# Is Short Interest A Significant Indicator of Stock Price Movement?

Weijia Li*

April 2017

**Abstract**

It is now commonly agreed that stocks with high short interest underperform the market. In this paper, machine learning models are used to test the fundamental hypothesis that whether short interest is a significant indicator of stock return. Further, given the relationship between short interest and stock return, the predictive powers between different models under various conditions are compared, including random forest model, support vector machine, and linear regression model.

*keywords:* Stock market, short interest, stock return, machine learning.

*JEL classification:* D91, E21, H30

*University of Chicago, Masters in Computational Social Science, weijial@uchicago.edu.

# 1 Introduction

After the dot-com bubble period, a surge of interest in short selling has been notice along with the tremendous rise and fall of stock price. Accordingly, many researchers had looked into the relationship between short interest and stock price.(Lamont et al. 2004) A short interest is defined as the quantity of stock shares that investors have sold short but not yet covered or closed out. Thus it is a market-sentiment indicator that tells whether investors think a stock's price is likely to fall. (Investopedia, 2017)

It is now commonly agreed that stocks with high short interest underperform the market (Asquith and Meulbroek, 1995; Desai et al, 2002; Asquith, et al, 2004) More specifically, many empirical literatures on short selling show that short interest ratios have increased over time and stocks with high short interest ratios have poorer performance (Desai et al (2002)), and that there is a large negative correlation between market performance and short interest (Lamont and Stein, 2004). Prior to these papers, the conventional view was that, due to the flow demand from short sellers covering their positions, large short positions foreshadow positive future returns. (Asquith et al, 2004).

In the light of previous researches, I am using machine learning models to test a more fundamental hypothesis, that whether short interest is a significant indicator of stock return. Further, given the relationship between short interest and stock return, I am comparing the predictive power between different models under various conditions, including random forest model, support vector machine, and linear regression model.

# 2 Theoretical Framework

Earlier work that I reviewed majorly used two methods to predict stock return : a four-factor regression model (Fama and French, 1993 and Carhart, 1997) or a vector autoregression model.

Desai et al., 2002 and Asquith et al, 2004 investigated on the informational role of short interest in the Nasdaq market using a calendar-time portfolio approach to measure performance over long horizons. Both of the papers used OLS to estimate the regression of the monthly portfolio excess returns on four factors: market factor, size factor, book-to-market factor and a fourth momentum factor (Fama and French, 1993 and Carhart, 1997). The regressions suggested negative relationship between high level short interests and stock market performances. The advantage of this method is that the cross-sectional correlation among individual securities that comprise the portfolio is automatically taken into account when calculating variance of the event portfolios. (Desai et al, 2002)

Another method was used by Rapach et al.. They applied time series approach and unprecedentedly showed that aggregate short interest is the strongest known predictor of the equity risk premium. Rapach et al. regress short interest and firms? shares outstanding data with vector autoregression (VAR) model to estimate the significance of the explanatory variables where the SP 500 log excess return for each month is the response variable. (Rapach et al., 2016) They also compared the predictability of aggregated normalized short interest (raw short interest divided by firm′s shares outstanding) with 14 monthly predictor variables such as log dividend yield, log earnings-price ratio, excess stock return volatility, inflation, and others.

# 3 Models

In this research, new methods are introduced in examining the significance of short interest as a predictor of stock price movement. First, machine learning models are constructed to compare their performances with that of the linear regression. Two machine learning models are constructed in this research: a random forest model with 12 predictors and a support vector machine model. Second, instead of constructing time series regression of each variable to compare their predictive abilities, I am using random forest to rank the influence of 11 company fundamentals and daily security data and stock interest on stock return. The predictive ability of short interest on stock price movement under different conditions can then be identified.

## 3.1 Linear Regression

## 3.2 Random Forest

The random forest model is difficult to illustrate due to its nature as an ensemble method. Basically, the price of stocks or the percentage change in stock prices is the response variable whilst the change in stock price in the previous period, short-interest-volume ratio and their quadratic terms are the predictors.

