A Machine Learning Approach to Understanding Short Interest and Stock Returns

Ryan Huang[*]

Department of Economics

University of North Carolina

rhuang1@unc.edu

This Version: May 6, 2025

Abstract

We investigate how asset prices incorporate information through the lens of shorting activity, focusing on both rational explanations and potential behavioral biases among different types of market participants. We ask whether observed shorting patterns reflect rational updates to new information or instead reveal biases such as extrapolation, overconfidence, or prospect theory–type behavior. We adapt recent machine-learning-based approaches from empirical asset pricing by replacing future returns as the dependent variable with shorting volume. By leveraging a large cross-section of firm and macro-level predictors—encompassing both standard risk factors and variables linked to behavioral tendencies—we aim to uncover which signals dominate short-seller behavior and how strongly those signals propagate into market prices.

Keywords: Machine Learning, Big Data, Short Interest Prediction, Cross-Section of Returns, Ridge Regression, (Group) Lasso, Elastic Net, Random Forest, Gradient Boosting, (Deep) Neural Networks, Fintech

[*] We benefitted with discussions with Peter Hansen and Christopher Roark throughout the Spring 2025 semester at UNC Chapel Hill. We gratefully acknowledge the computing support from Research Computing at the University of North Carolina at Chapel Hill. We also are grateful for the cleaned and compiled data by Xin He used in Dacheng Xiu's recent paper on Empirical Asset Pricing via Machine Learning.

1 Introduction

Short selling plays a pivotal role in financial markets, often viewed with suspicion yet fundamentally linked to price discovery and market efficiency. Short Interest—the total amount of shares sold short—is a key gauge of aggregate shorting activity and investor sentiment. Prior research indicates that short sellers are generally informed investors who anticipate overvaluation and future downturns. Rapach, Ringgenberg, and Zhou (2016) document that aggregate Short Interest is among the strongest predictors of market-level stock returns, suggesting that elevated Short Interest captures critical negative information before other investors. However, short selling is often restricted by regulatory constraints and influenced by behavioral biases, creating ongoing debate about its implications for information efficiency and market behavior.

This study investigates how short-selling patterns incorporate information into asset prices using modern machine learning methods. Specifically, we ask: Do shorting patterns represent rational reactions to new information or reflect behavioral biases such as extrapolation, overconfidence, or prospect theory–type behaviors? Our research addresses this question by adapting recent methodologies from empirical asset pricing—specifically the influential approach of Gu, Kelly, and Xiu (2020)—replacing future returns with shorting activity as the dependent variable in machine learning prediction frameworks. While previous literatures have documented the predictive power of Short Interest using traditional econometric approaches, we extend this literature by leveraging these cutting-edge ML techniques to capture complex interactions that our standard models may miss.

Our main contribution to existing literature is twofold. First, we integrate machine learning with the analysis of short-selling activity to provide fresh insights into market efficiency and potential biases in investor behavior. Second, we will explore whether predictions derived from these advanced models enhance the understanding of how Short Interest translates into stock return predictability. By leveraging a large dataset of firm-specific characteristics and macroeconomic indicators, we aim to uncover nuanced, nonlinear relationships between short-selling signals and market outcomes – for example, perhaps Short Interest combined with certain accounting ratios or macro variables, which could have especially strong predictive power that only emerges in a nonlinear model.

2 Related Literature

2.1 Short Selling and Market Efficiency

Short selling has long been theorized to improve market efficiency by allowing pessimistic views and negative information to be reflected in prices. In an unconstrained market, prices should incorporate the average of investors' opinions, both optimistic and pessimistic. However, classic theory by Miller (1977) argues that when short-sale constraints exist, prices will be biased upward because only the most optimistic investors influence the price. Pessimistic investors who believe a stock is overvalued may be unable to short sell freely, so their information and beliefs are not fully impounded into the stock's price. As a result, overvaluation can occur under short-sale constraints - prices are set by optimists and can exceed fundamental value. This "overvaluation hypothesis" implies that stocks with restricted shorting tend to become overpriced and subsequently earn abnormally low returns when reality corrects the initial optimism. In contrast, Diamond and Verrecchia (1987) offer a rational expectations perspective, suggesting that short-sale constraints slow down the incorporation of negative information but do not permanently bias prices. According to their model, prices will eventually reflect fundamentals even if shorting is limited, but the adjustment is delayed, implying a temporary inefficiency.

Empirical evidence generally supports the notion that short-selling activity enhances informational efficiency, while constraints or high costs to short can lead to mispricing. For example, Bris, Goetzmann, and Zhu (2007) document that markets with greater short-selling restrictions tend to experience more overpricing and slower price adjustments, consistent with Miller's hypothesis. Experimental studies likewise show that banning short sales causes prices to deviate systematically from fundamental values, whereas allowing shorting leads to more accurate, efficient pricing. From a behavioral standpoint, the absence of short sellers means that behavioral biases such as over-optimism and overconfidence among long-only investors can go unchecked. In exuberant markets, pessimistic traders serve an invaluable function by preventing prices from "rising too far", counteracting the collective optimism. If short sellers are shut out, optimistic sentiment can dominate, creating a disconnect between prices and intrinsic value. This dynamic was evident during episodes like the dot-com bubble, where short-sale constraints and sentiment-driven buying led to extreme valuations that later crashed. Thus, both theory and evidence underscore that shorting activity is intimately linked to price efficiency and can either mitigate or exacerbate mispricing depending on the presence of constraints.

Behavioral finance research also highlights how investor psychology interacts with short-sale limitations. When betting against stocks is difficult, investors holding irrationally positive views face little resistance, so cognitive biases can have a larger impact on prices. Short sellers, often castigated as "speculators" or "vandals" during market booms, are actually a stabilizing force: their transactions inject negative information and skepticism into the market, which can correct overly rosy valuations. In essence, short selling provides a mechanism for the market to discipline hype and hubris. Without it, prices may reflect a one-sided view. This insight aligns with Miller's (1977) argument and is supported by observations that stocks with greater divergence of opinion (disagreement among investors) tend to be especially prone to overvaluation when shorting is difficult. On the other hand, when short selling is allowed and actively used, it can help align prices with fundamentals, thereby improving information efficiency in the spirit of the Efficient Market Hypothesis. The tug-of-war between optimistic bias and short sellers' corrective force is a central theme in understanding how information and beliefs get translated into asset prices.

2.2 Short Interest and Stock Return Predictability

A substantial body of empirical work examines the informational content of Short Interest – the aggregate volume of short positions in a stock – for subsequent stock returns. Cross-sectional studies have consistently found a negative relationship between Short Interest levels and future stock performance. In other words, stocks with unusually high Short Interest tend to underperform going forward. Early evidence by Asquith, Pathak, and Ritter (2005) shows that stocks facing short-sale constraints (proxied by high Short Interest coupled with limited stock loan supply) significantly underperform their peers. They report that the most constrained stocks earn abnormally low returns – about 215 basis points per month lower (equal-weighted) than unconstrained stocks – over the 1988–2002 period. This large performance gap supports the view that heavy shorting pressure flags overvaluation: when many traders are betting against a stock and the ability to short is limited, it likely indicates fundamental negatives that will eventually drive the price down.

