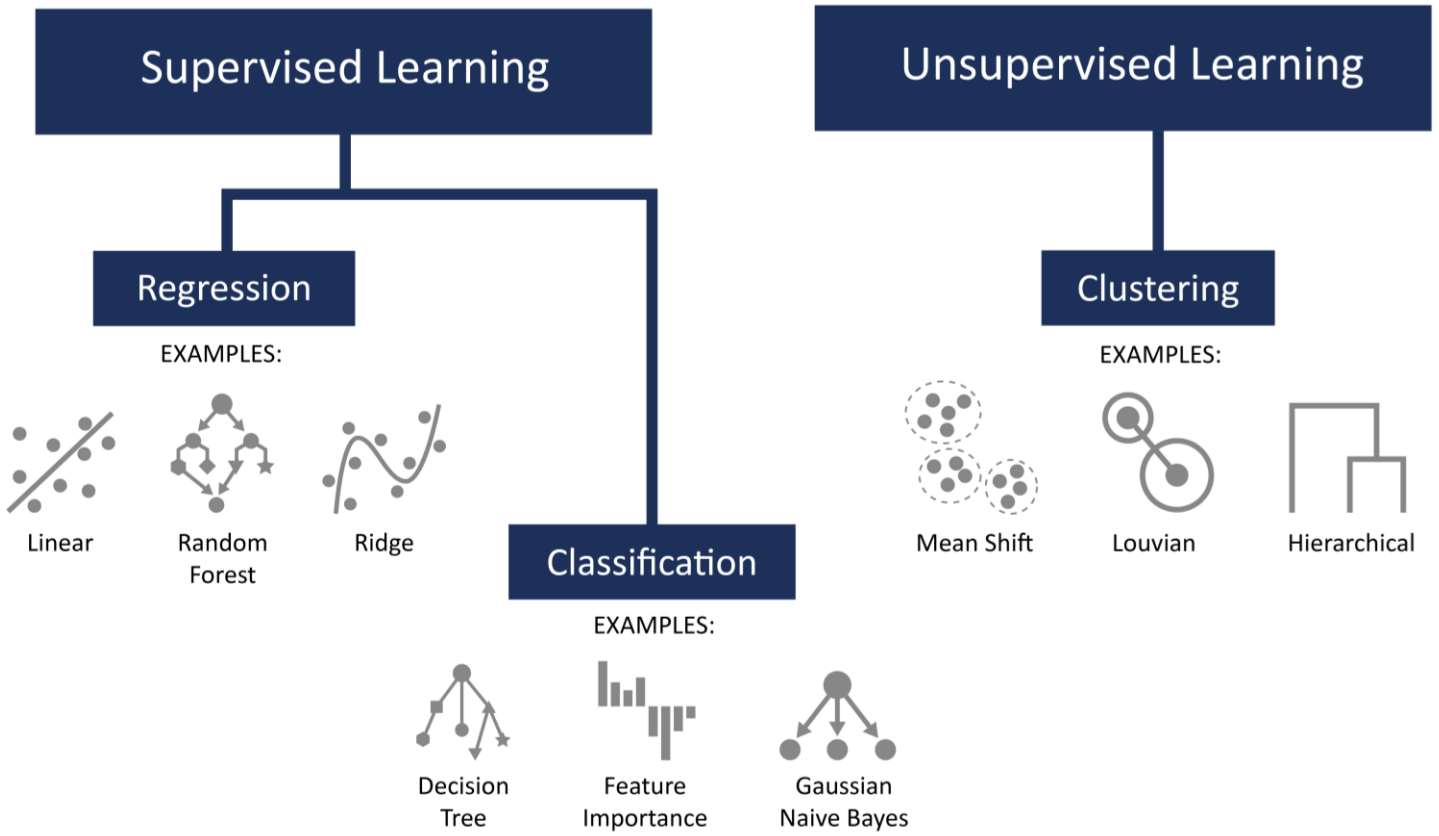
Entry2DA Course Series

Python Foundation – Python 101

Quynh-Anh Dang, Data Scientist

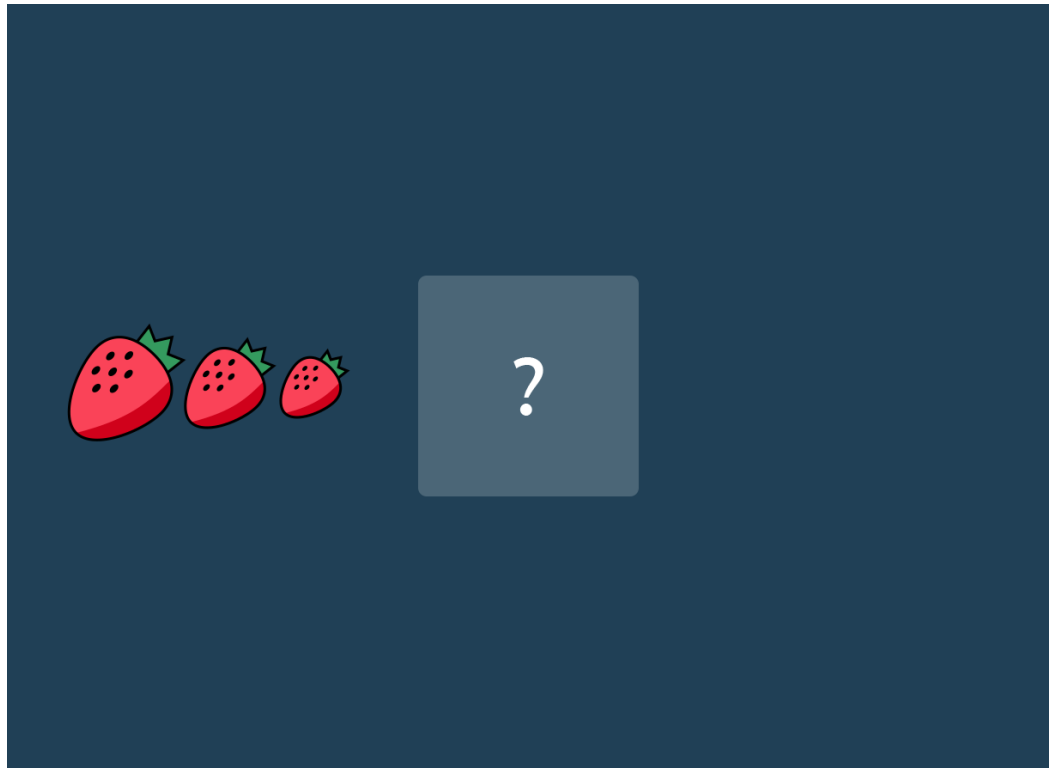
Section 1 : Review

Machine Learning approaches

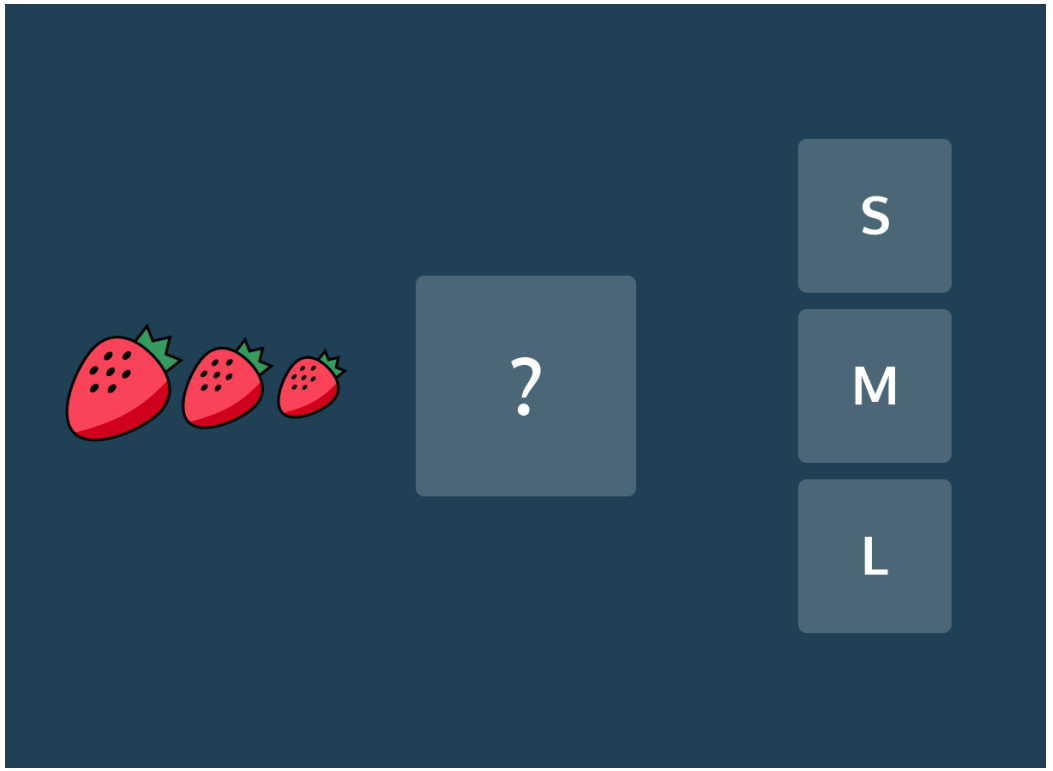


Regression vs Classification

