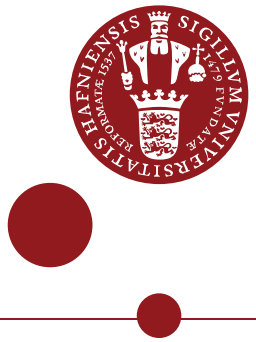
UNIVERSITY OF COPENHAGEN
DEPARTMENT OF ECONOMICS



Master Thesis

Ufuk Yasin & Jonas Theodor Ø. Schimdt

Empirical Asset Pricing via Machine Learning: The Predictive Content of Insider Trading

Stefan Voigt

22nd December 2025

Keystrokes: 142,748 (59.5 Normal Pages)

Empirical Asset Pricing via Machine Learning: The Predictive Content of Insider Trading

Ufuk Yasin & Jonas Theodor Ø. Schmidt*

December 22, 2025

Abstract

Insider trading anomalies are well documented in low-dimensional linear models, but their impact in high-dimensional machine learning is unclear. This thesis examines whether insider trading adds predictive value in the high-dimensional framework of Gu, Kelly, and Xiu (2020). We assess out-of-sample predictive performance for U.S. equities (2014–2021) using models ranging from simple linear regression to deep neural networks. Our findings are threefold. First, our results align with Gu, Kelly, and Xiu (2020), confirming that neural networks perform best. Second, we find no systematic evidence that insider trading provides predictive value within high-dimensional machine learning models. Third, we document heterogeneity across firm types. Insider trading enhances simple linear predictability across all firm sizes, particularly for large firms. In addition, among liquid stocks, insider trading adds predictive value in nonlinear models that appear robust to transaction costs. Overall, our results suggest that while insider signals contain information, this information is redundant when conditioned on the high-dimensional information set used in modern machine learning models.

Contributions:

Ufuk

Sections: 1, 2.1, 2.3, 3.1.1, 3.1.3, 4.1, 4.3.1, 4.3.3, 4.3.5, 4.5, 4.7.1, 5.1, 5.3, 5.4.2, 6.2, 7.2, 7.4, & 8

Jonas

Sections: 1, 2.2, 3.1.2, 3.1.4, 4.3.2, 4.3.4, 4.4, 4.6, 4.7.2, 4.7.4, 4.7.3, 5.2, 5.4.1, 5.4.3, 5.4.4 6.1, 7.1, 7.3, & 8

*We are deeply grateful to our supervisor, Stefan Voigt, for his invaluable guidance and expertise. His support and prior work were instrumental in shaping this research.

Contents

1. Introduction

Fifty years ago, Fama (1970) formalised the Efficient Market Hypothesis (EMH), in which the strong form posits that asset prices fully reflect all available information, leaving no scope for investors to systematically predict returns. While the EMH implies that return predictability should be negligible, subsequent research has challenged this view. Behavioural finance highlights market anomalies driven by psychological biases such as loss aversion and overconfidence (Daniel et al., 1998; Shiller, 2003; Kahneman and Tversky, 1979; Barber and Odean, 2001). This literature revived interest in return predictability, with empirical studies documenting predictive relationships between returns and macroeconomic variables or firm characteristics (Lettau and Ludvigson, 2001; Campbell, 2000). However, the evidence remains contested. Welch and Goyal (2008) argue that many predictive models fail out of sample, while Campbell and Thompson (2008) show that predictability only emerges under weak and economically plausible restrictions.

A key reason for this tension lies in how prediction is approached. Breiman (2001) distinguishes between a structural culture, which relies on parsimonious linear models to test economic theory, and a prediction culture, which prioritises forecasting accuracy and relies on machine learning. In financial returns, the structural culture faces severe limitations due to strong persistence and cross-sectional correlation among predictors, leading to unstable estimates and weak inference (Harvey et al., 2016). In contrast, machine learning models cast a wide net to identify drivers of returns by flexibly processing high-dimensional information and uncover nonlinear interactions that restrictive linear models ignore (Kelly et al., 2023).

Machine learning has proven effective at predicting returns (Gu et al., 2020; Drobetz and Otto, 2021; Leippold et al., 2022; Hanauer and Kalsbach, 2023), and the next step proposed is to expand the predictor set by including behavioural data (Giglio et al., 2022). Insider trading offers a natural starting point, as the literature suggests these trades have a behavioural edge in predicting returns (e.g. Lakonishok and Lee, 2001). Insider trading is an anomaly even Fama (1970) acknowledged as a challenge to strong-form market efficiency. Insiders can predict returns through their access to private information regarding firm fundamentals, investment opportunities, and future cash flows. Consequently, the predictive power of these trades varies across firms and markets in a nonlinear, noisy manner (Aboody and Lev, 2000; Piotroski and Roulstone, 2005; Cohen et al., 2012). While such complexities make insider trading difficult to exploit within a pre-specified linear framework, they are well-suited to the flexibility of machine learning methods. Therefore, it opens an exciting research question that this paper seeks to answer:

Does insider trading provide incremental predictive power for stock returns in a high-dimensional machine learning framework?

To answer this question, our paper applies the prediction culture. Building on the machine learning framework of Gu, Kelly, and Xiu (2020) with a high-dimensional predictor set referred to as the *Baseline*. We examine whether augmenting the Baseline with insider

trading improves return prediction. To structure the question, we distinguish between two perspectives. The *Corporate Insider* perspective captures private information available only to insiders at the time of trade execution. In contrast, the *Outsider* perspective reflects public information available to investors once trades are disclosed. This enables us to test whether the *Corporate Insider* can predict returns, while the *Outsiders* cannot. To evaluate the predictive value of insider trading, we organise our paper around five thematic dimensions and seven hypotheses.

- **Outsider:** Adding the *Outsider* information set to the *Baseline* improves both stock-level return predictability (H_{1a}) and economic performance, as measured by the Sharpe ratio (H_{1b}).
- **Corporate Insider:** Adding the *Corporate Insider* information set to the *Baseline* improves both stock-level return predictability (H_{2a}) and economic performance, as measured by Sharpe ratio (H_{2b}).
- **Information Advantage:** The *Corporate Insider* perspective provides a private informational advantage, allowing *Corporate Insiders* portfolios to outperform *Outsider* portfolios in terms of Sharpe ratio (H_3).
- **Heterogeneity:** Sharpe ratio gains from insider trading are more pronounced among small-cap firms, consistent with stronger information asymmetry (H_4).
- **Trading Frictions:** The economic performance gains from insider trading are robust once trading frictions are taken into account (H_5).

We begin by constructing a large cross-sectional dataset covering 6,870 publicly traded U.S. firms used to predict monthly returns. Our Baseline predictor set incorporates 94 firm characteristics, eight macroeconomic variables, and 70 industry classifications. Following the interaction scheme proposed by Gu, Kelly, and Xiu (2020), this yields a high-dimensional Baseline predictor set of 916 predictors. While Gu, Kelly, and Xiu (2020) studied the period from 1957 to 2016, our analysis is restricted to 2006–2021 due to the availability of digital insider trading records.

We construct 23 insider trading signals empirically hypothesised to predict returns. Using the Double Machine Learning (DML) framework of Chernozhukov et al. (2018), we isolate the unique informational content of these insider signals by controlling for the Baseline predictors. This screening procedure reduces the insider signal set to 14 signals, which expand to 30 variables once interaction terms are included, yielding a final merged dataset with 946 predictors. Our model suite ranges from simple OLS and penalised linear models (Elastic Net) to dimension reduction methods (PCR and PLS) and more complex nonlinear models, including Boosted Regression Trees, Random Forests, and Neural Networks.

Consistent with the findings of Gu, Kelly, and Xiu (2020), we find that neural networks deliver the strongest Baseline return predictions, with NN4 achieving the highest out-of-sample R^2_{OOS} of 0.37%. While the inclusion of insider signals provides statistically significant

improvements in NN2 and NN3 for the Outsider perspective, these gains are economically negligible, leading to the rejection of H_{1a} and H_{2a}.

When translating these predictions into economic performance, the inclusion of insider signals fails to systematically improve the Sharpe ratio. Consequently, we reject H_{1b} and H_{2b}. Moreover, we find that Corporate Insider has no informational advantage over Outsider in terms of the Sharpe ratio, leading us to reject H_3. Controlling for firm size, we find Sharpe ratio gains only for OLS across all size groups, with the gains being most pronounced among large-cap firms. Consequently, we reject H_4. Finally, restricting to the most liquid firms, as a proxy for low market frictions, reveals that NN4 achieves a significant Sharpe ratio gain of 0.22 under the Outsider set relative to the Baseline. This gain remains robust to transaction costs, suggesting the acceptance of H_5.

In summary, this paper contributes to the literature by confirming the results of Gu, Kelly, and Xiu (2020) and demonstrating that while insider trading signals appear informative in low-dimensional linear models, their incremental value vanishes within high-dimensional machine learning models. Our results align with the 'factor zoo' view that many documented anomalies are redundant once evaluated in a high-dimensional setting (Feng, Giglio, and Xiu, 2020). For investors using modern high-dimensional machine learning models, insider trading signals offer little incremental value beyond firm characteristics and macro variables, and are therefore unlikely to systematically improve investment strategies.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature. Section 3 describes the data processing, construction, and merging steps. Section 4 introduces the model suite and performance metrics. Section 5 presents the empirical results. Section 6 examines whether these results are robust to firm size and liquidity. Section 7 provides a discussion of our results, and Section 8 concludes.

2. Literature Review

In this section, we present the literature for our paper by describing the evolution of return prediction models from simple linear models to machine learning models. We then review empirical evidence on how insider trading predicts stock returns, thereby conceptualising why this is a compelling field of study. Finally, we conclude our literature review by motivating the decision to combine machine learning with insider trading.

2.1. From Linear to Machine Learning Prediction Models

Predicting stock returns is the cornerstone of empirical asset pricing. For decades, linear regression has been the workhorse for predicting returns due to its simplicity, transparency, and low computational cost. The foundation was built by Markowitz (1952), who introduced Modern Portfolio Theory (MPT) to model the risk-return trade-off and demonstrate how diversification reduces risk. Later, MPT led to the Capital Asset Pricing Model (CAPM) (Sharpe, 1964; Lintner, 1965; Mossin, 1966), which incorporated a risk-free asset and linked expected returns to systematic risk via the market beta, β. The CAPM proved to be useful,

but empirical tests showed that β alone could not explain returns (Black et al., 1972). This led to the development of multivariable models, such as the Fama–French three-factor model (Fama and French, 1992), which extended the framework to include firm size and valuation. Later, Fama and French (2015) further extended this to a five-factor model including profitability and investment.

The spread of multivariable models initially gave optimism that returns were predictable (Lettau and Ludvigson, 2001; Campbell, 2000). However, the robustness of these results was heavily debated. For example, Welch and Goyal (2008) argues that most factors fail to provide out-of-sample prediction, while Campbell and Thompson (2008) demonstrates that imposing weak restrictions can yield modest but economically meaningful predictive gains.

This ongoing effort to identify predictive variables eventually led to a "factor zoo" (Cochrane, 2011; Harvey et al., 2016; Feng et al., 2020), where hundreds of variables appeared significant in-sample but failed to replicate out-of-sample. Hou et al. (2020) shows that 64% of the 447 market anomalies fail in replication after rigorous testing. While linear models provided early insights, they are structurally rigid. They struggle with high-dimensional data, regime shifts, multicollinearity, and the complex, nonlinear interactions inherent in modern markets (Kelly et al., 2023).

These limitations motivate the use of machine learning, which is explicitly designed to handle large predictor sets, capture nonlinearities, and prioritise out-of-sample forecasting performance. Given that the definition of machine learning varies by context, we follow Gu et al. (2020) in describing it as

> *"We use the term to describe (a) a diverse collection of high-dimensional models for statistical prediction, combined with (b) so-called "regularization" methods for model selection and mitigation of overfit, and (c) efficient algorithms for searching among a vast number of potential model specifications."*

By employing flexible functional forms to approximate unknown data-generating processes, machine learning aligns more closely with the empirical reality of financial markets. Furthermore, by using regularisation to manage hundreds of correlated predictors, these techniques guard against overfitting while capturing the interactions that linear models miss.

While machine learning has recently gained relevance in predicting return, the potential of nonlinear methods was recognised early in the literature. White et al. (1988) pioneered the use of neural networks for stock returns, but early applications were restricted by data limitations and overfitting. Later, Bansal et al. (1993) and Dittmar (2002) confirmed that nonlinear specifications capture return variations that are invisible to linear models. Advances in computing and data culminated in the landmark study by Gu et al. (2020), which systematically demonstrated that machine learning significantly outperforms linear models in a high-dimensional setting in the U.S. from 1957 to 2016. This outperformance has been shown to be globally robust after transaction costs, documented in Europe (Drobetz and Otto, 2021), China (Leippold et al., 2022), and emerging markets (Hanauer and Kalsbach, 2023). Furthermore, the efficacy of machine learning extends beyond stocks

to bonds (Bianchi et al., 2021) and cryptocurrencies (Akyildirim et al., 2021), establishing it as a superior predictive framework across asset classes.

Despite this methodological progress, return prediction remains dominated by financial and accounting characteristics. This leaves a gap regarding behavioural predictors, specifically insider trading, which this research aims to address.

2.2. Insider Trading and Return Prediction

Investors often rely on a broad set of information to distinguish firm quality. Insider trading serves as a credible and costly signal in this regard, as insiders possess private information and commit personal wealth to their trades. The literature agrees that *Corporate Insiders* predict future returns. At the same time, it is less clear whether *Outsiders* can profitably mimic these trades after accounting for disclosure lags and transaction costs.

Empirical evidence has demonstrated that Corporate Insiders can predict returns. For example, Seyhun (1986, 1992); Lakonishok and Lee (2001) find that Corporate Insiders, particularly purchases in smaller firms, predict returns. Similarly, Jeng et al. (2003) shows that Corporate Insiders purchases yield abnormal returns exceeding 6% annually, though they find that sales are less predictive. Subsequent studies confirm that insider trades convey private information not yet incorporated into market prices (e.g., Meulbroek, 1992; Ke et al., 2003). On the contrary, international evidence shows no return predictability in Norway (Eckbo and Smith, 1998) and little to none in most European markets (Gębka et al., 2017), suggesting that the anomaly is largely a U.S. phenomenon.

The predictive power of insider trading is largely attributed to information asymmetry, in which insiders trade on their private information. For example, insider trading predicts higher returns in environments with low analyst coverage and non-transparent financial statements (Frankel and Li, 2004; Wu, 2019). Higher returns are also observed in firms with high R&D intensity because the productivity or value of R&D is not disclosed and materialises alongside long-horizon uncertainty (Aboody and Lev, 2000). Insiders know the value and timeline of R&D, creating information asymmetry between Corporate Insiders and Outsiders. Moreover, Piotroski and Roulstone (2005) demonstrates that insiders are knowledgeable, exploiting mispricing through both contrarian beliefs and their insight into future cash flows. Finally, the timing of insider trading further reveals private information as trades often lead up to price-relevant events such as dividend announcements (John and Lang, 1991), stock repurchases (Lee et al., 1992), M&A bids (Seyhun, 1990), bankruptcies (Seyhun and Bradley, 1997), and earnings declines (Ke et al., 2003; Karpoff and Lee, 1991).

The literature suggests that Corporate Insiders can predict returns, but remains divided on whether Outsiders can effectively mimic insider activity. Seyhun (1986, 1988); Lakonishok and Lee (2001) argue that the return predictability of insider trading is eroded by trading costs for the Outsider. On the other hand, Bettis et al. (1997) find that mimicking Corporate Insiders can be profitable net of costs. But Seyhun (2000) offers a more nuanced perspective, suggesting that Outsiders should not view insider trading in isolation, but rather as one component within a broader investment process. These conflicting results highlight

the limitations of linear models and motivate our use of a flexible and high-dimensional framework to reassess the predictive value of insider trading.

Harnessing the predictive power of insider activity requires separating signal from noise, as insider trading is notoriously noisy (Cohen et al., 2012). The foundational literature is quite dated, but recent research confirms that insider predictability persists but has become more nuanced. Akbas et al. (2020) shows that "unexpected" trades deviating from historical benchmarks remain predictive, while Cziraki and Gider (2021) finds that although insiders possess superior information, regulatory scrutiny limits their absolute dollar profits. This implies that the predictive power of insider trading lies in the timing and nature of the signals.

One method for isolating these signals is to classify "opportunistic" trades. Cohen et al. (2012) identifies opportunistic trades as deviations from routine patterns and shows that they accurately forecast future news, such as analyst recommendations and earnings announcements. Half of this predictive gain stems from opportunistic sales, challenging the traditional view that sales are liquidity-driven (Lakonishok and Lee, 2001). A different way to assess signal strength is through trade clustering, as Alldredge and Blank (2019) find that clustered purchases lead to monthly abnormal returns exceeding 2%. The absence of insider trading also conveys information. Hong and Li (2019) demonstrate that routine insiders strategically remain silent when they possess private information, and that such silence following a routine sell or buy predicts annual abnormal returns of 6% to 10%. Finally, an informational hierarchy has signal quality. Executives and directors are generally equally informed (Ravina and Sapienza, 2010), while Fidrmuc et al. (2008) suggests CEOs may provide less signal content than other officers due to intense regulatory scrutiny.

Because this evidence rests primarily on traditional linear models, it remains an open question whether the predictive power of insider trading can be more effectively harnessed within a high-dimensional machine learning framework.

2.3. Marrying Machine Learning with Insider Trading

While many studies employ machine learning to detect illegal insider trading or predict insider trading, few focus on return prediction. To our knowledge Safer (2002) provided the earliest example, and his core premise is that standard linear specifications do not easily capture the information content of insider trades. To address this, he provided the first application of neural networks in this field. He used a single-hidden-layer neural network with 13 insider trading variables. Analysing 343 U.S. stocks from 1993 to 1997, they identified significant abnormal returns in small- and mid-cap firms, particularly within the electronics and business services sectors.

Recently, Chakravorty and Elsayed (2025) demonstrated that combining insider signals with machine learning can enhance price predictability. However, their one-stock (Tesla) and short-horizon focus leaves open the question of whether these gains persist across a broad cross-section of firms.

Thus, we identify a gap in the current research landscape between two established

frontiers. On one hand, the machine learning frontier (e.g., Gu et al., 2020; Drobetz and Otto, 2021; Leippold et al., 2022; Hanauer and Kalsbach, 2023) has demonstrated that high-dimensional nonlinear models are superior predictive tools, yet these studies largely restrict their predictor space to standard firm characteristics. On the other hand, the traditional asset pricing frontier has long identified insider trading as a predictor of returns (e.g., Lakonishok and Lee, 2001; Piotroski and Roulstone, 2005; Cohen et al., 2012), but these findings remain primarily constrained by the rigidity of low-dimensional linear models. We exploit this gap by integrating insider trading into a high-dimensional nonlinear framework to process these behavioural inputs alongside traditional characteristics.

3. Data

This section describes the construction of the dataset used in the thesis, which combines the *Baseline* predictor set of Gu et al. (2020) with our insider trading dataset. It outlines the data sources, cleaning procedures, and variable construction for both the replication sample and the insider trading variables. It concludes with a description of the final merged dataset.

3.1. Baseline Replication Dataset

We construct our Baseline predictor set following Gu et al. (2020) using three data sources. First, we obtain monthly stock-level data from the Center for Research in Security Prices (CRSP) covering all listed and delisted stocks on the NYSE, AMEX, and NASDAQ from 1957 through 2023. We access this data via an SQLite database provided by Stefan Voigt, which includes stock identifiers, returns, and market capitalisation. Second, we extract eight macroeconomic predictors from the same database, constructed following the definitions of Welch and Goyal (2008). Third, we obtain 94 stock characteristics and 74 standard industry classifications (SIC) from the website of Dacheng Xiu (nd), constructed by Gu et al. (2020) and based on the foundational work of Green et al. (2013). The dataset covers the period from 1957 through 2021 and is fully updated through 2016. For firms entering the sample after 2016, the only missing information is the SIC.

To avoid look-ahead bias and ensure that information is surely available to investors, we lag all predictors. The eight macroeconomic variables are lagged by one month to account for publication delays. The 94 firm characteristics are lagged by one, four, and six months for monthly, quarterly, and annual variables, respectively, to account for reporting delays [1].

To limit the influence of outliers and ensure numerical stability for our machine learning models, we rank the 94 firm characteristics stock-by-stock each month. Missing values are excluded to avoid distorting the ranking, and ties are assigned the same rank for fairness.

[1] The firm characteristics obtained from Xiu's website (Dacheng Xiu, nd) are pre-lagged according to their specifications.

The ranks are subsequently scaled to the interval $c_{ij,t} \in [-1, 1]$ and defined as

$$c_{ij,t} = \begin{cases} 2\left(\dfrac{\text{rank}(\bar{c}_{ij,t}) - 1}{N - 1} - \dfrac{1}{2}\right), & \text{if } N > 1, \\ \text{undefined}, & \text{if } N \leq 1. \end{cases} \tag{3.1}$$

Where $rank(\bar{c}_{ij,t})$ is the rank of stock i for predictor j at month t, while N is the total number of firms in month t. Missing values are handled using a two-step procedure described by Stefan Voigt (ndb). First, missing values are replaced with the cross-sectional median. Second, if the median is undefined, missing values are set to zero.

Interaction terms between the 94 firm characteristics and the eight macroeconomic predictors are constructed to capture how firm characteristics depend on the level of macroeconomic conditions. After including interaction terms, the 94 non-interacted firm characteristics are kept, while the eight macroeconomic predictors are dropped, as their information is now embedded within the interactions. In total, the replicated dataset of Gu et al. (2020) consists of $94 \times (8 + 1) + 74 = 920$ features, as given by the Kronecker product:

$$z_{i,t} = (1, x_t)' \otimes c_{i,t}, \tag{3.2}$$

where $(1, x_t)'$ is the macroeconomic information vector and $c_{i,t}$ is the firm-specific characteristic vector.

With the Baseline predictor set in place, we now proceed to construct the insider trading dataset.

3.2. Insider Trading Dataset

Insider trading is defined as transactions made by executives, directors, and large shareholders who own more than 10% of a company's shares. Insiders must report their trades by filing Forms 3, 4, and 5 at SEC, within two business days. We obtain insider trading data from 2006 through 2021 from SEC. While insider transaction records exist for earlier years, consistently digitised coverage on the SEC's EDGAR system is available only from 2006 onward. In line with the literature (e.g., Seyhun, 1988; Lakonishok and Lee, 2001), we only consider open-market purchases and sales. These insider transactions reflect genuine trading decisions and are therefore most relevant for predicting returns. Other filings primarily capture administrative actions, derivative exercises, or compensation-related transactions.

To map insider trades to individual stocks using their separate identifiers, the Central Index Key (CIK) and the Permanent Stock Number (PERMNO), respectively, we retrieve linking tables from the GitHub repository of Ding (nd).

The data cleaning process is extensive, as insider transactions are manually reported and therefore prone to human error.

Table 1: Data Cleaning Process for Insider Transactions

Filter	Removed	Remaining
Reported Insider Transactions	–	3,754,776
(-) Outside Analysis Window	12,024	3,742,752
(-) Transaction Not on Market-Day	7,054	3,735,698
(-) Unavailable CIK to PERMNO Mapping	1,335,743	2,399,955
(-) Invalid Ticker	182	2,399,773
(-) Total >$10,000 and Non-Missing	768,801	1,630,972
(-) Small Trades (<100 shares)	4,030	1,626,942
(-) Market-Cap Filter (>5%)	11,771	1,615,171
(-) Exact Duplicates	28,964	1,586,207
(-) Near Duplicates (Name/ID only)	274,199	1,312,008
Total Insider Transactions	**2,442,768**	**1,312,008**

Notes: The table summarises the cleaning steps applied to insider filings to separate noisy trades from informative signals.

Table 1 provides an overview of the data cleaning process. The raw dataset contains 3,754,776 insider transactions. We first remove 12,024 trades whose transaction or filing month falls outside our analysis window, 2006 through 2021, and then remove 7,054 transactions recorded on non-market days. Next, we remove 1,335,743 filings with no valid CIK to PERMNO mapping and 182 observations with invalid ticker codes. We then remove 768,801 trades with missing or sub-$10,000 transaction sizes, as such trades are unlikely to represent information-driven insider transactions. Following Lakonishok and Lee (2001), we remove 4,030 trades involving fewer than 100 shares and 11,463 trades exceeding 5% of the firm's market capitalisation. To prevent double-counting, we remove 28,964 exact duplicates and 274,199 near duplicates that differ solely in the reported insider name/ID. After all filtering steps, the dataset is reduced to 1,312,008 insider transactions, with 2,442,768 observations removed. As a result of the cleaning process, some potentially informative trades may have been excluded.

