X Education
Lead Prediction and Rating
(Machine Learning CASE STUDY)

By:-

Kamal Kumar

Table of Contents

Business Understanding

- X Education is an online courses company that sells online courses to industry professionals.
- The primary business of the organization depends on the number of leads being generated.
- A generated leads has chances of being converted into a customer by buying a course or being non-converted.
- These Leads can be arranged on their probability of getting converted. A Lead with higher chances of being converted are termed as Hot Leads.

”The only profitable leads are the ones which gets converted”

Working of Lead Conversion Process

An ideal lead conversion process followed at any online course selling company is completed in multiple steps:

1. Social Media Marketing and other medium pulls the visitors on the website
2. The Visitor fills the form with their contact and interest details
3. Sales team collects the initial pool of the data
4. Lead Nurturing
5. Lead Conversion

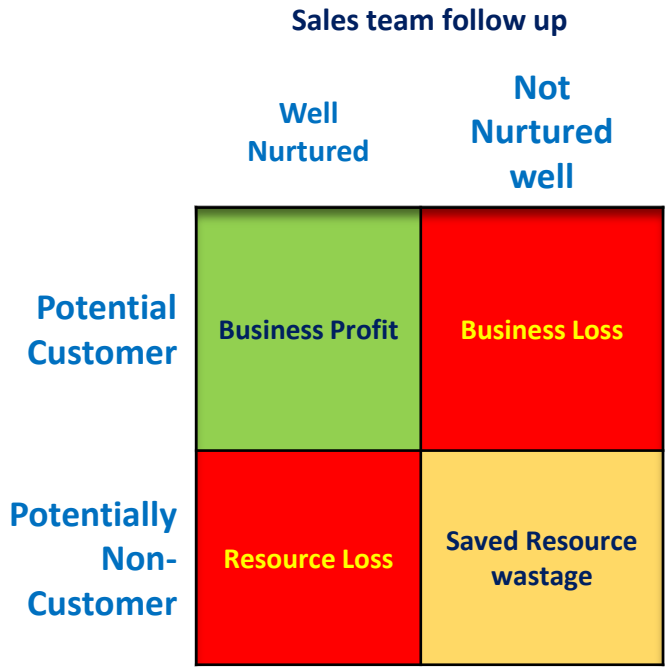
As an Analyst, we are supposed to involve in Lead Nurturing phase with Sales team to identify the high potential leads and term them as Hot Leads. These Leads are of high probability to get converted.

Here, as we progress in the stages, we are left only with concrete leads that have higher chances of getting converted.

Since we are already given with the lead pool, the focus of our case study will revolve around the Lead Nurturing / Hot Leads Identification.

Business Profitability

- The X Education will make profit if a set of probable customers are identified as soon as possible and hence they can be contacted and well Nurtured with all the course benefits.
  - **Customer Loss**
    If a Website Visitor / Lead is likely to join the course and the sales team didn't follow up appropriately, it is a potential Customer loss for the company.
  - **Resource Loss**
    If a Website Visitor / Lead is not likely to join the course and the sales team followed up assuming him/her to be potential customer, it is a potential Resource loss for the company.



It is evident from matrix that there are very less chances of resources being spent on actual conversions and hence can benefit the organization.
So, identification of Hot Leads are very important!

- With this case study, we aim to **develop a model to identify the Hot Leads.**
- We need to identify the driving variables towards a customer opting for the curse.
- Build a model which can be partially triggered to identify when to go for the aggressive marketing as well as time to slow down and hence not wasting the resources.

The company can utilize this model to rank the leads based on their merits, plan their marketing campaign, decide on the new course launch, offers, and multiple other business decisions.

- A Summery of the dataset and their treatments are as follows:
  - ✓ **Lead.csv** contains all the information of the leads. It contains the data about how they were referred to the website, their past interactions with the sales team, and their conversion status. This is a mix of numerical and categorical columns.

