Machine Learning Modeling pipeline

Machine Learning Architects Basel

Bassem Ben Hamed

December 2022

ML ARCHITECTS
BASEL

Agenda

- Data Preprocessing
- Imbalanced Data
- Machine Learning Modeling

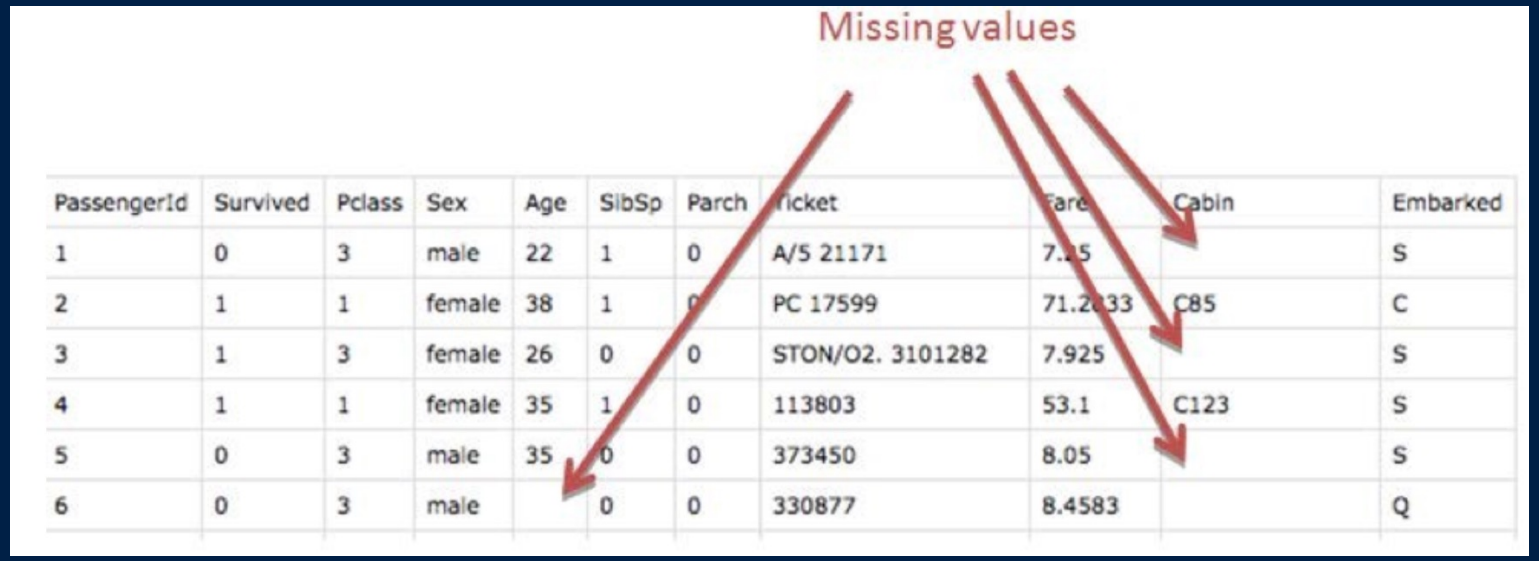
ML ARCHITECTS BASEL

Data Preprocessing

6 easy steps

1. Import all the crucial libraries
2. Import the dataset
3. Identify and handling the missing values
4. Encoding the categorical data
5. Splitting the dataset
6. Feature scaling

Missing Values



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Find the number of NaN:
$ df.isnull().sum()

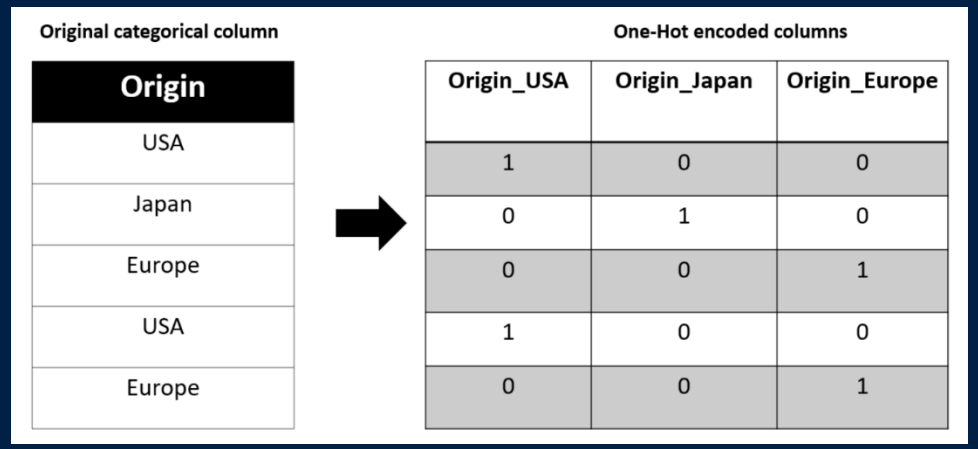
Replace NaN using fillna:
$ df.col = df.col.fillna(strategy)

