Building an Insider Trading Database and Predicting Future Equity Returns

By

John J Ryle

Thesis Project

To Be Submitted in partial fulfillment of the Requirements for the degree of

MASTER OF SCIENCE IN PREDICTIVE ANALYTICS

NORTHWESTERN UNIVERSITY

May 2017

Ernest Chan, First Reader

Laurence Fulton, Second Reader

Abstract

This paper considers the impact of legal insider trading on subsequent investment returns. With filings data gathered directly from the Security Exchange Commission, various machine learning algorithms are applied, scored and compared for their return predictability relative to the Russell 3000 index. Predictor variables are derived directly from the insider filings, as well as from metadata from related corporate filings. The filings data, spanning years 2005 through 2015, is transformed into a relational database and analyzed using statistical analysis and predictive modeling libraries within the R programming language. The data is separated into training, validation and test sets. Modeling was performed using a variety of algorithms. Performance in the validation set appeared quite predictive, but results were less conclusive when scored on the test set.

Table of Contents

Introduction

This paper aims to address the feasibility of utilizing insider trade filings in forecasting futures stock returns. This topic has been widely investigated in the past, and has been shown to be predictive. However, given the increased availability of information, low-cost computing power, regulations that require corporate transparency, and a more skilled pool of market participants, can insider trading behavior still prove to be predictive, or has it withered away? While there are numerous vendors which parse through SEC filings and sell their aggregated data, I've decided to gather the raw filings directly from the SEC. Perhaps by gathering this data independently, and engineering a set of predictors from this raw set, new insights can be gathered from the filings, providing an alternate lens on the data that could shed some new understandings.

1. Overview

1.1 A Brief History of Insider Trading

For as long as there have been financial markets, inside information has, naturally, provided an edge – a material advantage over other market participants in assessing the future prospects of a company. Whether that advantage is fair or not has long been the subject of great debate. A great many scandals in the history of finance involved insider trading at their core. Since the aftermath of the Great Depression, the Securities Exchange Commission has been the primary regulator of financial markets in the United States. The SEC requires corporate insiders of publicly traded companies to submit a Form 4 Filing which details their recent transactions in the firm's securities. As of June 2003[1], they've been required to file electronically through the SEC's

Electronic Data Gathering, Analysis and Retrieval System, EDGAR, within two days of the transaction.

The EDGAR website is entirely available to the public. In addition to Insider filings, it stores a plethora of information such as corporate annual 10K and quarterly 10Q financial reports, as well as mutual fund information, hedge fund holding reports, and key corporate event notifications via 8-K filings.

1.2 The Evolution of Market Efficiency

The investment industry keeps getting smarter. Ever since Eugene Fama put forth the Efficient Market Hypothesis (EMH), the efficiency concept has been questioned by the investing establishment. Over time, firms have invested in increasingly sophisticated methods of analysis, dismissing the EMH, and showing that firms, in many cases, could consistently beat the market. Ironically, however, this led to researchers Dwight Lee and James Verbrugge[2] in 1996 pointing out: "the Efficient Market theory is practically alone among theories in that it becomes more powerful when people discover serious inconsistencies between it and the real world. If a clear efficient market anomaly is discovered, the behavior (or lack of behavior) that gives rise to it will tend to be eliminated by competition among investors for higher returns." These researchers are pointing out what Michael Maubossin refers to as the paradox of skill[3], whereby as market participants get smarter, it becomes more difficult to outperform the market as a whole.

After mentioning how his firm, Oaktree Capital, took advantage of dislocations in high yield debt in the 1970s due to market inefficiencies, Howard Marks emphasized that he expected to see less low-hanging fruit going forward, in Barron's in 2014 stating: "If efficiency should be the going-in presumption, so should "efficientization." That's my term for the process through which

a market becomes more efficient. In short, over time the actions of diligent investors should have the effect of driving out bargains. If at first bargains exist, their holders will enjoy superior risk-adjusted returns, other investors will take note, and they'll study them and bid them up enough to eliminate the bargain element and thus the potential for further excess returns." He goes on to say: "the conditions of today are less propitious for inefficiency than those of the past. In short, it makes sense to accept that most games are no longer as easy as they used to be, and that as a result free lunches are scarcer. Thus, in general, I think it will be harder to earn superior risk-adjusted returns in the future, and the margin of superiority will be smaller." [4]

1.3 Why Look at Insider Trading Data?

The utilization of Insider Transaction filings has long been studied by researchers and top firms. In 1998, H. Nejat Seyhun published *Insider Intelligence*, which elaborated on many of his prior papers on insider trading, and the market impact subsequent to such trading. He showed that, during the period of his study (1975-1994), Trading based on Insider Buys as well as Sells was very profitable[5].

However, much has changed since 1994, and, as Marks referred to above, the market is more efficient now than in the past. Information is more readily available. Computing power has grown exponentially. Corporations are required to disclose material information publicly.

Specific to insider trading, the SEC now has all of this data available online, for free. In the 1990s, you'd need to request paper copies of the filings. Now, it is all stored in a standardized xml format. With free and low-cost software, individuals can mine the data using sophisticated data extraction libraries. In addition, vendors make this data available at increasingly competitive

prices. Given the explosion of accessibility to such data, can the anomaly of insider trading profitability persist? Or has it withered away? The purpose of this paper is to try to figure this out.

