Can Machines Learn from Corporate Insiders?

Tongyang Kong Chishen Wei Lei Zhang

November 14, 2025

Abstract

Using publicly available corporate insider trades from SEC Form 4 filings, we train a gradient-boosted decision tree model on a comprehensive set of trading records to generate a machine learning signal (MLS) to predict future stock returns. The long-short portfolio earns alphas of over 1.06% per month and a Sharpe ratio of 1.01 that is not explained by standard risk factors. Our machine-generated signals provides significant incremental information beyond human-generated insider trading signals in the existing literature. MLS positively predicts future earnings reactions and exhibits stronger predictive ability in hard-to-value stocks, stocks without managerial guidance, and more conservative financial reporting. Our findings demonstrate that machine learning can extract economically significant signals from trade data that are orthogonal to traditional human-generated signals.

Keywords: Machine learning, insider trading, firm complexity, managerial guidance, financial reporting
JEL Classification Codes: G11, G12, G14

*Kong (t.kong@my.cityu.edu.hk): City University of Hong Kong; Renmin University of China. Wei (chishen.wei@polyu.edu.hk): Faculty of Business, Hong Kong Polytechnic University. Zhang (lzhan29@cityu.edu.hk): City University of Hong Kong. We thank Weikai Li, Alberto Rossi, K.C. John Wei, Jia Zhai (discussant), and seminar participants at the CAFRE conference and HKPU for helpful comments. All remaining errors are our own.

1 Introduction

The accounting and financial services sector is undergoing a technological revolution with the integration of artificial intelligence (AI) and machine learning (ML) algorithms. These are transformative tools for accessing and processing information, and their capabilities now extend to more sophisticated applications such as regulatory compliance, financial analysis, and credit risk assessment. These rapid advancements in synthesizing information raise a natural question: Can machines generate *undiscovered* trading signals directly from raw data that are informative and distinct from human intelligence? Financial markets provide a challenging test setting because they are adaptive and quickly integrate new information. Consequently, the capabilities of AI in this area are not guaranteed.

This paper tests the hypothesis that machine learning can analyze trading data with minimal human intervention to uncover signals embedded in the data. We instruct a gradient-boosted decision tree (GBDT) model, a flexible nonparametric algorithm, to learn the complex mapping from these trading characteristics to future stock returns. The model is trained out-of-sample to synthesize these features into a single predicted return that we call the Machine Learned Signal (MLS). Our objective is to assess whether machines can move beyond human reasoning and prespecified heuristics to let the data speak for itself. To evaluate machine against human intelligence, we horse-race MLS against established insider trading signals in the existing literature.

To perform our exercise, we require trading data that are readily accessible for researchers to analyze. We select publicly disclosed trades from U.S. corporate insiders (executives, directors, and large shareholders) of publicly traded companies. There are several advantages of these trading data. First, the data are comprehensive. The SEC requires all insiders to disclose their trade by filing Form 4 whenever there is a material change in their holdings of the company's shares. This mitigates concerns that the trading data might be a selective sample or possibly not representative of typical trades. Second,

Form 4 filings are free to download from the Security and Exchange Commission (SEC) EDGAR system and have been digitalized in recent years. This transparency and open availability ensure that the data are available to researchers without restriction. Third, the data are continuously updated each trading day, so that analysis can continue after our initial sample period. Fourth, a policy change in 2002 requires insiders to disclose within two business days. Hence, we can design real-time trading strategies to exploit valuable signals that are potentially embedded in the trading data.

Importantly, for our research question, the information content of insider trades has been intensively studied since Jaffe (1974). Thus, we have 50 year corpus of presumably human-generated research that examines whether insiders, with their privileged access to non-public information, can systematically outperform the market. Insiders may trade for a multitude of reasons, many of which are not related to private information. Purchases may be part of pre-arranged compensation plans, and sales are often motivated by portfolio diversification, liquidity needs, or tax planning. Researchers have developed astute methods to filter the truly informed trades from liquidity-driven transactions. Our results may also inform regulation and firm policies around insider trading. Although insider trading has been regulated for more than 90 years since the Securities Exchange Act of 1934, insider trading regulations are constantly under scrutiny and were amended as recently as 2023 (Kim, Kim, and Rajgopal, 2025).

To implement the GBDT algorithm, we first construct a set of seven features for each insider i in month t, capturing the direction, size, and economic significance of trades, as well as the trading history of the insider involved. The three transaction-related variables are the change in the shares of the net transactions as a percentage of shares outstanding, the change in the dollar value of the net transactions, and the number of transactions. The three variables relating to insider i's trading history are the estimated return of transactions conducted in the prior 3, prior 6, and prior 12 months. The seventh variable is the resulting shares held as a percentage of shares outstanding after all transactions

are conducted in the month. We intentionally keep our set of features simple to avoid imparting human intervention. In robustness tests, we consider alternative data structures and find that our inferences are not sensitive to our methodological choices.

The GBDT algorithm is among the best machine learning techniques for forecasting stock returns (Gu, Kelly, and Xiu, 2020). Through its step-wise function, the algorithm manages common issues in trading data such as missing values and outliers. We utilize the latest available open-source version, light gradient-boosting machine (LightGBM), created by Microsoft Research. We use an expanding window approach to generate predicted returns and evaluate the out-of-sample performance. For the training period, we require a two-year estimation window to generate the GBDT prediction model for the next month return. After training is complete, we use the subsequent month as the test asset by applying the GBDT model to predict the future month return. We repeat the training-testing process with an expanding window for each successive month such that our predict returns start from January 2002 and end in December 2023. As multiple insiders i can trade the same stock j in month t, we compute the equal weighted average of each insider's predicted return to construct the insider trading signal (MLS) for each stock-month $\hat{Ret}_{j,t+1}^{MLS}$.

We start by forming long-short portfolios that buy stocks with high $\hat{Ret}_{j,t+1}^{MLS}$ and short stocks with low $\hat{Ret}_{j,t+1}^{MLS}$. The long-short portfolio generates monthly returns of 1.06% (t=3.43) and a Sharpe ratio of 1.01 using equal-weighting. The performance exists among small and medium market capitalization stocks, and is non-existent among large cap stocks. We find similar results using out-of-sample monthly alphas computed with respect to the Fama and French (2015) five-factors plus the momentum factor(1.10%, t=5.21) and Q-factor plus momentum factor (1.29, t=5.82) models. The resulting estimates from Fama-MacBeth regression indicate that the top quintile portfolio earns 0.97% per month (t=7.24) after controlling for firm and stock characteristics. The results are driven by the highest decile, with no significant results on the short side. It is possible that insider sales are likely

confounded by liquidity trades such that it is more difficult for the machine to separate signal from noise.

To horse-race the MLS machine-generated signal against human intelligence, we estimate a series of Fama and MacBeth (1973) regressions to compare the predictive power of MLS against existing measures in the literature. We find that the MLS measure is orthogonal to the following measures: 1) the insider trade imbalance in Lakonishok and Lee (2001), 2) the non-routine measure in Cohen, Malloy, and Pomorski (2012), 3) pre-earnings trading profitability in Ali and Hirshleifer (2017), and 4) the cancelation of routine trades in Hong and Li (2019). The inclusion of these measures does not materially weaken the machine-generated signal as the top quintile portfolio continues to earn 0.90% (t=6.60) per month.

We investigate the statistical and economic underpinnings of the MLS measure. First, we "open the black box" to understand what the machine has learned. Feature importance analysis reveals that the model prioritizes the economic value of the insider trade measured by the net trading percentage and the dollar value. Partial dependence plots report nonlinear step-function-like relationships, confirming that MLS derives its power from capturing patterns that linear models would likely miss. Decision tree plots are another method to visualize interaction effects in ML models. We find that MLS is particularly high in instances when 1) net trading % is high and 2) recent trading performance was poor. An interpretation of this scenario is that as an insider purchase trade is a strong signal of future returns if their current trade is large but their past trading performance was not particularly profitable, suggesting that they have high confidence in this current trade. Finally, we evaluate whether the effectiveness of MLS is dominated by a single variable, however, $\hat{Ret}^{MLS}_{j,t+1}$ remains a predictor of future returns across all specifications, even when each variable is individually omitted.

Next, we aim to better understand how MLS relates to firm fundamentals and financial reporting. We begin by examining the subsequent earnings announcement because it

releases new fundamental information to investors. The results indicate that *MLS buy* predicts future earnings reactions, suggesting that *MLS* detects trades where insiders have better information than outside investors. Therefore, we hypothesize and test the role of the information environment on the predictive ability of *MLS*.

First, we consider fundamental business operations such as operational complexity and high R&D that make the firm hard-to-value for investors. Our evidence suggests that the predicted returns associated with *MLS* are nearly twice as large in complex firms (Loughran and McDonald, 2024) and high R&D firms (Aboody and Lev, 2000) compared to their low complexity and low R&D counterparts. Second, we consider voluntary disclosure using managerial forecasts. *MLS* predicts much larger future returns in the sample of firms without manager guidance compared to the sample with managerial guidance, which implies that an information void may provide insiders with an opportunity to trade profitability. Third, to evaluate the importance of financial reporting, we choose accounting conservatism because it reflects the standard setting principle of recognizing bad news earlier. A wedge between insiders and outside investors can arise if investors cannot unwind the conservatism inherent in financial reporting. Consistent with this view, the *MLS* measure performs best for firms with the highest levels of accounting conservatism.

As described earlier, insider data must be disclosed within two business days according to SEC rules. Therefore, we can design a real-time trading strategy as the data is reported. Occasionally, insider trades are reported with delays due to technical issues or human error. We repeat our analysis with two modifications. First, we sort the data so that SEC file date < month=0. Second, we estimate a skip-a-month strategy such that use the *MLS* measure from the prior month $(t-1)$ to predict the future month's return $(t+1)$. We continue to find that *MLS* predicts future returns using these alternative measures.

We conclude by performing a series of additional tests. First, we evaluate and find that other types of ML algorithms such as Ridge, Lasso, Random forest, and neural networks also predict future returns based on our set of insider trading features, but the predictive

power of these models is weaker using equal-weighted portfolios and mostly non-existent using value-weighted portfolios. Second, we separately create MLS measures based on the insider's position in the company such as senior management, directors, and independent directors and continue to find that the MLS measure predicts future returns using all types of insiders. One interpretation is that the information content is general, not concentrated in the C-suite. Third, we find that the predictive power of MLS concentrates among non-10b5-1 rule trades (Fich, Parrino, and Tran, 2023).

A growing literature examines the ability of AI and ML technologies to perform fundamental valuation (Chen, Cho, Dou, and Lev, 2022; Geertsema and Lu, 2023; Jones, Moser, and Wieland, 2023). Studies show that 'AI analysts' are valuable (Van Binsbergen, Han, and Lopez-Lira, 2023; deHaan, Lee, Liu, and Noh, 2025) and complement human intelligence in the context of equity research (Grennan and Michaely, 2021; Cao, Jiang, Wang, and Yang, 2024). Building on this literature, we investigate whether ML can uncover signals in a trading database that has already been extensively analyzed by human experts. Our findings suggest that ML is capable of detecting patterns, possibly due to its ability to model complex non-linear relationships, that may elude human analysts.[1]

Studies show that new ML algorithms are useful in broader accounting practices such as predicting misreporting (Brown, Crowley, and Elliott, 2020) and future tax rates(Guenther, Peterson, Searcy, and Williams, 2023). These finding have implications for rule-setting and regulatory actions. Similarly, our results may also help evaluate policies for insider trading. Although insider trading has been regulated since the Securities Exchange Act of 1934, insider trading regulations are constantly under scrutiny and were amended as recently as 2023 (Kim et al., 2025). Using ML methods, we discover profitable trading signals in insider purchases that are beyond those found by human researchers. Our results are potentially useful in informing future insider trading regulations.

[1]Our study shares methodological similarities with Bogousslavsky, Fos, and Muravyev (2024), who use ML to predict informed trading primarily from Schedule 13D activist trades. Our study uses trading records from SEC Form 4 insiders to ask a different question regarding trading signals embedded in these insider trades.

2 Data and methodology

2.1 Data

We collect data from various sources. Our primary data on insider trades are drawn from the Thomson Reuters Insider Filing Data Feed, which includes all trades by corporate insiders reported on SEC Form 4.[2] The U.S. Securities and Exchange Commission (SEC) mandates that all officers and directors, large shareholders (those who own 10% or more of the outstanding shares), and affiliated shareholders report their transactions to the SEC within 10 days after the end of the transaction month. This deadline was changed to two days in 2002. The dataset contains the name and position(s) of each insider, the transaction date, the transaction price and quantity, and the date the filing was received by the SEC. The sample period of our main analysis is January 2002 to November 2023, so we collect insider trading data from 2000 to ensure at least 24 months of data to generate the insider trading signal measure. Thomson data: 2000 2023. The sample consists of 1,482,452 filings containing 4,457,504 trades, such that 593,082 (40.01%) filings contain more than one trade.

We obtain trading data for US common share stocks (with a share code of 10 or 11) listed on the NYSE, AMEX, and NASDAQ from the Center for Research in Securities Prices (CRSP). The accounting variables and earnings announcement data are obtained from Compustat. The monthly stock-level anomaly data for US stocks are obtained from Open Source Asset Pricing.[3] We obtain Fama-French factors and the momentum factor from WRDS and Hou, Xue, and Zhang (2015) Q-factors from Lu Zhang's website.[4]

All variables definitions are available in Appendix. The test assets consists of all common stocks from January 2002 to November 2023, with price above $1 at the end of each month, excluding stocks with a negative book value of equity.

[2]We exclude records with a cleanse code of "S" or "A." And we focus on open market purchases and sales with a trancode of "P" or "S."

[3]See https://www.openassetpricing.com/data/.

[4]The q-factors can be downloaded from http://global-q.org/index.html.

2.2 *Insider trading signal measure construction*

In this section, we describe how we construct the machine learning signal measure using insider trading filings. Traditional research often relies on linear regression to explore the relationships between variables. Although linear regression is useful for understanding basic relationships, it has limitations in capturing the complex interactions and nonlinearities that may exist between variables. In the context of our insider trading dataset, the underlying relationships are likely complex, including nonlinearities and variable interactions that cannot be fully explained by linear regression. For example, factors such as trading volume or the trading times could have predictive power conditional on past trading history such that they jointly influence the power of the signal. These interactions suggest that more flexible methods are promising.

We consider several commonly used machine learning methods. Lasso is a popular technique in economics, as it can shrink coefficients of irrelevant features to zero. Lasso is easy to interpret, but it may struggle with highly correlated variables and can be less effective in capturing complex relationships. Decision tree models model nonlinear relationships and variable interactions. Random Forest is an ensemble method that takes an average over many random decision trees. It is robust and less prone to overfitting, but it does not perform as well on regression tasks, where its predictive power is somewhat limited.