Imputing NaN:
$ sklearn.impute.SimpleImputer as SI
$ imp = SI(strategy)

- Strategy 1: mean
- Strategy 2: median
- Strategy 3: most frequent
- Strategy 4: constant

Encoding Categorical Data

2 concepts to know: OHE vs LE vs LOOE



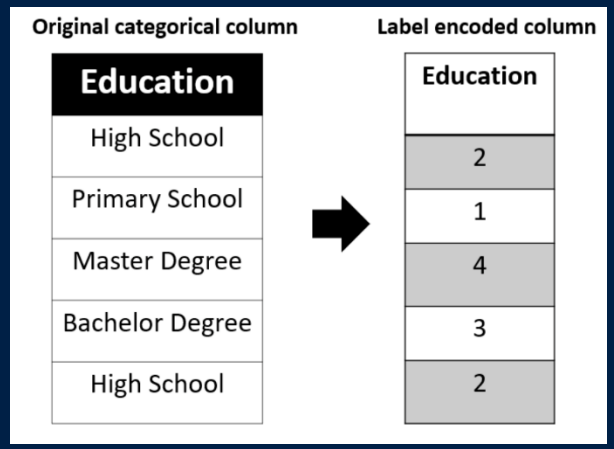
Original categorical column

Origin
USA
Japan
Europe
USA
Europe

One-Hot encoded columns

Origin_USA	Origin_Japan	Origin_Europe
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1

