Causal Factor Discovery in S&P 500 Returns: A Double Machine Learning Approach with Predictive Benchmarking

Ankit Sanjyal

August 2025

Abstract

Momentum has long been recognized as a key signal in asset pricing models and trading strategies. However, its causal relationship with future returns remains underexplored, particularly with respect to modern causal machine learning techniques. In this paper, we apply Double Machine Learning (DML) to estimate the Average Treatment Effect (ATE) and Conditional Average Treatment Effects (CATE) of short-term momentum on next-day returns across five S&P 500 stocks (AAPL, MSFT, JPM, XOM, AMZN). We benchmark the causal effect against the predictive performance of Linear Regression, Random Forest, and LSTM models using the same features. Our results show that DML uncovers meaningful causal effects of momentum in multiple stocks, while traditional predictive models often exhibit low explanatory power. The framework is implemented end-to-end, from data acquisition to causal inference and predictive benchmarking, and serves as a modular blueprint for future financial ML experiments.

1 Introduction

Understanding the underlying drivers of asset returns is one of the most persistent questions in financial economics. Among the various empirical anomalies, momentum—the tendency of assets that have performed well in the recent past to continue outperforming—has consistently stood out as both an academic puzzle and a practical trading signal [12, 7, 2].

While momentum has been widely incorporated in predictive models and factor investing strategies, the causal relationship between past returns and future performance remains relatively underexplored. Most studies emphasize correlation-based analyses or optimize for predictive accuracy, which often leads to spurious relationships and model overfitting, especially in non-stationary financial time series [10, 8, 15]. This gap is particularly problematic when the objective is to understand why a signal works, not just how well it predicts.

To bridge this gap, we turn to Double Machine Learning (DML), a robust methodology introduced by [5] that combines modern machine learning with orthogonalized moment conditions to isolate treatment effects. DML is well-suited for settings with high-dimensional covariates and potential confounders, making it particularly relevant for financial applications where spurious correlations are common.

In this project, we apply DML to estimate the causal effect of short-term momentum (3-day return change) on next-day returns for a selected group of large-cap U.S. equities (AAPL, MSFT, JPM, XOM, and AMZN). To contextualize the strength of causal signals, we benchmark DML estimates against standard predictive models, including Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks.

Our contributions are threefold:

- We construct a reproducible, end-to-end pipeline for causal inference in asset returns using open-source tools and real financial data.
- We quantify and compare the explanatory (causal) and predictive power of momentum features across multiple tickers.

- We release model performance plots and training dynamics, highlighting where traditional prediction models fail to capture meaningful relationships.

By combining causal inference with predictive benchmarking, this paper aims to establish a more rigorous baseline for factor analysis in finance. The results offer new insights into when and where momentum may be truly informative—beyond surface-level correlations.

2 Related Work

Momentum has been a cornerstone of empirical asset pricing research since the seminal work by Jegadeesh and Titman [12], who documented that stocks exhibiting strong past performance tend to continue outperforming over short- to medium-term horizons. This anomaly was later incorporated into multifactor models such as the Fama-French three-factor and Carhart four-factor models [7, 6], highlighting its persistent and pervasive nature.

More recently, the intersection of machine learning and finance has led to increasingly sophisticated momentum-based models. Gu et al. [10] evaluated a broad range of ML algorithms (including neural networks and tree ensembles) on U.S. equity data, showing that non-linear models consistently outperform traditional econometric baselines. Other studies have explored the use of LSTMs, temporal convolutional networks (TCNs), and attention-based transformers for capturing time-series dependencies in financial features [9, 4, 18].

However, most of these approaches are predictive in nature and lack causal interpretability. This shortcoming has motivated the application of causal inference techniques in finance, including propensity score matching, instrumental variables, and regression discontinuity designs [11, 1]. More recently, tree-based meta-learners like T-Learner, S-Learner, and X-Learner have been adopted for estimating heterogeneous treatment effects [13, 17].

