ARTEFACT
VALUE BY DATA

&

Jiao Tong DS Summer Class : Day 3

July 2020

Agenda

1 **Data Science Knowledge**

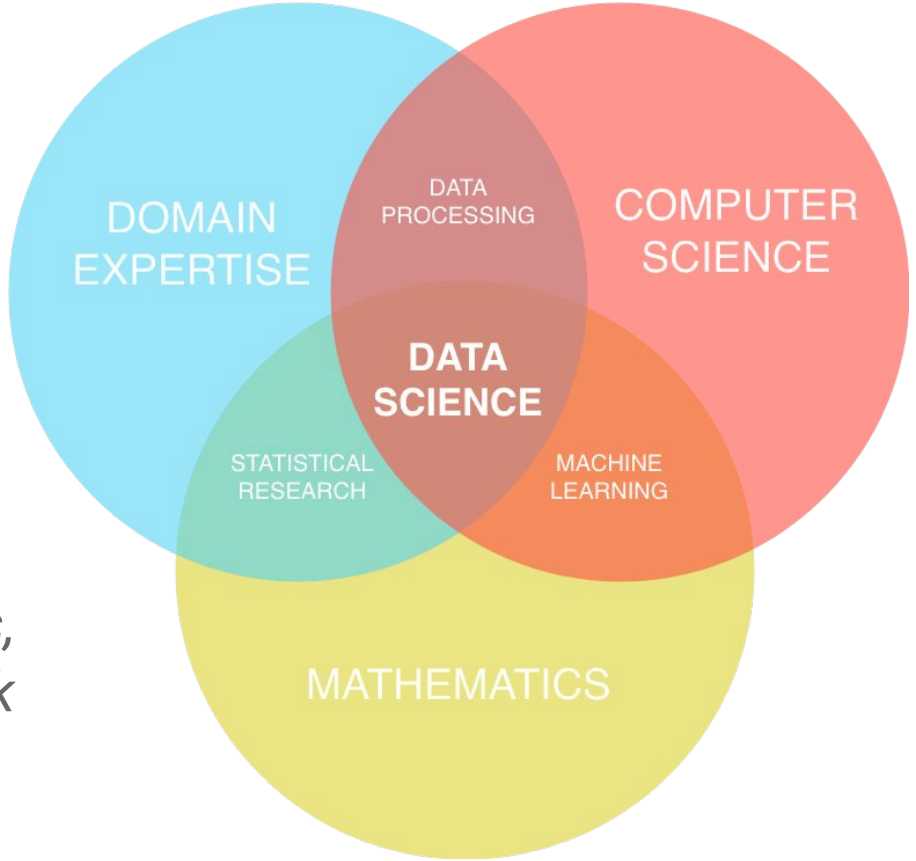
2 Data preparation and modelling approach

What is Data Science ?

Data scientists, rare profiles with highly valued skills

"A Data Scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

Josh Wills,
Director of Data Engineering at Slack



Data Science is a complete and comprehensive way of working with data

Data Science

Data Mining

Exploratory data analysis through the production of descriptive statistics and variable correlation analysis

e.g.: analysis of campaign performance

+

Machine Learning

Statistical methods designed to give a machine the ability to infer the rules defining a problem by itself.

e.g. recognizing a face in a picture

+

Scientific method

Working method aimed at producing reliable results, based on the formation of hypotheses and their confirmation by experience

e.g. research work in theoretical sciences

A growing need for data scientists

Quanthub estimates that there will be a **shortage of 250 000 data scientists** in the **World** in 2020.

Organizations are **beginning to develop internal data scientists**, even if they are still highly externalized.

What are the data scientist's tools ?

Data Scientist tools

Calculation frameworks
Definition of the calculation method adopted by the infrastructure

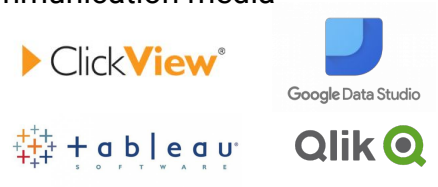
Programming languages
Allows operations to be translated into computer-interpretable instructions

Analysis tools
Organize the project and have a synthetic view of the sequence of operations performed

Data base
Storage and provision of data

SQL vs NoSQL

Visualization tools
Creation of visualization and communication media



Cloud computing tools
Ability to add machines to increase power in a simple way

How to choose between Cloud and On-premise?

7 evaluation criteria

	On-premise	Cloud
Data connection	Data hosted on the on-premise	Flows to put in place to the Cloud
Extra power available	Power limited by the number of cases in production	Horizontal Scalability
Operating costs	Costs already supported by the IS	Costs related to machines in service
Ease of implementation of tools	Limited toolbox	Open source tools available
Innovation	New machines and tools rarely available	Availability of all tools as soon as they are available. in open source
Compliance (Location, RGPD, traceability)	Full control	Ability to choose the location of the data, but trust the necessary cloud provider
security	Full control	Set up default security, access to additional solutions

How to Data Science ?

Machine Learning responds to problems that are too complex to be explicitly programmed

Classic problems

It is possible to explain the procedure to follow at the machine

Example

Programming an automaton of simple tasks

Binding problems

The problem is so complex that it is not adequate to specify the set of rules to solve it.

Example

Recognizing a dog in a photo

Problems that should be solved with Machine Learning!

The choice of the algorithm used is the result of the need defined beforehand

Need

What is the purpose of the analysis ?

Classification :
Predict belonging to different classes

Regression:
Predict the value that a continuous variable will take

Clustering :
Create groups of data points with similarities

Approach

What data do I have at my disposal ? How will my model learn ?

Supervised Learning :
The algorithm learns from qualified data (= examples)

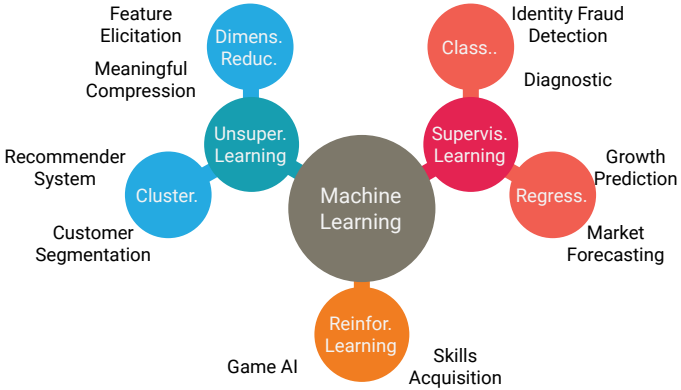
Unsupervised Learning :
Algorithm derives rules from unqualified data

Reinforcement Learning :
The algorithm finds the right way to deal with a problem in successive iterations

Algorithm Choice

Which algorithm should I choose consequently ?