Regression: outputs are continuous numbers

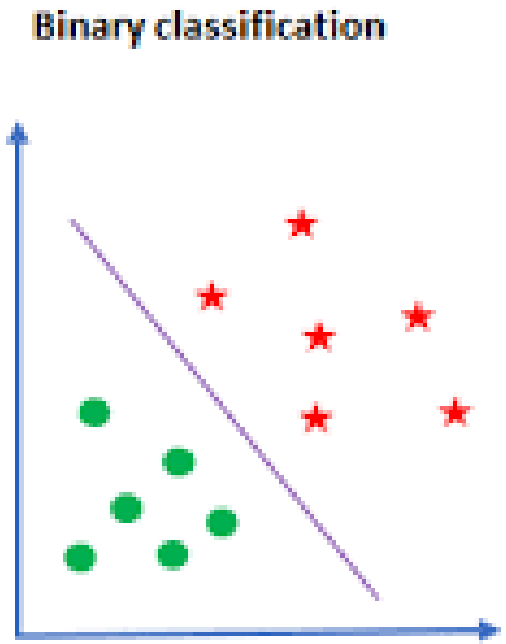


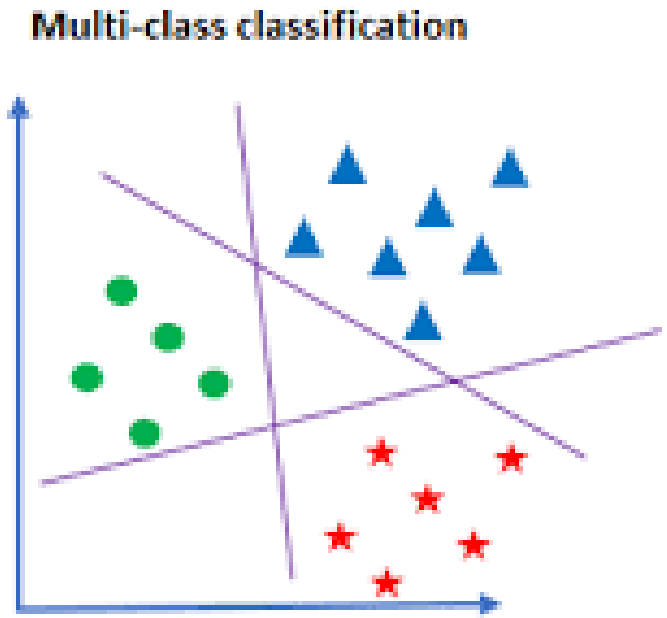
Classification: outputs are discrete labels



Section 2: Classification

Classification





Classification

Predict: if a customer would click an Ad or not?