Likewise, Desai et al. (2002) and Boehmer, Jones, and Zhang (2008) find that Short Interest and actual short-selling activity contain significant predictive power for individual-stock returns. Stocks in the top decile of Short Interest often experience negative abnormal returns in the following months, whereas stocks with little or no Short Interest (which short sellers avoid, presumably due to no glaring overvaluation) perform relatively better. Such patterns have been labeled the "Short Interest anomaly" or "Short Interest puzzle" – referring to the puzzle of why this return predictability persists if the information is publicly available.

One explanation for the Short Interest anomaly is that short sellers are uniquely informed or skilled, allowing them to collectively foresee which stocks are overvalued before the wider market corrects the mispricing. Engelberg, Reed, and Ringgenberg (2012) provide direct evidence of short sellers' information advantage. They show that short sellers dramatically increase their positions in response to public news that has negative implications, and that the well-known negative relation between shorting and next-period returns is twice as large on days with news releases and four times as large on days with negative news. This suggests short sellers are adept at processing public information (such as earnings announcements or news about a company) and trading on it faster or more efficiently than other investors. Notably, Engelberg et al. find that the most informed short trades are not driven by market makers (who may short for liquidity reasons), but by other market participants (e.g. hedge funds or sophisticated traders) who analyze news and fundamentals.

Overall, short sellers appear to be skilled information processors, and their trading conveys information that the market has not yet fully digested. This contributes to the predictive power of Short Interest: rising Short Interest often precedes the revelation of bad news or fundamentals, aligning with the idea that short sellers anticipate future cash-flow problems or price corrections. Rapach, Ringgenberg, and Zhou (2016) reinforce this interpretation in an aggregate context. They document that Short Interest aggregated across the market is an exceptionally strong predictor of aggregate stock returns, both in-sample and out-of-sample. In their study, changes in market-wide Short Interest predicted a significant portion of variation in future equity market premiums, yielding out-of-sample R^2 statistics above 12%– markedly higher than other well-known predictors like the dividend yield or term spread. A vector autoregression analysis in that paper indicates that the predictive power of Short Interest stems mainly from a cash-flow news channel, meaning short sellers collectively anticipate downturns in fundamentals (e.g. lower future earnings or dividends) which then lead to lower market returns. This finding portrays short sellers as informed traders at the macro level, not just stock-specific pessimists.

Importantly, the Short Interest-return relation highlights a tension between information efficiency and market frictions. If Short Interest is public (exchanges publish Short Interest data periodically), one might expect arbitrageurs to immediately trade on that information and erase any predictability in returns. Yet, the fact that Short Interest consistently forecasts returns – the Short Interest anomaly – implies the presence of limits to arbitrage or other frictions. One key friction is the cost and risk of shorting itself. Short sellers face constraints like loan fees (the cost of borrowing shares), the risk of recall (having to return borrowed shares on short notice), and potentially unlimited losses if the stock rises. These factors can deter even rational arbitrageurs from correcting mispricing.

Recent research by Engelberg, Reed, and Ringgenberg (2018) on "short-selling risk" shows that when the risk or uncertainty of staying in a short position is high, traders are less willing to short, and as a result mispricings persist longer. In their analysis, higher uncertainty in borrowing costs leads to significantly lower short-selling activity and greater price deviations from fundamentals. This helps explain why the predictive power of Short Interest is not arbitraged away: Short Interest may be publicly known, but not everyone can or will act on it due to the risks and costs involved. In sum, heavily shorted stocks often remain overpriced because many investors are either unaware of the information short sellers have, or unable/unwilling to trade on it aggressively (the so-called limits of arbitrage). This line of reasoning connects to behavioral finance as well – even if some investors recognize an overvaluation (the short sellers), others might ride the momentum due to exuberance or mandate constraints (e.g. long-only funds), allowing the mispricing to continue temporarily. Only over time, as fundamentals deteriorate or news becomes undeniable, do prices converge to reflect true value, rewarding those who shorted earlier. Thus, Short Interest serves as a harbinger of correction, and understanding it is crucial for anyone interested in return predictability and market anomalies.

2.3 Machine Learning and Empirical Asset Pricing

In recent years, machine learning (ML) techniques have been increasingly applied to finance, offering powerful tools to detect patterns and interactions in large datasets that traditional models might miss. Asset pricing researchers have begun to harness ML algorithms to improve the forecasting of stock returns and to better understand the pricing of risk. Gu, Kelly, and Xiu (2020) provide a seminal study in this domain, conducting an extensive comparison of various ML methods (including penalized regressions, tree-based models, random forests, and neural networks) for predicting stock returns. They demonstrate that ML-based approaches can deliver substantial gains in predictive accuracy and investor utility relative to conventional linear models. In their analysis, the best-performing methods (such as boosted trees and deep neural networks) roughly double the out-of-sample performance of leading regression-based strategies from the prior literature. These improvements are attributed to ML's ability to capture nonlinear relationships and variable interactions that are difficult to model with standard linear regressions.

3 Data

Our sample consists of all firms listed in the NYSE beginning in January 1973 and ending in December 2017, totaling 44 years. We obtain short interest by dividing the shares held short variable from Compustat by the total shares outstanding from CRSP. The periodicity of short interest is monthly. From CRSP, we also obtain stock returns, market equity, and other relevant variables. From Compustat, we can obtain data on firm fundamentals through accounting indicators. We then construct a large array of firm-level predictor variables and lag them to avoid look-ahead bias. We will also incorporate macro-finance predictors that are used in return-prediction papers and may be relevant in showing short-seller behavior rationality. We directly use the dataset compiled by Gu, Kelly, and Xiu (2020), which includes 98 firm characteristics, which also contains the periodicity of each variable.

Additionally, we gather information on the beginning of month price, market cap, and 20 percent of NYSE market cap. We lag these variables to maintain accuracy. We use these variables to filter our Short Interest dataset by filtering out firms which have a beginning of month price less than five dollars, and firms which are under the 20 percent of NYSE market cap. This results in our final dataset which contains the 94 variables listed (due to some columns having been constructed later than the start date of our dataset), the shorting volume (in percent), and identifier variables (PERMNO, YYYYMM). We also have outliers in our original shorting volume dataset pulled from WRDS, so we just filter out any entries that exceed 100% since it is unlikely (but not impossible) to have more shares being shorted than outstanding. Table 1 contains the summary statistics for shorting volume (SHTSHR). We use this final dataset to run all of our analysis (Table 2).

4 Methodology

We develop our methodology from Gu, Kelly, and Xiu (2020) in which we predict short volume per month. Suppose for each stock i, we have a vector $z_{i,t}$ of predictor variables (firm characteristics, macroeconomic indicators, etc.) and a target $s_{i,t+1}$ denoting the shorting volume over the next period. We want to approximate:

$$\mathbb{E}\big[s_{i,t+1} \mid z_{i,t}\big] = g\big(z_{i,t}\big),$$

where g captures complexities between the covariates and future shorting volume. When we forecast shorting volume, our signal-to-noise ratio can be low, just like returns. OLS regressions using all predictors is likely to fail due to the parameter space growing too large. Instead, we use methods that impose parameter parsimony: penalized linear regressions (Elastic Net, Lasso, Ridge); dimension reduction (PCR, PLS); and nonlinear methods (Regression Trees, ML, and NN).