- A basic cleaning of the data was carried out:
  - ✓ We identified the variables with single response and dropped them to the fact that they won't influence the model building.
  - ✓ Next, we dropped the leads identifier variables like prospect_ Id and Lead_number.
  - ✓ Further, few variables had a good number of "Select" in their value. This could be because a lead didn't want to reveal and hence, I converted them to Null values.

Overview structure of the datasets	
	Application Data
Number of Rows	**9240**
Number of Columns	**37**

- ✓ Dropped the columns that had more than 30% null values at this stage.
- ✓ Checked outliners using describe with percentile function and went ahead to validate the same with box plot. Dropped the above 99% data as felt like it can deviate the model output. A 97.72% of retention in data is still good enough for our studies.
- ✓ Next, we analyzed the remaining columns and focused on Data Imbalance.
  - ➢ If needed, the missing/null values were imputed based on their merits after checking the data distribution through plotting as well as checking the unique value counts.
  - ➢ After the Null value handling, data imbalance was checked. We dropped any variable which crossed the data imbalance threshold of 70% for a given value.
- ✓ Again, we re-checked the null values and handled them on case-to-case basis.

Now, the dataset is ready to pass through Data Preparation stage for Model Building.

- ✓ **Data Preparation** stage starts with creation of binary values against the string values, and next we created dummy for the categorical variables.
- ✓ Next, we checked the inter-correlation between variables and identified few **correlated variables** with coefficient above 0.7. But we didn't drop them at this stage as they can be taken care in RFE stage of Model Building.
- ✓ Further, the dataset was split into Train-Test dataset, after dropping the target variable **Converted** with **train size of 0.7 and test size of 0.3**.
- ✓ At this stage, we **Rescaled** the Numerical features using **MinMax Scaling**.
- ✓ Again, we re-checked the null values and handled them on case-to-case basis.

Now, the dataset is ready to pass through Data Preparation stage for Model Building.

- ✓ **`Used RFE` feature Selection to select 20 most important features/variables**
  - ➢ **Recursively evaluated the model for p-value and VIF**
  - ➢ **The threshold values for model building are:**

    P ≤ 0.05

    VIF ≤ 5
- ✓ **The final model includes a model where the lead prediction is dependent on following variables:**