Another ensemble technique is boosting, which builds each new tree to correct the errors of previous ones. Boosted Regression Trees can process large, high-dimensional datasets without overfitting, producing more accurate forecasts, are also robust to missing values and outliers (Hastie, Friedman, and Tibshirani, 2009). XGBoost is a gradient boosting algorithm known for its strong predictive performance (Chen and Guestrin, 2016), but requires careful tuning, and can be computationally costly and slow for large datasets. Our preferred method is LightGBM, an efficient gradient boosting framework that builds on the BRT (Ke et al., 2017). It is designed for large-scale machine learning

tasks, leveraging techniques such as histogram-based decision tree learning and leaf-wise tree growth to accelerate training. LightGBM requires less training time and lower memory usage compared to other BRT implementations like XGBoost. We also consider a neural network model. Neural networks are flexible and powerful tools capable of capturing complex, nonlinear relationships in data. Our dataset contains only a few features, therefore the neural network may not be the most suitable choice for this task. The limited number of inputs constrains the model's ability to learn rich patterns, and may lead to overfitting or unstable results. Our baseline analysis focuses on LightGBM because it is the most suitable algorithm for our dataset and yields the best performance.

We select seven features as described in the Appendix. Three features are current transaction characteristics: net change in shares as a percentage of shares outstanding ($\%Nettrade$), the net dollar value of transactions ($\$Nettrade$), and the net number of transactions ($\#Nettrade$). Three features are related to the past performance of insiders: the profitability of transactions in the prior 3, 6, and 12 months ($EstRet_{m-3}$, $EstRet_{t-6}$, $EstRet_{t-12}$).[5] The last is the resulting shareholding of insiders as a percentage of shares outstanding ($NetOwnership\%$).

To design the data structure for our analysis, we evaluate several factors. The most straightforward method would be to use each individual transaction from the complete set of insider trades. However, over 40% of filings contain more than one trade because brokers typically separate the insider's trade into smaller sub-trades for better liquidity and to avoid price impact. Alternatively, we considered structuring the data at the filing-level, which would be conducive to an event-study approach. However, most existing studies use a portfolio approach at the monthly frequency to more precisely adjust for risk. To be comparable with these studies, we choose to aggregate each insider's trade

[5]The profitability of insider j's transactions of firm i in month t is calculated as: $EstRet_{j,t} = Return_{i,t+1} * Direction_{j,t}$, where $Return_{i,t+1}$ is the next month return of firm i, $Direction_{j,t}$ equals to 1 if the insider only makes buy trades, and -1 if the insider only makes sell trades. If an insider makes multiple trades in a month, we aggregate the trades and classify them as a buy (sell) trade if the number of shares bought is greater (less) than the number of shares sold by the insider, following Ali and Hirshleifer, 2017.

at the monthly-level. We could also structure the data so that purchases and sales are separate observations. However, the vast majority (99.5%+) of insiders conduct trades in the same direction each month.[6] For these reasons, the data structure consists of trading observations at the insider i, month t level.

We use the insider trading features from prior months to forecast stock returns in the following month. The regression is as follows:

$$Ret_{i,j,t+1} = f\left(\mathbf{x}_{i,j,t} \mid \theta\right) + \epsilon_{i,j,t+1} \tag{1}$$

where $Ret_{i,j,t+1}$ is the one-month-ahead stock return forecast based on insider $i's$ trades on firm j. $x_{i,j,t}$ denotes insider j's trading features at month t. θ denotes the parameters for the prediction function $f(\cdot)$. To minimize the impact of outliers within the model, we winsorize the continuous features at the 1% level and follow Bogousslavsky et al. (2024) to standardize all the features by subtracting their average and dividing by their standard deviation over the prior years. The standardization makes features comparable across stocks and easier to interpret in our later analysis

We select the hyperparameters of the machine learning models using cross-validation: a data-driven method that does not have look-ahead bias by design. To perform cross-validation, we select two sub-periods of data. The first sub-period is the initial training set, which consists of the 23 months of data from the beginning of the sample until November 2001. The second sub-period is the testing set, which is a single month: December 2001. We train the model using the training set for different configurations of the hyperparameters. We evaluate the results in the testing set and pick the parameters that result in the best performance. We summarize the key parameters of LightGBM model as follows: the learning rate is set to 0.05, the maximum depth is set to 5. Our inferences are unchanged using hyperparameters defaults.

After determining the optimal hyperparameters, we use the remaining 263 months

[6]Our results are virtually identical using separate features for purchases and sales.

(January 2002 to November 2023) for out-of-sample testing. We begin the out-of-sample period in 2002 since the SEC shortened the reporting deadline for corporate insiders from ten business days to two business days following open-market transactions starting in 2002. Therefore, beginning our analysis in 2002 can help minimize look-ahead bias.[7]

We implement our machine learning models using expanding windows to incorporate all available information in generating forecasts, keeping the hyperparameters fixed.[8] In each month t, we train the models using historic data up to t, and use the data at month t to predict $\hat{Ret}_{i,j,t+1}$. After generating the insider-firm-level returns $\hat{Ret}_{i,j,t+1}$, we build a firm-level Insider Trading Intensity (MLS). At the end of each month t, we measure the firm-level MLS by aggregating the $\hat{Ret}_{i,j,t+1}$ for all insiders j in firm i:

$$MLS_{j,t} = \hat{Ret}^{MLS}_{j,t+1} = \frac{1}{j}\sum_{j} Ret_{i,j,t+1} \tag{2}$$

in which subscript j denotes firm, and t indicates the month when forecasts are made.

2.3 Summary statistics

Table 1 presents the summary statistics. We analyze the pairwise correlation of MLS with several prominent human-derived insider trading measures. The correlation between our MLS and the NonRoutine signal of Cohen et al. (2012) is 0.35, a moderate value reflecting that while both signals leverage an insider's trading history, MLS uses a continuous measure of past profitability rather than a simple binary rule based on calendar-month predictability. We document a similar correlation of 0.28 with the pre-QEA signal of Ali and Hirshleifer (2017), indicating that while MLS successfully captures information related to the timing of trades around earnings announcements, it is not exclusively defined by this single event window. The relationship with the NPR from

[7] In additional tests, we also consider a tradable strategy to ensure that all the data used is publicly available at the time.

[8] The result remains similar when using rolling windows, which is shown in the Online Appendix.

Lakonishok and Lee (2001) is also modest at 0.32, as our model incorporates more nuanced measures of economic significance, such as dollar value and percentage of shares outstanding, rather than relying on a simple ratio of trade counts. In each case, the moderate correlation coefficients—all falling well below 0.40—demonstrate that MLS shares a common informational basis with these well-established heuristics but is not beholden to any single dimension of insider behavior. The correlation table is available in the Internet Appendix.

3 Return patterns

We perform a series of asset pricing tests to examine whether the MLS measure predicts future returns.

3.1 Portfolio sorts

We begin by conducting portfolio sorts. At the end of each month, we rank stocks into 10 groups according to their MLS measure and construct a long-short portfolio that buys stocks in the highest MLS decile and sells stocks in the lowest MLS decile. Stocks are held in each portfolio for one month, and the portfolios are rebalanced at the end of each month. We report Newey-West adjusted t-statistics with 12 lags.

Panel A of Table 2 presents the results. The first row reports the average predicted values of MLS in each decile. The predicted return $\hat{Ret}_{j,t+1}^{MLS}$ ranges from −0.06% to +2.38% from decile 1 to decile 10. The long-short portfolio (10-1) has a predicted return of +2.44%. It is worth noting that for decile 1, MLS predicts a modest small negative return of -0.06, suggesting that algorithm cannot identify potential negative content embedded in insider trades.

The next row reports the realized returns of MLS-sorted decile portfolios using equal-weighting. Portfolio 1 aside, we observe a monotonically increasing realized return $Ret_{j,t+1}^{MLS}$

from portfolio 3 +0.80% (t=2.61) to portfolio 10 +2.01% (t=3.81). The predicted return is surprising close to realized return in portfolio 10 (+2.38% vs. +2.01%). The long-short portfolio earns +1.06% (t=3.43).

Figure 1 presents a visualization the performance of the equal-weighted long-short MLS portfolio. We observe that the portfolio outperforms the CRSP value-weighted index substantially. This observation is consistent with the reported Sharpe ratio of 1.01 in the subsequent row. The next set of rows shows that the value-weighted long-short portfolio also earned a significant 0.82% (t=2.37). It confirms that the MLS signal is not just driven by illiquid micro-cap stocks, but also holds when larger, more prominent firms are given more weight.

Panel B performs a double sort by market capitalization and MLS. This panel investigates whether the MLS signal's effectiveness varies with firm size. The results show that the signal is most potent among smaller firms, where information asymmetry is typically higher. For small stocks, the MLS long-short strategy is effective, yielding a monthly return of 1.41% (t=4.59). For medium-sized stocks, the strategy still generates a statistically significant return of 0.75% per month (t-statistic of 1.98). For the large stocks, however, the effect disappears. The long-short portfolio returns a statistically insignificant 0.06% per month.

Overall, the evidence suggests that MLS provides a robust predictor of future stock returns. The signal works for both equal- and value-weighted portfolios, although its predictive power is concentrated in small- and medium-sized firms and diminishes for the largest companies.

3.2 Factor model analysis

It is possible that the portfolio sorts reflect known risks such that the model simply learns to pick small value firms, or high-momentum stocks. To address this concern, we perform factor model analysis to estimate risk-adjusted performance of MLS portfolios

based on the CAPM, Fama-French 3-factor model, Carhart 4-factor model, Fama-French 5-factor model, Fama-French 5-factor with momentum factor, Q-factor model, and the Q-factor model with momentum factor.

Table 3 presents the results. The "H-L" reports the long-short portfolio alphas. The MLS strategy consistently generates a large and highly statistically significant positive alpha, regardless of the risk model used. Panel A reports equal-weighted results showing a monthly alpha ranging from 1.05 (FF5) to 1.31 (Q). Panel B reports the results for value-weighted portfolios, where each stock is weighted by its market capitalization. Although the alphas in the value-weighted panel are slightly lower, they continue to remain positive and statistically significant, ranging from 0.58 (CAPM) to 0.80 (Q).

Overall, the factor model tests indicate that the MLS measure is not a phenomenon driven by loadings on size, value, profitability, investment or momentum factors.

3.3 Fama–MacBeth cross-sectional regressions

We estimate cross-sectional Fama-MacBeth regressions to ensure that our results do not reflect risk premiums associated with firm characteristics. The regressions include the standard set of controls for log(size), log(BM), $ret_{t-12,t-1}$, $ret_{t=0}$, *asset growth*, *profitability*, and *illiquidty*. The dependent variable is the monthly return *Ret* $t+1$. t-statistics are adjusted using Newey and West (1987) with 12 lags.

Table 4 presents the results. We create two indicator variables to denote extreme predictions of the MLS measure because the test assets contain the cross-section of stocks, but MLS can be estimated each month only for stocks with a reported insider trade in $t = 0$. $\mathbb{1}$(*MLS buy*) and $\mathbb{1}$(*MLS sell*) is an indicator variable equal to 1 if MLS is in the top 20% (bottom 20%), and 0 otherwise. In both column (1) and column (3), we observe a significant loading on $\mathbb{1}$(*MLS buy*). The estimate in column (3) implies that, after accounting for all other factors, stocks in the top quintile of the MLS are predicted to earn an additional 0.97% (t=7.24) in the next month compared to other stocks. In contrast, the sell signal has

no predictive power. In column (2), the coefficient estimate on $\mathbb{1}(MLS\ sell)$ is economically small and insignificant 0.090 (t=1.03). This result suggests that insider sales are "noisier" signals than purchases. Insiders might sell for many reasons unrelated to the company's future performance, such as diversifying their personal portfolio, planning for a large purchase, or exercising stock options.

The loadings on the control variables are as expected, lending credibility to the model's specification. log(BM), capturing the value effect, has a positive and significant coefficient (0.244, t=2.38). $Ret_{t=0}$ captures short-term reversal effect. It has is negative and significant coefficient estimate (−1.727, t=−4.52). Profitability and illiquidity both have positive and statistically significant coefficients, respectively. Asset growth has a negative and significant coefficient (−0.658, t=−5.47), in line with the asset growth anomaly. *Size* and $Ret_{t-12,t-1}$ are not statistically significant in this specification.

3.4 Horse-racing machine and human intelligence

We conduct a "horse race" to determine whether MLS generates new information or if it is subsumed by existing, human-derived insider trading measures from the literature. We employ Fama-MacBeth regressions to test whether MLS can predict next-month stock returns after controlling for these well-established signals.

Table 5 presents the results. The evidence suggests that MLS performs well against the established insider trading variables. In column (1), we confirm that the Nonroutine buy signal in Cohen et al. (2012) is a significant predictor on its own. However, when we include our MLS in column (2), the coefficient estimate on $\mathbb{1}$(Nonroutine buy) becomes insignificant, suggesting that its predictive information is subsumed by our measure. We observe a similar pattern with the *pre-QEA buy* signal in Ali and Hirshleifer (2017). The SSN and PPN signals of Hong and Li (2019) survive the inclusion of our MLS measure. The final regression in column (7), which includes all variables simultaneously, shows that $\mathbb{1}$(MLS buy) continues to predict future monthly returns, but some of the other

human-derived signals lose their explanatory power.

Overall, MLS survives in the presence recently developed signals based on insider trading filings. The analysis provides evidence that our machine learning framework can generate a novel and independent signal from the universe of insider trades.

4 Statistical and economic mechanisms

This section examines the statistical and economic mechanism behind the MLS measure. To "open the black box," we use several statistical diagnostic tools to dissect the model's internal logic. Then, based on this data dashboard, we devise tests to explore the underlying economic mechanisms.

4.1 Statistical mechanisms

Our first step analyzes the pieces of information that the model considers most important. Figure 2 presents the feature importance analysis, which ranks the input variables based on how much each feature contributes to the model's predictive accuracy. The analysis reveals that the model places the highest importance on $\%NetTrade$ (the number of shares traded as a percentage of the company's total shares outstanding) and $\$NetTrade$ (the total dollar value of the transaction). The insider's recent trading profitability ($EstRet_{t-3}$) is also highly ranked.

The result is economically intuitive. It suggests that an insider making a large dollar purchase, or buying a quantity of stock that significantly increases their ownership percentage, is sending a much stronger signal of conviction than someone making a small trade. This "skin in the game" is a costly and credible signal that the model correctly identifies as being of primary importance.

Partial dependence plots in Figure 3 presents a visual diagram of how the model's prediction changes as a single feature is varied, as all other features are held constant.

The plots reveal highly non-linear, step-function-like relationships. This means the relationship between a feature (like trade size) and the predicted return is not a smooth, straight line. Instead, the predicted return might remain flat for a range of small trade sizes and then suddenly jump upwards once the trade size crosses a certain critical threshold. This confirms our conjecture that the signals in trading data are complex and not easily captured by traditional linear models (like a standard regression). A linear model assumes that doubling the trade size would double its predictive impact. The machine, however, has learned that such extrapolation may not represent reality by identifying specific thresholds that separate uninformative trades from highly informative ones. This ability to capture these non-linear jumps in predictive power is a primary source of the model's performance.

Next, we attempt understand how the model combines different features. A single feature might be uninformative on its own but powerful when combined with another. The decision tree visualization in Figure 4 provides a simplified map of the model's "if-then" logic and illustrates a key interaction. MLS produces a strong buy signal when an insider makes a trade that is large as a percentage of the company ($\%NetTrade$ is high), but their recent trading performance has performed poorly ($EstRet_{t-3}$ is low). This statistical diagnostic potentially provides a new insight, suggesting that the machine has learned a more subtle pattern. A large trade from an insider who has not been particularly profitable recently is a signal of a change. It suggests this specific trade is driven by a strong conviction that gives the insider a high degree of confidence right now. This combination of high current conviction and a lack of recent success makes the trade stand out as being highly unusual and, therefore, highly informative.