To emulate the time series nature of our monthly shorting volume data, we will divide our data into a training, validation, and test window. The predictor set can also be similarly large to the return prediction proposed in Gu, Kelly, and Xiu (2020). Before we fit our model, we cross-sectionally rank-transform the variables to mitigate outliers and scale differences. On the result front, we want to minimize the MSE in predicting the next period shorting volume. We use robust loss functions to reduce the influence of outliers. Our primary performance metrics are the root MSE and R^2 relative to historical average shorting volume. We also analyze how these results align with events such as earnings announcements, short-squeeze indicators, etc. We show how each predictor affects our forecasts and the marginal effects of specific predictors.

4.1 ML Methods

4.1.1 Random Forest

Random Forest (RF) is a robust ensemble learning algorithm that combines multiple decision trees trained on bootstrap samples of the data. Each decision tree contributes independently, and the final prediction is typically obtained through averaging (for regression) or majority voting (for classification). RF is particularly advantageous for predicting shorting volume due to its capability to capture complex nonlinear relationships and interactions between a large number of financial predictors without overfitting. Additionally, RF effectively mitigates issues related to multicollinearity and is robust against outliers—common challenges encountered in financial market data. Formally, the Random Forest prediction function for Short Interest $s_{i,t+1}$ given predictor vector $z_{i,t}$ is defined as:

$$\hat{s}_{i,t+1}^{RF}(z_{i,t}) = \frac{1}{B}\sum_{b=1}^{B} T_b(z_{i,t})$$

where $T_b(z_{i,t})$ denotes the prediction from the b-th decision tree, and B is the total number of trees in the Random Forest ensemble.

Our implementation sets the number of trees to vary between 250 and 500, our tree depths vary between three and five, a learning rate of 0.01 or 0.1, and the we allow the trees to use all the features for maximum complexity.

4.1.2 XGBoost

XGBoost (eXtreme Gradient Boosting) is a powerful ensemble machine learning algorithm known for its efficiency, scalability, and strong predictive accuracy, particularly suitable for financial applications such as forecasting Short Interest. Similar to Random Forest, XGBoost constructs an ensemble of decision trees; however, it does so sequentially, where each new tree corrects the residual errors of previous trees. Its strengths lie in its sophisticated regularization mechanisms, optimized handling of large datasets, and parallelization, making it particularly adept at managing noisy, high-dimensional financial data.

Formally, the XGBoost prediction function for Short Interest $s_{i,t+1}$ given predictor vector $z_{i,t}$ is represented as an additive ensemble:

$$\hat{s}_{i,t+1}^{XGB}(z_{i,t}) = \sum_{k=1}^{K} f_k(z_{i,t}),$$

where $f_k(z_{i,t})$ denotes the prediction from the k-th boosted decision tree, and K represents the total number of trees sequentially added to the model.

In our implementation, the number of boosted trees K ranges from 250 to 500, tree depths vary between one, three, and five, and we use all available features to maximize predictive complexity while controlling overfitting through built-in regularization parameters (L1 and L2 regularization terms).

4.1.3 Neural Network

Neural Networks (NN) are powerful machine learning algorithms capable of capturing highly nonlinear and complex interactions between predictors, making them particularly suitable for predicting financial outcomes such as shorting volume. Unlike traditional linear models, NNs explicitly construct nonlinear mappings through interconnected layers of neurons, allowing them to adapt flexibly to subtle data patterns.

Our implemented neural network is a feedforward multilayer perceptron (MLP) with multiple hidden layers, ReLU activations, and dropout regularization to prevent overfitting. Formally, the output of each neuron j in hidden layer l is defined as:

$$x_j^{(l)} = f\left(\theta_{j,0}^{(l-1)} + \sum_{k=1}^{n_{l-1}} x_k^{(l-1)} \theta_{j,k}^{(l-1)}\right),$$

where: $x_j^{(l)}$ is the output of neuron j in hidden layer l; $f(\cdot) = \text{ReLU}(\cdot) = \max(0, \cdot)$ is the activation function; $\theta_{j,k}^{(l-1)}$ are the parameters (weights) connecting neuron k in layer $l-1$

to neuron j in layer l; and n_{l-1} is the number of neurons in the previous layer $(l-1)$. For example, the second neuron in the first hidden layer transforms inputs into an output as follows:

$$x_2^{(1)} = f\left(\theta_{2,0}^{(0)} + \sum_{j=1}^{p} z_j\, \theta_{2,j}^{(0)}\right),$$

where z_j denotes the j-th predictor, and p is the total number of predictors in the input vector $z_{i,t}$. Finally, the outputs from the last hidden layer (L) are aggregated linearly into an ultimate short-volume prediction:

$$\hat{s}_{i,t+1}^{NN}(z_{i,t}) = \theta_0^{(L)} + \sum_{j=1}^{n_L} x_j^{(L)} \theta_j^{(L)}.$$

Explicitly, our neural network has the following structure: an input layer with the number of neurons equal to number of predictors; hidden layers starting with 64 or 32 neurons (tuned via validation) and decaying to 8 neurons by the fourth layer; and an output layer with a single neuron predicting our desired Short Interest. We incorporate dropout regularization between hidden layers, randomly dropping neuron activations with a dropout rate between 0.1 and 0.2. The network is trained using the Adam optimizer to minimize Mean Squared Error (MSE) over 500 epochs, with a learning rate of 0.01.

4.1.4 Performance Evaluation

Model performance is systematically evaluated through several key metrics. Primarily, we focus on predictive accuracy using the out-of-sample adjusted R^2, measuring how well our models predict future shorting volume relative to the historical mean. It is defined by:

$$R_{OOS}^2 = 1 - \frac{\sum_{i,t}(s_{i,t+1} - \hat{s}_{i,t+1})^2}{\sum_{i,t}(s_{i,t+1} - \bar{s}_{t+1})^2}$$

where: $s_{i,t+1}$ is the observed shorting volume for stock i at time $t+1$; $\hat{s}_{i,t+1}$ is the predicted shorting volume from the model; and $\bar{s}_{t+1}$ is the historical average shorting volume up to time t. Models exhibiting higher R^2 demonstrate superior explanatory power. Additionally, we rigorously assess the statistical significance of predictors and predicted short volume, employing Fama-MacBeth regressions and computing Newey-West adjusted t-statistics to account for serial correlation and heteroscedasticity.

4.1.5 Regression Models

We first estimate predictive regression models for short interest to establish baseline results regarding our predicted short interest. Our specification regresses next-month short interest on the current short interest (with and without the lag in the training data):

$$SI_{i,t+1} = \alpha + \beta_{SI}\, SI_{i,t} + \varepsilon_{i,t+1}, \tag{1}$$

$$SI_{i,t+1} = \alpha + \beta_{SI}\, SI_{i,t} + \beta_{SI}^{(1)}\, SI_{i,t-1} + \varepsilon_{i,t+1}, \tag{2}$$

This helps us understand how well predictions of short interest capture the persistence (private information) in actual shorting behavior and improve forecasting accuracy.

Then, we estimate predictive regression models for stock returns to establish baseline results regarding short interest. Our basic specification regresses next-month return on the current short interest ratio and controls:

$$r_{i,t+1} = \alpha + \beta_{SI}\, SI_{i,t} + \beta_Z^\top Z_{i,t} + \varepsilon_{i,t+1}, \tag{3}$$

where $r_{i,t+1}$ is the excess return for stock i in month $t+1$, $SI_{i,t}$ is the short interest ratio (in decimal form, so $0.10 = 10\%$ of float) for stock i at the end of month t, and $Z_{i,t}$ denotes a vector of control variables (such as momentum, size, value, etc., as described in Section 2) for stock i at time t. The coefficient β_{SI} captures the predictive effect of short interest on next-month returns, holding other factors constant. We expect β_{SI} to be negative if high short interest predicts lower future returns (i.e. short sellers are correct on average).

