Machine Learning for Financial Prediction

Theme II: Algorithmic Finance

Research Question: Can machine learning beat the market, and if so, why doesn't arbitrage eliminate the opportunity?

PhD Seminar in Digital Finance

Efficient Markets and Machine Learning

Efficient Market Hypothesis
Fama (1970): Prices reflect all available information.

$$P_t = \mathbb{E}[V|\mathcal{F}_t]$$

Implications for ML

- Weak form: Past prices uninformative
- Semi-strong: Public info priced
- Strong: All info (including private)

Grossman-Stiglitz Paradox
If markets perfectly efficient:

- No return to information acquisition
- No one acquires information
- Prices cannot be efficient

Proposition

In equilibrium, prices are partially revealing:

$$\textit{Return to ML} = \textit{Cost of ML}$$

Implication: ML can earn competitive returns, not excess returns long-term.

Grossman & Stiglitz (1980), "On the Impossibility of Informationally Efficient Markets"

Empirical Asset Pricing via Machine Learning

Gu, Kelly & Xiu (2020)
Predict individual stock returns:

$$r_{i,t+1} = g(z_{i,t}) + \epsilon_{i,t+1}$$

where $z_{i,t} = 94$ firm characteristics.

Methods Compared

Model	OOS R^2
OLS (all)	-1.2%
LASSO	0.4%
Random Forest	0.6%
Neural Network	0.8%
Ensemble	1.0%

Key Findings

- Nonlinear models dominate
- Interaction effects matter
- Variable importance: momentum, liquidity, volatility

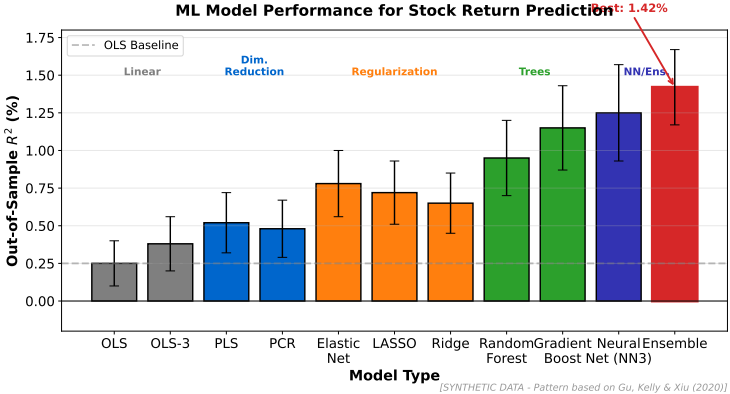
Economic Significance
Long-short portfolio:

- Sharpe ratio: 1.8 (vs 0.4 market)
- But: High turnover, transaction costs
- Net alpha: Debated

Critique: In-sample period tuning?

Gu, Kelly & Xiu (2020), "Empirical Asset Pricing via Machine Learning," RFS

ML Model Performance: Out-of-Sample R-Squared



Neural networks achieve highest OOS R-squared but require careful regularization and training.

The Overfitting Problem in Financial ML

Sources of Overfitting

1. **Low signal-to-noise**: $R^2 < 1\%$
2. **Non-stationarity**: Regime changes
3. **Multiple testing**: Factor zoo
4. **Data snooping**: Look-ahead bias

Bias-Variance Decomposition

$$\text{MSE} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

In finance: Noise dominates.

Statistical Challenges
Non-Stationarity

- Training: 1960-2000
- Test: 2000-2020
- But: Financial crisis, QE, meme stocks

Structural Breaks

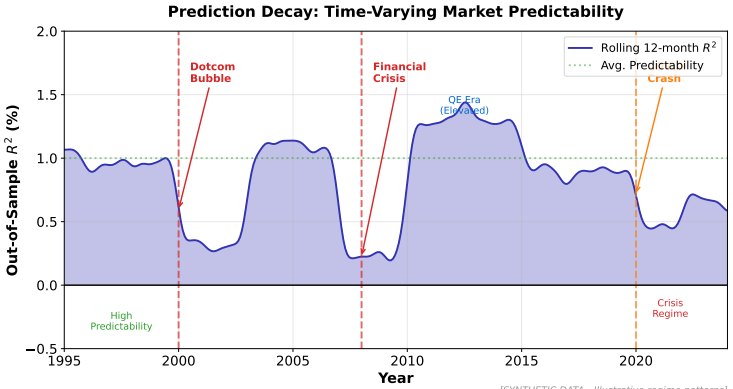
$$\beta_{pre} \neq \beta_{post}$$

Decay of Predictability
Predictable returns erode as:

- Strategies crowd
- Arbitrage capital enters
- Anomaly becomes common knowledge

McLean & Pontiff (2016): Anomaly returns decline 58% post-publication

Predictability Decay: Structural Breaks and Regime Changes



Model performance degrades during regime changes; continuous retraining is essential.

Theoretical Limits to Return Predictability

Fundamental Limits
Even with perfect model:

$$R^2_{max} \leq \frac{\sigma^2_{signal}}{\sigma^2_{signal} + \sigma^2_{noise}}$$

Campbell & Thompson (2008)
Monthly stock return R^2:

- Expected: 0.5% (theoretical max $\approx$ 2%)
- Achieved: 0.1-1.0%
- Economic value: Modest Sharpe gains

Why Limits Exist

1. **Arbitrage**: Eliminates predictable patterns
2. **Uncertainty**: Future is genuinely random
3. **Information costs**: Acquisition expensive
4. **Capacity constraints**: Strategies don't scale

Martin & Nagel (2022)
Lower bound on expected returns:

$$\mathbb{E}[R] \geq \text{Risk premium}$$

Excess returns require risk or market failure.

