QUANTITATIVE TRADING

QUANTITATIVE INVESTMENT AND TRADING IDEAS, RESEARCH, AND ANALYSIS.

FRIDAY, JULY 21, 2017

Building an Insider Trading Database and Predicting Future Equity Returns

By John Ryle, CFA

===

I've long been interested in the behavior of corporate insiders and how their actions may impact their company's stock. I had done some research on this in the past, albeit in a very low-tech way using mostly Excel. It's a highly compelling subject, intuitively aligned with a company's equity performance - if those individuals most in-the-know are buying, it seems sensible that the stock should perform well. If insiders are selling, the opposite is implied. While reality proves more complex than that, a tremendous amount of literature has been written on the topic, and it has shown to be predictive in prior studies.

In generating my thesis to complete Northwestern's MS in Predictive Analytics program, I figured employing some of the more prominent machine learning algorithms to insider trading could be an interesting exercise. I was concerned, however, that, as the market had gotten smarter over time, returns from insider trading signals may have decayed as well, as is often the case with strategies exposed to a wide audience over time. Information is more readily available now than at any time in the past. Not too long ago, investors needed to visit SEC offices to obtain insider filings. The standard filing document, the form 4 has only required electronic submission since 2003. Now anyone can obtain it freely via the SEC's EDGAR website. If all this data is just sitting out there, can it continue to offer value?

I decided to inquire by gathering the filings directly by scraping the EDGAR site. While there are numerous data providers available (at a cost), I wanted to parse the raw data directly, as this would allow for greater "intimacy" with the underlying data. I've spent much of my career as a database developer/administrator, so working with raw text/xml and transforming it into a database structure seemed like fun. Also,

since I desired this to be a true end-to-end data science project, including the often ugly 80% of the real effort – data wrangling, was an important requirement. That being said, mining and cleansing the data was a monstrous amount of work. It took several weekends to work through the code and finally download 2.4 million unique files. I relied heavily on Powershell scripts to first parse through the files and shred the xml into database tables in MS SQL Server.

With data from the years 2005 to 2015, the initial 2.4 million records were filtered down to 650,000 Insider Equity Buy transactions. I focused on Buys rather than Sells because the signal can be a bit murkier with sells. Insider selling happens for a great many innocent reasons, including diversification and paying living expenses. Also, I focused on equity trades rather than derivatives for similar reasons -it can be difficult to interpret the motivations behind various derivative trades. Open market buy orders, however, are generally quite clear.

After some careful cleansing, I had 11 years' worth of useful SEC data, but in addition, I needed pricing and market capitalization data, ideally which would account for survivorship bias/dead companies. Respectively, Zacks Equity Prices and Sharadar's Core US Fundamentals data sets did the trick, and I could obtain both via Quandl at reasonable cost (about $350 per quarter.)

For exploratory data analysis and model building, I used the R programming language. The models I utilized were linear regression, recursive partitioning, random forest and multiplicative adaptive regression splines (MARS). I intended to make use of a support vector machine (SVM) models as well, but experienced a great many performance issues when running on my laptop with a mere 4 cores. SVMs have trouble with scaling. I failed to overcome this issue and abandoned the effort after 10-12 crashes, unfortunately.

For the recursive partitioning and random forest models I used functions from Microsoft's RevoScaleR package, which allows for impressive scalability versus standard tree-based packages such as rpart and randomForest. Similar results can be expected, but the RevoScaleR packages take great advantage of multiple cores. I split my data into a training set for 2005-2011, a validation set for 2012-2013, and a test set for 2014-2015. Overall, performance for each of the algorithms tested were fairly similar, but in the end, the random forest prevailed.

MY TRADING WORKSHOPS

Algorithmic Options Strategies

PARTNER CENTER

SUBSCRIBE TO MY BLOG

Enter your Email

Subscribe me!

TWITTER

For my response variable, I used 3-month relative returns vs the Russell 3000 index. For predictors, I utilized a handful of attributes directly from the filings and from related company information. The models proved quite predictive in the validation set as can be seen in exhibit 4.10 of the paper, and reproduced below:

Exhibit 4.10 Mean Actual Returns Per Prediction Quintile, Random Forest, Validation Set

Quintile, Predicted Relative Return	Actual Relative Return
1	-1.29%
2	0.02%
3	0.42%
4	1.91%
5	3.96%

The random forest's predicted returns were significantly better for quintile 5, the highest predicted return grouping, relative to quintile 1(the lowest). Quintiles 2 through 4 also lined up perfectly - actual performance correlated nicely with grouped predicted performance. The results in validation seemed very promising!