There are a multitude of algorithms that can meet different needs / approaches!



I want to use

Machine Learning

to improve my performances on Mario Kart, how can I do it?

3 different AI approaches, each of which answers a different question

Supervised approach

Can I predict who will be the best driver?

We first try to determine what are the **characteristics that make a player win**.

If we know the characteristics of a new player, we are able to predict the **probability of winning a race**.

Unsupervised approach

What criteria differentiate pilots?

We seek to identify **homogeneous groups of pilots**.

From the **strengths and weaknesses of each group**, we can build game strategies!

Enhancement

Without any assumptions, how to win the race?

We try to learn **how to win a race,** according to very limited assumptions: on a given field, with a given driver, and with two choices of action: banana skins and turbo

The chosen approach depends on the problem to be addressed, but also on technical requirements.

Do I have a **data history**?

NO

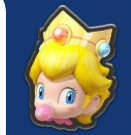
I have to use an AI method called **"Reinforcement"** that does not need a learning base to work.

YES

Can I base my decision on an **objective and known performance criterion** such as the victory rate of my riders?

YES

I will try to **predict based on past win rates** who is most likely to win tomorrow! I will use **a supervised approach**



?

NO

I will create **groups of runners** to identify their strengths. This is an **unsupervised approach**

Unsupervised approach: I am interested in knowing the different pilot groups and their characteristics (1/2)

I have a clean database with the different features of my drivers.

I do not know the victory rate or other variable that would allow me to correlate the characteristics of my drivers to a victory.

Character	Speed	Acceleration	Weight	Maneuverability	Adherence	Mini-Turbo
Baby Peach	9	13	9	19	18	12
Bowser Jr	13	12	12	15	16	10
Daisy	13	11	13	15	16	10
Donkey Kong	17	9	17	11	14	8
Mario	15	10	15	13	15	9
Roi Boo	17	8	19	11	13	7
Toad	11	12	11	17	17	11
Wario	19	8	19	9	13	7

Unsupervised approach: I am interested in knowing the different pilot groups and their characteristics (2/2)

2

I apply on this basis an unsupervised algorithm that defines groups of pilots who are alike.

Character	Speed	Accel.	Weight	Maneuverability	Adher.	Mini-Turbo
Baby Peach	9	13	9	19	18	12
Toad	11	12	11	17	17	11
Bowser Jr	13	12	12	15	16	10
Daisy	13	11	13	15	16	10
Mario	15	10	15	13	15	9
Roi Boo	17	9	17	11	14	8
Donkey Kong	17	8	19	11	13	7
Wario	19	8	19	9	13	7

3

I analyze my groups of drivers called clusters a posteriori. I "name" my groups manually.

Character	
Baby Peach	Light weights, with a strong acceleration
Toad	
Bowser Jr	Averages, with a correct level everywhere
Daisy	
Mario	
Roi Boo	Heavy, fast but unmanageable
Donkey Kong	
Wario	

Depending on the terrain, I decide which group of riders I will call.

Supervised approach: I'm looking for the best driver!

I have a clean database with the different features of my drivers.

I know the victory rate of each player.

Character	Speed	Acceleration	Weight	Maneuverability	Reason.	Mini-Turbo	Victory rate
Baby Peach	9	13	9	19	18	12	32 %
Bowser Jr	13	12	12	15	16	10	27 %
Daisy	13	11	13	15	16	10	12 %
Donkey Kong	17	9	17	11	14	8	31 %
Mario	15	10	15	13	15	9	40 %
Roi Boo	17	8	19	11	13	7	7 %
Toad	11	12	11	17	17	11	18 %
Wario	19	8	19	9	13	7	23 %

Supervised approach: I'm looking for the best driver!

2

I first seek to understand what are the factors explaining the success of a pilot

3

Thanks to a supervised type algorithm I try to predict what is the expected rate of victory

Character	Speed	Accel.	Weight	maneuverability	Reason.	Mini-Turbo	Victory rate
Iggy	15	10	15	13	15	9	???

Reinforcement learning: AI will learn to win races

Character	Speed	Accel.	Weight	maneuverability	Reason.	Mini-Turbo
Mario (IA)	12	12	12	15	15	10

- Mario (AI) participates in races, he can:
- Use the turbo
- Release banana peels

Reinforcement learning: 1st race

Mario (AI) uses the turbo from the first second.

He is quickly caught up.

Mario (AI) throws a banana, which is avoided by Daisy.

Mario (AI) receives a banana, which makes him lose places in the classification.

Mario (AI) finished 8th / 8th.

=> He lost

- Mario (AI) learns from:
-
- His tests
- The strategies of its competitors
- His final ranking (success / failure)

This race is a learning base

Reinforcement learning: 2nd run

Mario (AI) uses the turbo at the end.

He wins a place in the standings.

Mario (AI) avoids a banana.

He does not start.

Mario (AI) finished 7th / 8.

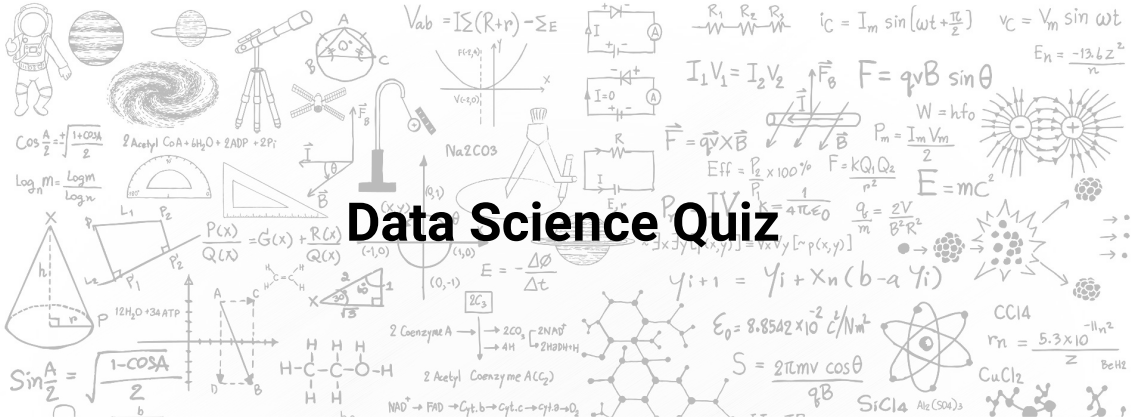
=> He lost
BUT less than the previous time.
This strategy is therefore better.

The goal of Mario (AI) is to improve his ranking.

Strategies that allow it are therefore valued.

AI improves with races

Quizizz Game : Data Science Quiz



Rules : Classic Multiple Answer Quiz about Data Science

Ready ? Go !

But what about actual data science algorithm ?

