Predicting Next-Day Stock Price Direction in the U.S. Steel Sector

Yangyang Wan
Department of Statistics
University of Michigan
Ann Arbor, USA
yywan@umich.edu

***Abstract*—This project is focusing on** whether next-day price direction (Up/Down) **for U.S. Steel and related companies can be predicted using related features**

Keywords—stock price, prediction, returns

I. Introduction

Short-horizon stock return predictability has been widely studied in empirical finance. While daily equity returns are typically close to unpredictable, certain weak structures—such as momentum, mean reversion, and volatility clustering—appear consistently across markets. These properties motivate the use of autoregressive features, moving-average trends, and volatility-based indicators in statistical forecasting models.

This project investigates whether **next-day price direction** (Up/Down) can be predicted for equities in the U.S. steel and metals sector using only historical price and volume information. The goal is not to construct a trading strategy, but to evaluate whether measurable predictive structure exists in these return dynamics and how different machine-learning methods compare in this setting.

Daily OHLCV data from **May 31, 2020 to May 31, 2025** were obtained through the **Alpaca Markets** historical equities API. Five tickers were analyzed: U.S. Steel (X), Nucor (NUE), Steel Dynamics (STLD), Cleveland-Cliffs (CLF), and the VanEck Steel ETF (SLX). Because U.S. Steel was delisted following an acquisition in 2025, its data terminate naturally near the end of the sample; this reflects real-world market events and does not disrupt the learning formulation.

The project contributes a full supervised-learning pipeline applied to financial time series, including feature engineering, model specification, time-series-aware hyperparameter tuning, and out-of-sample evaluation. Models studied include **Logistic Regression, Support Vector Machines, and Random Forests**, chosen to represent linear, margin-based nonlinear, and ensemble tree-based decision boundaries.

The predictability of asset returns has been a longstanding question in financial economics. While the Efficient Market Hypothesis (Fama, 1970) suggests returns should be unpredictable, empirical evidence (Lo and MacKinlay, 1988) has found short-term autocorrelations that hint at modest predictability. Building on this, recent studies have used machine learning (ML) models to uncover nonlinear and complex relationships in financial data. Gu, Kelly, and Xiu (2020) demonstrated that tree-based models and deep learning could outperform traditional linear models in cross-sectional return prediction for equities.

II. Method

A. Problem Formulation

For each ticker independently, daily log return is computed as

$$R_t = \log(P_t) - \log(P_{t-1}),$$

and the binary classification target is

$$Y_t = \mathbf{1}(R_t > 0),$$

indicating whether the next-day return is positive.

Each observation corresponds to a specific (ticker, date) pair, forming a unified panel dataset. The models predict Y_t from lagged and smoothed price-based features.

A chronological split is used to avoid look-ahead bias:

- **Training window:** 2020-05-31 to 2024-02-29
- **Testing window:** 2024-03-01 to 2025-05-31

This reflects a realistic forecasting problem: learning from past data and evaluating on a strictly later, unseen period.

B. Data Source and Preprocessing

Data were downloaded through Alpaca with split-adjusted prices. For each ticker, the data were sorted by date and cleaned

to remove any incomplete rows. Missing values arise only at the beginning of rolling windows and are dropped after feature construction.

C. Feature Engineering

The following predictors were computed **per ticker** using pandas group-operations:

1. **Lagged returns:** $R_{t-1}, \ldots, R_{t-5}$
2. **Moving averages:**
   - MA5 = 5-day simple moving average of Close
   - MA20 = 20-day simple moving average
   - MA_diff = MA5 − MA20
3. **Rolling volatility:** 10-day standard deviation of returns
4. **Volume change:** ΔVolume_t

These features capture short-term autocorrelation, local trend, regime volatility, and changes in liquidity.

D. Models

1) Logistic Regression

A linear classifier modeling

$$P(Y_t = 1 \mid X_t) = \sigma(X_t^\top \beta).$$

Serves as an interpretable baseline.

2) Support Vector Machine (RBF kernel)

Constructs a nonlinear decision boundary via the radial basis function kernel. Hyperparameters include C(margin penalty) and γ(kernel width).

3) Random Forest Classifier

An ensemble of decision trees capturing nonlinear interactions among features. Tuned using tree depth and number of estimators.