## 3.3 Support Vector Machine

SVM has become a rather popular machine learning technique to deal with non-linear regression and classification problems, especially with smaller datasets. Thus it is reasonable to consider comparing SVM with random forest since the number of observations is reduced to a much smaller value when looking at monthly or bi-weekly return. The initial predictors used in SVM are the same for random forest. Further model selection specific to each method, however, leads to inclusion of different predictors for the optimized models.

# 4 Data

Short interest data, daily market data and company fundamental data were obtained from Compustat spanning 2010 to 2017. Compustat data is provided by Standard Poor?s, the world?s foremost provider of independent credit ratings, risk evaluation, investment research, indices, data and valuations. The data sources is abundant, including Securities and Exchange Commission (SEC), annual and quarterly reports to shareholders, company contacts, HSBC, Frank Russell Company and others. Standard Poor?s removes reporting variability and bias in data collection and presentation process to ensure comparability. From the statistics below **??**, most of the companies in our sample are in financial sector while communication sector has the least companies included.

**Table 1:** Industry statistics of companies

| | Sectors.Var1 | Sectors.Freq |
|---|---|---|
| 1 | Financials | 15,976 |
| 2 | Consumer Cyclicals | 7,170 |
| 3 | Capital Goods | 5,081 |
| 4 | Basic Materials | 3,514 |
| 5 | Consumer Staples | 3,415 |
| 6 | Energy | 2,720 |
| 7 | Utilities | 2,577 |
| 8 | Health Care | 2,165 |
| 9 | Technology | 2,136 |
| 10 | Transportation | 1,120 |
| 11 | Communication | 869 |

From Compustat, I obtained daily security data of all North American companies in the dataset, including stock open price, stock close price, highest trade price for the trade date, and others, and short interest data. Some company fundamental data are also included as predictors for my model to compare the explanatory power across variables, such as total current asset, long-term debt due in one year, total long-term debt, total revenue and others.

In addition, in order to model a relatively clean and complete dataset, a subset of the dataset is used, which contains daily security data and supplementary bi-weekly short interest data for 3529 companies listed on the NYSE only (4852377 rows times 15 columns, the full dataset contains companies on other stock exchanges as well). Some features are engineered; for example, bi-weekly percent change is calculated between each date of release of short interest information for each company. A short-interest-volume ratio is calculated by dividing the amount of short interest by 60-day average daily volume. These two engineered variables are essential to the following models.

# 5 Result

## 5.1 Model fitting on single stocks

First, two arbitrarily picked single stocks, Bio-Rad Laboratories, Inc. and Alamo Group, Inc., are used to illustrate the limitations of using single stocks and closing prices for the purpose of this study. Bio-Rad Laboratories (ticker: BIO) is a healthcare company with 6.51 billion market capitalization and 233.47 trailing P/E. Alamo Group (ticker: ALG) is a farm and construction machinery firm whose market capitalization is 965.94 million and its trailing P/E is 22.24.

From 1 below, we can see that there are large variability between the short interest and stock prices over time for these two stocks.The short interests of BIO has roughly the same upward trend with its closing price, both have a significant rise in 2015 and dipped in 2016. However, short interest of ALG has a very different pattern with its closing price, where the closing price has a clear upward trend yet short interest has significant fluctuations.