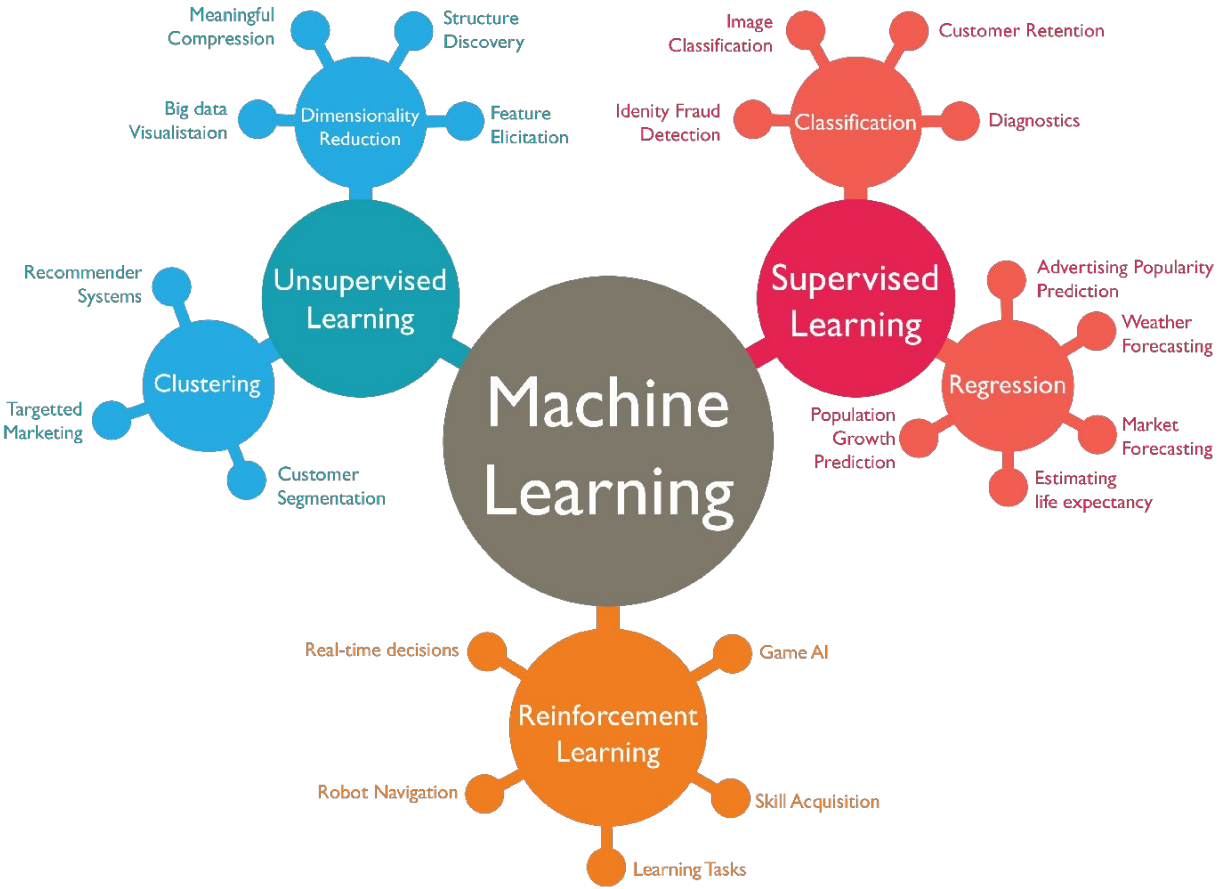
Zoom on the main algorithms of Machine Learning

The approaches	Supervised approach		Unsupervised approach				Enhancement
The main families of algo	Regression	Classification	Clustering	Dimension reduction	Anomaly detection	Association	
The main use cases	Sales prediction Forecasting Process optimization	Maintenance Predictive Churn Scoring	Customer base segmentation	Image compression	Fraud detection	Shopping Basket Analysis	Autonomous car

Reminder of the Mario Kart example

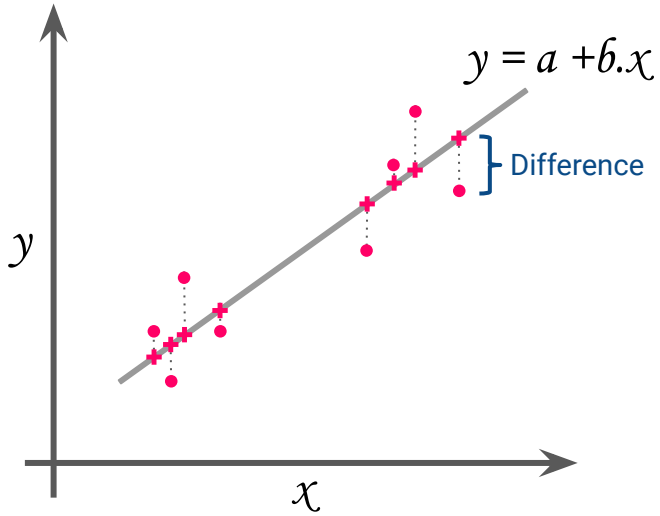
?

Let's dig into different algorithm



Linear Regression

Supervised | Regression



Pros

- Simple to implement mathematically and therefore fast prediction calculation
- Simple model to interpret

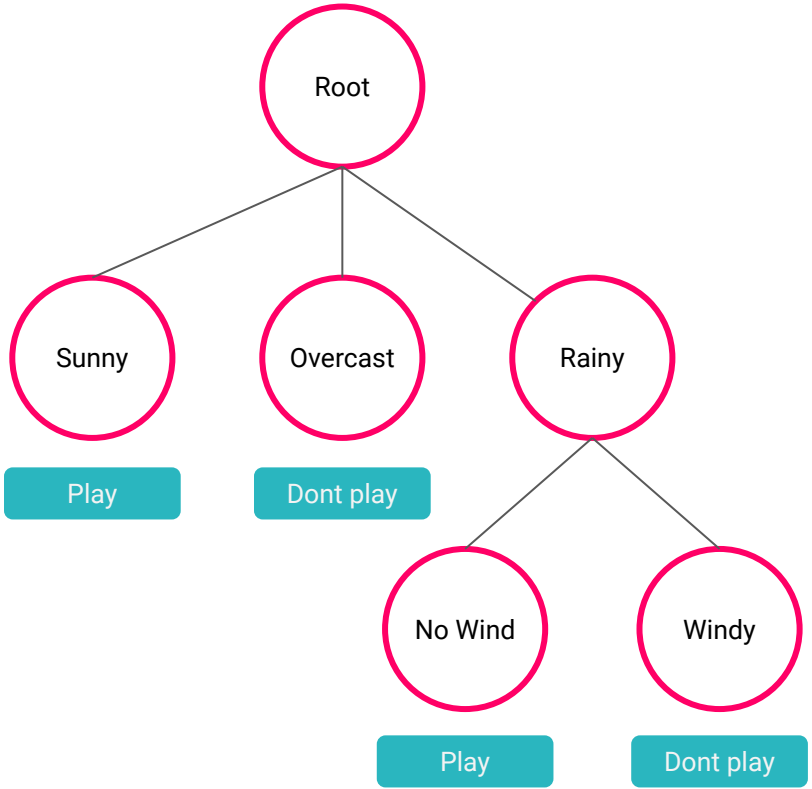
Cons

- Imposes a linear relationship
- Sensitive to outliers without regularization
- No relationship between predictor variables

Used to predict stock market prices, apartment rents, ...