E. Hyperparameter Tuning

Because shuffling violates time ordering, tuning uses **TimeSeriesSplit** with 5 folds. Each fold trains on an expanding historical window and validates on the immediately subsequent period. This avoids leakage and yields realistic generalization estimates. Parameter grids:

- Logistic Regression: $C \in \{0.01, 0.1, 1, 10\}$
- SVM (RBF): $C \in \{0.1, 1, 10\}$, $\gamma \in \{0.001, 0.01, 0.1\}$
- Random Forest: depth ∈ {3, 5, 7}, estimators ∈ {100, 200}

The best model of each type is then refit on the full training set before evaluation on the test set.

III. Results

Using the chronological split at 1 March 2024, the training period spans 2020–2024 and the test period covers March 2024 through May 2025. The experimental results show that all three classifiers achieve performance close to random-chance baseline, which is expected in short-horizon equity return prediction.

A. Test Accuracy

The out-of-sample accuracies of the tuned models are:

Model	Accuracy
Logistic Regression	**0.50096**
SVM (RBF kernel)	**0.48051**
Random Forest	**0.50224**

Random Forest achieves the highest test-set accuracy, though only marginally above 50%. Logistic Regression performs similarly, while SVM underperforms slightly. These values indicate that next-day direction prediction remains extremely challenging and consistent with the near-martingale behavior of daily equity returns.

B. Confusion Matrix of Best Model

The Random Forest classifier achieves the following confusion matrix:

$$\begin{bmatrix} 199 & 614 \\ 165 & 587 \end{bmatrix}$$

Interpreting this:

- **True Negatives (199):** Correctly predicted down days
- **False Positives (614):** Predicted up but actual was down
- **False Negatives (165):** Predicted down but actual was up
- **True Positives (587):** Correctly predicted up days

The classifier shows a tendency to **overpredict upward movements**, leading to a higher number of false positives. This is common in equity datasets with asymmetric or noisy return distributions.

C. Interpretation

The results confirm that:

- Models capture **very limited predictive structure**, consistent with efficient-market behavior at the daily level.
- Random Forest detects slightly more signal than linear or margin-based methods, suggesting mild nonlinear relationships among lagged returns and volatility features.
- However, the magnitude of improvement is extremely small, indicating that short-term forecastability is minimal.

These findings align with the empirical finance literature: daily return direction for liquid equities is notoriously difficult to predict using only past prices and volumes.

IV. Conclusion

This project implemented a complete machine-learning pipeline to forecast next-day stock price direction for five U.S. steel-sector equities using Alpaca historical data from 2020–2025. After constructing lagged-return, trend, volatility, and volume-based features and tuning three classifiers with time-series cross-validation, we evaluated performance on a strictly future test window beginning 1 March 2024.

The empirical results show that **all models perform at or near random-chance accuracy**. Logistic Regression achieves 50.1% accuracy, SVM achieves 48.1%, and Random Forest achieves 50.2%. Although Random Forest technically attains the highest score, the difference relative to Logistic Regression is extremely small and not statistically meaningful. The confusion matrix further confirms that the model exhibits no reliable ability to separate up from down movements.

These findings are consistent with the well-documented difficulty of predicting daily stock returns using only past price and volume information. While nonlinear models can represent more complex decision boundaries, the underlying market signal at this frequency appears too weak and too noisy to extract with the features and sample sizes available. Future work could explore longer-horizon predictions, additional cross-asset signals (e.g., iron ore futures), or sequence models such as LSTMs.

References

[1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. Journal of Finance, 25(2), 383–417.

[2] Feng, G., He, Q., & Polson, N. G. (2019). Deep learning in asset pricing.

[3] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273.

[4] Kolm, P. N., Tütüncü, R., & Fabozzi, F. J. (2014). 60 years of portfolio optimization: Practical challenges and current trends. European Journal of Operational Research, 234(2), 356–371.

[5] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689–702.

[6] Lim, M. H., Zohren, S., & Roberts, S. J. (2021). Enhancing time-series momentum strategies using deep neural networks. Journal of Financial Data Science, 3(1), 28–44.

[7] Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. The Review of Financial Studies, 1(1), 41–66.

[8] Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. Quantitative Finance, 19(9), 1449–1459.

[9] Zhang, W., Zhang, J., Ma, Y., & Zhong, Y. (2020). Machine learning for portfolio optimization. Journal of Risk and Financial Management, 13(8), 180.