AI for Business: Insights from Corporate Data

Topic 7: Applications of Machine Learning in Business

Miao Liu

Boston College

March 8, 2026

Overview: Topic 7

1. Can Machine Learn Finance?
2. Strengths and Limitations of ML in Business: Zillow iBuying

Overview: Topic 7

Fundamental Questions

- **Are asset returns predictable?**
- **If yes, should machine learning be the tool of choice?**

Is Finance Different?

- ML thrives in large data and high signal-to-noise environment
- To set realistic expectations for ML in asset management, it is crucial to understand what makes finance fundamentally different

Small Data in Finance

- **Return Prediction as a Small Data Problem:**
  - Consider the model:

$$r_t = \sum_{i=1}^{N} \beta_i x_{i,t-1}, \quad t = 1, \ldots, T.$$

  - r_t represents future returns and $x_{i,t-1}$ the predictor variables.
- **Key Insight:**
  - The richness of a model is determined not by the number of predictors N but by the number of independent return observations T.
  - Even if N is large, if T is small (e.g., 100 observations), the number of parameters you can reliably estimate is effectively capped.

Small Data in Finance

- **Frequency and Data Availability:**
  - The appropriate frequency of returns (e.g., monthly) is dictated by trading costs and investor rebalancing ability
  - For macro strategies (currencies, bonds, commodities), a few decades of monthly data yield only a few hundred observations
  - For assets like equities, while there are thousands of assets, but high cross-sectional correlation means the effective number of observations is much smaller
- **Implications for Finance:**
  - In finance, unlike other ML domains, additional data cannot be generated by experimentation—the only way to grow T is to wait.
  - High-frequency trading (HFT) firms do have more data, but their strategies are limited by trading costs and market impact.

Low Signal-to-Noise Ratios: Unanticipated News

- **Weak Predictive Signal because the target is driven by:**
  - Financial markets are heavily influenced by unanticipated news, leading to drastic performance swings
  - Liquidity trading orthogonal to fundamentals

Low Signal-to-Noise Ratios: Economic Forces

Economic Forces at Work:

- Profit maximization and competitive trading force informed traders to exploit any predictable signals
- This trading quickly drives prices to levels that "exhaust" the available information, leaving minimal predictable variation
- As posited by Fama (1970), efficient markets rapidly price in available information, thereby reducing predictable return components

Low Signal-to-Noise Ratios: Evolving Markets

- Markets are adaptive and dynamic.
- Technological innovations and structural changes (e.g., financial crisis, pandemic, etc.) can alter the structure of the economy and reshape market interactions.

Need for Interpretability

- Many machine learning models are notorious as "black boxes," making it challenging to extract meaningful interpretations of the underlying mechanisms
- In asset management, interpretability is critical—not only for understanding model risks but also to meet fiduciary duties and effectively communicate strategies
- There is an inherent trade-off between predictability and interpretability; even if a model is highly predictive, asset managers may prefer models whose inner workings they can understand

Can Machine Learn Finance? Evidence

- We will take a look at the state-of-the-arts practice on Wall Street

Data

- **Data Overview:**
  - Monthly individual equity returns from CRSP (NYSE, AMEX, NASDAQ) from March 1957 to December 2016 (60 years).
  - Approximately 30,000 stocks in total, with an average of over 6,200 stocks per month.
- **Predictor Variables:**
  - **Stock-level Characteristics:** 94 features (61 updated annually, 13 quarterly, 20 monthly) plus 74 industry dummies
  - **Macroeconomic Predictors:** 8 variables (e.g., dividend-price ratio, earnings-price ratio, book-to-market, etc.)
  - Total number of predictors is 920 = 94*8+94+74