Decision Tree

Supervised | Classification Regression



Pros

- Simple to present and intuitive (especially when the tree is not deep)
- Implicitly allows selection of explanatory variables
- Input variables can be quantitative or qualitative

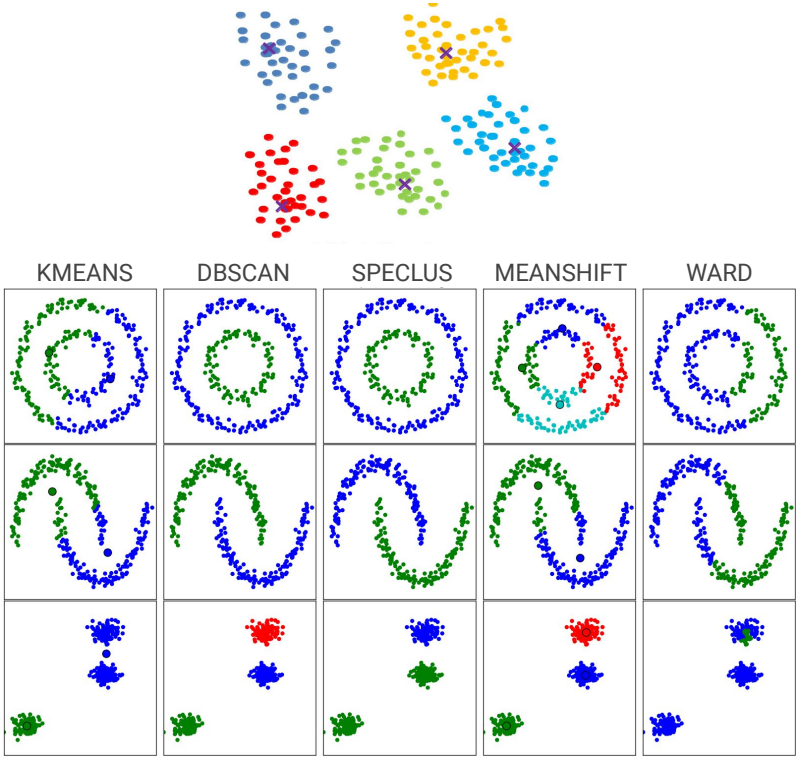
Cons

- Overfit if the tree is not trimmed
- First node of the capital tree and a wrong choice can lead to errors

Used for just about everything, classification, prediction and clustering.

K-Means

Unsupervised | Clustering



Pros

- Very quick to turn
- Implementable on large data volumes

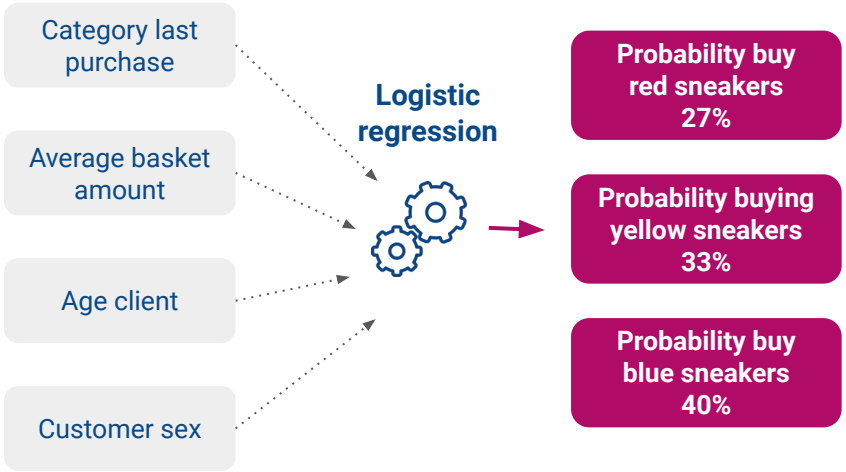
Cons

- Choice of K - number of class (take into account the natural partitioning)
- Sensitive to choice of initial centroids
- Trouble classifying complex shapes (½ moons)

Used for audience segmentation, text classification, and referral systems.

Logistic regression

Supervised | Classification



Logistic regression provides a probability of belonging to each class

Pros

- Accepts all explanatory variables (quantitative and qualitative)
- Easy to interpret

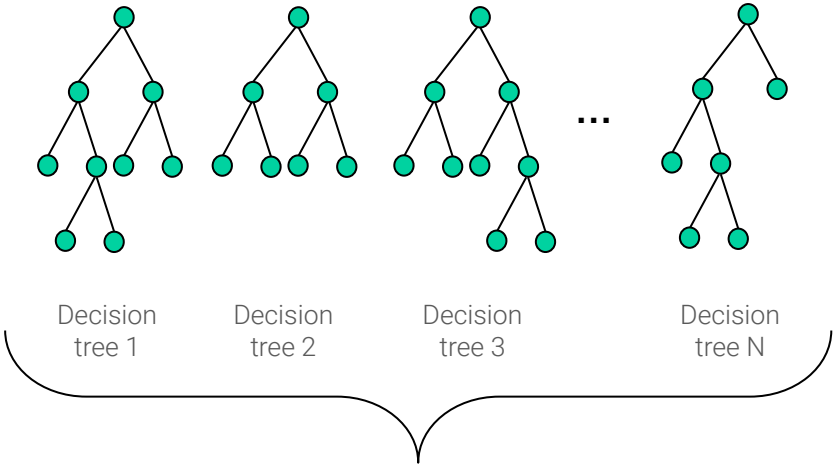
Cons

- Sensitive to correlated variables
- Do not process missing values

Used to predict a qualitative variable

Random Forest

Supervised | Classification Regression



N estimates
Final estimate = average of N estimates

Pros

- High and robust performance
- Generalize the results well to a new data sample
- Provides the importance of the variables
- Takes all types of explanatory variables