One-Hot Encoding



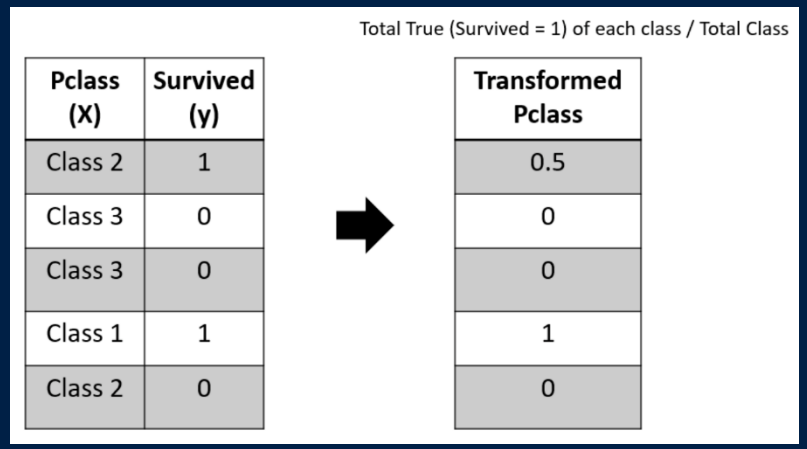
Original categorical column

Education
High School
Primary School
Master Degree
Bachelor Degree
High School

Label encoded column

Education
2
1
4
3
2

Label Encoding



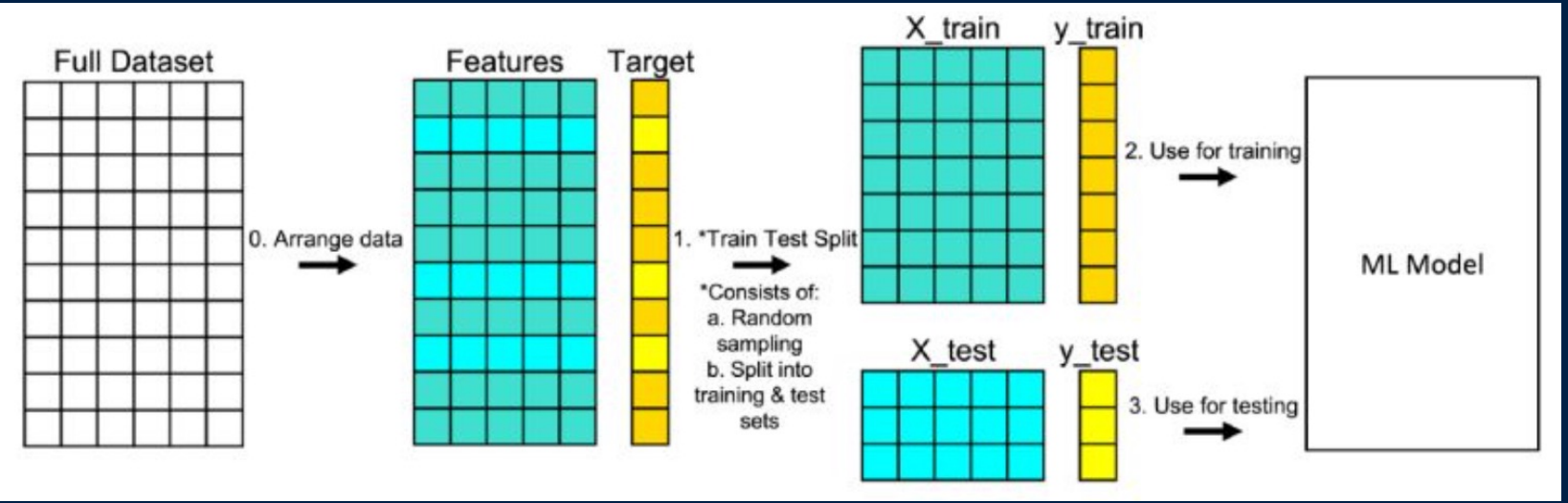
Total True (Survived = 1) of each class / Total Class