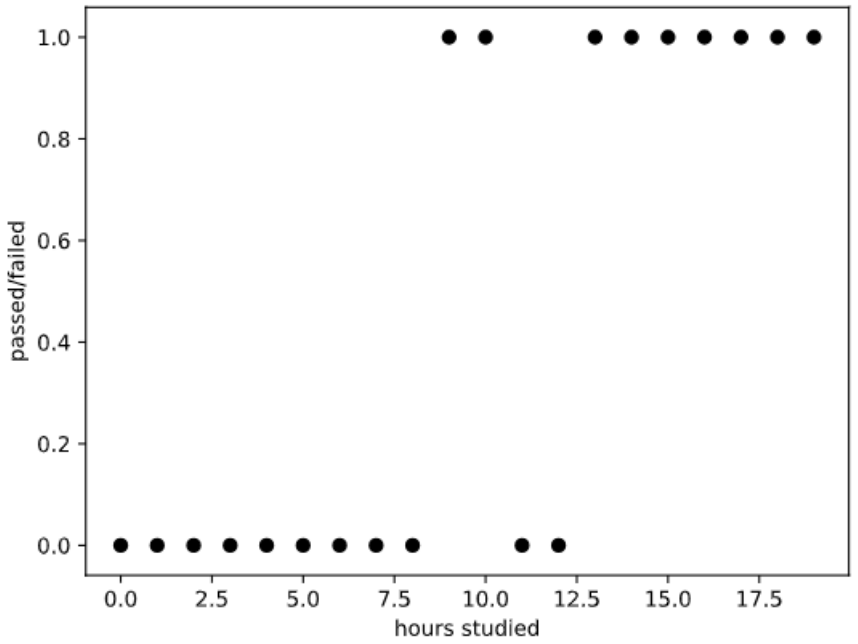
	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
0	68.95	35	61833.90	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	2016-03-27 00:53:11	0
1	80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	2016-04-04 01:39:02	0
2	69.47	26	59785.94	236.50	Organic bottom-line service-desk	Davidton	0	San Marino	2016-03-13 20:35:42	0
3	74.15	29	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	2016-01-10 02:31:19	0
4	68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	2016-06-03 03:36:18	0
...	...	...	...	...	...	...	...	...	...	...
995	72.97	30	71384.57	208.58	Fundamental modular algorithm	Duffystad	1	Lebanon	2016-02-11 21:49:00	1
996	51.30	45	67782.17	134.42	Grass-roots cohesive monitoring	New Darlene	1	Bosnia and Herzegovina	2016-04-22 02:07:01	1
997	51.63	51	42415.72	120.37	Expanded intangible solution	South Jessica	1	Mongolia	2016-02-01 17:24:57	1
998	55.55	19	41920.79	187.95	Proactive bandwidth-monitored policy	West Steven	0	Guatemala	2016-03-24 02:35:54	0
999	45.01	26	29875.80	178.35	Virtual 5thgeneration emulation	Ronniemouth	0	Brazil	2016-06-03 21:43:21	1

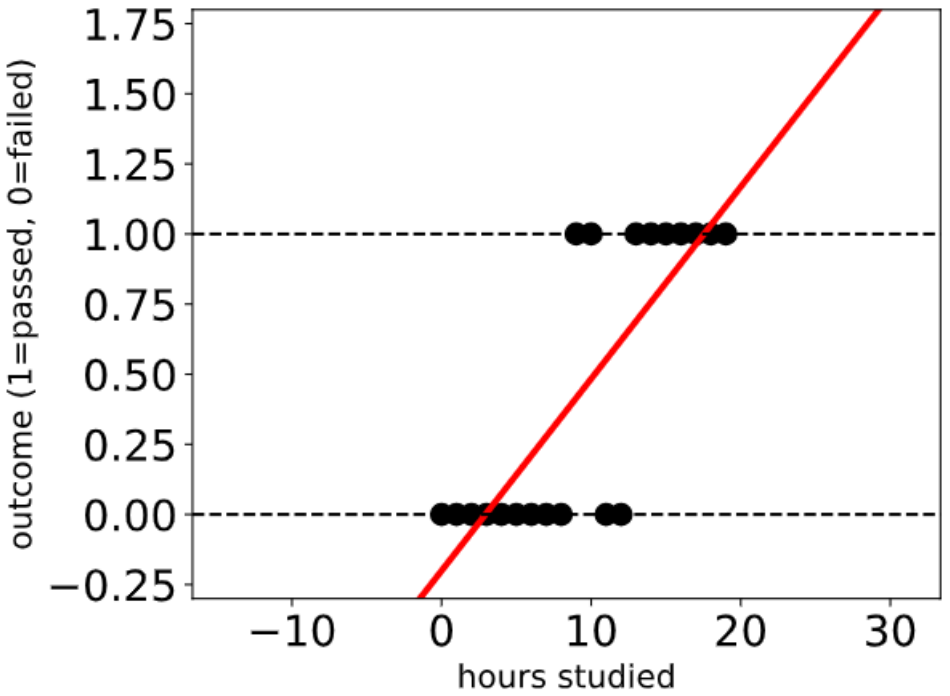
Classification algorithms

- Logistic Regression
- Decision Tree
- Random Forest
- Naive Bayes
- K-Nearest Neighbors
- Support Vector Machine
- Gradient Boosting
- …