Officer titles in SEC filings are often inconsistent, so we standardise and group them into seven categories. Each insider is assigned to exactly one category using a fixed hierarchy: CEO, CFO, COO, Other Officer, Vice President, Director, and Ten Percent Owner. This hierarchy hides cases where insiders hold multiple roles (e.g., a CEO who is also a Ten Percent Owner). Still, it ensures mutually exclusive categories, which is crucial for a clean empirical analysis.

This insider trading dataset enables the construction of the insider trading signals introduced in the next section.

3.3. Construction of Insider Trading Signals

In this section, we construct insider trading signals proposed in the literature to predict stock returns. Given the inherently noisy nature of insider trades, careful signal design

is essential to avoid diluting predictive content. We aim to isolate insider activity that is more likely to reflect informational motives rather than routine or mechanical trading.

Net Purchase Ratio (NPR) introduced by Lakonishok and Lee (2001), measures the extent to which insiders are net buyers or sellers of their firm's stock over the past six months. Lakonishok and Lee (2001) motivate the use of a six-month window by noting that insider trades are often infrequent and many firms have months with no reported activity, making a longer aggregation period necessary to obtain a reliable signal. Moreover, insider trading has been shown to predict returns persistently for up to twelve months (Seyhun, 2000). The six-month rolling window, therefore, provides a way to capture this persistence while maintaining a monthly prediction framework. The original measure is based on dollar amounts, whereas we extend it by also constructing a count-based version. Formally, the six-month rolling NPR for stock i traded in month t is defined as

$$\mathrm{NPR}_{i,t}^{\mathrm{Volume}} = \frac{\sum_{j=t-5}^{t} \text{Dollar Purchases}_{i,j} - \sum_{j=t-5}^{t} \text{Dollar Sales}_{i,j}}{\sum_{j=t-5}^{t} \text{Dollar Purchases}_{i,j} + \sum_{j=t-5}^{t} \text{Dollar Sales}_{i,j}}. \tag{3.3}$$

Similarly, the count-based NPR is defined as

$$\mathrm{NPR}_{i,t}^{\mathrm{Count}} = \frac{\sum_{j=t-5}^{t} \text{Purchases}_{i,j} - \sum_{j=t-5}^{t} \text{Sales}_{i,j}}{\sum_{j=t-5}^{t} \text{Purchases}_{i,j} + \sum_{j=t-5}^{t} \text{Sales}_{i,j}}. \tag{3.4}$$

By construction, both measures take values in the interval $[-1, 1]$, where -1 indicates that all insider activity in the past six months consisted solely of sales, while 1 means that all activity consisted solely of purchases.

Trade Clustering, introduced by Alldredge and Blank (2019), measures coordinated insider behaviour by identifying instances in which multiple insiders from the same firm trade within a short time window. Such clustering is interpreted as reflecting shared private information rather than individual motives. Alldredge and Blank (2019) show that clustered purchases, but not clustered sales, predict higher abnormal returns. Following their definition, we classify an insider cluster as two or more trades by at least two different insiders occurring within two market days, separately for purchases and sales. We then aggregate these clusters to the firm–month level and construct a measure named Net Clustering, defined as the monthly difference between the counts of purchase and sale clusters.

Routine versus Opportunistic Trades Following Cohen et al. (2012), we classify insider trades as either routine or opportunistic. Routine trades are predictable and typically driven by diversification motives, whereas opportunistic trades are irregular and more likely to reflect value-relevant information. Cohen et al. (2012) show that only opportunistic trades predict abnormal returns, and that both opportunistic buys and sells contain predictive information. Under their classification scheme, an insider becomes eligible for classification once she has traded at least once in each of the three preceding years. An insider is classified as routine if she trades in the same calendar month for at least three consecutive years; all

remaining insiders are classified as opportunistic. Classifications are determined at the start of each calendar year using the prior three years of trading history, and all subsequent trades by a routine insider during that year are labelled routine. Based on these classifications, we construct four firm–month dummy variables indicating the presence of each trade type: Routine Buy, Routine Sell, Opportunistic Buy, and Opportunistic Sell. Multiple dummies may equal one in a given month, as insiders may both buy and sell within the same month, or numerous insiders may trade in the same firm-month.

Insider Silence, introduced by Hong and Li (2019), captures the idea that the absence of an expected trade can in itself be highly informative. Hong and Li (2019) show that when routine sellers (buyers) unexpectedly refrain from trading, future returns tend to be positive (negative). Following this framework, an insider who sells in the same calendar month for three consecutive years is classified as Sell–Sell–Sell SSS. If that insider does not trade in the third year, she is classified as Sell–Sell–No Trade (SSN). The same logic applies to purchases, producing the Purchase-Purchase-Purchase (PPP) and Purchase-Purchase-No Trade (PPN) categories. We aggregate these insider-level silence signals to the firm–month level by defining SSN and PPN dummies that equal one when at least one insider at the firm meets the corresponding condition in that month, and zero otherwise.

Insider Roles have been shown to possess differentiated informational value (Ravina and Sapienza, 2010; Fidrmuc et al., 2008). To capture these differences, we classify insiders into seven groups as defined earlier: CEO, CFO, COO, Other Officer, Vice President, Director, and Ten Percent Owner. For each role and trade type, we construct firm–month dummies, such as CEO-Purchase and CEO-Sell, which equal one if at least one insider in that role executes a purchase or sale in that month, and zero otherwise.

This section allows us to derive 23 insider trading signals, summarised in Table 2. The following section describes how the Baseline predictor set is merged with the 23 insider trading signals.

3.4. Merging Baseline and Insider Trading

To ensure comparability when merging the Baseline predictor set of Gu et al. (2020) with the insider trading signals, we restrict the sample period from 2006 through 2021, the longest span for which both stock characteristics and insider trading data are jointly available. We further exclude firms without a valid link between CIK and PERMNO identifiers and impose a strict one-to-one CIK to PERMNO mapping, thereby removing ambiguous cases such as firms with multiple share classes.

Table 2: Overview of Insider Trading Signals

No.	Signal	Short description	Paper
1	npr_volume	Net purchase ratio by dollar volume rolling 6 months	Lakonishok and Lee (2001)
2	npr_count	Net purchase ratio by trade count rolling 6 months	Lakonishok and Lee (2001)
3	opp_buy	Opportunistic insider buy	Cohen et al. (2012)
4	opp_sell	Opportunistic insider sell	Cohen et al. (2012)
5	rtn_buy	Routine insider buy	Cohen et al. (2012)
6	rtn_sell	Routine insider sell	Cohen et al. (2012)
7	net_cluster	Net clusters	Alldredge and Blank (2019)
8	ppn	Silence after three routine buys	Hong and Li (2019)
9	ssn	Silence after three routine	Hong and Li (2019)
10	purchase_ceo	CEO purchase	Own construction
11	purchase_cfo	CFO purchase dummy	Own construction
12	purchase_coo	COO purchase dummy	Own construction
13	purchase_director	Director purchase dummy	Own construction
14	purchase_vp	Vice president purchase dummy	Own construction
15	purchase_10pct	10% owner purchase dummy	Own construction
16	purchase_other	Other officer purchase dummy	Own construction
17	sell_ceo	CEO sale dummy	Own construction
18	sell_cfo	CFO sale dummy	Own construction
19	sell_coo	COO sale dummy	Own construction
20	sell_director	Director sale dummy	Own construction
21	sell_vp	Vice president sale dummy	Own construction
22	sell_10pct	10% owner sale dummy	Own construction
23	sell_other	Other officer sale dummy	Own construction

Notes: Signals are defined at the firm–month level. The variables npr_volume, npr_count, and net_cluster are scaled cross-sectionally each month to the range $[1, 1]$, while the remaining variables are left unscaled because they are dummy indicators.

After these restrictions, the Baseline dataset contains 916 variables. Five SIC industry indicators are excluded due to restricting the sample, while one SIC is deliberately added to identify stocks that enter the dataset after 2016.

We then merge the Baseline dataset with the insider-trading signals. All insider signals are lagged by one month to avoid look-ahead bias and merged on CIK and month. Missing insider observations are set to zero for dummy variables, while continuous signals are cross-sectionally ranked and scaled to the interval $[-1, 1]$ as defined in Equation (3.1).

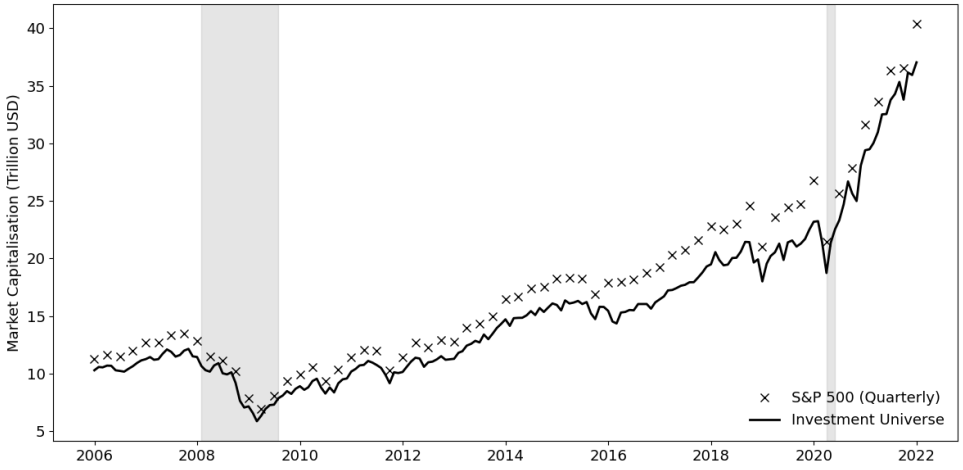
The merged dataset adds 23 insider signals, comprising three continuous signals and 20 dummy variables. The continuous insider signals are further interacted with the eight macroeconomic predictors, yielding 24 additional interaction terms. As a result, the feature set expands from 916 to 963 variables.

In total, the final insider trading augmented dataset comprises one target variable (excess returns), seven variables used for identification and conditioning, 963 predictive features (see Appendix A for an overview), and 579,056 firm–month observations spanning from 2006 through 2021.

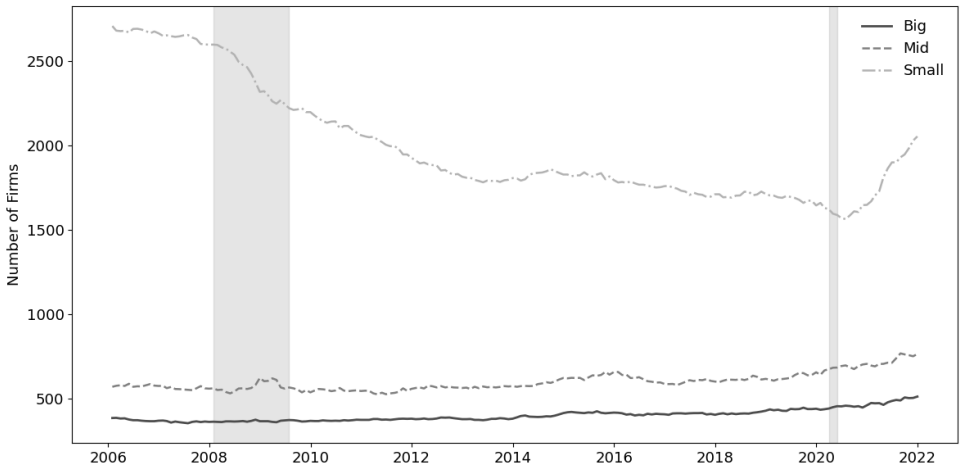
3.5. Data Visualisation and Descriptives

This section presents visualisations of the investment universe and descriptive statistics of insider trading activity. Figure 1 illustrates the scale and representativeness of the stock universe, its cross-sectional composition by firm size, and variation in insider trading activity across firm size groups. These figures provide context for the empirical analysis and help assess the realism of the investment setting.

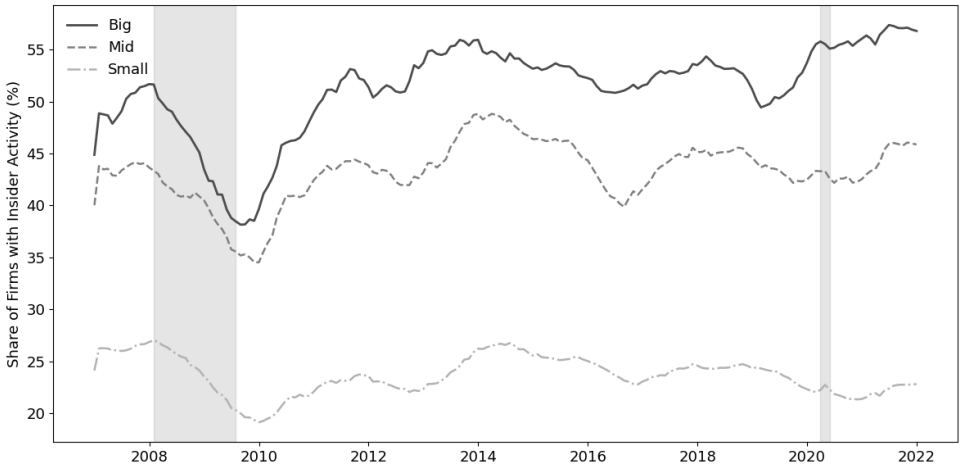
Figure 1: Investment Universe Overview (2006–2022)



(a) Market Capitalisation of the Investment Universe and the S&P500



(b) Firm Size Composition of the Investment Universe



(c) Share of Insider Trading Activity by Firm Size Composition (12-Month Moving Average)

Panel (a) plots the aggregate monthly market capitalisation of our investment universe and the S&P 500 from 2006 through 2021. The market capitalisation of our investment universe closely tracks that of the S&P 500, though at a slightly lower level, indicating that our universe is broadly representative in terms of both market size and market movements. The remaining gap in market capitalisation likely reflects our restriction to common stocks, excluding firms with multiple share classes (e.g., A- and B-shares). Overall, we find our stock universe to be representative of the U.S. equity market.

Panel (b) shows the number of stocks in the investment universe across small-, mid-, and large-cap firms using NYSE breakpoints from 2006 through 2021. The distribution is highly skewed, with relatively few large-cap firms, only marginally more mid-cap firms, and a disproportionately large number of small-cap firms. This essentially reflects our use of stocks from all available exchanges. In particular, NASDAQ and AMEX have many small-cap stocks.

Panel (c) reports insider trading activity across these size groups, measured as the fraction of firms in each category that experience at least one insider transaction in a given month from 2007 through 2021. Large-cap firms exhibit the highest insider-trading share, with a mean of 52% of firms having at least one insider transaction per month. Mid-cap firms follow closely, with an average of 43% of firms experiencing insider trading activity per month. By contrast, small-cap firms display markedly lower insider trading activity, with only about 24% of firms experiencing insider transactions per month. This pattern is consistent with higher liquidity in larger firms, which facilitates insider transactions and reduces their price impact.

Table 3: Descriptive Statistics for Insider Trading Signals

Panel (a): Continuous Insider Variables

Variable	Mean	Std. dev.	$P(x > 0)$	$P(x < 0)$
npr_volume	-0.18	0.49	0.06	0.25
npr_count	-0.17	0.48	0.06	0.24
net_cluster	-0.07	0.37	0.01	0.06

Panel B: Equal-weighted

Variable	Mean	Variable	Mean
opp_buy	0.055	purchase_vp	0.006
opp_sell	0.18	purchase_10pct	0.013
rtn_buy	0.008	purchase_other	0.054
rtn_sell	0.056	sell_ceo	0.068
ppn	0.006	sell_cfo	0.056
ssn	0.045	sell_coo	0.028
purchase_ceo	0.018	sell_director	0.009
purchase_cfo	0.008	sell_vp	0.089
purchase_coo	0.003	sell_10pct	0.017
purchase_director	0.002	sell_other	0.155

Notes: Panel (a) reports continuous firm–month insider trading signals, while Panel (b) reports dummy indicators for insider transactions. Both panels are constructed after merging with the *Baseline* dataset and imputing missing values with zero. The continuous variables are reported before cross-sectional scaling.

Table 3 presents descriptive statistics after merging the insider trading variables with the Baseline dataset and imputing missing values with zero, but before cross-sectional scaling of the continuous variables. Panel (a) shows that insider activity is structurally skewed

toward selling, as both NPR measures have negative means of around -0.18, confirming that firm–month insider trading exhibits net insider sales on average. This pattern is consistent with managers' liquidity and diversification motives. The clustering measure is also negative (-0.07), but substantially less so than the NPR measures, indicating a more balanced mix of buy and sell clusters on average. Panel (b) reports the means of the dummy indicators. It reinforces this asymmetry between purchases and sales: sell indicators exhibit means that are roughly an order of magnitude larger across all insider groups.

4. Methodology

In this section, we present our methodological approach, following the framework of Gu et al. (2020). First, we describe the overall research design, optimisation, and sample splitting. We then present the machine learning models we use, ranging from linear to nonlinear. Next, we detail our systematic procedure for selecting insider trading variables. To evaluate predictive power, we report statistical performance metrics and variable importance measures that identify the drivers of predictions. Finally, we translate forecasting performance into economic terms via portfolio construction and risk-adjusted performance.

4.1. Research Design

This paper investigates whether insider trading signals add incremental predictive power for stock returns by leveraging machine learning in a high-dimensional dataset.

Gu et al. (2020) describe realised excess returns within an additive prediction error framework, defined as

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1}, \tag{4.1}$$

where stocks are indexed by $i = 1, \ldots, N_t$ and months by $t = 1, \ldots, T$. Where the potentially nonlinear function is given as

$$E_t(r_{i,t+1}) = g^*(z_{i,t}). \tag{4.2}$$

Equation (4.1) states that the expected return for time $t+1$, computed at time t, can be decomposed into a predictable component given the information set available at time t, $E_t(r_{i,t+1})$, and an unpredictable noise component, $\epsilon_{i,t+1}$. Equation (4.2) defines expected returns as a potentially nonlinear function of the P-dimensional vector $z_{i,t}$, which includes predictor variables such as firm characteristics, macroeconomic variables, market information, and insider trading activity.

Our objective is to find a representation of $E_t(r_{i,t+1})$ that best explains realised returns $r_{i,t+1}$ in out-of-sample tests. More specifically, we aim to predict excess returns as a flexible function of the predictor variables, $g^*(z_{i,t})$, without restricting it to a specific parametric form (e.g., linear, quadratic, or logarithmic), by minimising the mean squared error (MSE).

Significantly, the function $g^*(\cdot)$ does not depend on i or t, meaning the model maintains the same functional form across both stocks and time. This allows the model to learn patterns from the entire dataset, rather than treating each stock or period in isolation. This

contrasts with traditional asset-pricing models, which typically estimate a cross-sectional regression at each period t (e.g., the Fama-French model). Moreover, $g^*(\cdot)$ depends solely on $z_{i,t}$, meaning predictions use only the current characteristics of stock i at time t, excluding both past values and information from other firms.

Our analysis relies on two nested datasets. The first is the high-dimensional dataset from Gu et al. (2020), which contains 916 firm-level and firm-macro interacted predictors. We refer to this as the *Baseline* set, and it serves as our control when evaluating the contribution of insider trading information. The second dataset extends the Baseline to 963 predictors by adding 23 insider trading variables and 24 interaction terms, enabling us to test whether insider activity provides incremental predictive power. The insider trading dataset is viewed through two distinct lenses. The *Outsider* information set defines insider activity based on the filing month, capturing information publicly available once trades are disclosed. The *Corporate Insider* set, in contrast, uses the transaction month, reflecting the private information insiders possess at the time of trade execution. Approximately 8% of all trades appear only in the Corporate Insider set because they are not publicly disclosed by month-end, providing information that Outsiders can never exploit. Therefore, comparing the Corporate Insider and Outsider sets allows us to assess whether private information offers predictive value beyond publicly disclosed insider trades.

4.2. Optimisation and Sample Splitting

The core of financial machine learning is forecasting future returns $r_{i,t+1}$ from a high-dimensional predictor set $z_{i,t}$ via a potentially nonlinear function $g(\cdot)$, as defined in Equation (4.2). Given the low signal-to-noise ratio typical of financial data, flexible models are prone to overfitting, capturing noise rather than signal. We mitigate this through regularisation and sample splitting. The data is split into training, validation, and test sets. The training data is used to estimate model parameters, the validation data to tune hyperparameters that control model complexity, and the test data to measure true out-of-sample performance.

In the validation set, optimal hyperparameters are selected by minimising the mean squared error (MSE) using a grid search over values based on Gu et al. (2020). Table 15 in Appendix B provides an overview of the hyperparameter grids used for all models. Once the optimal hyperparameters are identified, we deviate from Gu et al. (2020) by re-training the model on the combined training and validation set. Given the limited sample size, this approach allows us to make use of all available information. Furthermore, to reduce computational load, we tune hyperparameters only on the Baseline specification and fix them when estimating insider trading augmented models. Finally, the retrained model is evaluated on the test set. Since the model has never seen this data, it provides a truly out-of-sample evaluation of predictive performance.

Splitting the sample into training, validation, and test sets is essential for ensuring true out-of-sample performance. In our empirical analysis, we follow Gu et al. (2020) by adopting a hybrid design, in which the training sample expands recursively, and the model is re-estimated annually. For each estimation, we retain a fixed-size rolling window for

validation and generate forecasts for the subsequent year. This approach preserves the data's temporal structure and efficiently incorporates new information while remaining computationally feasible.

Figure 2: Sample Splitting

Notes: Sample splitting strategy for 2006–2021. The figure outlines an expanding training window (initial length of five years and expands by one year in each window) scheme with a rolling validation window (fixed at three years) and a rolling test window (fixed at one year). This yields 8 out-of-sample forecast years (2014–2021). Own creation based on Gu et al. (2020).

Figure 2 illustrates our sample splitting scheme for the 2006–2021 period. Our sample is only about one-quarter as long as the 1957–2016 period used by Gu et al. (2020). Each dot represents a calendar year, and each row corresponds to one out-of-sample forecast window. We employ an expanding window approach for training. The initial training window spans five years and increases by one year in each subsequent iteration. The validation window size is kept fixed at three years, and the test set is fixed at one year. Both the validation and test sets are rolled forward by one year. For example, the first iteration (first row in the figure) trains on 2006–2010, validates on 2011–2013, and produces an out-of-sample forecast for 2014. The final iteration (last row in the figure) trains on 2006–2017, validates on 2018–2020, and forecasts for 2021. This process yields 8 years of out-of-sample forecasts from 2014 through 2021.

4.3. Models

Following the framework of Gu et al. (2020), we outline the machine learning models used to estimate the function $g^*(\cdot)$ as defined in Equation (4.2). While all models share the goal of maximising predictive accuracy, they differ in the structural constraints they impose. Specifically, the simple linear model has no regularisation, whereas all other models employ different regularisation tools to manage complexity. We begin with the simple linear benchmark and progressively increase model complexity, ultimately reaching the complex nonlinear neural network.

4.3.1. Simple Linear

The simple linear model is the standard approach in empirical asset pricing, due to its transparency and interpretability. While we expect the simple linear model to underperform, it serves as a benchmark to measure the predictive gains offered by more complex models.

We define the simple linear model as

$$g\left(z_{i,t};\theta\right) = z_{i,t}'\theta, \tag{4.3}$$

where the function $g^{\star}(\cdot)$ predicts returns as a linear combination of the P-dimensional raw predictor variables $z_{i,t}$ and the parameter vector θ. This model relies on a simple regression specification that excludes nonlinear effects and interactions among predictors.

Model parameters are estimated via OLS, which minimises the l_2 loss function. We define the objective function as

$$\mathcal{L}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(r_{i,t+1} - g\left(z_{i,t};\theta\right)\right)^2. \tag{4.4}$$

The pooled OLS estimator is obtained by minimising the objective function $\mathcal{L}(\theta)$, which represents the sum of squared residuals. In practice, the OLS estimator chooses the parameter vector θ that minimises the squared differences between the observed returns, $r_{i,t+1}$, and the predicted returns of the linear combination of predictors, $g(z_{i,t};\theta)$.

OLS is convenient for several reasons. First, the minimisation problem has a closed-form solution, given by

$$\hat{\theta} = (Z'Z)^{-1}Z'r, \tag{4.5}$$

where Z is the $NT \times P$ design matrix constructed by stacking all predictor vectors $z_{i,t}$, and r is the $NT \times 1$ vector of stacked returns. This eliminates the need for iterative optimisation. Second, estimation is computationally inexpensive since it only involves standard linear algebra operations. Moreover, under the Gauss-Markov assumptions, OLS is the Best Linear Unbiased Estimator (BLUE).

