_notable_activity_Email Received	0.707051
last_notable_activity_Email Received	last_activity_Email Received	0.707051
last_activity_Page Visited on Website	last_notable_activity_Page Visited on Website	0.693902
last_notable_activity_Page Visited on Website	last_activity_Page Visited on Website	0.693902

# Model with essential features

- Model has 11 important features highlighted along with their corresponding coefficients (Higher values indicate more impact), p-value and VIFs.

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:	Wed, 08 Dec 2021	Deviance:	4158.1
Time:	01:46:20	Pearson chi2:	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

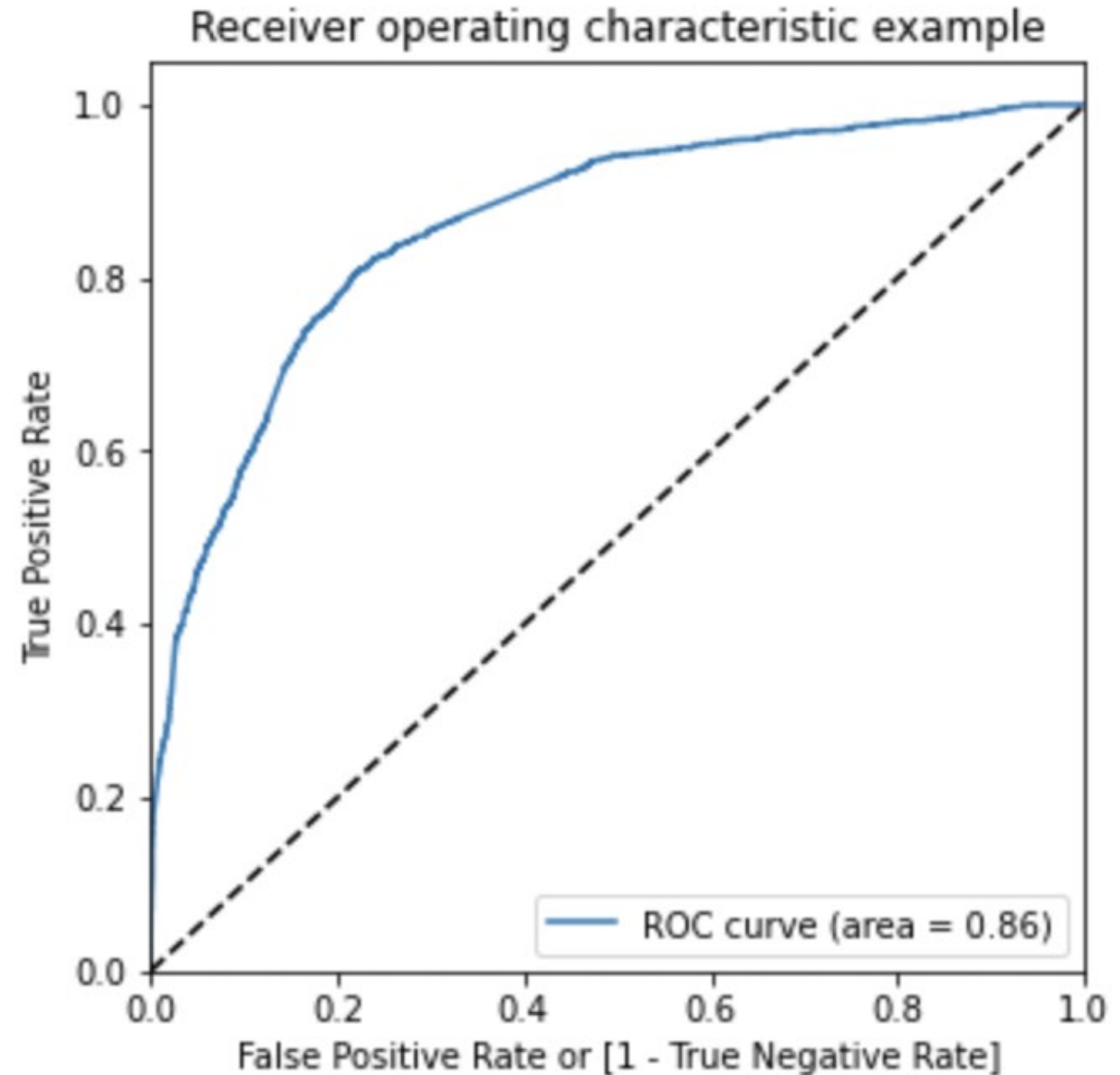
	coef	std err	z	P> z	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
do_not_email	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
totalvisits	11.1489	2.665	4.184	0.000	5.926	16.371
total_time_spent_on_website	4.4223	0.185	23.899	0.000	4.060	4.785
lead_origin_Lead Add Form	4.2051	0.258	16.275	0.000	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
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

