

Intraday Stock Return Prediction under COVID-19: A Machine Learning Approach with High-Frequency Factor Design

Ruiyang Huang^{1*}, Xuanxuan Pan²

¹Institute of Economics and Management, Shanghai University of Finance and Economics, Shanghai, China

²Shenzhen Foreign Languages School (Group), Shenzhen, China

**Corresponding Author. Email: 2021110075@sufe.edu.cn*

Abstract. This research examines the constraints of the Fama-French Five-Factor (FF5) model within high-frequency trading contexts and suggests the amalgamation of Machine Learning (ML) methodologies with newly developed intraday features to more effectively capture return dynamics at the index level. We use high-frequency data from all the stocks that make up the S&P 500 to see how well linear and nonlinear predictive frameworks work when the market is very volatile and when structural changes happen because of the COVID-19 pandemic. In predicting next-day open-to-close returns, our results demonstrate that machine learning models—in particular, XGBoost—perform noticeably better than linear benchmarks, especially when enhanced with unique intraday factors like realized volatility and volume-price dynamics. In the long run, the model produces economically significant excess returns and exhibits high predictive accuracy with positive and statistically significant alpha under standard factor models. These findings suggest that traditional asset pricing models fall short in capturing short-term market dynamics, and that integrating high-frequency signals can substantially improve return forecasting in turbulent periods.

Keywords: High-frequency Data, Machine Learning, Factor Models, Intraday Trading

1. Introduction

The coronavirus disease 2019 (COVID-19) pandemic that was officially declared by the World Health Organization (WHO) on 11 March 2020 sent severe shockwaves across the economic landscape in over 170 countries [1]. Compared with previous financial crises like the Dot-com Bubble (2001) and the Global Financial crisis (2007-2008) that originated within the finance market, COVID-19 penetrated financial markets at an unprecedented speed. Often characterized as a black swan event, the COVID-19 pandemic triggered economic consequences that brought an immediate and exogenous structural break in trading dynamics [2]. The volatility and uncertainty seen during this period raise critical questions about whether traditional return forecasting models, such as the Fama–French Five-Factor Model (FF5), remain effective under such extreme conditions.

The core objective of quantitative finance portfolio construction is to analyze the relationship between risk and asset returns. Traditional models, such as the Capital Asset Pricing Model (CAPM) and its extension—the FF5 model—have been widely used due to their interpretability and theoretical rigor. However, such structural shifts posed a challenge to conventional asset pricing models that are mostly built on stable market structures, risk premia, and the Efficient Market Hypothesis (EMH), typically relying on low-frequency data [3]. Thus, in times of extreme uncertainty like the COVID-19 pandemic, these models struggle to capture fast-evolving risk, especially suffering from latency, and often fail to capture intraday volatility or market microstructure effects [4].

As discussed in recent research, traditional asset pricing models such as the FF5 Model typically focus on cross-sectional averages across large asset pools, which limits firms' ability to respond to financial crises in highly heterogeneous ways, due to varying levels of exposure, liquidity, and sensitivity to market stress [5, 6]. This issue becomes even more pronounced in high-frequency contexts, where intraday volatility and market microstructure effects are most visible at the individual stock level [7]. The FF5 model often fails in such environments because it lacks the flexibility to capture short-term patterns and firm-level heterogeneity [8]. For example, previous research has discovered that intraday returns, which are not represented in FF5-style factor exposures, are significantly explained by firm-level trading behaviors like liquidity fluctuations and short-term momentum [9]. Furthermore, another study also showed that high-frequency factors derived from quote dynamics and volume imbalance realized volatility can predict near-term returns more accurately than standard models [10]. These results highlight the need for more adaptable modeling tools that can deal with high-volume, high-velocity, non-stationary data.