Pclass (X)	Survived (y)
Class 2	1
Class 3	0
Class 3	0
Class 1	1
Class 2	0

Transformed Pclass
0.5
0
0
1
0

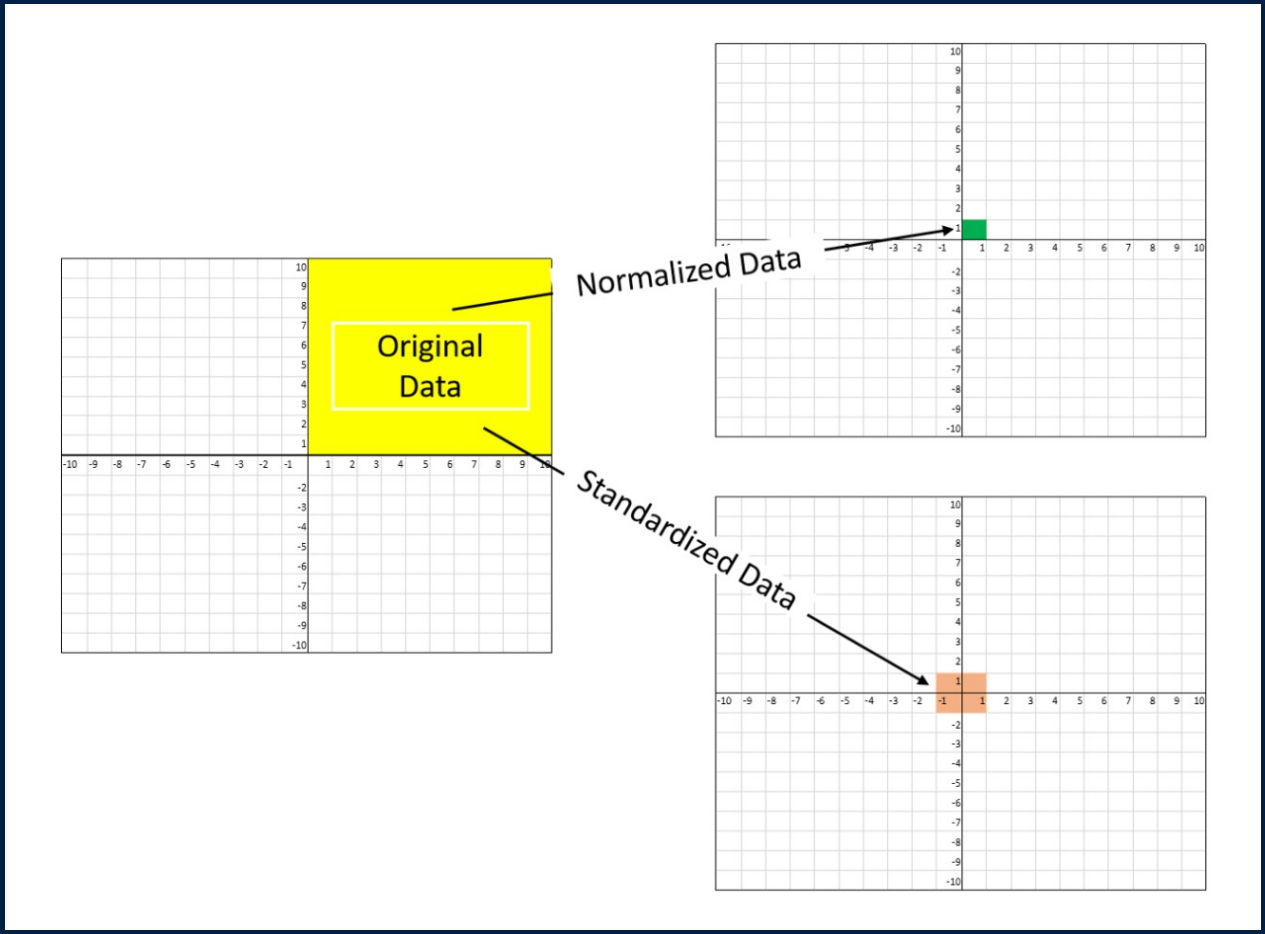
Leave One Out Encoding

Splitting Dataset



1. Arrange the data
2. Split the data
3. Train the model
4. Test the model

Feature Scaling