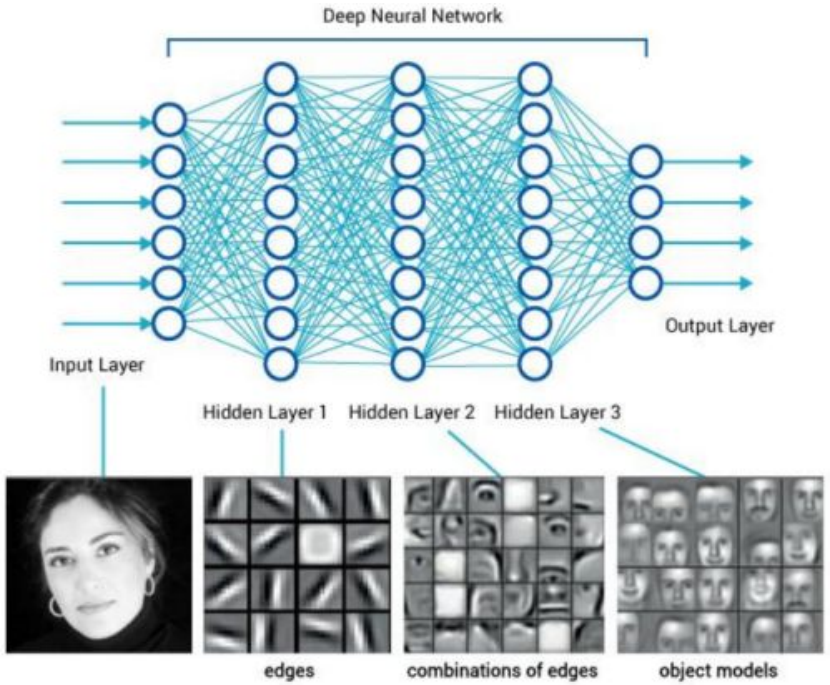
Cons

- Black box effect
- Requires high computing power

Used to predicts a qualitative or quantitative variable.

Neural Network

Supervised Unsupervised | Classification Regression



Pros

- Performs well on complex data (image, sound)

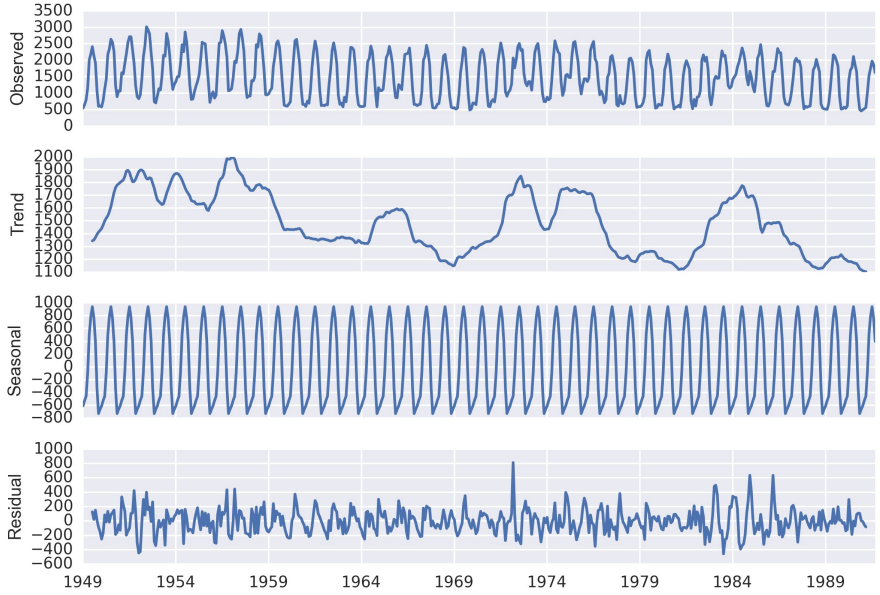
Cons

- Currently requires lots of data
- Requires high computing power
- Difficult to optimize
- Black box effect, harder to understand

Used to predicts a variable or detects patterns

Time series

Supervised | Regression



Pros

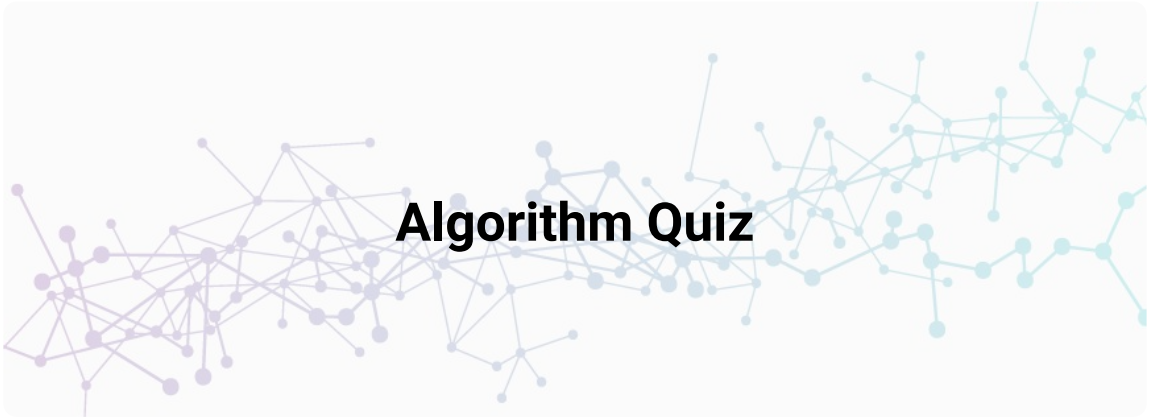
- Enter seasonality
- Do not need explanatory variables
- Deal naturally with time series data

Cons

- Requires a lot of historical data
- Requires complex models to include exogenous variables

Used to predicts future values of a quantitative variable

Quizizz Game : Algorithm Quiz



Rules : Guess which algorithm is the best fit for the described problem.

Ready ? Go !

Agenda

1 Data Science Knowledge

2 **Data preparation and modelling approach**

Why do we need to prepare the data ?

A wide variety of data format and issues

In-house data

Data already collected and available in the company.

E.g: CRM Customer data

Usual Format : **Tables**

Scrapping data

Data extracted using a computer program.

E.g: Website Scrapping

Usual Format : **Raw HTML**

Open data

Data that anyone can access, use and share.

E.g: Gov. Demographic data

Usual Format : **Tables**

Log data

Data passively generated by the running of an application.

E.g: Website connection log

Usual Format : **Raw Text**

APIs data

Data collected from another app, using HTTP request.

E.g: Weibo API

Usual Format : **JSON**

Self-collected data

Non-findable data collected to respond to specific question.

