The University of Chicago

MACS 30200 Methods And Initial Results:

Is Short Interest A Significant Indicator of Stock Price Movement?

Weijia Li

May 15th, 2017

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. 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)). Inspired by these literatures, this research is examining whether short interest is a significant indicator of stock price movement.

Methodology

Earlier works investigating the relationship between short interest and stock return majorly used two methods: 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)).

Another method was used by Rapach et al.(2016). They applied vector autoregression (VAR) model to regress short interest and firms' shares outstanding data to estimate the significance of the explanatory variables where the S&P 500 log excess return for each month is the response variable. A one-sided alternative hypothesis is used for a more powerful test of predictability. 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.

In this research, new methods are introduced in examining the significance of short interest as a predictor of stock price movement.

First, a machine learning model is constructed to improve the fitness of model. At this stage, only random forest model is constructed which outperformed linear regression model. Second, instead of constructing time series regression of each variable to compare their predictive abilities, I am using random forest to rank the importance of 14 variables from Rapach et al. and stock interest as the 15th variable.-The relationship between short interest and stock performances under different conditions can then be identified and a threshold that a short interest above this point will cause the stock underperformance with a very high probability can also be found.

Data

Short interest data, daily market data and company fundamental data were obtained from Compustat spanning 1975 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 Compustat, I requested daily security data, 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 planned to be included as predictors for my model to compare the explanatory power across variables, but the data is yet to be approved.

In addition, due to the time constraint and computational complexity for the initial analysis, 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). A couple of 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.

Model

At this stage, we have mainly compared two models to illustrate our research question--the linear regression model and random forest regression model. For the linear model, the basic equation is

$$Y = \beta_0 + \beta_1 X + e,$$

where beta includes the price of stocks or the percentage change in stock prices is the response variable while the change in stock price in the previous period, short-interest-volume ratio, their quadratic terms and two-way interactions are the predictors.

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

Initial results

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 figure 1 below, we can see that there are large variability between the short interest and stock prices over time for these two stocks. For the sake of simplicity, in the initial analysis, the observations are made only at bi-weekly time window, and we want to study the effects of predictors in the previous period to the response variables in the current periods. Thus, we can treat each biweekly period as independent observations and avoid complex time-series analysis (it is generally accepted to treat stock returns as i.i.d. variables, **Figure 3**). Training data and testing data are randomly selected and approximate 70 percent and 30 percent of total data. The results from linear regression are shown in **Table 1 and 2**. Notice that the estimates for the two companies are slightly different. In-sample fit (**Figure 2, 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 2, B).** Random forest models perform slightly worse compared to linear model in-sample and scale-dependent (**Figure 2, A-C**). However, the models are not stable across observations as the mean squared errors from cross-validation bumps around a lot. (**Figure 2, D**). Thus, fitting either linear regression model or random forest model on single stock price data does not seem to be a good idea.

