La finance quantitative : à la recherche de la pierre philosophale

Dorian Lagadec

23 mars 2021

Sommaire

Problématique

Disclaimer : Cours d'ouverture et d'introduction à la fois

Une fois qu'on a bâti une compréhension d'un marché (actions, obligations, taux, devises, ...), peut-on utiliser cette expertise pour générer de la performance ?

- Si non : à quoi servent les *asset managers* ?
- Si oui : comment faire ?

Construction d'un processus d'investissement systématique

- Identifier des signaux forts (données *leading*, signaux techniques, règles d'investissement, ...)
- Modéliser la stratégie pour couvrir tous les cas possibles (étape complexe), souvent selon une modélisation paramétrique (seuils entre plusieurs états, paramètres de constructions de signaux, ...)
- On aboutit à un **ensemble de stratégies** qui correspondent toutes à la méta-stratégie initiale
- **Comment choisir ? Le stratégie est-elle bonne ?**

Qu'est-ce qu'un *backtest*

"Le backtesting (France) ou test rétro-actif de validité (Canada) consiste à tester la pertinence d'une modélisation ou d'une stratégie en s'appuyant sur un large ensemble de données historiques réelles. [...] En finance, il permet de vérifier la validité et la rentabilité d'une stratégie d'investissement." (Wikipedia)

Concept séduisant :

- Validation statistique et historique ;
- Test immédiat n'impliquant aucun risque en capital ni en temps ;
- Possibilité d'en dériver des métriques (rendement, Sharpe, drawdown, etc.) permettant de comparer des stratégies entre elles, de discriminer, ...

Backtesting fraud 1/4

Pseudo-Mathematics and Financial Charlatanism : The Effects of Backtest Overfitting on Out-of-Sample Performance, Bailey, Borwein, Lopez de Prado, Jim Zhu [1]. *Les bullets points ci-dessous sont des slides de De Prado*

- **Survivorship bias** : Using as investment universe the current one, hence ignoring that some companies went bankrupt and securities were delisted along the way.
- **Look-ahead bias** : Using information that was not public at the moment the simulated decision would have been made. Be certain about the timestamp for each data point. Take into account release dates, distribution delays, and backfill corrections.
- **Storytelling** : Making up a story ex-post to justify some random pattern.

Backtesting fraud 2/4

- **Data mining and data snooping** : Training the model on the testing set.
- **Transaction costs** : Simulating transaction costs is hard because the only way to be certain about that cost would have been to interact with the trading book (i.e., to do the actual trade).
- **Outliers** : Basing a strategy on a few extreme outcomes that may never happen again as observed in the past.
- **Shorting** : Taking a short position on cash products requires finding a lender. The cost of lending and the amount available is generally unknown, and depends on relations, inventory, relative demand, etc.

