



# LEAD SCORING CASE STUDY

---

# PROBLEM STATEMENT

---

X Education is company that sells online courses to the industry professionals. It want to build a machine learning model that can assign score between 0 to 100 to potential leads for better targeting.

A score of 100 would mean probability of lead getting converted is very high



# DATA WRANGLING AND EDA

---

- We have built a logistic regression model for better interpretability.
- The data is taken for basic data cleaning, imputation and outlier treatment and dummy encoding.
- After the data wrangling, EDA is done on the dataset.
- With the results of the EDA, we can have a little bit of understanding of the data and the influence of the features.



# MODEL BUILDING AND EVALUATION

---

- After EDA model building is done, the logistical regression model is built.
- Using RFE, 16 features are first selected and after that the model is evaluated on VIF and P-Value. High P-Value and High VIF columns are dropped one by one until final model is arrived.
- The final model is then evaluated on Precision, Recall, ROC curve and the optimal threshold is found.



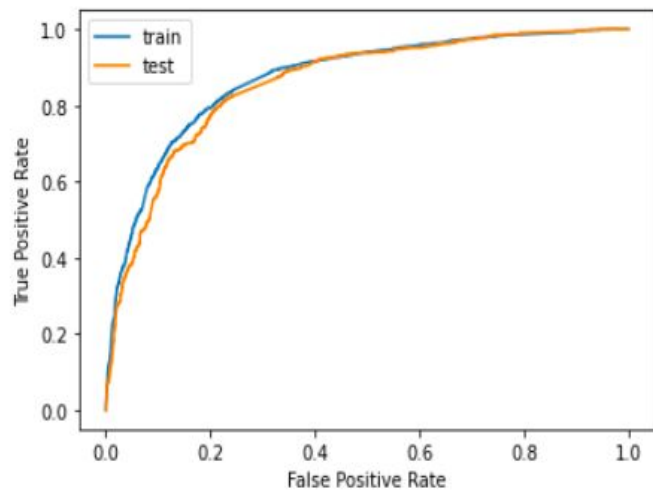
# MODEL RESULTS

---

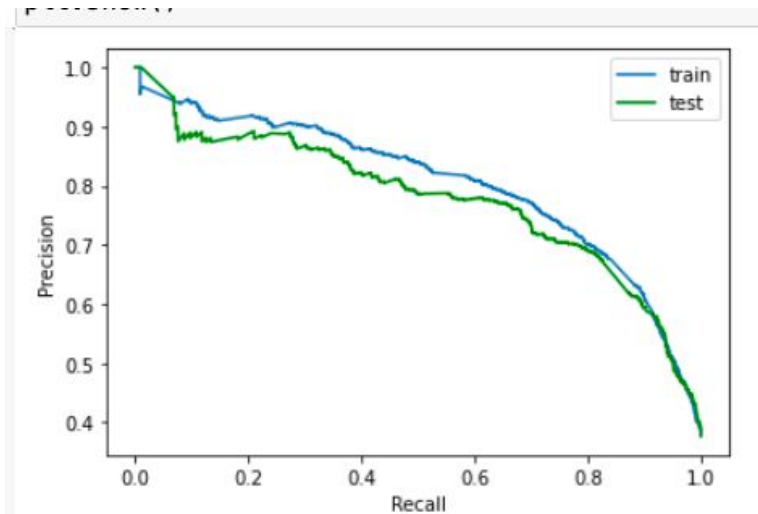
train metrics:		precision		recall	f1-score
	0	0.82	0.88	0.85	4118
	1	0.78	0.68	0.73	2487
accuracy				0.81	6605
macro avg		0.80	0.78	0.79	6605
weighted avg		0.80	0.81	0.80	6605
test metrics:		precision		recall	f1-score
	0	0.81	0.87	0.84	1366
	1	0.76	0.67	0.71	833
accuracy				0.79	2199
macro avg		0.79	0.77	0.78	2199
weighted avg		0.79	0.79	0.79	2199