Following Gu et al. (2020), we experimented with the robust Huber loss. However, it consistently worsened out-of-sample relative to plain OLS, possibly because downweighting extreme returns removes informative variation. We therefore report results for standard OLS only. Additionally, not using the Huber loss ensures comparability with models for which the loss cannot be applied, i.e., PLS and PCA (Drobetz and Otto, 2021).

Including all 916 Baseline predictors in an OLS model is likely to create unstable estimates and overfitting. Therefore, we estimate a simpler OLS model using only three predictors: size, book-to-market, and momentum, which we refer to as "OLS-3". The model is implemented using `LinearRegression` from the `scikit-learn` library. Because linear regression has no hyperparameters to tune, we use an initial training window of eight years rather than five.

4.3.2. Penalised Linear

High dimensionality and low signal-to-noise ratios in financial data frequently compromise the stability of OLS, leading to models that capture noise rather than underlying signals. To enhance predictive accuracy, model complexity can be constrained. Penalised linear regression methods achieve this by adding a penalty term to the loss function, thereby enforcing parameter sparsity or shrinkage. This methodology directly addresses the bias-variance trade-off by intentionally reducing in-sample accuracy to yield more stable models with improved out-of-sample performance (James et al., 2023).

The same statistical model as the simple linear model in Equation (4.3) is used, but the loss function in Equation (4.4) is augmented with a penalty term defined as

$$\mathcal{L}(\theta;\cdot) = \underbrace{\mathcal{L}(\theta)}_{\text{Loss Function}} + \underbrace{\phi(\theta;\cdot)}_{\text{Penalty Term}} . \tag{4.6}$$

Two penalization techniques are implemented: ridge regression (shrinkage) and lasso (selection). Ridge regression applies an l_2 penalty that shrinks coefficients θ towards zero without forcing exact zeros defined as

$$\phi(\theta;\lambda) = \lambda \sum_{j=1}^{P} \theta_j^2, \tag{4.7}$$

where λ controls the degree of shrinkage. When $\lambda = 0$, the OLS estimates are recovered; as λ increases, the coefficients are shrunk towards zero. Ridge regression is particularly valuable when $P > T$, a setting in which the OLS estimator is not uniquely defined and exhibits high variance and numerical instability, as is common in high-dimensional financial datasets (James et al., 2023).

Ridge regression retains all predictors, which can complicate interpretation when P is large. This limitation motivates the use of lasso, which employs an l_1 penalty that both shrinks coefficients and sets some exactly to zero, thereby enabling variable selection. Lasso is defined as

$$\phi(\theta;\lambda) = \lambda \sum_{j=1}^{P} |\theta_j|, \tag{4.8}$$

where λ again controls shrinkage intensity.

The elastic net combines both l_1 and l_2 penalties

$$\phi(\theta;\lambda,\rho) = \lambda(1-\rho) \sum_{j=1}^{P} |\theta_j| + \frac{1}{2}\lambda\rho \sum_{j=1}^{P} \theta_j^2, \tag{4.9}$$

where ρ controls the balance between ridge ($\rho = 1$) and lasso ($\rho = 0$). The hyperparameters estimated in the elastic net are ρ and λ controlling model regularisation. Implementation is conducted using `ElasticNet` from the `scikit-learn` library.

4.3.3. Dimension Reduction

While penalisation methods control variance through selection and shrinkage, their effectiveness may be limited by strong correlations present in financial data. When predictors represent the true target plus random noise, averaging these predictors frequently yields more accurate forecasts than selecting a subset using lasso. Dimension reduction addresses this limitation by projecting the high-dimensional predictor space onto a lower-dimensional subspace composed of linear combinations. This process constructs fewer, uncorrelated predictors that filter noise and improve forecast accuracy. This section examines two dimension reduction methods: principal component regression (PCR) and partial least squares (PLS).

PCR employs a two-step approach. The first step involves principal component analysis (PCA), which reduces dimensionality by transforming correlated predictors into a smaller set of linear combinations that preserve the original covariance structure. The first component captures the maximum available variance, and each subsequent component captures the maximum remaining variance while remaining uncorrelated with previous components. This process ensures that all principal components are uncorrelated. In the second step, the constructed principal components serve as predictors in a linear regression model. However, because PCA is unsupervised, it reduces dimensions by considering only the covariation among predictors and does not account for their relationship with future returns. As a result, low-variance variables that may be highly informative for future returns can be omitted (James et al., 2023).

The limitations of PCR motivate the adoption of PLS, a supervised alternative that constructs components based on both predictor covariance and their relationship with future returns. Initially, PLS estimates the univariate coefficient ϕ_j for each predictor j, reflecting its sensitivity to future returns R. These predictors are then aggregated into a single component, weighted in proportion to ϕ_j, thereby prioritising variables with the strongest predictive power. Subsequent components are constructed sequentially using the same supervised approach, ensuring that both the predictors and the returns are uncorrelated with all previously constructed components (James et al., 2023).

The mathematical definitions of both PCR and PLS are expressed in terms of a vectorised linear model

$$R = Z\theta + E, \tag{4.10}$$

where R is the $NT \times 1$ vector of future returns r_{t+1}, Z is the $NT \times P$ matrix of stacked predictors $z_{i,t}$, and E is the $NT \times 1$ vector of residuals $\epsilon_{i,t+1}$. Dimension reduction is achieved by shrinking the dimension of the predictors from P to K via a linear combination. The forecasting model can be described as

$$R = (Z\Omega_K)\,\theta_K + \tilde{E}. \tag{4.11}$$

where Ω_K is a $P \times K$ matrix with columns $w_1, \ldots, w_K$. Each w_j is the linear combination weight used to construct the jth predictive component. Therefore, $Z\Omega_K$ is the dimension-

reduced version of the original predictor set, implying that the predictive coefficient θ_K is a $K \times 1$ vector instead of $P \times 1$.

The objective functions for PCR and PLS differ primarily in their treatment of future returns R. PCR constructs weights Ω_K recursively by maximising predictor variance, thereby disregarding R. The jth linear combination is determined by solving

$$w_j = \arg\max_{w} \ \mathrm{Var}(Zw) \quad \text{s.t.} \quad \begin{cases} w'w = 1, \\ \mathrm{Cov}(Zw, Zw_l) = 0, \quad l = 1, 2, \ldots, j-1. \end{cases} \tag{4.12}$$

This identifies K orthogonal combinations that capture the maximum variation in Z, computed efficiently via singular value decomposition. In contrast, PLS seeks linear combinations that maximise predictive association with R. The jth PLS weights solve

$$w_j = \underset{w}{arg\,max}\ \mathrm{Cov}^2(R, Zw) \quad \text{s.t.} \quad \begin{cases} w'w = 1, \\ \mathrm{Cov}(Zw, Zw_l) = 0, \quad l = 1, 2, \ldots, j-1. \end{cases} \tag{4.13}$$

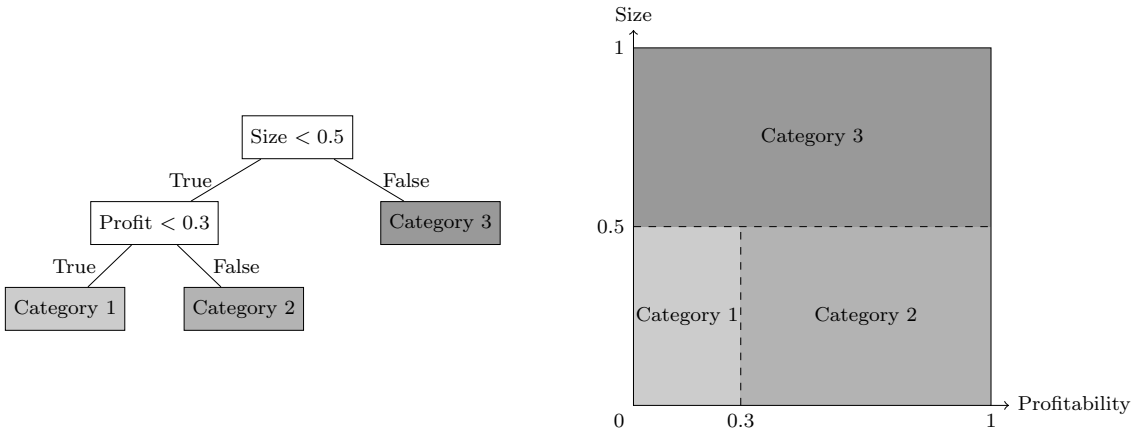
Thus, whereas PCR is unsupervised, PLS explicitly selects weights that maximise the squared covariance between the components and the target returns.

Model complexity for both PCR and PLS is regularised by the number of principal components K. In Python, PCR is constructed using `PCA` and `LinearRegression`, while PLS uses `PLSRegression`, all from the `scikit-learn` library.

4.3.4. Gradient Boosted Regression Tree and Random Forest

While linear models impose a fixed functional form, tree-based methods partition the predictor space to capture complex nonlinear interactions without requiring pre-specified assumptions (James et al., 2023). The tree structure "grows" iteratively, with each step splitting the data along a predictor selected for its capacity to minimise the loss function at that stage. This recursive process continues until a stopping criterion is met or each observation is isolated in a terminal node. The resulting model predicts the unknown estimates $g^*(z_{i,t})$ by computing the average return within each partition.

Figure 3: Regression Tree Example



(a) Tree

(b) Partitions

Notes: The left panel shows the regression tree where the terminal nodes (leaves) represent the final groups. The right panel shows the partition of the predictor space into corresponding regions labelled Category 1, Category 2, and Category 3. Own creation inspired by Gu et al. (2020).

Figure 3 demonstrates the growth of a decision tree using two firm characteristics: Size and Profitability. The initial split is based on size at the 0.5 threshold. Firms with a size greater than 0.5 are assigned to Category 3. Firms with a size less than 0.5 are further divided by profitability at the threshold of 0.3. Firms with profitability less than 0.3 are placed in Category 1, while those with profitability greater than 0.3 are placed in Category 2. This process produces three terminal leaves, and the forecast for each leaf corresponds to the average return of the firms within that group.

Formally, the prediction of tree $\mathcal{T}$ with K leaves (terminal nodes), and depth L is defined as

$$g(z_{i,t};\theta,K,L)=\sum_{k=1}^{K}\theta_k\,\mathbf{1}_{\{z_{i,t}\in C_k(L)\}}, \tag{4.14}$$

where $C_k(L)$ represents one of the K regions of the data. Each region is created through up to L binary splits of the predictors. The constant θ_k associated with region k is simply the average outcome of all observations that fall into that leaf. We focus on recursive binary trees because of their relative simplicity. The example in Figure 3 defines its prediction equation as

$$g(z_{i,t};\theta,3,2)=\theta_1\,\mathbf{1}_{\{\text{size}_{i,t}<0.5\}}\,\mathbf{1}_{\{profit_{i,t}<0.3\}}+\theta_2\,\mathbf{1}_{\{\text{size}_{i,t}<0.5\}}\,\mathbf{1}_{\{profit_{i,t}\geq 0.3\}}+\theta_3\,\mathbf{1}_{\{\text{size}_{i,t}\geq 0.5\}}. \tag{4.15}$$

The goal of the tree is to find regions that minimise forecast error. Since it is impossible to evaluate every possible partition, we use a greedy algorithm. This means the tree acts myopically, considering only the best choice at the current node without looking ahead (James et al., 2023). At every step, the algorithm selects a predictor and a threshold that best separate observations based on their outcomes. The quality of a split is judged by the

squared-error impurity (l_2), which is minimised when observations within a branch have similar outcomes. We define this loss function as

$$H(\theta, C) = \frac{1}{|C|} \sum_{z_{i,t} \in C} (r_{i,t+1} - \theta)^2, \tag{4.16}$$

where $|C|$ refers to the number of observations in a given region C. The best choice of θ for a region is the average of the outcomes within that region. In practice, this means that the algorithm always assigns to each branch the value that minimises the local prediction error (impurity). The tree keeps splitting nodes until it reaches a stopping rule, which can be selected adaptively using a validation sample, such as a maximum number of leaves or a maximum depth.

The flexibility of trees is both their strength and their weakness. Trees allow us to capture complex nonlinear relationships and interactions, but this same flexibility makes them prone to overfitting. Moreover, they are unstable, since small changes in the data can lead to significant changes in the final estimated tree (James et al., 2023). Thus, trees must be heavily regularised. While trees are weak on their own, combining them into a single model can potentially yield powerful results. We consider two methods: Boosting and Random Forests.

Boosting refers to training trees sequentially, with each new tree learning from the mistakes of previous trees to correct weaknesses. The final prediction is a weighted combination of all trees called an ensemble model, which exhibits a greater stability than a single complex tree. The algorithm is referred to as Gradient Boosted Regression Trees (GBRT). The first step involves fitting a shallow tree, e.g., with depth $L = 1$, to ensure an oversimplified tree that serves as a weak predictor with large bias or a poor fit to the training data. Next, another shallow tree (with depth L) is fitted to the residuals from the first tree. Its prediction is added to the ensemble, scaled by a shrinkage factor ν to limit overfitting and slow down the process. Statistical learning approaches that learn slowly tend to perform well (James et al., 2023). At each step b, a new shallow tree is trained on the residuals from the ensemble of $b-1$ trees, and its contribution is weighted by ν. This process continues until B trees are combined, producing an ensemble of shallow trees governed by three tuning parameters: tree depth L, learning rate ν, and the total number of trees B.

Like GBRT, a Random Forest is an ensemble method that combines predictions from multiple trees, but it differs in its construction. In contrast to the sequential, error-correcting nature of boosting, a Random Forest grows trees independently using bagging (bootstrap aggregation). This method utilises deep, high-variance trees that are individually noisy and sensitive to the training sample. Bagging addresses this instability by drawing B bootstrap samples, fitting a deep tree to each in parallel, and averaging the results (Breiman, 2001). However, standard bagging faces a limitation. If one strong predictor dominates the dataset, all trees will likely split on it first. This leads to highly correlated predictions and poor variance reduction, a particular concern with financial data. Random Forests overcome this by decorrelating the trees. At each split, the algorithm is forced to consider only a

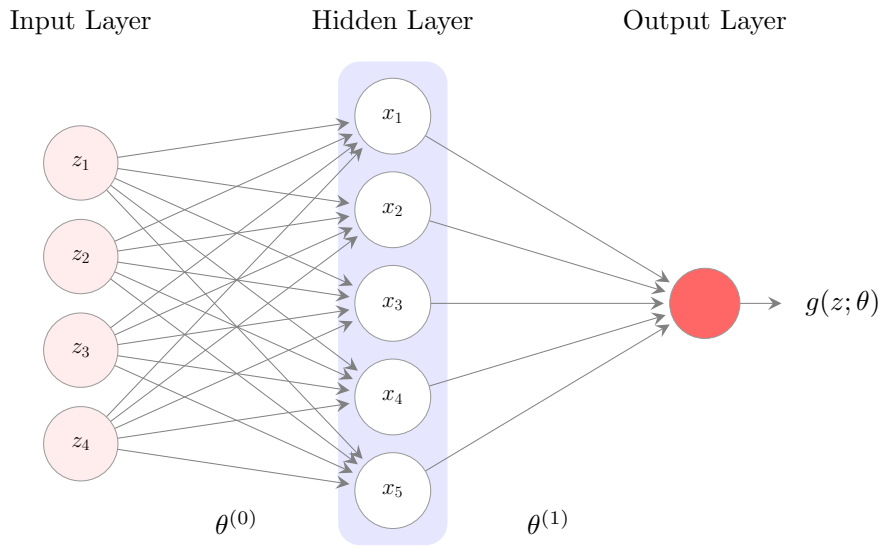
random subset of predictors (James et al., 2023), ensuring the trees capture diverse patterns. The method relies on three key tuning parameters: tree depth L, the number of features considered per split, and the total number of trees B.

In Python, we implement GBRT using `HistGradientBoostingRegressor` and Random Forest using `RandomForestRegressor`, both from the `scikit-learn` library.

4.3.5. Neural Networks

Neural networks are regarded as among the most sophisticated models due to their ability to approximate complex nonlinear functions through multiple layers, a structure known as 'deep learning'. This complexity facilitates their widespread application in empirical asset pricing (Kumbure et al., 2022). However, the complexity of neural networks poses key challenges when applied to financial data. These challenges include a high risk of overfitting, sensitivity to noise, and limited interpretability.

Figure 4: Feed-Forward Neural Network with One Hidden Layer



Notes: The network has inputs $z_1, \ldots, z_4$, a hidden layer with five neurons $x_1, \ldots, x_5$, and an output $g(z;\theta)$. Weights $\theta^{(0)}$ and $\theta^{(1)}$ are estimated during training. Own creation based on Gu et al. (2020).

Figure 4 illustrates a feed-forward neural network with four predictor variables $(z_1, \ldots, z_4)$ in the input layer, and one hidden layer with five neurons $(x_1, \ldots, x_5)$. Each neuron works in two steps. First, for each hidden neuron k, it computes a linear signal as a function of the inputs plus a bias term as

$$v_k(z) = \theta_{k,0}^{(0)} + \sum_{j=1}^{4} z_j \theta_{k,j}^{(0)}. \tag{4.17}$$

Second, this linear signal is transformed by applying a nonlinear activation function $f(\cdot)$ to

produce the hidden activation as

$$x_k^{(1)} = f(v_k(z)) = f\left(\theta_{k,0}^{(0)} + \sum_{j=1}^{4} z_j \theta_{k,j}^{(0)}\right), \tag{4.18}$$

where $\theta_{k,0}^{(0)}$ is the bias parameter (analogous to an intercept), and $\theta_{k,j}^{(0)}$ is the weight that scales the influence of predictor z_j on the signal of neuron k. Finally, the output layer linearly combines the hidden activations to produce the prediction

$$g(z;\theta) = \theta_0^{(1)} + \sum_{k=1}^{5} x_k^{(1)} \theta_k^{(1)}. \tag{4.19}$$

The trainable parameters are the set of weights θ, which is determined by an algorithm of choice such that return predictions $g(z;\theta)$ are approximated with the observed return. In this example, the network yields $31 = (4+1)\cdot 5 + 6$ trainable parameters. The first term, $(4+1)\cdot 5 = 25$, accounts for the connections between the input layer and the hidden layer, as each of the five hidden neurons has four weights (one for each input unit) and one bias. The second term, $6 = (5+1)$, comes from the output layer, which requires one weight for each of the five hidden activations plus a single bias term.

The architectural choices of structuring a neural network, such as the number of hidden layers, the number of neurons in each hidden layer, and which units should be connected, are plentiful. It has been established that a neural network with a single hidden layer is sufficient to approximate functions. However, recent literature challenges this view. For example, Eldan and Shamir (2016) demonstrates that depth (number of hidden layers) is exponentially more valuable than width (number of neurons) for feed-forward neural networks. The simplest setup is a single hidden layer of 32 neurons dubbed NN1. From there, each network is deeper, with fewer neurons added at each step following the geometric pyramid rule (Masters, 1993). NN2 has two layers (32 and 16 neurons), NN3 has three layers (32, 16, and 8), NN4 has four layers (32, 16, 8, and 4), and NN5 with five layers (32, 16, 8, 4, and 2).

Among the many possible activation functions, recent literature (e.g. Glorot et al., 2011; Krizhevsky et al., 2010) highlights the Rectified Linear Unit (ReLU) as an effective choice for the nonlinear function $f(\cdot)$ introduced in Equation (4.18) and is defined as

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise.} \end{cases} \tag{4.20}$$

The idea of ReLU is to add nonlinearity to the network in a way that is fast, simple, and efficient compared to other older methods, such as sigmoid and tanh.

We now define the general form of the multi-layer neural network. We denote $K^{(l)}$ the number of neurons in layer $l \in [1, 2, \ldots, L]$ and the vector of output of neuron k in layer l is denoted as $x_k^{(l)}$. The vector of activations for this layer is then defined as

$x^{(l)} = (1, x_1^{(l)}, \ldots, x_{K^{(l)}}^{(l)})'$. Similarly, the definition of the input layer using N raw predictors is $x^{(0)} = (1, z_1, \ldots, z_N)'$. This allows us to define the recursive output at each neuron in layer $l > 0$ as

$$x_k^{(l)} = \text{ReLU}\left(x^{(l-1)'}\,\theta_k^{(l-1)}\right), \tag{4.21}$$

with the final output given by

$$g(z;\theta) = x^{(L-1)'}\,\theta^{(L-1)}. \tag{4.22}$$

For each hidden layer l, the number of weight parameters is $K^{(l)}(1+K^{(l-1)})$, and the output layer requires an additional $1 + K^{(L-1)}$ weights. The neural network finds the optimal weights by minimising the penalised l_2 objective function of prediction errors. Though neural networks are heavy and hard to interpret, an advantage compared to trees, which use greedy optimisation, is that the neural network allows for joint updates of the parameters along each step, leading to a potentially globally best solution. Training of a neural network is costly due to the formation of nonlinear relationships that give rise to multiple local optima. A common workaround is to apply Stochastic Gradient Descent (SGD) to train a neural network. Instead of using the entire dataset at each step, SGD draws a small random subset of the data.

Given that a neural network is highly parameterised and nonlinear, extra care is needed for regularisation. Alongside l_1 penalties on the weight parameters, we also use four regularisation techniques in our estimation: *learning rate shrinkage*, *early stopping*, *batch normalisation*, and *ensembles*.

First, a key tuning parameter in SGD is the *learning rate shrinkage*, which decides how big each step of the descent is. As the gradient approaches zero, the learning rate needs to shrink toward zero, otherwise random noise overwhelms the direction of learning. To handle the learning rate, we use the learning rate shrinkage method from Kingma (2014).

Second, *early stopping* works by taking an initial parameter value guess that imposes parsimonious parametrisation, for example, setting all θ values close to zero. In each step of the optimisation, parameter estimates are gradually updated to reduce prediction errors in the training sample, thus updating weights to make the model fit the training data better. At the same time, we track prediction errors on a separate validation set. Even if the model improves on the training data, once validation sample errors start to rise, the search is stopped as this flags overfitting. The early stopping usually happens before training errors are fully minimised, which keeps parameters closer to their initial values, but does not let the model train until it fits the training data perfectly. Early stopping is widely used as a cheaper computationally alternative to l_2 regularisation of θ parameters. Early stopping can be used alone or be combined with l_1 regularisation, as we do here.

Third, *batch normalisation* by Ioffe and Szegedy (2015) addresses the problem of internal covariate shifts when training. Internal covariate shifts refer to when the distribution of inputs to hidden layers changes as parameters update, which often causes instability in training and slows down the training. Therefore, at each training step (a batch), the

algorithm normalises the inputs of each layer of a neural network to make the process more stable and to make the network faster.

Finally, we employ an *ensemble* strategy for training our neural network (Hansen and Salamon, 2002; Dietterich, 2000). This means that we estimate multiple networks using different random seeds and construct predictions by averaging their forecasts. Averaging across models reduces prediction variance as the stochastic nature of the optimization can cause different seeds to produce different forecasts. Thus, the ensemble strategy leads to more stable and reliable predictions.

The neural networks are implemented in Python with `TensorFlow` and `Keras`.

4.4. Variable Selection

To avoid blindly expanding the Baseline predictor space established by Gu et al. (2020), we estimate the causal impact of insider signals on returns using the Double Machine Learning (DML) framework developed by Chernozhukov et al. (2018). This framework allows for control of the high-dimensional Baseline predictor set by targeting insider predictors within a Partially Linear Regression (PLR) structure

$$Y_{i,t+1} = D_{i,t}\theta + g(X_{i,t}) + \epsilon_{i,t}, \tag{4.23}$$

where $D_{i,t}$ is the insider signal defined as

$$D_{i,t} = m(X_{i,t}) + v_{i,t}. \tag{4.24}$$

Here, $Y_{i,t+1}$ is future return, θ is the causal parameter of interest (the insider signal's effect on returns, independent of the Baseline predictors), $g(X_{i,t})$ is the nonlinear function of the Baseline predictor set, and $m(X_{i,t})$ is the nonlinear function of insider activity given the Baseline predictor set. The error terms are $\epsilon_{i,t+1}$ and $v_{i,t}$. Standard linear regression is unsuitable because the high-dimensional and nonlinear variables lead to functional form misspecification. However, the application of machine learning models introduces regularisation bias, as they shrink coefficients and prioritise predictive performance over structural parameter estimation. To overcome these issues and safely draw inference on θ, we apply the orthogonalisation technique proposed by Robinson (1988). This process removes the influence of the high-dimensional Baseline predictors from both the future return and the insider signal, and is structured into three steps.