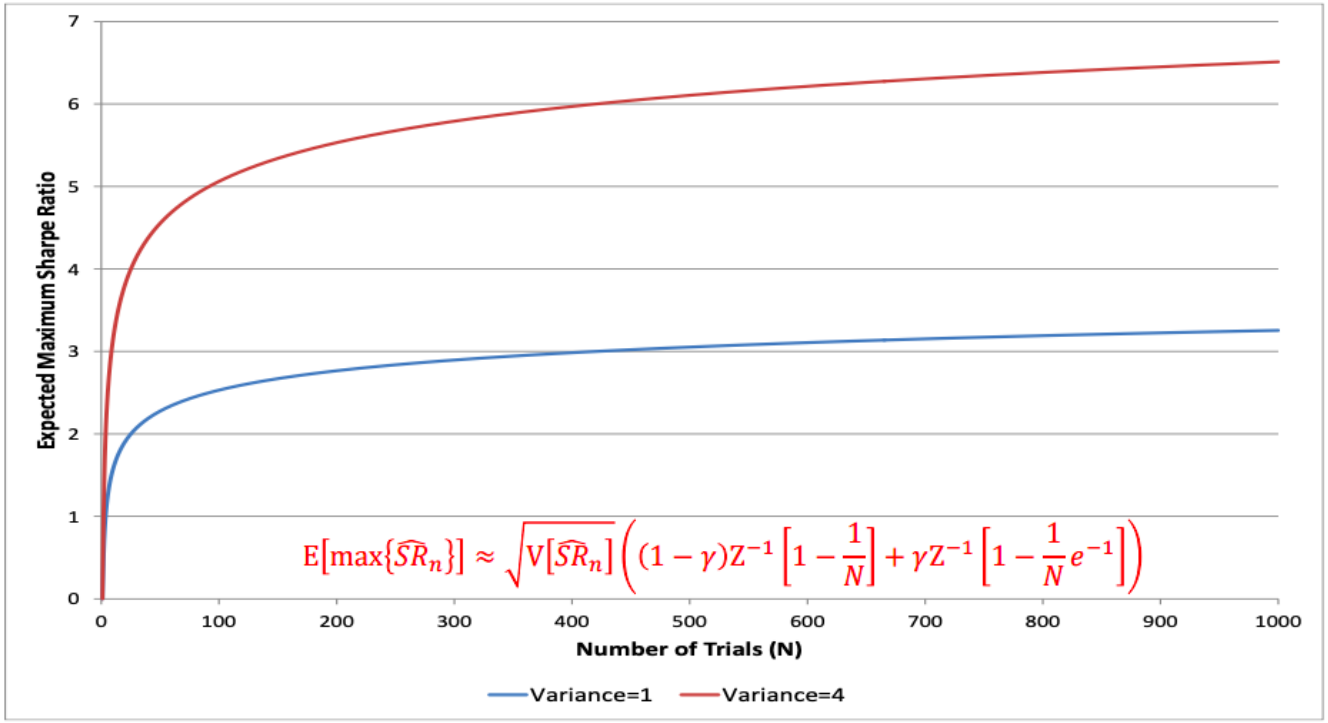
Backtesting fraud 3/4

Même en étant très pointilleux, la persévérance mène tout droit à l'*overfitting*.

Fama, dans une interview récente : "Chance alone can generate big winners when there are a lot of people playing the game. For somebody to throw 15 heads in a row, it's very low probability to identify, say, "That person will throw 15 heads in a row." But if you say, "Yeah, I looked at 25,000 investors, and this guy threw 10 heads in a row," well, that's not so unlikely anymore, because with that many throws of the die, somebody is going to come up and do it, strictly by chance."

Backtesting fraud 4/4

Deflated Sharpe Ratio, Minimum Backtest Length, ...



Expected Maximum Sharpe Ratio as the number of independent trials ***N*** grows, for $\mathrm{E}[\widehat{SR}_n] = 0$ and $\mathrm{V}[\widehat{SR}_n] \in \{1,4\}$.

Data Dredging: Searching for empirical findings regardless of their theoretical basis is likely to magnify the problem, as $\mathrm{V}[\widehat{SR}_n]$ will increase when unrestrained by theory.

We must always control for *N* and $\mathrm{V}[\widehat{SR}_n]$.

This is a consequence of pure random behavior. We will find high SR strategies *even if there is no investment skill associated with this strategy class* ($\mathrm{E}[\widehat{SR}_n] = 0$).

Conclusion intermédiaire

- Malgré le caractère séduisant d'une méthode d'investissement systématique, automatique, infaillible, le chemin est semé d'embûches ;
- Le mésusage des mathématiques sous-jacentes à l'inférence statistiques conduit à une impasse au mieux, à une fraude au pire ;
- Comment remédier à tout ça ?

Machine Learning : "rappels"

