# Can a Machine Learn from Behavioral Biases?

# Evidence from Stock Return Predictability of Deep Learning Models

Suk-Joon Byun[†] Sangheum Cho[‡] Da-Hea Kim[§]

November 14, 2021

**Abstract**

We examine how the return predictability of deep learning models varies with stocks' vulnerability to investors' behavioral biases. Using an extensive list of anomaly variables, we find that the long-short strategy based on deep learning signals generates greater returns for stocks more vulnerable to behavioral biases: stocks that are small, young, illiquid, unprofitable, volatile, non-dividend-paying, close to default or extreme growth, far from the 52-week high, and lottery-like. Such performance of deep learning models becomes more pronounced for stocks held by less sophisticated investors. These results suggest that deep learning models accommodating time-varying nonlinear factor exposures are useful in capturing mispricing induced by behavioral biases.

**JEL classification:** To be included

**Keywords:** Deep Learning, Behavioral Biases, Empirical Asset Pricing

[†] Korea Advanced Institute of Science and Technology (e-mail: sjbyun99@kaist.ac.kr)
[‡] Korea Advanced Institute of Science and Technology (e-mail: sh.cho@kaist.ac.kr)
[§] Sungkyunkwan University Business School (e-mail: daheakim@skku.edu)

# 1. Introduction

In the face of the "zoo" of asset pricing anomalies as noted by Cochrane (2011) and Harvey, Liu, and Zhu (2016), a growing body of work applies innovative machine-learning techniques in an effort to identify a better-performing asset pricing model with a modest number of factors. In particular, recent studies by Chen, Pelger, and Zhu (2020) and Gu, Kelly, and Xiu (2020, 2021) show that deep neural network algorithms allowing for dynamic nonlinear factor exposures outperform competing linear factor models. Given that much of the anomaly literature points to investors' behavioral biases as being responsible for the documented anomalous returns (Zhang 2006; Stambaugh, Yu, and Yuan 2012; Birru 2018; Daniel, Hirshleifer, and Sun 2020), how well such machine-learning-based asset pricing models capture mispricing caused by behavioral biases needs to be explored.

To answer this, we employ the conditional autoencoder (CA) deep learning model suggested by Gu et al. (2021) and investigate how its return predictability varies with stocks' vulnerability to investors' behavioral biases. Among various machine-learning-based models, we focus on the CA model for three main reasons. First of all, Gu et al. (2021) show that its pricing performance dominates that of other popular asset pricing models such as Fama-French models, principal components analysis (PCA), and instrumented PCA (IPCA) of Kelly, Pruitt, and Su (2019). Second, since the CA model can be thought of as a nonlinear neural network counterpart to IPCA, we can easily gauge the impact of nonlinearity on asset pricing performance by comparing the performance between the CA and IPCA models. Third, unlike the deep learning model of Chen et al. (2020) that is designed to estimate the stochastic discount factor, the CA model generates the stock-level signals, which can be directly used for cross-sectional analysis of stock returns.

To identify stocks most vulnerable to behavioral biases, we use the extensive list of stock characteristics that proxy for speculativeness compiled by Birru (2018). Birru (2018) defines speculative stocks as stocks that are hard to value properly and/or stocks that have the greatest impediments to arbitrage, and characterizes them as small, young, unprofitable, volatile, non-dividend-

paying, potentially close to distress or extreme growth, lottery-like, illiquid, and with high beta. Many previous studies as well as Birru (2018) suggest that such speculative stocks are most subject to behavioral biases, thereby leading to mispricing (Baker and Wurgler 2006; Zhang 2006; Jiang, Lee, and Zhang 2005; Cohen and Lou 2012; Seybert and Yang 2012; Hirshleifer, Hsu, and Li 2018; Lee, Sun, Wang, and Zhang 2019).

To obtain the CA deep learning signals, we train the model using US equity data with 92 firm characteristics by Green, Hand, and Zhang (2017) with the assumption of 6 latent factors and no pure idiosyncratic mispricing term. The model produces out-of-sample factor loadings $\beta_{i,t}(z_{i,t-1})$ for each firm-month observations using the lagged 92 firm characteristics $z_{i,t-1}$ and out-of-sample expected returns of 6 latent factors $\lambda_t$ using associated 92 characteristics-managed portfolios.[1] Thus, the final deep learning signals are expected returns $\beta_{i,t}(z_{i,t-1})'\lambda_t$ obtained by the factors loadings and the latent factors. We construct four types of the CA signals—CA0, CA1, CA2, and CA3—by using zero, one, two, and three layers of nonlinear activation functions, respectively. We consider the CA0 signal as a linear signal benchmark. We also estimate the IPCA model of Kelly et al. (2019) to obtain an alternative linear signal benchmark.

To examine the return predictability of the CA and IPCA models, we form portfolios by sorting stocks based on various CA and IPCA signals into deciles and calculate portfolio returns by value-weighting the returns of constituent stocks in each portfolio. We calculate the returns to the long-short strategy based on each signal by buying the stocks in the highest decile of the signal and shorting the stocks in the lowest. We find that the returns to the long-short strategies based on the nonlinear CA signals are substantially higher than those based on the CA0 signal or IPCA signal, thereby confirming that nonlinear deep learning models outperform the linear models for stock return predictability.

Most importantly, we find that such superior performance of nonlinear deep learning models

[1] Kelly et al. (2019) and Gu et al. (2021) introduce characteristics-managed portfolio perspectives for the latent factors.

over the linear models is mainly driven by stocks more vulnerable to behavioral biases. For illustration, for non-speculative stocks, the long-short strategy based on the CA3 signal outperforms that based on the IPCA signal only by 0.36% per month on average, which is statistically insignificant. By contrast, for speculative stocks, the long-short strategy based on the CA3 signal outperforms that based on the IPCA signal by 1.18% per month, highly statistically significant. A difference of 0.83% is both economically and statistically significant. When we use the CA0 signal as a linear signal benchmark, we find even more dramatic results. We also find that such performance of deep learning models driven by speculative stocks becomes more pronounced for stocks held by less institutional investors. These results hold true regardless of risk adjustment methods. Thus, these findings suggest that deep learning models allowing for time-varying nonlinear factor exposures are useful in capturing mispricing induced by behavioral biases.

This paper contributes to the emerging literature on empirical asset pricing using deep learning methods. Recently, there have been many studies on deep learning methods in various contexts (Chen et al. 2020; Cong, Tang, Wang, and Zhang 2020; Gu et al. 2020, 2021). Although those models show good performance in predicting asset returns, it is not easy to know the source of such good performance because, in general, the nonlinearity of neural networks makes the in-sample parameter inference difficult (Mullainathan and Spiess 2017). Karolyi and Van Nieuwerburgh (2020) point out that understanding the economic meaning of patterns that neural networks extract is also very important for a more credible usage of deep learning models. In this vein, Avramov, Cheng, and Metzker (Forthcoming) show that deep learning models are particularly good at predicting returns for firms with high trading frictions proxied by firm size and credit ratings. The idea is adopted in several other studies (Chen et al. 2020; Nagel 2021). Our study extends the literature by suggesting deep neural networks capture individual stock returns' covariances to the systematic behavioral factors, particularly instrumenting the covariances using firm characteristics related to speculativeness.

This paper is organized as follows. Section 2 describes the construction of our machine learning signals. Section 3 provides our main empirical results that the interaction of machine learning

signals with proxies of speculation. Finally, section 4 concludes our paper.

## 2. Performance of Nonlinear Machine Learning Signals

### 2.1. Construction of Machine Learning Signals

We construct the machine learning signals using the Instrumented Principal Component Analysis (IPCA) model of Kelly et al. (2019) and the Conditional Autoencoder (CA) model of Gu et al. (2021). In the training of the IPCA and CA models, we use the NYSE, AMEX, and NASDAQ stocks that are ordinary common shares incorporated in the US. Our full sample begins in July 1972 and ends in December 2020, covering around 50 years. The stock return data is obtained from the Center for Research in Security Prices (CRSP) monthly stock file. Financial statement items are obtained from the COMPUSTAT. We drop stocks with missing market capitalization, the book value of assets, quarterly earnings announcement dates, and prior twelve-month returns. Stocks with SIC code classified as the financial sector and utility sector are also dropped.

We use 92 firm characteristics to train our machine learning models. To mitigate the effects of outliers, we rank-transform the firm characteristics into unit intervals by following Gu et al. (2021). The firm characteristics are largely adopted from Green et al. (2017).[2] However, we drop some variables having high correlation with each other because those variables can be problematic in estimating the IPCA model.[3] Instead, we augment the list with several anomaly variables, including nearness to 52-week high, previous month stock price, and many others. As a result, we have 92 rank-transformed firm characteristics. The updating frequencies of the firm-level characteristics are annual (e.g., firm age,

[2] The set of firm characteristics is widely used in the machine learning literature (Green et al. 2017; Gu et al. 2020, 2021; Avramov et al. Forthcoming). Details are listed on the Appendix of Green et al. (2017), and the Internet Appendix of Gu et al. (2020).

[3] For example, percentage change in sales-to-inventory has 99% of correlations with percentage change in sales minus percentage change in inventory when they are rank-transformed. They provide redundant information and can incur noise in the estimation of IPCA model or training of CA model.

gross profitability), quarterly (e.g., returns on equity, revenue surprise), or monthly (e.g., beta, stock return volatility).

The initial training sample is 8 years (1972-1980), the validation sample is 6 years (1981-1986), and the out-of-sample test sample is the remaining 34 years (1987-2020). The machine learning models are re-trained for every year in the test sample by increasing the size of the training sample by one year while maintaining the size of the validation sample as 6 years by rolling it forward by following Gu et al. (2021).