Step one employs machine learning methods to estimate two nonlinear models. First, the expected return, $\hat{g}(X_{i,t})$, capturing the component of $Y_{i,t+1}$ explained by $X_{i,t}$. Second, the expected insider activity, $\hat{m}(X_{i,t})$, capturing the component of $D_{i,t}$ explained by $X_{i,t}$.

Step two removes the Baseline influence by computing the residuals to isolate the unexpected component. This yields the unexpected future return, defined as

$$\tilde{Y}_{i,t+1} = Y_{i,t+1} - \hat{g}(X_{i,t}), \tag{4.25}$$

and the unexpected insider signal, defined as

$$\tilde{D}_{i,t} = D_{i,t} - \hat{m}(X_{i,t}). \tag{4.26}$$

These residuals represent the variation in the future return and the insider signal, respectively, that the Baseline predictor set could not explain.

Step three establishes the causal relationship by regressing future return residuals ($\tilde{Y}$) on insider signal residuals ($\tilde{D}$). This process effectively isolates the effect of insider signals on future returns by controlling for Baseline influences. This method provides an unbiased estimate of the causal parameter θ with valid inference.

The functions $\hat{g}$ and $\hat{m}$ are estimated using a feed-forward neural network. To limit computational costs, we employ a standard off-the-shelf architecture consisting of a single hidden ReLU layer, as detailed in Section 4.3.5. Furthermore, we restrict this operation to the first 8 years of the sample. This ensures variable selection relies solely on ex-ante information, preventing look-ahead bias during the evaluation period.

4.5. Model Evaluation

This section outlines the framework for evaluating model performance. We first measure the out-of-sample predictive accuracy of each specification using R^2_{OOS}. We then assess whether differences in predictive performance are statistically meaningful by applying the Diebold-Mariano test for non-nested comparisons and the Clark-West test for nested models.

The out-of-sample prediction accuracy is measured as the model's prediction error relative to that of a benchmark forecast of zero, defined as

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}}. \tag{4.27}$$

Here, $\mathcal{T}_3$ denotes the set of observations in the out-of-sample test period, ensuring that performance is evaluated strictly on data not used for training or tuning. The variable $r_{i,t+1}$ represents the realised excess return, while $\hat{r}_{i,t+1}$ is the return predicted by the model. Standard forecasting metrics typically benchmark models against the historical average. While this approach is reasonable for aggregate portfolios, Gu et al. (2020) point out that it is flawed when analysing individual stocks. The issue is that historical means for single stocks are incredibly noisy, so much so that they typically perform worse than a simple forecast of zero. Using such a weak benchmark would effectively 'lower the bar' for success, artificially inflating our R^2 results. To avoid this and ensure a conservative evaluation, we follow Gu et al. (2020) and benchmark our predictions against a zero forecast.

For non-nested model comparisons, we employ the Diebold–Mariano (DM) test to evaluate whether the difference in predictive accuracy between two models is statistically significant (Diebold and Mariano, 1995). However, the standard test assumes independent errors, which is unlikely to hold in our setting, given the strong cross-sectional correlations in stock returns. To address this, Gu et al. (2020) adapt the test by aggregating performance across

the panel. Instead of comparing individual stock errors, they compare the cross-sectional average of the squared prediction errors at each time step. Specifically, to evaluate the performance of model (1) relative to model (2), the test statistic is defined as

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}}, \tag{4.28}$$

for which

$$d_{12,t+1} = \frac{1}{n_{3,t+1}} \sum_{i=1}^{n_{3,t+1}} \left(\left(\hat{e}_{i,t+1}^{(1)} \right)^2 - \left(\hat{e}_{i,t+1}^{(2)} \right)^2 \right), \tag{4.29}$$

where $\hat{e}_{i,t+1}^{(1)}$ and $\hat{e}_{i,t+1}^{(2)}$ denote the prediction errors for stock i at time $t+1$ for models (1) and (2), respectively. The term $n_{3,t+1}$ represents the number of stocks in the test sample for month $t+1$. Furthermore, $\bar{d}_{12}$ and $\hat{\sigma}_{\bar{d}_{12}}$ correspond to the sample mean and the Newey-West standard error of $d_{12,t+1}$ over the testing period, respectively. This modified Diebold-Mariano test statistic relies on the time series $d_{12,t+1}$ of average error differences with low autocorrelation. Therefore, it is more likely to satisfy the mild regularity conditions required for asymptotic normality.

While the Diebold–Mariano test is suitable for non-nested model comparisons, our primary interest lies in nested models. Specifically, we aim to determine whether expanding the Gu et al. (2020) Baseline predictor set with insider trading variables improves forecast performance within the same model class. Since the Baseline model is nested within the larger insider model, standard MSE comparisons are biased. To address this, we employ the method proposed by Clark and West (2007) to test for equal predictive accuracy in nested settings. We treat model (1) as the parsimonious benchmark and model (2) as the larger specification that nests model (1). The Clark-West test statistic is defined as

$$CW_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}}, \tag{4.30}$$

where the adjusted loss differential, that is aggregated across the panel, is given by

$$d_{12,t+1} = \frac{1}{n_{3,t+1}} \sum_{i=1}^{n_{3,t+1}} \left[\left(\hat{e}_{i,t+1}^{(1)} \right)^2 - \left(\left(\hat{e}_{i,t+1}^{(2)} \right)^2 - \left(\hat{r}_{i,t+1}^{(2)} - \hat{r}_{i,t+1}^{(1)} \right)^2 \right) \right]. \tag{4.31}$$

The Clark–West framework is closely related to the Diebold–Mariano framework, but differs in how the loss differential is adjusted when comparing nested models. The key adjustment involves the squared difference between the forecasts, $(\hat{r}_{i,t+1}^{(2)} - \hat{r}_{i,t+1}^{(1)})^2$, which accounts for the additional estimation noise introduced by the larger, unrestricted model. Intuitively, in nested comparisons, the unrestricted model often exhibits higher out-of-sample MSE, not because it has inferior predictive power, but because estimating additional parameters inflates forecast variance. The Clark-West adjustment removes the expected noise component from the loss differential, enabling a fair comparison between models.

The statistics $\bar{d}_{12}$ and $\hat{\sigma}_{\bar{d}_{12}}$ denote the sample mean and Newey-West standard error of the adjusted loss differential $d_{12,t}$ over the test period.

4.6. Variable Importance

Beyond predictive performance, this paper also utilises tools to uncover what drives return predictions. More specifically, to identify which characteristics are most important and how insider trading variables alter that dynamic. Consequently, understanding the underlying drivers is an essential exercise as it allows us to verify the economic rationale behind the results.

We adopt the approach of Gu et al. (2020), where variable importance is measured by the drop in panel R^2 when a specific predictor j is set to zero, holding all other parameters constant. This effectively isolates the variable's marginal contribution to predictive accuracy. Specifically, for each window w, we evaluate the fully fitted model on the combined training and validation sample to obtain a baseline performance measure, $R^2_{\text{Full},w}$. We then construct a new dataset in which predictor j is set to zero and re-evaluate model performance, yielding $R^2_{\text{zero}(j),w}$. The importance of predictor j in window w is defined as the performance loss,

$$\Delta R^2_{j,w} = R^2_{\text{Full},w} - R^2_{\text{zero}(j),w}.$$

We aggregate $\Delta R^2_{j,w}$ across all windows to obtain a global measure of predictor importance.

4.7. Portfolio Forecast

Until now, our analysis has focused on predicting monthly returns at the individual stock level. We extend the scope by forming machine-learning-based portfolios and evaluating their aggregate returns to assess economic performance. Portfolio analysis offers several advantages: it reflects economically meaningful aggregates held by investors. It accounts for dependence across stock returns, recognising that strong stock-level predictability does not necessarily translate into economic value at the portfolio level.

In the following section, we describe how model forecasts are translated into investment portfolios and how their economic performance is evaluated. Portfolio performance is assessed using standard measures such as the Sharpe ratio and alpha. We also outline the construction of stock universes using NYSE-based size and illiquidity breakpoints.

4.7.1. Portfolio Construction

We construct machine-learning portfolios following the bottom-up approach of Gu et al. (2020). At the end of each month, we sort stocks into deciles based on their one-month-ahead out-of-sample return forecasts and form a long–short portfolio that buys the highest-predicted decile (decile 10) and sells the lowest (decile 1).

We consider both equal- and value-weighted portfolios to evaluate model performance. Value-weighted portfolios serve as our primary measure of economic performance because

they reflect the distribution of invested capital and are less sensitive to trading costs.[2] However, they may understate predictive performance when signals are strongest among small-capitalisation stocks. Equal-weighted portfolios provide a valuable complement, as weighting schemes do not distort month-ahead predictions and align with the models' objective functions that minimise equally weighted forecast errors (Gu et al., 2020).

4.7.2. Portfolio Performance and Economic Evaluation

Portfolio performance is evaluated using measures such as Sharpe ratio, information ratio, maximum drawdown, portfolio turnover, and alpha.

The Sharpe ratio measures risk-adjusted performance by comparing a portfolio's excess return to its volatility, defined as

$$SR_p = \frac{E[r_p]}{\sigma_p}, \tag{4.32}$$

where $E[r_p]$ is the expected excess return of portfolio p and σ_p its standard deviation.

The information ratio (IR) measures the risk-adjusted performance of an actively managed portfolio relative to a benchmark. We use the excess return of the S&P 500 as our benchmark portfolio. Formally, the IR is defined as

$$IR_p = \frac{E[r_p - r_b]}{\sigma(r_p - r_b)}, \tag{4.33}$$

where r_p denotes the portfolio's excess return, r_b the benchmark's excess return, and $\sigma(r_p - r_b)$ the standard deviation of their return differential.

The maximum drawdown measures the largest cumulative loss (maximum percentage drop from peak to valley) that a portfolio could experience during an investment period before potentially recovering. Formally, the maximum drawdown is defined as

$$\text{Max DD} = \max_{0 \le t_1 \le t_2 \le T} \left(Y_{t_1} - Y_{t_2} \right), \tag{4.34}$$

where Y_t denotes the cumulative return at time t.

Portfolio turnover measures how much a portfolio's composition changes from one month to the next, i.e., the fraction of the portfolio that is bought and sold. Higher turnover implies greater trading activity, which increases transaction costs and can erode excess returns. Formally, the average monthly turnover is defined as

$$\text{Turnover} = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_i \left| w_{i,t+1} - \frac{w_{i,t}(1 + r_{i,t+1})}{1 + \sum_j w_{j,t} r_{j,t+1}} \right| \right), \tag{4.35}$$

where $w_{i,t}$ is the portfolio weight of stock i at month t, $r_{i,t+1}$ is the excess return of stock i from t to $t+1$, and j indexes all assets in the portfolio when computing the denominator.

Alpha (α) measures the component of a portfolio's excess return that is not explained by common risk factors. An α of zero indicates that the factor model fully accounts

[2] Value-weighted portfolios use the previous month's market capitalisation to avoid look-ahead bias.

for the portfolio's excess returns, while a positive (negative) α implies outperformance (underperformance) relative to the factors. We estimate α by regressing excess portfolio returns on the Fama and French (2015) five-factor model augmented with the Carhart (1997) momentum factor, given by

$$R_{p,t} - R_{f,t} = \alpha_p + \beta_p^{SP500}(R_{SP500,t} - R_{f,t}) + \beta_p^{SMB} SMB_t + \beta_p^{HML} HML_t + \beta_p^{RMW} RMW_t + \beta_p^{CMA} CMA_t + \beta_p^{MOM} MOM_t + \varepsilon_{p,t}, \tag{4.36}$$

where α_p represents the component of the portfolio's excess return not explained by the factor model, the coefficients β_p measure the portfolio's exposures to the systematic risk factors, and $\varepsilon_{p,t}$ captures idiosyncratic return innovations[3]. The market factor $(R_{\text{SP500},t} - R_{f,t})$ denotes the excess return on the S&P 500. The remaining factors follow the Fama–French five-factor model and the Carhart momentum factor, capturing size (SMB_t), value (HML_t), profitability (RMW_t), investment (CMA_t), and momentum (MOM_t) premia.

4.7.3. Sharpe Ratio Difference Tests

Following Ledoit and Wolf (2008), we conduct Sharpe ratio tests using the studentised circular block bootstrap, implemented via the `PeerPerformance`[4] package (Ardia and Boudt, 2018). The procedure does not rely on parametric assumptions about the return distribution and is robust to serial dependence and conditional heteroscedasticity.

Let $r_{1,t}$ and $r_{2,t}$ denote the excess returns of two portfolios over $t = 1, \ldots, T$. The Sharpe ratio difference is defined as

$$\Delta S = S_1 - S_2, \tag{4.37}$$

where the Sharpe ratio of portfolio i is given by

$$S_i = \frac{E[r_{i,t}]}{\sqrt{V[r_{i,t}]}}. \tag{4.38}$$

Inference is conducted on the studentized statistic

$$T = \frac{\Delta S}{\widehat{\sigma}_{\Delta S}}, \tag{4.39}$$

where $\widehat{\sigma}_{\Delta S}$ denotes a consistent estimate of the standard error of the Sharpe ratio difference.

Bootstrap samples are generated using a circular block bootstrap applied to the joint return series $(r_{1,t}, r_{2,t})$, ensuring that both temporal dependence and cross-sectional dependence between the two portfolios are preserved. Blocks are drawn with replacement and are allowed to wrap around the end of the sample. The block length is selected using the data-driven procedure proposed by Ledoit and Wolf (2008), which is designed to deliver an accurate test size for Sharpe ratio inference.

[3] Standard errors are adjusted for heteroscedasticity and autocorrelation using the Newey and West (1987) estimator.

[4] https://CRAN.R-project.org/package=PeerPerformance

For each bootstrap replication $i = 1, \ldots, 499$, the Sharpe ratio difference $\Delta S^{(i)}$ is computed from the resampled data, along with a bootstrap standard error $\widehat{\sigma}^{(i)}_{\Delta S}$. The latter is estimated within each bootstrap replication using a block-bootstrap-consistent long-run variance estimator. The bootstrap analogue of the test statistic is given by

$$T^{*(i)} = \frac{\Delta S^{*(i)} - \Delta S}{\widehat{\sigma}^{*(i)}_{\Delta S}}. \tag{4.40}$$

Inference is conducted using the empirical distribution of $T^{(i)}$. Two-sided p-values are obtained by comparing the observed statistic $|T|$ to the bootstrap distribution of $|T^{(i)}|$. The null hypothesis of equal Sharpe ratios is rejected at the 5 per cent level if the observed statistic lies in the tails of the bootstrap distribution.

4.7.4. Heterogeneity by Firm Size and Liquidity

Motivated by Lakonishok and Lee (2001), who document that insider trading profits are most pronounced among small-cap firms, we examine whether the predictive value of insider signals varies with firm size. Each month, stocks are sorted into size terciles based on market capitalisation. Size breakpoints are constructed using NYSE stocks and applied with a one-month lag to avoid look-ahead bias, following Fama and French (1993). Firms are thus classified into small-, mid-, and large-cap groups.

In addition to size-based heterogeneity, we impose a liquidity restriction to assess the implementability of the portfolios, using the illiquidity measure of Amihud (2002). Formally, the measure is computed from daily data and aggregated to the monthly level as

$$\text{ILLIQ}_{i,t} = \frac{1}{D_{i,t}} \sum_{d=1}^{D_{i,t}} \frac{|R_{i,d}|}{\text{Dollar Volume}_{i,d}}, \tag{4.41}$$

where $R_{i,d}$ is the daily return of stock i, $\text{Dollar Volume}_{i,d}$ is the daily dollar trading volume, and $D_{i,t}$ is the number of trading days in month t. Lower values indicate higher liquidity. Although the measure captures the price impact per unit of trading volume, it may still classify very small stocks with low return volatility but steady dollar trading volume as relatively liquid. To mitigate this limitation, we restrict the investment universe to large, liquid firms using NYSE breakpoints.

5. Empirical Results

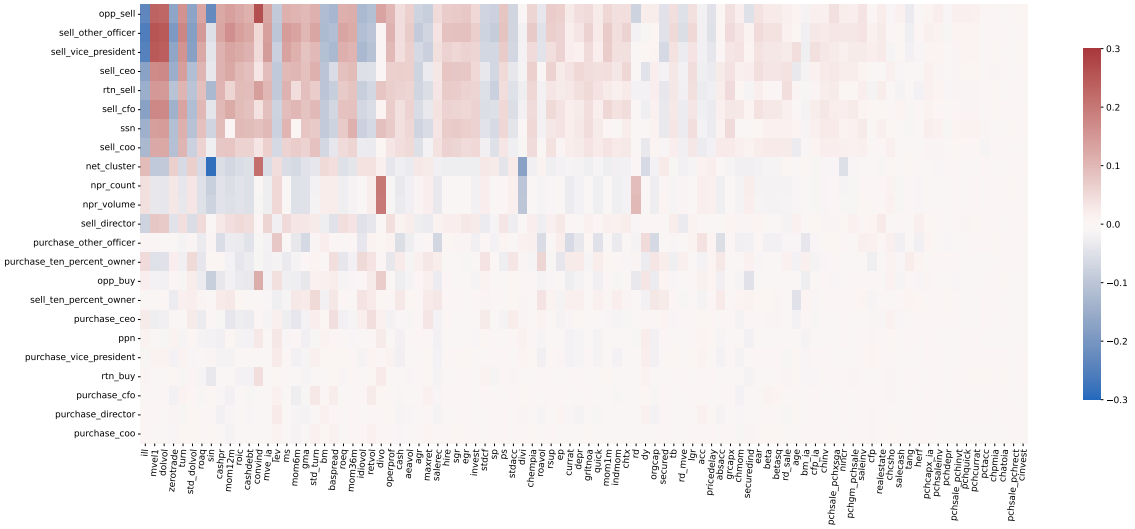
In this section, we present our empirical analysis examining whether augmenting the *Baseline* predictor set of the machine-learning models with *Outsider* and *Corporate Insider* information sets improves return prediction. We proceed in four steps. First, we implement a systematic variable selection procedure to identify insider signals that provide incremental information beyond the *Baseline* specification. Second, we evaluate out-of-sample predictive accuracy to assess whether the insider trading extensions deliver statistically significant

improvements. Third, we analyse variable importance to identify the primary drivers of the predictions. Finally, we consider economic significance by constructing long–short portfolios and comparing risk-adjusted performance measures, including Sharpe ratios, to determine whether the extended models generate superior returns.

5.1. Selecting Insider Trading Variables

In this section, we examine whether the 23 insider trading signals add predictive power beyond the 94 the Baseline characteristics. First, to gauge the correlation structure among the variables, we visually inspect their pairwise correlations. Next, we apply double machine learning (DML) to identify which insider signals provide incremental predictive power after controlling for the the Baseline characteristics.

Figure 5: Correlation Heatmap of Baseline and Insider Trading Variables



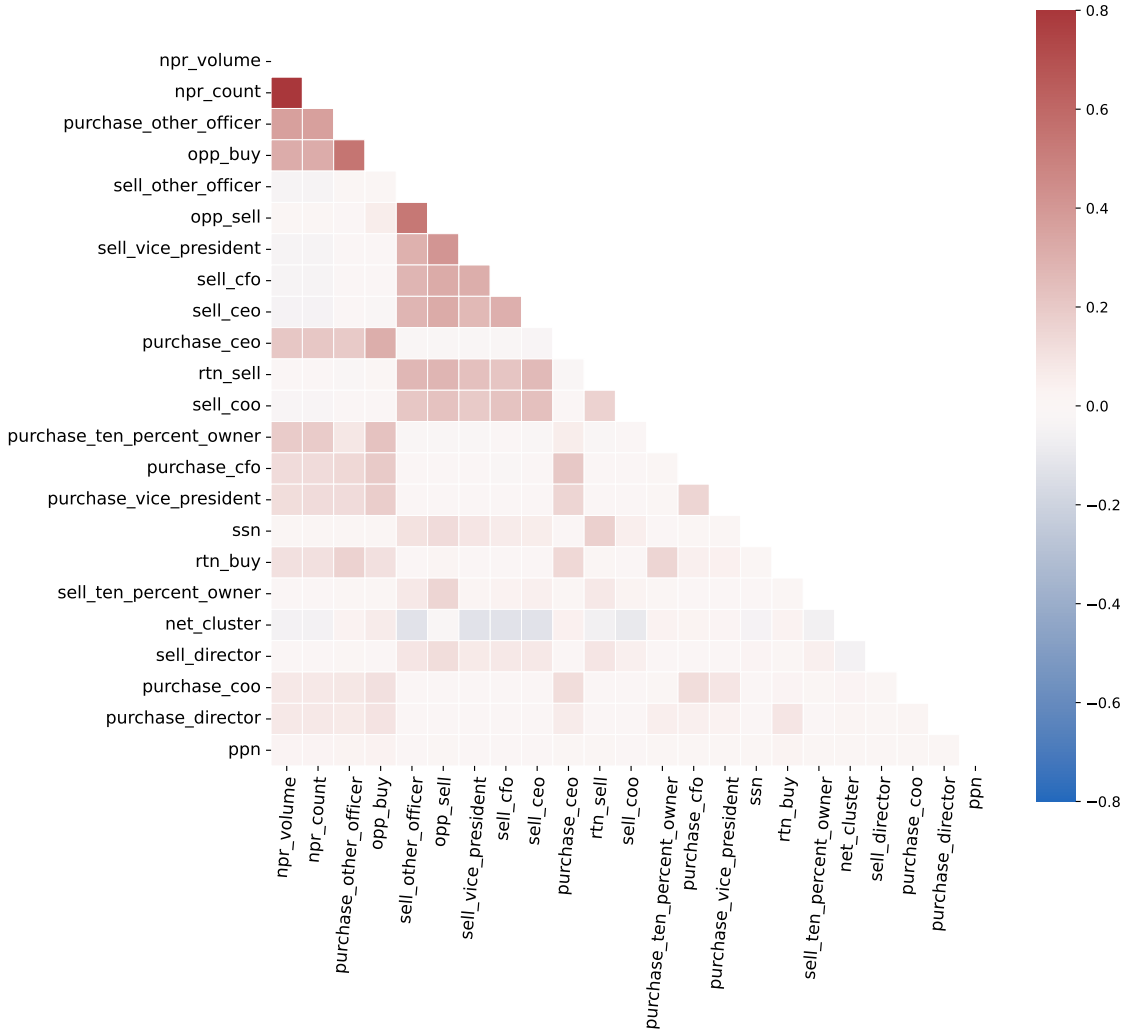
Notes: The 23 Corporate Insider trading variables are ordered by their average correlation (from highest to lowest) with the 94 Baseline predictors, which themselves are ordered according to their correlation-based relevance to the insider signals. The sample covers the first eight-year training window to avoid any look-ahead bias.

Figure 5 shows the correlation heatmap between the insider trading signals and the Baseline characteristics. The heatmap reveals that most correlations cluster around zero (white space), with absolute values rarely exceeding 0.3, which is considered low. This indicates that the insider signals are largely uncorrelated with the Baseline characteristics, thereby minimising concerns about multicollinearity. Digging deeper, insider purchases exhibit near-zero correlation with the Baseline characteristics, whereas insider sales show stronger correlations. The fact that purchase signals are largely uncorrelated with the Baseline suggests that they contain unique information and are therefore more likely to predict returns, consistent with the insider trading literature (Lakonishok and Lee, 2001).

Moreover, the heatmap reveals a clear economic hierarchy. Insider trading is most strongly correlated with trading-activity variables. Insider sales are negatively correlated with illiquidity (*ill*) and zero-trading days (*zerotrade*), and positively correlated with firm

size ($mvel1$), dollar volume ($dolvol$), and turnover ($turn$). This pattern indicates that insiders prefer to sell in large, liquid, and actively traded stocks where trades can be executed with lower price impact. The correlations also reveal patterns across fundamentals. Sales are negatively correlated with sin stocks (sin) but positively correlated with cash productivity ($cashpr$), momentum ($mom12m$), return on invested capital ($roic$), and cash-to-debt ratios. This suggests that insiders tend to sell in financially strong firms following periods of strong price performance. Such behaviour is consistent with the evidence that insiders act as contrarian traders (Piotroski and Roulstone, 2005).

Figure 6: Correlation Heatmap of Insider Trading Variables



Notes: The 23 Corporate Insider insider trading variables on the x-axis are ordered by their average correlation (from highest to lowest). The sample covers the first eight-year training window to avoid any look-ahead bias.

Figure 6 shows the pair-wise correlations among the 23 insider trading variables. Overall, the insider variables are only weakly correlated, with most correlations below 0.5, indicating that they capture distinct dimensions of insider behaviour. The main exception is NPR count and volume, which are strongly correlated at 0.8.