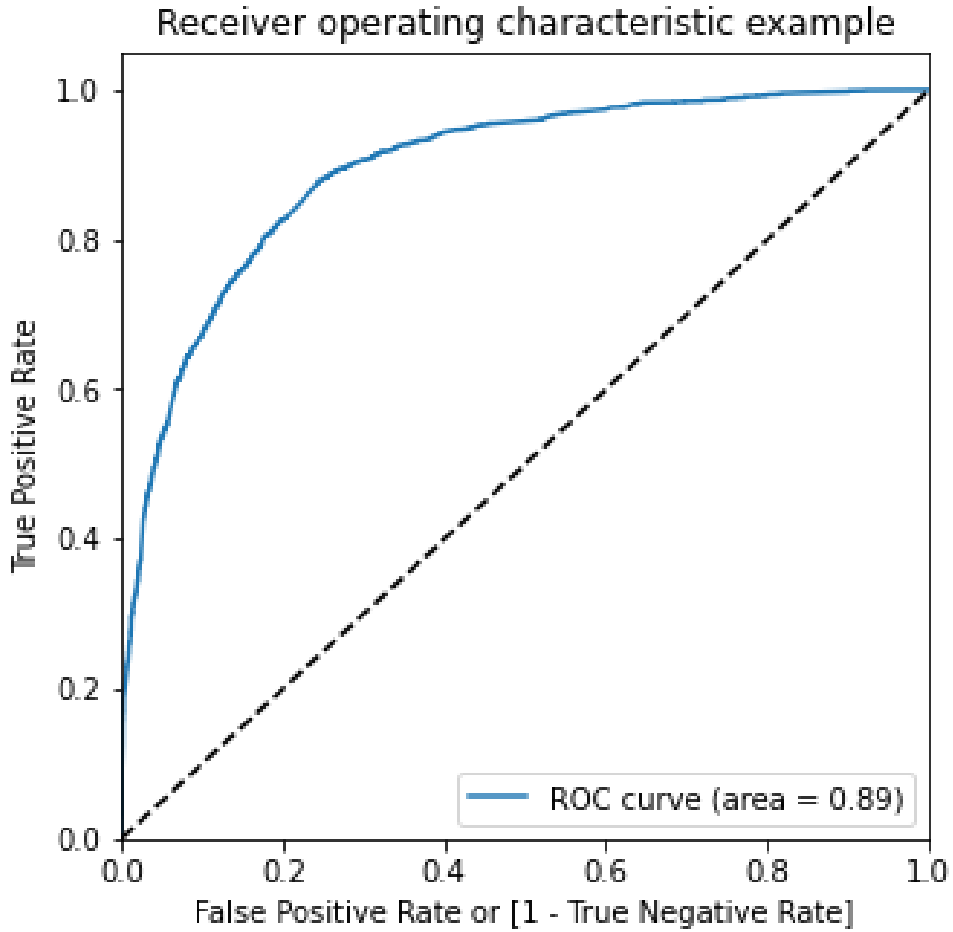
- TotalVisits
- Total Time Spent on Website
- Last Activity_Email Bounced
- Do Not Email
- Lead Source_Olark Chat
- Lead Origin_Lead Add Form
- Last Activity_Unreachable
- Last Notable Activity_Unreachable
- What is your current occupation_No Information
- Last Activity_Olark Chat Conversation
- Last Notable Activity_SMS Sent
- Lead Source_Welingak Website
- Last Notable Activity_Email Bounced
- What is your current occupation_Working Profes...
- Last Activity_Not Sure
- Last Activity_Converted to Lead
- Last Notable Activity_Had a Phone Conversation

Confusion Matrix		
Predicted Actual	Not Converted	Converted
Not Converted	3397	461
Converted	725	1737

Model Evaluation Parameters on Train Dataset	
	Train
Accuracy	0.8158
Sensitivity	0.7152
Specificity	0.8799
False Positive Rate	0.1200
positive predictive value	0.7918
Negative predictive value	0.8288

The Confusion Matrix and Model evaluation parameter indicates that model is good to test.

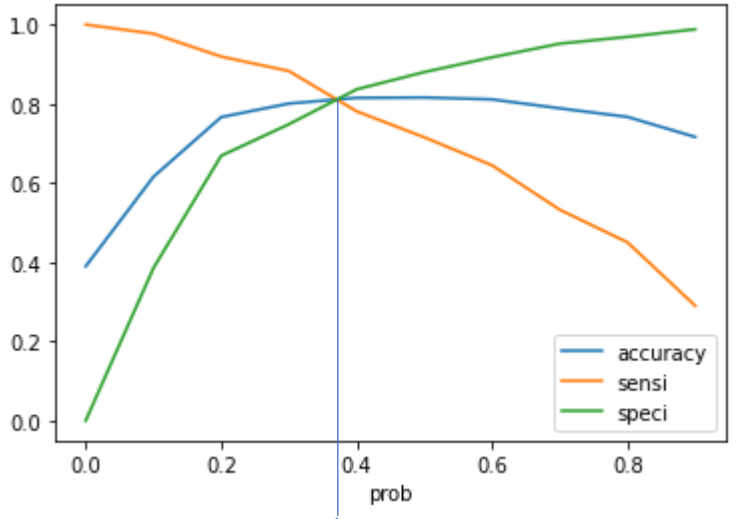
ROC Curve



Further, the ROC curve affirms that the model is good to test.