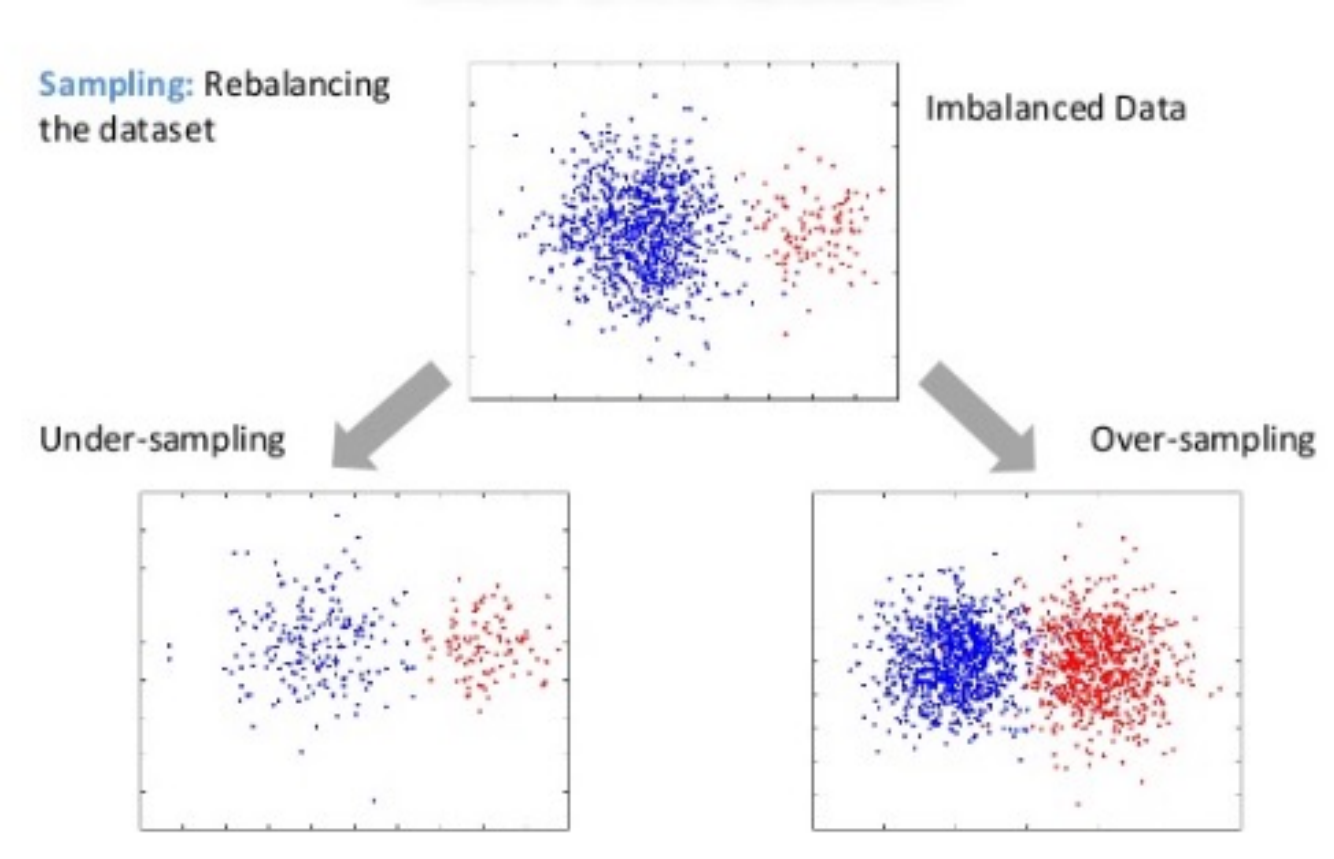


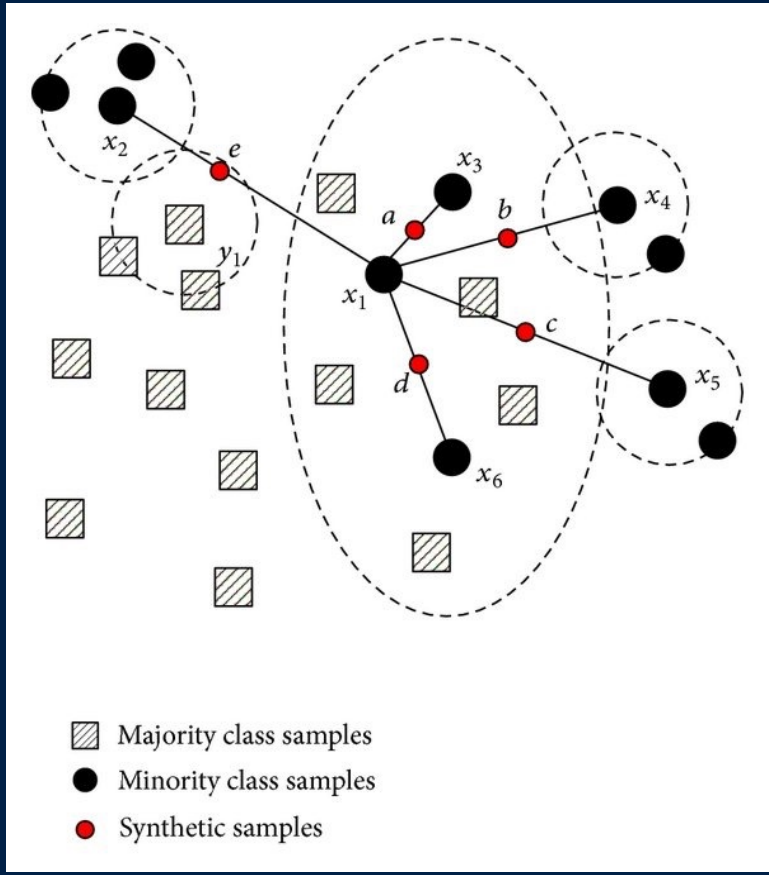
Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

1. Min Max Scaler
2. Standard Scaler
3. Max Abs Scaler
4. Robust Scaler
5. Quantile Transformer Scaler
6. Power Transformer Scaler
7. Unit Vector Scaler