In some specifications, we include lagged short interest to assess how long the short interest signal persists:

$$r_{i,t+1} = \alpha + \beta_{SI}^{(0)}\, SI_{i,t} + \beta_{SI}^{(1)}\, SI_{i,t-1} + \beta_Z^\top Z_{i,t} + \varepsilon_{i,t+1}. \tag{4}$$

Here $SI_{i,t-1}$ is the short interest from one month prior (lagged one period further), with coefficient $\beta_{SI}^{(1)}$. If $\beta_{SI}^{(1)}$ is near zero once current $SI_{i,t}$ is included, it would indicate that the impact of short interest is mostly realized within one month and does not carry over beyond that. On the other hand, a significant $\beta_{SI}^{(1)}$ might suggest a more prolonged effect or a delayed reaction by the market to short interest information.

We estimate the regressions using panel data across all stocks and months in the sample. To account for the strong contemporaneous correlation in stock returns (e.g., market-wide movements affecting many stocks in the same month), we include month fixed effects in some specifications. However, including month dummies is equivalent to de-meaning returns by the market's return each month, which is similar to using excess returns as we do; thus, our excess return specification already focuses on idiosyncratic performance. We do not include stock fixed effects because our goal is to exploit cross-sectional differences (a stock's short interest relative to others at a given time), and including stock fixed effects would absorb all time-invariant differences in average returns.

For inference, we report t-statistics based on robust standard errors clustered by month. Clustering by month addresses the cross-sectional dependence issue by effectively treating each month's cross-section as one cluster. This adjustment, which is analogous to a Fama-MacBeth two-step approach (averaging monthly regression slopes and using their time-series variability to gauge significance), ensures that our statistical significance is not overstated due to many stocks moving together. In all regression tables, t-statistics (in parentheses) are based on these month-clustered standard errors.

4.2 Training Approaches

One-Shot Approach

The one-shot analysis complements our rolling-window results by fitting the Random Forest model once on a static training and validation set (comprising 60% of the data) and subsequently evaluating predictive performance on the remaining 40% test set. Similar to the rolling approach, standardization of predictor variables utilizes statistics derived solely from the training-validation set. We perform hyperparameter optimization via TimeSeriesSplit cross-validation, selecting the optimal model configuration based on the lowest validation mean squared error (MSE).

Rolling-Window Approach

The rolling-window approach partitions the data sequentially into training, validation, and testing sets. Specifically, we employ a training window of 48 months up to 60 months, followed by a 12-month validation window, and finally a 1, 6, 12, or 24-month test window. These windows are iteratively advanced throughout the dataset. Within each iteration, predictor variables are standardized using the training data only to avoid look-ahead bias. Variables are appropriately lagged to avoid lookahead bias. Hyperparameter tuning is conducted via TimeSeriesSplit cross-validation within the combined training and validation dataset, ensuring temporal consistency. The out-of-sample model performance is evaluated using the root mean squared error (RMSE) and R^2 metrics, reflecting predictive accuracy and explanatory power.

Historical Window Approach

In addition to the above methodologies, we implement a historical window approach, which incrementally expands the training window as more historical data becomes available after each test period. Specifically, each year of testing data is preceded by all available historical data up to that point, thus reflecting the real-world scenario wherein an investor uses all past information when making decisions. This approach differs from the rolling-window by continuously enlarging the training set rather than moving it forward in time. Similar to previous approaches, we standardize predictors based only on historical data available at each prediction point, ensuring no future data leakage. Hyperparameter tuning and model validation are performed using TimeSeriesSplit cross-validation. The historical window approach enables us to leverage the increasing availability of historical data, potentially enhancing predictive accuracy and stability over time.

5 Results

5.1 Random Forest One-Shot and SHAP Analysis

Our one-shot Random Forest analysis achieved notable (not for the right reasons) predictive accuracy with a low RMSE and negative R^2 on the test set, highlighting the model's overfitting issues when considering data without rolling windows. To interpret predictor importance, we apply SHAP (SHapley Additive exPlanations), providing a nuanced understanding of variable contributions to Short Interest predictions.

The SHAP summary plot (Figure 1) reveals share turnover ('turn'), illiquidity ('ill'), and dollar volume traded ('dolvol') as the most influential predictors of Short Interest. High turnover and dollar volume are positively correlated with increased Short Interest, reflecting short sellers' preference for liquidity, while high illiquidity tends to decrease short-selling activity, likely due to higher trading frictions.

Other key predictors include volatility of turnover ('std_turn'), tangibility ('tang'), and firm age ('age'). High volatility in turnover signals greater uncertainty, thus attracting short sellers, while lower asset tangibility is associated with increased perceived risk, also elevating Short Interest. Younger firms are more frequently targeted, suggesting greater informational asymmetry and perceived mispricing.

The SHAP analysis further highlights the nuanced effects of research and development measures ('rd_sale' and 'rd_mve'), with increased R&D intensity typically associated with reduced Short Interest, indicative of investors interpreting these investments positively.

5.2 Rolling and Historical Window Models

In Table 2, we show results from a comparative analysis of various rolling and historical data-based Random Forest and XGBoost models to determine the optimal specification for predicting Short Interest, ultimately guiding our regression analysis. Parameters from different model specifications show varying predictive accuracy, as measured by their respective R^2 values. Among the models evaluated, the 12-month rolling Random Forest model with 60 months of training and the 12-month historical data XGBoost model emerged as particularly notable due to their higher predictive performance and statistical significance.

The enhanced predictive accuracy from these models suggests that a balanced approach in training size and historical window effectively captures the evolving market dynamics influencing Short Interest. The statistical significance observed further justifies their selection, indicating robustness in identifying meaningful relationships within the data.

Given these insights, we decided to utilize the 12-month rolling Random Forest model with 60 months of training and the 12-month historical data XGBoost model in our regressions. This decision is driven by the objective of leveraging statistically significant models capable of robustly capturing short-term shifts in market conditions, thereby ensuring the subsequent regression analysis is grounded in somewhat reliable predictive inputs.

5.3 Short Interest and Predicted Short Interest Regressions

Table 3 presents the results from regressions that explain next month's short interest (SI_{t+1}) using current short interest (SI_t) and predicted short interest ($\widehat{SI}_t$). Columns (1) through (3) compare the specifications without lagged short interest, while columns (4) through (6) include lagged short interest in the training of the prediction model.

In the no-lag specification, current short interest (SI_t) strongly predicts next month's short interest with a coefficient of 0.918 (column 1) and remains highly significant when both actual and predicted short interest are included (column 3). Predicted short interest alone also has strong predictive power (0.867 in column 2). The adjusted R^2 values indicate that actual short interest alone (0.933) slightly outperforms the combination of actual and predicted short interest (0.934).

When lagged short interest is included in the ML model (columns 4-6), predicted short interest becomes more informative, achieving a coefficient of 0.908 alone (column 5) and remains statistically significant when combined with actual short interest (0.163, column 6). However, the coefficient on actual short interest drops to 0.777, suggesting predicted short interest captures part of its explanatory power. The R^2 slightly improves (0.935) when combining both actual and predicted short interest.