However, when I ran the random forest model on the test set (2014-2015), the relationship broke down substantially, as can be seen in the paper's Exhibit 5.2, reproduced below:

Exhibit 5.2 Mean Actual Returns, Market Cap Per Prediction Decile, Random Forest, Test Set

Decile	Relative Return	Market Cap
1	-4.10%	3,938,765,296
2	-1.49%	12,410,729,439
3	-0.89%	16,886,486,468
4	-0.51%	21,436,082,514
5	-0.04%	16,973,231,878
6	-0.48%	10,933,652,622
7	-0.46%	5,530,582,160
8	-1.05%	2,883,277,390
9	-1.70%	1,845,858,414
10	-1.01%	582,953,531

Fortunately, the predicted 1st decile was in in fact the lowest performing actual return grouping. However, the actual returns on all remaining prediction deciles appeared no better than random. In addition, relative returns were negative for every decile.

While disappointing, it is important to recognize that when modeling time-dependent financial data, as the time-distance moves further away from the training set's time-frame, performance of the model tends to decay. All market regimes, gradually or abruptly, end. This represents a partial (yet unsatisfying) explanation for this relative decrease in performance. Other effects that may have impaired prediction include the use of price, as well as market cap, as predictor variables. These factors certainly underperformed during the period used for the test set. Had I excluded these, and refined the filing specific features more deeply, perhaps I would have obtained a clearer signal in the test set.

In any event, this was a fun exercise where I learned a great deal about insider trading and its impact on future returns. Perhaps we can conclude that this signal has weakened over time, as the market has absorbed the informational value of insider trading data. However, perhaps further study, additional feature engineering and clever consideration of additional algorithms is worth pursuing in the future.

John J Ryle, CFA lives in the Boston area with his wife and two children. He is a software developer at a hedge fund, a graduate of Northwestern's Master's in Predictive Analytics program (2017), a huge tennis fan, and a machine learning enthusiast. He can be reached at john@jryle.com.

===

Upcoming Workshops by Dr. Ernie Chan

July 29 and August 5: Mean Reversion Strategies

In the last few years, mean reversion strategies have proven to be the most consistent winner. However, not all mean reversion strategies work in all markets at all times. This workshop will equip you with basic statistical techniques to discover mean reverting markets on your own, and describe the detailed mechanics of trading some of them.

September 11-15: City of London workshops

These intense 8-16 hours workshops cover Algorithmic Options Strategies, Quantitative Momentum Strategies, and Intraday Trading and Market Microstructure. Typical class size is under 10. They may qualify for CFA Institute continuing education credits.

===
Industry updates

- scriptmaker.net allows users to record order book data for backtesting.
- Pair Trading Lab offers a web-based platform for easy backtesting of pairs strategies.

POSTED BY ERNIE CHAN AT 4:13 PM

5 COMMENTS:

Anonymous said...

why is the market cap so small. current us mktcap is ~20 trillion and all 10 deciles total up to something < 100 billion according to your table, i.e. well under 1% of total mktcap. additionally, i would guess that it's the larger companies that have more insiders buying and selling and are also more diligent and timely in reporting. what am i missing?

FRIDAY, JULY 21, 2017 AT 8:32:00 AM EDT

John Ryle said...

Thanks for your comment! I probably could've clarified the The Market Cap column more effectively. The Decile column did not represent a grouping by Market Cap. It merely represented the Predicted Return groupings based on the model, as built upon the Training set. The 2nd column represents actual avg return per decile in the Test set(2014-2015). The Market Cap column represents the Avg Mkt Cap within each decile group. Each of these deciles can contain large or small cap stocks. By averaging them, I wanted to see if those with a higher avg had a different return profile. As it turns out, they did..a little bit. Deciles 8,9,10 were the lowest 3 average market caps per decile. The other Decile Mkt Caps were all over the place, unfortunately.

FRIDAY, JULY 21, 2017 AT 8:43:00 PM EDT

Eduardo Gonzatti said...

This comment has been removed by the author.
SATURDAY, JULY 22, 2017 AT 8:50:00 AM EDT

Eduardo Gonzatti said...

Thanks for the article, John,
Did you actually tried to run this on short sellings? Thanks

SATURDAY, JULY 22, 2017 AT 8:51:00 AM EDT

John Ryle said...

LINKS

QTS Managed Accounts
RSS Site Feed
Quantitative Research & Trading
Quantocracy
Quant News
Insider Monkey
FactorWave
Eran Raviv (Quant Analyst at large pension fund)
Advertise With Us
High frequency historical data
MATLAB automated trading course

BLOG ARCHIVE

Hey Eduardo,
I ran this only for Insider Buy trades of equity securities. I do plan to revisit this data set in the future and explore other scenarios.

MONDAY, JULY 24, 2017 AT 7:57:00 AM EDT

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)

► 2016 (4)
► 2015 (7)
► 2014 (8)
► 2013 (11)
► 2012 (12)
► 2011 (15)
► 2010 (17)
► 2009 (32)
► 2008 (28)
► 2007 (50)
► 2006 (24)