The Case for Financial Machine Learning

Zichao Yang

Zhongnan University of Economics & Law

Date: March 12, 2025

Roadmap

In this chapter, authors try to justify why machine learning tools are suitable for financial studies.

What are the differences between ML and econometrics, between structural models and reduced form models?

- Prices are Predictions
- Information Sets are Large
- Function Forms are Ambiguous
- Machine Learning VS. Econometrics
- Challenges of Applying Machine Learning in Finance
- Two Cultures of Financial Economics

Prices

How do we define the price of an asset?

$$P_{i,t} = E[M_{t+1} X_{i,t+1} | \mathcal{I}_t]$$

where $X_{i,t+1}$ reflects investors' valuation of this asset's future payoffs, M_{t+1} denotes the realized marginal rates of substitution, and it encapsulates investors' different preferences, $\mathcal{I}_t$ is the available information set at time t.

Returns

It is common to analyze prices in an equivalent format: expected returns, or "discount rate".

$$E[R_{i,t+1}|\mathcal{I}_t] = \beta_{i,t}\lambda_t$$

where $R_{i,t+1} = X_{i,t+1}/P_{i,t} - R_{f,t}$ is the asset's excess return.
$R_{f,t} = E[M_{t+1}|\mathcal{I}_t]^{-1}$ is the one-period risk-free rate,
$\beta_{i,t} = \frac{Cov[M_{t+1},R_{i,t+1}|\mathcal{I}_t]}{Var[M_{t+1}|\mathcal{I}_t]}$ is the asset's covariance with M_{t+1}, and
$\lambda_t = -\frac{Var[M_{t+1}|\mathcal{I}_t]}{E[M_{t+1}|\mathcal{I}_t]}$ is the price of risk.

We can ask economic questions in terms of either prices or returns but the literature typically opts for returns, why?

Prices VS. Returns

- Prices are often non-stationary while returns are often stationary, so when the statistical properties of estimators rely on stationarity assumptions it is advantageous to work with returns.
- Differences in the scale of assets' payoffs will lead to uninteresting scale differences in prices. But returns are typically unaffected by differences in payoff scale so the researcher need not adjust for them.
- Returns are also predictions, and their interpretation is especially simple and practically important.

Information Sets are Large

The information set that can be used for predicting prices (returns) is enormous. Professional managers routinely pore over troves of news feeds, data releases, and expert predictions in order to inform their investment decisions.

The expanse of price-relevant information is further compounded by the panel nature of financial markets:

- The price of any given asset tends to vary over time in potentially interesting ways (time series dimension).
- At a given point in time, prices differ across assets in interesting ways (cross section dimension).

New forms of data can be utilized in financial studies (i.e. text, voice, facial recognition)

Function Forms are Ambiguous

The traditional econometric approach to financial market research:

- Specify a **functional form** for the return forecasting model motivated by a **theoretical economic model**
- Estimates **parameters** to understand how candidate information sources associate with observed market prices within the confines of the chosen model

Then which economic model should we choose?

Function Forms are Ambiguous

- Consumption-based models: fail to match market price data by most measures (e.g., Mehra and Prescott, 1985).
- Structural models: match price data somewhat better, but only on an in-sample basis. (e.g., Chen et al., 2022a)
- Reduced-form models: the current popular approach. It avoids imposing detailed economic structure, but it typically imposes statistical structure.

But there are many potential choices for statistical structure in reduced-form models, which one should we choose?

Function Forms are Ambiguous

Recall how we conduct linear regressions and how we introduce nonlinear relationships into our regression models.

Not only do market participants impound rich information into their forecasts, they do it in potentially complex ways that leverage the **nuanced powers of human reasoning and intuition**.

Investors use information in ways that we as researchers **cannot know explicitly** and thus **cannot exhaustively specify in a parametric statistical model**.

Comprised of diverse non-parametric estimators and large parametric models, machine learning methods are explicitly designed to approximate unknown data generating functions.

Machine Learning VS. Econometrics

At its core, machine learning **need not be differentiated** from econometrics or statistics more generally.

Gu et al. (2020b) describe machine learning as:

- (1) a diverse collection of high-dimensional models for statistical prediction
- (2) "regularization" methods for model selection and mitigation of overfitting
- (3) efficient algorithms for searching among a vast number of potential model specifications

On Definition Part (1)

Econometrics methods prefer small models that can have comparatively **precise parameter estimates** and **ease of interpretation**. But these models can be rigid and oversimplified.

Machine learning is preferred when the analyst is **unsure which specific structure their statistical model should take**. These models are much more flexible, but can also be more sensitive and suffer from poor out-of-sample performance when they overfit noise in the system.

Machine learning can be viewed as **nonparametric (or semi-parametric) modeling**. Researchers turn to large models when the benefits from more accurately describing the complexities of real world outweigh the costs of potential overfitting.

On Definition Part (2)

Machine learning puts model selection at the heart of its empirical design.

However, this idea has a rich history in econometrics under the heading of model selection (and, relatedly, model averaging).

Machine learning research processes are accompanied by **regularization**, which is a blanket term for constraining model size to encourage stable performance out-of-sample.

Regularization methods encourage smaller models; richer models are only selected if they are likely to give a genuine boost to out-of-sample prediction accuracy.

On Definition Part (3)

Definition Part (3) is perhaps its clearest differentiator from traditional statistics.

Traditional econometric estimators only cease the parameter search when the routine **converges**.

Machine learning models halt a search before convergence to **reduce computation** and do so with **little loss of accuracy**.

Challenges of Applying Machine Learning in Finance

- While machine learning is often viewed as a "big data" tool, many foundational questions in finance are frustrated by the decidedly "small data" reality of economic time series (i.e., macro data).
- Financial research often faces weak signal-to-noise ratios. For example, price variation is expected to emanate predominantly from the arrival of unanticipated news. Returns are expected to be small and fiercely competed over.
- Investors learn and markets evolve. This creates a moving target for machine learning prediction models. Previously reliable predictive patterns may be arbitraged away.

Two Cultures of Financial Economics

"structural model/hypothesis test" culture VS. "prediction model" culture

For **structural models**, the constraints come in the form of:

- specific functional forms/distributions
- limited variables admitted into the conditioning information set

These models often "generalize" poorly in the sense that they have weak explanatory power for asset price behaviors outside the narrow purview of the model design or beyond the training data set.

Two Cultures of Financial Economics

For **prediction models**:

This culture willingly espouses model specifications that might lack an explicit association with economic theory, so long as they produce meaningful, robust improvements in data fit versus the status quo.

However, prediction models are often labeled as *"there is no economics"*. Does that mean prediction models are inferior to structural models or even useless?

Two Cultures of Financial Economics

Structural model learns economics by probing specific economic mechanisms. But economics is not just about testing theoretical mechanisms.

Atheoretical (for lack of a better term) prediction models survey the empirical landscape in broader terms, charting out new empirical facts upon which theories can be developed, and for which future hypothesis tests can investigate mechanisms.

Even if details of the economic mechanisms remain shrouded, economic actors—financial market participants in particular—can always benefit from improved empirical maps.

Distance Between Theory and Reality

There is a wedge between the efficiency of allocations achievable by economic agents when the data generating process (DGP) is known (*first-best*), versus when it must be estimated by the model we impose (*second-best*).

There is another wedge due to sampling variation. Even if we knew the functional form of the DGP, we still must estimate it and noise in our estimates produces deviations from first-best (*third-best*).

Hence, in reality we must always live with *third-best* allocations; i.e., mis-specified models that are noisily estimated.