5.4 Short Interest and Stock Return Regressions

Table 4 explores the predictive power of short interest and predicted short interest for next month's stock returns (R_{t+1}). Columns (1)-(3) present results without lagged short interest, and columns (4)-(6) include lagged short interest in the ML training.

Actual short interest negatively predicts future returns consistently across all specifications, with significant negative coefficients of -0.0476 (column 1) and -0.0523 (column 3). Predicted short interest alone also negatively forecasts returns (-0.0285, column 2), but when combined with actual short interest, its effect diminishes and becomes insignificant (0.0178, column 3).

Including lagged short interest in the ML training (columns 4-6) yields similar results. Actual short interest remains negatively significant (-0.0461, column 4; -0.0272, column 6), while predicted short interest is significantly negative alone (-0.0478, column 5) but loses significance when combined (-0.0217, column 6). The R^2 values remain stable around 0.185, suggesting limited incremental predictive ability from predicted short interest.

Table 5 directly contrasts actual next-month short interest (SI_{t+1}) and predicted short interest ($\widehat{SI}_t$) in forecasting returns. Without lagged short interest, actual next-period short interest significantly predicts lower returns (-0.0291, column 1; -0.0280, column 3), while predicted short interest alone significantly forecasts negative returns (-0.0285, column 2) but loses significance when combined (-0.0042, column 3).

With lagged short interest, the sign and significance notably shift. Actual next-period short interest shows a significant positive coefficient (0.0489, column 6), indicating a potential reversal or informational interpretation when predicted short interest is considered. Predicted short interest strongly predicts negative returns alone (-0.0478, column 5) and retains substantial predictive power combined with actual short interest (-0.0922, column 6).

Overall, actual short interest remains the most consistent predictor of negative future returns. Predicted short interest adds informational value primarily when trained on lagged data and appears valuable in identifying systematic market expectations rather than capturing purely private information.

5.5 Economic Insights

The finding that actual short interest has stronger predictive power for stock returns than a machine-learned prediction suggests that short sellers' real-time positions carry unique information that a model's expected value cannot fully capture. In other words, the actual short interest observed in the market reflects the convictions of informed traders, whereas the model-predicted short interest reflects only what one would anticipate based on publicly known factors.

Our prediction represents a synthesized expectation of short interest based on observable inputs (fundamentals, technical indicators, macro variables). One can think of it as the level of short interest one would predict if all market participants used a very sophisticated, data-driven model. If a stock has declining earnings, poor credit indicators, high valuations, and negative momentum, both human analysts and a machine-learning model would expect higher short interest. But if those problems are already public, much of that expectation is baked into the stock's price. Hence, the predicted component may simply capture known risk factors rather than mispricing, explaining its weaker return-predictive power.

In contrast, deviations of actual short interest from its prediction are likely driven by idiosyncratic, unmodeled reasons: qualitative insights (e.g. proprietary research, fraud rumors, sector-specific information) or nonlinear combinations of factors that even a complex model misses. Economically, predicted short interest serves as a proxy for investor expectations—what the broader market thinks SI_t "should" be given all obvious data. When actual short interest diverges upward from this baseline, it signals that informed traders collectively perceive more downside risk than one would infer from public metrics alone. Such a surprise often precipitates price declines as the market updates on new information.

Notably, even gradient-boosted trees—powerful at uncovering complex patterns—fall short of capturing this divergence, indicating that short sellers' advantage stems not only from past data but from forward-looking judgment: news, rumors, and investor interpretation. Indeed, when periodic short interest disclosures exceed expectations, the market reacts sharply, aligning with the notion that actual short interest embeds unexpected, price-moving information that a backward-looking model cannot fully anticipate. Even when including the lagged short interest in training, we find that the return predictability does not improve, suggesting that short interest is persistent.

5.5.1 Extension to Stock Returns

When we include the ML-predicted short interest $\widehat{SI}_t$ as an additional regressor in our Fama–MacBeth return regressions, the coefficient on $\widehat{SI}_t$ is economically small and statistically insignificant. This indicates that the nonlinearities and higher-order interactions captured by the ML model add almost no incremental return-predictive power beyond established linear effects (momentum, size, value, volatility, and actual short interest). Another indication here is that the information used by short traders is based on private information rather than purely public information.

6 Next Steps

We have a meeting scheduled with Professor Matthew Ringgenberg (Utah), to discuss our results and see where we can improve and tweak the framework to gain some more significant results. In addition, we intend to expand the dimensionality of our predictor set by incorporating a broader universe of firm characteristics and macro-financial indicators. Specifically, we will utilize the Open Source Asset Pricing (OSAP) dataset, which includes over 200 standardized predictors commonly used in asset pricing literature. This expansion will allow us to investigate whether the inclusion of more nuanced signals can uncover latent structures or nonlinear relationships relevant to short interest prediction. By increasing the richness of our input space, we hope to improve the signal-to-noise ratio and test the robustness of our model architectures to larger, more diverse feature sets.

Lastly, once all predictive models and datasets are finalized, we will evaluate the practical utility of our predicted shorting volume by constructing and analyzing long-short portfolios. These portfolios will be formed by taking long positions in stocks with low predicted shorting activity and short positions in stocks with high predicted shorting volume. Performance metrics such as average returns, Sharpe ratios, and alphas relative to standard factor models will be computed. This analysis will allow us to assess whether the information extracted by our machine learning models has not only statistical but also economic value in forming profitable trading strategies.

References

Asquith, Paul, Parag A. Pathak, and Jay R. Ritter. 2005. "Short Interest, Institutional Ownership, and Stock Returns." *Journal of Financial Economics* 78 (2): 243–276.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang. 2008. "Which Shorts Are Informed?" *Journal of Finance* 63 (2): 491–527.

Bris, Arturo, William N. Goetzmann, and Ning Zhu. 2007. "Efficiency and the Bear: Short Sales and Markets Around the World." *Journal of Finance* 62 (3): 1029–1079.

Diamond, Douglas W., and Robert E. Verrecchia. 1987. "Constraints on Short-selling and Asset Price Adjustment to Private Information." *Journal of Financial Economics* 18 (2): 277–311.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg. 2012. "How Are Shorts Informed? Short Sellers, News, and Information Processing." *Journal of Financial Economics* 105 (2): 260–278.

Engelberg, Joseph E., Adam V. Reed, and Matthew C. Ringgenberg. 2018. "Short-selling Risk." *Journal of Finance* 73 (2): 755–786.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *Review of Financial Studies* 33 (5): 2223–2273.

Miller, Edward M. 1977. "Risk, Uncertainty, and Divergence of Opinion." *Journal of Finance* 32 (4): 1151–1168.

Rapach, David E., Matthew C. Ringgenberg, and Guofu Zhou. 2016. "Short Interest and Aggregate Stock Returns." *Journal of Financial Economics* 121 (1): 46–65.

Wang, Xue, Xuemin Yan, and Lingling Zheng. 2020. "Shorting Flows, Public Disclosure, and Market Efficiency." *Journal of Financial Economics* 135 (1): 191–212.

