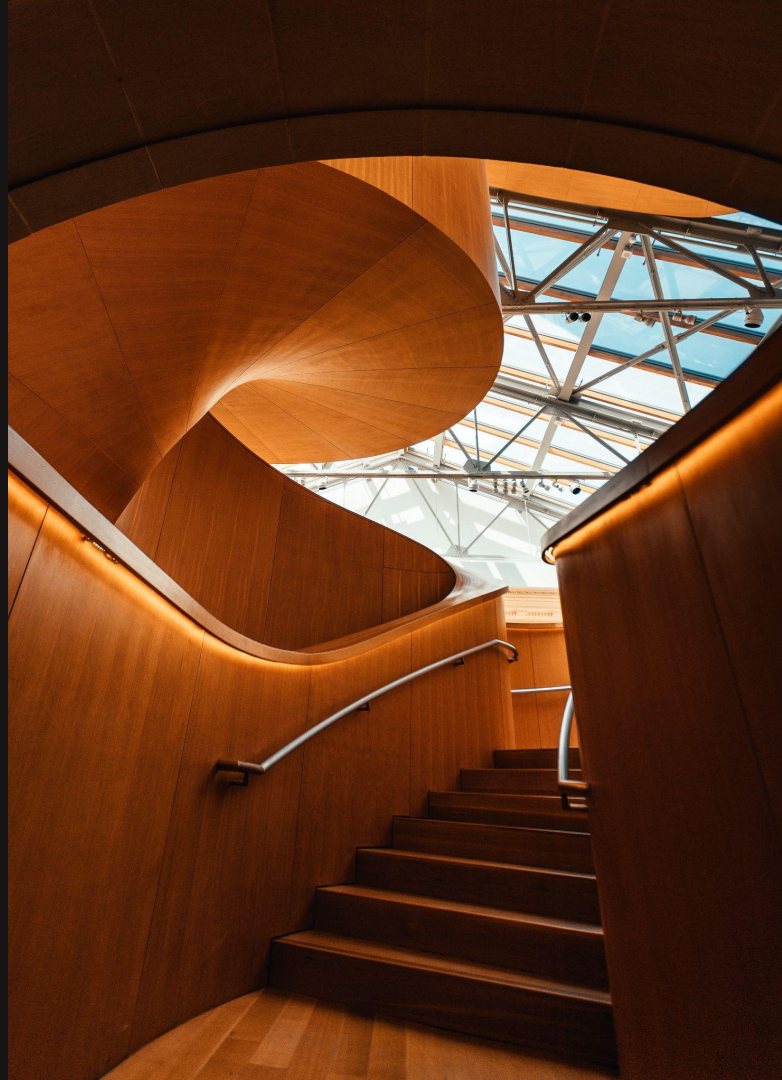
SEQUENTIAL BOOTSTRAPPING IN FINANCE:

APPROACHING THE TRUE IID SAMPLING

HUDSON AND THAMES



ABOUT ME:
VALERIIA PERVUSHYNA

Quantitative Researcher at Hudson & Thames

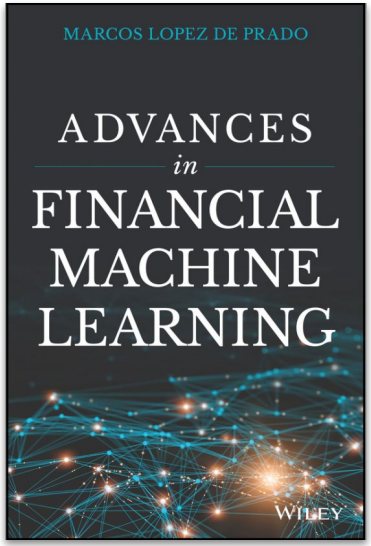
M.Sc. in Quantitative Finance at the University of Warsaw

Twitter: @VPervushyna

Git: @vlrie

CONTENTS

THE KEY INSPIRATION FOR THIS LECTURE



Advances in Financial Machine Learning, Chapter 4

- *Marcos López de Prado 2018*

Professor Marcos López de Prado

- *He has helped modernize finance for the past 20 years, by advancing the adoption of machine learning and supercomputing*