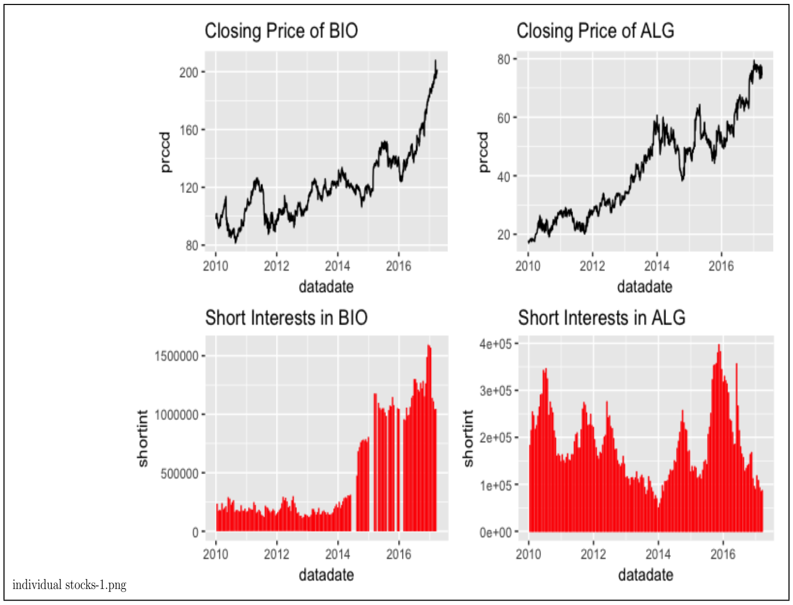


**Figure 1: Closing Price and Short interest plot for BIO and ALG. Clockwise from top left: A, B, C and D.**

The observations are made at bi-weekly time window to study the effects of predictors in the previous periods to the response variables in the current periods. Thus, each biweekly period can be treated as independent observations and avoid complex time-series analysis (it is generally accepted to treat stock returns as i.i.d. variables, Figure 2.

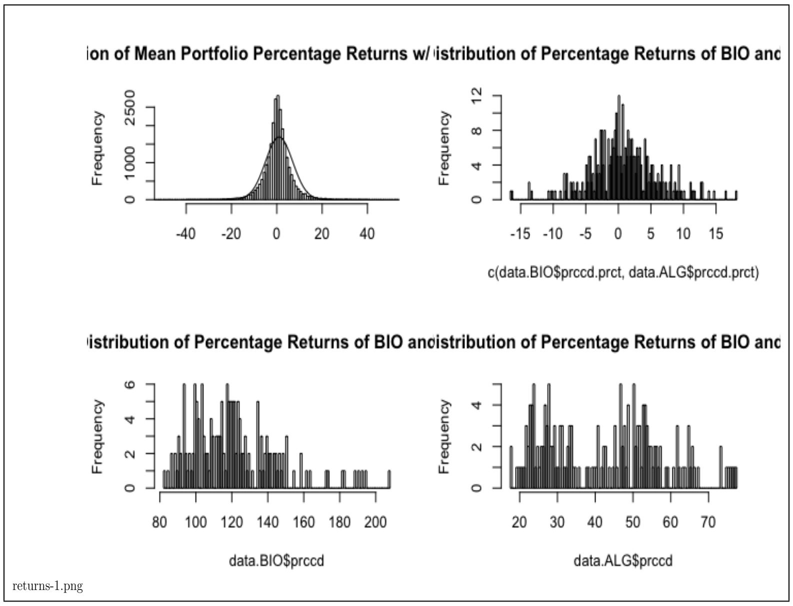


**Figure 2: Distribution of Percentage Returns**

Total data is randomly splitted into training data (70%) and testing data (30%). The results from linear regression are shown in Table 1 and 2, note that the estimates for the two companies are slightly different. In-sample fit (Figure 3, A and C) of the linear model is very close to actual value. Predicted prices of ALG using model from BIO also shows good fit (Figure 3, B). Random forest models perform slightly worse compared to linear model in-sample and scale-dependent (Figure 3, A-C). However, the models are not stable across observations as the mean squared errors from cross-validation bumps around a lot. (Figure 3, D). Thus, fitting either linear regression model or random forest model on single stock price data does not seem reliable.