Dan Ariely

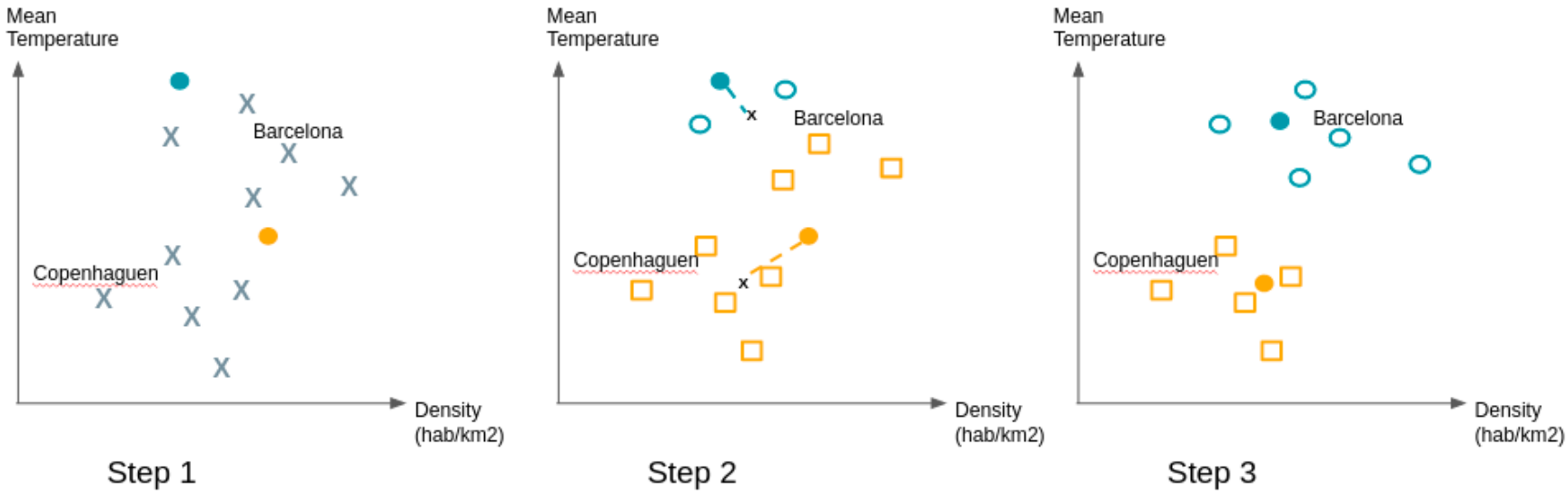
6. Januar 2013 ·

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

2451 „Gefällt mir"-Angaben 115 Kommentare 1176 geteilte Inhalte

- Big Data : Volume, Variété, Vélocité
- Machine Learning : l'apprentissage machine
- Apprentissage supervisé vs Apprentissage non supervisé
- Deep Learning, Online Learning, Reinforcement Learning, Transfer Learning, ...
- Intelligence artificielle

Clustering - exemple du K-means :

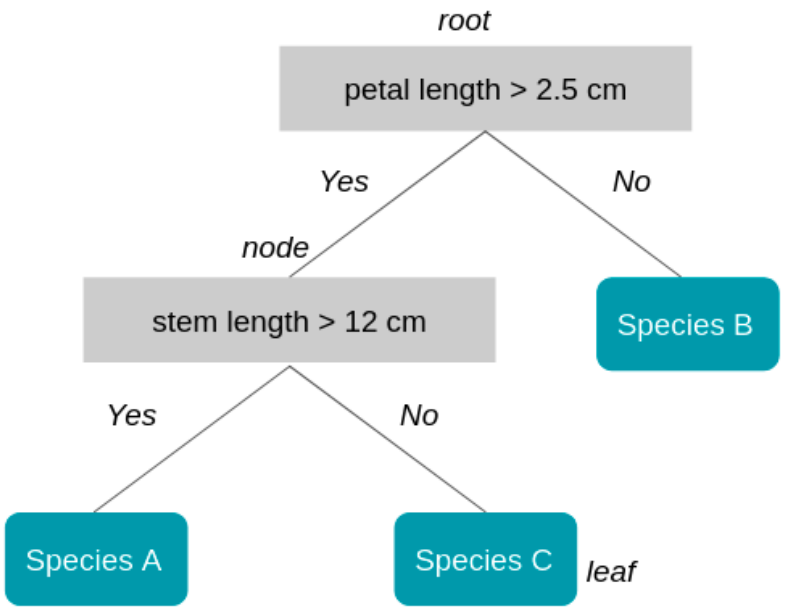


Video : ici : 5:48

Arbre de décision :

Observation	Petal length	Stem length	Species
1	2.7 cm	20 cm	A
2	2.6 cm	12 cm	C
3	2.1 cm	21 cm	B
4	1.9 cm	20 cm	B