DATA LIMITATIONS IN FINANCIAL MACHINE LEARNING

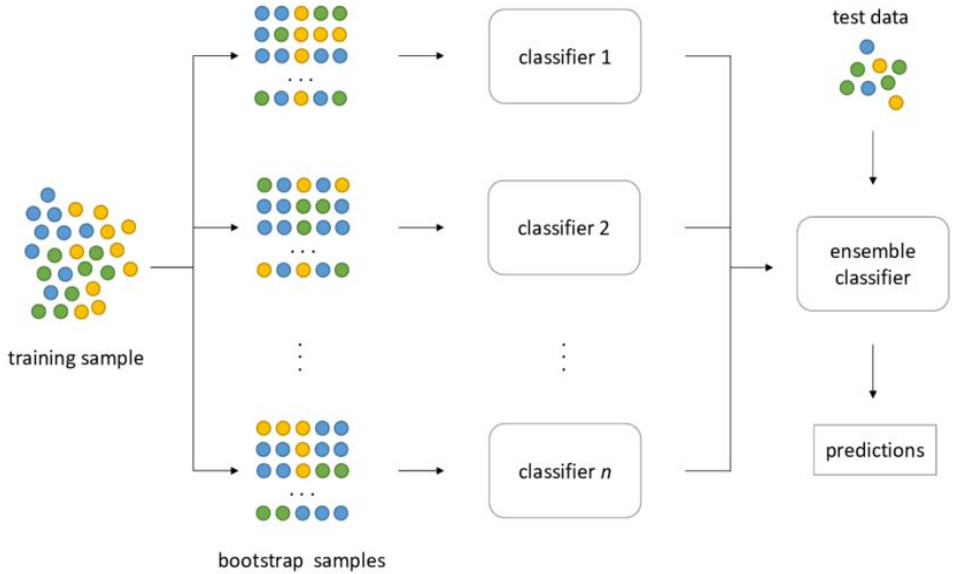
One of the biggest constraints for quantitative finance is how limited the historical financial data is.

Date	Open	High	Low	Close	Volume
2012-05-18	42.049999	45.000000	38.000000	38.230000	573576400
2012-05-21	36.529999	36.660000	33.000000	34.029999	168192700
2012-05-22	32.610001	33.590000	30.940001	31.000000	101786600
2012-05-23	31.370001	32.500000	31.360001	32.000000	73600000
2012-05-24	32.950001	33.209999	31.770000	33.029999	50237200
...	...	...	...	...	...
2021-03-31	289.989990	296.500000	288.609985	294.529999	19498200
2021-04-01	298.399994	302.399994	296.600006	298.660004	17590700
2021-04-05	300.890015	310.769989	300.679993	308.910004	28237000
2021-04-06	308.839996	311.350006	305.250000	306.260010	17335200
2021-04-07	306.339996	314.250000	305.500000	313.089996	22828800

BOOTSTRAPPING: THE SOLUTION TO SCARCITY

Bootstrapping allows to generate new samples from an already existing dataset.

"Under certain conditions such as large sample sizes, the sampling distribution will be approximately normal, and the standard deviation of the distribution will be equal to the standard error"



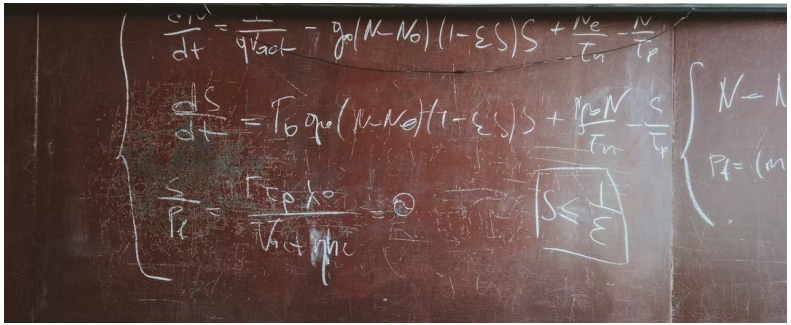
CORE POINT: ASSUMING THE IID

The phrase:

"Assuming our samples are independent and identically distributed"

prefaces almost every ML research. In fact, it is used so often that saying this almost became habitual for the practitioners.

IID assumption allows to create generalized algorithms and is crucial for such core theorems in data science as **central limit theorem** and **law of large numbers**. For face and voice recognition, spam classification, etc. assuming the iid will be a no-brainer, but for financial data, unfortunately, it is not the case.

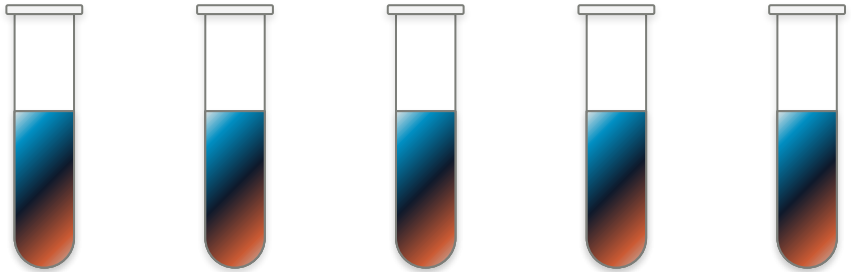


WHAT HAPPENS IF THE MAIN ASSUMPTION IS NOT TRUE?