Table 1 describes machine learning signals generated by the IPCA model and CA model. Panel A reports the baseline machine learning signals. The IPCA signal is generated using the IPCA model. The IPCA signal is a linear signal since the IPCA model has no nonlinear structures. The CA0 signal is also a linear signal because it is generated by the CA model having no nonlinear structures. On the other hand, we generate the CA1, CA2, and CA3 signals which are nonlinear signals. The CA1 (CA2, CA3) signals are generated by the CA model having one (two, three) layers of nonlinear layers.

A crucial part of our study is to compare the performance of the nonlinear signals and the linear signals in portfolio sorting. Therefore, we measure the differences in performance of the portfolio returns constructed by the nonlinear and linear machine learning signals. Panel B describes the measures of differences. NML_IPCA is the difference in portfolio returns of the nonlinear CA3 and the linear IPCA signal. Similarly, NML_CA0 is the difference between the CA3 and CA0 signals.

## 2.2. Returns of Portfolios Sorted by Machine Learning Signals

Table 2 reports the returns of portfolios sorted by machine learning signals. There are linear and nonlinear machine learning signals. The linear signals are the IPCA and CA0 signals, and the nonlinear signals are the CA1, CA2, and CA3 signals. For each month, stocks are sorted into a decile by the NYSE breakpoints of each machine learning signal to form a value-weighted portfolio. The excess

returns of each decile are reported. We also report returns of the long-short strategy of buying the highest decile (D10) and shorting the lowest decile (D1) as D10-D1.

The returns of the long-short strategy by the linear IPCA signal generate 0.82% (t=3.52) per month. The returns of the long-short strategy by the linear CA0 signal also generate a similar magnitude of returns as 0.84% (t=3.55) per month. The portfolio returns constructed by the nonlinear signals are higher than the linear signals. In the case of the nonlinear CA1 signal, the long-short strategy generates 1.09% (t=4.11) per month. For the nonlinear signals having more nonlinearity, the CA2 and CA3 signals, the long-short strategy by the CA2 signal generates 1.44% (t=5.48) per month, and the CA3 signal generates 1.40% (t=5.34).

### 2.3. The Differences of Returns of Portfolios Sorted by the Nonlinear and Linear Machine Learning Signals

Table 2 also reports the differences in performance of portfolios sorted by the nonlinear and linear machine learning strategies. Since there are two linear signals, the IPCA and CA0 signals, we report the differences of portfolio returns between the nonlinear signals, the CA1, CA2, and CA3 signals, with each of the linear signals. We denote the differences in portfolio returns sorted by the nonlinear signal and the linear signal as the nonlinear-minus-linear (NML). Especially, the difference between the portfolio returns by the nonlinear CA3 signal and the linear IPCA signal is denoted by the NML_IPCA. Similarly, NML_CA0 represents the differences in returns of the CA3 signal and the CA0 signal.

The difference between the long-short strategy by the nonlinear CA1 signal and the linear IPCA signal is 0.28% per month but statistically insignificant. However, as the nonlinear signals have more nonlinearity than the linear signal, the differences between the long-short strategies become significant. The difference of long-short strategy between CA2 and IPCA is 0.63% (t=3.68) per month, and it is statistically significant and economically sizable. In the case of NML_IPCA, the difference between the CA3 and IPCA signal, the difference between the long-short strategy is 0.58% (t=3.35)

which is also statistically significant and economically meaningful. When we measure the differences between the portfolio returns using the CA0 signal as the linear signal, similar patterns emerge. The NML_CA0 signal generates 0.56% (t=3.40) per month.

## 2.4. Risk-Adjusted Returns of the Differences in the Portfolios by the Nonlinear and Linear Signals

Table 3 reports the risk-adjusted returns of the differences in the decile portfolios constructed by the nonlinear and linear signals using various factor models. For brevity, we report the risk-adjusted returns of the lowest decile (D1), the highest decile (D10), and the long-short strategy of buying the highest decile and the shorting the lowest decile denoted by D10-D1.

When we use the IPCA as the linear signal, the difference in the CAPM risk-adjusted return of the long-short strategy by the CA1 and IPCA signals shows 0.30% per month, which is statistically insignificant. However, in the case of the nonlinear signal as CA2, the difference in the CAPM risk-adjusted returns of the long-short strategy by the CA2 and IPCA signals is 0.68% (t=3.62), which is statistically and economically significant. In the case of NML_IPCA, the difference in the CAPM risk-adjusted returns of the long-short strategy is 0.57% (t=3.04), which is also statistically and economically significant. When the CA0 signal is used as the linear signal, the CAPM alpha of the long-short strategy is also sizeable. NML_CA0 is 0.62% (t=3.50) per month.

Similar patterns emerge when we use different factor models. For example, when returns are adjusted by the Fama-French five-factor model augmented by Carhart's momentum factor by Fama and French (2018) (FF6), the FF6 alpha of the long-short strategy of NML_IPCA generates 0.66% (t=3.30), and NML_CA0 generates 0.56% (t=2.82). Using other factor models also produces similar results.[4]

[4] The factor models include the Fama-French three-factor model by Fama and French (1993) (FF3), the Fama-French five-factor model by Fama and French (2015) (FF5), the long- and short-horizon behavioral factor model by Daniel et al. (2020) (DHS), the q-factor model by Hou, Xue, and Zhang (2015), and the Fama-French three-

NML_IPCA and NML_CA0 consistently produce significant risk-adjusted returns ranging from 0.35% to 0.67%.

## 2.5. Fama-Macbeth Regression of Machine Learning Signals

The Fama-Macbeth regression further supports the portfolio sorting results. Table 4 reports the Fama-Macbeth regression results. The dependent variable is the one-month ahead returns (%). The main independent variables are baseline machine learning signals: the IPCA (or CA0) signal and the CA3 signal.[5] The stock-level nonlinear-minus-linear by the CA3 and IPCA signal (NML_IPCA) is constructed by taking the difference between the nonlinear CA3 signal and the linear IPCA signal. The NML_CA0 is the difference between the CA3 and CA0 signals.

Panel A reports the results using the IPCA as the linear signal. Column (1) reports the result using the IPCA signal as the main independent variable. The IPCA signal predicts returns, and the predictability is statistically significant. In terms of economic significance, as a stock's IPCA rank increases from the 25th percentile to the 75th percentile, then the realized return increases by 1.45% (2.908*0.5=1.454). Column (2) reports the results using the CA3 signal. The CA3 signal also has statistically significant return predictability. In terms of economic significance, as a stock's CA3 rank increases from the 25th percentile to the 75th percentile, the realized return increases by 1.77% (3.534*0.5=1.767). Both the linear and nonlinear signals have significant return predictability. It is consistent with the portfolio sorting results that each of the linear and nonlinear signals produces sizable excess returns and alphas.

Column (3) reports the return predictability of the NML_IPCA signal that is the predictive signal for future returns mainly attributable to the nonlinear structure of neural networks. NML_IPCA

factor model augmented by the liquidity factor of Pástor and Stambaugh (2003) (PS).

[5] The CA3 and IPCA, CA0 signals are rank-transformed into a unit interval to mitigate the outliers' effect.

also has statistically significant return predictability. In terms of economic significance, as a stock's NML_IPCA rank increases from the 25th percentile to the 75th percentile, the realized return increases by 0.86% (1.711*0.5=0.8555). This result echoes the portfolio sorting results that the differences in portfolio returns constructed by the CA3 and IPCA signals are sizable.

Column (4) reports the result using both the IPCA signal and the NML_IPCA signal. When the linear IPCA signal is controlled, the NML_IPCA signal generates higher and more significant return predictability. In terms of economic significance, as a stock's NML_IPCA rank increases from the 25th percentile to the 75th percentile, the realized return increases by 1.74% (3.488*0.5=1.744). This result indicates that the nonlinear NML_IPCA signal has distinct return predictability compared to the linear IPCA signal.

The return predictability of machine learning signals remains similar after controlling various stock characteristics. Columns (5) to (8) include stock-month level control variables including market beta, firm size, book-to-market ratio, momentum, asset growth, return on equity, illiquidity, bid-ask spread, and short-term reversal, which are also rank-transformed. Column (8) indicates that when those variables are controlled, the NML_IPCA signal even generates higher returns predictability.

Panel B reports the results using the CA0 signal as the linear signal to construct the NML_CA0 signal. The return predictability of the NML_CA0 signal is also statistically significant and economically meaningful. Column (8) shows that when a stock's NML_CA0 rank increases from the 25th percentile to the 75th percentile, the 2.04% (4.088*0.5=2.044) after controlling various firm characteristics.

The Fama-Macbeth regression results support the portfolio sorting results. The linear and nonlinear machine learning signal produces significant return predictability. However, the nonlinear signal has superior return predictability than the linear signal. The portfolios sorted by our nonlinear CA3 signal produce higher risk-adjusted value-weighted returns than the portfolios sorted by the linear signals. Moreover, the return predictability of the nonlinear signal has separate return predictability than

the linear signal after controlling various firm characteristics.

# 3. Performance of Nonlinear Machine Learning Signals and Speculative Characteristics

In this paper, our purpose is to study whether machine learning models show superior performance in capturing anomalous return behaviors by behavioral biases. Birru (2018) suggests 20 firm characteristics proxying stocks' vulnerability to behavioral biases and speculative demands. He argues that stocks that are volatile, young, unprofitable, or small are highly subjective or difficult to value, so they are prone to speculation.