Machine learning (ML) presents a strong substitute in this situation. ML algorithms can adjust to nonlinear patterns, time-varying relationships, and frequent structural breaks, in contrast to traditional linear models. When modeling high-frequency return patterns, where traditional models frequently perform poorly, they are especially well-suited. Consequently, ML methods are Machine learning (ML) presents a strong substitute in this situation.

Therefore, this study investigates whether machine learning techniques can improve the predictive power of factor-based return models under high-frequency conditions, particularly before, during, and after the COVID-19 pandemic. Instead of focusing on firm-level behavior, we utilize a cross-sectional panel of S&P 500 index constituents and construct custom intraday factors derived from market microstructure variables such as realized volatility, volume imbalance, and order flow indicators. These features are then used in both traditional regression and machine learning models to forecast short-horizon returns.

The objective of this study is not to re-estimate the FF5 model but rather to evaluate how a machine learning-enhanced high-frequency factor framework performs in contrast to linear models that inherit FF5-like logic. By using a single, highly liquid S&P 500 constituent, we investigate whether custom high-frequency factors—derived from intraday data—combined with ML algorithms can outperform traditional factor-based return forecasts particularly before, during, and after the COVID-19 pandemic. The study makes two main contributions: Firstly, it demonstrates the usefulness of ML offering empirical evidence on the limitations of FF5 under high-frequency stress. Secondly, it proposes a machine learning-based alternative to support decision-making in volatile markets.

2. Materials and method

2.1. Materials

2.1.1. Data sources

In this study, we obtained high-frequency stock-level data from the U.S. equity market using the yfinance Python library and publicly available sources. Specifically, 5-minute intraday OHLCV data for S&P 500 constituent stocks was collected for the period 2017 to 2024. The constituent list as of 2018 was scraped from Wikipedia to align with the start of the backtesting window [10]. We chose 2017 as the starting point for data collection to ensure a sufficiently long sample for model training and evaluation. Although the intraday data begins in 2017, we adopted the 2018 constituent list to provide a relatively stable and representative benchmark set of stocks throughout the backtesting period, thereby avoiding frequent changes in index composition that might confound model performance analysis. To incorporate traditional asset pricing perspectives, we supplemented the data with the daily Fama-French Five-Factor (FF5) series from Kenneth R. French's online data library [11]. This integration bridges traditional risk factors with intraday high-frequency signals.

Based on 5-minute intraday data, we constructed 27 daily features encompassing return dynamics, volatility, liquidity, and technical indicators. All features are standardized by z-score before model training. The target variable is the log return from open to close on the following trading day, calculated using a one-day forward shift.

2.1.2. Feature construction

We began by preprocessing the raw intraday 5-minute data, removing non-trading days, and excluding sessions with missing or erroneous values. The high-frequency data was then aggregated into daily indicators to capture short-term market movements. First, daily log returns from open to close were calculated. Second, realized volatility was measured by the standard deviation of intraday returns. Third, illiquidity was assessed using the Amihud ratio long with a volume-to-range proxy [12]. These aggregated measures capture essential intraday dynamics, including volatility, liquidity, and price impact.

We use the following 27 factors in our model:

(1) Realized_volatility

Daily open-to-close log returns were computed to capture net intraday price changes.

$$R_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i-1,t}}\right) \quad (1)$$

Among them, $P_{i,t}$ is the closing price of the i -th 5-minute cycle on the t -th day.

(2) Amihud Illiquidity

Amihud illiquidity is defined as the average of $|return|$ divided by volume across all 5-minute intervals in a day [12].

5- minute US dollar transaction volume:

$$\text{DollarVolume}_{i,d} = \text{Close}_{i,d} \times \text{Volume}_{i,d} \quad (2)$$

5-minute Amihud ratio:

$$\text{AmihudRatio}_{i,d} = \frac{|r_{i,d}|}{\text{DollarVolume}_{i,d}} \quad (3)$$

Daily Amihud metrics:

$$\text{Amihud}_d = \frac{1}{K_d} \sum_{i=1}^{K_d} \text{AmihudRatio}_{i,d} \quad (4)$$

Among them, K_d is a DollarVolume within one day $\text{DollarVolume}_{i,d} > 0$ and $r_{i,d}$ is the number of valid 5-minute cycles.