Imbalanced Data

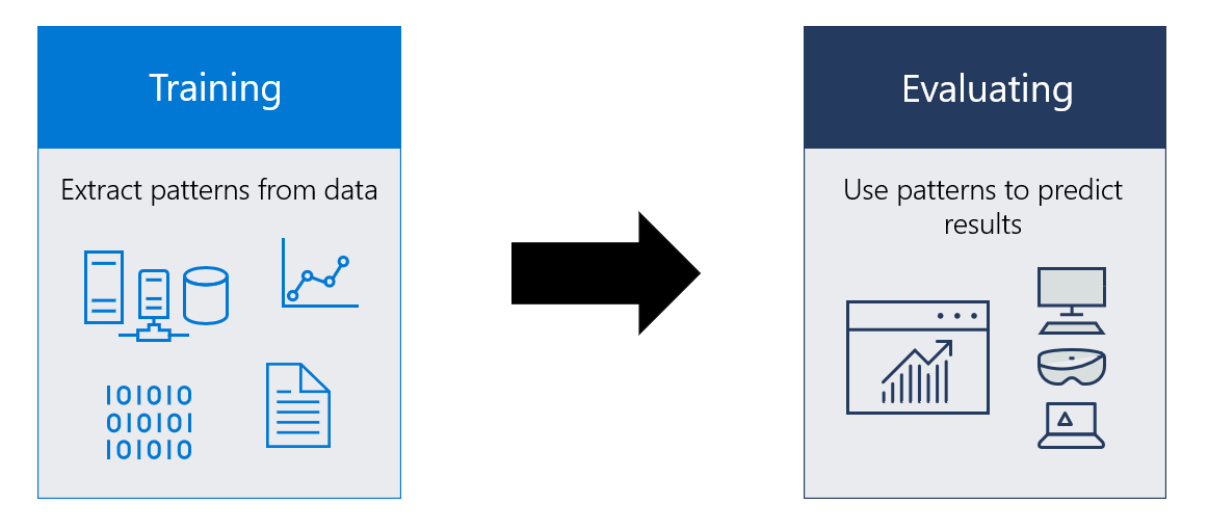
Synthetic Minority Oversampling Technique (SMOTE)