## 3.1. Speculative Anomaly Variables

Table 5 describes those speculative anomaly variables. In our empirical analysis, we divide stocks into quintiles by their speculative characteristics. For example, stocks in the highest quintile (Q5) of high return volatility (SIGMA) are prone to speculative demand than the stocks in the lowest volatility quintile (Q1). Therefore, for the SIGMA, the speculative leg is Q5. The speculative anomaly variables have one speculative leg. Speculative anomaly variables from stock return volatility (SIGMA) to analyst dispersion (DISP), the speculative leg is Q5. In contrast, from firm size (SIZE) to operating profitability (OP), the speculative leg is Q1. In the case of dummy variables including earnings (E), cash flow (CF), and dividend (D), the speculative legs are negative earnings, negative cash flow, and non-payers of dividends, respectively.

We further construct a stock-month-level composite signal of speculativeness, SPEC, from the 20 speculative anomaly variables by following the methodology of Maskara and Mullineaux (2011). The SPEC signal is the average of rank-transformed 20 speculative anomaly variables aligned in the

same direction to have higher values to indicate more speculative characteristics. Therefore, for the SPEC signal, the highest quintile (Q5) is the speculative leg.

## 3.2. Portfolio Sorting

### 3.2.1. Performance of Nonlinear Machine Learning Signals in Speculative Stocks

Table 6 reports the differences of risk-adjusted returns of long-short strategy constructed by nonlinear and linear machine learning signals in quintiles of speculative anomaly variables. The NML_IPCA (NML_CA0) signal indicates the differences of risk-adjusted returns of long-short portfolios constructed by the nonlinear CA3 signal and the linear IPCA (CA0) signal in each speculative anomaly quintile.

The long-short portfolio returns in each of the speculative anomaly quintiles are constructed by following steps. First, stocks are sorted into quintiles by the NYSE breakpoints of each of the speculative anomaly variables. Second, within each quintile, stocks are further sorted into decile portfolios by the NYSE breakpoints of each machine learning signal. Third, a long-short strategy is implemented by buying the highest decile and shorting the lowest decile in each of the speculative anomaly quintiles. The portfolio returns are value-weighted and adjusted by the Fama-French five-factor model augmented by Carhart's momentum factor (Fama and French 2018).

When using SIGMA as the speculative anomaly variable, NML_IPCA shows 0.07% (t=0.35) per month in Q1, the lowest volatility quintile. It means that the long-short strategy by the nonlinear CA3 signal and linear IPCA signal is not statistically different in the non-volatile stocks. However, NML_IPCA shows 1.57% (t=3.70) per month in the Q5, the highest volatility quintile. The result indicates that the long-short portfolio returns by the nonlinear CA3 signal are about 1.5% higher than the long-short portfolio returns by the linear IPCA signal in volatile stocks. The difference of NML_IPCA in Q5 and Q1 is 1.51%, and it is highly statistically significant that t-statistics is around 3.

The results of NML_CA0 also show similar patterns. In the lowest quintile of volatility, NML_CA0 only shows 0.03% per month with t-statistics near zero. However, in the highest quintile of volatility, NML_CA0 shows substantial returns of 1.90% (t=5.52) per month. The outperformance of NML_CA0 in the volatile universe than the non-volatile universe (Q5-Q1) is 1.87% (t=4.91) per month.

By using different measures of speculative anomaly also produces similar results. None of the NML_IPCA or NML_CA0 is significant in the non-speculative leg (Q1) when the speculative characteristics are measured by IVOL, BETA, MAX, BIDASK, ILLIQ, O-Score, FP, CFV, NXF, and DISP. However, all of NML_IPCA and NML_CA0 are positive and significant in the speculative leg (Q5). When the speculative characteristics are measured by SIZE, PRICE, AGE, NH, ROA, OP, E, CF, and D, NML_IPCA and NML_CA0 exhibit positive and significant returns in the speculative leg (Q1), while mostly insignificant in the non-speculative leg (Q5). The result of using the composite signal, SPEC, summarizes the results. NML_IPCA shows only 0.36% in the non-speculative leg (Q1). However, it shows 1.18% per month, which is statistically significant, in the speculative leg (Q5). Their difference (Q5-Q1) is 0.83% and is also statistically significant. NML_CA0 also shows a clear pattern that NML_CA0 is only 0.07% in the non-speculative leg (Q1) but shows a substantial return of 1.85% per month in the speculative leg (Q5). Their difference (Q5-Q1) is 1.78% and statistically significant.

### 3.2.2. The Effect of Shareholder Sophistication on the Performance of Nonlinear Machine Learning Signals in Speculative Stocks

Less sophisticated investors such as retail investors are more prone to speculation, for example, reliance on lottery-like properties (Kumar 2009). We measure the stock-month-level shareholder sophistication by the number of institutional investors (Lam and Wei 2011). Since the nonlinear machine learning models outperform the linear models in stocks more prone to speculation, the magnitude of outperformance would be more pronounced in firms with less-sophisticated shareholders.

Table 7 reports the results of the effect of shareholder sophistication on the performance of

nonlinear machine learnings to exploit mispricing signals in speculative stocks. In each month, stocks are divided into two groups using the NYSE median value of the number of institutional investors. In each of the low and high groups, stocks are double sorted using the speculative anomaly variables and machine learning signals: stocks are into quintiles using one of the speculative anomaly variables, and then further sorted into deciles using the nonlinear CA3 signal and the linear IPCA (or CA0) signals. The long-short strategy is implemented using the decile portfolios by buying the highest decile and shorting the lowest decile. We report the differences between the risk-adjusted returns of the long-short strategy by the CA3 and IPCA signal.

In the group of the low number of institutional investors, when the speculative characteristics are measured by the stock return volatility (SIGMA), NML_IPCA shows 0.32% (t=2.19) in the lowest quintile (Q1). However, in the highest quintile (Q5) of stock return volatility, NML_IPCA shows 2.06% (t=5.14), which is substantially higher than the long-short returns of Q1. The difference between the Q5 and Q1 (Q5-Q1) is 1.74% per month, and it is statistically significant. The results of NML_CA0 also exhibit similar patterns. NML_CA0 is around 2.5% in the highest quintile (Q5), and it is significantly higher than that of Q1 by 2.19% per month. On the other hand, in the group of a high number of institutional investors, none of NML_IPCA is significant regardless of stock return volatility. Although NML_CA0 shows 0.76% per month in the highest quintile of volatility (Q5), it is way smaller than the risk-adjusted returns in the group with the low number of institutional investors that is 2.50%.

Similar patterns emerge using other measures of speculative characteristics. NML_IPCA and NLM_CA0 mostly show significant risk-adjusted returns in the speculative leg (Q5) in the group of stocks with a low number of institutional investors. Moreover, the differences between the speculative leg and the non-speculative leg (Q5-Q1) are mostly significant. However, almost none of NML_IPCA and NML_CA0 shows significant returns in the speculative leg (Q5) in the group of stocks with a high number of institutional investors. The results of using the composite signal of speculativeness, SPEC, summarize the patterns. In the group of stocks with a low number of institutional investors, NML_IPCA shows a statistically significant 2.17% per month in the speculative leg (Q5). In comparison,

NML_IPCA only shows -0.03% in the non-speculative leg (Q1). Their difference (Q5-Q1) is 2.20% per month and highly significant. On the other hand, in the group of stocks with a high number of institutional investors, NML_IPCA is insignificant in both the speculative leg and non-speculative leg. The triple-sorting results using the proxy of shareholder sophistication further support that nonlinear machine learning signals are useful in exploiting stocks that are prone to speculative demand.

### 3.3. Fama-Macbeth Regression of Machine Learning Signals Interacted with Speculative Characteristics

Table 8 reports the Fama-Macbeth regression results of examining the return predictability of the nonlinear machine learning signals in speculative stocks. The dependent variable is the one-month ahead individual stock returns (%). There are several independent variables. SPEC_MID is a dummy variable equal to 1 when the composite signal of speculative characteristics, SPEC, is on the 20th percentile to 80th percentile of NYSE breakpoints of SPEC. SPEC_HIGH is a dummy variable equal to 1 when SPEC is above the 80th percentile. The main independent variables are the interaction of nonlinear machine learning signals, NML_IPCA (or NML_CA0), with the dummy variables of speculative characteristics, SPEC_MID and SPEC_HIGH. The machine learning signals are rank-transformed to mitigate the effects of outliers.

Panel A uses the IPCA signal to construct the differences between the nonlinear and linear machine learning signal, NML_IPCA. Column (1) reports the baseline results. The IPCA signal and NML_IPCA signal show substantial return predictability. Column (2) reports that both the IPCA signal and the NML_IPCA signal have higher return predictability in stocks with more speculative characteristics. In the base group, the below 20th percentile of SPEC, the return predictability of the IPCA signal is positive and statistically significant. In terms of economic significance, the increase of stock's rank of the NML_IPCA signal from the 25th percentile to the 75th percentile predicts 0.70% higher returns (1.402*0.5=0.701). In the intermedium range of SPEC, the coefficients on the IPCA

signal is not significant. However, in the SPEC_HIGH group, the increase of rank from the 25th percentile to the 75th percentile predicts 1.16% higher returns (2.323*0.5=1.1615), and the additional predictability is highly statistically significant. In other words, the return predictability of the nonlinear signal is stronger in firms exhibiting more speculative characteristics. Columns (3) and (4) control for various firm characteristics. The results remain the same.

Panel B uses the CA0 signal in the construction of the NML_CA0 signal. The results indicate even stronger return predictability of the NML_CA0 signal in firms with higher speculative characteristics. Column (2) indicates that the increase of stocks' rank of NML_CA0 from the 25th percentile to the 75th percentile predicts 1.90% higher returns (3.794*0.5=1.987) in the stocks with the highest quintile of speculative characteristics (SPEC_HIGH). Column (4) shows that the significance of interaction terms between the NML_CA0 signal and SPEC_HIGH is robust after controlling various firm characteristics. The results support the double-sorting results that the superior performance of nonlinear machine learning signals is generated in stocks with speculative characteristics.