Figures and Tables

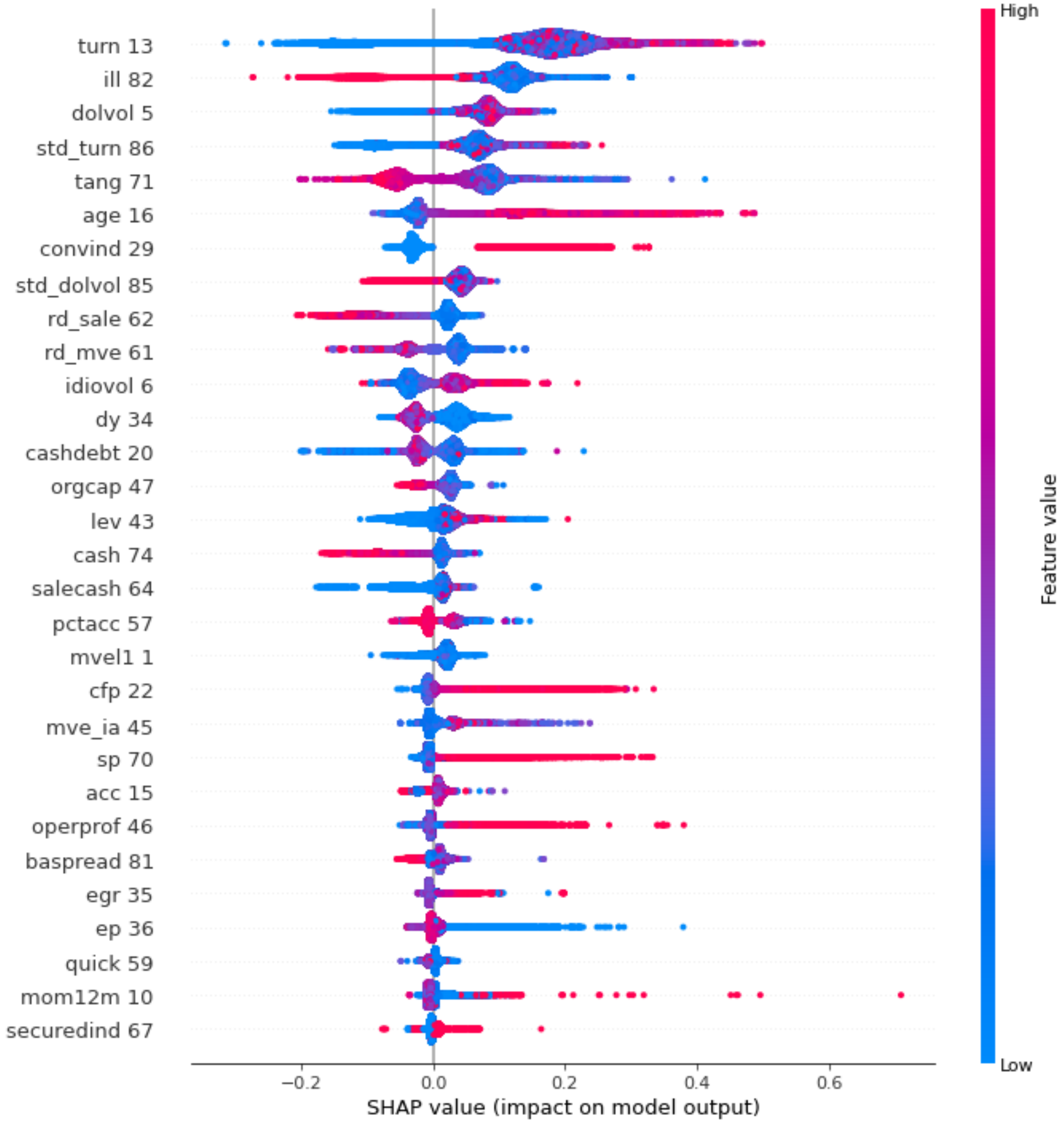

Figure 1: SHAP Values for Random Forest Model

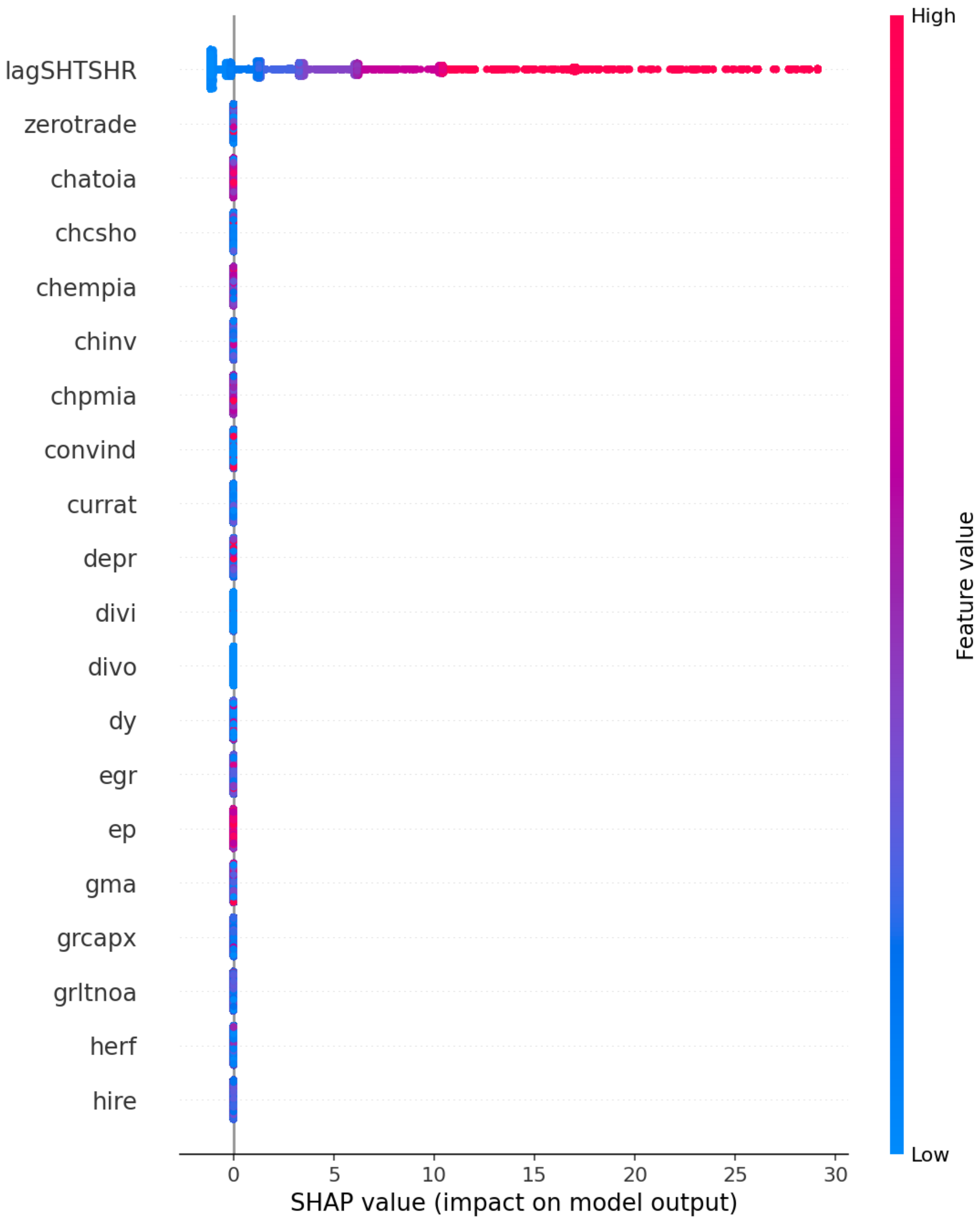

Figure 2: SHAP Values for Random Forest Model (with lag)