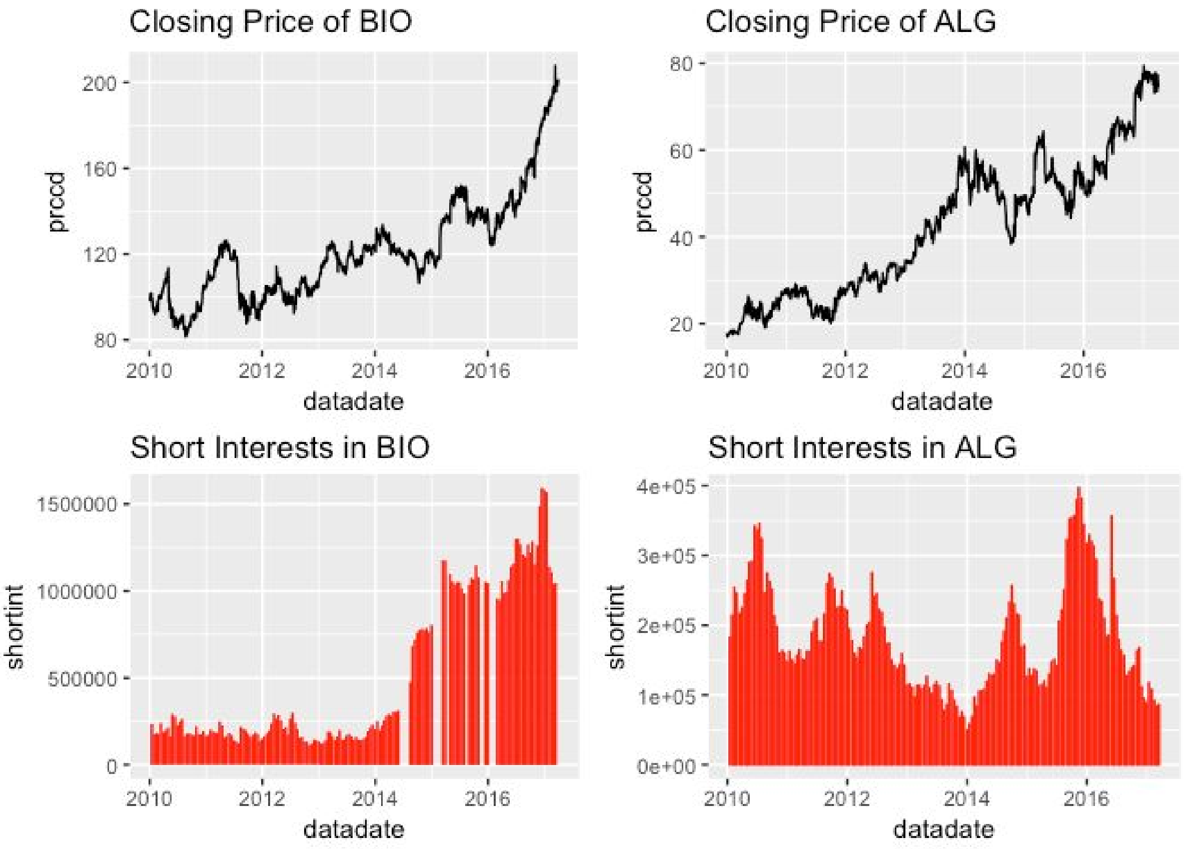


Fig. 1 Closing Price and Short interest plot for BIO and ALG. Clockwise from top left: A, B, C and D.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.567024	1.730037	0.906	0.367	
prccd	0.992764	0.035038	28.334	<2e-16	***
shsh.ratio	-0.21142	0.349011	-0.606	0.546	
prccd:shsh.ratio	0.002033	0.007833	0.26	0.796	

Table 1: Summary Statistics of Linear Regression of ALG

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.235898	6.415633	0.349	0.728	
prccd	0.983899	0.058081	16.94	<2e-16	***
shsh.ratio	-0.535407	0.971633	-0.551	0.583	
prccd:shsh.ratio	0.004906	0.008256	0.594	0.554	