## 4. Conclusion

In this paper, we study the source of the superior performance of machine learning models. We find that the nonlinear machine learning models are superior in capturing anomalous returns in speculative stocks than the linear models. We find that the risk-adjusted returns of the long-short strategy by the nonlinear machine learning signals by the nonlinear machine learning model generate substantially higher risk-adjusted returns than the linear machine learning signals generated by the linear machine learning model in stocks with more speculative characteristics. Furthermore, the superior performance of nonlinear models becomes more pronounced in stocks with less-sophisticated shareholders as those stocks are more subjective to speculation. The results are also supported by Fama-Macbeth regression.

This paper contributes to the literature by suggesting that machine learning models'

nonlinearity plays an important role by capturing mispricing induced by behavioral biases. The flexible structure of machine learning models helps exploit the mis-valuation in stocks vulnerable to behavioral biases.

## References


Avramov, D., S. Cheng, and L. Metzker. Forthcoming. Machine learning versus economic restrictions: Evidence from stock return predictability. *Management science*.

Baker, M., and J. Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *The journal of Finance* 61: 1645-80.

Birru, J. 2018. Day of the week and the cross-section of returns. *Journal of financial economics* 130: 182-214.

Chen, L., M. Pelger, and J. Zhu. 2020. Deep learning in asset pricing. *Available at SSRN 3350138*.

Cochrane, J. H. 2011. Presidential address: Discount rates. *The journal of Finance* 66: 1047-108.

Cohen, L., and D. Lou. 2012. Complicated firms. *Journal of financial economics* 104: 383-400.

Cong, L. W., K. Tang, J. Wang, and Y. Zhang. 2020. AlphaPortfolio for investment and economically interpretable AI. *SSRN*, *https://papers.ssrn.com/sol3/papers.cfm*.

Daniel, K., D. Hirshleifer, and L. Sun. 2020. Short-and long-horizon behavioral factors. *The Review of Financial Studies* 33: 1673-736.

Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics* 33: 3-56.

———. 2015. A five-factor asset pricing model. *Journal of financial economics* 116: 1-22.

———. 2018. Choosing factors. *Journal of financial economics* 128: 234-52.

Green, J., J. R. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average US monthly stock returns. *The Review of Financial Studies* 30: 4389-436.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33: 2223-73.

———. 2021. Autoencoder asset pricing models. *Journal of Econometrics* 222: 429-50.

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

Hirshleifer, D., P.-H. Hsu, and D. Li. 2018. Innovative originality, profitability, and stock returns. *The Review of Financial Studies* 31: 2553-605.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting anomalies: An investment approach. *The Review of Financial Studies* 28: 650-705.

Jiang, G., C. M. Lee, and Y. Zhang. 2005. Information uncertainty and expected returns. *Review of Accounting Studies* 10: 185-221.

Karolyi, G. A., and S. Van Nieuwerburgh. 2020. New methods for the cross-section of returns. *The Review of Financial Studies* 33: 1879-90.

Kelly, B. T., S. Pruitt, and Y. Su. 2019. Characteristics are covariances: A unified model of risk and return. *Journal of financial economics* 134: 501-24.

Kumar, A. 2009. Who gambles in the stock market? *The journal of Finance* 64: 1889-933.

Lam, F. E. C., and K. J. Wei. 2011. Limits-to-arbitrage, investment frictions, and the asset growth anomaly. *Journal of financial economics* 102: 127-49.

Lee, C. M., S. T. Sun, R. Wang, and R. Zhang. 2019. Technological links and predictable returns. *Journal of financial economics* 132: 76-96.

Maskara, P. K., and D. J. Mullineaux. 2011. Information asymmetry and self-selection bias in bank loan announcement studies. *Journal of financial economics* 101: 684-94.

Mullainathan, S., and J. Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31: 87-106.

Nagel, S. 2021. *Machine Learning in Asset Pricing*. Princeton University Press.

Pástor, Ľ., and R. F. Stambaugh. 2003. Liquidity risk and expected stock returns. *Journal of political economy* 111: 642-85.

Seybert, N., and H. I. Yang. 2012. The party's over: The role of earnings guidance in resolving sentiment-driven overvaluation. *Management science* 58: 308-19.

Stambaugh, R. F., J. Yu, and Y. Yuan. 2012. The short of it: Investor sentiment and anomalies. *Journal of financial economics* 104: 288-302.

Zhang, X. F. 2006. Information uncertainty and stock returns. *The journal of Finance* 61: 105-37.

## Table 1. Descriptions of Machine Learning Signals

This table describes various machine learning signals. Our sample consists of 1,330,719 stock-month-level observations from 1987 to 2020. Panel A describes the construction and details of our baseline stock-month-level machine learning signals. IPCA is generated by the Instrumented Principal Component Analysis (IPCA) of Kelly et al. (2019). CA0, CA1, CA2, and CA3 are generated by the Conditional Autoencoder (CA) model of Gu et al. (2021). The CA3 signal is our main nonlinear signal of interest. The IPCA and CA0 signals are our main linear signals. We construct machine learning signals in the out-of-sample (OOS) fashion by following Gu et al. (2021). For the initial test year, 1987, the training sample consists of stock-month observations until 1980. The validation sample ranges from 1981 to 1986. In the training sample, parameters of machine learning signals are trained using 92 firm characteristics largely adopted from Gu et al. (2021). The validation sample is used to prevent the overfitting of models. As the test year incremented by one year from 1987, the training sample is also incremented by one year while the validation sample is rolled to make the length to be fixed for six years. For each test year, stock-month-level input variables of 92 firm characteristics generate machine learning signals using machine learning models trained in the training sample. In our sample, stocks without firm size, previous month returns, asset size, quarterly earnings announcement dates are dropped, and financial sector and utility sector are also dropped. Panel B describes the measures of representing the difference between the portfolio returns constructed nonlinear and linear signals. In each month, stocks are sorted into a decile and construct value-weighted portfolios by using the NYSE breakpoints of each machine learning signal. NML_IPCA (NML_CA0) is the difference between portfolio returns constructed by the nonlinear CA3 signal and the linear IPCA (CA0) signal.

Panel A. Baseline Signals

| Signal | Construction | Linearity |
|---|---|---|
| IPCA | IPCA model | Linear |
| CA0 | CA model, no nonlinear layer | Linear |
| CA1 | CA model, 1 nonlinear layer | Nonlinear |
| CA2 | CA model, 2 nonlinear layers | Nonlinear |
| CA3 | CA model, 3 nonlinear layers | Nonlinear |

Panel B. Nonlinear-minus-linear (NML)

| Signal | Construction |
|---|---|
| NML_IPCA | Differences in portfolio returns by the CA3 and IPCA signal |
| NML_CA0 | Differences in portfolio returns by the CA3 and CA0 signal |

**Table 2. Performance of Portfolios Sorted by Machine Learning Signals**

This table reports the excess returns (%) of portfolios sorted by machine learning signals generated by the Instrumented Principal Component Analysis (IPCA) of Kelly et al. (2019) and the Conditional Autoencoder (CA) model of Gu et al. (2021). Our sample ranges from 1987 to 2020. The baseline machine learning signals are the IPCA, CA0, CA1, CA2, and CA3 signals. The IPCA signal is a linear signal generated by the IPCA model of Kelly et al. (2019). The CA0 signal is a signal generated by the CA model without neural network structures. The CA1 (CA2, CA3) signal is the nonlinear signal generated by the CA model having one (two, three) layers of nonlinear activation functions. In each month, stocks are sorted into a decile using NYSE breakpoints of each IPCA, CA0, CA1, CA2, and CA3 signal. The portfolio returns are value-weighted. The portfolio-level differences of returns are denoted as nonlinear-minus-linear. For each decile portfolio returns constructed by the nonlinear signals and the linear signals, we take the differences of portfolio returns between the portfolios constructed by the nonlinear signals and the linear signals. CA1-IPCA is the difference between portfolio returns sorted by the CA1 signal and IPCA signal. CA2-IPCA and CA3-IPCA are similarly constructed. CA1-CA0 is the difference between portfolio returns sorted by the CA1 signal and the CA0 signal. CA2-CA0 and CA3-CA0 are also similarly constructed. NML_IPCA (NML_CA0) is the difference of portfolio returns between CA3 and IPCA (CA0). D1 is the lowest decile, and D10 is the highest decile. D10-D1 is the difference between the returns of D10 and D1. The standard errors are estimated by Newey-West adjustment. t-statistics are reported in the parentheses. ***, **, * denoted 1%, 5%, and 10% statistical significance.