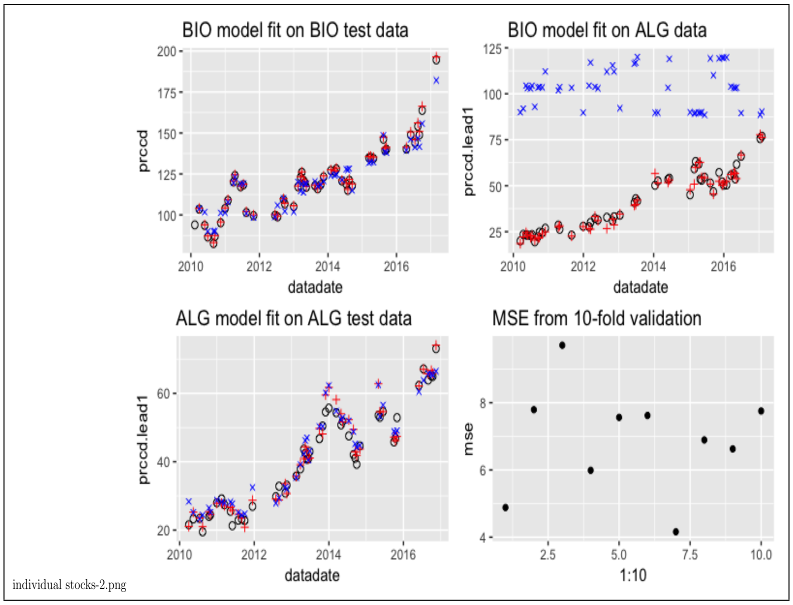


**Figure 3: Distribution of Percentage Returns**

## 5.2 Cross-Sectional Study of Portfolio Returns

In order to reduce the variance in data from individual stock and the computational complexity of fitting historical data for more than 30,000 stocks, a portfolio of 200 randomly selected stocks are constructed as a market proxy, of which 150 stocks constitute the training data while the rest 50 stocks constitute the out-of-sample testing data. The pattern of cumulative returns of this portfolio from 2010 to 2017 is highly correlated with Dow Jones Index and SP 500 Index. Numerical values are first consolidated and normalized at the stock level and then averaged across the stocks. Percentage changes from periods to periods are calculated for the fundamental indicators and price-volume data to normalize the scales across different periods of different stocks. A new feature, short interest to average trading shares ratio (sh-sh ratio), is calculated by dividing each dates short interest by the previous two-week amount of shares traded for each. The period-to-period portfolio return is calculated by averaging the individual period-to-period returns of the 200 stocks (assuming equal weighting of the stocks in the portfolio). Other fundamental indicators of the portfolio are calculated in the same fashion. This time, percentage change in closing prices (or periodically realized return) is treated as the response variable, since returns follow normal distribution much more than prices (Figure 2). A new metric, directional accuracy of model predictions, is used to access the performance of each model. Direction accuracy means the percentage of predictions that share the same sign as the actual portfolio return. Mean squared errors and directional accuracy are shown in Table 4.

## 5.3 Time Series Study of Portfolio Returns

To test whether an event will cause a systematically change along time series, I further trained my models using data prior to 2015 and test the model using data after 2015 to see the robustness of my models. The variable significant plot of random forest model below (Figure 4) suggests previous period percentage change in closing price (pre.pc.pre) is the strongest predictor with 50% increase in MSE when removing it and the second order term of short interest ratio is the third significant indicator of the model. Thus, in time series study, short interest ratio is indeed a strong predictor of stock return yet return in the previous two-week period is the strongest indicator according to my model.