Table 2: Summary Statistics of Linear Regression of BIO

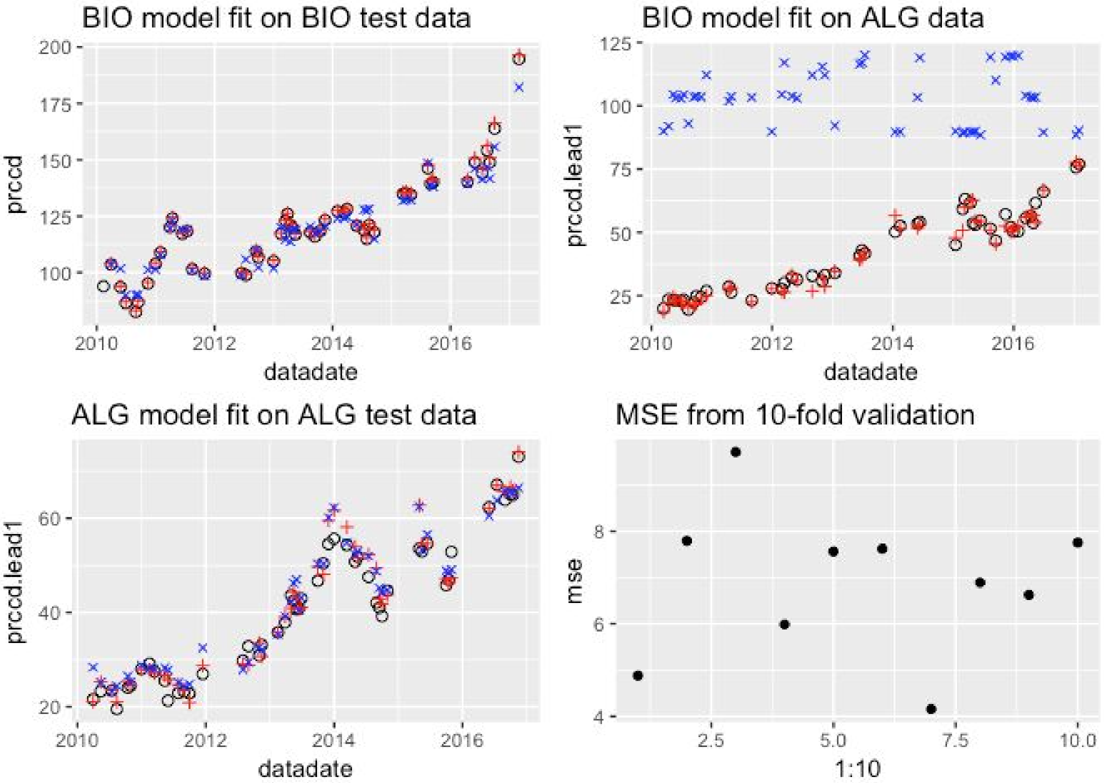
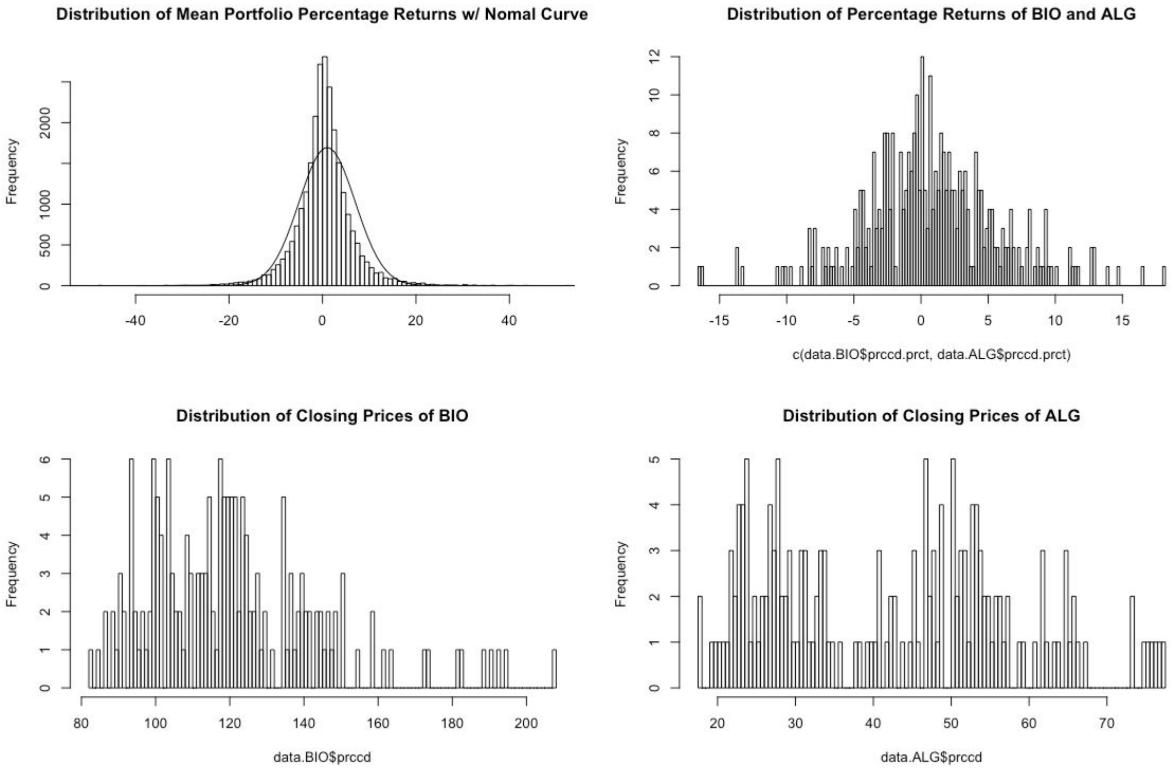


Fig. 2 Model Fitting for single stocks and MSE from cross validation.

Next, 200 randomly selected stocks are constructed as a portfolio to reduce variance among individual stocks. 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). 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 3**). A another portfolio composed of another set of 100 stocks is used as validation dataset. Linear regression model and random forest model are fitted to the training data. Mean squared error of the linear regression model on the validation dataset is 30.56 while that of the random forest model is 18.41. A comparison of the directional prediction accuracy is summarized in **Table 3**. Since the accuracy of the linear regression model on the training data is higher than the testing data, it is possible that the linear model overfits the data and thus random forest actually performs better overall, and its performance is also better than naive guessing. Based on the variance importance plot shown below, the ratio of short interest over average daily volume is indeed an important variable.

	Model	Data	Accuracy
1	LinReg	Train	0.5976
2	RndmFrst	Train	0.5385
3	LinReg	Test	0.5621
4	RndmFrst	Test	0.5621