Model Evaluation on Predictive Probability Values

- ✓ **Let's create a dataframe with the probability value and its binary indicator.**
- ✓ **Based on the predictive data, plotted the accuracy, sensitivity, and specificity versus probability plot to find the cutoff probability.**



Taking the point of intersection, i.e. 0.38 as the cutoff probability value (optimum point) to decide the merit of the lead. Any lead with probably of conversion more than 0.38 will be considered as a Hot Lead.

Model Evaluation based on Optimum Point

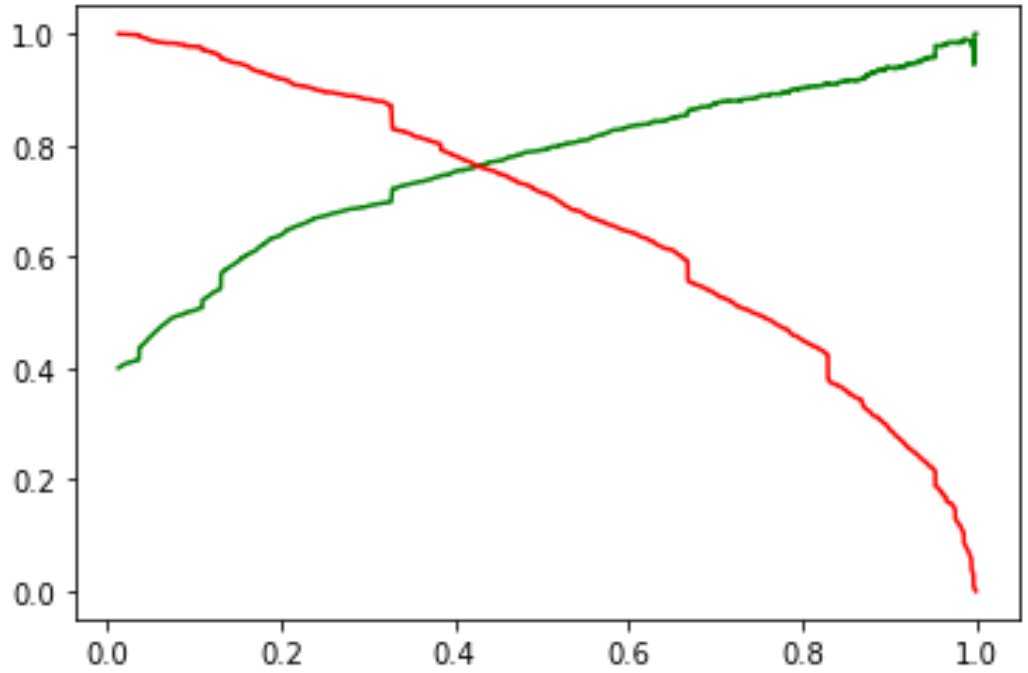
- ✓ **Now, let's create a new dataframe with the probability value and its binary indicator (0 or 1) where all the probability > 0.38 indicates 1 otherwise 0 .**
- ✓ **Based on this, we get correct prediction percentage of 80.38% which is above 80% threshold given by CEO:**

```
final_predicted conversions percentage = 80.38%
```

Confusion Matrix		
Predicted Actual	Not Converted	Converted
Not Converted	3174	684
Converted	483	1979

Model Evaluation Parameters on Probability Prediction Data	
	Value
Accuracy	0.8153
Sensitivity	0.8038
Specificity	0.8227
False Positive Rate	0.1772
positive predictive value	0.7431
Negative predictive value	0.8679

Precision and Recall



Confusion Matrix		
Predicted / Actual	Not Converted	Converted
Not Converted	3395	463
Converted	701	1761