Campbell & Thompson (2008), "Predicting Excess Stock Returns," RFS

Recent Advances: Transformers and Alternative Data

Attention-Based Models

Transformer architecture for time series:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Finance Applications

- Cross-sectional attention (stock interactions)
- Temporal attention (regime detection)
- Multi-modal (text + prices)

Challenges

- Data hungry
- Computational cost
- Interpretability

Alternative Data (2024 Papers)

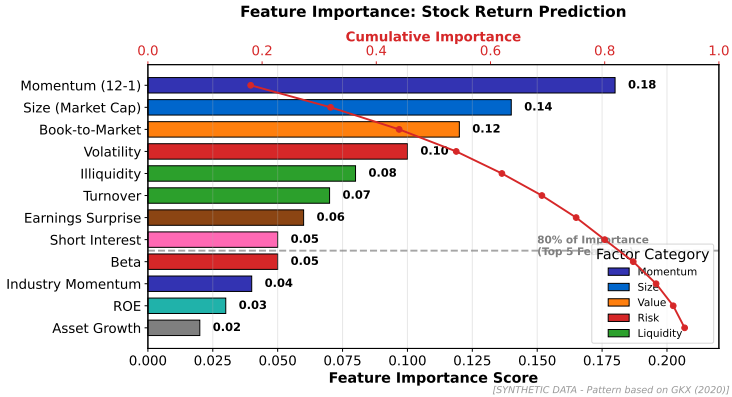
Data Source	Claimed α
Satellite imagery	2-3%
Credit card txns	1-2%
Social media	0.5-1%
Job postings	1-2%
Web traffic	0.5-1%

Skeptical View

- Publication bias
- Backtest overfitting
- Capacity limits
- Data vendor marketing

Alternative data market: $7B by 2025 (Opimas estimate)

Feature Importance: What Drives Predictions?



SHAP-style decomposition reveals which features drive individual predictions.

Large Language Models in Finance

Lopez-Lira & Tang (2023)
GPT-3.5 for stock prediction:

- Input: News headlines
- Output: Sentiment score
- Result: Significant predictability

Methodology

$$\text{Score} = \text{LLM}(\text{“Is this good or bad for stock price?”})$$

Long-short portfolio Sharpe: 0.8-1.2

Critique and Concerns

1. **Data contamination**: Training data includes backtests?
2. **Temporal validity**: Knowledge cutoff issues
3. **Causal inference**: Correlation vs. causation
4. **Capacity**: Can results scale?

Research Frontier

- Fine-tuned financial LLMs
- Multi-modal (10-K + charts)
- Agent-based trading

Key Question: Are LLMs a new factor or noise?

Lopez-Lira & Tang (2023), “Can ChatGPT Forecast Stock Price Movements?” – SSRN

Open Research Questions

Publishable Research Directions

1. **Decay of ML Alpha**
   - RQ: How quickly do ML strategies lose efficacy after deployment?
   - Method: Track strategy returns over time, control for crowding
   - Gap: Limited longitudinal studies of strategy lifecycle
2. **Feature Importance Stability**
   - RQ: Are ML-identified predictors stable across regimes?
   - Method: Rolling window SHAP analysis, regime detection
   - Gap: Most papers report point-in-time importance
3. **LLM Information Content**
   - RQ: What information do LLMs extract that traditional NLP misses?
   - Method: Residual analysis, attention visualization
   - Gap: Black box nature of LLM predictions

Replication studies are valuable; many ML finance papers don't replicate

Exercises

Mathematical
Derive the bias-variance tradeoff for ridge regression:

$$\hat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'y$$

Show:

- Bias increases with λ
- Variance decreases with λ
- Optimal λ^* minimizes MSE

Empirical
Replicate Gu, Kelly & Xiu:

1. Download CRSP/Compustat
2. Construct 20 characteristics
3. Compare OLS vs. LASSO
4. Report OOS R^2

Data: WRDS (Wharton)

Research Proposal
Draft 1-page proposal:

- "ML Strategy Crowding"
- Measure crowding proxy
- Test return erosion
- Identification challenge

Due: Week 5 – Empirical exercise is substantial; start early

Replication is the best way to understand ML finance research

Required Readings

Core Papers (Read Before Class)

1. **Gu, Kelly & Xiu** (2020). "Empirical Asset Pricing via Machine Learning." *RFS*, 33(5), 2223-2273.
   - Focus: Sections 1-3, Tables 3-5, methodology
2. **Campbell & Thompson** (2008). "Predicting Excess Stock Returns Out of Sample." *RFS*, 21(4), 1509-1531.
   - Focus: Section 2 (theory), economic significance

Supplementary

- Grossman & Stiglitz (1980): Information and efficiency
- McLean & Pontiff (2016): Anomaly decay – important for skepticism
- Lopez-Lira & Tang (2023): LLM frontier – SSRN

Gu et al. is methodologically influential; Campbell-Thompson provides economic grounding