Data

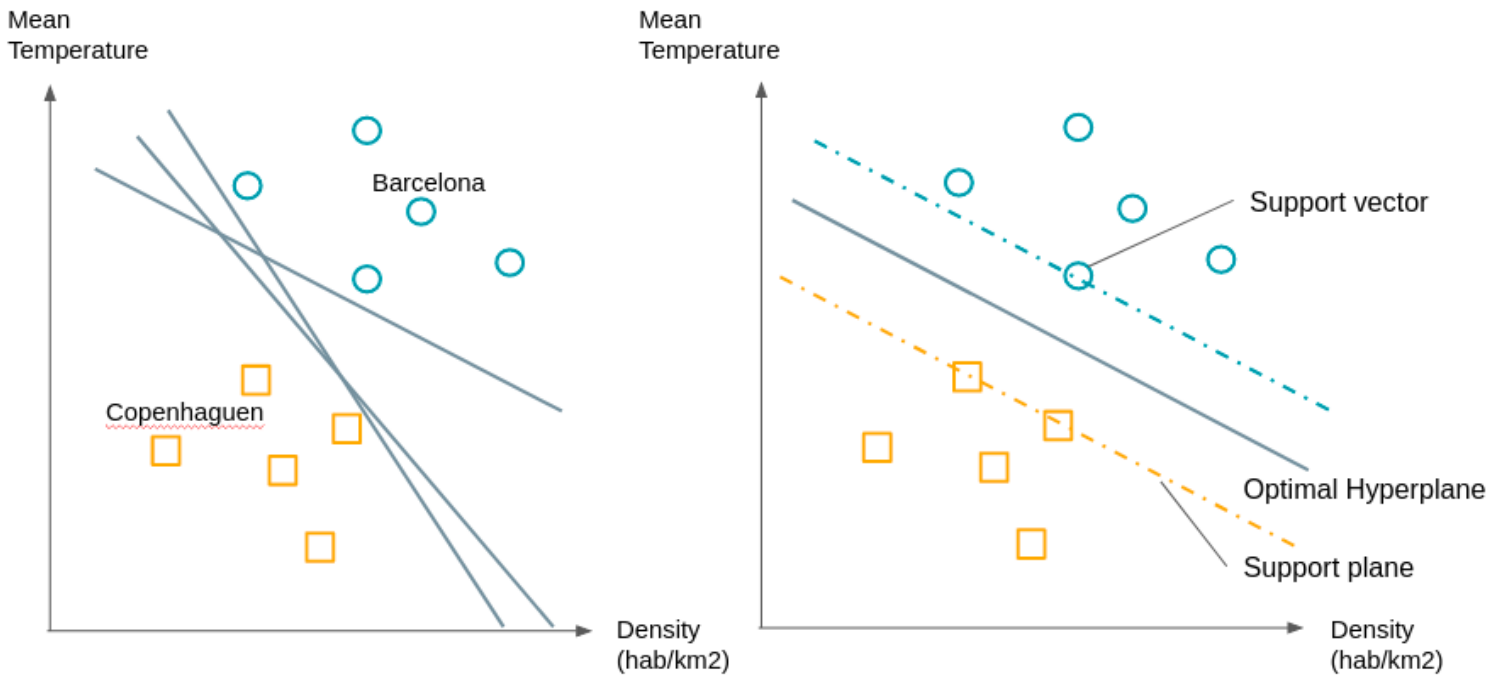


Quelques méthodes de Machine Learning

Forêt aléatoire - bagging - et XGBoost :

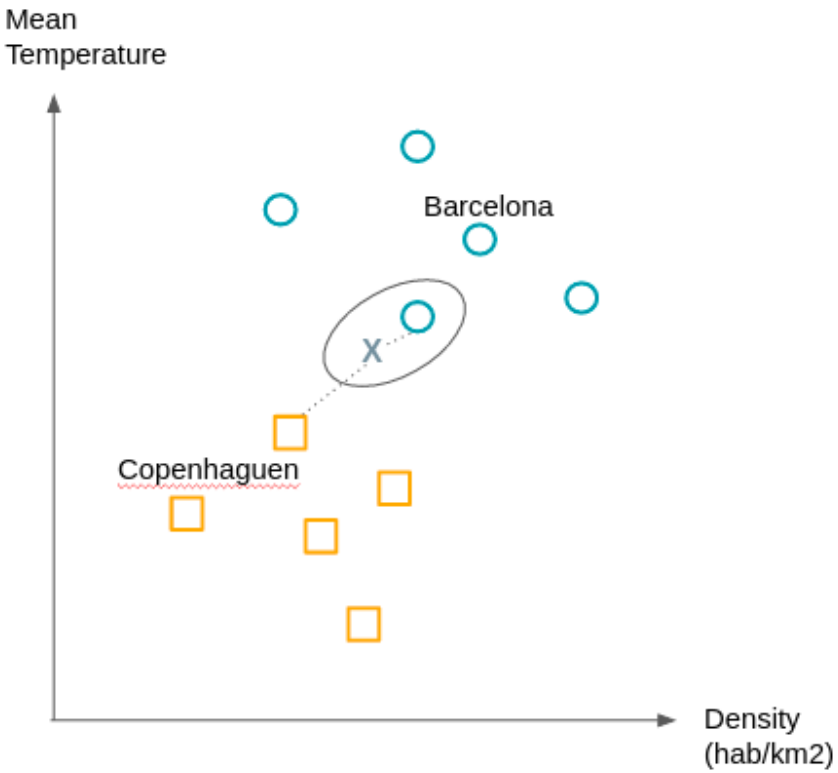
Age	Prediction Tree 1	Difference	Prediction Tree 2
13	19	-6	15
68	55	+13	63