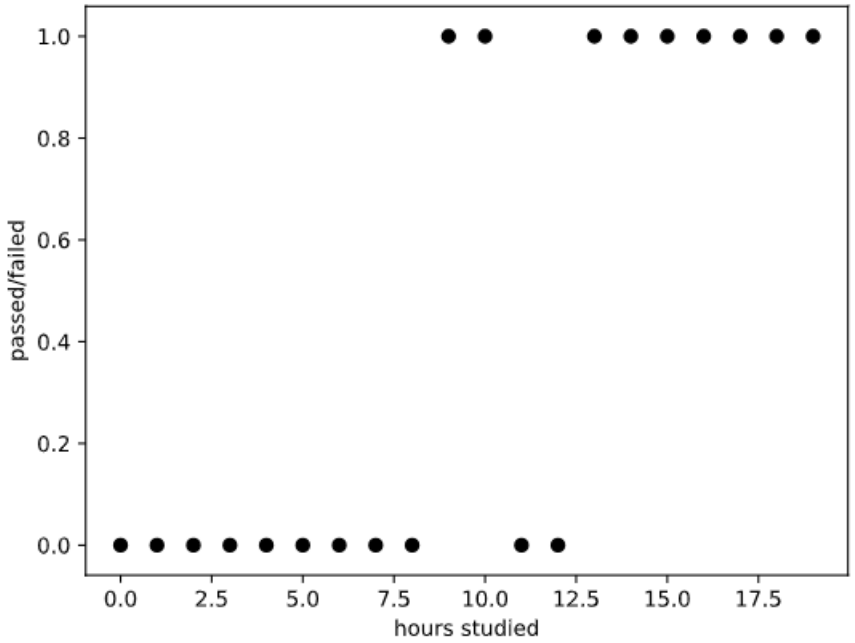
Logistic regression

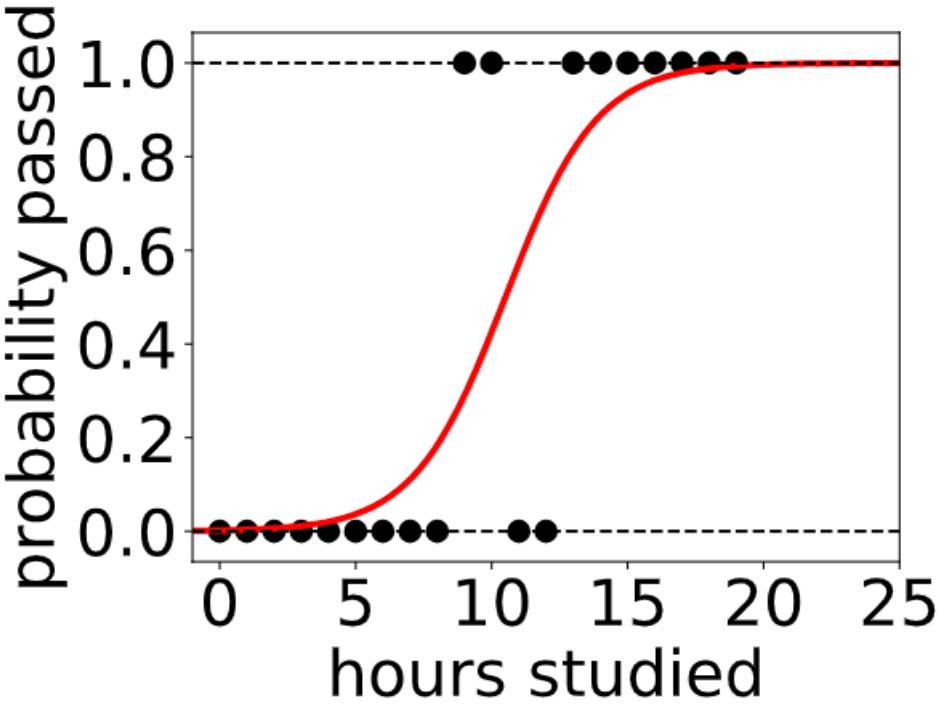
Logistic Regression





Logistic Regression



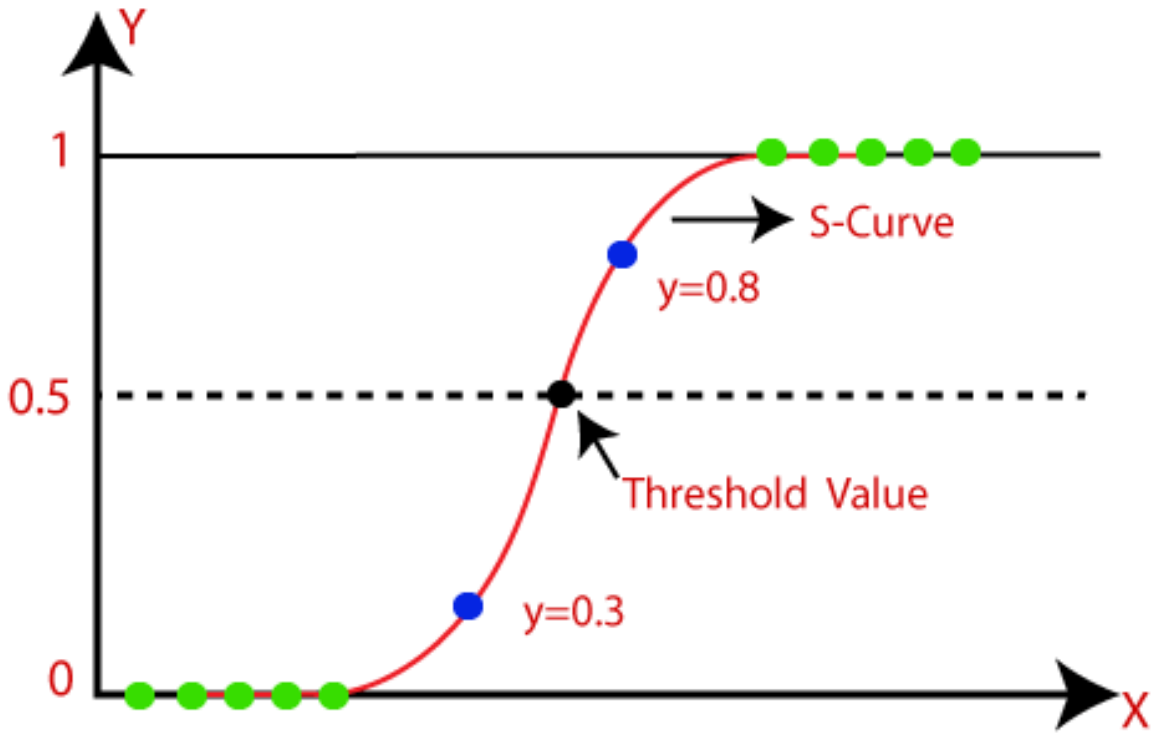


Logistic regression = find the best-fitting S-shaped curve that maps the input features to the probability of the positive class (class 1)

Logistic Regression

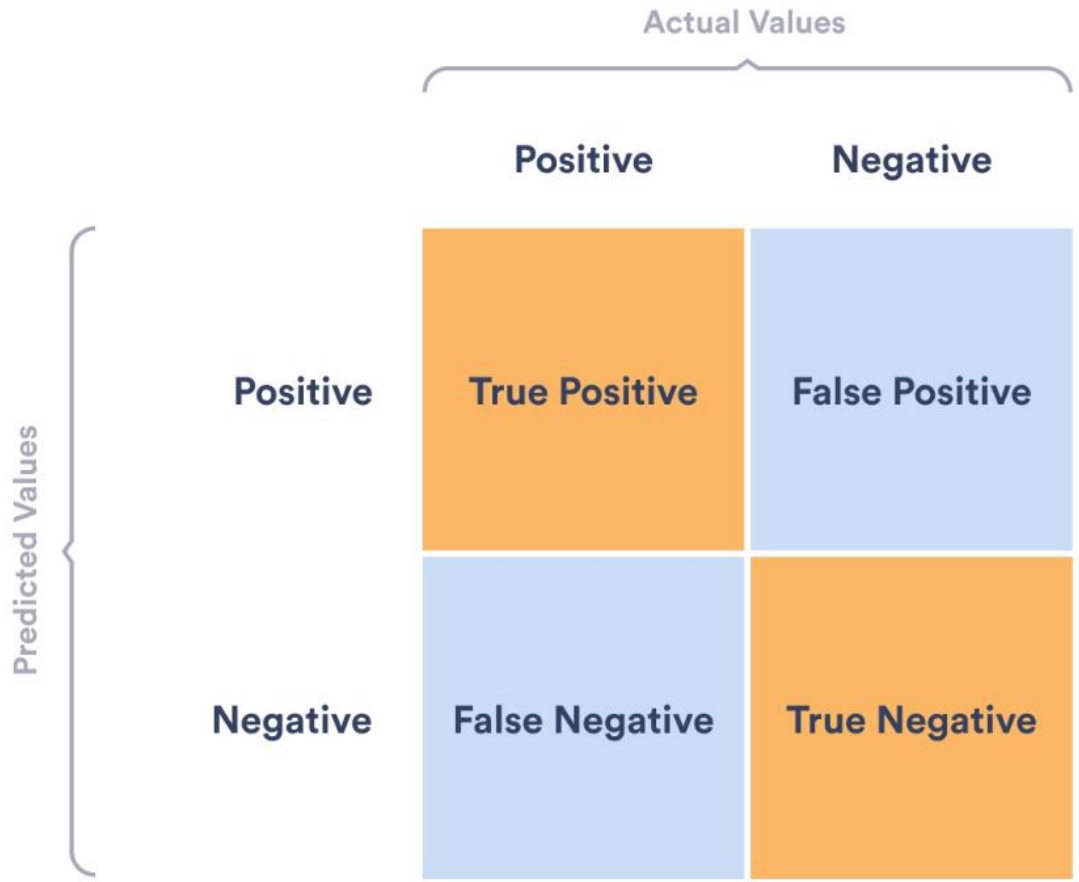
X1	X2
…	
…	

$$P(Y = 1) = ?$$



Evaluation

Confusion Matrix



Basic Terminology:

- True Positives (TP): dự đoán là 1 và thực sự là 1
- True Negatives (TN): dự đoán là 0 và thực sự là 0
- False Positives (FP): dự đoán là 1 thực sự 0
- False Negatives (FN): dự đoán là 0 thực sự là 1

Accuracy

$$Accuracy = \frac{TP + TN}{Total\ number\ observation}$$

➔ Out of the predictions made by the model, what percentage is correct?