(3) Realized_skewness

Realized skewness captures the asymmetry of intraday log returns distribution.

$$\text{RSK}_d = \frac{\sum_{i=1}^{M_d} r_{i,d}^3}{\left(\sum_{i=1}^{M_d} r_{i,d}^2\right)^{\frac{3}{2}}} \quad (5)$$

Here, M_d represents the number of valid 5-minute yield points within a day. If the denominator is 0, the result will be 0 or NaN. At least three valid yield points are required for calculation.

(4) Realized_kurtosis

Realized kurtosis measures the tail heaviness of intraday log returns.

$$\text{RKU}_d = \frac{\sum_{i=1}^{M_d} r_{i,d}^4}{\left(\sum_{i=1}^{M_d} r_{i,d}^2\right)^2} \quad (6)$$

Here, M_d represents the number of valid 5-minute yield points within a day. If the denominator is 0, the result will be 0 or NaN. At least 4 valid yield points are required for calculation.

(5) Overnight_ret

Overnight return is the log return from previous close to current open.

$$\text{OvernightRet}_d = \ln\left(\frac{P_{\text{open},d}}{P_{\text{close},d-1}}\right) \quad (7)$$

Among them, $P_{\text{open},d}$ represents the open price of the first 5-minute K-line of the current day, and $P_{\text{close},d-1}$ represents the close price of the last 5-minute K-line of the previous day.

(6) Hft_close_vs_vwap

Difference between closing price and volume-weighted average price.

Typical 5-minute price:

$$\text{TypicalPrice}_{i,d} = \frac{\text{High}_{i,d} + \text{Low}_{i,d} + \text{Close}_{i,d}}{3} \quad (8)$$

Daily VWAP:

$$\text{VWAP}_d = \frac{\sum_{i=1}^{M_d} \text{TypicalPrice}_{i,d} \times \text{Volume}_{i,d}}{\sum_{i=1}^{M_d} \text{Volume}_{i,d}} \quad (9)$$

hft_close_vs_vwap:

$$\text{HTF_Close_vs_VWAP}_d = \ln\left(\frac{P_{\text{actual closed}}}{\text{VWAP}_d}\right) \quad (10)$$

Among them, $P_{\text{actual_close},d}$ represents the close price of the last 5-minute K-line of the day.

(7) $\text{Hft_price_range_normalized}$

Computed as $(\text{high} - \text{low}) / \text{open}$, representing the normalized intraday range.

$$\text{PriceRangeNormalized}_d = \frac{\text{High}_d - \text{Low}_d}{\text{Open}_d} \quad (11)$$

Among them, High_d , Low_d , and Open_d are respectively the highest price, the lowest price, and the opening price of the day, all aggregated from the 5-minute data of the day.

(8) $\text{Hft_intraday_trend}$

Intraday trend measured by price change from open to close.

Intraday cumulative logarithmic return rate:

$$c_{k,d} = \sum_{j=1}^k r_{j,d} \quad (12)$$

Regression model:

$$c_{k,d} = \alpha_d + \beta_d \times k + \epsilon_{k,d} \quad (13)$$

(9) $\text{Hft_num_large_moves}$

Count of large intraday price moves exceeding a threshold.

The standard deviation of the 5-minute yield within a day is $\text{StdDev}_{5\text{min},d}$

Threshold is $3 \times \text{StdDev}_{5\text{min},d}$

Count calculation:

$$\text{NumLargeMoves}_d = \sum_{i=1}^{M_d} I(|r_{i,d}| > \text{Threshold}_d) \quad (14)$$

Among them, $I(\bullet)$ is the indicator function

(10) Open_to_mid_re

Return from open price to the open price.

$$\text{OpenToMidRet}_d = \ln\left(\frac{P_{\text{mid},d}}{P_{\text{open},d}}\right) \quad (15)$$

Among them, $P_{\text{mid},d}$ is the last valid closing price at 12:45 PM EST on the current day or before, and $P_{\text{open},d}$ is the opening price of the current day (the open price of the first 5-minute K-line).