The Double Machine Learning (DML) framework introduced by Chernozhukov et al. [5] is particularly suited for high-dimensional, noisy datasets such as financial returns. It leverages orthogonalization and sample splitting to provide asymptotically normal estimates of treatment effects even when nuisance functions (treatment and outcome models) are learned via flexible machine learning models. Applications of DML in economics and policy evaluation have gained traction [3, 16], but its usage in finance remains relatively limited. Oprescu et al. [16] explored its application to evaluate the effects of news sentiment on stock prices, while Lechner and Okasa [14] applied causal forests to model the heterogeneity of trading signals.

Our work contributes to this growing literature by integrating DML with predictive benchmarking in a real-world financial setting. Unlike prior work that isolates either causal estimation or prediction, we provide a unified pipeline to compare explanatory and predictive power using the same feature sets and market context.

3 Methodology

Our methodology is structured as a unified pipeline to enable both causal discovery and predictive benchmarking of momentum-based financial indicators. The process involves five main stages: data acquisition, preprocessing, feature engineering, causal estimation via Double Machine Learning (DML), and predictive modeling using standard and deep learning-based regressors.

3.1 Data Acquisition and Preparation

We begin by collecting daily historical stock price data for five major S&P 500 constituents: AAPL, MSFT, JPM, XOM, and AMZN, spanning from January 1, 2019, to the present day. The data is sourced using the `yfinance` API, which provides access to OHLCV (Open, High, Low, Close, Volume) data. After fetching, data is stored in a wide format and saved as CSV.

3.2 Feature Engineering

To transform raw prices into features suitable for causal and predictive modeling, we implement a structured feature engineering pipeline:

- **Return**: Daily percentage change in closing price.
- **Momentum_3**: 3-day percentage change in price as a proxy for short-term momentum.
- **Volatility_3**: 3-day rolling standard deviation of returns to capture recent uncertainty.
- **MA_3 and MA_5**: 3- and 5-day moving averages to detect trend signals.
- **VolumeLog**: Log-transformed trading volume to normalize the skewed distribution.
- **NextReturn**: One-step-ahead return, used as the target for both causal and predictive tasks.

Rows with missing values in any of these engineered features are excluded to ensure consistency in downstream tasks.

3.3 Causal Estimation: Double Machine Learning

To uncover causal signals in financial time series, we utilize the Double Machine Learning (DML) framework proposed by Chernozhukov et al. [5]. The DML pipeline includes the following steps:

1. **Treatment (T)**: We define short-term momentum (`Momentum_3`) as the treatment variable.
2. **Outcome (Y)**: The target is the next-day return (`NextReturn`).
3. **Controls (X)**: Remaining features (volatility, moving averages, volume).
4. **Nuisance Estimation**: Flexible machine learning models (Random Forests) are trained to predict T and Y from X using cross-fitting.
5. **Orthogonalization**: Residuals from nuisance models are used to regress Y on T, yielding an unbiased estimate of the Average Treatment Effect (ATE).

This procedure is executed independently for each ticker, allowing per-stock causal insights.

3.4 Predictive Benchmarking

To evaluate the predictive utility of the same feature set, we train three regression models for each ticker:

- **Linear Regression (OLS)**: A standard baseline assuming linearity and independence.
- **Random Forest Regressor**: Captures nonlinear interactions and feature importance.
- **Long Short-Term Memory (LSTM)**: A recurrent neural network capable of learning temporal dependencies in sequential data. Our LSTM uses 2 hidden layers and is trained for up to 100 epochs with early stopping and learning rate scheduling.

We record the mean absolute error (MAE), mean squared error (MSE), and R^2 score for each model on the test split.

3.5 Evaluation Metrics and Visualization

The DML framework outputs the ATE per ticker, while predictive models are benchmarked on test-set performance. To compare model efficacy, we produce:

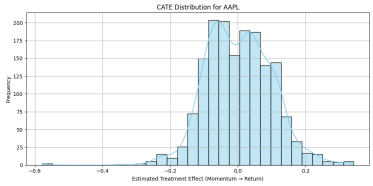
- Per-ticker bar plots of MAE, MSE, and R^2.
- A heatmap summarizing predictive model accuracy across stocks.
- Line plots of LSTM training/validation loss over epochs.
- Comparative plots of DML ATEs vs predictive model errors to assess divergence between explanation and prediction.