E.g: Survey

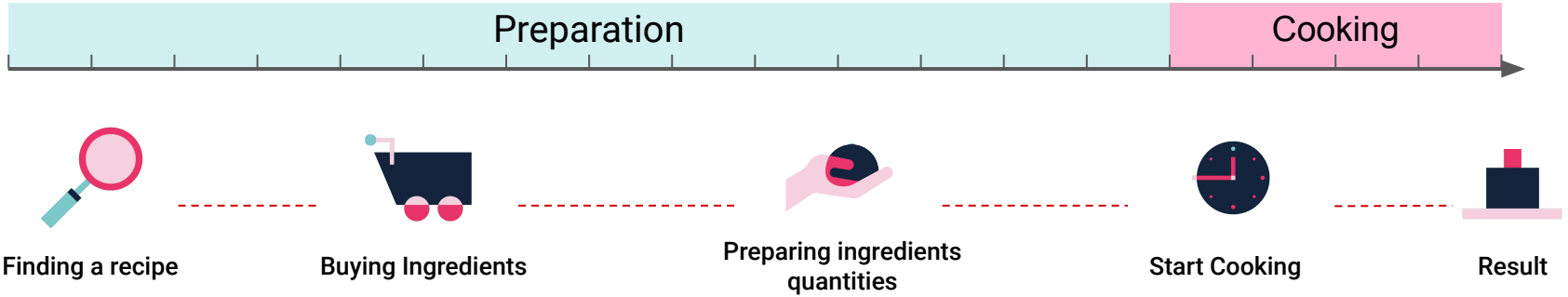
Usual Format : **Anything**

Common Problems

- Missing values
- Misspelled text
- Numbers as text
- Wrong date format
- Outliers
- Duplicates
- Accessibility
- Incorrect data
- Mislabeled data
- Illegal data ...

The preparation stage takes up 80% of the data scientist's work time

Cooking a cake ...



... is like Building a model

The goal of the data preparation : store example

*A data lake is a central storage repository that holds big data from many sources in a raw, granular format.

How to prepare data ?

Visualise the data

By simply observing the tables

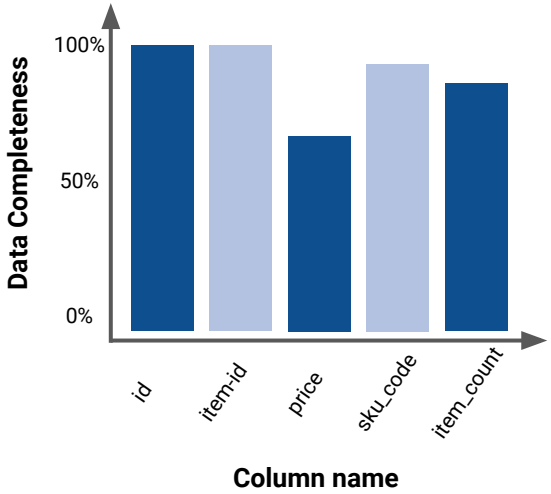
with the df.head() command

id	item_id	price	sku_code	item_count
1001	3001	54.99	AL34DE	122
1002	3002	23.99	DD99LQ	18
1003	3003		DD45KP	ERR
1004	3004	0	GH23LO	939
1005	3005	10.00	EW01EW	12

we observe some missing value for the column "price"

With global charts

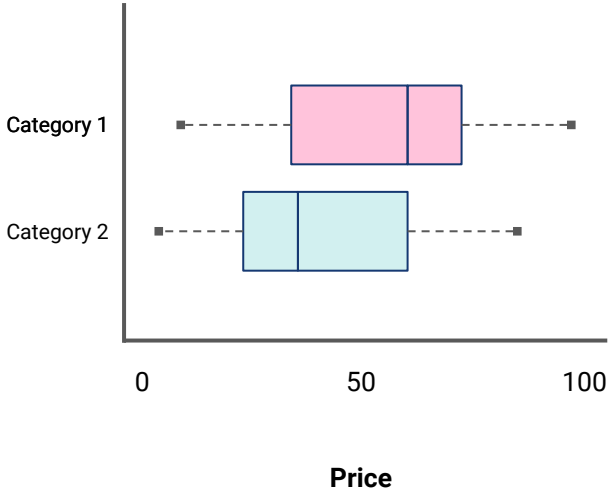
such as a data completeness chart



this observation is confirmed, around 30% of our item price are missing

With more specific charts

such as boxplots



no outliers detected, price are incomplete but looks clean

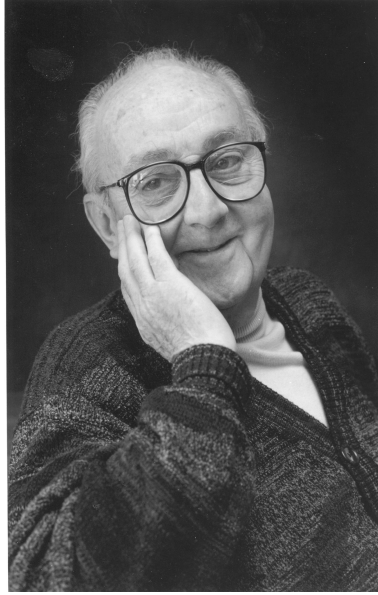
Find and apply the right solution

Practical Exercise

students_performance_data_preparation.ipynb

What is a model ?

Let’s start with a quote



“All models are wrong, but some are useful”

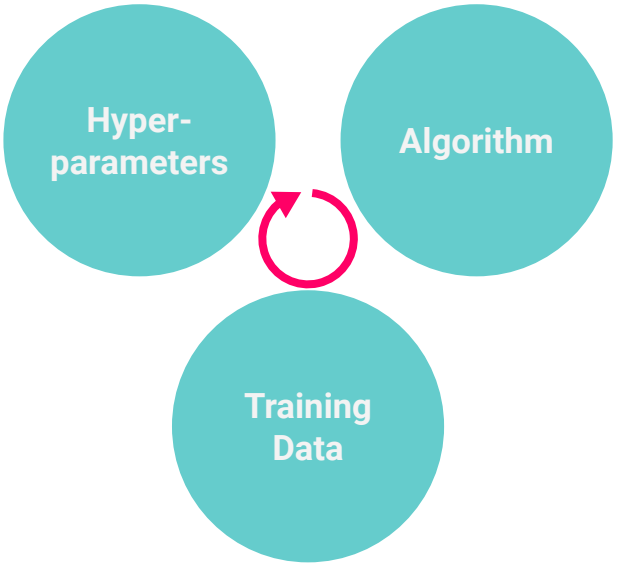
Georges BOX,

English statistician who has contributed to the fields of quality control, time series and Bayesian inference

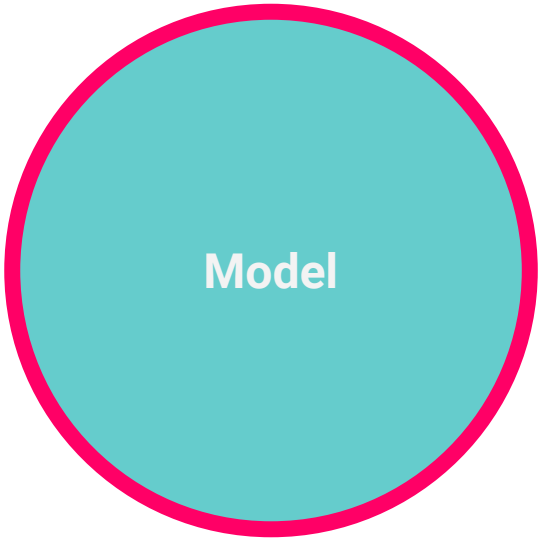
What is exactly a model ?