	Features	VIF
9	what_is_your_current_occupation_Unemployed	2.82
2	total_time_spent_on_website	2.00
1	totalvisits	1.54
7	last_activity_SMS Sent	1.51
3	lead_origin_Lead Add Form	1.45
4	lead_source_Olark Chat	1.33
5	lead_source_Welingak Website	1.30
0	do_not_email	1.08
8	what_is_your_current_occupation_Student	1.06
6	last_activity_Had a Phone Conversation	1.01
10	last_notable_activity_Unreachable	1.01



# ROC Curve

- Area Under the Curve as 0.86

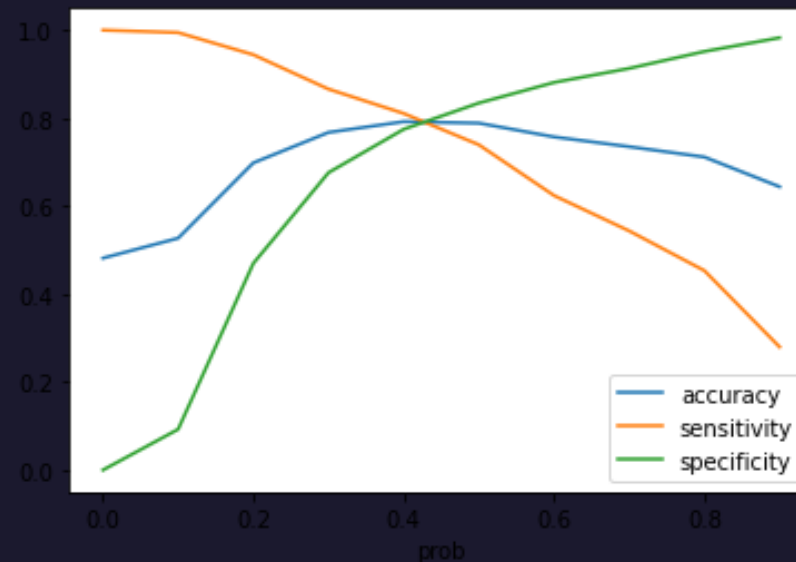


# Optimal cut-off (Sensitivity vs Specificity vs Accuracy)

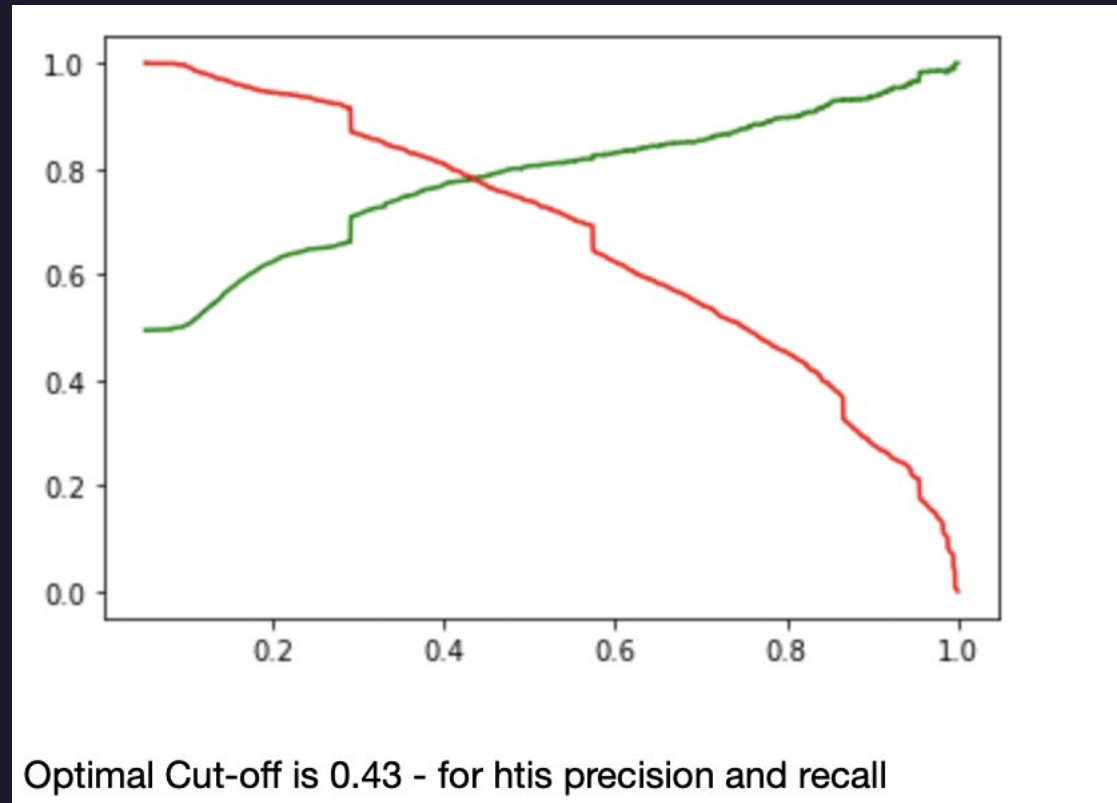
The optimal cut-off point is 0.43 based on the curve's intersection.