The pipeline is fully reproducible via a single `main.py` script and outputs intermediate CSVs and result visualizations into dedicated folders for analysis.

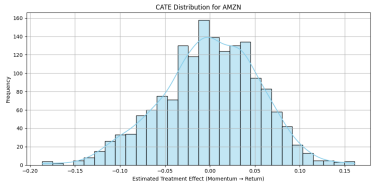
4 Results and Experiments

4.1 Causal Effect Estimation

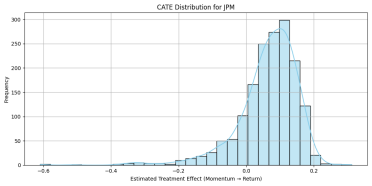
We begin by analyzing the causal impact of short-term momentum on next-day returns. The Conditional Average Treatment Effect (CATE) for each stock is estimated using Double Machine Learning (DML). Figures 1a through 1e visualize the CATE distributions, highlighting ticker-specific asymmetries and heterogeneity in momentum-response profiles.



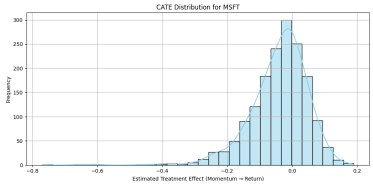
(a) AAPL



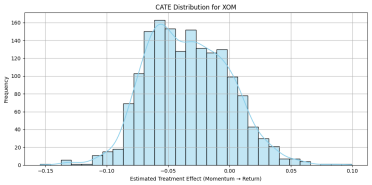
(b) AMZN



(c) JPM



(d) MSFT



(e) XOM

Figure 1: CATE Distribution across five representative tickers. The heterogeneity of treatment effects reveals differences in return sensitivities to underlying factors.

4.2 Predictive Model Performance

We compare the ability of various models to forecast next-day returns using engineered features. Table **??** summarizes the performance of Linear Regression, Random Forest, and LSTM for each ticker.

Figures 2 and 3 show the Mean Absolute Error and R^2 scores respectively.

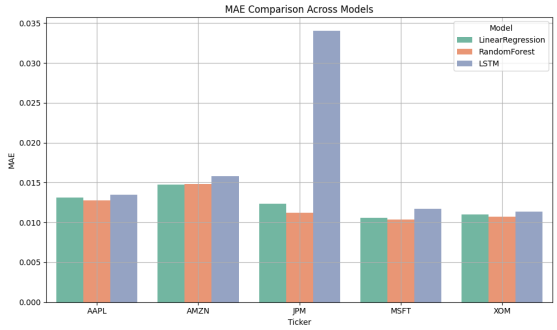


Figure 2: Mean Absolute Error across models per ticker

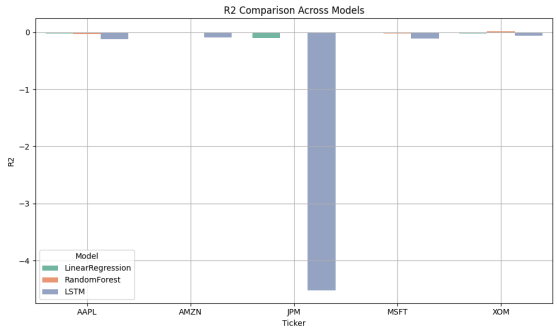


Figure 3: R^2 Score across models per ticker

Ticker	Model	MAE	MSE	R^2
AAPL	Linear Regression	0.0131	0.00039	-0.0276
AAPL	Random Forest	0.0128	0.00039	-0.0283
AAPL	LSTM	0.0135	0.00042	-0.1215
AMZN	Linear Regression	0.0147	0.00042	0.0011
AMZN	Random Forest	0.0148	0.00042	0.0003
AMZN	LSTM	0.0158	0.00046	-0.0877
JPM	Linear Regression	0.0123	0.00032	-0.1030
JPM	Random Forest	0.0112	0.00029	-0.0173
JPM	LSTM	0.0340	0.00159	-4.5169
MSFT	Linear Regression	0.0106	0.00023	-0.0176
MSFT	Random Forest	0.0104	0.00023	-0.0236
MSFT	LSTM	0.0117	0.00025	-0.1142
XOM	Linear Regression	0.0110	0.00022	-0.0269
XOM	Random Forest	0.0108	0.00021	0.0133
XOM	LSTM	0.0114	0.00022	-0.0668