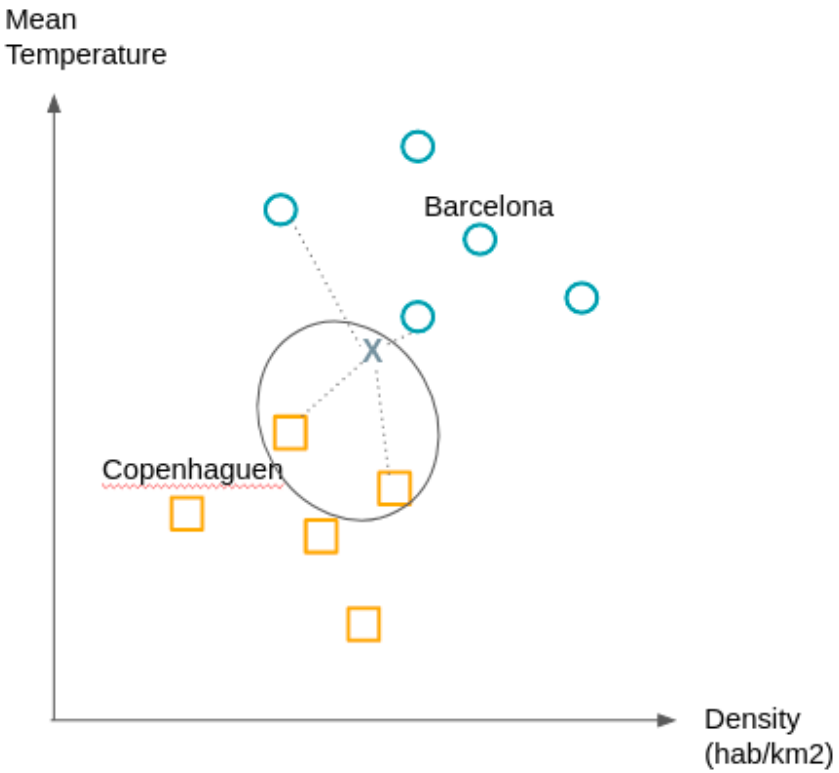
Machines à vecteurs de support :



Quelques méthodes de Machine Learning

KNN - K plus proches voisins :





Importance des variables et explicabilité

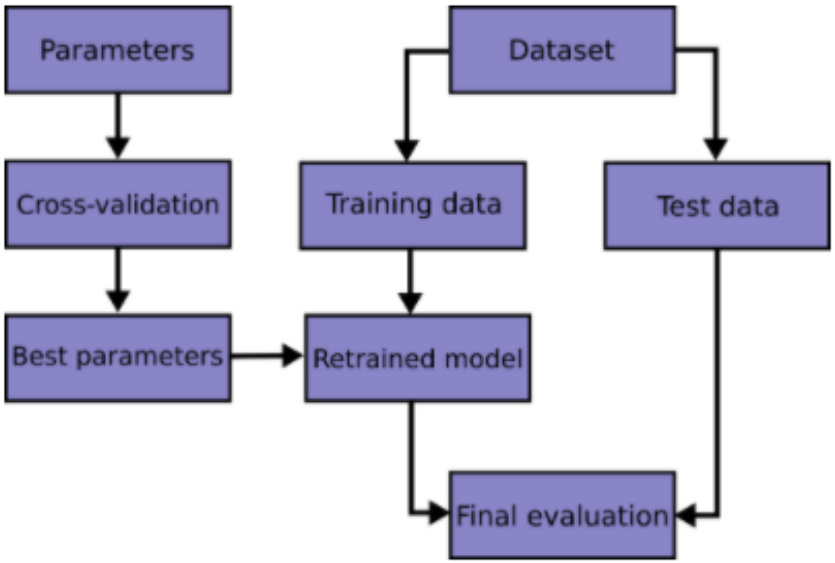
Dans une régression linéaire, l'importance des variables est extrêmement simple à calculer, et le modèle est transparent à l'utilisateur :