Finally, we evaluate whether the effectiveness of MLS is dominated by a single variable. To do so, we re-estimate the entire GBDT model after dropping each feature individually. The alternative construction of $\hat{Ret}_{j,t+1}^{MLS}$ continues to predict future returns. This result provides further support for the view that the model learns of the complex mapping from

trading features such that any single feature is not the source of predictive power. The results are available in the Internet Appendix.

By systematically examining feature importance, partial dependencies, and decision tree logic, we create a statistical dashboard to interpret the model's complex calculations into economically grounded narrative. In our view, MLS has learned to prioritize high-conviction trades, recognize critical non-linear thresholds, and identify powerful interactions between variables that signal a significant change in an insider's information set.

4.2 Mechanisms: Market reaction to earnings announcement

To evaluate the possibility that MLS reflect private information content embedded within insider trades, we examine the market reaction to earnings announcements. If insiders possess private information, we expect insider purchases to positively predict future earnings announcement reactions. We estimate the following equation.

$$CAR_{i,t-k} = \alpha_{i,t} + \beta_{i,t}\mathbb{1}(MLS) + \beta_{i,t}\Theta_{t-k} + \beta_{i,t}\omega_{t-k} + \varepsilon_{i,t-k} \tag{3}$$

where CAR is 3-day announcement period abnormal return calculated as daily stock return minus return on the CRSP value-weighted portfolio. $\mathbb{1}$(MLS) represents indicators for $\mathbb{1}(\text{MLS}\, buy)$ and $\mathbb{1}(\text{MLS}\, sell)$. Θ is a vector of the following firm characteristics. CAR_{t-1} is the lagged earnings announcement return. $Ret_{t-12,t-1}$ is stock return between month $t-12$ and $t-1$. $Ret_{t=0}$ is the stock return in the month before the earnings announcement. NPR is the insider net purchase ratio defined in Lakonishok and Lee, 2001. Industry and quarter fixed effects are included as indicated. Standard errors are double-clustered at the firm and the quarter level.

Table 6 reports the results. $\mathbb{1}(\text{MLS}\ buy)$ significantly predicts future earnings announcement reactions. Column (1) shows that stocks in the top quintile of our MLS measure—our $\mathbb{1}(\text{MLS}\ buy)$ dummy—exhibit a subsequent earnings announcement CAR

that is 48.2 basis points higher (t=6.89). Conversely, when we examine the sell signal in column (2), we find that the coefficient on $\mathbb{1}(MLS_{t-k})$ is small (9 basis points) and statistically insignificant (t = 1.46). Column (3) reports similar results with the inclusion of both indicators. This evidence suggests that the predictive power of MLS is concentrated on the long side, where insiders trade on impending good news.

The coefficient estimates of the control variables are largely consistent with the prior literature. The prior announcement return, $CAR\ t-1$, is positive and significant (t= 3.38), which is consistent with auto-correlated earnings surprise. The coefficient on the log book-to-market ratio is also positive and significant. Furthermore, our model includes controls for short-term return reversal ($Ret_{t=0}$) and momentum ($Ret_{t-12,t-1}$), although they are not significant in this specification. We find that NPR is statistically significant in column (2), when we omit $\mathbb{1}$(MLS buy), but insignificant in all specifications where $\mathbb{1}$(MLS buy) is included. The inclusion of these controls, in addition to industry and quarter fixed effects, ensures that the predictive ability of the MLS measure is not simply capturing known firm characteristics but is indeed providing a novel signal about future fundamental information.

Finally, to evaluate whether $\mathbb{1}$(MLS buy) can *predict* future earnings announcement returns, we re-estimate the model using Fama and MacBeth (1973) regressions. Column (4) reports a significantly postive loading of 0.543 (t=7.23) on $\mathbb{1}$(MLS buy), suggesting that the MLS measure does predict future earnings announcement returns.

4.3 Mechanisms: Information environment

We expect predictive power of MLS will be pronounced in settings where information asymmetry between insiders and outside investors is highest. Therefore, we partition firms based on three dimensions of the information environment including (1) firm complexity, which speaks to the underlying business operations, (2) Managerial guidance, which captures voluntary disclosure policy, and (3) accounting conservatism, which reflects the

firm's financial reporting style. For each dimension, we sort firms into three groups each year and then, within each tercile, we form decile portfolios based on our measure.

4.3.1 Complexity of business operations

To measure firm complexity, we use the textual-based firm complexity measure developed by Loughran and McDonald (2024). This approach builds on studies showing that the complexity of a firm's operations and disclosures is a primary source of information asymmetry (e.g., Bushman et al., 2004; Li, 2008).

Panel A of Table 7 shows a clear, monotonic relationship. The long-short portfolio's monthly return increases from a significant 0.85% (t-stat = 2.20) for the least complex firms to a remarkable 1.67% (t-stat = 3.30) for the most complex firms. This result supports our hypothesis that the machine-learned signal is most valuable precisely when valuation is most challenging, demonstrating that the MLS effectively isolates informative trades in opaque environments.

We examine the effectiveness of MLS across firms with varying levels of research and development (R&D) intensity. We first sort stocks by R&D expenditure and then by the MLS measure. For firms with high R&D spending, the long-short portfolio—buying stocks with the highest MLS decile and selling those with the lowest—yields a monthly return of 1.96% (t=3.88). The profitability of this strategy is lower for firms with low R&D (1.07%) and for those with missing R&D data (0.90%), indicating that the signal is strongest in R&D-intensive environments.

These results support the findings of Aboody and Lev (2000), who argue that R&D is a primary source of information asymmetry between corporate insiders and outside investors. The opaque nature and uncertain outcomes of R&D projects create a significant information gap that insiders can exploit. The fact that MLS generates the highest returns in high-R&D firms suggests it is successfully identifying the trades that are most information-laden. This aligns with the conclusion from Aboody and Lev (2000) that

insider gains are substantially larger in firms with significant R&D activities. The results are available in Table IA.7 in the Internet Appendix.

4.3.2 Managerial disclosure

Next, we investigate the role of voluntary disclosure, measured by the frequency of managerial guidance over the prior six months. When managers frequently provide forward-looking information, they reduce uncertainty and lower information asymmetry. We therefore expect the MLS to be less potent in such environments. The results in Panel B confirm this intuition. There is a stark inverse relationship between disclosure and the profitability of the MLS strategy. For firms that provide no managerial guidance, the long-short portfolio generates its highest return of 1.64% per month (t-stat = 3.91). This profitability declines as guidance increases, falling to just 0.72% (t-stat = 2.28) for firms with the most frequent disclosures. This finding suggests that insider trades, as interpreted by our algorithm, act as a substitute for managerial disclosure. In an information vacuum left by silent management, the MLS becomes an exceptionally powerful predictor of future returns.

4.3.3 Financial reporting

Finally, we investigate how financial reporting style affects the MLS signal by sorting on accounting conservatism. We argue that conservatism, by requiring a higher verification standard for recognizing good news than bad news (Basu, 1997), can create a gap between reported accounting performance and the firm's true economic prospects (Watts, 2003). Conservative accounting, by requiring a higher threshold for recognizing good news than bad news, can temporarily mask a firm's underlying economic prospects. This creates opportunities for insiders, who have a clearer view of the firm's true performance, to trade on unrecorded economic gains. The results in Panel C are striking. The effectiveness of the MLS is greatest among firms with the most conservative accounting, where the long-

short strategy yields a monthly return of 1.69% (t-stat = 4.38). In contrast, for firms with the least conservative (i.e., most aggressive) reporting, the strategy's return of 0.31% is statistically insignificant. This suggests that our model is particularly adept at identifying insider purchases that signal latent good news that has not yet been reflected in the financial statements due to conservative reporting principles.

Collectively, the evidence that it has learned a sophisticated and intuitive economic relationship. The MLS signal is systematically more powerful in firms characterized by higher information asymmetry—whether that asymmetry arises from operational complexity, a lack of voluntary disclosure, or a conservative financial reporting style. These findings anchor our primary results in established accounting theory and enhance the credibility of the MLS as a genuine information signal.

5 Additional tests

5.1 Tradable strategy

We assess whether it is possible to use MLS to create a real-time tradable strategy. According to SEC regulations, insider trades must be disclosed within two business days according to SEC rules. Given the relatively timely disclosure of insider trades, we can design a real-time trading strategy as the trades are reported. Occasionally, insider trades are reported with delays due to technical issues or human error. Therefore, to implement a real time strategy, we re-construct the MLS measure using only trades available on the SEC Edgar system in month=0. Then, we repeat our Fama-Macbeth analysis.

Table 8 reports the result. Columns (1) to (3) report significantly positive loadings on $\mathbb{1}(\text{MLS } buy)^{real-time}$ constructed using the real-time version of MLS, suggesting that MLS could be used as a tradable strategy. In columns (4) to (6), we perform a skip-a-month strategy such that we use the MLS predicted return from the month prior $t-1$. We continue to find significant loadings on $\mathbb{1}(\text{MLS } buy)_{t-1}$, suggesting that the prior month's

MLS measure contains valuable information on future stock returns.

5.2 Using alternative machine learning techniques

Table 9 address the concern of whether our findings are specific to our chosen Light GBM algorithm or if they represent a more generalizable phenomenon. To this end, we repeat our portfolio sorting analysis using the standard suite of machine learning techniques, ranging from simple linear models such as ordinary least squares to other complex, non-linear methods.

Panel A presents results using equal-weighting. We observe that most models can generate positive long-short portfolio returns, although some models are not statistically significant at conventional levels. Light GBM dominates the rest of the models in terms of portfolio return (1.01%) and Sharpe ratio (1.10). The closest performer is Random forest with a monthly return of 0.98% (t=3.24) and Sharpe ratio of 0.98. When we use simpler, linear methods like OLS, Ridge, and Lasso regression, the resulting long-short portfolios, while still directionally positive, are substantially weaker. The best-performing linear model, Lasso, produces a long-short portfolio return of only 0.46% (t= 2.14), which is less than half the magnitude of the returns generated by the tree-based methods. Furthermore, the performance of our neural network models is solid but does not consistently outperform the tree-based models, which aligns with our initial intuition that for a dataset with a limited number of highly potent features, gradient-boosted trees are an exceptionally well-suited tool. Ultimately, the consistent outperformance of the non-linear models over the linear ones supports our conjecture that the machine learning signal is driven by complex interactions and thresholds that simpler models are incapable of capturing, supporting our methodological approach.

Panel B shows that the linear models fail to generate economically meaningful returns using value-weighting. Of the three models, Lasso performs the best, but the resulting monthly return of 0.11% is not statistically significant (t=0.37). Random forest is again

the best performer with a month return of 1.03% (t=3.37), which is larger than the Light GBM monthly return of 0.82% (t=2.37). Hence, Random forest is a solid alternative ML method in our setting. It is comforting that alternative tree-based models (Random forest, XGBoost) generate results that are comparable to our preferred algorithm, which quells 'model-mining' concerns.

5.3 Firm valuation

Our results related to accounting conservatism suggest that under-valuation may facilitate the profitability of *MLS*. To examine the role of firm valuation, we conduct a double-sort analysis, first partitioning firms based on a valuation metric and then sorting within those partitions by *MLS*. Our valuation metric is the industry-adjusted price-to-sales (PS) ratio. Firms with the lowest PS ratios as classified as "value" stocks (Low) and those with the highest are classified as "glamour" stocks (High). For value stocks, the long-short portfolio constructed using *MLS* yields a monthly return of 1.91% (t=4.71). For the glamour stocks—those most likely to be overvalued—*MLS* is ineffective, producing a statistically insignificant return of 0.23%. We find similar results using the anomaly score measure developed in Hou et al. (2015). Overall, the results suggest that *MLS* exhibits the strongest predictability in undervalued stocks. We present the results in Table IA.8 in the Internet Appendix.

5.4 Insider seniority

Table 10 separately estimates *MLS* values based on the insider's employment role within the firm. Generally, we find that *MLS* signal remains statistically significant across all insiders. For instance, the *MLS* buy signal for senior officers is economically similar to non-senior employees (1.002% vs. 0.995%). The strongest signal originates from non-directors (1.344, t=4.59), but non-director insiders (0.847, t=5.47) also strongly predict future returns, indicating that the machine learning model successfully identifies

informative trades across different levels of the corporate hierarchy.

These results also add a new dimension to the findings in Ravina and Sapienza (2010), who investigated the trading performance of independent directors. We find that trades by both independent and dependent directors are highly informative. This aligns with the conclusion from Ravina and Sapienza (2010) that independent directors earn substantial abnormal returns, suggesting they are not at an informational disadvantage compared to executive insiders. The machine learning model confirms that trades by these supposedly more detached insiders still contain significant predictive power, reinforcing the idea that they possess and trade on valuable, non-public information.

5.5 Rule 10b5-1

SEC Rule 10b5-1 allows corporate insiders to establish a pre-arranged, written trading plan for their company's stock at a time when they are not in possession of any material nonpublic information. Rule 10b5-1 is intended to provide an affirmative defense against insider trading by allowing insiders to transact based on a pre-arranged schedule, thus removing the influence of immediate, private information. This plan specifies the future dates, prices, and amounts of shares to be traded, or provide a fixed formula for doing so, thereby removing the insider's direct influence over the transactions once the plan is active. To address concerns about potential misuse, the SEC introduced significant amendments in 2022, which now mandate a "cooling-off" period between establishing a plan and executing the first trade, restrict the use of multiple overlapping plans, and require greater public disclosure of these arrangements. This framework enables insiders to systematically sell their shares for personal financial planning while maintaining compliance with securities laws.

We separate our trades into Rule 10b5-1 and non-Rule 10b5-1 trades. For trades conducted outside of these pre-scheduled plans, the MLS measure demonstrates significant predictive power. The long-short portfolio, which buys stocks with the highest MLS and

sells those with the lowest, generates a robust return of 1.12% per month (t=3.71). The increase in returns from the lowest to the highest decile underscores the model's ability to successfully identify discretionary, information-driven trades made by insiders. In contrast, the signal's predictive ability vanishes for trades executed under a Rule 10b5-1 plan. The long-short portfolio for these trades yields a statistically insignificant monthly return of −0.15% (t=−0.24).

This finding aligns with Fich et al. (2023), who investigate how these plans are used. Although the study finds that insiders can still be opportunistic, for instance by timing the initiation or cancellation of plans, our results suggest that individual trades within the plans are not as informative as spontaneous trades. The inability find a signal in 10b5-1 trades demonstrates its ability to distinguish between pre-scheduled, less-informative trades and discretionary trades that are more likely to be based on valuable inside information. The results are available in Table IA.9 in the Internet Appendix.

6 Conclusion

This paper investigates whether machine learning can move beyond human-devised heuristics to extract economically significant signals from complex trading data. We address this question by using a gradient-boosted decision tree model to analyze corporate insider trades filed with the SEC. This model synthesizes a high-dimensional set of features—including trade size, direction, and the insider's past trading performance—into a single machine learning signal based on insider trading files. We test the out-of-sample performance of this machine-generated signal to determine if it can systematically predict stock returns and provide information beyond existing insider-trading based signals.