Table 1: Summary Statistics for Short Interest

Statistic	**Value**
Observations	753,248
Mean	2.957
Standard Deviation	19.861
Minimum	0.000
25th Percentile	0.069
Median	0.881
75th Percentile	3.363
Maximum	4,373.080

Table 2: Summary Statistics for ML Models

Training Parameters	Beta	N-W	$adj.R^2$
1 Month Rolling Random Forest	-0.0017	-1.01	0.0160
12 Month Rolling Random Forest	-0.0023	-1.28	0.0166
12 Month Rolling Random Forest (more validation)	-0.0032	-1.57	0.0156
12 Month Rolling Random Forest (more training)	-0.0015	-1.90	0.0094
6 Month Historical Data Random Forest	-0.0028	-1.52	0.0147
12 Month Historical Data Random Forest	-0.0028	-1.53	0.0148
24 Month Historical Data Random Forest	-0.0026	-1.37	0.0152
12 Month Historical Data XGBoost	-0.0015	-1.91	0.0100

Note: We use winsorized returns at the top and bottom 1% in the Fama MacBeth regression.

Table 3: Explaining Short Interest SI_{t+1}

	No Lag			**With Lag**		
	(1)	(2)	(3)	(1)	(2)	(3)
SI_t	0.918***		0.900***	0.918***		0.777***
	(0.0054)		(0.0069)	(0.0053)		(0.0322)
$\widehat{SI}_t$		0.867***	0.069***		0.908***	0.163***
		(0.0148)	(0.0061)		(0.0036)	(0.0320)
Constant	0.247***	0.704***	0.131***	0.244***	0.319***	0.196***
	(0.0151)	(0.0357)	(0.0068)	(0.0144)	(0.0097)	(0.0047)
Observations	717,608	717,608	717,608	694,637	694,637	694,637
R^2	0.933	0.684	0.934	0.934	0.896	0.935

* $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 4: Explaining Returns R_{t+1} (percentage-point units)

	No Lag			**With Lag**		
	(1)	(2)	(3)	(1)	(2)	(3)
SI_t	-0.0476^{***}		-0.0523^{***}	-0.0461^{***}		-0.0272^{**}
	(0.0060)		(0.0065)	(0.0060)		(0.0129)
$\widehat{SI}_t$		-0.0285^{***}	0.0178		-0.0478^{***}	-0.0217
		(0.0106)	(0.0118)		(0.0062)	(0.0132)
Constant	0.962^{***}	0.898^{***}	0.932^{***}	0.952^{***}	0.955^{***}	0.959^{***}
	(0.0161)	(0.0254)	(0.0258)	(0.0162)	(0.0164)	(0.0168)
Observations	717,608	717,608	717,608	694,637	694,637	694,637
R^2	0.183	0.183	0.183	0.185	0.185	0.185

* p<0.10, ** p<0.05, *** p<0.01

Table 5: Predicting Returns R_{t+1} Using Future SI vs. Predicted SI

	No Lag			**With Lag**		
	(1)	(2)	(3)	(1)	(2)	(3)
SI_{t+1}	-0.0291^{***}		-0.0280^{***}	-0.0263^{***}		0.0489^{***}
	(0.0060)		(0.0067)	(0.0061)		(0.0120)
$\widehat{SI}_t$		-0.0285^{***}	-0.0042		-0.0478^{***}	-0.0922^{***}
		(0.0110)	(0.0120)		(0.0061)	(0.0120)
Constant	0.911^{***}	0.898^{***}	0.918^{***}	0.899^{***}	0.955^{***}	0.939^{***}
	(0.0170)	(0.0250)	(0.0260)	(0.0170)	(0.0160)	(0.0170)
Observations	717,608	717,608	717,608	694,637	694,637	694,637
R^2	0.183	0.183	0.183	0.185	0.185	0.185

* p<0.10, ** p<0.05, *** p<0.01

Table 6: Table A.6: Details of the Characteristics

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
1	absacc	Absolute accruals	Bandyopadhyay, Huang & Wirjanto	2010, WP	Compustat	Annual
2	acc	Working capital accruals	Sloan	1996, TAR	Compustat	Annual
3	aeavol	Abnormal earnings announcement volume	Lerman, Livnat & Mendenhall	2007, WP	Compustat+CRSP	Quarterly
4	age	# years since first Compustat coverage	Jiang, Lee & Zhang	2005, RAS	Compustat	Annual
5	agr	Asset growth	Cooper, Gulen & Schill	2008, JF	Compustat	Annual
6	baspread	Bid-ask spread	Amihud & Mendelson	1989, JF	CRSP	Monthly
7	beta	Beta	Fama & MacBeth	1973, JPE	CRSP	Monthly
8	betasq	Beta squared	Fama & MacBeth	1973, JPE	CRSP	Monthly
9	bm	Book-to-market	Rosenberg, Reid & Lanstein	1985, JPM	Compustat+CRSP	Annual
10	bm_ia	Industry-adjusted book to market	Asness, Porter & Stevens	2000, WP	Compustat+CRSP	Annual
11	cash	Cash holdings	Palazzo	2012, JFE	Compustat	Quarterly
12	cashdebt	Cash flow to debt	Ou & Penman	1989, JAE	Compustat	Annual
13	cashpr	Cash productivity	Chandrashekar & Rao	2009, WP	Compustat	Annual
14	cfp	Cash flow to price ratio	Desai, Rajgopal & Venkatachalam	2004, TAR	Compustat	Annual
15	cfp_ia	Industry-adjusted cash flow to price ratio	Asness, Porter & Stevens	2000, WP	Compustat+CRSP	Annual
16	chatoia	Industry-adjusted change in asset turnover	Soliman	2008, TAR	Compustat	Annual
17	chcsho	Change in shares outstanding	Pontiff & Woodgate	2008, JF	Compustat	Annual
18	chempia	Industry-adjusted change in employees	Asness, Porter & Stevens	1994, WP	Compustat	Annual
19	chinv	Change in inventory	Thomas & Zhang	2002, RAS	Compustat	Annual
20	chmom	Change in 6-month momentum	Gettleman & Marks	2006, WP	CRSP	Monthly
21	chpmia	Industry-adjusted change in profit margin	Soliman	2008, TAR	Compustat	Annual
22	chtx	Change in tax expense	Thomas & Zhang	2011, JAR	Compustat	Quarterly
23	cinvest	Corporate investment	Titman, Wei & Xie	2004, JFQA	Compustat	Quarterly
24	convind	Convertible debt indicator	Valta	2016, JFQA	Compustat	Annual
25	currat	Current ratio	Ou & Penman	1989, JAE	Compustat	Annual
26	depr	Depreciation / PP&E	Holthausen & Larcker	1992, JAE	Compustat	Annual