Sample Splitting and Model Fitting Strategy

- **Data Partitioning:**
  - **Training Sample:** 18 years (1957–1974).
  - **Validation Sample:** 12 years (1975–1986) for hyperparameter tuning.
  - **Testing Sample:** 30 years (1987–2016) for out-of-sample performance evaluation.
- **Model Fitting:**
  - Machine learning methods approximate the model $\mathbb{E}_t(r_{i,t+1}) = f(x_{i,t})$ by minimizing mean squared prediction error (MSE) with regularization to avoid overfitting
  - Models are refit once every year
  - Each refit expands the training sample by one year and rolls the 12-month validation window forward
- **Objective:** Max out-of-sample predictive power for individual stock excess returns

Sharpe Ratio Definition

- **Definition:** Excess return per unit of risk for an asset or portfolio
- **Purpose:** It evaluates how effectively an investment compensates the investor for the risk taken
- **Equation:**

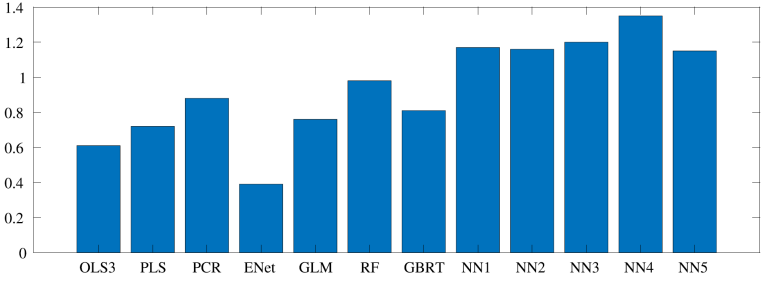
$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

where:

- R_p is the return of the portfolio
- R_f is the risk-free rate
- σ_p is the standard deviation of the portfolio's excess returns

Model Performance: Sharp Ratios

Exhibit 3. Sharpe ratio comparison of machine learning strategies.



Note. Out-of-sample annualized Sharpe ratios for value-weighted decile spread long–short portfolios based on predictions from 12 machine learning models including a three-predictor OLS (OLS3) model, partial least squares (PLS), principal components regression (PCR), elastic net (ENet), generalized linear model (GLM), random forest (RF), gradient-boosted regression trees (GBRT), and feed forward neural networks with one to five hidden layers (NN1–NN5).
Source: Gu *et al.* (forthcoming).

Model Performance: Cumulative Portfolio Returns

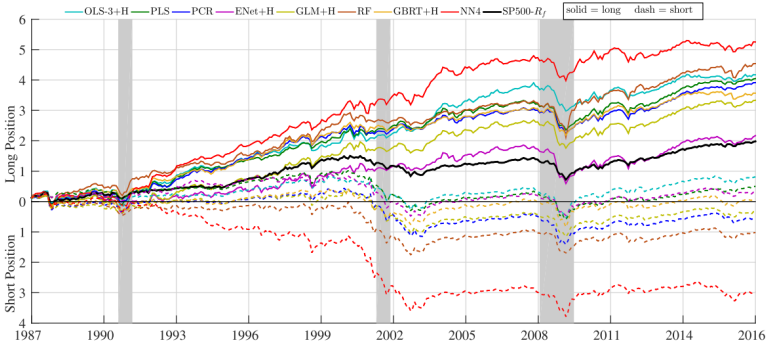


Figure 9
Cumulative return of machine learning portfolios
The figure shows the cumulative log returns of portfolios sorted on out-of-sample machine learning return forecasts. The solid and dashed lines represent long (top decile) and short (bottom decile) positions, respectively. The shaded periods show NBER recession dates. All portfolios are value weighted.

Takeaways

- Stock returns are a complex function of firm characteristics, macro environments, and their interactions - RF and NN with deep layers work the best
- Returns are difficult to predict for various reasons
- State-of-the-art ML models performed well pre-2005 but not afterwards, suggesting that widespread adoption of ML-based predictions has made markets more efficient at incorporating signals.

Overview: Topic 7

Assignment for Next Week

- Lab Assignment 6
- Required Reading: *Can Machines "Learn" Finance?* sections 1 to 5.1