Is Short Interest A Significant Indicator of Stock Price Movement

Weijia Li | weijial@uchicago.edu
Computational Social Science, Social Science Division, University of Chicago

Background

•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

1. Using machine learning models to improve the fitness of model
2. Using random forest to rank the importance of 14 variables from earlier works and stock interest as the 15th variable to verify the significance of short interest
3. Identifying the relationship between short interest and stock performances under different conditions
4. Identifying a threshold that a short interest above this point will cause the stock underperformance with a very high probability

Data

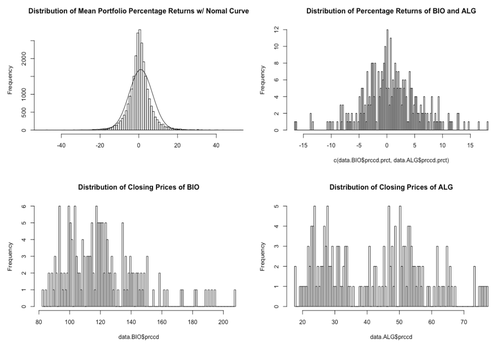
Short interest data, daily market data and company fundamental data were obtained from Compustat spanning 2010 to 2017.

Information on industry distribution is shown below

	Sectors.Var1	Sectors.Freq
1	Financials	15976
2	Consumer Cyclicals	7170
3	Capital Goods	5081
4	Basic Materials	3514
5	Consumer Staples	3415
6	Energy	2720
7	Utilities	2577
8	Health Care	2165
9	Technology	2136
10	Transportation	1120
11	Communication	869

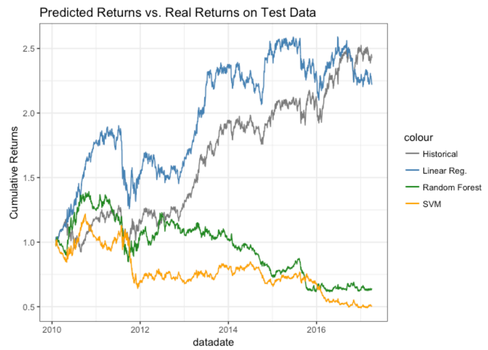
Later, 200 randomly selected stocks are constructed as a portfolio to reduce variance among individual stocks. The period-to-period portfolio return is calculated by averaging the individual period-to-period return (assuming equal weighting of the stocks).

Two arbitrarily picked single stocks, Bio-Rad Laboratories, Inc. and Alamo Group, Inc., are used to illustrate the distribution of percentage return on single stocks

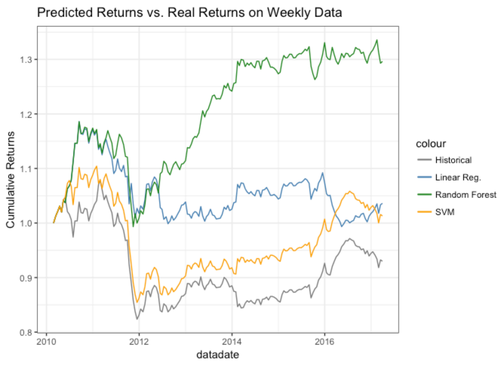


Compare model performances

Assuming trading once per day, i.e. buying if the model predicts a rise in stock price in the next day and selling if the model predicts a negative trending, linear model outperforms random forest model and SVM model.



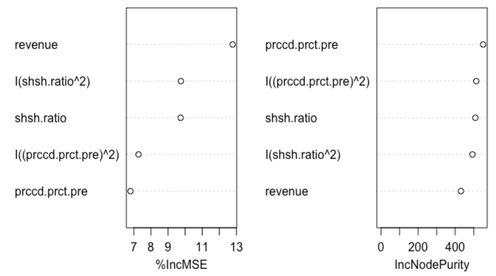
However, when using previous biweekly data to predict next week's trending, random forest model makes better predictions than other models.



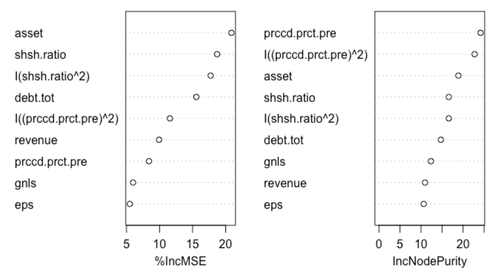
The table on the right shows the accuracy and predicted returns of each model using daily data and biweekly data respectively.

The models presenting now have not had their coefficients adjusted; tuned models in the final report are expected to have better results.

Variable Significance Plot



The variable significance plot of daily basis prediction model shows short interest ratio is indeed an important predictor of the model



The variable significance plot of the model using previous biweekly data to make prediction of the next week stock movement also shows that short interest ratio is an significant predictor

Model	Time	Accuracy	Return
LinReg	Daily	0.5343	2.2190
RndmFrst	Weekly	0.5343	1.0330
SVM	Daily	0.4801	0.6421
LinReg	Weekly	0.5647	1.2930
RndmFrst	Daily	0.4889	0.5079
SVM	Weekly	0.5353	1.0160

Conclusion

For daily trading strategy, linear regression gives more accurate predictions than random forest or SVM model. In other words, predicted return of random forest model and SVM model underperforms the market when making daily prediction.

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 using prior biweekly data to predict next week's movement. More specifically, though all models outperforms the market, random forest has a better performance.

Limitations

1. 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
2. The variables considered in this research are limited. Since there are too many factors that have potential effect on stock price movement, it is impossible to cover all of them.
3. Though performed back testing on the result, it is not easy to give significance
4. Predictions may not be accurate as the model sets to buy at closing price.