The primary findings demonstrate that machine learning can indeed uncover potent, economically significant signals. A long-short portfolio constructed based on *MLS* generates a significant alpha of 1.06% per month, with a Sharpe ratio of 1.01. *MLS* provides

incremental information that is orthogonal to well-established, human-derived insider trading signals, proving that the model is not merely rediscovering existing knowledge. We further link MLS to fundamental corporate news, showing it can predict future earnings announcement reactions, which suggests the signal is capturing genuine information about firm undervaluation.

These results have potential policy implications and may open new avenues for future research. For regulators like the SEC, the finding that machine learning can systematically identify profitable trading patterns within the current disclosure framework highlights the persistent challenge of mitigating informational advantages. This methodology could be adapted by regulators to enhance surveillance and identify trades that warrant closer scrutiny. A natural extension of this work would be to combine the machine learning techniques with the signals developed in extant studies to potentially improve the predictive signal. Future research could also incorporate alternative data sources, like news sentiment or satellite imagery, to see if they can further refine the model's predictive accuracy and provide an even deeper understanding of market dynamics.

References

Aboody, D. and Lev, B. (2000). "Information Asymmetry, R&D, and Insider Gains". Journal of Finance 55, 2747–2766.

Ali, U. and Hirshleifer, D. (2017). "Opportunism as a firm and managerial trait: Predicting insider trading profits and misconduct". Journal of Financial Economics 126, 490–515.

Amihud, Y. (2002). "Illiquidity and stock returns: cross-section and time-series effects". Journal of Financial Markets 5, 31–56.

Bogousslavsky, V., Fos, V., and Muravyev, D. (2024). "Informed trading intensity". Journal of Finance 79, 903–948.

Brown, N. C., Crowley, R. M., and Elliott, W. B. (2020). "What are you saying? Using topic to detect financial misreporting". Journal of Accounting Research 58, 237–291.

Cao, S., Jiang, W., Wang, J., and Yang, B. (2024). "From man vs. machine to man+ machine: The art and AI of stock analyses". Journal of Financial Economics 160, 103910.

Chen, T. and Guestrin, C. (2016). "XGboost: A scalable tree boosting system". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, X., Cho, Y. H., Dou, Y., and Lev, B. (2022). "Predicting future earnings changes using machine learning and detailed financial data". Journal of Accounting Research 60, 467–515.

Cohen, L., Malloy, C., and Pomorski, L. (2012). "Decoding inside information". Journal of Finance 67, 1009–1043.

deHaan, E., Lee, C., Liu, M., and Noh, S. (2025). "The Shadow Price of". Information (May 15, 2025).

Fama, E. F. and French, K. R. (2015). "A five-factor asset pricing model". Journal of Financial Economics 116, 1–22.

Fama, E. F. and MacBeth, J. D. (1973). "Risk, Return, and Equilibrium: Empirical Tests". Journal of Political Economy 81, 607–636.

Fich, E. M., Parrino, R., and Tran, A. L. (2023). "When and how are rule 10b5-1 plans used for insider stock sales?" Journal of Financial Economics 149, 1–26.

Geertsema, P. and Lu, H. (2023). "Relative valuation with machine learning". Journal of Accounting Research 61, 329–376.

Grennan, J. and Michaely, R. (2021). "Fintechs and the market for financial analysis". Journal of Financial and Quantitative Analysis 56, 1877–1907.

Gu, S., Kelly, B., and Xiu, D. (2020). "Empirical asset pricing via machine learning". Review of Financial Studies 33, 2223–2273.

Guenther, D. A., Peterson, K., Searcy, J., and Williams, B. M. (2023). "How useful are tax disclosures in predicting effective tax rates? A machine learning approach". The Accounting Review 98, 297–322.

Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning: data mining, inference, and prediction*.

Hong, C. Y. and Li, F. W. (2019). "The Information Content of Sudden Insider Silence". Journal of Financial and Quantitative Analysis 54, 1499–1538.

Hou, K., Xue, C., and Zhang, L. (2015). "Digesting anomalies: An investment approach". Review of Financial Studies 28, 650–705.
Jaffe, J. F. (1974). "Special information and insider trading". The Journal of Business 47, 410–428.
Jones, S., Moser, W. J., and Wieland, M. M. (2023). "Machine learning and the prediction of changes in profitability". Contemporary Accounting Research 40, 2643–2672.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). "LightGBM: A highly efficient gradient boosting decision tree". Advances in neural information processing systems 30.
Kim, S., Kim, S., and Rajgopal, S. (2025). "Insider Trading After the 2022 Rule 10b5-1 Amendment". Available at SSRN 5362431.
Lakonishok, J. and Lee, I. (2001). "Are insider trades informative?" Review of Financial Studies 14, 79–111.
Loughran, T. and McDonald, B. (2024). "Measuring firm complexity". Journal of Financial and Quantitative Analysis 59, 2487–2514.
Newey, W. K. and West, K. D. (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". Econometrica 55, 703–708.
Ravina, E. and Sapienza, P. (2010). "What do independent directors know? Evidence from their trading". The Review of Financial Studies 23, 962–1003.
Van Binsbergen, J. H., Han, X., and Lopez-Lira, A. (2023). "Man versus machine learning: The term structure of earnings expectations and conditional biases". The Review of financial studies 36, 2361–2396.

Appendix. Variable definitions

Insider trading features

Variable	Description
%NetTrade	Total number of shares purchased minus the number of shares sold divided by shares outstanding for insider j in month $t = 0$.
$NetTrade	Total dollar value of shares purchased minus dollar value of shares sold for insider j in month $t = 0$.
#NetTrade	Total number of purchase transactions minus the number of sale transactions for insider j in month $t = 0$.
$EstRet_{t-3}$	Estimated return of $\#NetTrade$ for insider j conducted in month $t-3$ computed as the signed subsequent month Ret_{t-2}. For purchases, $EstRet_{t-3}$=Ret_{t-2}. For sales, $EstRet_{t-3}$=$-Ret_{t-2}$. The value is set to missing if insider j did not conduct a trade in month $t-3$.
$EstRet_{t-6}$	Estimated return of $\#NetTrade$ for insider j conducted in month $t-6$ computed as the signed subsequent month Ret_{t-5}. The value is set to missing if insider j did not conduct a trade in month $t-6$.
$EstRet_{t-12}$	Estimated return of $\#NetTrade$ for insider j conducted in month $t-12$ computed as the signed subsequent month Ret_{t-11}. The value is set to missing if the insider did not conduct a trade in month $t-12$.
$Ownership\%$	Resulting shares held by insider j divided by shares outstanding after conducting all transactions in month $t = 0$.

Variable descriptions

Variable	Description
MLS	Insider trading signal defined in Section 2.
$\mathbb{1}$(MLS buy)	Indicator equal to one if MLS in a given month is in the top 20 quintile, and zero otherwise.
$\mathbb{1}$(MLS $buy)^{realtime}$	Indicator equal to one if MLS in a given month is in the top 20 quintile, and zero otherwise using trades available on the SEC Edgar system at $t = 0$.
$\mathbb{1}$(MLS $sell$)	Indicator equal to one if MLS in a given month is in the bottom 20 quintile, and zero otherwise.
$\mathbb{1}$(Nonroutine buy)	Indicator equal to one if there are any buys on a given firm by a nonroutine insider classified by Cohen, Malloy, and Pomorski (2012).
$\mathbb{1}$(Nonroutine sell)	Indicator equal to one if there are any sells on a given firm by a nonroutine insider classified by Cohen, Malloy, and Pomorski (2012).
$\mathbb{1}$(pre-QEA buy)	Indicator equal to one if there are any buys on a given firm by an insider in quintile 5 classified by Ali and Hirshleifer (2017).
$\mathbb{1}$(pre-QEA sell)	Indicator equal to one if there are any sells on a given firm by an insider in quintile 5 classified by Ali and Hirshleifer (2017).
$\mathbb{1}$(SSN)	Indicator equal to one if a firm has any insider who sells consecutively in the same calendar month for the previous two years, but does not trade in the last month Hong and Li (2019).
$\mathbb{1}$(PPN)	Indicator equal to one if a firm has any insider who purchases consecutively in the same calendar month for the previous 2 years, but does not trade in the last month Hong and Li (2019).
NPR	Insider net purchase ratio over the past 6 months following Lakonishok and Lee (2001), calculated as $NPR = \frac{\#Trade^{Buy}_{t-1,t-6} - \#Trade^{Sell}_{t-1,t-6}}{\#Trade^{Buy}_{t-1,t-6} + \#Trade^{Sell}_{t-1,t-6}}$.
Ret $t+1$	Stock return in month $t+1$.
CAR	3-day cumulative abnormal return around a quarterly earnings announcement.
log(Size)	Natural logarithm of the market value of equity.
log(BM)	Natural logarithm of the ratio of book value of equity and market value of equity.
$Ret_{t-12,t-1}$	Stock return between month $t-12$ and $t-1$.
$Ret_{t=0}$	Stock return in month $t = 0$.
Asset growth	The annual growth rate of total assets.
Profitability	Firm gross profits to assets.
Illiquidity	Past 12 months average of daily return divided by turnover Amihud (2002).
PS	Market value of equity to sales, minus industry mean.
Anomaly score	The sum of ternary signal for 27 anomalies in Hou, Xue, and Zhang (2015).
Cash	The ratio of cash and short-term investment and total assets.
ROA	The ratio of operating income before depreciation and total assets.
Leverage	The ratio of total liabilities and total assets.

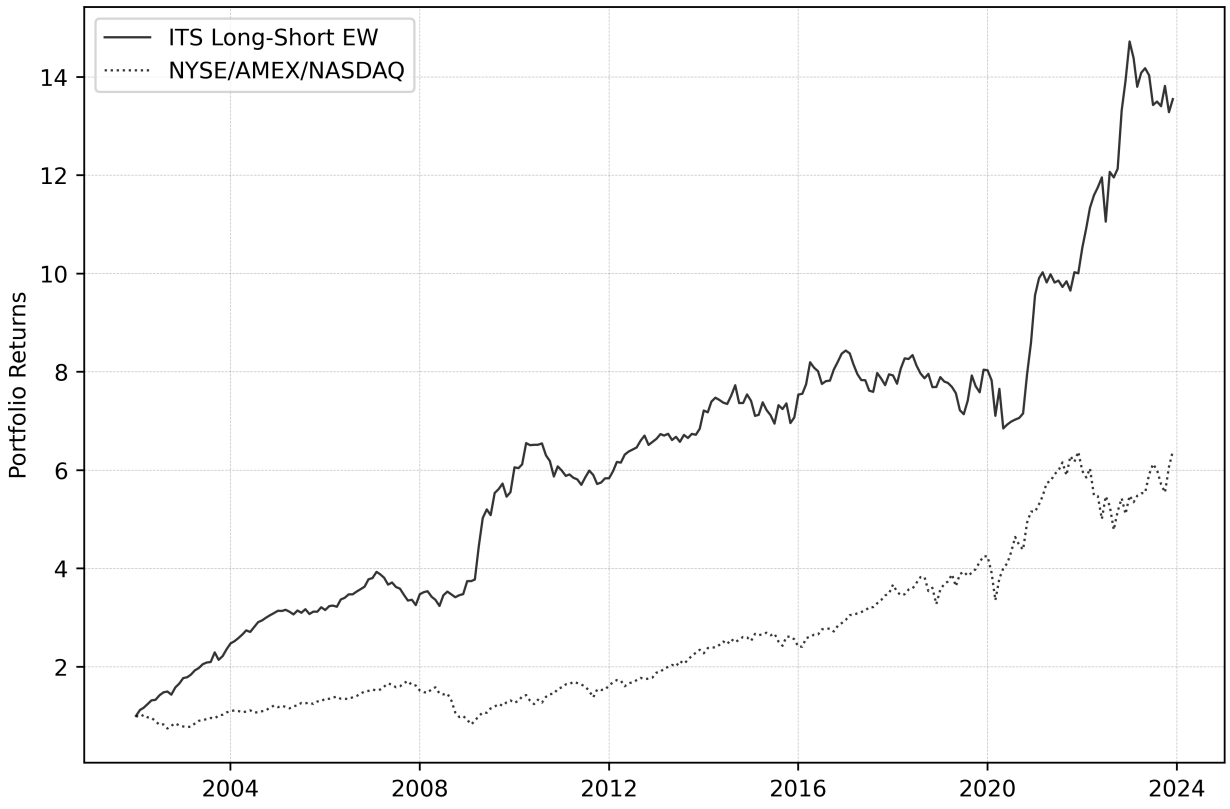

Figure 1. Cumulative performance of the portfolios sorted on MLS

This figure presents the cumulative returns of the equal-weighted *MLS* Long-Short portfolio and the CRSP value-weighted portfolio from 2002 to 2023.

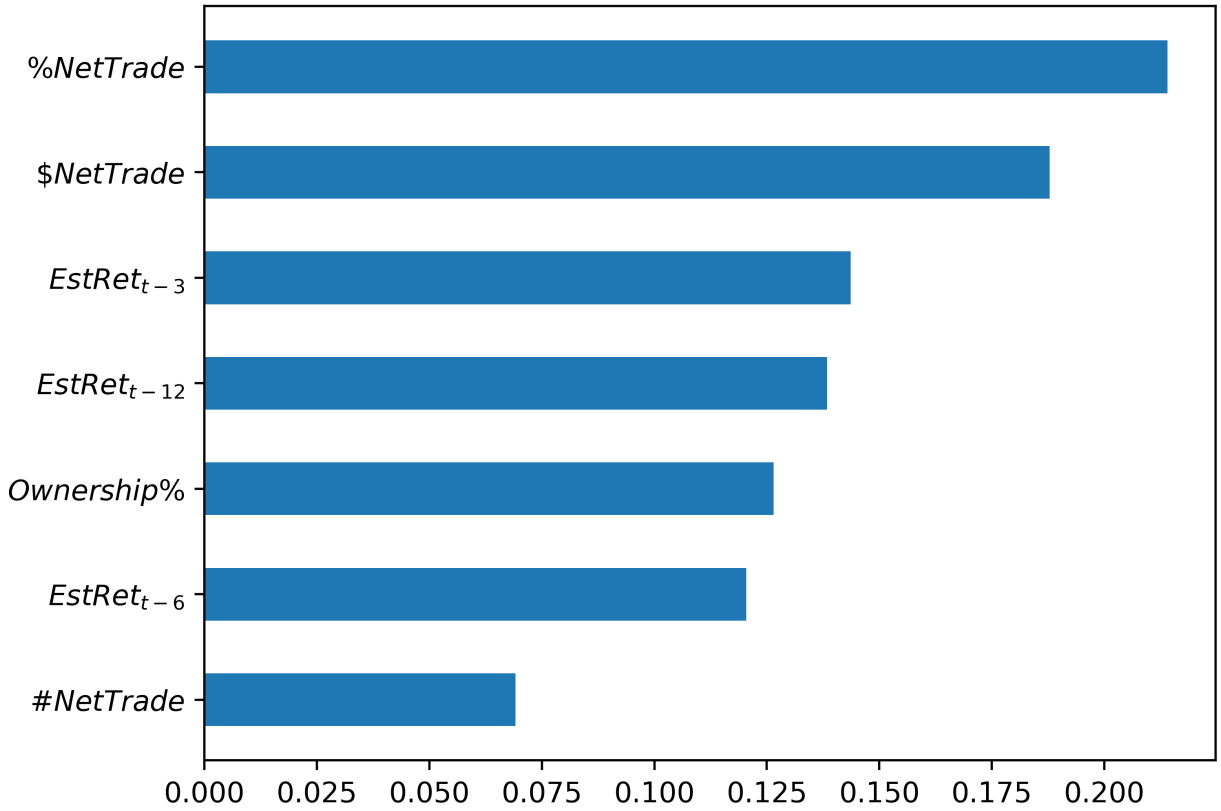

Figure 2. Feature importance

This figure presents the time-series average of feature importance of seven insider trading characteristics used to predict stock return. The feature importance measures the number of times a feature is used to split the data across all trees in the model. The relative importance measure across all features sums to 1.