| | Baseline Machine Learning Signal | | | | | Nonlinear-minus-linear | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Linear Signal | | Nonlinear Signal | | | IPCA as Linear Signal | | | CA0 as Linear Signal | | |
| Decile | IPCA | CA0 | CA1 | CA2 | CA3 | CA1-IPCA | CA2-IPCA | CA3-IPCA (NML_IPCA) | CA1-CA0 | CA2-CA0 | CA3-CA0 (NML_CA0) |
| D1 | 0.399 | 0.435 | 0.291 | 0.144 | 0.150 | -0.108 | -0.255** | -0.249* | -0.144 | -0.291** | -0.285** |
| | (1.31) | (1.49) | (0.77) | (0.38) | (0.40) | (-0.85) | (-2.01) | (-1.89) | (-1.07) | (-2.23) | (-2.08) |
| D2 | 0.518** | 0.563** | 0.583** | 0.577** | 0.602** | 0.065 | 0.059 | 0.084 | 0.020 | 0.015 | 0.040 |
| | (2.22) | (2.42) | (2.18) | (2.13) | (2.40) | (0.60) | (0.54) | (0.82) | (0.18) | (0.13) | (0.39) |
| D3 | 0.521** | 0.742*** | 0.641*** | 0.699*** | 0.639*** | 0.120 | 0.178* | 0.118 | -0.101 | -0.043 | -0.103 |
| | (2.14) | (3.12) | (2.62) | (3.22) | (3.14) | (1.11) | (1.69) | (1.04) | (-0.82) | (-0.39) | (-0.93) |
| D4 | 0.822*** | 0.910*** | 0.670*** | 0.785*** | 0.800*** | -0.152 | -0.036 | -0.022 | -0.240*** | -0.125 | -0.110 |
| | (3.72) | (4.25) | (3.04) | (3.94) | (4.11) | (-1.38) | (-0.38) | (-0.21) | (-2.64) | (-1.50) | (-1.20) |
| D5 | 0.822*** | 0.792*** | 0.812*** | 0.763*** | 0.847*** | -0.010 | -0.059 | 0.025 | 0.019 | -0.030 | 0.054 |
| | (3.60) | (3.45) | (4.29) | (3.54) | (3.65) | (-0.09) | (-0.57) | (0.24) | (0.16) | (-0.25) | (0.51) |
| D6 | 0.936*** | 0.881*** | 0.938*** | 1.018*** | 1.068*** | 0.002 | 0.082 | 0.132 | 0.057 | 0.137 | 0.186 |
| | (4.40) | (3.81) | (4.23) | (4.77) | (4.68) | (0.02) | (0.85) | (1.12) | (0.58) | (1.32) | (1.64) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D7 | 0.947*** | 0.830*** | 0.999*** | 1.099*** | 1.048*** | 0.052 | 0.152 | 0.102 | 0.169* | 0.269** | 0.219* |
| | (3.97) | (3.42) | (4.37) | (4.33) | (4.18) | (0.45) | (1.16) | (0.82) | (1.67) | (1.98) | (1.85) |
| D8 | 1.019*** | 1.095*** | 0.961*** | 1.116*** | 1.156*** | -0.058 | 0.097 | 0.137 | -0.134 | 0.022 | 0.061 |
| | (4.46) | (4.40) | (3.74) | (4.22) | (4.21) | (-0.46) | (0.79) | (1.11) | (-1.00) | (0.16) | (0.45) |
| D9 | 1.190*** | 0.908*** | 1.236*** | 1.214*** | 1.226*** | 0.046 | 0.024 | 0.036 | 0.328** | 0.306** | 0.318** |
| | (4.56) | (3.27) | (4.40) | (4.22) | (3.90) | (0.39) | (0.21) | (0.26) | (2.21) | (2.35) | (2.22) |
| D10 | 1.214*** | 1.272*** | 1.381*** | 1.584*** | 1.545*** | 0.168 | 0.370*** | 0.331** | 0.110 | 0.312** | 0.273* |
| | (4.29) | (4.46) | (4.01) | (4.74) | (4.26) | (1.15) | (2.71) | (2.23) | (0.77) | (2.38) | (1.76) |
| D10-D1 | 0.815*** | 0.837*** | 1.090*** | 1.440*** | 1.395*** | 0.275 | 0.625*** | 0.581*** | 0.253 | 0.603*** | 0.558*** |
| | (3.52) | (3.55) | (4.11) | (5.48) | (5.34) | (1.56) | (3.68) | (3.35) | (1.50) | (3.95) | (3.40) |

**Table 3. Risk-Adjusted Returns of the Differences in Decile Portfolios Sorted by the Nonlinear and Linear Machine Learning Signals**

This table reports the differences of risk-adjusted returns (%) of decile portfolios sorted by nonlinear and linear machine learning signals. The returns are adjusted by various factor models. The machine learning signals are generated by the Instrumented Principal Component Analysis (IPCA) model of Kelly et al. (2019) and the Conditional Autoencoder (CA) model of Gu et al. (2021). The nonlinear machine learning signals are CA1, CA2, and CA3 generated using the CA model with 1, 2, and 3 nonlinear layers, respectively. The linear machine learning signals are IPCA and CA0, generated by using the IPCA and the CA model having no nonlinear layers, respectively. Our sample ranges from 1987 to 2020. To construct the portfolio-level differences of returns between nonlinear and linear signals, we first construct the decile portfolios using NYSE breakpoints of each machine learning signal. Portfolio returns are value-weighted. Then, we take the difference of portfolio returns by nonlinear signal and linear signal. CA1-IPCA is the difference of portfolio returns sorted by the CA1 and IPCA signal. CA2-IPCA and CA3-IPCA are similarly constructed. CA3-IPCA is also denoted by NML_IPCA. CA1-CA0 is the difference between portfolio returns sorted by CA1 and CA0 signal. CA2-CA0 and CA3-CA0 are similarly constructed. CA3-CA0 is also denoted by NML_CA0. The portfolio returns are adjusted by various factor models, including the CAPM, the Fama-French three-factor model by Fama and French (1993) (FF3), the Fama-French five-factor model by Fama and French (2015) (FF5), the Fama-French five-factor model augmented by Carhart's momentum factor by Fama and French (2018) (FF6), the long- and short-horizon behavioral factor model by Daniel et al. (2020) (DHS), the q-factor model by Hou et al. (2015), and the Fama-French three-factor model augmented by the liquidity factor of Pástor and Stambaugh (2003) (PS). For brevity, we report only the lowest decile (D1), and the highest decile (D10), and the difference between D10 and D1 (D10-D1). The standard errors are estimated by Newey-West adjustment. t-statistics are reported in the parentheses. ***, **, * denoted 1%, 5%, and 10% statistical significance.

| Factor | Decile | Nonlinear-minus-linear by IPCA | | | Nonlinear-minus-linear by CA0 | | |
|---|---|---|---|---|---|---|---|
| | | CA1-IPCA | CA2-IPCA | CA3-IPCA (NML_IPCA) | CA1-CA0 | CA2-CA0 | CA3-CA0 (NML_CA0) |
| CAPM | D1 | -0.282** | -0.441*** | -0.430*** | -0.356*** | -0.515*** | -0.504*** |
| | | (-2.50) | (-3.87) | (-3.68) | (-3.22) | (-4.73) | (-4.39) |
| | D10 | 0.021 | 0.240* | 0.141 | -0.005 | 0.214 | 0.115 |
| | | (0.15) | (1.70) | (1.02) | (-0.04) | (1.55) | (0.79) |
| | D10-D1 | 0.303 | 0.681*** | 0.571*** | 0.351** | 0.729*** | 0.619*** |
| | | (1.64) | (3.62) | (3.04) | (2.02) | (4.34) | (3.50) |
| FF3 | D1 | -0.291*** | -0.442*** | -0.432*** | -0.368*** | -0.519*** | -0.509*** |
| | | (-2.64) | (-3.82) | (-3.60) | (-3.51) | (-4.72) | (-4.32) |
| | D10 | 0.014 | 0.219* | 0.131 | -0.009 | 0.195 | 0.107 |
| | | (0.11) | (1.74) | (1.05) | (-0.07) | (1.49) | (0.77) |
| | D10-D1 | 0.305* | 0.660*** | 0.563*** | 0.359** | 0.714*** | 0.616*** |
| | | (1.65) | (3.66) | (3.12) | (2.04) | (4.22) | (3.48) |
| FF5 | D1 | -0.274** | -0.410*** | -0.421*** | -0.302*** | -0.438*** | -0.449*** |
| | | (-2.45) | (-3.52) | (-3.19) | (-2.80) | (-3.94) | (-3.51) |
| | D10 | 0.102 | 0.307** | 0.249** | 0.020 | 0.225* | 0.167 |
| | | (0.76) | (2.29) | (2.00) | (0.15) | (1.86) | (1.30) |
| | D10-D1 | 0.376** | 0.716*** | 0.670*** | 0.322* | 0.663*** | 0.616*** |
| | | (2.02) | (3.85) | (3.37) | (1.75) | (3.97) | (3.13) |
| FF6 | D1 | -0.178 | -0.339*** | -0.342*** | -0.175 | -0.336*** | -0.338*** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | (-1.54) | (-2.97) | (-2.71) | (-1.64) | (-3.25) | (-2.93) |
| | D10 | 0.157 | 0.367*** | 0.314** | 0.061 | 0.271** | 0.218 |
| | | (1.16) | (2.75) | (2.51) | (0.44) | (2.21) | (1.58) |
| | D10-D1 | 0.335* | 0.706*** | 0.655*** | 0.236 | 0.606*** | 0.556*** |
| | | (1.75) | (3.69) | (3.30) | (1.23) | (3.60) | (2.82) |
| DHS | D1 | -0.086 | -0.261** | -0.214* | -0.082 | -0.257** | -0.211 |
| | | (-0.79) | (-2.26) | (-1.71) | (-0.72) | (-2.21) | (-1.62) |
| | D10 | 0.149 | 0.357** | 0.261* | 0.027 | 0.235* | 0.140 |
| | | (1.07) | (2.29) | (1.76) | (0.20) | (1.66) | (0.88) |
| | D10-D1 | 0.234 | 0.618*** | 0.476** | 0.109 | 0.492*** | 0.350* |
| | | (1.29) | (3.05) | (2.32) | (0.61) | (2.73) | (1.68) |
| q-factor | D1 | -0.185 | -0.361*** | -0.365*** | -0.164 | -0.340*** | -0.344** |
| | | (-1.47) | (-2.94) | (-2.63) | (-1.27) | (-2.76) | (-2.47) |
| | D10 | 0.130 | 0.321** | 0.299** | 0.073 | 0.264** | 0.242* |
| | | (0.93) | (2.27) | (2.39) | (0.51) | (2.03) | (1.73) |
| | D10-D1 | 0.314 | 0.683*** | 0.665*** | 0.236 | 0.605*** | 0.587*** |
| | | (1.54) | (3.42) | (3.29) | (1.12) | (3.27) | (2.79) |
| PS | D1 | -0.304*** | -0.450*** | -0.438*** | -0.391*** | -0.537*** | -0.525*** |
| | | (-2.73) | (-3.79) | (-3.59) | (-3.76) | (-4.85) | (-4.49) |
| | D10 | 0.011 | 0.202 | 0.133 | -0.008 | 0.183 | 0.114 |
| | | (0.08) | (1.53) | (1.04) | (-0.06) | (1.30) | (0.79) |
| | D10-D1 | 0.315* | 0.653*** | 0.571*** | 0.383** | 0.720*** | 0.639*** |
| | | (1.67) | (3.40) | (3.06) | (2.14) | (3.98) | (3.53) |