# MODEL RESULTS

---



ROC CURVE



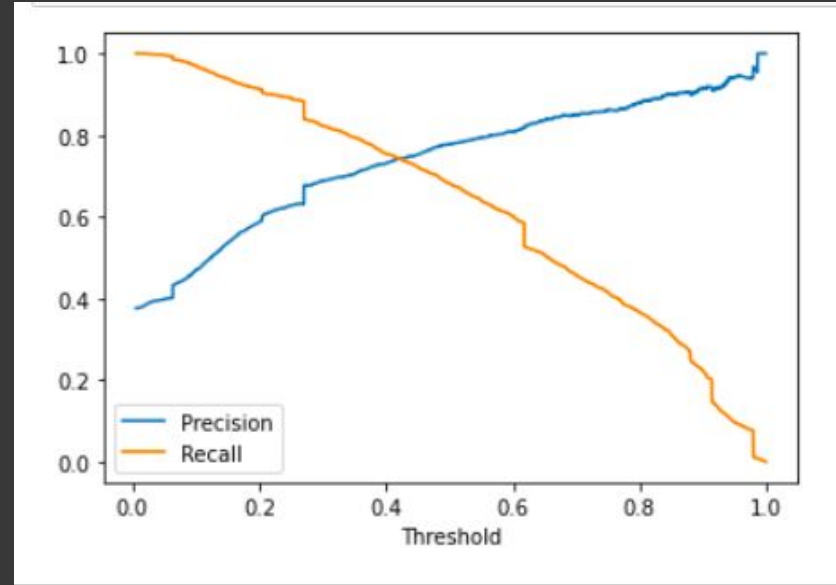
PR CURVE



# MODEL RESULTS

---

Optimal Threshold value for the model would be 0.40.




	features	coef
11	Lead_Origin_Lead Add Form	4.722356
2	Last_Notable_Activity_SMS Sent	1.841438
0	Do_Not_Email_Yes	-1.578327
4	Last_Notable_Activity_Others	1.439108
10	Lead_Source_Olark Chat	1.371932
7	Last_Activity_Olark Chat Conversation	-1.329764
5	Total_Time_Spent_on_Website	1.118691
1	Lead_Origin_Lead Import	0.996951
9	Last_Activity_Email Opened	0.367218
6	TotalVisits	0.307385
8	Last_Activity_Page Visited on Website	-0.274283
3	Lead_Source_Google	0.273872
12	Page_Views_Per_Visit	-0.223143

# TOP FEATURES

## TOP COLUMNS IN THE DATASET ARE:

- Lead Origin
- Last Notable Activity
- Do Not Email
- Lead Source
- Total Time spent on website
- Last Activity
- Total Visits
- Pages viewed per visit.





TOP THREE VARIABLES IN MODEL  
WHICH CONTRIBUTE MOST TOWARDS  
THE PROBABILITY OF A LEAD GETTING  
CONVERTED ARE:

---

- Lead Origin
- Last Notable Activity
- Lead Source

TOP THREE CATEGORICAL/DUMMY  
VARIABLES IN THE MODEL WHICH  
SHOULD BE FOCUSED THE MOST ON  
IN ORDER TO INCREASE PROBABILITY  
OF LEAD CONVERSION ARE:

---

- Lead Origin - Lead Add Form
- Last Notable Activity - SMS Sent
- Last Notable Activity - Olark Chat



TO MAXIMISE CONVERSION RATE, WE  
MUST PRIORITISE LEADS BASED ON

- 
- Who had sent a sms.
  - Who had allowed us to send them an email.
  - Who have originate from Olark Chat.
  - Who come from lead add form.
  - Whose last activity was not Olark Chat Conversation

TO MINIMISE THE RATE OF USELESS  
PHONE CALLS, WE DO NOT CALL  
PEOPLE WHO,

- 
- Who had not allowed us to sent emails to them.
  - Whose last activity is Olark chat conversation with us.
  - Whose last activity is page visited on website.



THANK YOU

---