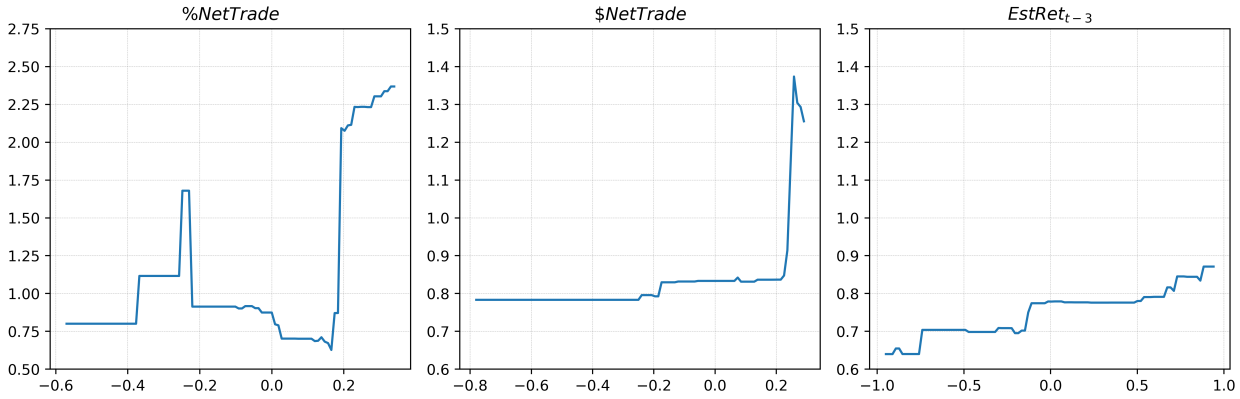

Figure 3. Partial dependence plots

This figure presents the partial dependence on 2002 2023 for three most important features. The partial dependence plots are calculated from the Light GBM model, which regresses stock return on the features described in the Appendix.

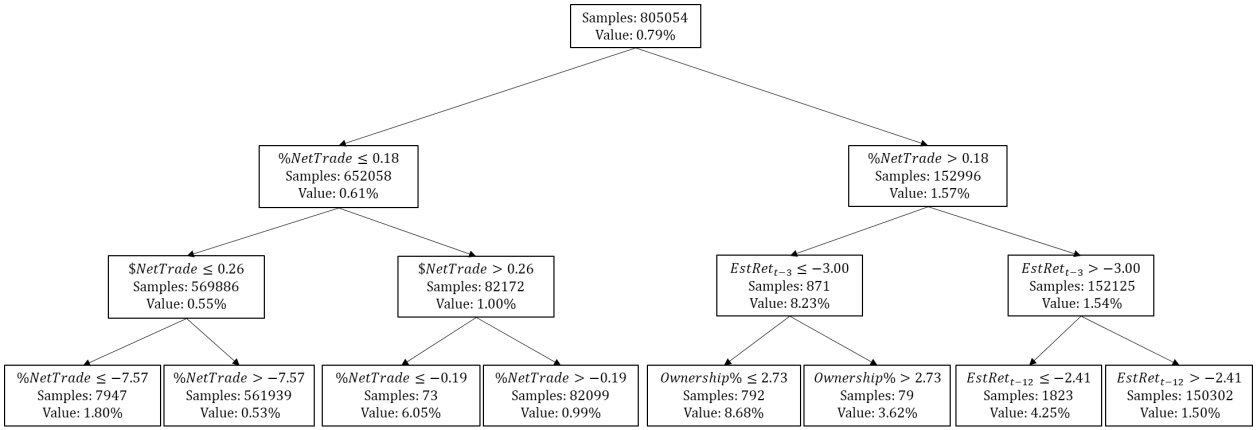

Figure 4. Example decision tree

The figure presents the decision tree from the Light GBM model.

Table 1. Summary statistics

This table reports summary statistics of the main variables used in this study. Panel A reports summary statistics of the MLS measure and its component features. Panel B reports the difference in firm characteristics between $\mathbb{1}$(MLS buy) and $\mathbb{1}$(MLS $sell$) stocks. Panel C reports summary statistics of the firm characteristics used in the analysis. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above \$1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11.

Panel A. MLS & component features

	N	Mean	10%	Median	90%	SD
$MLS(\%)$	367,252	0.847	0.220	0.641	1.824	0.950
$\%NetTrade(\%)$	878,287	-0.075	-0.144	-0.006	0.025	0.474
$\$NetTrade(Million)$	878,287	-1.520	-2.668	-0.140	0.056	5.917
$\#NetTrade$	878,287	-2.047	-5.000	-1.000	2.000	6.294
$EstRet_{t-3}$	376,938	-0.015	-0.128	-0.012	0.097	0.103
$EstRet_{t-6}$	471,814	-0.012	-0.112	-0.011	0.088	0.093
$EstRet_{t-12}$	583,231	-0.009	-0.099	-0.009	0.083	0.086
$Ownership\%$	774,382	1.565	0.004	0.080	3.183	4.874

Panel B. $\mathbb{1}$(MLS buy) and $\mathbb{1}$(MLS $sell$)

	MLS buy	MLS $sell$	Difference
Ret $t+1$ (%)	1.933	0.932	1.001***
Size (Billion)	1.179	12.126	-10.947***
BM	0.832	0.411	0.420***
$Ret_{t-12,t-1}$	0.061	0.332	-0.271***
$Ret_{t=0}$	0.012	0.030	-0.018***
Asset growth	0.128	0.202	-0.075***
Profitability	0.242	0.382	-0.140***
Illiquidity	2.691	0.120	2.571***
PS	8.466	9.339	-0.873*
PB	2.543	5.593	-3.050***

Panel C. Firm characteristics

	N	Mean	10%	Median	90%	SD
$\mathbb{1}$(MLS buy)	853,110	0.053	0.000	0.000	0.000	0.225
$\mathbb{1}$(MLS $sell$)	853,110	0.068	0.000	0.000	0.000	0.251
Ret $t+1$ (%)	853,110	0.963	-14.008	0.466	15.346	15.959
Size (Billion)	853,110	4.671	0.043	0.564	9.851	13.695
BM	853,110	0.668	0.159	0.540	1.278	0.548
$Ret_{t-12,t-1}$	853,110	0.152	-0.419	0.062	0.684	0.790
$Ret_{t=0}$	853,110	0.013	-0.137	0.005	0.157	0.167
Asset growth	853,110	0.130	-0.137	0.056	0.417	0.368
Profitability	853,110	0.296	0.027	0.257	0.691	0.310
Illiquidity	853,110	1.351	0.000	0.006	1.393	5.971
NPR	853,110	-0.297	-1.000	-0.111	1.000	0.710
$\mathbb{1}$(Nonroutine buy)	853,110	0.065	0.000	0.000	0.000	0.247
$\mathbb{1}$(Nonroutine $sell$)	853,110	0.206	0.000	0.000	1.000	0.404
$\mathbb{1}$(pre-QEA buy)	853,110	0.008	0.000	0.000	0.000	0.088
$\mathbb{1}$(pre-QEA $sell$)	853,110	0.032	0.000	0.000	0.000	0.176
$\mathbb{1}$(SSN)	853,110	0.056	0.000	0.000	0.000	0.230
$\mathbb{1}$(PPN)	853,110	0.012	0.000	0.000	0.000	0.109

Table 2. Portfolio sorts: Excess returns

This table reports the raw returns (%), t-statistics, and the Sharpe ratio of portfolios constructed using the *MLS* measure. At the end of each month, we rank stocks into 10 groups based on *MLS* and construct the long-short (10−1) portfolio. Stocks are held in the portfolios for one month, and the portfolios are rebalanced at the end of each month. Panel A reports the portfolio returns sorted based on deciles of *MLS*. Panel B reports portfolio returns double-sorted by market capitalization and deciles of *MLS*. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

Panel A. MLS-sorted portfolios

		MLS deciles										
		1	2	3	4	5	6	7	8	9	10	10−1
MLS	$\hat{Ret}_{j,t+1}^{GBM}$	-0.06	0.32	0.37	0.45	0.54	0.67	0.83	1.07	1.47	2.38	2.44
Equal-weighted	Ret	0.95	0.89	0.80	0.83	0.88	0.98	1.00	1.24	1.27	2.01	1.06
	t-stat	(2.12)	(2.65)	(2.61)	(2.53)	(2.66)	(2.89)	(2.79)	(3.35)	(2.99)	(3.81)	(3.43)
	SR	0.45	0.53	0.47	0.47	0.49	0.56	0.54	0.67	0.67	0.94	1.01
Value-weighted	Ret	1.07	0.81	0.73	0.85	0.72	1.05	1.16	0.88	0.79	1.89	0.82
	t-stat	(2.83)	(2.42)	(2.68)	(2.97)	(2.41)	(2.99)	(3.93)	(2.35)	(1.85)	(3.87)	(2.37)
	SR	0.60	0.52	0.50	0.54	0.45	0.64	0.67	0.48	0.36	0.79	0.51

Panel B. Double sort by market capitalization and MLS

		MLS deciles										
Market capitalization		1	2	3	4	5	6	7	8	9	10	10−1
Small	Ret	0.76	0.77	0.88	0.80	1.37	0.93	1.61	1.72	1.90	2.17	1.41
	t-stat	(1.50)	(1.76)	(2.03)	(2.08)	(3.12)	(2.19)	(3.57)	(3.36)	(3.63)	(3.43)	(4.59)
Medium	Ret	0.85	0.99	0.58	1.13	0.83	1.06	0.90	1.19	1.20	1.59	0.75
	t-stat	(1.79)	(2.41)	(1.67)	(3.00)	(2.38)	(3.27)	(2.54)	(3.41)	(3.42)	(3.30)	(1.98)
Large	Ret	1.02	0.89	0.80	0.83	0.88	0.98	1.00	0.99	1.01	1.08	0.06
	t-stat	(2.27)	(2.75)	(3.30)	(2.55)	(2.48)	(2.69)	(3.23)	(3.31)	(3.08)	(2.92)	(0.20)

Table 3. Portfolio sorts: Risk-adjusted performance of MLS portfolios

This table reports the risk-adjusted performance of *MLS* portfolios based on the CAPM model, the Fama-French 3-factor model, the Carhart 4-factor model, the Fama-French 5-factor model, the Fama-French 5-factor model augmented with momentum factor, the Q-factor model, and the Q-factor model augmented with momentum factor. Panel A reports results for equal-weighted portfolios, and Panel B reports results for value-weighted portfolios. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

Panel A. Equal-weighted MLS-sorted portfolios

	CAPM		FF3		Carhart		FF5		FF5+MOM		Q		Q+MOM	
	Alpha	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*
L	-0.10	-0.51	-0.11	-0.95	-0.11	-0.93	0.11	1.07	0.10	1.03	0.17	1.35	0.18	1.54
2	0.02	0.16	0.01	0.14	0.00	0.01	0.11	1.51	0.10	1.48	0.15	2.48	0.16	2.73
3	-0.07	-0.80	-0.05	-1.01	-0.08	-1.54	-0.07	-1.19	-0.08	-1.42	-0.05	-0.79	-0.04	-0.68
4	-0.08	-0.73	-0.10	-1.07	-0.12	-1.22	-0.09	-0.98	-0.10	-1.09	-0.10	-0.92	-0.10	-0.86
5	-0.04	-0.34	-0.08	-1.30	-0.09	-1.52	-0.04	-0.71	-0.05	-0.76	-0.02	-0.31	-0.02	-0.24
6	0.09	0.52	0.09	0.78	0.11	0.92	0.13	1.22	0.14	1.29	0.18	1.45	0.18	1.45
7	0.07	0.38	0.08	0.77	0.14	1.43	0.18	1.85	0.21	2.54	0.28	2.58	0.27	2.67
8	0.33	1.48	0.34	2.81	0.40	3.23	0.36	3.02	0.39	3.41	0.50	4.47	0.49	4.45
9	0.33	1.26	0.35	1.95	0.44	2.37	0.46	2.43	0.50	2.83	0.64	3.43	0.63	3.46
H	0.99	2.87	1.00	4.19	1.13	4.86	1.15	4.83	1.20	5.78	1.48	6.15	1.47	6.22
H−L	1.08	3.55	1.11	4.27	1.23	4.95	1.05	4.22	1.10	5.21	1.31	5.40	1.29	5.82

Panel B. Value-weighted MLS-sorted portfolio

	CAPM		FF3		Carhart		FF5		FF5+MOM		Q		Q+MOM	
	Alpha	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*	*Alpha*	*t*
L	0.17	1.06	0.15	1.14	0.14	1.03	0.24	1.68	0.23	1.59	0.29	1.61	0.29	1.65
2	0.02	0.15	0.00	0.04	-0.02	-0.20	0.08	0.75	0.06	0.64	-0.00	-0.03	0.00	0.04
3	-0.04	-0.44	-0.04	-0.53	-0.08	-1.00	-0.07	-0.73	-0.08	-0.93	-0.11	-1.23	-0.10	-1.18
4	0.03	0.30	0.03	0.28	0.03	0.27	0.02	0.17	0.02	0.17	0.04	0.34	0.04	0.36
5	-0.09	-1.02	-0.09	-1.07	-0.11	-1.24	-0.12	-1.29	-0.12	-1.37	-0.14	-1.38	-0.13	-1.33
6	0.19	1.33	0.19	1.37	0.21	1.59	0.15	1.06	0.16	1.18	0.14	0.99	0.14	0.97
7	0.26	1.94	0.27	2.10	0.30	2.16	0.25	1.64	0.26	1.71	0.32	2.46	0.32	2.45
8	-0.04	-0.22	-0.02	-0.14	-0.00	-0.02	-0.03	-0.16	-0.02	-0.10	0.01	0.04	-0.00	-0.00
9	-0.21	-0.83	-0.19	-0.90	-0.11	-0.53	-0.14	-0.57	-0.10	-0.44	0.16	0.68	0.14	0.66
H	0.75	2.64	0.77	3.33	0.89	3.94	0.88	3.72	0.93	4.27	1.09	4.56	1.07	4.53
H−L	0.58	1.71	0.62	2.19	0.75	2.71	0.64	2.21	0.70	2.51	0.80	2.53	0.78	2.49

Table 4. Fama-MacBeth regression

This table reports the the average cross-sectional coefficient estimates from Fama-MacBeth regressions of stock returns on MLS. The dependent variable is next month stock return $t+1$. MLS in the Machine Learned Signal. $\mathbb{1}$(MLS buy) and $\mathbb{1}$(MLS $sell$) are indicator variables equal to one if MLS is in the top 20% (bottom 20%), and 0 otherwise. *Size* and *log(BM)* are the natural logarithms of the firm market equity and book-to-market. $Ret_{t-12,t-2}$ is stock return from month $t-12$ to $t-2$. $Ret_{t=0}$ is the return in month $t=0$. *AssetGrowth* is annual growth rate of total assets. *Profitability* is firm gross profits to assets. *Illiquidity* is Amihud illiquidity measure. We winsorize *Size*, *log(BM)*, *AssetGrowth*, *Profit* and *Illiquidity* at 1% and 99% levels. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above \$1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. 1%, 5%, and 10% statistical significance is indicated with ***, **, and *, respectively.