## Table 4. Fama-Macbeth Regression of Machine Learning Signals

This table reports the Fama-Macbeth regression results of returns on machine learning signals. Our sample ranges from 1987 to 2020, consisting of 408 months. The main dependent variable is one-month ahead returns (%). The baseline machine learning signals are the IPCA (or CA0) and CA3 signals. The linear IPCA signal is generated by Kelly et al. (2019). The linear CA0 signal and the nonlinear CA3 signal are generated by Gu et al. (2021). The CA3 and IPCA, CA0 signals are rank-transformed into a unit interval to mitigate the outliers' effect. The stock-level nonlinear-minus-linear by the CA3 and IPCA signal (NML_IPCA) is constructed by taking the difference between the nonlinear CA3 signal and the linear IPCA signal. The NML_CA0 is the difference between the CA3 and CA0 signals. Panel A (Panel B) uses the IPCA (CA0) signal as the linear signal. Control variables are as follows: market beta, firm size, book-to-market ratio, momentum, asset growth, return on equity, illiquidity, bid-ask spread, and short-term reversal. The standard errors are estimated by Newey-West adjustment. t-statistics are reported in the parentheses. ***, **, * denoted 1%, 5%, and 10% statistical significance.

| Panel A. IPCA as Linear Signal | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | Returns (%) | | | | |
| IPCA | 2.908*** | | | 3.553*** | 3.501*** | | | 4.604*** |
| | (11.06) | | | (11.73) | (13.81) | | | (14.62) |
| CA3 | | 3.534*** | | | | 4.074*** | | |
| | | (11.91) | | | | (14.31) | | |
| NML_IPCA | | | 1.711*** | 3.488*** | | | 2.124*** | 3.537*** |
| | | | (5.98) | (9.60) | | | (8.52) | (11.49) |
| Constant | -0.486 | -0.800** | 0.968*** | -0.809** | -0.452 | -0.885 | 3.134*** | -1.410* |
| | (-1.30) | (-2.19) | (2.89) | (-2.16) | (-0.55) | (-1.15) | (3.37) | (-1.79) |
| Controls | No | No | No | No | Yes | Yes | Yes | Yes |
| Months | 408 | 408 | 408 | 408 | 408 | 408 | 408 | 408 |
| Adjusted $R^2$ | 0.006 | 0.007 | 0.002 | 0.009 | 0.047 | 0.048 | 0.047 | 0.049 |
| Panel B. CA0 as Linear Signal | | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | | | | Returns (%) | | | | |
| CA0 | 2.496*** | | | 3.311*** | 3.841*** | | | 3.993*** |
| | (9.09) | | | (10.77) | (13.87) | | | (14.17) |
| CA3 | | 3.534*** | | | | 4.074*** | | |
| | | (11.91) | | | | (14.31) | | |
| NML_CA | | | 2.839*** | 4.494*** | | | 3.995*** | 4.088*** |
| | | | (9.45) | (12.31) | | | (11.92) | (12.07) |
| Constant | -0.281 | -0.800** | 0.968*** | -0.688* | -0.604 | -0.885 | 2.939*** | -0.889 |
| | (-0.78) | (-2.19) | (2.89) | (-1.87) | (-0.72) | (-1.15) | (3.22) | (-1.10) |
| Controls | No | No | No | No | Yes | Yes | Yes | Yes |
| Months | 408 | 408 | 408 | 408 | 408 | 408 | 408 | 408 |
| Adjusted $R^2$ | 0.006 | 0.007 | 0.003 | 0.009 | 0.047 | 0.048 | 0.048 | 0.049 |

## Table 5. Descriptions of Speculative Anomaly Variables

This table describes speculative anomaly variables suggested by Birru (2018). There are 20 anomaly variables and a stock-month-level composite signal of speculative characteristics, SPEC, constructed by using the methodology of Maskara and Mullineaux (2011). Each anomaly has a leg of featuring speculative characteristics. In our empirical analysis, we sort stocks into quintiles. The leg featured with the speculative characteristics is denoted by either Q1 or Q5. In the case of dummy variables, E (earnings), CF (cash flow), and D (dividend), stocks are divided into two groups. For E (CF), the groups are negative or positive earnings (cash flow). For D, the groups are non-payer of dividends or payers. SPEC is the stock-month-level composite signal of speculative characteristics constructed using the 20 anomaly variables. We rank-transform each of the speculative anomaly variables into a unit interval. We further adjust the direction of speculative anomaly variables to represent higher ranks as more speculative characteristics. Therefore, the rank of speculative anomaly variables are reversed: firm size (SIZE), the stock price (PRICE), firm age (AGE), nearness to 52-week high (NH), return-on-asset (ROA), and operating profitability (OP) is reversed. In the dummy variables, including E, CF, and D, we assign 0.25 to the non-speculative and 0.75 to the speculative leg. Then, we take the average of those 20 rank-transformed and direction-adjusted anomaly variables by following the methodology of

| Anomaly | Description | Speculative Characteristics | Speculative Leg |
|---|---|---|---|
| SIGMA | Stcok return volatility | Volatile | Q5 |
| IVOL | Idiosyncratic volatility | Volatile | Q5 |
| BETA | Market beta | Speculative | Q5 |
| MAX | Maximum daily returns | Lottery | Q5 |
| BIDASK | Bid-ask spread | High transaction cost | Q5 |
| ILLIQ | Illiquidity | Illiquid | Q5 |
| O-Score | Bankruptcy probability | Close-to-default | Q5 |
| FP | Default probability | Close-to-default | Q5 |
| CFV | Cash flow volatility | Volatile cash flow | Q5 |
| NXF | Net external financing | Extreme growth | Q5 |
| DISP | Analyst dispersion | Dispersed opinion | Q5 |
| SIZE | Firm size | Small | Q1 |
| PRICE | Stock price | Lottery | Q1 |
| AGE | Firm age | Young | Q1 |
| NH | Nearness to 52-week high | Anchoring bias | Q1 |
| ROA | Return on assets | Unprofitable | Q1 |
| OP | Operating profitability | Unprofitable | Q1 |
| E | Earnings | Negative earnings | Negative |
| CF | Cash flow | Negative cash flow | Negative |
| D | Dividend | No dividend | Non-payers |
| SPEC | Speculative index | Speculative | Q5 |

**Table 6. The Differences of Risk-Adjusted Returns of Long-Short Strategy by Nonlinear and Linear Machine Learnings Signals in Quintiles of Speculative Anomaly Variables**

This table reports the differences of risk-adjusted returns (%) of the long-short portfolios sorted by the nonlinear and linear signals in quintiles of speculative anomaly variables. There are 20 speculative anomaly variables suggested by Birru (2018). Our sample ranges from 1987 to 2020. For each month, stocks are sorted into quintiles based on the NYSE breakpoints of one of the speculative anomaly variables. Then within each quintile, stocks are further sorted into deciles using the NYSE breakpoints of the machine learning signals. The returns are value-weighted. We implement the long-short portfolio strategy of buying the highest decile and shorting the lowest decile in each quintile are reported. Then, we report the differences of long-short portfolio returns by the nonlinear CA3 signal and the linear IPCA (or CA0) signal. NML_IPCA (NML_CA0) is the difference of long-short portfolio returns between the long-short portfolios constructed by CA3 and IPCA (CA0) signals in each of the quintiles of speculative anomaly variables. For dummy variables that are E (positive earnings), CF (positive cash flow), and D (dividend payer), stocks are sorted into two groups rather than quintiles. The speculative leg is denoted in parentheses. The returns are adjusted by the six-factor model of Fama and French (2018). SPEC is the rank-transformed average of 20 speculative anomaly variables aligned to have higher values as more speculative characteristics by following the methodology of Maskara and Mullineaux (2011). The standard errors are estimated by Newey-West adjustment. t-statistics are reported in the parentheses. ***, **, * denoted 1%, 5%, and 10% statistical significance.