"Finance is **not** a plug-and-play subject as it relates to ML applications"

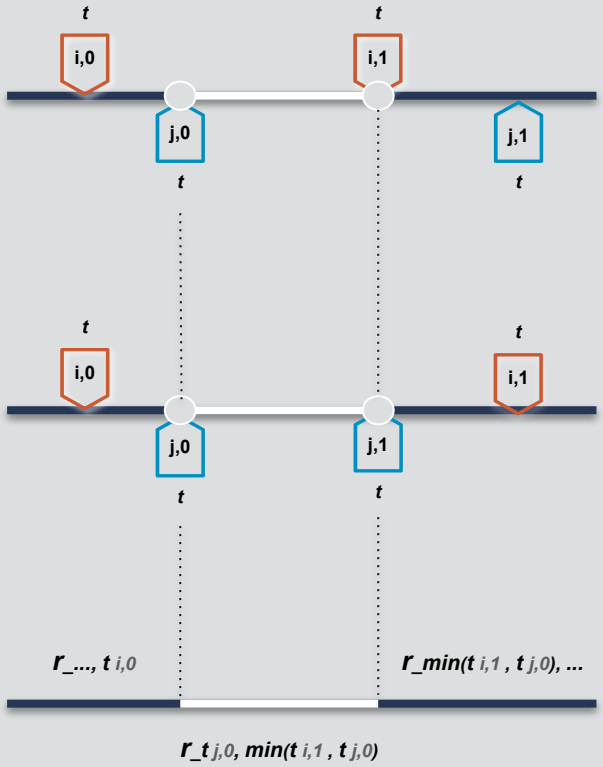
- Marcos Lopez de Prado

"… someone in your laboratory spills blood from each tube into the following none tubes to the right. … Now you need to determine the features predictive of high cholesterol (diet, exercise, age, etc.) without knowing for sure the cholesterol level of each patient. That is the equivalent challenge that we face in financial ML, with the additional handicap that the spillage pattern is non-deterministic and unknown."
- example described by prof Marcos Lopez de Prado



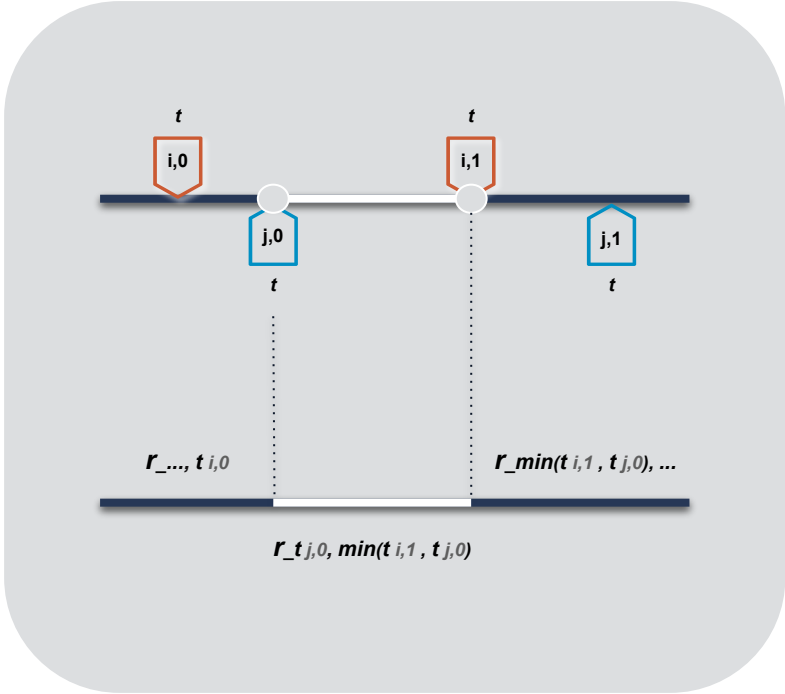
THE ISSUE OF THE OVERLAPPING OUTCOMES

Assume the outcome ***yi*** assigned to a feature ***Xi*** is a function of price bars that occurred over an interval ***[ti,0, ti,1]***. The series of labels, ***{yi}i=1,…,I***, are **not IID** whenever there is an overlap between any two consecutive outcomes, ***∃ i |ti,1 > ti+1,0***



THE CONCEPT OF CONCURRENCY

Two labels y_i and y_j are **concurrent** at t when both are a function of at least one common return. The opposite characteristic to concurrency is **uniqueness**.



THREE SOLUTIONS FOR LABEL UNIQUENESS

During the bootstrapping process, incorrectly assuming IID draws leads to oversampling, and it becomes increasingly likely that our in-bag and out-of-bag observations will be very similar, making the whole process redundant.