ML ARCHITECTS BASEL

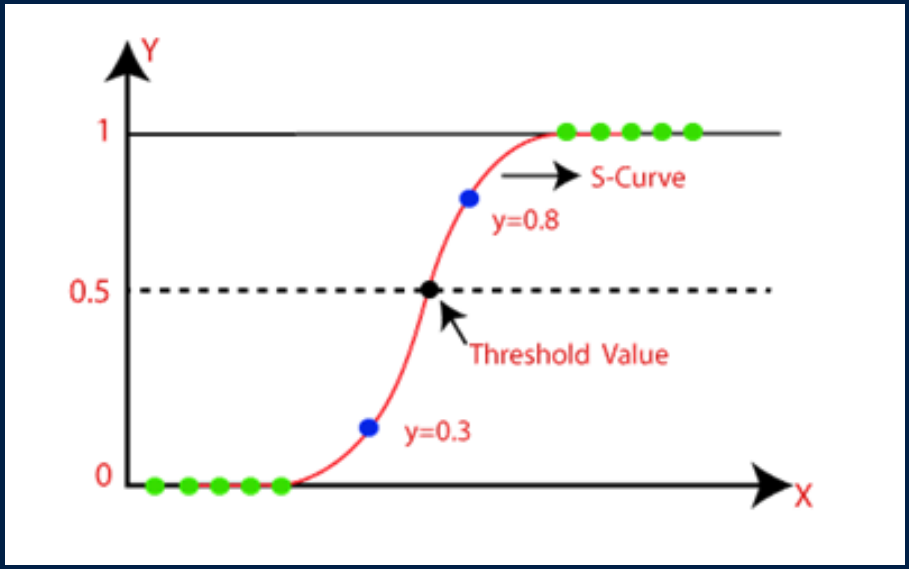
Machine Learning Modeling

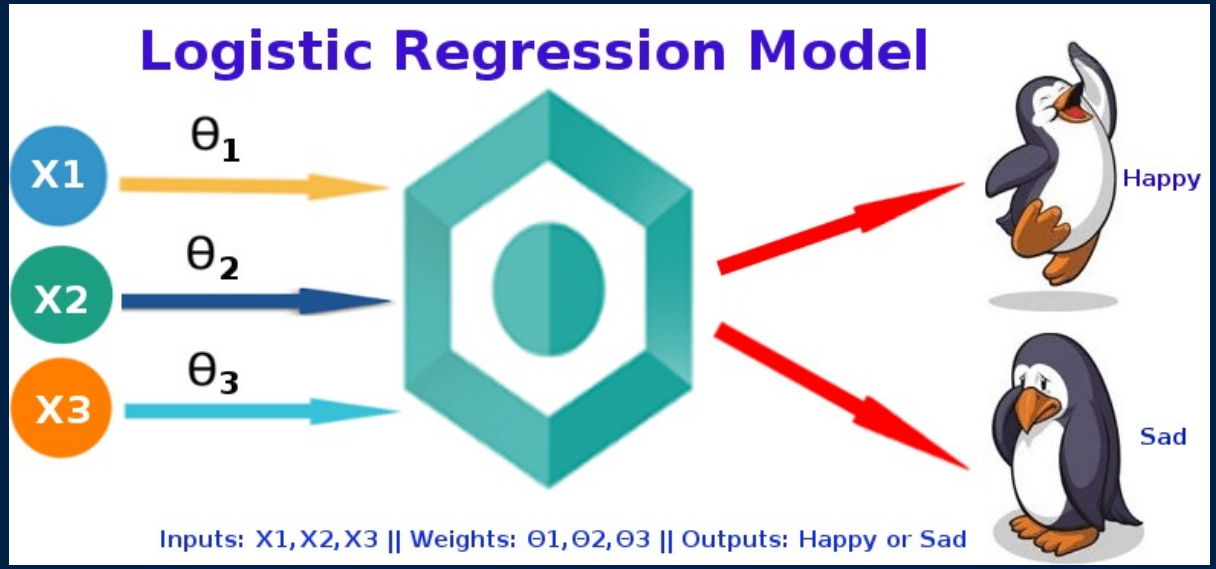


Good Machine Learning scenarios:

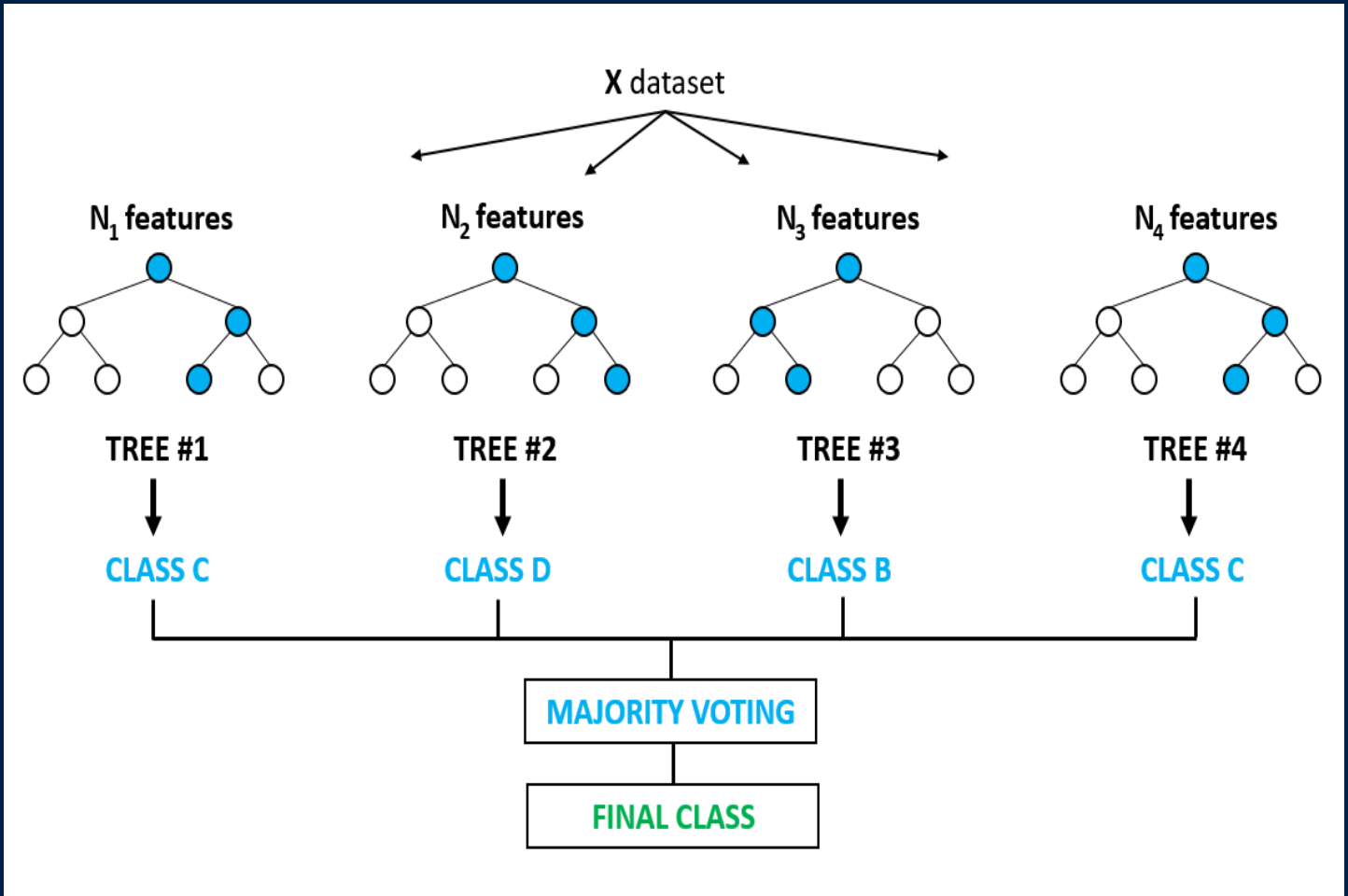
1. They involve a repeated decision or evaluation which you want to automate and need consistent results.
2. It is difficult or impossible to explicitly describe the solution or criteria behind a decision.
3. You have labeled data, or existing examples where you can describe the situation and map it to the correct result.