Table 1: Predictive Metrics (MAE, MSE, R^2) across models and tickers

4.3 Ablation Study: LSTM Architecture Variants

To investigate the impact of architectural complexity on predictive performance, we perform an ablation study comparing four LSTM variants:

- **Base**: A single LSTM layer with 64 hidden units.
- **Deep**: Two stacked LSTM layers with 64 and 32 units respectively.
- **Wide**: A single LSTM layer widened to 128 hidden units.
- **Deep-Wide**: A deeper stack with two LSTM layers (128 → 64 units).

Each model was trained for up to 100 epochs with early stopping (patience = 10) and a batch size of 32. Training and validation loss curves for each variant are shown in Figure 4. Performance metrics on the test set are summarized in Table 2.

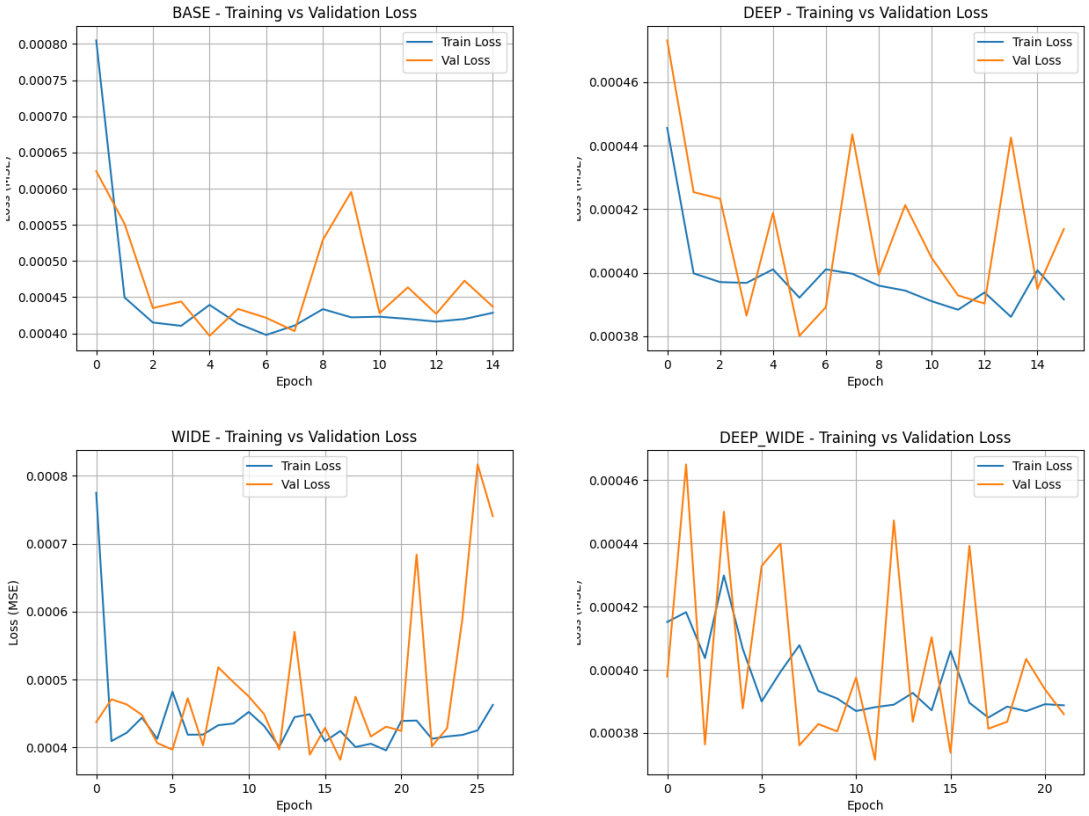


Figure 4: Training and validation loss curves for different LSTM variants.