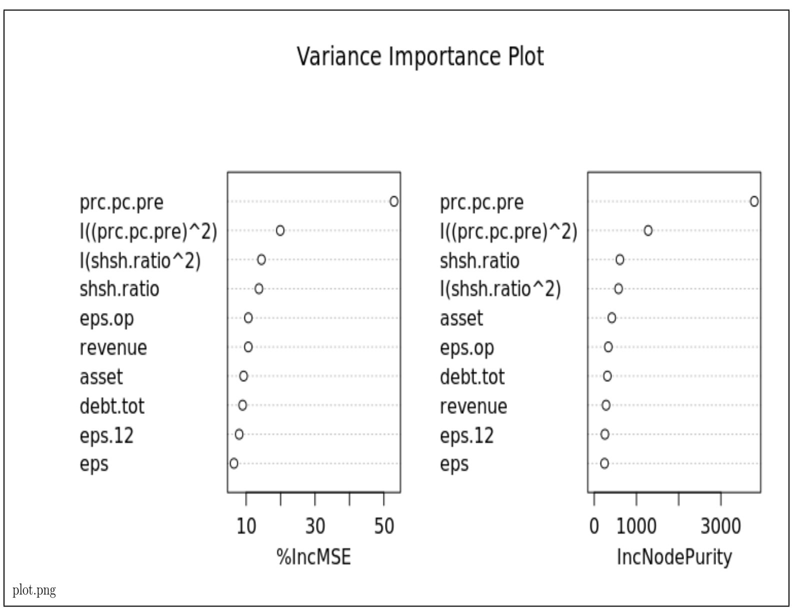


**Figure 4: Variable importance plot of time series model**

The correlation coefficients of predictions of the models with the actual return in the test period are 0.818, 0.816, 0.809,0.793 (Figure 5, clockwise from top-left to bottom-left), respectively. Further, from figure below (Figure 6), all models gives positive returns significantly above historical data, particularly, random forest give the best return across all models and SVM has the poorest performance. Since stepwise selection process dropped short interest as a predictor, I construct another linear model with short interest as an explanatory variable and compare the performance of theses two models. Also, as figure 6 illustrated, the linear model with short interest as a predictor indeed slightly outperforms that without.