The Table indicates the sensitivity / specificity and accuracy for various probabilities.

	prob	accuracy	sensitivity	specificity
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699



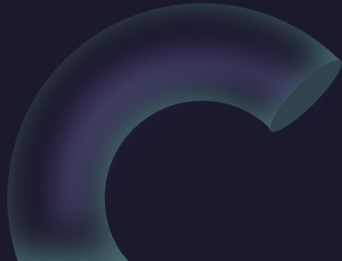

# Precision vs Recall trade-off





# Model Performance

Train/Test	Accuracy	Sensitivity	Specificity
Train	0.7895	0.7854	0.7932
Test	0.7850	0.7751	0.7941



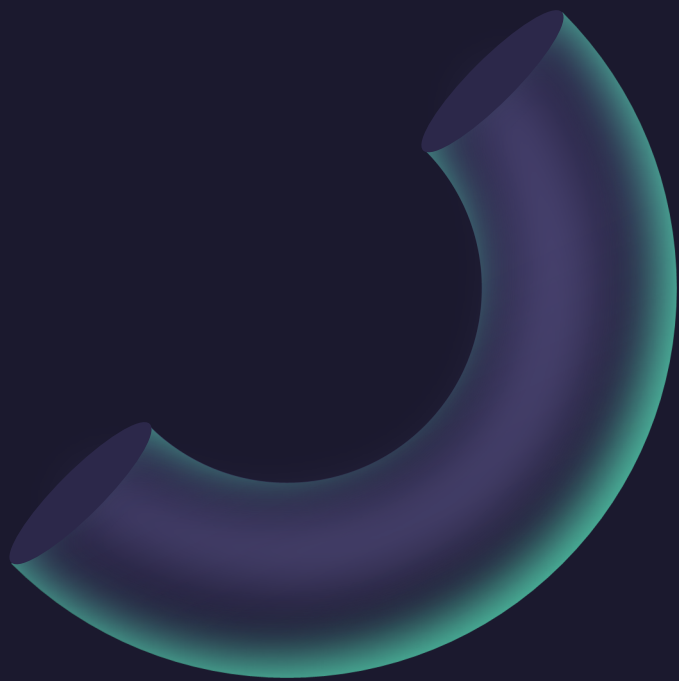
# Lead Score

- This table has the Lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

	converted	Conversion_Prob	Final_Predicted_Conv	Lead Score
0	1	0.996296	1	99.6
1	0	0.129992	0	13.0
2	0	0.703937	1	70.4
3	1	0.299564	0	30.0
4	1	0.720796	1	72.1
5	1	0.792250	1	79.2
6	0	0.704038	1	70.4
7	1	0.464521	1	46.5
8	0	0.282978	0	28.3
9	1	0.786460	1	78.6
10	1	0.987981	1	98.8
11	0	0.351053	0	35.1
12	0	0.189840	0	19.0
13	1	0.472712	1	47.3
14	1	0.876810	1	87.7

# Recommendations

- Below are the attributes help us to determine the potential leads that convert into paying customers-
  - Total number of visits to the website
  - Total time spent on the website
  - Origin identified as Lead Add Form
  - Last Activity was a phone conversation
  - Source of the lead - Welingak Website and Olark Chat
  - Last Activity performed- SMS Sent
  - Customer is a student
  - Current Occupation is Unemployed



# Modeling for different scenarios

#### Scenario 1

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.



	prob	accuracy	sensitivity	specificity
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699

- **As we have more people, one of the best strategies is to target people with good successful conversion rate. We can choose the cut off of prob as 0.4 where we can have sensitivity (% of customers to call) as 0.810610 and Specificity (% of Customers that make a successful lead conversion) as 0.774654.**



## Scenario 2

Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e., they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage

	prob	accuracy	sensitivity	specificity
0.0	0.0	0.481731	1.000000	0.000000
0.1	0.1	0.527012	0.994416	0.092561
0.2	0.2	0.698274	0.944160	0.469723
0.3	0.3	0.767541	0.865984	0.676038
0.4	0.4	0.791975	0.810610	0.774654
0.5	0.5	0.788612	0.739414	0.834343
0.6	0.6	0.757229	0.624011	0.881055
0.7	0.7	0.735037	0.543509	0.913062
0.8	0.8	0.711500	0.453234	0.951557
0.9	0.9	0.644026	0.279665	0.982699

- In order to focus on making only essential phone calls, X education should concentrate on people with high conversion rate. Based on the above table, we can choose the cut off prob as 0.8 where we can have sensitivity (% of customers to call) as 0.45 and Specificity (% of Customers that make a successful lead conversion) as 0.95.