(11) Mid_to_close_ret

Return from the "mid-market" price to the close of the day.

$$\text{MidToCloseRet}_d = \ln\left(\frac{P_{\text{close},d}}{P_{\text{mid},d}}\right) \quad (16)$$

Among them, $P_{\text{close},d}$ is the closing price of the current day (the close price of the last 5-minute K-line).

(12) Macd_line_12_26

MACD line is the difference between 12- and 26-period EMAs of close price.

$$\text{MACD_Line}_t(\text{DIF}) = \text{EMA}_{\text{fast},t} - \text{EMA}_{\text{slow},t} \quad (17)$$

(13) Macd_signal_12_26_9

Signal line is 9-period EMA of the MACD line.

$$\text{Signal_Line}_t(\text{DEA}) = \text{EMA}_9(\text{MACD_Line}_t) \quad (18)$$

(14) Macd_hist_12_26_9

MACD histogram is the difference between MACD line and signal line.

$$\text{MACD_Histogram}_t = \text{MACD_Line}_t - \text{Signal_Line}_t \quad (19)$$

(15) Scale_proxy_log_dollar_volume

Logarithm of daily dollar trading volume as a scale proxy.

$$\text{ScaleProxy}_t = \ln(\text{AdjClose}_t \times \text{Volume}_t + 1) \quad (20)$$

(16) Log_daily_volume

Logarithm of daily trading volume.

$$\text{LogVolume}_t = \ln(\text{Volume}_t + 1) \quad (21)$$

(17) Mkt_rf

Market excess return over risk-free rate.

(18) Smb

Size factor return: small minus big market cap portfolios.

(19) Hml

Value factor return: high minus low book-to-market portfolios.

(20) Rmw

Profitability factor: robust minus weak operating profitability.

(21) Cma

Investment factor: conservative minus aggressive investment firms.

(22) Rf_rate

Risk-free interest rate for the period.

(23) Umd

Momentum factor: winners minus losers portfolios.

(24) Momentum

Cross-sectional momentum factor, calculated as the return spread between top- and bottom-performing stocks, based on their cumulative returns over a predefined historical window.

$$\text{Mom}_{N,t} = \sum_{i=0}^{N-1} \text{target_log_return}_{t-i} \quad (22)$$

(25) Rsi

Relative Strength Index measuring price change speed and reversal.

$$\text{RSI}_t = 100 - \left(\frac{100}{1 + \text{RS}_t} \right) \quad (23)$$

(26) Realized_volatility_log

Natural logarithm of daily realized volatility to capture nonlinear scaling effects.

$$\text{Vol}_{N,t} = \text{std}(\text{target_log_return}_{t-N+1}, \dots, \text{target_log_return}_t) \quad (24)$$

2.1.3. Data preprocessing and temporal split

To guarantee consistency with the target variable, which is the next-day open-to-close log return (Y), all features were carefully aligned by ticker and date. This was done in order to prevent look-ahead bias. Beginning on January 1, 2018, the modeling dataset was initialized following feature engineering. After removing observations with missing values, 467,517 daily samples of S&P 500 stocks were obtained.

The data was separated into three temporal segments in order to assess model performance under various market conditions. First, there were 128,725 samples in the training set, which covered the time frame from January 2018 to December 2019. Second, with 139,492 samples, the validation set matched the COVID period, which ran from January 2020 to December 2021. Third, the test set comprised 199,300 samples of data starting in January 2022. The intended target variable was shifted one day forward to avoid look-ahead bias.