| | NML_IPCA | | | NML_CA0 | | |
|---|---|---|---|---|---|---|
| Anomaly | Q1 | Q5 | Q5-Q1 | Q1 | Q5 | Q5-Q1 |
| SIGMA | 0.065 | 1.573*** | 1.508*** | 0.030 | 1.904*** | 1.874*** |
| | (0.35) | (3.70) | (3.15) | (0.15) | (5.52) | (4.91) |
| IVOL | 0.267 | 1.088*** | 0.821** | 0.022 | 1.744*** | 1.723*** |
| | (1.17) | (3.44) | (2.19) | (0.12) | (5.71) | (5.34) |
| BETA | 0.148 | 0.891*** | 0.743* | 0.120 | 0.959*** | 0.839** |
| | (0.73) | (2.88) | (1.88) | (0.65) | (3.20) | (2.23) |
| MAX | 0.031 | 1.324*** | 1.293*** | -0.251 | 1.831*** | 2.083*** |
| | (0.13) | (3.65) | (2.63) | (-1.08) | (4.75) | (4.30) |
| BIDASK | -0.099 | 1.335*** | 1.434*** | -0.211 | 1.544*** | 1.755*** |
| | (-0.49) | (3.44) | (3.09) | (-1.14) | (4.64) | (4.43) |
| ILLIQ | 0.219 | 0.817*** | 0.598** | 0.318 | 1.232*** | 0.914*** |
| | (1.06) | (3.87) | (2.10) | (1.61) | (6.48) | (3.17) |
| O-Score | 0.198 | 1.180*** | 0.982*** | 0.341 | 1.474*** | 1.133*** |
| | (0.70) | (4.21) | (2.90) | (1.31) | (5.07) | (2.99) |
| FP | 0.217 | 0.582* | 0.366 | 0.307 | 1.324*** | 1.018** |
| | (0.91) | (1.66) | (0.90) | (1.20) | (3.84) | (2.56) |
| CFV | 0.381 | 0.822*** | 0.441 | 0.333 | 1.008*** | 0.675* |
| | (1.60) | (2.60) | (1.13) | (1.40) | (3.32) | (1.65) |
| NXF | 0.385 | 0.707*** | 0.323 | -0.018 | 1.005*** | 1.023** |
| | (1.19) | (2.59) | (0.73) | (-0.06) | (3.77) | (2.45) |
| DISP | 0.045 | 0.479 | 0.434 | 0.160 | 0.378 | 0.218 |
| | (0.15) | (1.40) | (1.00) | (0.60) | (1.13) | (0.59) |
| SIZE | 0.749*** | 0.189 | -0.559* | 1.282*** | 0.298 | -0.984*** |
| | (3.39) | (0.88) | (-1.87) | (6.65) | (1.47) | (-3.67) |
| PRICE | 1.089*** | 0.105 | -0.984** | 1.861*** | 0.171 | -1.690*** |

| | | | | | | |
|---|---|---|---|---|---|---|
| | (3.69) | (0.44) | (-2.39) | (5.88) | (0.61) | (-3.58) |
| AGE | 0.891*** | -0.108 | -0.999** | 0.739*** | 0.020 | -0.718* |
| | (3.18) | (-0.41) | (-2.57) | (2.59) | (0.09) | (-1.91) |
| NH | 1.714*** | -0.087 | -1.801*** | 1.984*** | 0.227 | -1.757*** |
| | (3.60) | (-0.32) | (-3.40) | (5.01) | (0.84) | (-3.71) |
| ROA | 1.456*** | 0.538** | -0.918** | 1.572*** | 0.473** | -1.099*** |
| | (3.61) | (2.17) | (-2.12) | (4.73) | (2.01) | (-2.80) |
| OP | 0.806** | -0.128 | -0.934** | 1.269*** | -0.121 | -1.389*** |
| | (2.29) | (-0.45) | (-2.02) | (3.66) | (-0.42) | (-3.13) |
| E | 1.275*** | 0.420** | -0.855** | 1.794*** | 0.402** | -1.392*** |
| | (3.76) | (2.17) | (-2.34) | (5.30) | (2.20) | (-3.64) |
| CF | 1.980*** | 0.377** | -1.603*** | 2.519*** | 0.325* | -2.194*** |
| | (4.32) | (1.98) | (-3.38) | (5.69) | (1.71) | (-4.45) |
| D | 1.029*** | 0.267 | -0.763** | 1.296*** | 0.109 | -1.187*** |
| | (3.70) | (1.21) | (-2.10) | (4.46) | (0.51) | (-3.06) |
| SPEC | 0.357 | 1.183*** | 0.826** | 0.071 | 1.846*** | 1.775*** |
| | (1.50) | (3.75) | (2.45) | (0.36) | (5.84) | (4.92) |

**Table 7. The Effect of Shareholder Sophistication on the Performance of Nonlinear Machine Learning Signals**

This table reports the differences of risk-adjusted returns (%) of long-short portfolios by the nonlinear and linear machine learning signals conditional on the shareholder sophistication and speculative characteristics. We proxy the shareholder sophistication by using the number of institutional investors (Lam and Wei 2011). For each month, stocks are divided into two groups based on the median value of their number of institutional investors. A low number of institutional investors represent low shareholder sophistication. Then, for each of the low and the high groups divided by the number of institutional investors, stocks are further double-sorted by a speculative anomaly variable and a machine learning signal. The speculative anomaly variables are suggested by Birru (2018). By using the speculative anomaly variables, stocks are sorted into quintiles. In each quintile, stocks are sorted into deciles using machine learning signals, including the IPCA, CA0, and CA3 signals. We implement the long-short strategy of buying the highest decile and shorting the lowest decile.
The differences of returns of long-short portfolios between the nonlinear CA3 signal and the linear IPCA (or CA0) signal are reported for each of low and high groups of a number of institutional investors and quintiles of speculative anomaly variables. The speculative leg is denoted in parentheses as either Q1 or Q5. For dummy variables E, Q1 refers to negative earnings, for CF, negative cash flow, and for D, non-payers of dividends. NML_IPCA (NML_CA0) is the difference of risk-adjusted returns between the long-short strategy using the CA3 and IPCA (CA0) signals. All portfolios are sorted based on the NYSE breakpoints. Returns are value-weighted, and the risk-adjusted returns are calculated by the six-factor model of Fama and French (2018). The standard errors are estimated by Newey-West adjustment. t-statistics are reported in the parentheses. ***, **, * denoted 1%, 5%, and 10% statistical significance.