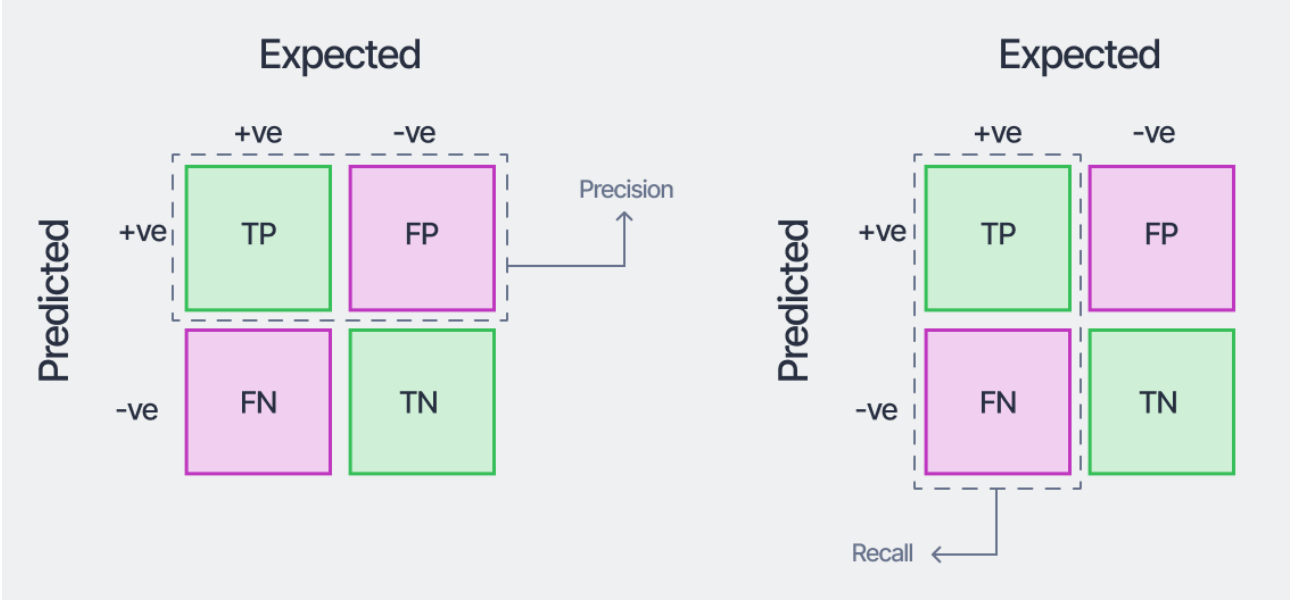
Precision, Recall

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

➔ Out of all the YES predictions, how many of them were correct?

➔ How good was the model at predicting real YES events?