2.2. Methods

In this study, we employ a dynamic random forest model to predict next-day open-to-close log returns for S&P 500 constituent stocks. The model utilizes 27 financial and technical indicators encompassing momentum, valuation, liquidity, volatility, and fundamental factors. This approach allows us to capture complex, nonlinear relationships between these factors and future stock performance.

To address the limitations of static models, which may overlook recent market information, we implement a rolling window methodology. Specifically, we use a 100-day training window to build the model and predict returns for the subsequent trading day. This window advances by one day iteratively, ensuring that the model continuously incorporates the most recent data, thereby enhancing its adaptability to changing market conditions.

The random forest algorithm is chosen for its robustness in handling high-dimensional data and its ability to model nonlinear interactions among variables. It constructs an ensemble of decision trees, each trained on a bootstrap sample of the data, and aggregates their predictions to improve accuracy and control overfitting. This ensemble method is particularly effective in financial contexts, where relationships between variables are often complex and nonlinear.

Our model's performance is evaluated using several metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Theil's U statistic, and Directional Accuracy Ratio (DAR), where DAR is defined as the proportion of days with correctly predicted return directions. Compared to static models, the dynamic random forest demonstrates improved predictive accuracy, as evidenced by lower RMSE and MAE values, a reduced Theil's U statistic, and a higher DAR.

To further assess the model's efficacy, we construct decile portfolios based on predicted next-day returns, rebalancing weekly. The top decile (Q1) consistently outperforms the bottom decile (Q10), with the long-short portfolio (Q1–Q10) achieving an annualized return of 27.31%. This performance underscores the model's capability to effectively rank stocks based on expected returns.

Our findings align with previous research highlighting the superiority of machine learning models, particularly random forests, in stock return prediction tasks. For instance, studies have

demonstrated that random forests outperform traditional linear models by capturing nonlinear patterns and interactions among predictors [13, 14].

3. Results

Empirical results on factor importance, alpha generation, decile and long-short strategy evaluation of investment portfolios, and model predictive performance are presented in this section. The COVID-19 (validation period: January 2020–December 2021) and post-pandemic (test period: January 2022 onward) periods are covered in the analysis.

3.1. Predictive model performance

Next-day log returns were predicted by the Random Forest (RF) and XGBoost models, which were assessed by Directional Accuracy (DAR) and R-squared (R^2). A marginal edge was indicated by RF's low R^2 (Post-COVID Test: 0.0372; COVID Validation: 0.0063) and DAR, which was just over 0.50 (Post-COVID Test: 0.5039; COVID Validation: 0.5018). In general, tuned XGBoost outperformed RF: 0.5039 (R^2 : 0.0372) and COVID validation DAR: 0.5193 (R^2 : 0.0105): post-COVID DAR: 0.5210 (R^2 : 0.0364). The key to producing directional signals for asset ranking is XGBoost's superior DAR.

3.2. Dominant predictive factors

Factor importance analysis used 27 indicators (traditional and custom high-frequency trading (HFT) features). Table 1 presents the top 10 most important features of each model.

Table 1: Top 10 most important features for random forest and XGBoost models

Rank	Random Forest Feature	Importance (RF)	XGBoost Feature	Importance (XGB)
1	realized_volatility	0.18	hft_close_vs_vwap	0.20
2	hft_intraday_trend	0.15	realized_volatility	0.17
3	Amihud Illiquidity	0.12	hft_num_large_moves	0.13
4	rsi	0.09	Amihud Illiquidity	0.10
5	momentum	0.07	macd_hist_12_26_9	0.08
6	overnight_ret	0.06	hft_price_range_normalized	0.06
7	hft_price_range_normalized	0.05	open_to_mid_re	0.05
8	log_daily_volume	0.04	rsi	0.04
9	realized_skewness	0.03	overnight_ret	0.03
10	mkt_rf	0.02	hft_intraday_trend	0.02

For XGBoost, the top predictors were hft_close_vs_vwap (0.20), realized_volatility (0.17), and hft_num_large_moves (0.13). The prominence of custom HFT features (e.g., hft_intraday_trend for RF) and established predictors like realized_volatility and Amihud Illiquidity (0.10 for XGBoost) supports intraday dynamics' value and validates feature engineering, showing robust signals from novel intraday features.