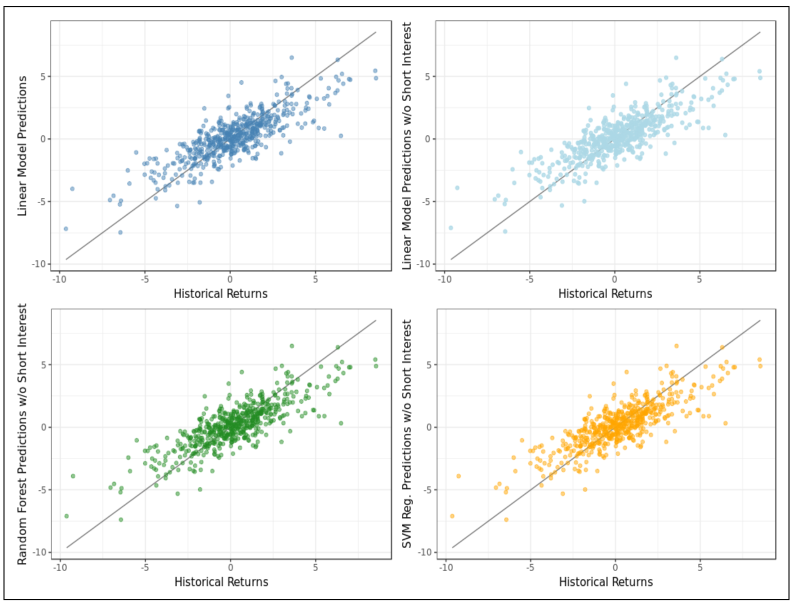


**Figure 5: Variable importance plot of time series model**

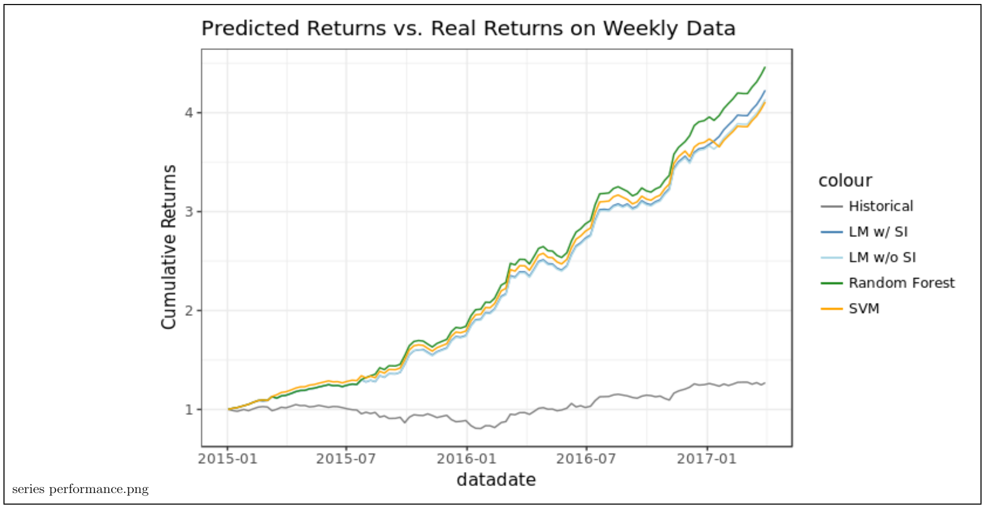


**Figure 6: Predicted returns from different models**

# 6 Conclusion and Limitations

From above, linear regression gives more accurate predictions than random forest or SVM model only when using daily data to make predictions. In other words, predicted return of random forest model and SVM model underperforms the market when following predictions made using stock closing price one day prior to trading day. Yet under such circumstance, none of the models output returns that outperforms the market. However, since variance on daily trading is relatively large and the transaction cost is high, the result is doubtfully reliable.

On the contrary, random forest gives better results both when using prior biweekly data to predict next weeks movement and when performing time series study of portfolio return. More precisely, though all models outperforms the market, random forest model generate better results.

Moreover, all variable significance plots shows second order short interest ratio and short interest ratio are significant predictors on stock return yet percentage return of prior two weeks is the strongest indicator.

The results may be biased in the for the following reasons, first, the data I am using is not large enough, only span 2010 to 2017. I was able to get data for a longer period, from 1975 to 2017, yet it is too large to run on my local machine. Secondly, the variables considered in this research are limited. Since there are too many factors that have potential effects on stock return, it is impossible to cover all of them. Furthermore, though performed back testing on the result, it is not easy to give significance. Finally, predictions may not be accurate as the model sets to buy at closing price while in the real life trading occurs at every prices in a day.

# 7 References


Lamont, Owen and Jeremy C. Stein. 2004. "Aggregate Short Interest and Market Valuations." American Economic Review, 94(2): 29-32.

Investopedia, "Short Interest", 2017, April 20, http://www.investopedia.com/terms/s/shortinterest.

Asquith, Paul, and Lisa Meulbroek, 1995, An empirical investigation of short interest, Working paper, Harvard Business School, Harvard University.

Desai, H, Ramesh, K, Ramu Thiagarajan, S  Balachandran, BV 2002, 'An investigation of the informational role of short interest in the Nasdaq market' Journal of Finance, vol 57, no. 5, pp. 2263-2287. DOI: 10.1111/0022-1082.00495

Asquith, Paul, Parag A. Pathak and Jay R. Ritter, 2005, "Short Interest, Institutional Ownership, And Stock Returns," Journal of Financial Economics, v78(2,Nov), 243-276.

Lamont, O.A., Stein, J.C., 2004. Aggregate short interest and market valuations. American Economic Review 94, 2932.

Fama, E.F., French, K.R., 1988. Dividend yields and expected stock returns. Journal of Financial Economics 91, 389406.

Carhart, Mark, 1997, On Persistence in Mutual Fund Performance, Journal of Finance 52, 57- 82.

Rapach, David and Ringgenberg, Matthew and Zhou, Guofu, 2016, Short Interest and Aggregate Stock Returns, Journal of Financial Economics (JFE), Forthcoming. Available at SSRN: https://ssrn.com/abstract=2474930 or http://dx.doi.org/10.2139/ssrn.2474930