MACHINE LEARNING IN ASSET PRICING:

Exploring Predictability in Stock Returns

Ximing Luo • CS (DMD) & Econ, 2027

Mentor: Yiwen Lu • Finance PhD 2nd Year

CONTENT

INTRODUCTION

Challenge:

- Despite the efficient market hypothesis suggesting that asset prices fully reflect all available information, empirical evidence shows subtle, persistent anomalies in returns—patterns too faint for traditional models to reliably capture

Motivation:

- Asset prices reflect predictions based on vast, evolving information sets
  - Traditional models struggle to keep up with their complexity and volume
- Machine learning offers flexibility in uncovering patterns from large, ambiguous, and nonlinear data without requiring strict functional assumptions, unlike traditional econometric models

BACKGROUND

- **Efficient Market Hypothesis (EMH):**
  - Weak, Semi-strong, and Strong forms.
- **Asset Pricing Fundamentals:**
  - Asset Pricing Equation:

$$P_{i,t} = E\left[M_{t+1}\, X_{i,t+1} \mid \mathcal{I}t\right]$$

  - Expected Return Representation:

$$E\left[R_{i,t+1} \mid \mathcal{I}t\right] = \beta_{i,t}\lambda_t$$

  - CAPM Equation (Benchmark)

$$E\left[R_i\right] = R_f + \beta_i\left(E\left[R_m\right] - R_f\right)$$

$P_{i,t}$ = asset price
M_{t+1} = stochastic discount factor
$X_{i,t+1}$ =future payoff
$\mathcal{I}_t$ = available information

$\beta_{i,t}$ =asset sensitivity
λ_t =price of risk

R_f = risk-free rate
R_m = market return

Large, complex information sets imply that “prices are predictions” built on much richer data than what classical models assume

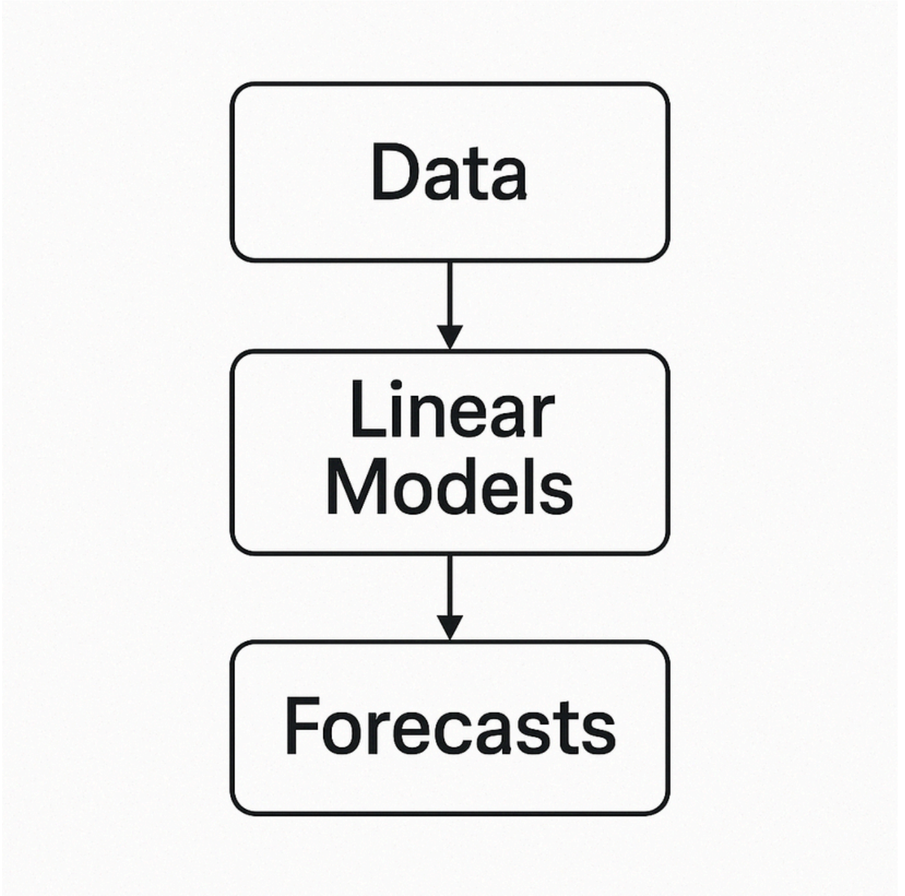
CLASSICAL ECONOMETRIC MODELS

Traditional Models:

- Linear regression, ARIMA, and factor models (Ex: Fama–French)
- Concepts such as the random walk hypothesis

Benchmarking Role:

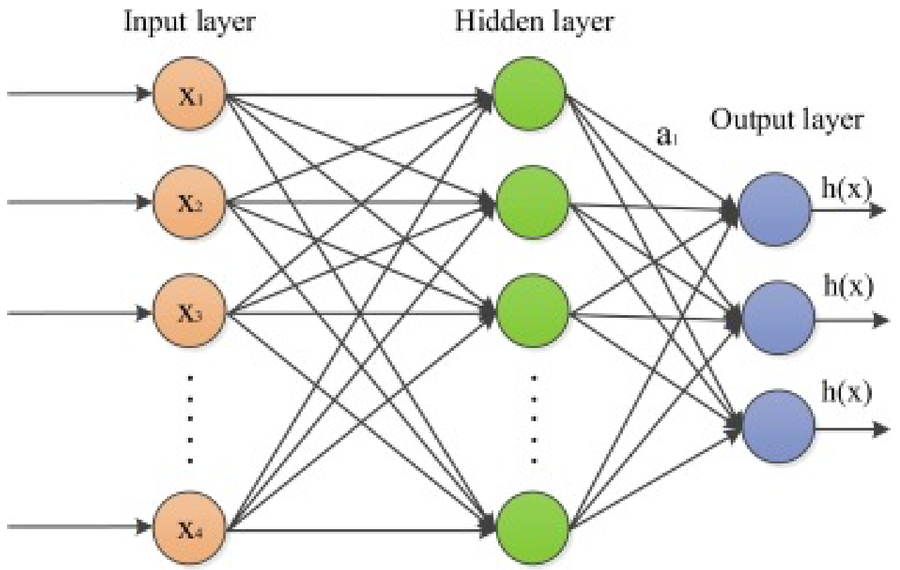
- These methods provide a baseline against which modern ML methods are compared.



MODERN MACHINE LEARNING TECHNIQUES

Modern ML Approaches:

- Penalized Regression Models (Ridge/Elastic Net)
- Tree-Based Methods:
  - Decision Trees, Random Forests, Gradient Boosting
- Deep Learning Architectures:
  - Feed-Forward Neural Networks (FFNN)
  - Recurrent Neural Networks & LSTMs
  - Convolutional Neural Networks (CNNs)
    - Used to extract spatial features from chart images
- Alternative Data Integration:
  - Textual analysis using supervised topic models for sentiment
  - Image feature extraction from candlestick charts using CNNs



Key Advantages: Capture nonlinear functions, handle high-dimensional and heterogeneous data, regularize effectively to avoid overfitting

DATA, ANALYSIS & ALTERNATIVE DATA SOURCES

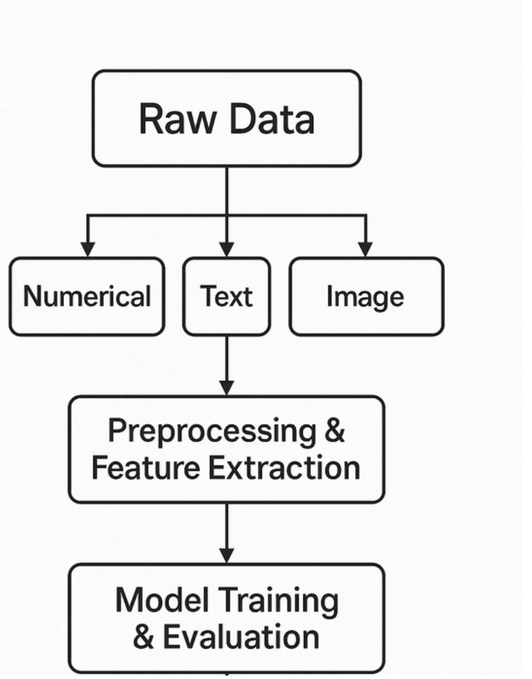
Data Sources:

- Traditional: Historical asset prices; technical and fundamental signals
- Alternative:
  - Textual data from news articles and social media
  - Image data from OHLC candlestick charts

Processing Pipeline:

- Feature Selection & Extraction:
  - Use PCA or Principal Component Regression
- Model Training & Validation:
  - Cross-validation selects models based on their predictive performance in the pseudo-out-of-sample data

Integration: Merge numerical features with alternative data streams using NLP for text and CNNs for images.