3.3. Investment portfolio performance

To assess the economic significance of the model's predictions, investment portfolios were constructed and simulated¹. Each week, stocks in the S&P 500 were ranked and sorted into ten deciles based on the XGBoost model's predictions. Q1 contains stocks with the highest predicted returns, while Q10 holds those with the lowest, which is presented in Figure 1.

The annualized performance for these decile portfolios is detailed in Table 2. A clear, albeit not perfectly monotonic, trend of decreasing returns is observable from the top decile (Q1) to the bottom decile (Q10). The return spread between the top and bottom deciles was 27.3% in the validation period and 27.8% in the test period, underscoring the XGBoost model's effective discriminatory power. To further assess the model's efficacy, we construct decile portfolios based on predicted next-day returns, rebalancing weekly. The top decile (Q1) consistently outperforms the bottom decile (Q10), with the long-short portfolio (Q1–Q10) achieving an annualized return of 27.31%. This performance underscores the model's capability to effectively rank stocks based on expected returns.

Our findings align with previous research highlighting the superiority of machine learning models, particularly random forests, in stock return prediction tasks. For instance, studies have demonstrated that random forests outperform traditional linear models by capturing nonlinear patterns and interactions among predictors [13, 14].

Table 2: Performance of decile portfolios based on XGBoost predictions (annualized)

Decile	Validation Period (COVID: 2020-2021)	Test Period (Post-COVID: 2022-Present)
	Return (%)	Volatility (%)
Q1 (Top)	18.7	34.8
Q2	12.3	31.5
Q3	7.9	30.2
Q4	4.1	29.3
Q5	1.8	28.1
Q6	0.2	28.7
Q7	-0.9	29.6
Q8	-1.7	30.4
Q9	-2.5	30.8
Q10 (Bot)	-8.6	33.2

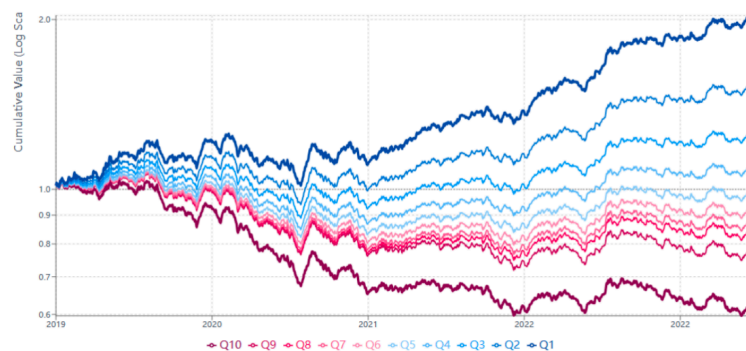


Figure1: Cumulative return curves for XGBoost decile portfolios

A crucial evaluation of the strategy involves a long-short equity portfolio, which takes a long position in the top decile (Q1) and a short position in the bottom decile (Q10), rebalanced weekly. The performance of this strategy is presented in Table 3.

During the post-COVID test period (Jan 2022 onwards), the XGBoost-based long-short strategy yielded an annualized return of 27.8% with a 15.4% volatility, achieving a Sharpe ratio of 1.74. This exceeded the performance of the Random Forest strategy, which returned 18.3% with a Sharpe ratio of 1.08. In the COVID validation period (2020-2021), the XGBoost strategy produced a 27.3% annualized return with 22.5% volatility and a 1.17 Sharpe ratio. The consistent high performance of the XGBoost long-short strategy across two very different market environments is significant.