| | Low Number of Institutional Investors | | | | | | High Number of Institutional Investors | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NML_IPCA | | | NML_CA0 | | | NML_IPCA | | | NML_CA0 | | |
| Anomaly | Q1 | Q5 | Q5-Q1 | Q1 | Q5 | Q5-Q1 | Q1 | Q5 | Q5-Q1 | Q1 | Q5 | Q5-Q1 |
| SIGMA (Q5) | 0.322** | 2.060*** | 1.738*** | 0.309* | 2.497*** | 2.188*** | 0.038 | 0.400 | 0.362 | -0.031 | 0.755** | 0.786** |
| | (2.19) | (5.14) | (4.04) | (1.87) | (6.04) | (4.76) | (0.15) | (1.09) | (0.87) | (-0.13) | (2.04) | (2.00) |
| IVOL (Q5) | 0.109 | 1.873*** | 1.764*** | 0.181 | 2.276*** | 2.095*** | 0.246 | 0.089 | -0.157 | -0.132 | 0.402 | 0.533 |
| | (0.67) | (5.49) | (4.74) | (1.09) | (6.56) | (5.73) | (1.06) | (0.31) | (-0.42) | (-0.56) | (1.36) | (1.58) |
| BETA (Q5) | 0.499*** | 0.844** | 0.344 | 0.647*** | 1.447*** | 0.800* | 0.148 | 0.200 | 0.052 | 0.079 | 0.387 | 0.308 |
| | (2.67) | (2.38) | (0.91) | (3.34) | (3.55) | (1.79) | (0.60) | (0.71) | (0.14) | (0.33) | (1.31) | (0.79) |
| MAX (Q5) | 0.526*** | 2.293*** | 1.766*** | 0.338* | 2.918*** | 2.579*** | 0.233 | 0.357 | 0.124 | -0.100 | 0.716* | 0.817 |
| | (2.74) | (5.69) | (4.25) | (1.88) | (7.15) | (5.52) | (1.02) | (1.02) | (0.27) | (-0.48) | (1.70) | (1.61) |
| BIDASK (Q5) | 0.085 | 2.042*** | 1.957*** | 0.254 | 2.333*** | 2.079*** | 0.069 | 0.443 | 0.374 | -0.234 | 1.014*** | 1.248*** |
| | (0.50) | (5.10) | (4.24) | (1.51) | (5.54) | (4.41) | (0.30) | (1.23) | (0.87) | (-1.05) | (2.73) | (2.94) |
| ILLIQ (Q5) | 0.361 | 1.099*** | 0.737* | 0.894*** | 1.518*** | 0.624 | 0.234 | 0.321 | 0.087 | 0.274 | 0.535** | 0.261 |
| | (1.46) | (3.26) | (1.70) | (3.39) | (5.16) | (1.53) | (1.11) | (1.23) | (0.24) | (1.39) | (2.23) | (0.87) |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O-Score (Q5) | 0.101 | 1.543*** | 1.442*** | 0.452** | 1.655*** | 1.203*** | 0.007 | 0.376 | 0.368 | 0.031 | 0.583** | 0.553 |
| | (0.50) | (4.18) | (3.62) | (2.26) | (4.59) | (2.97) | (0.02) | (1.41) | (0.94) | (0.12) | (2.26) | (1.62) |
| FP (Q5) | 0.345* | 0.621** | 0.276 | 0.556*** | 1.110*** | 0.554 | 0.091 | 0.473 | 0.382 | 0.197 | 0.566 | 0.369 |
| | (1.79) | (2.14) | (0.76) | (2.71) | (4.05) | (1.62) | (0.32) | (1.36) | (0.96) | (0.82) | (1.56) | (0.94) |
| CFV (Q5) | -0.132 | 1.079*** | 1.211*** | 0.237 | 1.355*** | 1.118*** | 0.363 | 0.443 | 0.081 | 0.186 | 0.759** | 0.573 |
| | (-0.66) | (3.67) | (3.68) | (1.25) | (3.97) | (3.23) | (1.50) | (1.29) | (0.20) | (0.87) | (2.50) | (1.62) |
| NXF (Q5) | 0.635** | 0.545* | -0.090 | 1.056*** | 1.114*** | 0.058 | 0.027 | 0.532 | 0.505 | 0.048 | 0.869*** | 0.821* |
| | (2.47) | (1.91) | (-0.26) | (3.91) | (3.96) | (0.17) | (0.08) | (1.53) | (1.08) | (0.15) | (2.58) | (1.70) |
| DISP (Q5) | -0.032 | 0.469 | 0.500 | 0.086 | 0.612* | 0.526 | 0.336 | -0.042 | -0.378 | 0.506* | 0.370 | -0.136 |
| | (-0.14) | (1.28) | (1.28) | (0.36) | (1.66) | (1.23) | (1.04) | (-0.14) | (-0.88) | (1.84) | (1.28) | (-0.36) |
| SIZE (Q1) | 1.253*** | 0.188 | -1.066** | 1.836*** | 0.489** | -1.347*** | 0.327 | 0.150 | -0.177 | 0.532** | 0.225 | -0.307 |
| | (4.06) | (0.75) | (-2.32) | (7.09) | (2.04) | (-3.53) | (1.23) | (0.68) | (-0.49) | (2.05) | (1.11) | (-0.98) |
| PRICE (Q1) | 1.120*** | 0.152 | -0.968** | 1.945*** | 0.336 | -1.609*** | 0.991*** | 0.176 | -0.815* | 1.006*** | 0.316 | -0.691* |
| | (3.84) | (0.63) | (-2.29) | (7.29) | (1.41) | (-4.49) | (2.62) | (0.66) | (-1.68) | (3.18) | (1.22) | (-1.69) |
| AGE (Q1) | 0.426* | 0.653** | 0.228 | 0.746*** | 0.748*** | 0.003 | 0.268 | -0.139 | -0.407 | 0.511 | -0.269 | -0.780** |
| | (1.76) | (2.31) | (0.60) | (2.88) | (2.64) | (0.01) | (0.91) | (-0.53) | (-1.07) | (1.55) | (-1.12) | (-2.00) |
| NH (Q1) | 1.245*** | 0.211 | -1.034** | 1.794*** | 0.481** | -1.314*** | 0.815* | -0.077 | -0.891* | 0.695* | 0.259 | -0.435 |
| | (3.25) | (0.97) | (-2.32) | (4.86) | (2.02) | (-2.79) | (1.94) | (-0.28) | (-1.92) | (1.85) | (0.91) | (-0.94) |
| ROA (Q1) | 1.845*** | 0.167 | -1.678*** | 2.214*** | 0.289 | -1.925*** | -0.115 | 0.290 | 0.405 | 0.197 | 0.028 | -0.169 |
| | (4.98) | (0.79) | (-3.79) | (6.36) | (1.29) | (-4.22) | (-0.34) | (1.07) | (0.91) | (0.53) | (0.10) | (-0.35) |
| OP (Q1) | 1.295*** | 0.147 | -1.148*** | 1.510*** | 0.270 | -1.240*** | 0.449 | 0.432 | -0.017 | 0.667* | 0.311 | -0.356 |
| | (5.25) | (0.65) | (-3.40) | (6.52) | (1.13) | (-3.72) | (1.26) | (1.30) | (-0.03) | (1.73) | (0.96) | (-0.69) |
| E (Q1) | 1.333*** | 0.222 | -1.111*** | 1.819*** | 0.474** | -1.345*** | 1.114** | 0.175 | -0.940* | 1.280** | 0.216 | -1.064** |
| | (3.92) | (1.32) | (-3.14) | (5.51) | (2.55) | (-4.07) | (2.37) | (0.82) | (-1.90) | (2.51) | (1.03) | (-1.97) |
| CF (Q1) | 1.775*** | 0.300* | -1.475*** | 1.982*** | 0.567*** | -1.414*** | 1.218** | 0.265 | -0.953 | 1.748*** | 0.318 | -1.431** |
| | (4.23) | (1.71) | (-3.71) | (5.06) | (3.00) | (-3.80) | (2.14) | (1.29) | (-1.61) | (2.96) | (1.58) | (-2.25) |
| D (Q1) | 0.763*** | 0.521*** | -0.242 | 1.187*** | 0.519*** | -0.668*** | 0.024 | 0.383 | 0.358 | 0.505* | 0.265 | -0.240 |
| | (3.67) | (3.06) | (-0.89) | (6.17) | (3.01) | (-2.98) | (0.08) | (1.49) | (0.92) | (1.94) | (1.23) | (-0.65) |
| SPEC (Q5) | -0.034 | 2.167*** | 2.201*** | 0.080 | 2.409*** | 2.330*** | 0.191 | 0.428 | 0.237 | 0.027 | 1.041*** | 1.014** |
| | (-0.20) | (5.86) | (5.58) | (0.51) | (6.33) | (5.61) | (0.87) | (1.34) | (0.64) | (0.13) | (2.58) | (2.27) |

## Table 8. Fama-Macbeth Regression of Machine Learning Signals Interacted with Speculative Anomaly Variables

This table reports the Fama-Macbeth regression results of returns on machine learning signals interacted with the stock-level composite signal of speculative characteristics (SPEC). Our sample ranges from 1987 to 2020, consisting of 408 months. The main dependent variable is one-month ahead returns (%). The main independent variables are machine learning signals, including IPCA, CA0, NML_IPCA, NML_CA0, and their interactions with the stock-level composite signal of speculative characteristics (SPEC). The linear IPCA signal is generated by Kelly et al. (2019). The linear CA0 signal and the nonlinear CA3 signal are generated by Gu et al. (2021). The IPCA, CA0, and CA3 signals are rank-transformed to mitigate the effect of outliers. NML_IPCA (NML_CA0) is the stock-month level difference between CA3 and IPCA (CA0) signal. SPEC is the rank-transformed average of 20 speculative anomaly variables aligned to have higher values as more speculative characteristics by following the methodology of Maskara and Mullineaux (2011). SPEC_HIGH is a dummy variable that is 1 when SPEC is higher than the 80th percentile based on its NYSE breakpoints. SPEC_MID is a dummy variable that is 1 when SPEC is between 20th to 80th percentile based on its NYSE breakpoints. In calculating the nonlinear-minus-linear signal, Panel A uses the IPCA signal as the linear signal, and Panel B uses the CA0 signal. Control variables are as follows: market beta, firm size, book-to-market ratio, momentum, asset growth, return on equity, illiquidity, bid-ask spread, and short-term reversal. The standard errors are estimated by Newey-West adjustment. t-statistics are reported in the parentheses. ***, **, * denoted 1%, 5%, and 10% statistical significance.

| Panel A. IPCA as Linear Signal | | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | | Returns (%) | | |
| SPEC_MID | -0.079 | -0.181 | -0.219** | -0.348** |
| | (-0.65) | (-0.93) | (-2.07) | (-2.29) |
| SPEC_HIGH | 0.017 | -0.988** | -0.305* | -1.405*** |
| | (0.05) | (-2.18) | (-1.94) | (-5.46) |
| IPCA | 3.658*** | 1.922*** | 4.588*** | 2.882*** |
| | (12.43) | (6.18) | (14.52) | (9.57) |
| NML_IPCA | 3.458*** | 1.402*** | 3.599*** | 1.606*** |
| | (9.47) | (3.10) | (11.51) | (4.10) |
| IPCA × SPEC_MID | | 0.520** | | 0.525** |
| | | (2.13) | | (2.48) |
| IPCA × SPEC_HIGH | | 2.268*** | | 2.350*** |
| | | (6.32) | | (7.09) |
| NML_IPCA × SPEC_MID | | 0.591 | | 0.516 |
| | | (1.56) | | (1.51) |
| NML_IPCA × SPEC_HIGH | | 2.323*** | | 2.240*** |
| | | (4.59) | | (5.18) |
| Constant | -0.827*** | -0.085 | -1.217 | -0.572 |
| | (-3.56) | (-0.46) | (-1.52) | (-0.70) |
| Controls | No | No | Yes | Yes |
| Months | 408 | 408 | 408 | 408 |
| Adjusted $R^2$ | 0.021 | 0.022 | 0.050 | 0.050 |

Panel B. CA0 as Linear Signal

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Returns (%) | | | |
| SPEC_MID | -0.122 | -0.205 | -0.239** | -0.323** |
| | (-0.99) | (-1.03) | (-2.25) | (-2.08) |
| SPEC_HIGH | -0.014 | -0.802* | -0.319** | -1.189*** |
| | (-0.04) | (-1.81) | (-2.00) | (-4.66) |
| CA0 | 3.442*** | 1.904*** | 3.990*** | 2.578*** |
| | (11.65) | (6.09) | (14.12) | (8.85) |
| NML_IPCA | 4.356*** | 1.281*** | 4.136*** | 1.102*** |
| | (11.72) | (2.78) | (12.06) | (2.65) |
| CA0 × SPEC_MID | | 0.511** | | 0.476** |
| | | (2.07) | | (2.16) |
| CA0 × SPEC_HIGH | | 1.850*** | | 1.914*** |
| | | (5.17) | | (5.81) |
| NML_CA0 × SPEC_MID | | 1.065** | | 1.033*** |
| | | (2.35) | | (2.72) |
| NML_CA0 × SPEC_HIGH | | 3.794*** | | 3.667*** |
| | | (6.57) | | (7.24) |
| Constant | -0.686*** | -0.045 | -0.697 | -0.217 |
| | (-3.02) | (-0.24) | (-0.85) | (-0.26) |
| Controls | No | No | Yes | Yes |
| Months | 408 | 408 | 408 | 408 |
| Adjusted $R^2$ | 0.021 | 0.022 | 0.050 | 0.050 |