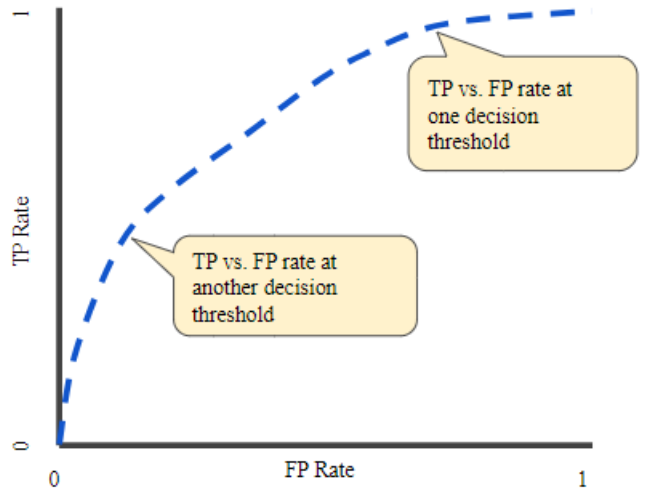


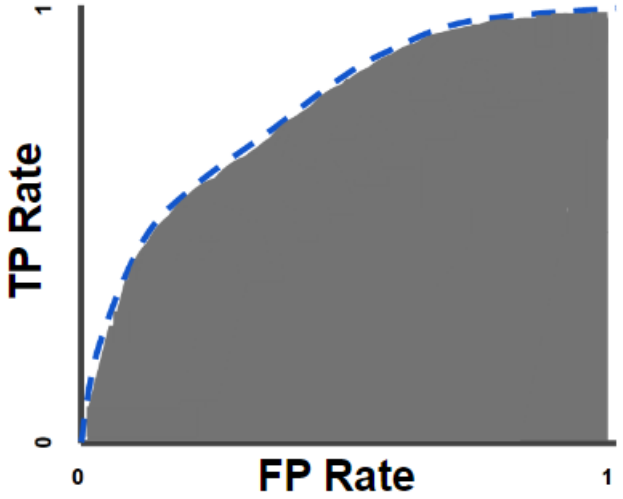
F1-Score

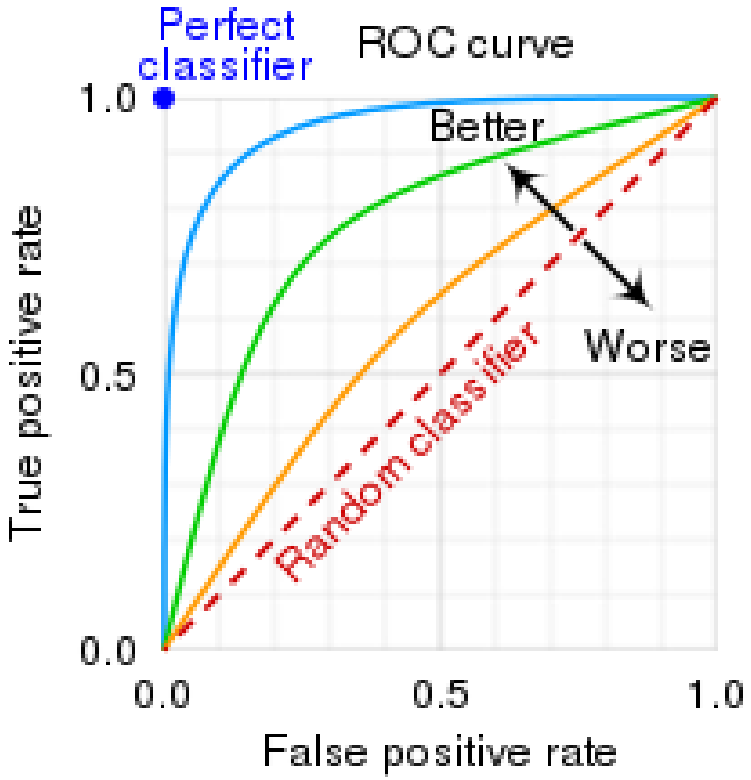
$$F1\ Score = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$$

➔ Harmonic mean of the precision and recall

AUC/ ROC Curve







- An **ROC** is a graph showing the performance of a classification model at all classification thresholds. An ROC curve plots TPR vs. FPR at different classification thresholds.
- **AUC** stands for "Area under the ROC Curve“, which measures the entire two-dimensional area underneath the entire ROC curve.

Choose the right metric

DATATOMIC

Choose accuracy
- The cost of FP and FN are roughly equal.
- The benefit of TP and TN are roughly equal.

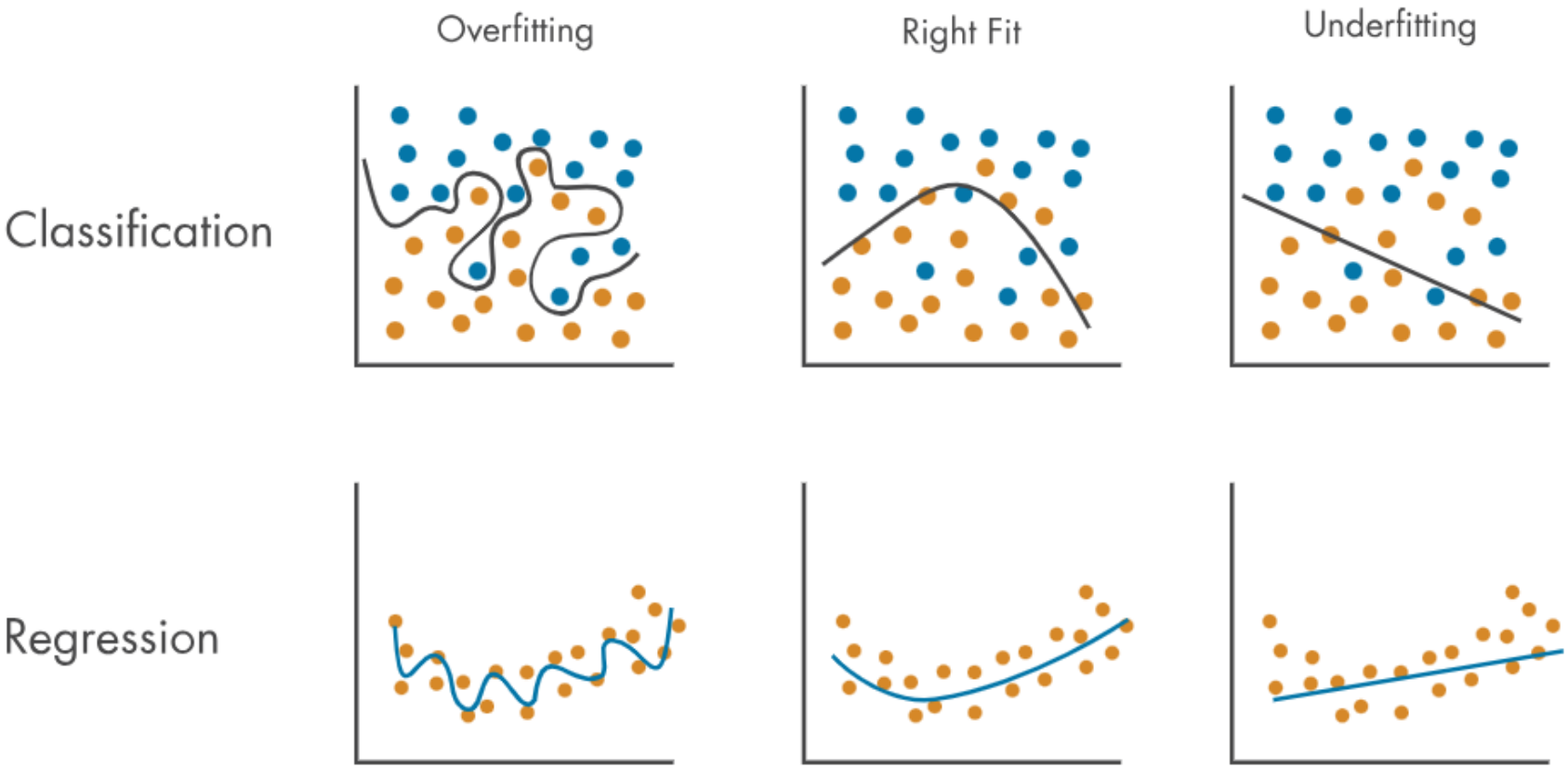
Choose Precision
- The cost of FP is much higher than a FN.
- The benefit of a TP is much higher than a TN.

Choose recall
- The cost of FN is much higher than a FP.
- The cost of a TN is much higher than a TP.