	(1)	(2)	(3)
$\mathbb{1}$(MLS buy)	0.967***		0.970***
	(7.19)		(7.24)
$\mathbb{1}$(MLS $sell$)		0.064	0.090
		(0.72)	(1.03)
log(Size)	0.021	0.008	0.020
	(0.41)	(0.15)	(0.38)
log(BM)	0.242**	0.247**	0.244**
	(2.34)	(2.41)	(2.38)
$Ret_{t-12,t-1}$	0.075	0.064	0.072
	(0.29)	(0.25)	(0.28)
$Ret_{t=0}$	-1.714***	-1.724***	-1.727***
	(-4.49)	(-4.52)	(-4.52)
Asset growth	-0.656***	-0.652***	-0.658***
	(-5.40)	(-5.43)	(-5.47)
Profitability	0.762***	0.745***	0.757***
	(3.99)	(3.91)	(3.97)
Illiquidity	0.021**	0.021**	0.021**
	(2.02)	(2.01)	(2.02)
N	853,110	853,110	853,110

Table 5. MLS versus existing insider trading signals

This table reports results from Fama-MacBeth regression to compare the performance of Machine Learned Signal (MLS) with *Non-Rountine buy/Sell* from Cohen, Malloy, and Pomorski (2012), *pre-QEA Buy/Sell* from Ali and Hirshleifer, 2017, and *SSN/PPN* from Hong and Li, 2019. The dependent variable is next month stock return $t+1$. *MLS* is defined in Section 1. *MLS buy* (*MLS sell*) is an indicator variable equal to 1 if *MLS* is in the top 20% (bottom 20%) and 0 otherwise. *NPR* is the insider net purchase ratio defined by Lakonishok and Lee, 2001. We winsorize *Size*, *log(BM)*, *AssetGrowth*, *Profitability* and *Illiquidity* at 1% and 99% levels. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. 1%, 5%, and 10% statistical significance is indicated with ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
𝟙(MLS *buy*)		0.906***		0.964***		0.963***	0.903***
		(6.66)		(7.35)		(7.35)	(6.60)
𝟙(MLS *sell*)		0.122		0.094		0.104	0.120
		(1.48)		(1.18)		(1.24)	(1.51)
𝟙(NonRoutine *buy*)	0.566***	0.090					0.118
	(5.52)	(0.77)					(1.03)
𝟙(NonRoutine *sell*)	-0.022	-0.034					-0.057
	(-0.39)	(-0.60)					(-1.05)
𝟙(pre-QEA *buy*)			0.377*	-0.136			-0.188
			(1.90)	(-0.68)			(-0.93)
𝟙(pre-QEA *sell*)			0.116	0.107			0.120
			(1.01)	(1.03)			(1.13)
𝟙(SSN)					0.203***	0.187**	0.191**
					(2.61)	(2.44)	(2.57)
𝟙(PPN)					-0.329*	-0.383**	-0.410**
					(-1.78)	(-2.05)	(-2.16)
log(Size)	0.017	0.023	0.016	0.023	0.014	0.020	0.022
	(0.33)	(0.46)	(0.32)	(0.44)	(0.27)	(0.40)	(0.42)
log(BM)	0.238**	0.239**	0.241**	0.240**	0.240**	0.241**	0.242**
	(2.36)	(2.38)	(2.38)	(2.39)	(2.37)	(2.40)	(2.42)
$Ret_{t-12,t-1}$	0.075	0.076	0.074	0.075	0.080	0.081	0.085
	(0.30)	(0.30)	(0.29)	(0.30)	(0.31)	(0.32)	(0.34)
$Ret_{t=0}$	-1.713***	-1.728***	-1.722***	-1.731***	-1.716***	-1.725***	-1.720***
	(-4.52)	(-4.54)	(-4.51)	(-4.52)	(-4.50)	(-4.51)	(-4.51)
Asset growth	-0.657***	-0.659***	-0.654***	-0.660***	-0.653***	-0.659***	-0.660***
	(-5.45)	(-5.49)	(-5.42)	(-5.49)	(-5.40)	(-5.49)	(-5.51)
Profitability	0.775***	0.767***	0.761***	0.761***	0.758***	0.757***	0.759***
	(4.05)	(4.01)	(3.98)	(3.98)	(3.97)	(3.97)	(3.99)
Illiquidity	0.021**	0.021**	0.021**	0.021**	0.021**	0.021**	0.021**
	(2.00)	(2.01)	(2.01)	(2.02)	(2.01)	(2.02)	(2.01)
NPR	0.035	0.026	0.073*	0.038	0.081**	0.044	0.035
	(0.96)	(0.71)	(1.89)	(1.03)	(2.07)	(1.17)	(0.96)
N	853,110	853,110	853,110	853110	853,110	853,110	853,110

Table 6. Information hypothesis: MLS and future earnings announcement return

This table reports the results of a three-day cumulative abnormal return CAR (in %) around a quarterly earnings announcement on *MLS* in the previous quarter. The dependent variable *CAR* is 3-day abnormal return calculated as daily stock return minus return on the CRSP value-weighted portfolio. *MLS* is the insider trading signal. $\mathbb{1}$(MLS *buy*) ($\mathbb{1}$(MLS *sell*)) is an indicator variable equal to 1 if MLS is in the top 20% (bottom 20%) and 0 otherwise. Columns (1)-(3) report the results using panel regressions. Standard errors are double-clustered at firm and quarter level. Industry and quarter fixed effects are included as indicated. Column (4) reports the result using Fama-Macbeth regression and standard errors corrected using Newey and West (1987) with 4 lags. *log*(Size) and *log*(BM) are the natural logarithms of the firm market equity and book-to-market. $Ret_{t-12,t-1}$ is stock return between month $t-12$ and $t-1$. $Ret_{t=0}$ is the stock return in the month before the earnings announcement. *NPR* is the insider net purchase ratio defined in Lakonishok and Lee (2001). We winsorize *Size* and *log*(BM) at 1% and 99% levels. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002 Q1 to 2023 Q4. 1%, 5%, and 10% statistical significance is indicated with ***, **, and *, respectively.

	(1) OLS	(2) OLS	(3) OLS	(4) Fama-Macbeth
$\mathbb{1}$(MLS *buy*)	0.481***		0.477***	0.543***
	(6.87)		(6.38)	(7.23)
$\mathbb{1}$(MLS *sell*)		0.091	0.067	0.068
		(1.46)	(1.14)	(1.24)
CAR $t-1$	0.012***	0.011***	0.012***	0.012**
	(3.43)	(3.27)	(3.43)	(2.40)
log(Size)	0.005	-0.001	0.002	0.037
	(0.24)	(-0.07)	(0.10)	(1.33)
log(BM)	0.200***	0.199***	0.201***	0.211***
	(4.92)	(4.89)	(4.47)	(4.26)
$Ret_{t=0}$	-0.346	-0.331	-0.346	-0.394
	(-1.28)	(-1.22)	(-1.23)	(-1.57)
$Ret_{t-12,t-1}$	0.042	0.044	0.041	0.122
	(0.52)	(0.54)	(0.49)	(1.45)
NPR	0.000	0.105***	0.010	-0.028
	(0.01)	(3.13)	(0.26)	(-0.84)
Quarter fixed effects	Yes	Yes	Yes	
Industry fixed effects	Yes	Yes	Yes	
N	310,499	310,499	310,499	310,499
$AdjustedR^2$	0.01	0.01	0.01	0.01

Table 7. Information environment: MLS and complexity/disclosure/and and firm information environment

This table reports the time-series average of equal-weighted returns (in percent) for portfolios formed by sorting dependently on the *MLS* and firm information environment, measured by firm complexity (Loughran and McDonald, 2024), managerial guidance, and accounting conservatism. At the end of each month, We first sort the stocks into 3 groups based on these three measures, and then in each group, we sort the stocks into 10 deciles based on the *MLS*. Long (Short) refers to stocks in the top (bottom) decile based on *MLS*. Panel A reports the results of double sorting on firm complexity. Panel B reports the results of double sorting on managerial guidance in past 6 month. And Panel C reports the results of double sorting on accounting conservatism. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period for Panel B and C is from 2002.01 to 2023.11, for Panel A is from 2002.01 to 2021.12.

Panel A. Sort on firm complexity

		MLS deciles										
Complexity		1	2	3	4	5	6	7	8	9	10	10−1
Low	Ret	1.06	1.11	0.68	0.79	0.95	0.92	1.20	1.16	1.41	1.90	0.85
	t-stat	(2.98)	(3.39)	(2.26)	(2.28)	(2.71)	(3.24)	(4.37)	(3.31)	(3.02)	(3.85)	(2.20)
Medium	Ret	1.17	1.16	0.83	0.97	0.89	1.03	0.93	1.82	1.78	2.08	0.92
	t-stat	(2.48)	(3.32)	(2.35)	(2.69)	(2.38)	(2.82)	(2.36)	(3.23)	(3.45)	(4.10)	(2.36)
High	Ret	0.88	0.96	0.63	0.82	0.89	0.83	1.21	1.08	1.36	2.50	1.67
	t-stat	(2.11)	(2.60)	(1.88)	(2.17)	(2.55)	(2.14)	(3.28)	(2.55)	(3.23)	(3.35)	(3.30)

Panel B. Sort on managerial guidance

		MLS deciles										
Guidance		1	2	3	4	5	6	7	8	9	10	10−1
≥ 5	Ret	0.97	0.99	0.96	0.93	0.83	0.94	1.19	1.27	1.47	1.69	0.72
	t-stat	(2.25)	(2.77)	(2.95)	(3.22)	(2.33)	(2.86)	(3.57)	(3.28)	(3.53)	(3.37)	(2.28)
≥ 1, < 5	Ret	0.88	0.76	0.90	0.92	0.73	1.03	0.74	1.22	1.27	2.02	1.14
	t-stat	(1.98)	(2.16)	(2.51)	(2.64)	(2.01)	(2.98)	(1.70)	(3.30)	(2.85)	(2.95)	(2.38)
0	Ret	1.19	0.98	0.74	0.84	0.77	0.60	1.20	1.40	1.84	2.83	1.64
	t-stat	(2.31)	(2.73)	(2.24)	(2.55)	(1.97)	(1.60)	(2.99)	(3.39)	(4.05)	(5.18)	(3.91)

Panel C. Sort on accounting conservatism

		MLS deciles										
Conservatism		1	2	3	4	5	6	7	8	9	10	10−1
Low	Ret	0.99	0.86	0.91	0.98	0.84	0.94	0.77	1.10	1.20	1.31	0.31
	t-stat	(2.29)	(2.48)	(2.81)	(3.33)	(2.23)	(2.85)	(2.49)	(3.53)	(3.23)	(3.11)	(0.91)
Median	Ret	1.00	0.93	0.57	1.02	0.74	0.99	0.86	1.16	1.23	1.89	0.90
	t-stat	(2.18)	(2.34)	(1.71)	(2.82)	(1.96)	(3.07)	(2.50)	(3.24)	(3.27)	(3.58)	(2.24)
High	Ret	0.77	0.89	0.64	1.09	0.99	1.35	1.34	1.53	1.96	2.46	1.69
	t-stat	(1.56)	(2.12)	(1.29)	(2.46)	(2.20)	(2.67)	(2.94)	(3.01)	(3.50)	(3.81)	(4.38)

Table 8. Fama-MacBeth regression: Tradeable investment strategy

This table reports the average coefficient estimates from cross-sectional Fama-MacBeth regressions of stock returns on alternative measures of MLS. The dependent variable is next month stock return $t+1$. $MLS^{real-time}$ in the machine learned signal measure that omits trades reported to the Edgar system after the month end $t=0$. MLS_{t-1} is the MLS measure from month $t-1$. *MLS buy (MLS sell)* is an indicator variable equal to 1 if MLS is in the top 20% (bottom 20%), and zero otherwise. *Size* and *log(BM)* are the natural logarithm of the firm market equity and book-to-market, respectively. $Ret_{t-12,t-1}$ is the stock return from $t-12$ to $t-1$. $Ret_{t=0}$ is the prior month return. *AssetGrowth* is annual growth rate of total assets. *Profitability* is firm gross profits to assets. *Illiquidity* is Amihud illiquidity measure. We winsorize *Size*, *log(BM)*, *AssetGrowth*, *Profitability* and *Illiquidity* at 1% and 99% levels. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. 1%, 5%, and 10% statistical significance is indicated with ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
$\mathbb{1}(\text{MLS } buy)^{real-time}$	0.710***		0.715***			
	(4.97)		(5.00)			
$\mathbb{1}(\text{MLS } sell)^{real-time}$		0.168	0.183*			
		(1.63)	(1.76)			
$\mathbb{1}(\text{MLS } buy)_{t-1}$				0.445***		0.445***
				(5.00)		(4.98)
$\mathbb{1}(\text{MLS } sell)_{t-1}$					0.044	0.055
					(0.53)	(0.67)
$log(\text{Size})$	0.015	0.006	0.013	0.012	0.004	0.011
	(0.30)	(0.12)	(0.25)	(0.24)	(0.09)	(0.22)
$log(\text{BM})$	0.243**	0.249**	0.246**	0.222**	0.225**	0.223**
	(2.35)	(2.42)	(2.40)	(2.13)	(2.18)	(2.16)
$Ret_{t-12,t-1}$	0.072	0.062	0.067	0.045	0.038	0.044
	(0.28)	(0.24)	(0.26)	(0.18)	(0.15)	(0.17)
$Ret_{t=0}$	-1.722***	-1.725***	-1.736***	-1.752***	-1.740***	-1.752***
	(-4.51)	(-4.52)	(-4.54)	(-4.42)	(-4.39)	(-4.42)
Asset growth	-0.653***	-0.654***	-0.657***	-0.643***	-0.646***	-0.648***
	(-5.39)	(-5.45)	(-5.47)	(-5.40)	(-5.44)	(-5.47)
Profitability	0.758***	0.743***	0.752***	0.781***	0.772***	0.778***
	(3.97)	(3.90)	(3.94)	(3.93)	(3.89)	(3.92)
Illiquidity	0.022**	0.021**	0.021**	0.023**	0.022**	0.022**
	(2.04)	(2.00)	(2.03)	(2.13)	(2.08)	(2.11)
N	853,110	853,110	853,110	827,430	827,430	827,430