Logistic Regression





Random Forest



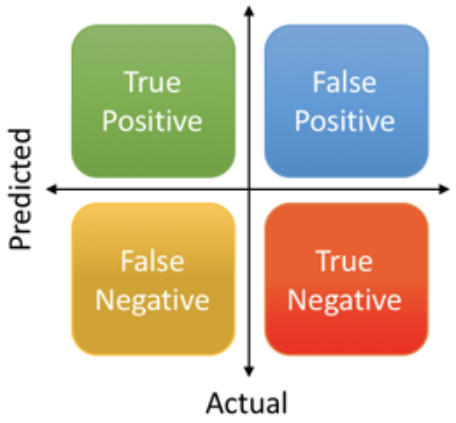
Evaluate Models

Confusion matrix and metrics

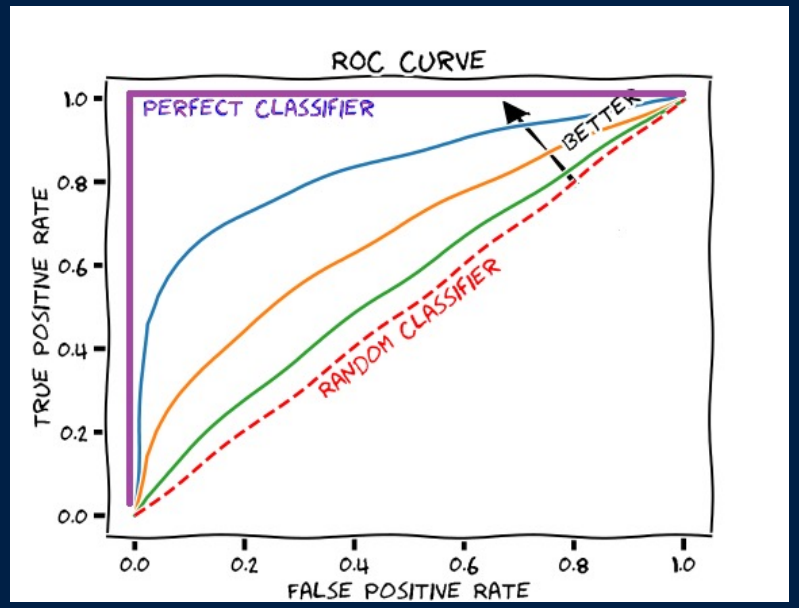
$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



AUC-ROC Curve



ML ARCHITECTS BASEL

Implementing reliable
machine learning solutions

ML ARCHITECTS
BASEL

Operating Models – Technologies – Culture & Skills

Consulting – Engineering - Training