The generally low correlations, both between insider signals and the Baseline characteristics, and among the insider signals themselves, suggest that insider variables may contain information not captured by the Baseline predictors. However, correlation patterns alone cannot establish incremental predictive value. We therefore proceed to systematically test for this using Double Machine Learning (DML), discussed in Section 4.4, which isolates the effect of each of the 23 insider trading signals on excess returns after removing the variation explained by the 94 Baseline factors.

Table 4: Double Machine Learning Coefficients

Insider Signal	Coefficient	p-value
opp_buy	0.029*	<0.001
opp_sell	0.008*	<0.001
purchase_other_officer	0.009*	<0.001
npr_volume	-0.004*	<0.001
rtn_buy	0.021*	<0.001
npr_count	-0.004*	<0.001
purchase_vice_president	0.015*	<0.001
purchase_ceo	0.009*	0.001
sell_ten_percent_owner	-0.008*	0.001
purchase_coo	0.018*	0.006
rtn_sell	0.004*	0.006
sell_ceo	-0.003*	0.008
net_cluster	0.003*	0.012
purchase_cfo	0.009*	0.021
purchase_director	0.013*	0.045
sell_vice_president	-0.001	0.068
purchase_ten_percent_owner	0.005	0.069
ppn	0.007	0.123
ssn	0.002	0.159
sell_cfo	-0.001	0.251
sell_director	-0.001	0.513
sell_other_officer	0.000	0.729
sell_coo	0.000	0.812

Notes: This table reports Double Machine Learning (DML) coefficients for the 23 Corporate Insider insider trading variables measuring their monthly impact on return while controlling for 94 Baseline characteristics. The coefficients are sorted by p-value that is a two-sided test of no effect, with * denoting significance at the 5% level. The sample covers the first eight-year training window to avoid any look-ahead bias.

Table 4 reports the coefficients and p-values for the 23 insider signals, showing their effect on future returns conditional on the 94 Baseline characteristics. We find that 15 out of the 23 variables are significant at the 5% level. As expected, the DML process primarily filters out insider sale signals that act as proxies for the Baseline characteristics. However, highly correlated sale signals, such as opportunistic sell and routine sell, remain significant. Furthermore, opportunistic buy and sell show strong predictive value, with large coefficients and strong statistical significance. Given the high correlation between NPR volume and NPR count shown in Figure 6, and their nearly identical descriptive statistics in Table 3, we drop the marginally less informative NPR count. This selection process results in a final set of 14 insider signals, which we use to assess out-of-sample predictive performance in the next section.

5.2. Predictive Performance

This subsection evaluates hypotheses H_{1a} and H_{2a}. Hypothesis H_{1a} tests whether augmenting the *Baseline* information set with *Outsider* information improves stock-level return predictability, while hypothesis H_{2a} tests whether augmenting the *Baseline* with *Corporate Insider* information yields similar improvements.

We assess out-of-sample return predictability using monthly R^2_{OOS} for ten machine-learning models estimated on the 916 Baseline characteristics of Gu et al. (2020), and augmented with the 14 insider trading signal variables identified as significant by the DML procedure, together with 16 interaction terms constructed by interacting the continuous insider trading variables with the macro predictors.

Changes in predictive power between the Baseline and the two insider-augmented information sets are assessed using Clark–West tests (see Section 4.5). The null hypothesis is that the insider-augmented model does not improve upon the Baseline. One-sided p-values are reported, with an asterisk (*) indicating rejection at the 5% level. The Clark–West statistic is asymptotically standard normal, $\mathcal{N}(0,1)$, and Newey–West standard errors with four lags are used following Drobetz and Otto (2021). Results are presented in Table 5.

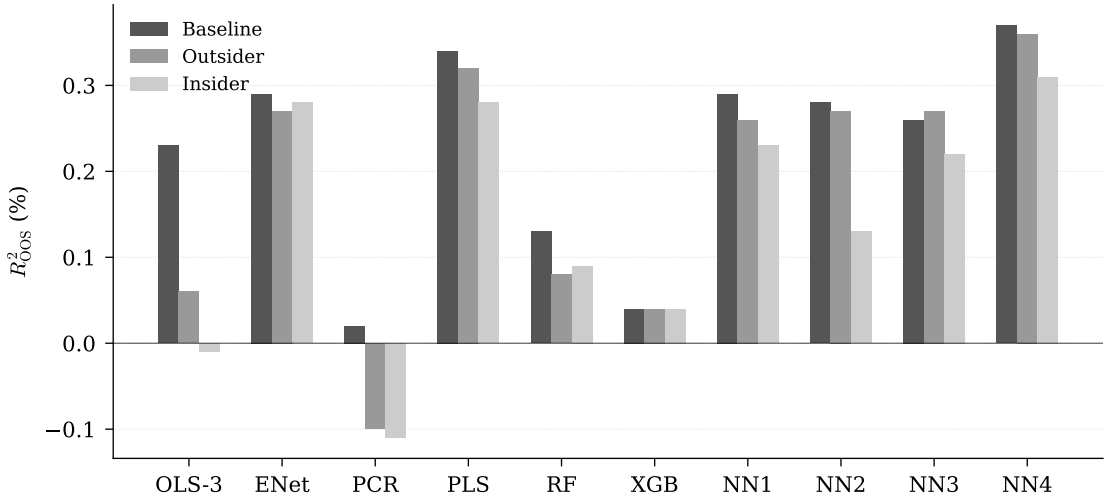
Panel A reports the Baseline results. We find the patterns to be broadly consistent with Gu et al. (2020). Neural networks deliver the strongest predictive performance, with NN4 achieving the highest R^2_{OOS} of 0.37%. Elastic Net (0.29%) and PLS (0.34%) outperform the OLS-3 benchmark (0.23%), while PCR performs poorly (0.02%), reflecting its inability to account for the covariance structure between predictors and returns. Tree-based methods underperform relative to both the linear benchmark and neural networks, likely due to their sensitivity to data variation combined with our shorter sample period (James et al., 2023).

Panel B shows results for the Outsider specification, reflecting public insider trading information. Predictive accuracy, as measured by R^2_{OOS}, varies considerably across models, with some exhibiting sharp performance declines. In particular, OLS-3 experiences a pronounced deterioration, as R^2_{OOS} falls from 0.23% to 0.06%, reflecting its tendency to overfit when additional predictors are included without regularisation. PCR also worsens substantially, declining from 0.02% to –0.10%. The other linear models are comparatively more stable. Elastic Net declines only marginally, from 0.29% to 0.27%, while PLS falls slightly from 0.34% to 0.32%. Among the neural networks, changes are small. NN4 decreases from 0.37% to 0.36%, whereas NN3 increases modestly from 0.26% to 0.27%. Despite these limited changes in R^2_{OOS}, Clark–West tests reveal statistically significant forecast improvements for NN2 ($p = 0.03$) and NN3 ($p = 0.02$), but these are economically insignificant. Overall, these findings provide no support for H_{1a}, that the Outsider is more informed than the Baseline.

Panel C evaluates the Corporate Insider specification, reflecting private information not yet available to the market. In contrast to the Outsider specification, predictive performance deteriorates across nearly all models. OLS-3 collapses from 0.23% to –0.01%, while PCR declines from 0.02% to –0.11%. For the other linear models, the declines are less pronounced. Elastic Net falls slightly from 0.29% to 0.28%, and PLS drops from

Table 5: Monthly Out-of-Sample Stock-Level Prediction Performance

	OLS-3	ENet	PCR	PLS	RF	XGB	NN1	NN2	NN3	NN4
Panel (a): Baseline										
R^2_{OOS} (%)	0.23	0.29	0.02	0.34	0.13	0.04	0.29	0.28	0.26	0.37
Panel (b): Outsider										
R^2_{OOS} (%)	0.06	0.27	-0.10	0.32	0.08	0.04	0.26	0.27	0.27	0.36
Clark-West (p-value)	0.63	0.19	0.40	0.28	0.98	0.42	0.17	0.03*	0.02*	0.16
Panel (c): Corporate Insider										
R^2_{OOS} (%)	-0.01	0.28	-0.11	0.28	0.09	0.04	0.23	0.13	0.22	0.31
Clark-West (p-value)	0.71	0.16	0.32	0.45	0.97	0.46	0.22	0.56	0.27	0.37



Notes: This table reports monthly out-of-sample predictive R^2_{OOS} (in %) for ten models under three information sets. Panel (a) Baseline shows the 94-characteristic specification of Gu et al. (2020), panel (b) Outsider is Baseline plus public insider trading, and panel (c) Corporate Insider is Baseline plus private insider trading at trade execution. Clark–West one-sided p-values test whether each insider augmented model improves forecast accuracy relative to the Baseline where * indicates significance at the 5% level. A graphical comparison of R^2_{OOS} appears below.

0.34% to 0.28%. Nonlinear models also fail to benefit from private insider information. Random Forest declines from 0.13% to 0.09%, while XGB remains flat at 0.04%. Among neural networks, performance declines materially: NN4 from 0.37% to 0.31%, NN3 from 0.26% to 0.22%, and NN2 from 0.28% to 0.13%. Crucially, none of the models exhibit statistically significant Clark–West test statistics. Overall, these findings provide no support for H_{2a} as the Corporate Insider information set does not outperform the Baseline. In a high-dimensional setting, private insider information processed nonlinearly appears to add complexity without providing an incremental signal, as much of it may be absorbed by the rich set of Baseline characteristics. The fact that the Corporate Insider perform worse than the Outsider is particularly unexpected, given that they are endowed with 8% more insider trades. This result stands in contrast to the traditional insider trading literature (Lakonishok and Lee, 2001), which finds that the Corporate Insider are more informative than the Outsider.

To compare predictive performance across models for each information set, we apply

the Diebold–Mariano (DM) test for equal predictive accuracy. Under the null of equal predictive accuracy, the DM test statistic is asymptotically standard normal, allowing us to directly translate its magnitude into p-values. We use Newey–West standard errors with four lags.

Table 6: Diebold-Mariano Statistics for Baseline, Outsider, and Corporate Insider Models

	OLS-3	ENet	PCR	PLS	RF	XGB	NN1	NN2	NN3	NN4
Panel (a): Baseline										
OLS-3		0.63	-1.50*	1.01	-0.90	-0.99	0.37	0.34	0.31	1.76**
ENet			-2.03**	0.68	-1.36	-1.23	0.04	-0.05	-0.21	0.79
PCR				2.41***	0.62	-0.03	1.56*	1.21	1.49*	2.14**
PLS					-1.39	-1.24	-0.49	-0.35	-0.52	0.27
RF						-0.91	0.93	0.86	0.99	1.79**
XGB							0.98	1.01	1.09	1.51*
NN1								-0.08	-0.20	0.73
NN2									-0.15	1.49*
NN3										1.50*
NN4										
Panel (b): Outsider										
OLS-3		1.68**	-1.15	2.14**	0.09	-0.12	1.32*	1.14	1.40*	2.26**
ENet			-2.32**	0.46	-1.34	-1.15	-0.11	0.04	0.02	0.83
PCR				3.23***	0.86	0.51	2.11**	1.66**	1.93**	2.67***
PLS					-1.41	-1.20	-0.63	-0.24	-0.32	0.30
RF						-0.50	1.07	1.21	1.51*	2.33***
XGB							0.95	1.21	1.36*	1.90**
NN1								0.12	0.11	1.09
NN2									-0.06	1.23
NN3										1.85**
NN4										
Panel (c): Corporate Insider										
OLS-3		2.45***	-0.59	2.29**	0.69	0.22	1.73**	0.93	1.65**	2.42***
ENet			-2.49***	0.03	-1.40	-1.28	-0.52	-1.42	-0.57	0.41
PCR				2.91***	0.95	0.52	2.02**	1.13	1.58*	3.11***
PLS					-1.18	-1.12	-0.49	-0.90	-0.36	0.32
RF						-0.73	0.93	0.35	1.23	1.92**
XGB							0.93	0.62	1.25	1.65**
NN1								-0.89	-0.11	0.81
NN2									1.99**	1.90**
NN3										1.02
NN4										

Notes: The table reports Diebold-Mariano (DM) test statistics for pairwise forecast comparisons. Entry (i, j) gives the DM statistic comparing model i (row) with model j (column). Positive values indicate that the column model performs better than the row model, while negative values indicate that the row model performs better. Significance at the 1%, 5%, and 10% levels is denoted by ***, **, and *, using Newey-West standard errors with four lags. Panel (a) reports Baseline models, Panel (b) adds public insider trading information (Outsider), and Panel (c) adds private insider trading information (Corporate Insider).

Table 6 reports the Diebold–Mariano test statistics. Under the Baseline specification in Panel A, we find no systematic evidence that high-dimensional models achieve lower out-of-sample mean squared error than the benchmark OLS-3 model. This contrasts with Gu et al. (2020), who document broad predictive improvements from high-dimensional methods. The only model that statistically outperforms OLS-3 is NN4 ($p < 0.05$), which also attains the lowest out-of-sample MSE across all specifications. Overall, the OLS-3 benchmark performs surprisingly well and, together with Elastic Net, statistically outperforms PCR ($p < 0.10$

and $p < 0.05$, respectively) while also exceeding the predictive accuracy of tree-based models. Notably, PLS significantly outperforms PCR ($p < 0.01$), indicating that supervised dimension reduction yields superior predictive performance relative to unsupervised PCA. Tree-based models perform poorly, in stark contrast to the findings of Gu et al. (2020).

Panels B and C augment the Baseline information set with Outsider and Corporate Insider information, respectively. While the inclusion of insider trading information increases dispersion in statistical significance, the conclusions for Panels B and C remain unchanged. Across both panels, PLS and Elastic Net consistently and statistically outperform OLS-3 and PCR, underscoring the advantages of supervised and regularised linear methods in high-dimensional settings. The relative performance of OLS-3 deteriorates compared to Panel A and is significantly outperformed by several models, while PCR weakens further. Tree-based models continue to exhibit comparatively limited predictive accuracy. By contrast, neural network models exhibit stronger relative predictive performance than under the Baseline, with Elastic Net no longer outperforming them in any specification and with PLS being the only model that outperforms them in some specifications. Along with PLS, neural network models achieve the lowest MSE across all specifications.

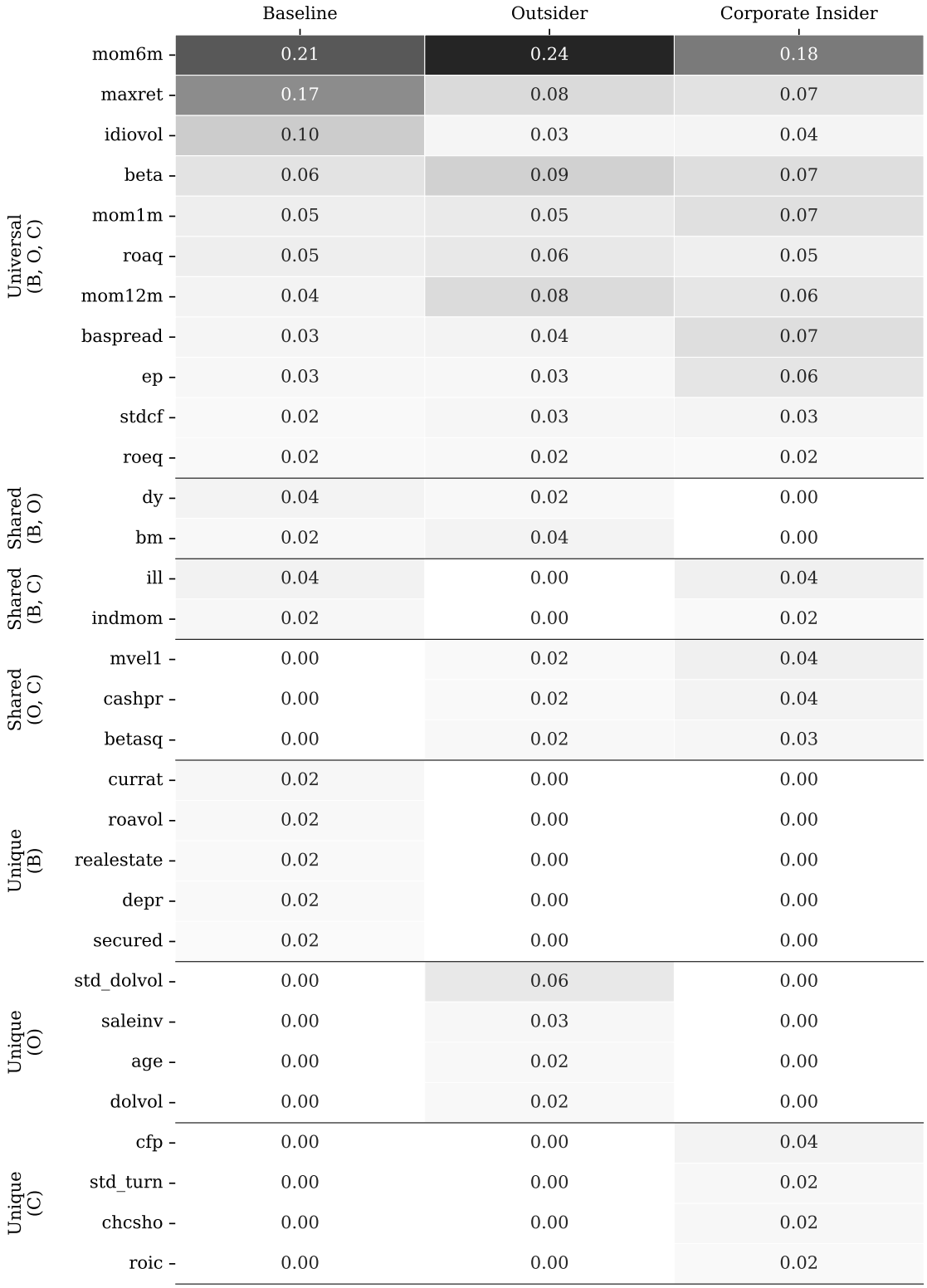
Overall, the inclusion of insider trading information does not lead to a large systematic change in relative model rankings, except for a deterioration in the performance of OLS-3.

The results in this subsection provide no evidence that insider trading information improves stock-level return predictability (R^2_{OOS}) after controlling for the high-dimensional Baseline characteristics. Consequently, we reject both (H_{1a}), that the Outsider is more informative than the Baseline, and (H_{2a}), that the Corporate Insider is more informative. Insider trading information appears to increase model variance rather than add signal, and might largely be explained by the Baseline characteristics. For the later portfolio analysis, we narrow our focus to four models motivated by the DM test results: OLS-3 (linear benchmark), PLS (best dimension-reduction model), and NN3 and NN4 (strongest nonlinear models). The tree-based models are excluded due to their consistently weak performance in our setting.

5.3. Variable Importance

Having established predictive performance, we examine predictor importance across the three information sets. We restrict our attention to NN3, since it is the only model in which adding insider trading variables produces a statistically significant improvement in out-of-sample R^2 under the *Outsider* information set. This allows us to isolate how insider information alters the predictive mapping. Following Gu et al. (2020), for insider trading signals, we restrict the variable importance analysis to the continuous predictors, *npr_volume* and *net_cluster*.

Figure 7: Variable Importance NN3



Notes: Top 20 most important variables in a NN3 setting. Variable importance values are normalised to sum to one. The model specifications are denoted on the y-axis as follows: B (Baseline) is the specification of Gu et al. (2020), O (Outsider) augments the Baseline with public insider trading signals, and C (Corporate Insider) augments the Baseline with private insider trading signals. Variables are categorised by their overlap across models. Universal denotes variables present in the top 20 of all three models (B, O, C), Shared denotes variables present in exactly two models (e.g., B, O), and Unique denotes variables present in only one specific information set.

Figure 7 reports the top 20 predictors for the *Baseline*, *Outsider*, and *Corporate Insider* specifications, with variable importances normalised to sum to one. Across all three information specifications, a stable core of predictors dominates. Return predictability is consistently driven by momentum ($mom1m$, $mom6m$, $mom12m$), risk and trading frictions ($maxret$, $idiovol$, $beta$, $baspread$, $stdcf$), and profitability ($roaq$, $roeq$). The persistence of these variables indicates that the fundamental drivers of return predictability are invariant to the inclusion of insider information. These variables are also found to be important in Gu et al. (2020).

The Baseline specification solely weights on static accounting and balance sheet characteristics, including current ratio ($currat$), earnings volatility ($roavol$), real estate holdings ($realestate$), depreciation ($depr$), and secured debt ($secured$). These variables are complemented by traditional valuation measures such as dividend yield (dy) and book-to-market (bm), as well as by variables capturing market frictions and industry trends, notably illiquidity (ill) and industry momentum ($indmom$). In this sense, the Baseline model most closely resembles a traditional fundamental investor using traditional market and industry information.

Moving from the Baseline to the Outsider specification, the importance structure shifts away from static balance sheet characteristics toward market-based, trading-related, and operational efficiency variables, including liquidity volatility (std_dolvol), sales-to-inventory ($saleinv$), firm age (age), and dollar trading volume ($dolvol$). At the same time, the Outsider shares two key predictors with the Baseline: dividend yield (dy) and book-to-market (bm). The Outsider shares firm size ($mvel1$), cash productivity ($cashpr$), and squared market risk ($betasq$) with the Corporate Insider specification.

The Corporate Insider specification exhibits a departure from the Baseline, reflecting a shift away from static accounting structure toward firm size, cash flow, and operational efficiency. Compared to the Baseline, the Corporate Insider model weights on illiquidity (ill) and industry momentum ($indmom$). Relative to the Outsider, the Corporate Insider emphasises firm size ($mvel1$), cash productivity ($cashpr$), and squared market risk ($betasq$). Finally, the Insider specification uniquely introduces cash flow and capital allocation measures such as cash-flow-to-price (cfp), return on invested capital ($roic$), change in shares outstanding ($chcsho$), and turnover volatility (std_turn). Overall, this pattern is consistent with the literature, which shows that insider trading activity is related to firm size (Lakonishok and Lee, 2001), predicts cash flow (Piotroski and Roulstone, 2005), and is associated with stock repurchases (Lee et al., 1992).

To summarise, the Baseline variable importance pattern largely reflects traditional return predictors. Augmenting the Baseline with insider trading primarily shifts the emphasis toward firm size, cash flow, and capital allocation. However, insider trading variables do not enter the top 20 drivers of return predictability. This suggests that in a high-dimensional machine learning setting, insider information is either muted by or absorbed through the existing Baseline characteristics.

5.4. Machine Learning Portfolios

Leitch and Tanner (1991) argue that statistical predictive performance does not necessarily translate into economic profitability. This motivates us to evaluate our models from an economic perspective by testing hypotheses H_{1b}, H_{2b}, and H_3. Hypothesis H_{1b} examines whether augmenting the *Baseline* information set with the *Outsider* insider trading signals improves the Sharpe ratio, while H_{2b} tests whether adding the *Corporate Insider* trading signals yields a similar improvement. Finally, H_3 assesses whether private information in the Corporate Insider specification provides an incremental advantage over public Outsider information, as reflected in higher Sharpe ratios.

We construct decile-sorted portfolios for the OLS-3, PLS, NN3, and NN4 models. We first examine the mapping from predicted to realised returns across all deciles. For the economic evaluation, we focus on the corresponding long–short portfolios. These portfolios are used to compare the Baseline, Outsider, and Corporate Insider information sets under both value- and equal-weighting schemes. We assess performance using a range of risk metrics, including returns, volatility, maximum drawdown, turnover, and alpha. We then compare cumulative log returns of the long and short legs for each model. Finally, we test whether insider trading has significant predictive value by comparing Sharpe ratios using the method of Ledoit and Wolf (2008).

5.4.1. Portfolio Predictions

This section evaluates predictive performance by examining the relationship between predicted and realised stock returns across portfolio deciles, including long–short portfolios. We consider both value-weighted and equal-weighted schemes. This distinction is important as value-weighted portfolios accurately represent realistic investment strategies. In contrast, the models are trained to minimise prediction error in equal-weighted stock returns rather than to optimise value-weighted portfolio performance.

The portfolio predictive performance over the 2014–2021 out-of-sample period is evaluated using realised mean returns, return volatility, and annualised Sharpe ratios. In addition, we examine the mapping from predicted to realised returns across all portfolio deciles, including the long–short decile portfolios. Results for each model and the three information sets are reported in Tables 7 (value-weighted) and 8 (equal-weighted).