Table 9. Performance of other machine learning portfolios: Portfolio sorts

This table reports the performance of *MLS*-sorted portfolios. All stocks are sorted into deciles based on their predicted returns for the next month. Columns "MLS," "Ret," "NW-t," and "SR" provide the average *MLS* for each decile, the average realized monthly returns, their Newey-West adjusted t-statistics, and Sharpe ratios, respectively. Panel A reports the equal-weighted returns and Panel B reports the value-weighted returns. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is 2002.01 – 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

Panel A. Equal Weighted

	OLS				Ridge				Lasso				Random forest				XGBoost				NN1				NN2			
	Ret	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*
L	1.06	2.48	0.54	0.19	1.06	2.48	0.53	0.19	1.03	2.52	0.53	0.32	0.96	2.15	0.46	0.09	0.98	2.25	0.47	0.02	1.01	2.52	0.53	-0.76	0.90	2.27	0.47	-0.30
2	0.87	2.46	0.48	0.80	0.87	2.46	0.48	0.80	0.98	2.55	0.54	0.86	0.91	2.31	0.54	0.41	0.86	2.50	0.50	0.38	0.89	2.64	0.49	0.18	0.84	2.37	0.46	0.36
3	0.98	2.82	0.55	0.95	0.98	2.80	0.55	0.95	0.80	2.28	0.44	0.98	1.07	2.78	0.64	0.42	0.82	2.46	0.47	0.48	0.85	2.50	0.48	0.50	0.90	2.53	0.52	0.59
4	0.83	2.45	0.46	1.02	0.83	2.46	0.47	1.02	0.78	2.18	0.43	1.04	0.82	2.41	0.45	0.51	0.88	2.76	0.53	0.58	0.94	2.63	0.53	0.72	0.94	2.87	0.55	0.76
5	1.16	3.41	0.65	1.07	1.16	3.41	0.65	1.07	0.95	2.39	0.52	1.08	0.90	2.61	0.51	0.64	0.86	2.51	0.48	0.69	0.87	2.46	0.48	0.90	0.91	2.68	0.53	0.91
6	1.03	3.02	0.59	1.11	1.03	3.01	0.59	1.11	0.94	2.21	0.48	1.12	1.01	3.09	0.56	0.79	0.97	2.86	0.54	0.84	1.11	3.14	0.63	1.08	1.00	3.05	0.58	1.06
7	1.05	3.19	0.59	1.16	1.05	3.20	0.59	1.16	0.91	2.45	0.46	1.16	0.99	2.81	0.53	1.01	1.02	2.91	0.54	1.04	1.05	3.09	0.61	1.28	1.02	2.95	0.58	1.25
8	1.17	3.15	0.66	1.22	1.17	3.14	0.66	1.22	1.25	3.21	0.67	1.20	1.23	3.35	0.66	1.36	1.26	3.30	0.68	1.34	1.23	3.50	0.69	1.53	1.26	3.51	0.68	1.51
9	1.27	3.16	0.68	1.30	1.27	3.16	0.68	1.30	1.40	3.24	0.69	1.25	1.34	3.21	0.71	1.96	1.27	3.08	0.66	1.82	1.36	3.53	0.73	1.86	1.34	3.32	0.69	1.85
H	1.39	3.04	0.70	1.60	1.39	3.04	0.70	1.60	1.48	3.20	0.76	1.46	1.94	3.60	0.91	2.92	1.94	3.71	0.93	2.82	1.50	3.10	0.72	3.11	1.69	3.31	0.78	2.86
H–L	0.34	1.80	0.44	1.41	0.34	1.80	0.44	1.41	0.45	2.07	0.55	1.15	0.98	3.24	0.98	2.84	0.96	3.06	0.94	2.80	0.49	2.21	0.57	3.87	0.79	3.09	0.84	3.16

Panel B. Value Weighted

	OLS				Ridge				Lasso				Random forest				XGBoost				NN1				NN2			
	Ret	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*	*Ret*	*t*	*SR*	*MLS*
L	0.87	2.51	0.51	0.19	0.87	2.51	0.51	0.19	0.98	2.68	0.58	0.32	0.93	2.40	0.50	0.09	1.14	2.94	0.65	0.02	0.68	1.95	0.38	-0.76	0.79	1.98	0.45	-0.30
2	0.84	2.34	0.55	0.80	0.84	2.34	0.55	0.80	0.94	2.67	0.59	0.86	0.89	2.24	0.57	0.41	0.69	2.21	0.44	0.38	1.05	3.38	0.67	0.18	0.85	2.79	0.52	0.36
3	0.79	2.66	0.51	0.95	0.79	2.65	0.51	0.95	0.83	2.69	0.53	0.98	0.91	2.74	0.60	0.42	0.77	2.71	0.49	0.48	0.78	2.67	0.50	0.50	0.81	2.50	0.51	0.59
4	0.72	2.60	0.47	1.02	0.72	2.60	0.47	1.02	0.71	2.62	0.49	1.04	0.78	2.71	0.47	0.51	0.80	2.80	0.55	0.58	0.72	2.16	0.45	0.72	0.76	2.60	0.49	0.76
5	0.86	3.17	0.58	1.07	0.86	3.17	0.58	1.07	0.55	1.61	0.31	1.08	0.81	2.80	0.53	0.64	0.76	2.62	0.48	0.69	0.71	2.57	0.47	0.90	0.93	3.48	0.61	0.91
6	0.84	2.90	0.54	1.11	0.84	2.90	0.54	1.11	0.84	2.58	0.47	1.12	1.00	3.26	0.62	0.79	1.01	3.17	0.63	0.84	0.90	2.65	0.55	1.08	0.81	2.59	0.53	1.06
7	1.06	3.17	0.61	1.16	1.06	3.18	0.61	1.16	0.65	1.73	0.31	1.16	1.12	3.27	0.64	1.01	1.08	3.31	0.60	1.04	0.92	2.83	0.56	1.28	0.94	2.90	0.56	1.25
8	0.99	2.67	0.55	1.22	0.99	2.67	0.55	1.22	0.72	1.80	0.37	1.20	1.13	3.18	0.62	1.36	0.98	2.60	0.51	1.34	0.95	3.13	0.56	1.53	0.98	2.46	0.51	1.51
9	0.96	2.83	0.52	1.30	0.96	2.82	0.52	1.30	1.11	2.87	0.55	1.25	0.93	2.33	0.46	1.96	0.96	2.42	0.48	1.82	1.04	2.46	0.56	1.86	0.90	2.41	0.46	1.85
H	0.90	2.04	0.42	1.60	0.90	2.04	0.42	1.60	1.06	2.93	0.53	1.46	1.96	4.34	0.85	2.92	1.85	3.93	0.79	2.82	1.22	3.26	0.63	3.11	1.30	4.26	0.64	2.86
H–L	0.03	0.10	0.03	1.41	0.03	0.09	0.03	1.41	0.08	0.30	0.07	1.15	1.03	3.37	0.69	2.84	0.71	1.81	0.44	2.80	0.54	2.11	0.44	3.87	0.51	1.59	0.35	3.16

Table 10. Fama-MacBeth regression: By insider types

This table reports the average coefficient estimates from cross-sectional Fama-MacBeth regressions. We decompose *MLS* into those containing trading signal from senior officers (defined as CEO or CFO) and those from other insiders; directors and other insiders; independent and other insiders. The dependent variable is the monthly return $t+1$. *MLS buy (MLS sell)* is an indicator variable equal to 1 if *MLS* is in the top 20% (bottom 20%) and 0 otherwise. The control variables include *Size*, *log(BM)*, $Ret_{t-12,t-1}$, $Ret_{t=0}$, *AssetGrowth*, *Profitability*, and *Illiquidity*. We winsorize *Size*, *log(BM)*, *AssetGrowth*, *Profitability* and *Illiquidity* at 1% and 99% levels. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. 1%, 5%, and 10% statistical significance is indicated with ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)
$\mathbb{1}$(MLS *buy* Senior)	0.953***		1.002***		
	(4.71)		(4.85)		
$\mathbb{1}$(MLS *sell* Senior)	0.178		0.196		
	(1.27)		(1.38)		
$\mathbb{1}$(MLS *buy* NonSenior)		0.966***	0.995***		
		(6.87)	(6.92)		
$\mathbb{1}$(MLS *sell* NonSenior)		-0.024	-0.004		
		(-0.30)	(-0.05)		
$\mathbb{1}$(MLS *buy* Director)				0.879***	
				(5.47)	
$\mathbb{1}$(MLS *sell* Director)				0.156*	
				(1.75)	
$\mathbb{1}$(MLS *buy* NonDirector)				1.344***	
				(4.59)	
$\mathbb{1}$(MLS *sell* NonDirector)				-0.007	
				(-0.06)	
$\mathbb{1}$(MLS *buy* Independent)					0.670***
					(4.00)
$\mathbb{1}$(MLS *sell* Independent)					0.052
					(0.61)
$\mathbb{1}$(MLS *buy* Dependent)					0.801***
					(5.70)
$\mathbb{1}$(MLS *sell* Dependent)					0.064
					(0.63)
Controls?	Yes	Yes	Yes	Yes	Yes
N	853,110	853,110	853,110	853,110	853,110

INTERNET APPENDIX
for
Can Machines Learn from Trading Data?

NOT INTENDED FOR PUBLICATION

Table IA.1. Correlations

This table reports Pearson correlations between the main variables and the insider trading measures. The sample period is from 2002.01 to 2023.11.

Panel A. Correlation of main variables

	$\mathbb{1}$(MLS buy)	$\mathbb{1}$(MLS $sell$)	Ret	Size	log(BM)	$Ret_{t-12,t-2}$	Ret_{t-1}	Asset growth	Profitability	Illiquidity	NPR
$\mathbb{1}$(MLS buy)	1										
$\mathbb{1}$(MLS $sell$)	-0.064	1									
Ret	0.014	-0.004	1								
Size	-0.135	0.220	-0.011	1							
log(BM)	0.068	-0.149	0.025	-0.328	1						
$Ret_{t-12,t-2}$	-0.024	0.063	0.005	0.061	-0.177	1					
Ret_{t-1}	-0.002	0.027	0.004	0.016	0.037	-0.011	1				
Asset growth	-0.002	0.049	-0.025	0.070	-0.109	-0.046	-0.031	1			
Profitability	-0.037	0.077	0.009	0.052	-0.222	0.017	-0.001	0.155	1		
Illiquidity	0.055	-0.055	0.014	-0.348	0.179	0.008	0.037	-0.077	-0.030	1	
NPR	0.177	-0.206	0.009	-0.387	0.261	-0.117	0.027	-0.040	-0.154	0.162	1

Panel B. Correlation of insider trading measures

	$\mathbb{1}$(MLS buy)	$\mathbb{1}$(MLS $sell$)	$\mathbb{1}$(Nonroutine buy)	$\mathbb{1}$(Nonroutine sell)	$\mathbb{1}$(pre-QEA buy)	$\mathbb{1}$(pre-QEA sell)	$\mathbb{1}$(SSN)	$\mathbb{1}$(PPN)	NPR
$\mathbb{1}$(MLS buy)	1.000								
$\mathbb{1}$(MLS $sell$)	-0.064	1.000							
$\mathbb{1}$(NonRoutine buy)	0.563	-0.059	1.000						
$\mathbb{1}$(NonRoutine $sell$)	-0.056	0.423	0.002	1.000					
$\mathbb{1}$(pre-QEA buy)	0.214	-0.020	0.279	-0.007	1.000				
$\mathbb{1}$(pre-QEA $sell$)	-0.019	0.168	-0.012	0.322	0.008	1.000			
$\mathbb{1}$(SSN	-0.021	0.098	-0.003	0.145	-0.003	0.076	1.000		
$\mathbb{1}$(PPN	0.047	-0.012	0.074	-0.008	0.029	-0.005	-0.004	1.000	
NPR	0.177	-0.206	0.172	-0.318	0.072	-0.131	-0.146	0.085	1.000

Table IA.2. Portfolio sorts: MLS using default parameters

This table reports the raw returns (%), risk adjusted return (Fama-French 5 factors + momentum factor), t-statistics, and the Sharpe ratio of portfolios based on MLS constructed from a default LightGBM model. At the end of each month, we rank stocks into 10 groups based on MLS and construct the long-short (10−1) portfolio. Stocks are held in the portfolios for one month, and the portfolios are rebalanced at the end of each month. Panel A reports the portfolio returns sorted based on deciles of MLS. Panel B reports portfolio returns double-sorted by market capitalization and deciles of MLS. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

		MLS deciles										
		1	2	3	4	5	6	7	8	9	10	10−1
MLS	$\hat{Ret}_{j,t+1}^{GBM}$	-0.06	0.35	0.40	0.48	0.60	0.73	0.91	1.18	1.60	2.70	2.76
Equal-weighted	Ret	1.02	0.85	0.75	0.69	0.80	1.02	1.04	1.16	1.27	1.98	0.96
	t-stat	(2.33)	(2.58)	(2.32)	(2.13)	(2.20)	(2.98)	(3.03)	(3.06)	(3.14)	(3.69)	(3.01)
	α	0.18	0.01	0.01	-0.14	-0.12	0.19	0.26	0.33	0.47	1.19	1.02
	t-stat	(1.80)	(0.10)	(0.13)	(-1.33)	(-0.97)	(1.81)	(2.72)	(2.79)	(3.09)	(5.34)	(4.20)
	SR	0.49	0.51	0.44	0.38	0.41	0.58	0.57	0.62	0.68	0.92	0.94
Value-weighted	Ret	1.06	0.67	0.59	0.73	0.76	1.15	1.11	0.78	1.00	1.57	0.51
	t-stat	(2.68)	(2.16)	(1.99)	(2.68)	(2.23)	(3.12)	(3.85)	(2.00)	(2.49)	(3.10)	(1.44)
	α	0.22	-0.11	-0.15	-0.02	-0.12	0.29	0.27	-0.13	0.04	0.66	0.44
	t-stat	(1.66)	(-1.19)	(-1.43)	(-0.12)	(-0.88)	(2.13)	(2.33)	(-0.87)	(0.21)	(2.92)	(1.71)
	SR	0.59	0.42	0.38	0.46	0.44	0.68	0.67	0.41	0.47	0.66	0.33

Table IA.3. Portfolio sorts: MLS using 10 features

This table reports the raw returns (%), risk adjusted return (Fama-French 5 factors + momentum factor), t-statistics, and the Sharpe ratio of portfolios based on MLS constructed from 10 features: $\%TradeBuy$, $\%TradeSell$, $\$TradeBuy$, $\$TradeSell$, $\#TradeBuy$, $\#TradeSell$, $EstRet_{t-3}$, $EstRet_{t-6}$, $EstRet_{t-12}$, $Ownership\%$. At the end of each month, we rank stocks into 10 groups based on MLS and construct the long-short (10−1) portfolio. Stocks are held in the portfolios for one month, and the portfolios are rebalanced at the end of each month. Panel A reports the portfolio returns sorted based on deciles of MLS. Panel B reports portfolio returns double-sorted by market capitalization and deciles of MLS. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