KEY FINDINGS

Empirical Results:

- Modern ML models improve out-of-sample return forecasting
- Higher-dimensional models achieve higher R^2 and improved trading metrics

Trading Performance Metric:

$$Sharpe\ Ratio = \frac{R_x - R_f}{std(R_x)}$$

$R_x = Average\ rate\ of\ return\ on\ investment\ x$

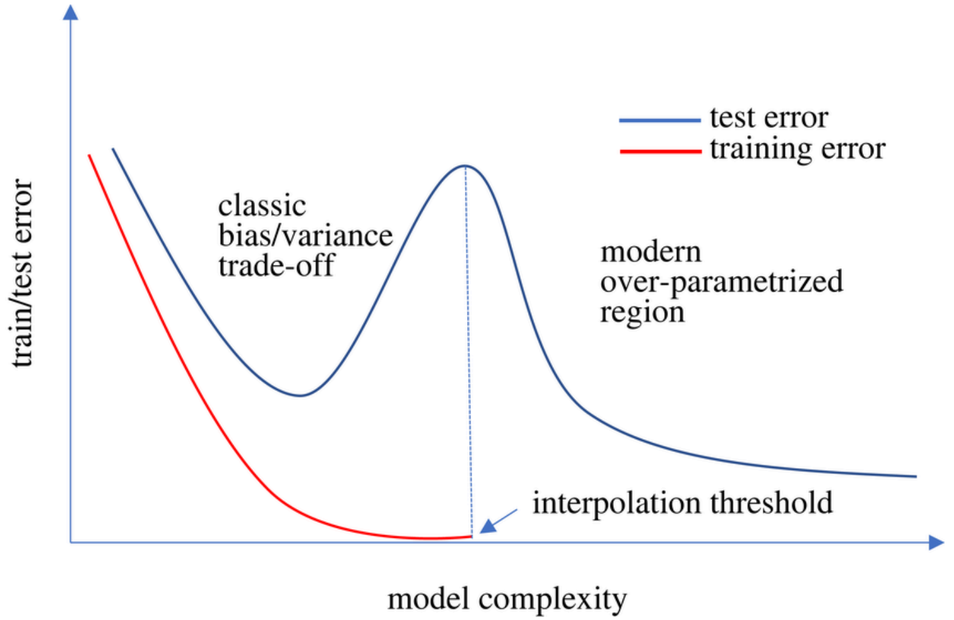
$R_f = Risk\ free\ rate$

$std(R_x) = standard\ deviation\ of\ R_x$

The Complexity Wedge:

- As model complexity increases, the gap between in-sample performance and true out-of-sample performance widens

Double Descent/Ascent Phenomenon: Overparameterized models (with P/T > 1) can generalize better than expected due to implicit regularization and shrinkage



IMPLICATIONS

Trading Applications

- Generation of enhanced, precise trading signals using high-dimensional, nonlinear models
- Design of adaptive long-short strategies that capitalize on refined return forecasts

Risk Management Enhancements

- More accurate volatility forecasts lead to improved stress testing and scenario planning
- Adaptive models that react to market regime shifts and incorporate real-time alternative data

Modern ML Contributions

- Real-time integration of textual and image data for agile decision-making
- Flexibility in adapting to evolving market conditions and rich investor information

TAKEAWAYS & FUTURE DIRECTIONS

Key Takeaways:

- Modern ML techniques, by leveraging extensive data and nonlinear approaches, enhances asset return predictability
- Conventional wisdom favors simple models to avoid overfitting. But research (e.g. Kelly et al. 2022a) shows that complex models—even those with more parameters than observations ($P/T > 1$)—can actually improve both statistical and economic performance when paired with shrinkage or regularization

Future Directions:

- Deeper integration of alternative data such as real-time sentiment analysis and image processing
- Hybrid frameworks that combine economic theory with data-driven predictions to better explain market dynamics

REFERENCES

Books

- Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.
- Kelly, B., & Xiu, D. (2023). *Financial Machine Learning*. Foundations and Trends® in Finance, 13(3–4), 205–363. now Publishers Inc.

Articles

- Campbell, J. Y. (2000). Asset Pricing at the Millennium. *Journal of Finance*, 55(4), 1515–1567.
- Fama, E. F. (1991). Efficient Capital Markets: II. *Journal of Finance*, 46(5), 1575–1617.

THANK YOU!