Chapter2-Introduction

2023 年 2 月 24 日

1 Introduction

1.1 Context

Blossoming of machine learning:

- Data size sufficiently large
- Computational Power
- Economic Framing: non-linearity in asset pricing, etc.

1.2 Portfolio Construction: the Workflow

The baseline equation in supervised learning

$$\mathbf{y} = f(\mathbf{X}) + \varepsilon$$

is translated in financial terms as

$$\mathbf{r}_{t+1,n} = f(\mathbf{x}_{t,n}) + \varepsilon_{t+1,n}$$

where $f(\mathbf{x}_{t,n})$ can be viewed as the **expected return** for time $t+1$ computed at time t, i.e. $\mathbb{E}_t[r_{t+1,n}]$. Note that the model is common to all assets (f is not indexed by n), thus it shares similarity with panel approaches.

How to make accurate predictions?

- Gather better data, include some classical predictors
- The choice and engineering of inputs are important
- An integrated process

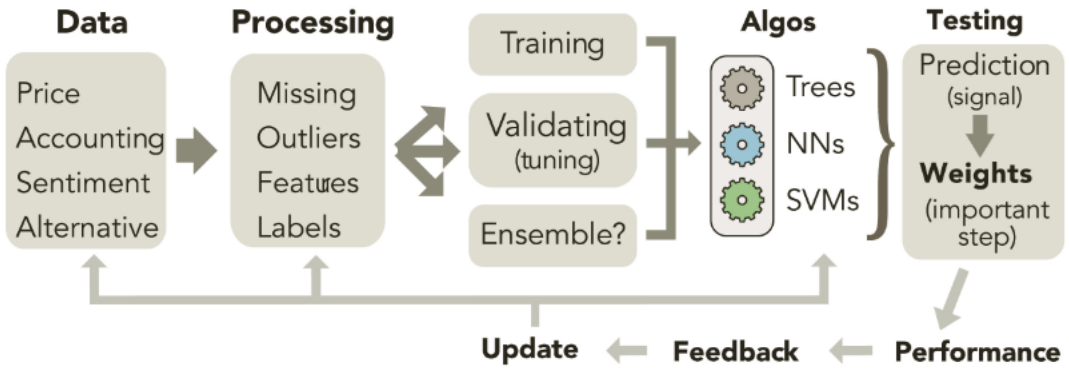

FIGURE 2.1: Simplified workflow in ML-based portfolio construction.

1.3 Machine Learning is No Magic Wand

In fact, heuristic guesses are often hard to beat. Below, we sum up some key points that we have learned through our exploratory journey in financial ML:

- **Causality** is key. If one is able to identify $X \to y$, where y are expected returns, then the problem is solved. Unfortunately, causality is incredibly hard to uncover.
- Thus, researchers have most of the time to make do with simple **correlation** patterns, which are far less informative and robust.
- Relatedly, financial datasets are extremely **noisy**. It is a daunting task to extract signals out of them. No-arbitrage reasonings imply that if a simple pattern yielded durable profits, it would mechanically and rapidly vanish.
- **Data is key**. The inputs given to the models are probably much more important than the choice of the model itself.
- **Persistent** series are more likely to unveil enduring patterns.
- What matters is to learn from those lapses.

Gathering and cleaning data, coding backtests, tuning ML models, testing weighting schemes, debugging, starting all over again: these are all absolutely indispensable steps and tasks that must be repeated indefinitely. There is no sustitute to experience.

Finally, this chapter emphasizes two key points:

- **Data is key.** Better data often saves your effort for tuning models. The choice of models is probably not as important as "inputs".
- **Practice makes perfect.** TRY MORE CASES!