A **machine learning model** is a function saved as a file that has been **trained** to recognize certain types of **patterns**. You **train** a model over a **set of data**, providing it an **algorithm** that it uses to reason over and **learn from that data**.

Choose an algorithm and train it using specific parameters ...

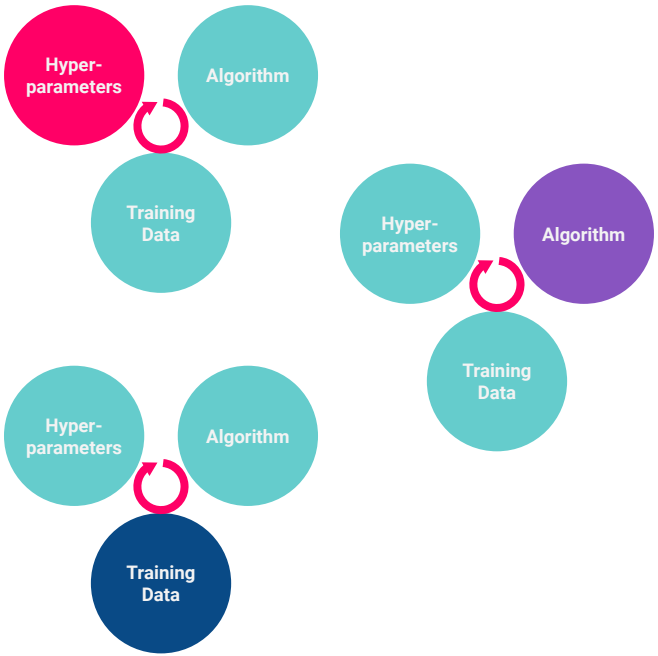


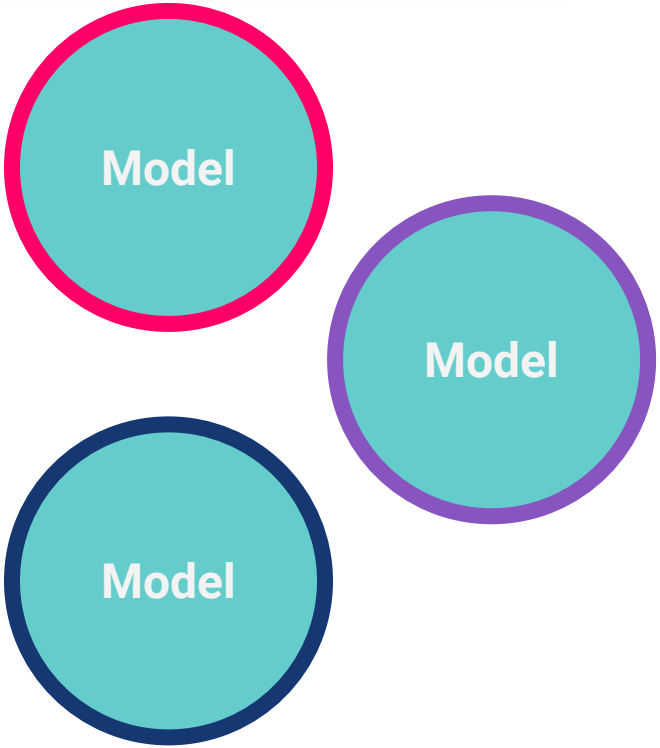
... and you obtain an unique model !



An infinite number of models

Change any ingredient of the recipe, and you get a whole **new model**.
This fact lead to an **infinite number** of potential models.

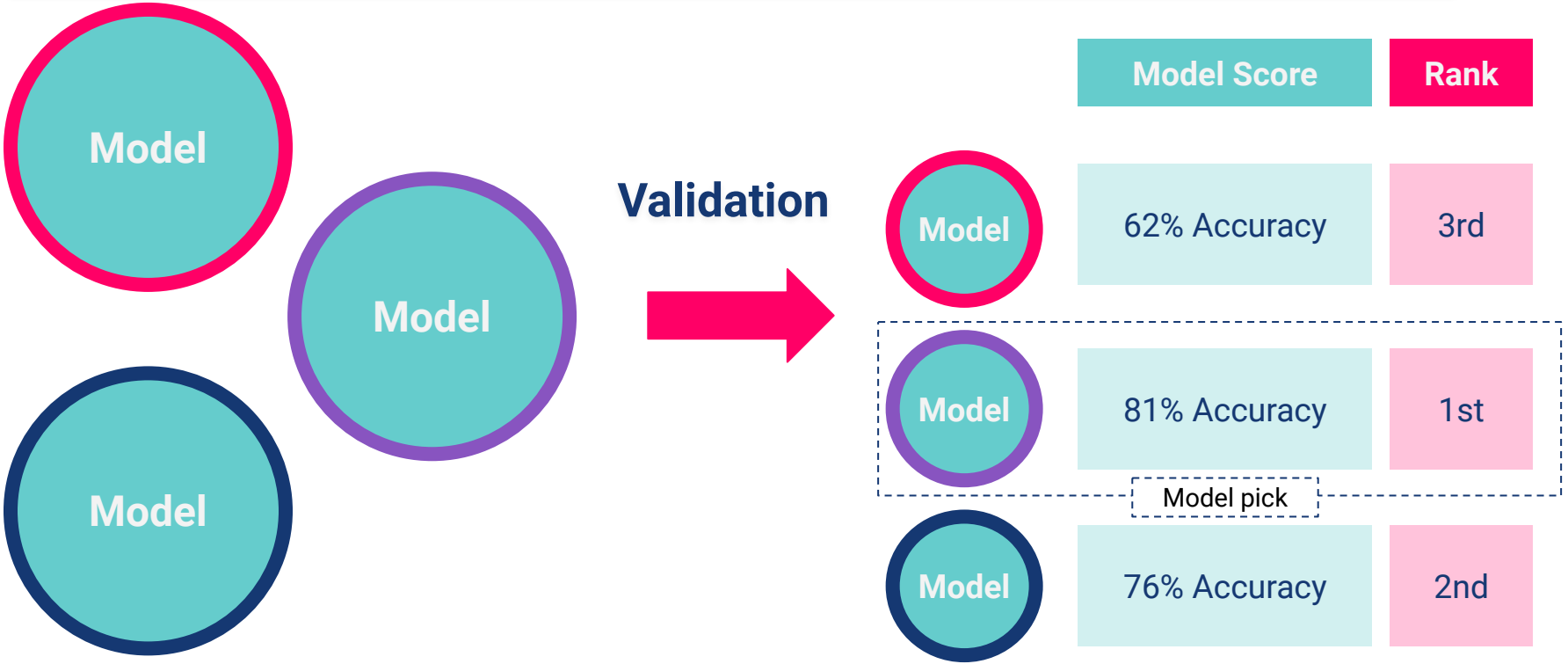




Then how to create the right model ?