Table 7: Predictive Performance of the Value-weighted Machine Learning Portfolios

	Baseline				Outsider				Corporate Insider			
Percentage	Pred	Avg	SD	SR	Pred	Avg	SD	SR	Pred	Avg	SD	SR
OLS-3												
Low (L)	0.54	1.84	7.20	0.89	0.31	1.74	4.54	1.33	0.18	1.60	4.35	1.27
2	0.59	1.81	6.71	0.94	0.50	1.14	4.60	0.86	0.52	1.24	4.41	0.97
3	0.62	1.52	6.40	0.82	0.56	1.07	4.57	0.81	0.60	0.74	4.80	0.54
4	0.64	1.13	5.38	0.73	0.61	0.54	4.82	0.39	0.66	0.60	4.94	0.42
5	0.67	0.95	5.20	0.63	0.66	0.64	4.78	0.46	0.71	0.86	5.52	0.54
6	0.69	1.30	4.45	1.01	0.71	0.95	5.73	0.57	0.77	0.78	5.99	0.45
7	0.71	1.11	4.52	0.85	0.82	0.66	5.08	0.45	0.91	1.01	5.42	0.65
8	0.74	0.87	4.39	0.69	1.09	0.87	4.21	0.72	1.20	0.98	4.87	0.70
9	0.77	0.90	4.79	0.65	1.54	1.43	4.57	1.08	1.64	1.41	4.69	1.05
High (H)	0.81	0.35	6.09	0.20	2.34	0.96	4.13	0.81	2.53	0.90	4.23	0.74
H–L	0.28	-1.49	6.15	-0.84	2.03	-0.78	2.55	-1.06	2.35	-0.70	2.15	-1.13
PLS												
Low (L)	-0.64	1.06	5.67	0.65	-0.64	1.09	5.55	0.68	-0.68	1.07	5.58	0.66
2	-0.11	0.69	4.94	0.48	-0.07	0.67	4.51	0.52	-0.07	0.69	4.61	0.52
3	0.18	0.82	4.47	0.64	0.24	1.08	5.28	0.71	0.26	1.08	4.48	0.83
4	0.40	1.11	4.69	0.82	0.49	1.05	4.63	0.79	0.51	1.14	4.90	0.80
5	0.61	1.15	4.83	0.82	0.70	0.96	4.25	0.78	0.73	1.05	4.56	0.79
6	0.80	0.91	4.47	0.70	0.92	1.33	4.62	1.00	0.96	1.31	4.53	1.00
7	1.00	1.69	4.25	1.38	1.15	1.29	4.56	0.98	1.20	1.31	4.54	1.00
8	1.22	1.54	4.69	1.13	1.41	1.34	4.42	1.05	1.46	1.23	4.23	1.00
9	1.51	1.76	5.40	1.13	1.74	1.83	5.08	1.25	1.80	1.75	5.10	1.19
High (H)	1.98	1.20	5.76	0.72	2.31	1.12	5.11	0.76	2.40	1.28	5.37	0.83
H–L	2.62	0.14	5.13	0.09	2.95	0.02	4.59	0.02	3.08	0.22	4.66	0.16
NN3												
Low (L)	-0.96	1.04	7.56	0.47	-0.94	0.62	9.01	0.24	-0.94	1.29	7.86	0.57
2	-0.22	1.71	8.11	0.73	-0.17	1.44	7.85	0.64	-0.13	1.06	8.20	0.45
3	0.14	0.69	6.17	0.39	0.18	0.95	5.36	0.61	0.24	0.79	5.96	0.46
4	0.38	1.20	4.84	0.86	0.42	1.07	6.17	0.60	0.48	0.71	5.61	0.44
5	0.57	0.93	5.86	0.55	0.61	0.99	5.44	0.63	0.67	0.97	5.10	0.66
6	0.73	1.10	4.35	0.87	0.77	1.34	5.12	0.91	0.84	0.95	5.04	0.65
7	0.88	0.77	4.47	0.60	0.92	1.18	4.50	0.91	1.01	1.21	4.06	1.03
8	1.04	1.09	4.16	0.91	1.09	1.22	4.17	1.01	1.19	1.08	4.57	0.81
9	1.23	1.21	4.23	0.99	1.30	1.19	4.05	1.02	1.42	1.09	4.22	0.90
High (H)	1.55	1.50	4.48	1.16	1.67	1.16	4.68	0.86	1.85	1.57	4.28	1.27
H–L	2.51	0.47	5.96	0.27	2.62	0.54	7.01	0.27	2.79	0.28	6.36	0.15
NN4												
Low (L)	-0.47	0.92	7.67	0.42	-0.52	1.23	8.15	0.52	-0.49	1.12	7.89	0.49
2	0.03	1.25	6.76	0.64	0.06	1.25	7.08	0.61	0.09	1.60	5.77	0.96
3	0.26	1.40	5.35	0.90	0.31	1.38	6.24	0.77	0.33	1.24	6.07	0.71
4	0.40	1.13	4.67	0.84	0.47	1.24	5.43	0.79	0.49	1.15	4.96	0.80
5	0.51	0.84	4.61	0.63	0.59	1.05	5.19	0.70	0.61	1.17	4.27	0.95
6	0.61	1.28	4.16	1.07	0.70	1.29	4.57	0.97	0.71	1.23	4.58	0.93
7	0.70	1.20	3.84	1.08	0.80	0.98	4.64	0.73	0.82	1.38	4.31	1.11
8	0.80	1.07	4.57	0.81	0.92	1.24	3.96	1.08	0.95	0.93	4.17	0.78
9	0.95	1.36	4.25	1.11	1.08	1.38	4.22	1.14	1.12	1.55	4.77	1.13
High (H)	1.29	2.03	7.06	1.00	1.42	1.77	4.91	1.25	1.54	1.21	5.86	0.71
H–L	1.75	1.11	5.84	0.66	1.94	0.54	6.47	0.29	2.04	0.09	5.95	0.05

Notes: This table reports the performance of prediction-sorted decile value-weighted portfolios over the 8-year out-of-sample period from 2014 to 2021 for the Baseline, Outsider, and Corporate Insider specifications. Columns "Pred", "Avg", "SD", and "SR" denote the predicted monthly return, the average realised monthly return, the standard deviation of realised returns, and the Sharpe ratio, respectively.

Table 8: Predictive Performance of the Equal-weighted Machine Learning Portfolios

	Baseline				Outsider				Corporate Insider			
Percentage	Pred	Avg	SD	SR	Pred	Avg	SD	SR	Pred	Avg	SD	SR
OLS-3												
Low (L)	0.52	1.60	8.00	0.69	0.33	1.66	6.60	0.87	0.09	1.58	6.37	0.86
2	0.58	1.42	6.65	0.74	0.50	1.19	5.97	0.69	0.53	1.31	5.77	0.78
3	0.62	1.19	6.06	0.68	0.56	0.99	6.25	0.55	0.60	1.02	6.26	0.57
4	0.64	0.96	5.79	0.58	0.61	0.77	5.90	0.45	0.66	0.67	5.85	0.40
5	0.67	0.99	5.75	0.60	0.66	0.98	6.00	0.57	0.71	0.94	6.01	0.54
6	0.69	1.00	5.71	0.61	0.71	1.20	6.23	0.67	0.77	1.14	6.36	0.62
7	0.71	0.92	5.61	0.57	0.80	0.99	6.53	0.52	0.88	0.91	6.46	0.49
8	0.74	0.93	5.30	0.61	1.07	0.87	5.66	0.53	1.18	1.02	5.70	0.62
9	0.77	1.03	5.86	0.61	1.51	1.00	5.36	0.65	1.62	0.99	5.44	0.63
High (H)	0.81	0.64	6.57	0.34	2.38	1.02	5.59	0.63	2.53	1.10	5.69	0.67
H–L	0.29	-0.96	6.47	-0.51	2.05	-0.64	2.68	-0.83	2.43	-0.48	2.20	-0.76
PLS												
Low (L)	-0.71	0.22	6.00	0.13	-0.70	0.15	5.98	0.09	-0.72	0.23	5.99	0.13
2	-0.12	0.51	5.67	0.31	-0.08	0.59	5.67	0.36	-0.07	0.49	5.65	0.30
3	0.18	0.85	5.88	0.50	0.24	0.90	5.85	0.53	0.25	0.90	5.88	0.53
4	0.41	0.88	5.50	0.56	0.49	0.96	5.61	0.59	0.51	1.02	5.64	0.63
5	0.61	1.16	5.70	0.70	0.71	0.87	5.73	0.52	0.74	0.95	5.86	0.56
6	0.80	0.84	5.83	0.50	0.92	1.12	5.95	0.65	0.96	1.07	5.92	0.63
7	1.00	1.21	5.79	0.73	1.15	1.16	5.69	0.71	1.20	1.14	5.60	0.70
8	1.23	1.46	6.02	0.84	1.41	1.42	6.00	0.82	1.46	1.49	5.98	0.87
9	1.52	1.58	6.47	0.85	1.74	1.66	6.52	0.88	1.81	1.54	6.39	0.83
High (H)	2.09	1.96	7.47	0.91	2.39	1.84	7.22	0.88	2.48	1.85	7.22	0.89
H–L	2.79	1.75	5.37	1.13	3.09	1.69	5.16	1.14	3.21	1.62	5.16	1.09
NN3												
Low (L)	-1.19	0.09	7.62	0.04	-1.13	-0.06	7.76	-0.03	-1.15	-0.02	7.72	-0.01
2	-0.25	0.70	7.09	0.34	-0.20	0.67	7.03	0.33	-0.17	0.58	6.91	0.29
3	0.13	0.70	6.39	0.38	0.17	0.90	6.34	0.49	0.22	0.85	6.67	0.44
4	0.38	1.21	5.89	0.71	0.42	1.00	6.12	0.57	0.47	0.94	5.92	0.55
5	0.56	1.13	5.79	0.68	0.61	1.28	5.68	0.78	0.67	1.18	5.54	0.74
6	0.72	1.13	5.36	0.73	0.77	1.18	5.60	0.73	0.84	1.35	5.75	0.81
7	0.88	1.13	5.25	0.75	0.92	1.18	5.20	0.79	1.00	1.24	5.32	0.81
8	1.04	1.20	4.98	0.84	1.09	1.29	5.08	0.88	1.18	1.36	5.48	0.86
9	1.23	1.41	4.86	1.00	1.30	1.33	4.93	0.93	1.42	1.27	4.88	0.90
High (H)	1.69	1.97	6.55	1.04	1.81	1.91	6.32	1.05	1.96	1.94	5.69	1.18
H–L	2.88	1.88	4.38	1.48	2.94	1.97	4.61	1.48	3.11	1.96	4.55	1.49
NN4												
Low (L)	-0.66	0.02	7.42	0.01	-0.67	-0.01	7.49	0.00	-0.67	0.30	7.57	0.14
2	0.02	0.71	6.41	0.38	0.04	0.67	6.93	0.34	0.06	1.01	6.99	0.50
3	0.25	0.77	5.81	0.46	0.30	1.04	6.49	0.56	0.33	1.00	6.19	0.56
4	0.40	0.96	5.67	0.59	0.46	0.96	5.68	0.58	0.49	0.82	5.70	0.50
5	0.51	0.91	5.29	0.59	0.59	1.18	5.71	0.72	0.61	0.99	5.31	0.65
6	0.60	1.13	5.16	0.76	0.69	1.02	5.27	0.67	0.71	1.23	5.14	0.83
7	0.70	1.12	5.11	0.76	0.80	1.14	5.11	0.77	0.82	0.83	5.18	0.56
8	0.80	1.28	5.17	0.86	0.92	1.22	5.18	0.81	0.95	1.26	5.21	0.84
9	0.96	1.54	6.33	0.84	1.09	1.35	5.47	0.85	1.14	1.41	5.83	0.84
High (H)	1.47	2.23	7.73	1.00	1.62	2.11	6.78	1.08	1.71	1.82	7.28	0.87
H–L	2.12	2.20	4.17	1.83	2.29	2.11	4.17	1.75	2.37	1.51	4.92	1.07

Notes: This table reports the performance of prediction-sorted decile equal-weighted portfolios over the 8-year out-of-sample period from 2014 to 2021 for the Baseline, Outsider, and Corporate Insider specifications. Columns "Pred", "Avg", "SD", and "SR" denote the predicted monthly return, the average realised monthly return, the standard deviation of realised returns, and the Sharpe ratio, respectively.

For the value-weighted portfolios reported in Table 7, the mapping from predicted to realised monthly returns is generally weak, as realised returns do not increase monotonically with predicted returns. This pattern is not unexpected, since the models are trained to minimise prediction error in equal-weighted stock returns rather than in value-weighted portfolio returns. Within this setting, OLS-3 emerges as a clear outlier. Under the Baseline specification, predicted and realised returns are inversely related: low-decile portfolios earn higher realised returns than high-decile portfolios (e.g., 1.84% versus 0.35% per month), implying that a contrarian long–short strategy (long Low, short High) would dominate the intended trade. Augmenting the information set with Outsider and Corporate Insider signals yields a modest improvement in ordering for OLS-3. The relations become closer to U-shaped rather than strictly reversed, but remain economically weak and far from monotonic. Focusing on the Corporate Insider specification, since the results for the Outsider closely align, the low-decile portfolio continues to earn monthly returns close to the Baseline at 1.60%. In contrast, the high-decile portfolio return increases from 0.35% to 0.90% relative to the Baseline. These partial improvements are accompanied by lower monthly return volatility, with volatility declining from 7.20% to 4.35% for the low portfolio and from 6.09% to 4.43% for the high portfolio, relative to the Baseline. Interestingly, the lower volatility of the long–short (H–L) portfolio (2.15%), combined with its negative mean return of −0.70%, reduces the Sharpe ratio from −0.84 under the Baseline to −1.13 under the Corporate Insider specification. This reflects that, when mean returns are negative, a reduction in volatility increases the persistence of losses and mechanically worsens the Sharpe ratio.

For PLS, NN3, and NN4, the mapping from predicted to realised returns is closer to monotonic, though still imperfect. For example, NN3 predicts a monthly return of −0.96% for the low-decile portfolio, compared with a realised return of 1.04%, for the mid-decile portfolio (decile 5) it predicts a return of 0.57% while the realised return is 0.93%, and for the high-decile portfolio it predicts a return of 1.55%, which closely matches the realised return of 1.50%. Augmenting the Baseline specification with Outsider and Corporate Insider signals does not systematically strengthen cross-decile ordering in the value-weighted results, nor does it materially affect return volatility or Sharpe ratios.

For the equal-weighted portfolios reported in Table 8, the mapping from predicted to realised returns is considerably stronger than for the value-weighted portfolios. For PLS, NN3, and NN4, realised returns increase close to monotonically across deciles. Augmenting the Baseline specification with Outsider and Corporate Insider signals does not, in general, induce systematic changes in mean returns, return volatility, or Sharpe ratios for these models. OLS-3 remains an outlier, exhibiting weak ordering between predicted and realised returns, though the ordering is generally less adverse than under value weighting. Notably, as in the value-weighted case, augmenting the Baseline with Outsider and Corporate Insider signals for OLS-3 reduces both long and short portfolio volatility, leading to smaller long–short Sharpe ratios.

Across both weighting schemes, augmenting the Baseline with Outsider and Corporate

Insider trading does not yield a consistent improvement in the predicted-to-realised mapping across deciles. In some cases, ordering improves modestly, while in others it deteriorates slightly.

To assess whether these insider trading extensions translate into improved economic performance, we next examine portfolio performance measures.

5.4.2. Portfolio Performance

We evaluate portfolio performance and how it changes as we move from the Baseline information set to the publicly available insider trading set (the Outsider) and, subsequently, to private insider information (the Corporate Insider). We first present descriptive performance evidence related to hypotheses H_{1b}, H_{2b}, and H_3, which are later tested using Sharpe ratio difference tests.

Table 9: Performance Statistics by Model: Value- and Equal-weighted Portfolios

	OLS-3			PLS			NN3			NN4		
	B	O	CI	B	O	CI	B	O	CI	B	O	CI
Panel A: Value-weighted Portfolios												
Max DD (%)	82.4	57.6	52.5	36.3	37.6	33.7	43.0	44.4	53.4	25.0	32.3	50.3
Max 1M Loss (%)	14.7	14.7	7.3	14.9	9.3	9.0	17.1	20.5	19.0	11.1	22.7	14.6
Turnover (%)	35.5	141.7	133.8	115.6	128.9	130.8	98.0	112.7	106.3	126.7	120.9	130.0
Mean Ret (%)	-1.5	-0.8	-0.7	0.1	0.0	0.2	0.5	0.5	0.3	1.1	0.5	0.1
FF5+Mom α (%)	-0.6	-0.7	-0.6	0.3	0.1	0.4	0.2	0.4	0.0	0.8	0.6	0.3
$t(\alpha)$	-1.3	-4.0	-4.2	0.7	0.2	0.8	0.3	0.8	0.0	1.8	1.2	0.5
R^2 (%)	33.1	3.3	4.2	20.4	13.0	18.5	32.2	55.3	37.3	20.7	38.7	26.5
IR	-1.0	-1.3	-1.3	-0.5	-0.6	-0.5	-0.3	-0.2	-0.3	0.0	-0.2	-0.4
Panel B: Equal-weighted Portfolios												
Max DD (%)	73.5	49.7	40.2	15.1	14.1	14.6	11.2	9.8	14.9	11.1	11.9	19.5
Max 1M Loss (%)	21.5	13.0	10.2	9.6	9.6	9.7	9.6	8.1	9.4	8.4	9.8	11.3
Turnover (%)	40.9	133.1	135.8	102.4	112.6	113.9	99.8	107.4	106.3	103.5	105.1	102.4
Mean Ret (%)	-1.0	-0.6	-0.5	1.7	1.7	1.6	1.9	2.0	2.0	2.2	2.1	1.5
FF5+Mom α (%)	-0.1	-0.2	-0.1	1.5	1.4	1.3	1.8	2.0	1.7	1.9	2.0	1.7
$t(\alpha)$	-0.2	-1.1	-0.8	3.5	3.3	3.0	5.9	5.8	8.2	5.4	5.6	4.7
R^2 (%)	28.0	30.6	26.1	23.5	19.6	22.0	17.7	31.7	32.7	15.4	22.4	16.0
IR	-0.8	-1.1	-1.0	0.4	0.4	0.3	0.5	0.5	0.5	0.7	0.6	0.2

Notes: This table reports the performance of the machine learning long-short portfolios over the 8-year out-of-sample period from 2014 to 2021. Information sets are denoted as Baseline "B", Outsider "O", Corporate Insider "CI". Panel (a) reports value-weighted and Panel (b) equal-weighted portfolio statistics. Max DD is the maximum peak-to-trough drawdown, Max 1M Loss is the most extreme negative monthly return, and Turnover is the average monthly percentage change in holdings. Mean Ret and FF5+Mom α are average monthly returns and abnormal returns, respectively, expressed in percentages. IR denotes the information ratio computed relative to the S&P 500 excess return benchmark ($R_{\text{SP500},t} - R_{f,t}$). R^2 is the coefficient of determination with respect to the Fama-French five-factor model augmented with the momentum factor.

Table 9 presents risk, return, and risk-adjusted performance metrics for long–short portfolios constructed under the Baseline (B), Outsider (O), and Corporate Insider (CI) specifications across our four models, using both value-weighted and equal-weighted portfolios.

In Panel (a), value-weighted portfolios display weak and unstable performance across models and information sets. For OLS-3, augmenting the Baseline with insider trading information reduces downside risk: maximum drawdowns fall from 82.4% under the Baseline

to 57.6% with Outsider information and further to 52.5% with Corporate Insider information. However, these improvements are accompanied by a sharp increase in turnover, which rises from 35.5% under the Baseline to 141.7% under Outsider and 133.8% under Corporate Insider, and do not translate into improved risk-adjusted performance. Mean returns remain negative, improving only from −1.5% under the Baseline to −0.8% with Outsider and −0.7% with Corporate Insider, while information ratios deteriorate from −1.0 to −1.3. FF5+Momentum alphas remain negative and become more significant, with t-statistics falling from −1.3 under the Baseline to −4.0 and −4.2 under Outsider and Corporate Insider, respectively.

For PLS, insider trading information modestly reduces maximum drawdowns from 36.3% under the Baseline to 33.7% under Corporate Insider, but turnover increases from 115.6% to 130.8%, and mean returns, alphas, and information ratios remain close to zero across all specifications. NN3 exhibits unstable behaviour: mean returns remain unchanged from 0.5% under the Baseline to 0.5% under Outsider, before declining to 0.3% under Corporate Insider, while maximum drawdowns increase from 43.0% to 53.4% from Baseline to Corporate Insider, and all information ratios remain around −0.3. For NN4, both insider specifications weaken performance, with mean returns falling from 1.1% under the Baseline to 0.5% under Outsider and 0.1% under Corporate Insider, alongside information ratios declining from 0.0 to −0.2 and −0.4, respectively. Overall, value-weighted portfolios provide no consistent evidence that insider trading signals improve economic performance after accounting for high-dimensional firm characteristics.

In Panel (b), equal-weighted portfolios display substantially stronger and more stable performance for all models except OLS-3. For OLS-3, insider trading information modestly reduces downside risk, with maximum drawdowns declining from 73.5% under the Baseline to 49.7% with Outsider and further to 40.2% with Corporate Insider, but mean returns remain negative, improving only from −1.0% to −0.6% and −0.5%, while information ratios deteriorate from −0.8 to −1.1 and −1.0, indicating that the fundamental performance issues persist.

By contrast, PLS, NN3, and NN4 achieve economically large and statistically significant abnormal returns across all information sets. Under the Baseline specification, PLS delivers a mean return of 1.7% with an alpha of 1.5% ($t = 3.5$), NN3 achieves a mean return of 1.9% with an alpha of 1.8% ($t = 5.9$), and NN4 reaches 2.2% with an alpha of 1.9% ($t = 5.4$). Augmenting the Baseline with insider trading information does not systematically improve performance. For NN4, performance declines noticeably under insider specifications, with mean returns falling from 2.2% under the Baseline to 2.1% under Outsider and 1.5% under Corporate Insider, and information ratios dropping from 0.7 to 0.6, then to 0.2. For PLS and NN3, alphas and information ratios remain broadly unchanged, with alphas moving only marginally from 1.5% to 1.4% and 1.3% for PLS, and from 1.8% to 2.0% and 1.7% for NN3, indicating no robust incremental value from insider trading signals.

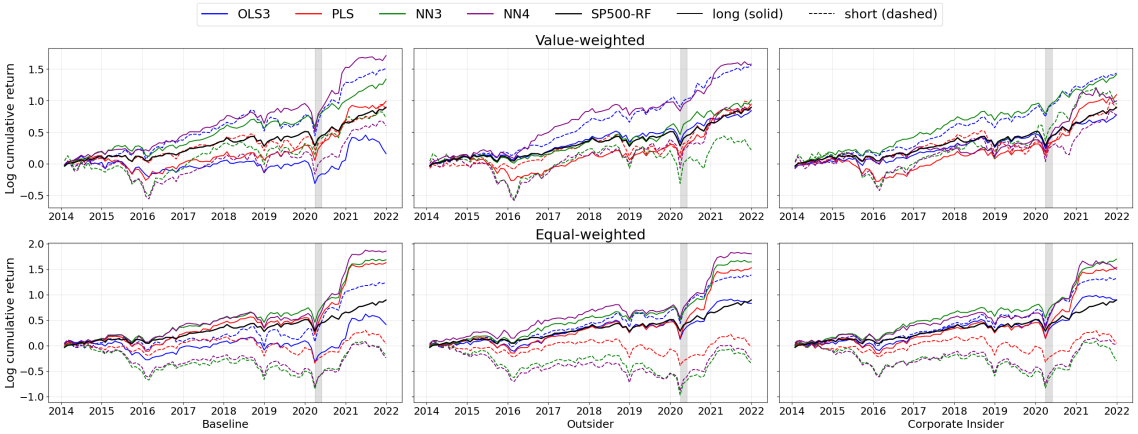
Overall, while equal-weighted portfolios confirm that high-dimensional models can generate strong economic performance, consistent with Gu et al. (2020), insider trading

information, public or private, does not deliver systematic additional gains beyond the Baseline characteristics.

5.4.3. Cumulative Returns

Figure 8 plots cumulative log out-of-sample returns from 2014 to 2021 for the long and short portfolios across our four models. This allows us to compare cumulative performance under the Baseline specification with that under the insider-augmented Outsider and Corporate Insider specifications, relative to the S&P 500 benchmark.

Figure 8: Cumulative log Returns of Machine Learning Based Long and Short Portfolios



Notes: This figure displays the cumulative log returns of portfolios sorted on out-of-sample machine learning return forecasts. The portfolios are constructed by taking a long position in stocks with the highest forecasts (top decile) and a short position in stocks with the lowest forecasts (bottom decile). The solid and dashed lines represent the long and short positions, respectively. Portfolios are value- and equal-weighted, respectively. Shaded periods indicate NBER recession dates.