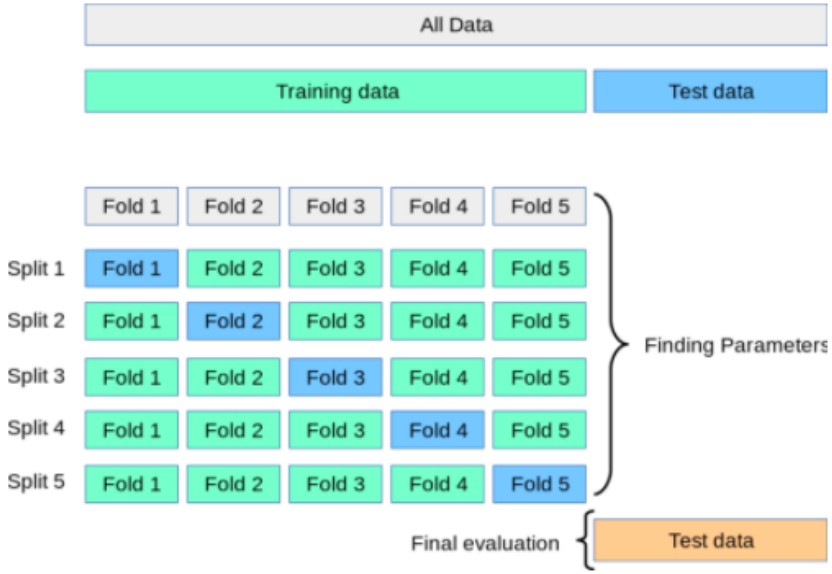
$$y = 1.2x_0 - 0.3x_1 + 0.02x_2$$

Ce n'est pas le cas de la plupart des algorithmes de machine-learning, aux variables plus intriquées et aux prédicteurs plus complexes et hautement non-linéaires.

- Feature Importance
- Shapley

Validation croisée





Mesures de performance des modèles : RMSE, Précision, Rappel, F1 Score, ...

Peut-on appliquer le ML ? La question des données financières

- Principalement des séries temporelles, qui nécessitent un traitement particulier (différenciation, désaisonnalisation, modélisation avancée sur l'excitation des variables, ...)
- Ratio *signal-to-noise* faible : la robustesse des prédictions sera toujours un problème
- Quasi colinéarité d'un ensemble de séries. Les données ne manquent pas, au contraire.

La plupart des algorithmes de ML ayant des hypothèses fortes sur les variables aléatoires modélisées, un travail important de *preprocessing* est important si l'on veut des résultats exploitables.

Exemple : différenciation fractionnée.

Où peut-on l'appliquer ?

Compréhension des données : Feature importance

Compréhension des données : Clustering

Modélisation d'une stratégie : Algorithmes de ML.
Prédiction d'un rendement d'une classe d'actifs, prédiction des actions qui vont surperformer le marché, prédiction d'un label "bull", "bear" sur un marché, ... de manière générale prédiction d'un événement de marché dont on espère tirer profit.

Comment peut-on l'appliquer ?

Validation croisée appliquée à la finance

Walk-Forward :

Avantages :

- Interprétation historique claire qui s'apparente à du *paper trading* ;
- Le test set est, par construction, *out-of-sample*

Désavantages :

- Un seul scénario testé ;
- Un biais peut exister dans l'ensemble du jeu de données (exemple : taux bas) ;
- Les premières décisions sont prises avec un échantillon très réduit

Comment peut-on l'appliquer ?

Validation croisée standard :

Avantages :

- Teste de nombreux scénarios de manière isolée
- Chaque décision est prise avec un nombre égal d'inputs, ce qui rend les résultats comparables

Désavantages :

- Un seul scénario testé ;
- Pas d'interprétation historique claire car le modèle peut prendre à date t des décisions en connaissance de $t+1$
- Les premières décisions sont prises avec un échantillon très réduit

Apprendre sur des données synthétiques

Autres cas d'usage

- Voir de nombreux exemples dans ce lien

QUESTIONS


David H Bailey et al. "Pseudo-mathematics and financial charlatanism : The effects of backtest overfitting on out-of-sample performance". In : *Notices of the American Mathematical Society* 61.5 (2014), p. 458-471.


Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.


Marcos M López de Prado. *Machine learning for asset managers*. Cambridge University Press, 2020.


Gaël Bonnardot. *8 Algorithmes de Machine Learning expliqués en Language Humain*. https://datakeen.co/8-machine-learning-algorithms-explained-in-human-language/. [Online ; accessed 21-March-2021]. 2017.


Scikit-Learn. *Cross-validation*. https://scikit-learn.org/stable/modules/cross_validation.html. [Online ; accessed 21-March-2021].