Table 3: Prediction Accuracy of different models


Distribution of Mean Portfolio Percentage Returns w/ Nomal Curve

Distribution of Percentage Returns of BIO and ALG

Distribution of Closing Prices of BIO

Distribution of Closing Prices of ALG

Fig. 3 Distribution of Percentage Returns

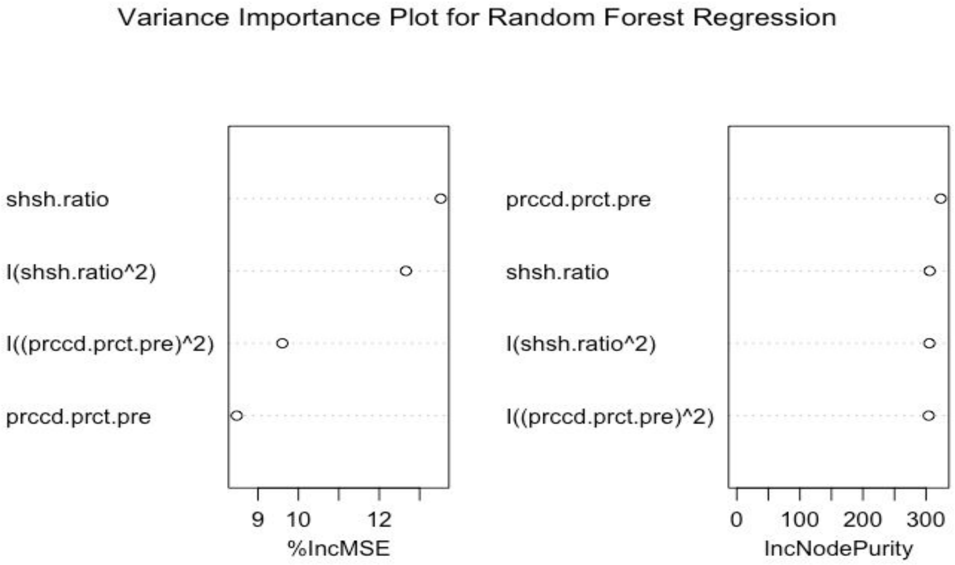


Fig. 4 Variance Importance Plot for Random Forest

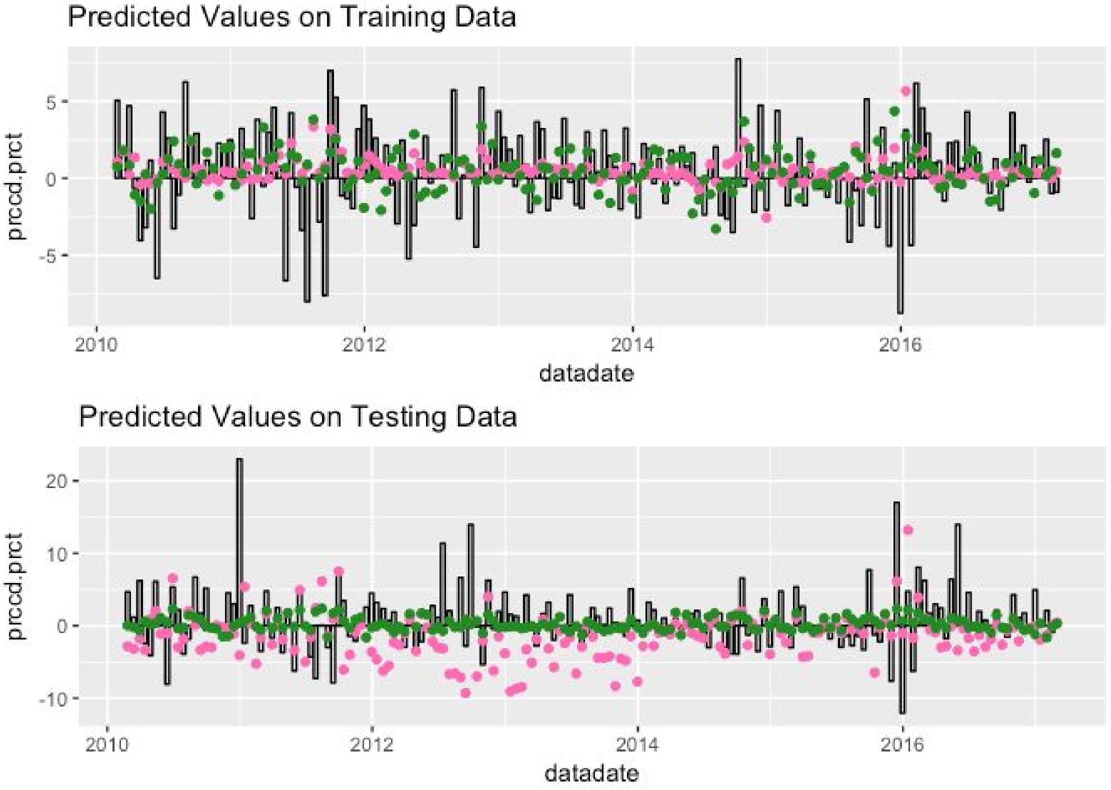


Fig. 5 Predicted Values on Training and Testing Data. Pink dots are linear regression predictions and green dots are random forest predictions.