The value-weighted portfolios in the top panel show that the more complex neural network architectures (NN3 and NN4) substantially outperform both OLS-3 and PLS across all specifications, for both the long and short legs. Importantly, with the exception of OLS-3, extending the Baseline with either insider trading information set does not yield systematically higher returns for the long leg or lower returns for the short leg. Instead, performance generally deteriorates with insider information, with lower returns on the long leg and higher returns on the short leg. By contrast, for OLS-3, the cumulative return of the long leg increases while that of the short leg decreases. These patterns are consistent across both the Outsider and Corporate Insider specifications, suggesting that the economic value of insider signals is concentrated in simpler linear models.

The equal-weighted portfolios in the bottom panel display stronger performance across all models than those in the top panel. This pattern is expected, as the models are trained to minimise prediction error in equal-weighted stock returns. Cumulative returns rise sharply after 2021, likely reflecting the effect of longer training windows. Except for OLS-3, for which both the long and short legs improve markedly when augmented with both Outsider

and Corporate Insider information, we find no systematic improvement or deterioration in returns from extending the Baseline with insider trading information.

Overall, these results indicate that including insider trading information, whether public or private, does not lead to a significant performance improvement across models, except for the long and short legs of the OLS-3 specification.

5.4.4. Sharpe Ratio Difference Tests

Having found no evidence that insider trading provides incremental predictive power once high-dimensional Baseline characteristics are included, we now evaluate whether they nonetheless improve risk-adjusted performance to answer our three economic hypotheses H_{1b}, H_{2b}, and H_3. To do so, we conduct Sharpe ratio difference tests following the methodology of Ledoit and Wolf (2008).

Table 10: Sharpe Ratio Difference Tests

	Δ Sharpe (O-B)	Δ Sharpe (CI-B)	Δ Sharpe (CI-O)
Panel (a): Value-weighted portfolios			
OLS-3	-0.21	-0.29	-0.07
PLS	-0.07	0.07	0.14
NN3	-0.00	-0.12	-0.11
NN4	-0.37	-0.61	-0.24
Panel (b): Equal-weighted portfolios			
OLS-3	-0.31	-0.25	0.07
PLS	0.01	-0.04	-0.05
NN3	-0.01	0.00	0.01
NN4	-0.08	-0.76**	-0.69**

Notes: The table reports annualised Sharpe ratio differences computed from monthly excess returns for value- and equal-weighted portfolios under three information sets: *Baseline (B)*, *Outsider (O)* using public insider trading, and *Corporate Insider (CI)* using private insider trading. Δ Sharpe (O–B) denotes the corresponding differences. Statistical significance is based on two-sided Ledoit and Wolf (2008) test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 10 reports monthly Sharpe ratio differences for value- and equal-weighted long-short portfolios when extending the Baseline information set with Outsider ($O - B$), Corporate Insider ($CI - B$), and Corporate Insider relative Outsider ($CI - O$). Thus, a positive difference indicates an improvement in the insider trading augmented model relative to the Baseline, while a negative difference signals deterioration.

Across models and weighting schemes, the estimated Sharpe ratio differences provide no evidence that either insider trading specification produces Sharpe ratios that are significantly higher than the Baseline. This leads us to reject H_{1b}, which posits that adding Outsider insider trading signals improves the Sharpe ratio relative to the Baseline.

On the other hand, NN4 exhibits significantly negative differences under equal weighting, with the Corporate Insider Sharpe ratio being 0.76 lower than the Baseline at the 5% significance level. Therefore, we reject H_{2b}, which states that augmenting the Baseline with Corporate Insider trading improves return performance.

Moreover, NN4 performs worse under the Corporate Insider specification than under Outsider by 0.69. Since private insider information should not underperform publicly

disclosed insider information, this finding contradicts the expected informational hierarchy. Consequently, we also reject H_3, which posits that the Corporate Insider information set provides incremental return predictability beyond the Outsider information set.

Economically, these results suggest that once a high dimension of firm characteristics is controlled for, neither public nor private insider trading information systematically yields higher economic value for investors. Private insider trading also does not exhibit greater predictive power than public insider trading; if anything, it performs worse.

Having established that insider trading variables do not provide incremental economic value in a high-dimensional machine learning framework, the next section examines whether these findings are robust across firm size and trading liquidity.

6. Robustness: Firm Size and Liquidity

In this section, we evaluate the robustness of our empirical results across firm size and liquidity heterogeneity. First, we examine whether our findings remain consistent across different firm size groups to account for potential heterogeneity. Next, we focus on the most liquid firms to determine whether insider trading affects returns persistently in highly liquid environments. While firm size and liquidity are positively correlated, they are not one-to-one. Large firms can still exhibit limited trading activity due to concentrated ownership, low free float, or market conditions. We focus on long-only, value-weighted High (H) portfolios, as they better reflect practical investment strategies.

This robustness section provides evidence that touches upon hypotheses H_4 and H_5. Hypothesis H_4 examines whether Sharpe ratio gains from insider trading are more pronounced among small-cap firms. H_5 tests whether any performance gains from insider trading are robust once low trading frictions (highly liquid environments) are considered.

6.1. Firm Size

Given evidence that the predictive power of insider trading is strongest among small stocks (Lakonishok and Lee, 2001), we examine whether the predictive content of insider trading signals varies across firm size groups. We therefore construct small-, mid-, and large-cap groups based on market capitalisation, see Section 4.7.4 for details. High portfolios are derived from quartile sorts on predicted returns to ensure diversified portfolios across size groups.

Table 11: Sharpe Ratios and Sharpe Ratio Differences by Information Set Across Firm Size Groups

	Sr (B)	Sr (O)	Sr (IC)	ΔSr (O–B)	ΔSr (IC–B)	ΔSr (IC–O)
Panel A: Large-cap firms (H)						
OLS-3	0.37	1.03	0.92	0.66**	0.55	-0.11
PLS	1.18	1.09	1.05	-0.09	-0.13	-0.05
NN3	1.18	1.10	1.17	-0.09	-0.01	0.07
NN4	1.00	1.24	1.04	0.24	0.04	-0.20
Panel B: Medium-cap firms (H)						
OLS-3	0.34	0.57	0.60	0.24	0.26*	0.03
PLS	0.67	0.67	0.69	0.00	0.02	0.02
NN3	0.66	0.73	0.81	0.08	0.15	0.07
NN4	0.68	0.72	0.70	0.05	0.02	-0.03
Panel C: Small-cap firms (H)						
OLS-3	0.32	0.50	0.54	0.19	0.22*	0.03
PLS	0.68	0.68	0.67	0.00	-0.00	-0.01
NN3	0.74	0.77	0.78	0.03	0.04	0.01
NN4	0.71	0.77	0.63	0.07	-0.08	-0.15

Notes: The table reports annualised Sharpe ratios and Sharpe ratio differences for value-weighted *High* (H) portfolios formed within firm-size groups. The three information sets are *Baseline (B)*, *Outsider (O)* using public insider trading, and *Corporate Insider (CI)* using private insider trading. Statistical significance of Sharpe ratio differences is based on two-sided Ledoit and Wolf (2008) tests applied to monthly excess return series: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 11 presents Sharpe ratios for long-only high (H) portfolios segmented by market capitalisation. The results show that insider trading leads to statistically significant improvements in Sharpe ratios only for portfolios constructed from the low-dimensional OLS-3 model, with gains most pronounced among large-cap stocks. This finding contrasts with the insider trading literature, which typically documents stronger predictive power among small-cap firms (Lakonishok and Lee, 2001).

This result also contrasts with Table 10, which evaluates Sharpe ratio differences for long–short portfolios over the full investment universe and finds no significant Sharpe ratio improvements for OLS-3. Focusing on long-only portfolios reveals significant improvements in Sharpe ratios across firm sizes and specifications. This result is consistent with our earlier finding that the mapping from predicted to realised returns for OLS-3 improves under insider information sets, but that these gains come at the cost of lower long–short Sharpe ratios, driven by reduced return volatility combined with negative mean returns.

We argue that the incremental value of insider trading information depends partly on firm size but is driven primarily by model complexity. Under OLS-3, Sharpe ratio improvements are most pronounced among large-cap stocks, with a statistically significant ($p < 0.05$) annualised increase of 0.66 in the Sharpe ratio relative to the *Baseline* when *Outsider* information is included. Because OLS-3 relies on only three predictors, augmenting it with insider trading information meaningfully expands the information set and partially mitigates omitted-variable bias. By contrast, as model complexity increases, the incremental contribution of insider trading diminishes. High-dimensional machine-learning models are

likely to have already internalised the economic content of insider trading through correlated firm characteristics, rendering insider trading signals largely redundant or potentially noisy.

Importantly, the economic gains observed for OLS-3 under insider information sets do not translate into generalised predictive improvements, as shown in Table 5.2. Extending the Baseline specification with insider trading information substantially reduces out-of-sample R^2 for OLS-3, from 0.22% under the Baseline to around 0.06% and −0.01% under the Outsider and *Corporate Insider* specifications, respectively. This suggests that sparse linear models improve economic performance primarily by altering factor exposures rather than by improving return forecasts. In contrast, high-dimensional machine-learning models exhibit little to no improvement in predictive accuracy or Sharpe ratios when insider trading information is added to the Baseline.

Overall, these findings indicate that earlier studies documenting predictive gains from insider trading information, particularly among small-cap firms, may not be robust once a rich, high-dimensional information set is employed. After controlling for firm size, we find no evidence that improvements in the Sharpe ratio from augmenting the information set with insider trading are more pronounced among small-cap firms. Accordingly, we reject H_4.

Next, we evaluate robustness to trading frictions by restricting the stock universe to liquid stocks.

6.2. Liquidity

Malkiel (2003) argues that many documented anomalies lose economic significance after accounting for market frictions. Since liquidity proxies for such frictions, we focus on the most liquid firms to better isolate the role of trading frictions in insider trading performance. This motivates the use of a direct liquidity filter for a cleaner assessment.

The liquidity-restricted universe is constructed by jointly sorting stocks on illiquidity and size. We retain only large-cap stocks in the lowest illiquidity tercile; see Section 4.7.4 for details. High (H) portfolios are formed using quartile sorts on predicted returns to ensure diversified exposure across size–illiquidity groups.

Table 12: Performance Statistics by Model

	OLS-3			PLS			NN3			NN4		
	B	O	CI	B	O	CI	B	O	CI	B	O	CI
Max DD (%)	30.45	18.50	19.65	17.71	14.99	15.81	14.01	15.33	15.52	17.17	16.04	14.93
Max 1M Loss (%)	19.42	11.15	12.46	11.04	10.16	10.51	9.31	8.84	8.56	11.17	8.95	9.68
Turnover (%)	31.42	141.12	139.18	87.61	108.14	107.20	61.35	81.26	77.39	72.26	70.34	80.70
Mean Ret (%)	0.45	1.19	1.11	1.36	1.34	1.30	1.43	1.36	1.36	1.19	1.41	1.31
FF5+Mom α (%)	-0.35	0.08	0.01	0.41	0.29	0.23	0.36	0.25	0.28	0.14	0.35	0.25
$t(\alpha)$	-2.83	0.62	0.11	3.49	2.47	2.26	3.37	2.54	3.14	0.97	4.89	2.09
R^2 (%)	91.10	93.38	93.86	90.46	91.73	92.41	89.82	92.01	93.23	92.55	94.16	93.11
SR	0.36	1.00	0.93	1.14	1.10	1.05	1.20	1.11	1.17	1.02	1.23	1.05
IR	-1.13	0.51	0.28	0.88	0.85	0.76	0.96	0.88	0.94	0.51	1.14	0.78

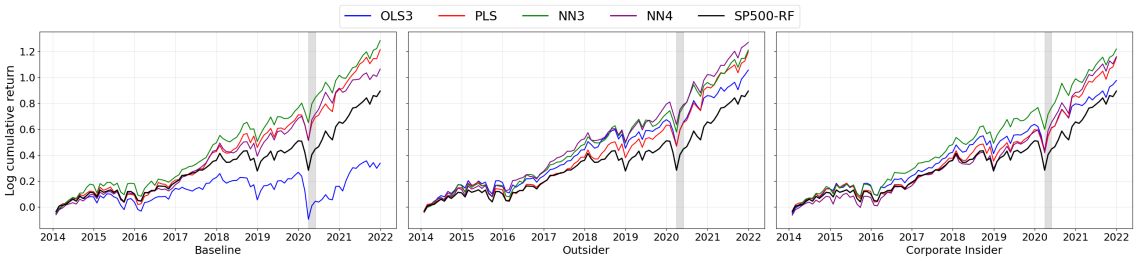
Notes: This table reports the performance of machine learning long (High) tercile value-weighted portfolios over the out-of-sample period from 2014 to 2021. The three information sets are *Baseline (B)*, *Outsider (O)* using public insider trading, and *Corporate Insider (CI)* using private insider trading. Max DD is the maximum peak-to-trough drawdown, Max 1M Loss is the most extreme negative monthly return, and Turnover is the average monthly percentage change in holdings. Mean Ret and FF5+Mom α are average monthly returns and abnormal returns, respectively, expressed in percentages. SR denotes the annualised Sharpe ratio computed from monthly excess returns. IR denotes the information ratio computed relative to the S&P 500 excess return benchmark. R^2 is the coefficient of determination from the Fama-French five-factor model augmented with momentum.

Performance statistics for the liquidity-restricted universe are presented in Table 12. The R^2 values from the five-factor plus momentum model exceed 90% across all specifications, indicating strong comovement with the aggregate market and market betas close to unity. This result is consistent with the composition of the liquidity-restricted universe, which closely resembles a large-cap market portfolio.

All models and specifications, except the Baseline OLS-3, deliver positive monthly abnormal returns, economically meaningful Sharpe ratios around one, and high information ratios. Augmenting the Baseline specification with Outsider or Corporate Insider information does not systematically improve performance across models. The main exceptions are OLS-3 and NN4. For OLS-3, the FF5+Mom alpha increases from -0.35% under the Baseline to 0.08% with Outsider information and 0.01% with Corporate Insider information, alongside an increase in the Sharpe ratio from 0.36 to approximately 1.00. For NN4, Outsider information raises the monthly alpha from 0.14% to 0.35%, with a corresponding increase in the Sharpe ratio from 1.02 to 1.23. In contrast, Corporate Insider information yields a smaller gain and is closer to the Baseline.

To address market frictions, we focus on the specification with the lowest turnover, the Baseline NN3 model, which has an average monthly turnover of 61.35%. Given a monthly abnormal return of 0.36%, a simple back-of-the-envelope calculation shows that transaction costs of at most $c \leq \alpha/\text{Turnover} = \frac{0.36\%}{61.35\%} \approx 0.59\%$ per month would fully offset abnormal performance. For comparison, Frazzini et al. (2018) report average transaction costs for large-cap stocks of 9.93 basis points (bps), with a standard error of 0.73 bps, over the 1998–2016 period. These costs are far below the 0.59% threshold, suggesting that the strategy is plausibly implementable at a profit.

Figure 9: Cumulative log Returns of Machine Learning Based Long Portfolios



Notes: This figure displays the cumulative log returns of portfolios sorted on out-of-sample machine learning return forecasts. The portfolios are constructed by taking a long position in stocks with the highest forecasts (top quartile). The investment universe is restricted to large liquid stocks only. Shaded periods indicate NBER recession dates.

Figure 9 plots cumulative log out-of-sample returns for the Baseline, Outsider, and Corporate Insider specifications in the liquidity-restricted universe. The figure visually corroborates the table-level evidence. Under the Baseline specification, except for OLS-3, all models closely track the S&P 500, reflecting strong market comovement while achieving slightly higher cumulative returns over the sample period. Augmenting the information set with Outsider and Corporate Insider trading signals leads to a marked improvement in cumulative performance for OLS-3. Among the high-dimensional models, NN4 is the only one to exhibit a notable improvement under both the Outsider and Corporate Insider specifications, with the highest cumulative returns achieved under the Outsider specification and a more minor but apparent increase relative to the Baseline under the Corporate Insider specification.

Table 13: Sharpe Ratio Differences by Information Set: Liquid-Universe Tercile High Portfolios

	ΔSr (O–B)	ΔSr (IC–B)	ΔSr (IC–O)
OLS-3	0.64	0.57	-0.07
PLS	-0.04	-0.08	-0.04
NN3	-0.09	-0.03	0.06
NN4	0.22*	0.03	-0.18

Notes: The table reports differences in annualised Sharpe ratios for *High* (H) portfolios formed in the restricted, liquid stock universe using tercile sorts. The three information sets are: *Baseline (B)*, *Outsider (O)* using public insider trading, and *Corporate Insider (CI)* using private insider trading. Δ Sharpe (O–B) denotes the corresponding differences. Statistical significance is based on two-sided Ledoit and Wolf (2008) test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Sharpe ratio differences for the liquidity-restricted universe are reported in Table 13. Overall, only the NN4 model augmented with the Outsider specification delivers a statistically significant and economically meaningful increase in the Sharpe ratio of 0.22. Consistent with this, Table 12 shows that NN4 under the Outsider specification achieves a monthly abnormal return of 0.35%, the highest Sharpe ratio (1.23), and the highest information ratio (1.14) across all models and information sets in the liquid universe. A natural question is whether this strong performance survives transaction costs. Given its average monthly

turnover of 70.34%, a monthly abnormal return of 0.35% implies that transaction costs of at most $c \leq \frac{\alpha}{\text{Turnover}} = \frac{0.35\%}{70.34\%} \approx 0.50\%$ per month would fully offset the excess return. Drawing on Frazzini et al. (2018), who estimate average large-cap transaction costs of 9.93 basis points, this threshold is above realistic trading costs, suggesting that NN4 with the Outsider specification could plausibly be implemented profitably. Therefore, this indicates that a significantly better insider-augmented machine learning model may generate abnormal returns after accounting for transaction costs. Thus, we find pieces of evidence supporting H_5.

Overall, the robustness analysis provides little evidence that expanding the information set improves performance after accounting for firm size. Any gains that do appear are confined to simple linear models and do not generalise to more flexible specifications. In the liquidity-restricted universe, we find modest evidence of improved return predictability for insider trading in the NN4 model, robust to transaction costs, providing partial support for H_5. Taken together, these results suggest that the strong insider effects documented in earlier studies largely reflect the limitations of low-dimensional modelling approaches.

7. Discussion

To facilitate an informed discussion, we synthesise the empirical findings. We find that insider trading provides no systematic incremental predictive value once machine learning models process a high-dimensional set of firm characteristics. Any economic value from insider trading is confined to the simple linear model in a low-dimensional setting and to a single case in which NN4 improves performance among the most liquid firms. The discussion addresses the limitations of machine learning approaches for predicting stock returns, relates our findings to the existing literature, explores their practical implications for investors, and outlines directions for future research.

7.1. Limitations of Machine Learning in Return Prediction

While stock returns are notoriously noisy, machine learning methods have proven effective in predicting stock returns in high-dimensional and flexible settings, as they are capable of capturing nonlinear relationships that pre-specified linear models fail to detect (Gu et al., 2020). These findings have subsequently been shown to be robust across different geographic markets and when accounting for transaction costs (Drobetz and Otto, 2021; Leippold et al., 2022; Hanauer and Kalsbach, 2023). Despite the success of machine learning in predicting stock returns, its application faces limitations. We comment on overfitting, interpretability, and economic viability.

Machine learning models are flexible, making them inherently susceptible to overfitting. This is particularly the case when the signal-to-noise ratio is low and data are limited, as is often the case for stock returns. Stock returns are heavily influenced by randomness, behavioural biases, non-repeating shocks, regime changes, and evolving market structures. Consequently, patterns learned during the in-sample training period often fail to generalise,

leading to a breakdown in predictive performance out-of-sample, known as overfitting. Overfitting is a plausible driver of the lack of systematic prediction gains when augmenting the high-dimensional Baseline predictor set with insider trading variables. Because the high-dimensional Baseline already captures a significant portion of the relevant predictive signals, the addition of insider trading variables likely introduces more noise than signal, leading the model to fit patterns that do not generalise out-of-sample.

The challenge of overfitting in machine learning performance is emphasised by Martin and Nagel (2022). Their central argument is that strong in-sample predictability in a "Big Data" setting is often a statistical illusion rather than something investors can exploit out-of-sample. From this perspective, our finding of out-of-sample predictability would be economically interesting, as it would suggest that our models have identified a genuine, persistent signal rather than 'hindsight' noise from the past. They show that in high-dimensional environments, strong in-sample performance can arise even when investors process information optimally in real time. The reason is that when researchers evaluate historical data ex-post, they can observe the precise mistakes investors made while learning the underlying parameters. A flexible machine learning model can then connect the dots between those historical errors and subsequent returns, creating the appearance of a stable predictive structure. However, these patterns largely reflect transient estimation errors rather than persistent economic relationships. As a result, they tend to disappear when evaluated out-of-sample, because random shocks and informational noise do not repeat in the future.

Another limitation is the limited economic interpretability of machine learning models, often referred to as "black boxes," especially neural networks, which makes it difficult to distinguish genuine risk compensation from spurious statistical regularities (Molnar, 2020). The rise of machine learning prompts a fundamental tension in finance: should one prioritise the complexity of neural networks or the transparency of linear regression? While neural networks can detect patterns that simpler models miss, they often lack an intuitive economic logic. For investors, predictions alone are insufficient; understanding the economic forces driving a trade is essential. These concerns highlight the difficulty of navigating high-dimensional environments and the importance of drawing careful conclusions from out-of-sample evaluations and economic performance metrics. This lack of interpretability underscores a key challenge of this paper, as the primary hurdle is diagnostic. Because the models lack transparency, we cannot easily explain the decline in predictive power after adding insider trading signals.

Recent work has therefore questioned the economic viability of machine learning based investment strategies. Avramov et al. (2023) show that much of the profitability attributed to machine learning is concentrated in micro-cap and distressed stocks that are costly or infeasible to trade. Our findings suggest otherwise. When the sample is restricted to the most liquid stocks and to long-only portfolios, typically larger firms with lower execution barriers, the abnormal returns across most models and information sets become significant and potentially robust to transaction costs. Moreover, it is in this setting that we find

the only instance in which a nonlinear model shows improvement when augmented with the Outsider insider trading dataset. This suggests that the models ability to extract meaningful signals from insider activity is not universal but instead confined to specific segments of the investment universe.

7.2. Relation to the Existing Literature

Relation to Gu, Kelly, and Xiu (2020) Our Baseline results are broadly consistent with the central insight of Gu et al. (2020) that return predictability is better captured by flexible, nonlinear models than by sparse linear specifications. However, we also find that more complex models, most notably regression trees, underperform the OLS-3 benchmark. In our setting, tree-based models struggle to cope with the low signal-to-noise ratio in stock-level return predictability. This likely reflects the shorter sample period, which substantially reduces the number of firm-month observations relative to Gu et al. (2020). With a limited sample size, validation loss can become flat and noisy, making it challenging to balance overfitting and underfitting.

Similar challenges are observed for Elastic Net regularisation. Early in the sample, the training window is too short for reliable variable selection, causing the LASSO component to dominate and shrink all slope coefficients to zero, leaving only the intercept. As the training sample expands, variable selection stabilises and predictive performance improves markedly. In later periods, the resulting gains are sufficiently strong that Elastic Net statistically outperforms the OLS-3 benchmark, consistent with Gu et al. (2020).

Furthermore, the composition of our sample period captures a regime of increased efficiency, in which technological advancements and greater liquidity have reduced the magnitude of market anomalies (Chordia et al., 2014). Significantly, the recent rise in machine learning has accelerated the search for complex return predictability. As these patterns are discovered and traded upon, they are rapidly arbitraged away, leaving less scope for persistent out-of-sample relationships in modern data (McLean and Pontiff, 2016). By contrast, Gu et al. (2020) evaluate return predictability over a long horizon starting in 1957, benefitting from earlier decades when markets were less efficient and offered greater scope for complex models to extract persistent patterns. Taken together, these factors might help explain why tree-based models and PCR, despite performing well on the larger sample in Gu et al. (2020), exhibit weaker and less stable out-of-sample performance in our more recent setting.

Relation to the Insider Trading Literature Our results partially confirm the economic intuition of the insider trading literature while highlighting the redundancy of these signals in high-dimensional settings. Consistent with Piotroski and Roulstone (2005), our variable importance analysis indicates that insider trading models put more weight on cash flow metrics. Unlike Lakonishok and Lee (2001), who document significant abnormal returns for Corporate Insiders, we find no systematic incremental value when insider trading is benchmarked against a high-dimensional predictor set. While adding insider variables to a

low-dimensional linear model improves performance (suggesting the presence of omitted variable bias), this advantage disappears in a high-dimensional setting. This contrast highlights that the bar for "economically meaningful" incremental value is significantly higher in complex models. Consequently, when comprehensive characteristics and macro variables are modelled flexibly, insider signals become largely redundant, making the predictive power of insider trading conditional on the complexity of the data benchmark.