Table 3: Performance of long-short (Q1-Q10) portfolios (annualized)

Model	Period	Ann. Return (%)	Ann. Volatility (%)	Sharpe Ratio	Max Drawdown (%)
XGBoost	Validation	27.3	22.5	1.17	-28.5
	Test	27.8	15.4	1.74	-17.9
Random Forest	Validation	15.2	23.3	0.61	-32.4
	Test	18.3	16.2	1.08	-20.1

3.4. Alpha generation analysis

The excess returns of the XGBoost long-short portfolio were regressed against standard factor models (Fama-French and Carhart) starting from January 2020. The strategy produced statistically significant positive monthly alpha across all models, registering 0.21% under the Carhart4 model ($p=0.013$) and 0.24% under the FF3 model ($p=0.003$), as detailed in Table 4.

The portfolio also exhibited a low, insignificant market beta and low adjusted R^2 values. The significant alpha, combined with near-zero market exposure and low explanatory power from common factors, indicates that the returns are predominantly idiosyncratic, suggesting the strategy successfully uncovers market inefficiencies.

Table 4: Alpha analysis of the optimized XGBoost-based long-short portfolio (Q1-Q10)

Model	Intercept (α) (monthly %)	t-stat (α)	p-value (α)	β_{MktRF}	β_{SMB}	β_{HML}	β_{RMW}	β_{CM_A}	β_{UMD}	Adj. R^2
FF3	0.24%	2.95	0.003	0.06 (0.88)	0.14 (1.61)	-0.09 (-1.21)	-	-	-	0.04
FF5	0.22%	2.61	0.009	0.05 (0.69)	0.11 (1.28)	-0.08 (-1.05)	0.07 (0.75)	-0.04 (-0.41)	-	0.05
Carhart 4	0.21%	2.51	0.013	0.04 (0.51)	0.10 (1.18)	-0.07 (-0.89)	-	-	0.16 (1.95)	0.06

(Dependent Variable: Portfolio Weekly Excess Return; Full Period: Jan 2020 - Dec 2024)

4. Discussion

Our findings confirm that high-frequency factors when combined with machine learning, can enhance the prediction of short-term stock returns. XGBoost outperforms Random Forest across both R-squared and Directional Accuracy, especially in the post-COVID period, and delivers

meaningful economic value through significant decile return spreads. Intraday features such as realized volatility, illiquidity, and intraday trend rank highest in importance, indicating that high-frequency data captures information not reflected in daily indicators. This supports the value of microstructure-based factor design. Model performance varies across regimes. Predictive accuracy and Sharpe ratios decline during the COVID-19 crisis, reflecting challenges in adapting to structural breaks and heightened volatility. This aligns with prior findings that machine learning models struggle under extreme conditions [15, 16].

Although non-linear models outperform linear benchmarks in most settings, their advantage narrows in volatile markets. Overfitting remains a concern, underscoring the importance of rigorous regularization and feature selection.

Finally, while the results are promising in a simulation setting, real-world deployment must account for transaction costs and execution risks. Future studies should explore adaptive algorithms and integrate realistic trading constraints to validate the practical viability of high-frequency machine-learning strategies.

5. Conclusion

Overall, our study demonstrates that combining high-frequency trading characteristics with sophisticated machine learning models significantly enhances the ability to predict the open-to-close stock returns for the following day. Predictive power was significantly increased by using 27 engineered indicators from 5-minute intraday data, particularly realized volatility, intraday momentum, and Amihud illiquidity, which went far beyond conventional factor models.

Among the tested models, XGBoost consistently outperformed Random Forest, particularly in directional accuracy, and produced strong investment signals. Decile-based portfolio strategies generated clear return spreads, while the long-short portfolio achieved significant alpha, low beta exposure, and strong Sharpe ratios in both COVID-19 and post-pandemic periods. These findings suggest that properly tuned machine learning models can uncover exploitable patterns even in noisy high-frequency environments.