Possible solutions for the issue

1. Dropping overlapping outcomes before performing the bootstrap.
2. Utilizing the average uniqueness, to reduce the influence of outcomes that contain redundant information.
3. Performing a sequential bootstrap, where draws are made according to a changing probability, controlling for redundancy.

THE IDEA BEHIND SEQUENTIAL BOOTSTRAPPING

Our goal is to reduce the probability of drawing the observation with a highly overlapping outcome.

We achieve this by updating the probabilities for drawing particular values for each new draw, based on the uniqueness calculated using the sequence of previously made draws.

MATHEMATICAL DEFINITION

The key advantage of sequential bootstrap is that overlaps (even repetitions) are still possible, but decreasingly likely.

Notation:

I - number of items in the initial set

φ - the sequence of draws so far

$u_{t,i}$ - uniqueness of drawing *i* at time *t*

The Procedure

1. An observation ***Xi*** is drawn from a uniform distribution with the original probability of drawing any value ***i*** denoted as $\delta_i^{(1)} = I^{-1}$

2. To reduce the probability of drawing an observation ***Xj*** with a highly overlapping outcome we calculate the $u_{t,j}^{(2)} = 1_{t,j}\left(1 + \sum_{k \in \varphi^{(1)}} 1_{t,k}\right)^{-1}$

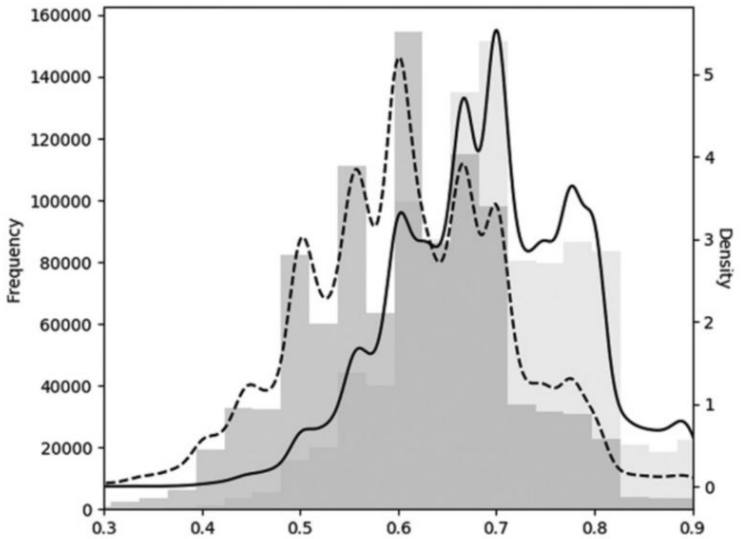
3. After that we calculate the average uniqueness of $u_{t,j}$ over ***j***'s lifespan $\bar{u}_j^{(2)} = \left(\sum_{t=1}^{T} u_{t,j}\right)\left(\sum_{t=1}^{T} 1_{t,j}\right)^{-1}$

4. A second draw then is made based on the updated probabilities $\delta_j^{(2)} = \bar{u}_j^{(2)}\left(\sum_{k=1}^{I} \bar{u}_k^{(2)}\right)$

5. The process is repeated until ***I*** draws have taken place.

EMPIRICAL RESULTS COMPARISON

Statistically speaking, samples from the sequential bootstrap method have an expected uniqueness that exceeds that of the standard bootstrap method, at any reasonable confidence level



Monte Carlo experiment of 1E6 iterations of standard vs. sequential bootstraps

Source: *Advances in Financial Machine Learning, Chapter 4 by Marcos Lopez de Prado.*

CONCLUSIONS

While there is no perfect cookie-cutter solution for the overlap problem in financial ML, the sequential bootstrap allows to efficiently and reliably increase the uniqueness of the labels providing a big improvement in the quality of the bootstrapping process.

REFERENCES

1. De Prado, M.L., 2018. Advances in financial machine learning. John Wiley & Sons.

2. De Prado, M.L., 2020. Machine learning for asset managers. Cambridge University Press.

BRING THE **POWER** TO YOUR TRADING

ML•FINLAB

BY HUDSON & THAMES

Use all the mentioned algorithms right away with MlFinlab

MlFinLab is a collection of production-ready algorithms (from the best journals and graduate-level textbooks), packed into a python library that enables portfolio managers and traders who want to leverage the power of machine learning by providing reproducible, interpretable, and easy to use tools.

THANK YOU

Does anyone have any questions?

@VPervushyna

@vlrie

www.linkedin.com/in/valeriia-pervushyna

HUDSON AND THAMES