ROC AUC & Precision – Recall curves
- Use ROC when the dealing with balanced data sets.
- Use precision-recall for imbalanced data sets.

Problems

Overfitting/ Underfitting

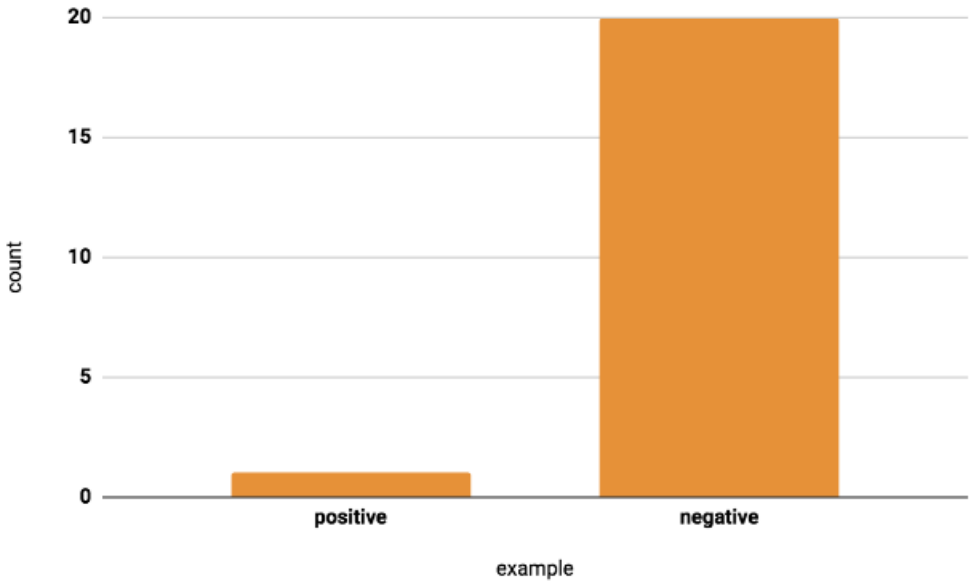


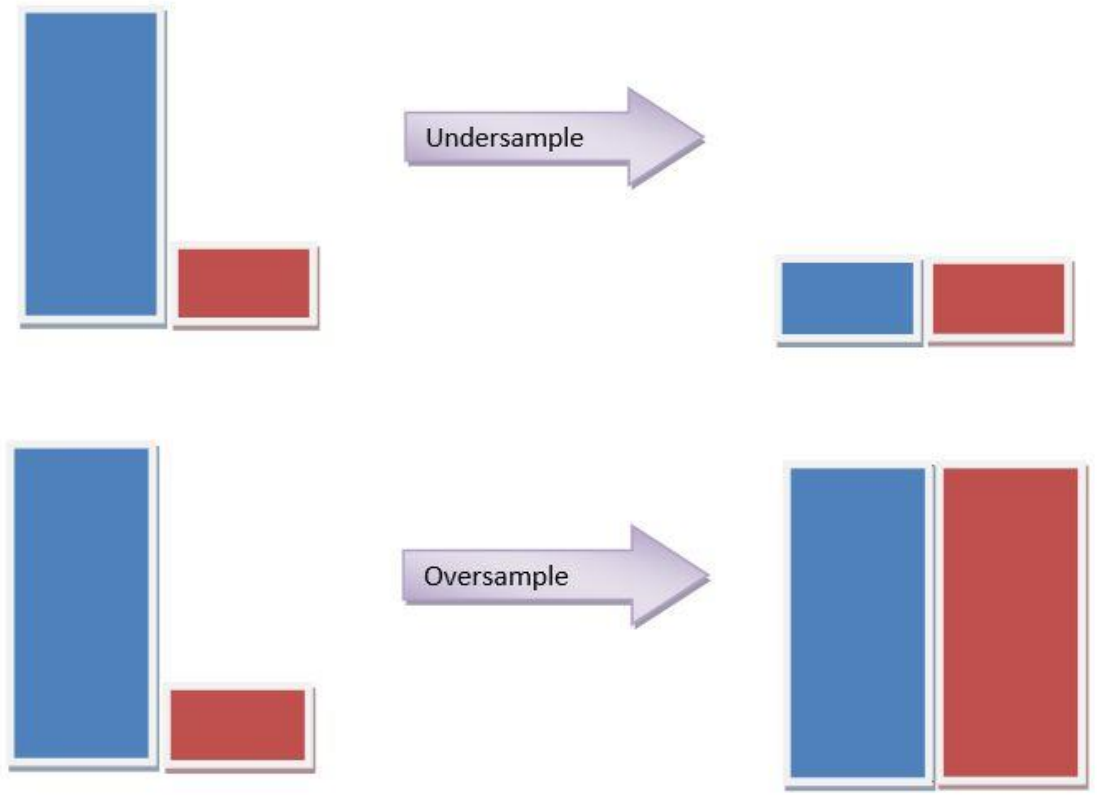
Overfitting/ Underfitting

DATATOMIC

	Underfitting	Overfitting
Reasons	• The model is too simple. • The input features which is used to train the model is not the adequate representations of underlying factors influencing the target variable. • The size of the training dataset used is not enough. • Excessive regularization are used to prevent the overfitting, which constraint the model to capture the data well. • Features are not scaled.	• The model is too complex. • The size of the training data.
Techniques to Reduce	• Increase model complexity. • Increase the number of features, performing feature engineering. • Remove noise from the data. • Increase the number of epochs or increase the duration of training to get better results.	• Increase training data. • Reduce model complexity. • Early stopping during the training phase (have an eye over the loss over the training period as soon as loss be • gins to increase stop training). Use regularization

Imbalanced dataset





Thank you