Continued on next page

Table 6 – *Continued from previous page*

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
27	divi	Dividend initiation	Michaely, Thaler & Womack	1995, JF	Compustat	Annual
28	divo	Dividend omission	Michaely, Thaler & Womack	1995, JF	Compustat	Annual
29	dolvol	Dollar trading volume	Chordia, Subrahmanyam & Anshuman	2001, JFE	CRSP	Monthly
30	dy	Dividend to price	Litzenberger & Ramaswamy	1982, JF	Compustat	Annual
31	ear	Earnings announcement return	Kishore, Brandt, Santa-Clara & Venkatachalam	2008, WP	Compustat+CRSP	Quarterly
32	egr	Growth in common shareholder equity	Richardson, Sloan, Soliman & Tuna	2005, JAE	Compustat	Annual
33	ep	Earnings to price	Basu	1977, JF	Compustat	Annual
34	gma	Gross profitability	Novy-Marx	2013, JFE	Compustat	Annual
35	grCAPX	Growth in capital expenditures	Anderson & Garcia-Feijoo	2006, JF	Compustat	Annual
36	grltnoa	Growth in long term net operating assets	Fairfield, Whisenant & Yohn	2003, TAR	Compustat	Annual
37	herf	Industry sales concentration	Hou & Robinson	2006, JF	Compustat	Annual
38	hire	Employee growth rate	Bazdresch, Belo & Lin	2014, JPE	Compustat	Annual
39	idiovol	Idiosyncratic return volatility	Ali, Hwang & Trombley	2003, JFE	CRSP	Monthly
40	ill	Illiquidity	Amihud	2002, JFM	CRSP	Monthly
41	indmom	Industry momentum	Moskowitz & Grinblatt	1999, JF	CRSP	Monthly
42	invest	Capital expenditures and inventory	Chen & Zhang	2010, JF	Compustat	Annual
43	lev	Leverage	Bhandari	1988, JF	Compustat	Annual
44	lgr	Growth in long-term debt	Richardson, Sloan, Soliman & Tuna	2005, JAE	Compustat	Annual
45	maxret	Maximum daily return	Bali, Cakici & Whitelaw	2011, JFE	CRSP	Monthly
46	mom12m	12-month momentum	Jegadeesh	1990, JF	CRSP	Monthly
47	mom1m	1-month momentum	Jegadeesh & Titman	1993, JF	CRSP	Monthly
48	mom36m	36-month momentum	Jegadeesh & Titman	1993, JF	CRSP	Monthly
49	mom6m	6-month momentum	Jegadeesh & Titman	1993, JF	CRSP	Monthly
50	ms	Financial statement score	Mohanram	2005, RAS	Compustat	Quarterly
51	mvel1	Size	Banz	1981, JFE	CRSP	Monthly
52	mve_ia	Industry-adjusted size	Asness, Porter & Stevens	2000, WP	Compustat	Annual
53	nincr	Number of earnings increases	Barth, Elliott & Finn	1999, JAR	Compustat	Quarterly
54	operprof	Operating profitability	Fama & French	2015, JFE	Compustat	Annual

Continued on next page

Table 6 – *Continued from previous page*

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
55	orgcap	Organizational capital	Eisfeldt & Papanikolaou	2013, JF	Compustat	Annual
56	pchcapx_ia	Industry adjusted % change in capital expenditures	Abarbanell & Bushee	1998, TAR	Compustat	Annual
57	pchcurrat	% change in current ratio	Ou & Penman	1989, JAE	Compustat	Annual
58	pchdepr	% change in depreciation	Holthausen & Larcker	1992, JAE	Compustat	Annual
59	pchgm_pchsale	% change in gross margin - % change in sales	Abarbanell & Bushee	1998, TAR	Compustat	Annual
60	pchquick	% change in quick ratio	Ou & Penman	1989, JAE	Compustat	Annual
61	pchsale_pchinvt	% change in sales - % change in inventory	Abarbanell & Bushee	1998, TAR	Compustat	Annual
62	pchsale_pchrect	% change in sales - % change in A/R	Abarbanell & Bushee	1998, TAR	Compustat	Annual
(Page break in the original text; continuing here in a single table.)						
63	pchsale_pchxsga	% change in sales - % change in SG&A	Abarbanell & Bushee	1998, TAR	Compustat	Annual
64	pchsaleinv	% change sales-to-inventory	Ou & Penman	1989, JAE	Compustat	Annual
65	pctacc	Percent accruals	Hafzalla, Lundholm & Van Winkle	2011, TAR	Compustat	Annual
66	pricedelay	Price delay	Hou & Moskowitz	2005, RFS	CRSP	Monthly
67	ps	Financial statements score	Piotroski	2000, JAR	Compustat	Annual
68	quick	Quick ratio	Ou & Penman	1989, JAE	Compustat	Annual
69	rd	R&D increase	Eberhart, Maxwell & Siddique	2004, JF	Compustat	Annual
70	rd_mve	R&D to market capitalization	Guo, Lev & Shi	2006, JBFA	Compustat	Annual
71	rd_sale	R&D to sales	Guo, Lev & Shi	2006, JBFA	Compustat	Annual
72	realestate	Real estate holdings	Tuzel	2010, RFS	Compustat	Annual
73	retvol	Return volatility	Ang, Hodrick, Xing & Zhang	2006, JF	CRSP	Monthly
74	roaq	Return on assets	Balakrishnan, Bartov & Faurel	2010, JAE	Compustat	Quarterly
75	roavol	Earnings volatility	Francis, LaFond, Olsson & Schipper	2004, TAR	Compustat	Quarterly
76	roeq	Return on equity	Hou, Xue & Zhang	2015, RFS	Compustat	Quarterly
77	roic	Return on invested capital	Brown & Rowe	2007, WP	Compustat	Annual
78	rsup	Revenue surprise	Kama	2009, JBFA	Compustat	Quarterly
79	salecash	Sales to cash	Ou & Penman	1989, JAE	Compustat	Annual
80	saleinv	Sales to inventory	Ou & Penman	1989, JAE	Compustat	Annual

Continued on next page

Table 6 – *Continued from previous page*

No.	Acronym	Firm characteristic	Paper's author(s)	Year, Journal	Data Source	Frequency
81	salerec	Sales to receivables	Ou & Penman	1989, JAE	Compustat	Annual
82	secured	Secured debt	Valta	2016, JFQA	Compustat	Annual
83	securedind	Secured debt indicator	Valta	2016, JFQA	Compustat	Annual
84	sgr	Sales growth	Lakonishok, Shleifer & Vishny	1994, JF	Compustat	Annual
85	sin	Sin stocks	Hong & Kacperczyk	2009, JFE	Compustat	Annual
86	sp	Sales to price	Barbee, Mukherji & Raines	1996, FAJ	Compustat	Annual
87	std_dolvol	Volatility of liquidity (dollar trading volume)	Chordia, Subrahmanyam & Anshuman	2001, JFE	CRSP	Monthly
88	std_turn	Volatility of liquidity (share turnover)	Chordia, Subrahmanyam & Anshuman	2001, JFE	CRSP	Monthly
89	stdacc	Accrual volatility	Bandyopadhyay, Huang & Wirjanto	2010, WP	Compustat	Quarterly
90	stdcf	Cash flow volatility	Huang	2009, JEF	Compustat	Quarterly
91	tang	Debt capacity/firm tangibility	Almeida & Campello	2007, RFS	Compustat	Annual
92	tb	Tax income to book income	Lev & Nissim	2004, TAR	Compustat	Annual
93	turn	Share turnover	Datar, Naik & Radcliffe	1998, JFM	CRSP	Monthly
94	zerotrade	Zero trading days	Liu	2006, JFE	CRSP	Monthly