Precision	0.7918
Recall	0.7152

Model Evaluation based on Test Data Set

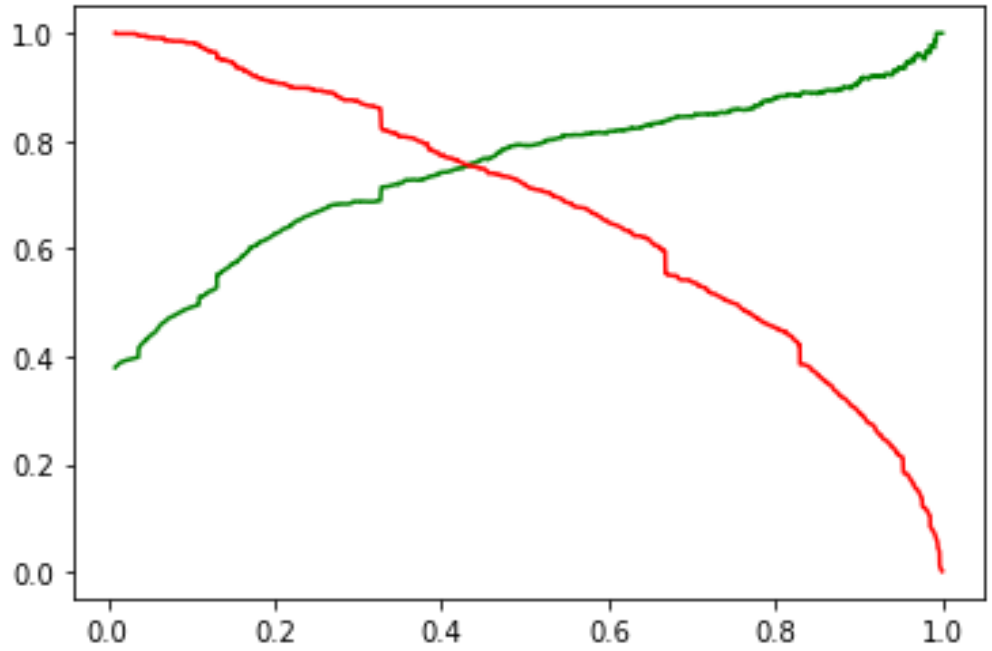
- ✓ **Based on this, we get correct prediction percentage of 79.50% which is near 80% threshold given by CEO:**

```
final_predicted conversions percentage = 79.50%
```

Confusion Matrix		
Predicted / Actual	Not Converted	Converted
Not Converted	1396	298
Converted	208	807

Model Evaluation Parameters on Test Data	
	Value
Accuracy	0.8132
Sensitivity	0.7950
Specificity	0.8240

Precision and Recall on Test Data Set



Confusion Matrix		
Predicted / Actual	Not Converted	Converted
Not Converted	1396	298
Converted	208	807

Precision	0.7303
Recall	0.7950

Prediction and Lead Scoring

	Converted	LeadId	Converted_Prob	final_predicted	lead_score
0	1	2296	0.91	1	91
1	0	8697	0.60	1	60
2	0	7378	0.19	0	19
3	0	8631	0.46	1	46
4	1	4799	0.95	1	95
5	0	4503	0.38	1	38
6	0	7129	0.94	1	94
7	0	1717	0.13	0	13
8	0	8398	0.90	1	90
9	1	5116	0.33	0	33
10	0	1838	0.16	0	16
11	1	5057	0.63	1	63
12	0	7015	0.01	0	1
13	0	6352	0.32	0	32
14	0	575	0.11	0	11
15	1	4597	0.21	0	21

Created a Lead score for each lead, which indicates the chances of it being converted.
Higher Lead score depicts a better chances of being converted

- Built the model with above 80.38% prediction rate.
- Tested the model with 79.5% prediction rate.
- Checked and evaluated the model on below parameters for train, probability, and test dataset:
  - ✓ Overall Accuracy
  - ✓ Sensitivity
  - ✓ Specificity
  - ✓ Confusion Matrix parameters
  - ✓ Precision and Recall Score
- Identified the optimum cut-off probability value of 0.38 for the lead to fall under Hot Leads Category.
- Created the final prediction dataframe and assigned a **Lead Score to each lead.**

Overall, the model looks good on all the parameters.