Model Validation

In order to choose the **best model**, we need to obtain an indicator on the model's performance. This indicator is obtained during the **model validation** phase.



Example : Mario Kart Validation System

Character	Speed	Accel.	Weight	Victory rate
Baby Peach	9	13	9	32 %
Bowser Jr	13	12	12	27 %
Daisy	13	11	13	12 %
Donkey Kong	17	9	17	31 %
Mario	15	10	15	40 %
Roi Boo	17	8	19	7 %
Toad	11	12	11	18 %
Wario	19	8	19	23 %

Sample of learning

Character	Speed	Accel.	Weight	Victory rate
Baby Peach	9	13	9	32 %
Bowser Jr	13	12	12	27 %
Daisy	13	11	13	12 %
Donkey Kong	17	9	17	31 %
Mario	15	10	15	40 %

The training sample is used to train the model.

This sample contains all the explanatory variables + the target variable

Validation sample

Character	Speed	Accel.	Weight	Variable to predict	Victory rate
Roi Boo	17	8	19	?? %	7 %
Toad	11	12	11	?? %	18 %
Wario	19	8	19	?? %	23 %

Estimate by model

Actual rates

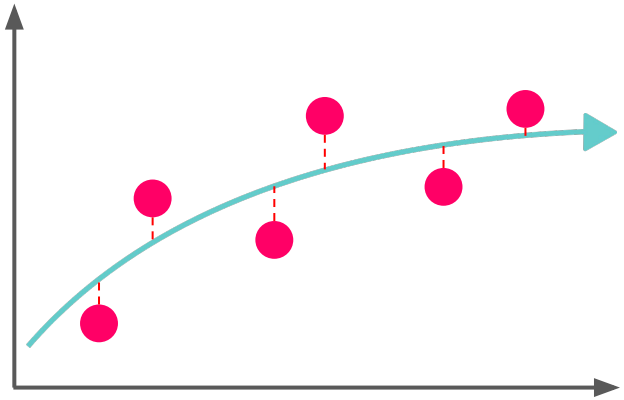
The validation sample is used to measure the quality of the model.

The target of the validation sample (victory rate) is initially hidden. It is then compared to the target calculated by the algorithm.

Some validation methods : RSS and AUC

Residual Sum of Squares (RSS)

RSS is a measure of the gap between the **data** and a **regression** **model**. A small RSS indicates a tight fit of the **model** to the **data**.

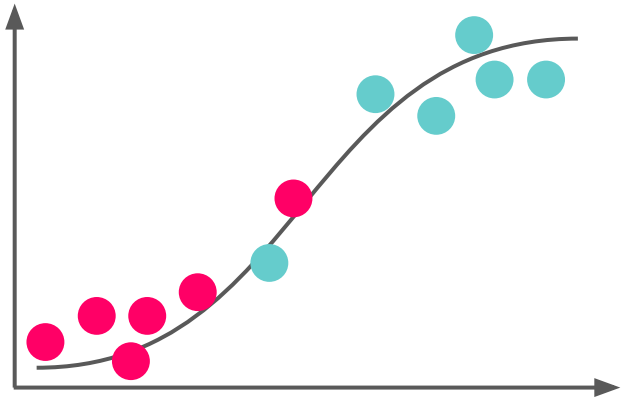


$$\text{Sum of Squares} = \frac{\text{distance}^2(-- + -- + --- + --- + -- + -)}{\text{number}(\bullet+\bullet+\bullet+\bullet+\bullet+\bullet)}$$

$$\text{Sum of Squares} = \frac{2^2 + 2^2 + 3^2 + 3^2 + 2^2 + 1^2}{6} \approx 5.17$$

Area Under the Curve (AUC)

AUC measure the probability that the **classification** model ranks a random **positive** example more highly than a random **negative** example.



5 **positive** have 100% probability of being ranked more highly than a **negative** and 1 **positive** have $\frac{5}{6}$ = 83% probability to be ranked more highly than a **negative**.

$$\text{Area Under the Curve} = \frac{1 + 1 + 1 + 1 + 1 + 0.83}{6} \approx 0.97$$

Depending on the objective function the evaluation metric should be picked wisely

REGRESSION

Mean Absolute Error

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root Mean Squared Logarithmic Error

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} log\left(\frac{y_i + 1}{\hat{y}_i + 1}\right)^2}$$

CLASSIFICATION

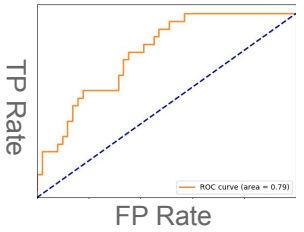
Accuracy

Number of correct predictions/ Number of predictions
(also *Recall*, *Precision* and *F1-Score*)

Logarithm loss

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

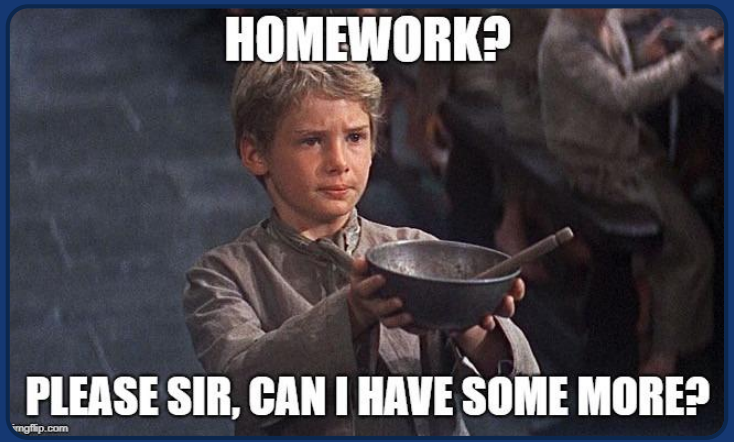
Area under ROC curve



Practical Exercise

students_performance_build_validate_model.ipynb

Homework time !



Thank you