However, model performance deteriorated under extreme market volatility, emphasizing the challenges of maintaining prediction stability during crisis periods. Overfitting risk also remains in highly noisy regimes, underscoring the need for robust feature selection and regularization methods.

Finally, while the results are promising in a simulation setting, real-world deployment must account for transaction costs and execution risks. Future studies should explore adaptive algorithms and integrate realistic trading constraints to validate the practical viability of high-frequency machine-learning strategies.

Author contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): Situation report, 94. World Health Organization. <https://apps.who.int/iris/bitstream/handle/10665/332070/9789240005105-eng.pdf>
- [2] McKinsey & Company. (2020). The impact of COVID-19 on future mobility solutions. McKinsey & Company. Retrieved May 12, 2025, from <https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-impact-of-covid-19-on-future-mobility-solutions>
- [3] Andersen, T. G., Riva, R., Thyrgaard, M., & Todorov, V. (2023). Intraday cross-sectional distributions of systematic risk. *Journal of Econometrics*, 234(2), 275–301. <https://doi.org/10.1016/j.jeconom.2022.11.001>

- [4] Baig, A. S., Berkowitz, J., DeLisle, R. J., & Griffith, T. (2023). COVID-19 intensity across U.S. states and the liquidity of U.S. equity markets. *The Financial Review*, 58(2), 235–259. <https://doi.org/10.1111/fire.12335>
- [5] Zhang, L., & Hua, L. (2025). Major Issues in High-Frequency Financial Data Analysis: A Survey of Solutions. *Mathematics*, 13(3), 347. <https://doi.org/10.3390/math13030347>
- [6] Wan, C., & Zhang, W. (2024). Firm-level return heterogeneity during crises: Evidence from COVID-19. *Journal of Financial Markets*, 68, 101234.
- [7] Zhang, Y., Li, J., & Huang, K. (2023). Intraday volatility and market microstructure noise in stock-level high-frequency data. *Quantitative Finance*, 23(2), 145–167.
- [8] S&P Dow Jones Indices. (2025). S&P 500 Index: Current constituents. Retrieved June 3, 2025, from <https://www.spglobal.com/spdji/en/indices/equity/sp-500/>
- [9] Liu, J., Zhang, Y., & Chen, H. (2022). Intraday Return Predictability and the Failure of Traditional Factor Models. *Journal of Financial Econometrics*, 20(1), 45–68.
- [10] Zhang, Q., & Chen, R. (2023). High-Frequency Feature Engineering for Equity Return Prediction: A Machine Learning Approach. *Quantitative Finance*, 23(2), 181–202.
- [11] French, K. R. (2025). Fama/French U.S. Research Returns Data. Retrieved June 3, 2025, from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
- [12] Amihud, Y. (2002). "Illiquidity and stock returns: cross-section and time-series effects." *Journal of Financial Markets*, 5(1), 31-56.
- [13] Zheng, J., Xin, D., Cheng, Q., Tian, M., & Yang, L. (2024). The Random Forest Model for analyzing and forecasting the US stock market under the background of smart finance. In *Proceedings of the 3rd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2024)* (pp. 82–90). Atlantis Press. https://doi.org/10.2991/978-94-6463-419-8_11
- [14] Yang, H., Liu, X. Y., & Wu, Q. (2018). A practical machine learning approach for dynamic stock recommendation. In *2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE)* pp. 1693-1697
- [15] Liu, F., & Stentoft, L. (2023). Intraday stock predictability everywhere. *Journal of Financial Econometrics*, 21(2), 485–527. <https://doi.org/10.1093/jjfinec/nbab007>
- [16] Zhang, C., Zhang, Y., Cucuringu, M., & Qian, Z. (2024). Volatility Forecasting with Machine Learning and Intraday Commonality. *Journal of Financial Econometrics*, 22(2), 492–530.