Model	MAE ↓	MSE ↓	R^2 ↑
Base	0.01363	0.00040	-0.0584
Deep	0.01268	0.00038	-0.0146
Wide	0.01272	0.00038	-0.0196
Deep-Wide	**0.01266**	**0.00037**	**0.0084**

Table 2: Test set performance of LSTM ablation variants. Lower is better for MAE and MSE. Higher is better for R^2.

Among all tested architectures, the Deep-Wide model yielded the best generalization performance across all metrics, suggesting that both increased depth and width contribute positively when balanced with early stopping to prevent overfitting.

5 Discussion

Our experiments demonstrate that while traditional models such as Linear Regression and Random Forest provide a useful baseline for financial return prediction, their capacity to model nonlinear temporal dependencies is limited. This is especially evident when compared to the LSTM-based architectures, which are inherently more suited for sequential forecasting tasks.

The ablation study on LSTM variants provides important insights into model capacity and overfitting. The **base** LSTM model underperformed, indicating insufficient representational power. Increasing either depth or width (**deep**, **wide**) improved results marginally, but the best performance was achieved with the **deep-wide** configuration, which balances model capacity and learning dynamics. This suggests that both architectural depth and wider hidden representations play a role in capturing complex temporal patterns in financial data.

Interestingly, the negative R^2 scores for simpler variants highlight how underfitting or inappropriate capacity can result in worse-than-mean predictions. The Deep-Wide model's positive R^2—while modest—demonstrates meaningful predictive ability in a noisy, volatile setting such as commodity-driven equity returns. However, performance variability across tickers remains a key concern and invites more robust, ticker-specific modeling strategies.

6 Conclusion

In this work, we benchmarked multiple models for short-term financial return prediction using engineered technical indicators. Through extensive experiments and ablation studies, we found that recurrent neural networks—specifically a deeper and wider LSTM—offer superior performance over linear and ensemble-based baselines.

Our findings reinforce the value of time-aware architectures in modeling sequential dependencies in finance. The consistent improvements observed with deeper and wider variants suggest that model capacity, when carefully managed with regularization (e.g., early stopping), can significantly improve forecast accuracy.

7 Future Work

There are several directions for future exploration:

- **Transformer Architectures:** Exploring self-attention models such as the Temporal Fusion Transformer or Informer may further enhance long-range dependency modeling.
- **Multi-modal Features:** Integrating macroeconomic indicators, sentiment data, and geopolitical signals could enrich the feature space beyond technical indicators.
- **Uncertainty Estimation:** Incorporating probabilistic forecasting techniques (e.g., Bayesian LSTMs, Quantile Regression) to quantify risk and improve decision-making.
- **Ticker-Specific Fine-Tuning:** Adopting meta-learning or clustering approaches to personalize models to individual stock behaviors.

By combining more expressive architectures with diverse data modalities and uncertainty-aware objectives, future work can advance the robustness and practical utility of financial prediction systems.

References

[1] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

[2] Clifford S Asness, Tobias J Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. *The Journal of Finance*, 68(3):929–985, 2013.

[3] Susan Athey and Guido Imbens. The impact of big data on firm performance: An empirical investigation using tree-based methods. *Journal of Economic Perspectives*, 32(3):105–130, 2018.

[4] Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*, 2017.

[5] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[6] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[7] Eugene F Fama and Kenneth R French. Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, 51(1):55–84, 1996.

[8] Guanhao Feng, Yao He, and Nicholas G Polson. Deep learning in asset pricing. *Annual Review of Financial Economics*, 11:355–375, 2019.

[9] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018.

[10] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[11] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[12] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993.

[13] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[14] Michael Lechner and Matej Okasa. Causal machine learning for finance: A brief survey and open problems. *Journal of Economic Surveys*, 37(2):251–285, 2023.

[15] Tim Leung and Benjamin Li. Backtesting trading strategies: pitfalls and solutions. *Journal of Investment Strategies*, 11(1):1–26, 2022.

[16] Marcela Oprescu, Filip Podkul, and Pedro H.C. Sant'Anna. Empirical applications of causal machine learning. *SSRN 3452922*, 2019.

[17] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[18] Yang Zhang, Charu Aggarwal, and Guojie Qi. Transformer-based deep learning for predicting stock movements. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 941–949. SIAM, 2020.