Another key distinction between our findings and the consensus view lies in the evaluation framework. Lakonishok and Lee (2001) utilise long holding periods and in-sample evaluation, finding the most substantial insider effects among small firms with high information asymmetry. By contrast, our framework relies on monthly rebalancing and strict out-of-sample evaluation. Under this regime, the "small firm" advantage disappears. Instead, we find that insider signals appear predictive among larger firms when using a simple linear model. A tangible explanation may lie in the distribution of insider trades. As shown in Figure 1, large-cap firms account for roughly 50% of all insider trades, despite representing a much smaller share of the total number of firms. This means that insider trading data are far more abundant for large-cap firms, giving simple linear models more information to learn from and naturally improving estimates in that segment. However, this alone does not explain why our results differ from the existing literature, as their conclusions also draw on similar observations.

One reason the Outsider specification systematically fails to predict returns may be delays in the information flow. In our framework, if an insider trade is filed or executed on, for example, 10 January, the model incorporates this information to predict returns in February, resulting in a roughly 20-day delay. Rogers et al. (2017) show that these trades are often realised within ten days, creating a disconnect between the trade and the model's reaction. They find that professional traders with direct database access enjoy an 81-second window of private advantage, capturing approximately 0.28% of the return before the public even observes the insider filing in EDGAR. The total return reaches 1.01% by the market close on the filing day, and prices drift by an additional 1.60% over the subsequent ten days (Rogers et al., 2017). Consequently, the vast majority of the price adjustment, approximately 2.61%, is fully realised within ten days of the filing. By the time our Outsider model rebalances in February, often more than ten days after the insider filing, this information has already been fully reflected in prices.

We initially hypothesised that the Corporate Insider specification would deliver superior return prediction compared to the Outsider, premised on the assumption that insiders possess and act on private information before it reaches the public. Paradoxically, our results indicate that the Corporate Insider generally performs worse than the Outsider specification. Two structural mechanisms can explain this counterintuitive finding. First, a potential explanation for the limited predictive power of Corporate Insider transactions is the misalignment between their investment horizons and our one-month target return. Insiders often trade based on long-term structural changes that may take months or years to be fully priced into the market Seyhun (2000). Thus, a one-month return window likely

fails to capture the full value of an insider's private information. However, this long-term perspective should partially be captured by the NPR variable, which aggregates insider activity over a six-month window. Furthermore, the immediate price reaction is often dampened by filing delays. Since Corporate Insiders cannot move prices in isolation, the market only reacts once Outsiders observe and act upon the public disclosure of these trades. This delay suggests that the explanatory power of Corporate Insider information is frequently absorbed by Outsider signals, which appear more robust because they reflect the market's eventual response to delayed disclosure. Second, the complex legal and regulatory environment in financial markets constrains Corporate Insiders' ability to trade aggressively on material non-public information.

7.3. Implications for Investors

Our findings offer several actionable insights for investors and portfolio managers, particularly those seeking to integrate alternative data such as insider trading into machine learning frameworks.

For investors using machine learning as a stand-alone tool, our results highlight an essential gap between statistical prediction and real-world implementation. Although our estimates suggest that these machine learning strategies applied within our most liquid stock universe remain profitable under implied trading costs, real-world frictions challenge this conclusion. As discussed earlier, the portfolios rely on patterns learned from historical data, which may not persist once market conditions change or once the signal becomes crowded. Consequently, there is implementation risk in real-world trading, and investors should view machine learning signals not as directly implementable sources of abnormal returns, but rather as inputs into a broader investment due diligence process. For example, they can serve as a modern filter for identifying complex nonlinear relationships that linear models miss.

For investors relying on insider trading, our results challenge the traditional "small-firm" anomaly. Unlike the consensus view that favours small-cap stocks due to information asymmetry, our strict out-of-sample evaluation shows that insider signals are most robust among larger firms. The fact that insider variables enhance economic performance in sparse linear benchmarks, such as OLS-3, for large firms suggests they can still play a role as supporting information. However, because this finding diverges from the existing literature, we would advise against implementing it without rigorous analysis to test its robustness before capital deployment.

In practice, this points to the use of insider trading data as part of the investment research and due diligence pipeline rather than as the basis for a stand-alone trading strategy. Investors might use insider transactions to corroborate valuation signals, assess managerial confidence, or validate investment theses. However, for those seeking to capture the immediate return of insider trading, the barrier remains high. The 81-second private advantage and subsequent ten-day drift we discussed limit the profits of insider trading (Rogers et al., 2017).

Overall, both machine learning and insider trading information are best viewed as complementary inputs for corroboration and risk management, rather than as independent generators of alpha. The reason is that effective regularisation is non-trivial and materially affects model performance, reducing the ease of implementation relative to standard linear benchmarks. High-dimensional machine learning models require large amounts of high-quality data, careful tuning, and substantial computational resources. Training, validating, and maintaining such models is time-consuming and costly. Consequently, even if machine learning strategies generate abnormal returns net of transaction costs, their overall economic value must also account for computing costs, data requirements, and human capital.

7.4. Future Research

Two avenues for further research emerge from the limitations identified in our analysis. Methodological refinement and the construction of insider signals must go hand in hand.

First, our Double Machine Learning (DML) analysis reveals the difficulty of identifying robust insider signals. Although DML successfully reduced the insider set from 23 to 14 variables, improving the insider-augmented NN3 R^2_{OOS} from 0.15 (with 23 signals) to 0.22 (with 14 variables), the resulting insider model still underperforms the Baseline (0.26). This performance gap is informative. The 14 signals selected by DML are likely false positives or insufficiently robust for out-of-sample prediction. If these signals contained genuine incremental value, the insider-augmented model should, at a minimum, match the Baseline. The fact that performance deteriorates suggests that even a rigorous selection method such as DML can struggle to distinguish accurate information from noise in high-dimensional settings. Consequently, standard DML approaches may be too lenient in this context. Future research should focus on more rigorous identification strategies and, crucially, on developing more creative signals to isolate genuinely informative insider trading.

Second, the scope should be expanded beyond monthly U.S. equity strategies, as the U.S. market is widely regarded as the world's most efficient equity market. Cross-country tests could therefore assess whether insider signals offer greater incremental value in markets characterised by lower informational efficiency or different regulatory constraints.

8. Conclusion

In this thesis, we investigate whether insider trading information improves return prediction in high-dimensional machine learning models using U.S. equity data from 2006 through 2021, relying on the framework of Gu, Kelly, and Xiu (2020). We extended the Gu, Kelly, and Xiu (2020) framework, hereafter referred to as the *Baseline*, by adding two insider based information sets: an *Outsider* set containing insider trades publicly disclosed during a given month, and a *Corporate Insider* set containing all insider trades executed during that month, regardless of when they are filed. Thus, the *Corporate Insider* set captures information that is private to insiders at the time of trading. In contrast, the *Outsider* set

captures information available to the market upon disclosure. These distinctions allows us to address the seven hypotheses outlined in Section 1:

- Hypotheses H_{1a} and H_{2a} propose that adding *Outsider* or *Corporate Insider* signals improves out-of-sample return predictability relative to the Baseline. We reject both hypotheses. Across all models, neither *Outsider* nor *Corporate Insider* signals increase predictive accuracy, indicating that insider trading signals do not generalise in high-dimensional machine-learning settings.

- Hypotheses H_{1b} and H_{2b} propose that portfolios formed using Outsider or Corporate Insider information achieve higher Sharpe ratios relative to the Baseline. These hypotheses are rejected. Neither information set yields a significant improvement in the Sharpe ratio relative to the Baseline, indicating that insider trading does not add economic value in high-dimensional machine-learning settings.

- Hypothesis H_3 proposes that the Corporate Insider information set outperforms the Outsider set. We find no support for this hypothesis. Corporate Insider portfolios do not outperform Outsider portfolios, and performance even declines when moving from public to private insider information.

- Hypothesis H_4 proposes that insider-based Sharpe ratio gains are strongest among small-cap firms. This hypothesis is rejected. Only OLS-3 shows improvement, and the gains are larger for large firms than for small firms.

- Hypothesis H_5 proposes that insider-based gains remain robust after transaction costs. We partially accept this hypothesis. While no systematic gains appear when augmenting high-dimensional models with insider signals, restricting the sample to the most liquid firms shows that NN4 under the Outsider information set achieves a Sharpe ratio that is 0.22 above the Baseline, and this improvement persists after transaction costs.

Overall, our findings confirm Gu, Kelly, and Xiu (2020) in showing that high-dimensional machine learning can predict returns from publicly available firm and macroeconomic information, providing evidence that challenges the semi-strong form of the EMH. By contrast, insider trading signals, whether public or private, add no systematic incremental predictive value in high-dimensional settings. This suggests that the informational content of insider trades is largely redundant, as it has already been incorporated into the rich public information set captured by the Baseline predictor set. Although we identify isolated improvements, particularly in low-dimensional linear models for large firms and in nonlinear models among the most liquid stocks, these exceptions do not overturn the broader conclusion. Fifty years after Fama (1970), our results challenge the semi-strong form of the EMH by demonstrating that public data predicts returns, yet we find that insider information itself offers negligible incremental value within this high-dimensional setting.

References

Aboody, D. and B. Lev (2000). Information asymmetry, r&d, and insider gains. *The journal of Finance 55*(6), 2747–2766.

Akbas, F., C. Jiang, and P. D. Koch (2020). Insider investment horizon. *The Journal of Finance 75*(3), 1579–1627.

Akyildirim, E., A. Goncu, and A. Sensoy (2021). Prediction of cryptocurrency returns using machine learning. *Annals of Operations Research 297*(1), 3–36.

Alldredge, D. M. and B. Blank (2019). Do insiders cluster trades with colleagues? evidence from daily insider trading. *Journal of Financial Research 42*(2), 331–360.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets 5*(1), 31–56.

Ardia, D. and K. Boudt (2018). The peer performance ratios of hedge funds. *Journal of Banking & Finance 87*, 351–368.

Avramov, D., S. Cheng, and L. Metzker (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science 69*(5), 2587–2619.

Bansal, R., D. A. Hsieh, and S. Viswanathan (1993). A new approach to international arbitrage pricing. *The Journal of Finance 48*(5), 1719–1747.

Barber, B. M. and T. Odean (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The quarterly journal of economics 116*(1), 261–292.

Bettis, C., D. Vickrey, and D. W. Vickrey (1997). Mimickers of corporate insiders who make large-volume trades. *Financial Analysts Journal 53*(5), 57–66.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1046–1089.

Black, F., M. C. Jensen, M. Scholes, et al. (1972). The capital asset pricing model: Some empirical tests.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Campbell, J. Y. (2000). Asset pricing at the millennium. *The Journal of Finance 55*(4), 1515–1567.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies 21*(4), 1509–1531.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance 52*(1), 57–82.

Chakravorty, A. and N. Elsayed (2025). A comparative study of machine learning algorithms for stock price prediction using insider trading data. *arXiv preprint arXiv:2502.08728*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters.

Chordia, T., A. Subrahmanyam, and Q. Tong (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics 58*(1), 41–58.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics 138*(1), 291–311.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance 66*(4), 1047–1108.

Cohen, L., C. Malloy, and L. Pomorski (2012). Decoding inside information. *The Journal of Finance 67*(3), 1009–1043.

Cziraki, P. and J. Gider (2021). The dollar profits to insider trading. *Review of Finance 25*(5), 1547–1580.

Dacheng Xiu (n.d.). Dacheng xiu. https://dachxiu.chicagobooth.edu/. Accessed September 12, 2025.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam (1998). Investor psychology and security market under-and overreactions. *the Journal of Finance 53*(6), 1839–1885.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & economic statistics 20*(1), 134–144.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer.

Ding, W. (n.d.). Insider trading data and linking tables. GitHub repository. Accessed September 2025.

Dittmar, R. F. (2002). Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *The Journal of Finance 57*(1), 369–403.

Drobetz, W. and T. Otto (2021). Empirical asset pricing via machine learning: evidence from the european stock market. *Journal of Asset Management 22*(7), 507–538.

Eckbo, B. E. and D. C. Smith (1998). The conditional performance of insider trades. *The Journal of Finance 53*(2), 467–498.

Eldan, R. and O. Shamir (2016). The power of depth for feedforward neural networks. In *Conference on learning theory*, pp. 907–940. PMLR.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance 25*(2), 383–417.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *the Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

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

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance 75*(3), 1327–1370.

Fidrmuc, J., M. Goergen, and L. Renneboog (2008). Insider trading, news releases, and ownership concentration. In *Insider Trading*, pp. 309–370. CRC Press.

Frankel, R. and X. Li (2004). Characteristics of a firm's information environment and the information asymmetry between insiders and outsiders. *Journal of accounting and economics 37*(2), 229–259.

Frazzini, A., R. Israel, and T. J. Moskowitz (2018). *Trading costs*, Volume 3229719. SSRN.

Gębka, B., A. Korczak, P. Korczak, and J. Traczykowski (2017). Profitability of insider trading in europe: A performance evaluation approach. *Journal of Empirical Finance 44*, 66–90.

Giglio, S., B. Kelly, and D. Xiu (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics 14*(1), 337–368.

Glorot, X., A. Bordes, and Y. Bengio (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings.

Green, J., J. R. Hand, and X. F. Zhang (2013). The supraview of return predictive signals. *Review of Accounting Studies 18*(3), 692–730.

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

Hanauer, M. X. and T. Kalsbach (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review 55*, 101022.

Hansen, L. K. and P. Salamon (2002). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence 12*(10), 993–1001.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Hong, C. Y. and F. W. Li (2019). The information content of sudden insider silence. *Journal of Financial and Quantitative Analysis 54*(4), 1499–1538.

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. Technical report, National Bureau of Economic Research.

Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr.

James, G., D. Witten, T. Hastie, R. Tibshirani, and J. Taylor (2023). An introduction to statistical learning: Python edition.

Jeng, L. A., A. Metrick, and R. Zeckhauser (2003). Estimating the returns to insider trading: A performance-evaluation perspective. *Review of Economics and Statistics 85*(2), 453–471.

John, K. and L. H. Lang (1991). Insider trading around dividend announcements: Theory and evidence. *The Journal of Finance 46*(4), 1361–1389.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific.

Karpoff, J. M. and D. Lee (1991). Insider trading before new issue announcements. *Financial Management*, 18–26.

Ke, B., S. Huddart, and K. Petroni (2003). What insiders know about future earnings and how they use it: Evidence from insider trades. *Journal of Accounting and Economics 35*(3), 315–346.

Kelly, B., D. Xiu, et al. (2023). Financial machine learning. *Foundations and Trends® in Finance 13*(3-4), 205–363.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., V. Nair, and G. Hinton (2010). Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html 5*(4), 1.

Kumbure, M. M., C. Lohrmann, P. Luukka, and J. Porras (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications 197*, 116659.

Lakonishok, J. and I. Lee (2001). Are insider trades informative? *The Review of Financial Studies 14*(1), 79–111.

Ledoit, O. and M. Wolf (2008). Robust performance hypothesis testing with the sharpe ratio. *Journal of Empirical Finance 15*(5), 850–859.

Lee, D. S., W. H. Mikkelson, and M. M. Partch (1992). Managers' trading around stock repurchases. *The Journal of Finance 47*(5), 1947–1961.

Leippold, M., Q. Wang, and W. Zhou (2022). Machine learning in the chinese stock market. *Journal of Financial Economics 145*(2), 64–82.

Leitch, G. and J. E. Tanner (1991). Economic forecast evaluation: profits versus the conventional error measures. *The American Economic Review*, 580–590.

Lettau, M. and S. Ludvigson (2001). Consumption, aggregate wealth, and expected stock returns. *the Journal of Finance 56*(3), 815–849.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The journal of finance 20*(4), 587–615.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of economic perspectives 17*(1), 59–82.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

Martin, I. W. and S. Nagel (2022). Market efficiency in the age of big data. *Journal of financial economics 145*(1), 154–177.

Masters, T. (1993). *Practical neural network recipes in C++*. Morgan Kaufmann.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Meulbroek, L. K. (1992). An empirical analysis of illegal insider trading. *The Journal of Finance 47*(5), 1661–1699.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, 768–783.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Piotroski, J. D. and D. T. Roulstone (2005). Do insider trades reflect both contrarian beliefs and superior knowledge about future cash flow realizations? *Journal of Accounting and Economics 39*(1), 55–81.

Ravina, E. and P. Sapienza (2010). What do independent directors know? evidence from their trading. *The Review of Financial Studies 23*(3), 962–1003.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: journal of the Econometric Society*, 931–954.

Rogers, J. L., D. J. Skinner, and S. L. Zechman (2017). Run edgar run: Sec dissemination in a high-frequency world. *Journal of Accounting Research 55*(2), 459–505.

Safer, A. M. (2002). The application of neural networks to predict abnormal stock returns using insider trading data. *Applied Stochastic Models in Business and Industry 18*(4), 381–389.

Seyhun, H. N. (1986). Insiders' profits, costs of trading, and market efficiency. *Journal of financial Economics 16*(2), 189–212.

Seyhun, H. N. (1988). The information content of aggregate insider trading. *Journal of Business*, 1–24.

Seyhun, H. N. (1990). Do bidder managers knowingly pay too much for target firms? *Journal of Business*, 439–464.

Seyhun, H. N. (1992). Why does aggregate insider trading predict future stock returns? *The Quarterly Journal of Economics 107*(4), 1303–1331.

Seyhun, H. N. (2000). *Investment intelligence from insider trading*. MIT press.

Seyhun, H. N. and M. Bradley (1997). Corporate bankruptcy and insider trading. *The Journal of Business 70*(2), 189–216.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance 19*(3), 425–442.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of economic perspectives 17*(1), 83–104.

Stefan Voigt (n.d.a). Equity market and trading data. Accessed September 15, 2025.

Stefan Voigt (n.d.b). Replicating gu, kelly, and xiu (2020): Empirical asset pricing via machine learning. https://www.tidy-finance.org/blog/gu-kelly-xiu-replication/. Accessed September 12, 2025.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies 21*(4), 1455–1508.

White, H. et al. (1988). Economic prediction using neural networks: The case of ibm daily stock returns. In *ICNN*, Volume 2, pp. 451–458.

Wu, W. (2019). Information asymmetry and insider trading. *Fama-Miller Working Paper, Chicago Booth Research Paper* (13-67).

A. Data

A.1. Overview of Characteristics

Table 14: Overview of Baseline Characteristics and Insider Trading Extension

No.	Acronym	Description
Baseline Predictor Set (94 characteristics from Gu et al. (2020))		
1	absacc	Absolute accruals
2	acc	Working capital accruals
3	aeavol	Abnormal earnings announcement volume
4	age	Years since first Compustat coverage
5	agr	Asset growth
6	baspread	Bid–ask spread
7	beta	Beta
8	betasq	Beta squared
9	bm	Book-to-market ratio
10	bm_ia	Industry-adjusted book-to-market
11	cash	Cash holdings
12	cashdebt	Cash flow to debt
13	cashpr	Cash productivity
14	cfp	Cash flow to price ratio
15	cfp_ia	Industry-adjusted cash flow to price
16	chatoia	Industry-adjusted change in asset turnover
17	chcsho	Change in shares outstanding
18	chempia	Industry-adjusted change in employees
19	chinv	Change in inventory
20	chmom	Change in 6-month momentum
21	chpmia	Industry-adjusted change in profit margin
22	chtx	Change in tax expense
23	cinvest	Corporate investment
24	convind	Convertible debt indicator
25	currat	Current ratio
26	depr	Depreciation to PP&E
27	divi	Dividend initiation
28	divo	Dividend omission
29	dolvol	Dollar trading volume
30	dy	Dividend-to-price ratio
31	ear	Earnings announcement return
32	egr	Equity growth
33	ep	Earnings-to-price ratio
34	gma	Gross profitability
35	grCAPX	Growth in capital expenditures
36	grltnoa	Growth in long-term net operating assets
37	herf	Industry sales concentration
38	hire	Employment growth
39	idiovol	Idiosyncratic return volatility
40	ill	Illiquidity
41	indmom	Industry momentum
42	invest	Capex + inventory investment
43	lev	Leverage
44	lgr	Long-term debt growth
45	maxret	Maximum daily return

Continued on next page

No.	Acronym	Description
46	mom12m	12-month momentum
47	mom1m	1-month momentum
48	mom36m	36-month momentum
49	mom6m	6-month momentum
50	ms	Financial statement score
51	mvel1	Market equity (size)
52	mve_ia	Industry-adjusted size
53	nincr	Number of earnings increases
54	operprof	Operating profitability
55	orgcap	Organizational capital
56	pchcapx_ia	Industry-adjusted change in capex
57	pchcurrat	Change in current ratio
58	pchdepr	Change in depreciation
59	pchgm_pchsale	Change in gross margin – Change in sales
60	pchquick	Change in quick ratio
61	pchsale_pchinvt	Change in sales – Change in inventory
62	pchsale_pchrect	Change in sales – Change in accounts receivable
63	pchsale_pchxsga	Change in sales – Change in SG&A
64	pchsaleinv	Change in sales-to-inventory
65	pctacc	Percent accruals
66	pricedelay	Price delay measure
67	ps	Financial statement score (alternative)
68	quick	Quick ratio
69	rd	R&D increase
70	rd_mve	R&D to market cap
71	rd_sale	R&D to sales
72	realestate	Real estate holdings
73	retvol	Return volatility
74	roaq	Return on assets
75	roavol	Earnings volatility
76	roeq	Return on equity
77	roic	Return on invested capital
78	rsup	Revenue surprise
79	salecash	Sales to cash
80	saleinv	Sales to inventory
81	salerec	Sales to receivables
82	secured	Secured debt
83	securedind	Secured debt indicator
84	sgr	Sales growth
85	sin	Sin-stock indicator
86	sp	Sales-to-price ratio
87	std_dolvol	Volatility of dollar volume
88	std_turn	Volatility of share turnover
89	stdacc	Accrual volatility
90	stdcf	Cash flow volatility
91	tang	Tangibility / debt capacity
92	tb	Tax income to book income
93	turn	Share turnover
94	zerotrade	Zero trading days
Insider Trading Extension (23 additional variables)		
95	npr_volume	Net purchase ratio (dollar volume, 6-month window)
96	npr_count	Net purchase ratio (trade count, 6-month window)

Continued on next page

No.	Acronym	Description
97	opp_buy	Opportunistic insider buy
98	opp_sell	Opportunistic insider sell
99	rtn_buy	Routine insider buy
100	rtn_sell	Routine insider sell
101	net_cluster	Net insider clusters
102	ppn	Post–routine-buy silence signal
103	ssn	Post–routine-sell silence signal
104	purchase_ceo	CEO purchase dummy
105	purchase_cfo	CFO purchase dummy
106	purchase_coo	COO purchase dummy
107	purchase_director	Director purchase dummy
108	purchase_vp	Vice president purchase dummy
109	purchase_10pct	10% owner purchase dummy
110	purchase_other	Other officer purchase dummy
111	sell_ceo	CEO sale dummy
112	sell_cfo	CFO sale dummy
113	sell_coo	COO sale dummy
114	sell_director	Director sale dummy
115	sell_vp	Vice president sale dummy
116	sell_10pct	10% owner sale dummy
117	sell_other	Other officer sale dummy

B. Machine Learning

B.1. Hyperparameters

Table 15: Specifications of Hyperparameters

Model	**Hyperparameter**	**Specification**
OLS-3	n.a.	n.a.
PLS	K	$\{1, 2, \ldots, 24\}$
PCR	K	$\{1, 2, \ldots, 128\}$
Elastic Net	λ	$(10^{-4}, 10^{-1})$
	ρ	$(0, 1)$
RF	L	$\{1, 2, 3\}$
	B	300
	mtry	$\{3, 5, 10, 20, 30, 50\}$
XGBoost	L	$\{1, 2\}$
	B	$\{32, \ldots, 320\}$
	η	$\{0.1\}$
Neural Network	batch size	$10,000$
	epochs	100
	patience	5
	L_1 penalty λ_1	$(10^{-5}, 10^{-3})$
	learning rate	$\{0.001, 0.01\}$
	ensemble	10

Note: This table details the hyperparameter specifications for our models. The selection is informed by the grid search procedure established by Gu et al. (2020), ensuring our models remain consistent.