		MLS deciles										
		1	2	3	4	5	6	7	8	9	10	10−1
MLS	$\hat{Ret}_{j,t+1}^{GBM}$	-0.07	0.31	0.36	0.44	0.53	0.66	0.84	1.09	1.48	2.37	2.44
Equal-weighted	Ret	0.95	0.86	0.90	1.03	0.85	0.96	0.99	1.17	1.35	1.98	1.03
	t-stat	(2.11)	(2.53)	(2.79)	(2.89)	(2.61)	(2.78)	(2.92)	(3.16)	(3.10)	(3.81)	(3.59)
	α	0.12	0.03	-0.07	0.07	-0.02	0.12	0.17	0.35	0.54	1.21	1.09
	t-stat	(1.20)	(0.35)	(-0.89)	(1.27)	(-0.35)	(1.23)	(2.12)	(3.05)	(3.18)	(5.64)	(5.20)
	SR	0.45	0.51	0.52	0.60	0.48	0.54	0.54	0.61	0.72	0.94	1.05
Value-weighted	Ret	1.01	0.71	0.82	0.96	0.82	1.27	1.00	0.87	0.89	1.75	0.75
	t-stat	(2.70)	(2.10)	(2.84)	(3.07)	(2.64)	(3.77)	(3.14)	(2.48)	(2.16)	(3.97)	(2.65)
	α	0.19	-0.05	-0.09	0.09	-0.02	0.35	0.10	-0.07	0.04	0.86	0.66
	t-stat	(1.44)	(-0.46)	(-0.78)	(1.00)	(-0.17)	(2.97)	(0.81)	(-0.60)	(0.18)	(4.45)	(3.14)
	SR	0.57	0.45	0.53	0.63	0.51	0.77	0.57	0.45	0.42	0.77	0.53

Table IA.4. Portfolio sorts: MLS using 13 features

This table reports the raw returns (%), risk adjusted return (Fama-French 5 factors + momentum factor), t-statistics, and the Sharpe ratio of portfolios based on MLS constructed from 13 features: $\%NetTrade$, $\$NetTrade$, $\#NetTrade$, $EstRet_{t-3}$, $EstRet_{t-6}$, $EstRet_{t-12}$, $\$Trade_{t-3}$, $\$Trade_{t-6}$, $\$Trade_{t-12}$, $\#Trade_{t-3}$, $\#Trade_{t-6}$, $\#Trade_{t-12}$, $Ownership\%$. At the end of each month, we rank stocks into 10 groups based on MLS and construct the long-short (10−1) portfolio. Stocks are held in the portfolios for one month, and the portfolios are rebalanced at the end of each month. Panel A reports the portfolio returns sorted based on deciles of MLS. Panel B reports portfolio returns double-sorted by market capitalization and deciles of MLS. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

		MLS deciles										
		1	2	3	4	5	6	7	8	9	10	10−1
MLS	$\hat{Ret}_{j,t+1}^{GBM}$	-0.08	0.32	0.37	0.45	0.54	0.66	0.82	1.06	1.47	2.37	2.45
Equal-weighted	Ret	1.00	0.88	0.87	0.74	0.90	0.95	0.99	1.21	1.27	2.04	1.04
	t-stat	(2.17)	(2.52)	(2.76)	(2.34)	(2.61)	(2.78)	(2.82)	(3.31)	(3.08)	(3.85)	(3.34)
	α	0.16	0.06	-0.02	-0.20	-0.03	0.11	0.21	0.37	0.48	1.24	1.08
	t-stat	(1.43)	(0.93)	(-0.26)	(-1.89)	(-0.43)	(1.02)	(2.79)	(3.35)	(2.93)	(5.81)	(4.77)
	SR	0.47	0.51	0.54	0.42	0.50	0.54	0.54	0.65	0.68	0.94	1.00
Value-weighted	Ret	1.09	0.79	0.77	0.63	0.73	1.03	1.16	0.79	1.01	1.81	0.72
	t-stat	(2.52)	(2.63)	(2.86)	(2.50)	(2.33)	(2.83)	(3.51)	(2.34)	(2.47)	(3.69)	(1.92)
	α	0.27	0.01	-0.07	-0.18	-0.17	0.22	0.31	-0.12	0.08	0.93	0.66
	t-stat	(1.71)	(0.17)	(-0.80)	(-1.73)	(-1.66)	(1.70)	(2.35)	(-0.79)	(0.36)	(3.90)	(2.28)
	SR	0.61	0.52	0.55	0.42	0.45	0.62	0.67	0.42	0.49	0.78	0.44

Table IA.5. Fama-MacBeth regression for 2010–2019

This table reports the the average cross-sectional coefficient estimates from Fama-MacBeth regressions of stock returns on MLS. The dependent variable is next month stock return $t+1$. MLS in the Machine Learned Signal. $\mathbb{1}$(MLS buy) and $\mathbb{1}$(MLS $sell$) are indicator variables equal to one if MLS is in the top 20% (bottom 20%), and 0 otherwise. *Size* and *log(BM)* are the natural logarithms of the firm market equity and book-to-market. $Ret_{t-12,t-2}$ is stock return from month $t-12$ to $t-2$. $Ret_{t=0}$ is the return in month $t=0$. *AssetGrowth* is annual growth rate of total assets. *Profitability* is firm gross profits to assets. *Illiquidity* is Amihud illiquidity measure. We winsorize *Size*, *log(BM)*, *AssetGrowth*, *Profit* and *Illiquidity* at 1% and 99% levels. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2010.01 to 2019.12. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. 1%, 5%, and 10% statistical significance is indicated with ***, **, and *, respectively.

	(1)	(2)	(3)
$\mathbb{1}$(MLS buy)	0.556***		0.561***
	(4.44)		(4.46)
$\mathbb{1}$(MLS $sell$)		0.110	0.127
		(1.41)	(1.60)
log(Size)	0.073	0.064	0.071
	(1.63)	(1.43)	(1.59)
log(BM)	0.062	0.067	0.064
	(0.61)	(0.66)	(0.64)
$Ret_{t-12,t-1}$	0.196	0.188	0.192
	(1.06)	(1.02)	(1.04)
$Ret_{t=0}$	-1.670***	-1.686***	-1.688***
	(-3.29)	(-3.31)	(-3.32)
Asset growth	-0.570***	-0.570***	-0.572***
	(-3.53)	(-3.55)	(-3.56)
Profitability	0.420*	0.407*	0.416*
	(1.73)	(1.67)	(1.71)
Illiquidity	0.016	0.016	0.016
	(1.62)	(1.62)	(1.62)
N	357,966	357,966	357,966

Table IA.6. Fama-MacBeth regression: Omitting features

This table reports the results of Fama-Macbeth regressions of stock returns on MLS. The dependent variable is next month stock return $t+1$. MLS(omit #Trade), MLS(omit %Trade), MLS(omit $Trade), and MLS(omit Ret_{t-1}) MLS(omit Ownership%) are versions of MLS that are trained using a subset of the explanatory variables. $\mathbb{1}$(MLS buy) $\mathbb{1}$(MLS sell) is an indicator variable equal to 1 if MLS is in the top 20% (bottom 20%) and 0 otherwise. The universe is all stocks listed on NYSE/AMEX/NASDAQ with price above $1 and negative book value firms are discarded. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. 1%, 5%, and 10% statistical significance is indicated with ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)
$\mathbb{1}$(MLS buy omit #$trade$)	0.944***				
	(7.13)				
$\mathbb{1}$(MLS $sell$ omit #$trade$)	0.120				
	(1.36)				
$\mathbb{1}$(MLS buy omit %$trade$)		0.821***			
		(6.41)			
$\mathbb{1}$(MLS $sell$ omit %$trade$)		0.064			
		(0.70)			
$\mathbb{1}$(MLS buy omit \$$trade$)			0.947***		
			(7.10)		
$\mathbb{1}$(MLS $sell$ omit \$$trade$)			0.071		
			(0.78)		
$\mathbb{1}$(MLS buy omit Ret)				0.970***	
				(6.08)	
$\mathbb{1}$(MLS $sell$ omit Ret)				0.048	
				(0.63)	
$\mathbb{1}$(MLS buy omit %$ownership$)					0.952***
					(7.05)
$\mathbb{1}$(MLS $sell$ omit %$ownership$)					0.081
					(0.84)
log(Size)	0.019	0.015	0.019	0.022	0.019
	(0.36)	(0.29)	(0.37)	(0.43)	(0.37)
log(BM)	0.245**	0.243**	0.243**	0.242**	0.244**
	(2.39)	(2.37)	(2.37)	(2.36)	(2.39)
$Ret_{t-12,t-1}$	0.073	0.072	0.073	0.076	0.072
	(0.28)	(0.28)	(0.29)	(0.30)	(0.28)
$Ret_{t=0}$	-1.731***	-1.717***	-1.722***	-1.722***	-1.733***
	(-4.54)	(-4.51)	(-4.53)	(-4.51)	(-4.54)
Asset growth	-0.657***	-0.659***	-0.660***	-0.655***	-0.657***
	(-5.46)	(-5.47)	(-5.48)	(-5.44)	(-5.47)
Profitability	0.757***	0.764***	0.761***	0.761***	0.755***
	(3.97)	(4.01)	(3.99)	(3.99)	(3.96)
Illiquidity	0.021**	0.022**	0.022**	0.021**	0.021**
	(2.01)	(2.06)	(2.04)	(1.99)	(2.00)
N	853110	853110	853110	853110	853110
R^2	0.04	0.04	0.04	0.04	0.04

Table IA.7. MLS and R&D

This table reports the time-series average of equal-weighted returns (in percent) for portfolios formed by sorting dependently on the MLS and the R&D expense over market value of equity. At the end of each month, We first sort the stocks into three groups: missing R&D, R&D below the monthly median, and R&D above the monthly median. Then in each group, we sort the stocks into 10 deciles based on the MLS. Long (Short) refers to stocks in the top (bottom) decile based on MLS. The sample consists of common stocks that are listed on NYSE/AMEX/NASDAQ greater than $1. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

		1	2	3	4	5	6	7	8	9	10	10−1
Missing R&D	Ret	0.88	0.77	0.94	0.73	0.78	0.88	1.18	1.02	1.11	1.77	0.90
	t-stat	(2.41)	(2.51)	(3.08)	(2.31)	(2.61)	(2.66)	(3.16)	(2.88)	(2.73)	(3.91)	(3.67)
Low R&D	Ret	0.98	0.95	0.94	0.79	1.00	0.73	1.02	0.86	1.42	2.05	1.07
	t-stat	(2.05)	(2.19)	(2.73)	(2.21)	(2.49)	(1.86)	(2.63)	(2.34)	(2.96)	(3.21)	(2.40)
High R&D	Ret	0.89	1.03	0.81	0.91	1.00	1.28	1.32	1.35	2.01	2.84	1.96
	t-stat	(1.52)	(2.34)	(2.01)	(2.58)	(2.34)	(2.55)	(2.42)	(2.59)	(3.21)	(3.76)	(3.88)

Table IA.8. Undervaluation hypothesis: MLS and firm valuation

This table reports the time-series average of equal-weighted returns (in percent) for portfolios formed by sorting dependently on the MLS and firm valuation. At the end of each month, We first sort the stocks into 3 groups based on industry-adjusted PS or the anomaly score, and then in each group, we sort the stocks into 10 deciles based on the MLS. Long (Short) refers to stocks in the top (bottom) decile based on MLS. Panel A reports the results of double sorting on PS. Panel B reports the results of double sorting on anomaly score, defined as the aggregation of 27 significant and robust anomalies considered in Hou, Xue, and Zhang (2015). The anomaly score is calculated as the sum of ternary signal for each anomaly. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags. The sample consists of all stocks listed on NYSE/AMEX/NASDAQ with price above $1, excluding firms with negative book value. The sample period is from 2002.01 to 2023.11.

Panel A. Sort on price-to-sales ratio

		MLS deciles										
P/S		1	2	3	4	5	6	7	8	9	10	10−1
Low	Ret	0.80	0.86	0.85	1.01	1.17	1.45	1.55	1.56	1.82	2.71	1.91
	t-stat	(1.59)	(2.39)	(2.70)	(2.73)	(3.02)	(3.38)	(3.00)	(3.39)	(3.41)	(3.85)	(4.71)
Medium	Ret	1.13	1.08	0.73	0.92	0.84	0.96	1.10	1.03	1.40	1.81	0.68
	t-stat	(3.06)	(3.69)	(2.06)	(2.74)	(2.45)	(3.01)	(3.36)	(2.78)	(3.56)	(3.76)	(2.33)
High	Ret	1.01	0.82	0.75	0.84	0.80	0.87	0.77	0.65	0.51	1.25	0.23
	t-stat	(1.76)	(1.86)	(1.62)	(2.31)	(2.19)	(2.07)	(1.83)	(1.57)	(1.40)	(2.36)	(0.71)

Panel B. Sort on anomaly score

		MLS deciles										
Anomaly		1	2	3	4	5	6	7	8	9	10	10−1
Overvaluation	Ret	0.67	0.70	0.58	0.54	0.69	0.68	0.61	0.45	0.75	1.33	0.66
	t-stat	(1.13)	(1.54)	(1.23)	(1.36)	(1.72)	(1.51)	(1.25)	(1.00)	(1.52)	(2.42)	(1.64)
Middle	Ret	1.00	0.88	0.86	1.01	1.03	1.10	1.01	1.37	1.23	1.97	0.97
	t-stat	(2.51)	(2.86)	(2.87)	(2.96)	(3.42)	(3.11)	(3.08)	(3.90)	(3.15)	(3.96)	(3.24)
Undervaluation	Ret	1.14	1.18	0.77	0.83	0.96	1.03	1.33	1.43	1.98	2.73	1.58
	t-stat	(2.95)	(4.18)	(2.33)	(2.79)	(2.84)	(3.29)	(3.97)	(3.44)	(4.72)	(4.24)	(3.64)

Table IA.9. MLS and Rule 10b5-1

This table reports the time-series average of equal-weighted returns (in percent) for portfolios sorted on the *MLS* constructed from transactions under 10b5-1 plan or others. The sample consists of common stocks that are listed on NYSE/AMEX/NASDAQ with price greater than $1. The sample period is from 2002.01 to 2023.11. t-statistics (in parentheses) are adjusted using Newey and West (1987) with 12 lags.

Panel A. MLS using non-Rule10b5-1 trades

	1	2	3	4	5	6	7	8	9	10	10−1
Ret	0.84	0.85	0.78	0.85	0.89	0.89	1.14	1.14	1.50	1.96	1.12
t-stat	(2.03)7	(2.80)	(2.57)	(2.41)	(2.71)	(2.64)	(2.94)	(3.12)	(3.21)	(3.80)	(3.71)
SR	0.41	0.52	0.45	0.46	0.49	0.49	0.60	0.61	0.75	0.91	1.01

Panel B. MLS using Rule10b5-1 trades

	1	2	3	4	5	6	7	8	9	10	10−1
Ret	1.33	1.58	1.13	0.65	0.89	0.89	1.32	0.99	0.60	1.18	-0.15
t-stat	(2.28)	(2.92)	(1.78)	(1.66)	(1.15)	(1.40)	(2.34)	(1.75)	(0.96)	(1.59)	(-0.24)
SR	0.46	0.62	0.34	0.31	0.36	0.36	0.52	0.38	0.20	0.43	-0.06