2. The Data

2.1 Sources

The most critical information required for this study are historic SEC Form 4 filings, which are stored on the SEC's EDGAR database, entirely accessible from the SEC.gov website. In addition, historic pricing data was needed, as well as market capitalization information. Each of these data sets were obtained via Quandl(Quandl.com), which aggregates data from various vendors. For the historic pricing data, the Zacks Equity Price data set was used. Zacks provides professional grade pricing information for active and delisted securities. The delisted security prices are critical to account for survivorship bias. The market capitalization information was obtained via Sharadar's US Fundamentals data set, also via the Quandl website. Market capitalization information is now available from with within a firm's 10-K filing standardized xml schema, but prior to 2011 this schema was not a standard.

2.2 Data Collection

Since the SEC does not provide any means of aggregation, obtaining 11 years of data (2005-2015), all kept in individual files, is challenging. Large quarterly index files are available, however, which provide directory location based on Central Index Keys, which are the SEC‘s primary identifier for filers, whether individuals or a corporations. To parse through each of these, I imported each index file into a Microsoft SQL Server 2016 database and searched the locations of each filing. From over 12 million files, 2.44 million represented unique Form 4 records. I used Powershell to loop through these records and obtain the 5 million files. The actual files were then imported into a relational database, again using Powershell and SQL Server’s xml shredding/normalization functions. There were approximately 1,100 exceptions from this list, generally due to formatting issues.

After elimination of transactions by corporate entities rather than individuals, data with integrity issues, prices less than $2 and/or market capitalization less than $1 million, and trades that represented derivatives only, we reduced the data set to 653,299 Insider Buy records.

2.3 Why We’ll Focus on Buys

This paper will focus on insider purchases, and specifically those buys involving open market equity orders of stock. The complexity of derivative transactions can send mixed signals which are troublesome to parse. Many insiders receive stock options they exercise shortly after the firm goes public. Insiders sell for many reasons other than the individual’s opinion of the future prospects of the stock[6]. They sell to pay their mortgage, their kids’ college funding, and, for those whose preponderance of wealth is tied up in one stock, to diversify. While some studies have shown predictiveness in insider selling events, others have shown this such results as far

less clear (Lakonishok, and Lee, 2001).[7] While there may be clever filters to identify the reasons to sell (or sell short) an investment based on insider sells, I have not been able to identify it. Buys provide a purer signal.

Period returns in this study were generated based on trading days. Each periods was approximated by utilizing the following mapping scheme:

5 Trading Days = 1 Week

21 Trading Days = 1 Month

63 Trading Days = 3 Months

126 Trading Days = 6 Months

252 Trading Days = 1 Year

504 Trading Days = 2 Years

756 Trading Days = 3 Years

Exhibit 2.1, below, shows the performance of stocks following sales over 1 week, 1, 3, and 6 months, as well as 1, 2, and 3 years following insider trades. Exhibit 2.2 shows same period performance for buys. Exhibit 2.3 and 2.4 reflect the tabular output, while Exhibit 2.5 compares Buys, Sells and Average performance.

Exhibit 2.1 Return Performance: Insider Sells vs Russell 3000

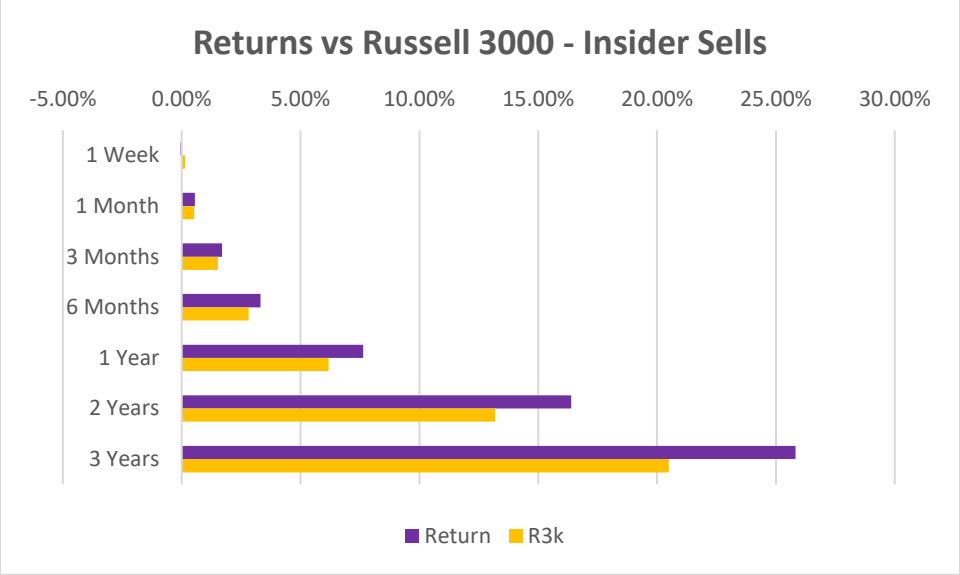


Exhibit 2.2 Return Performance: Insider Buys vs Russell 3000

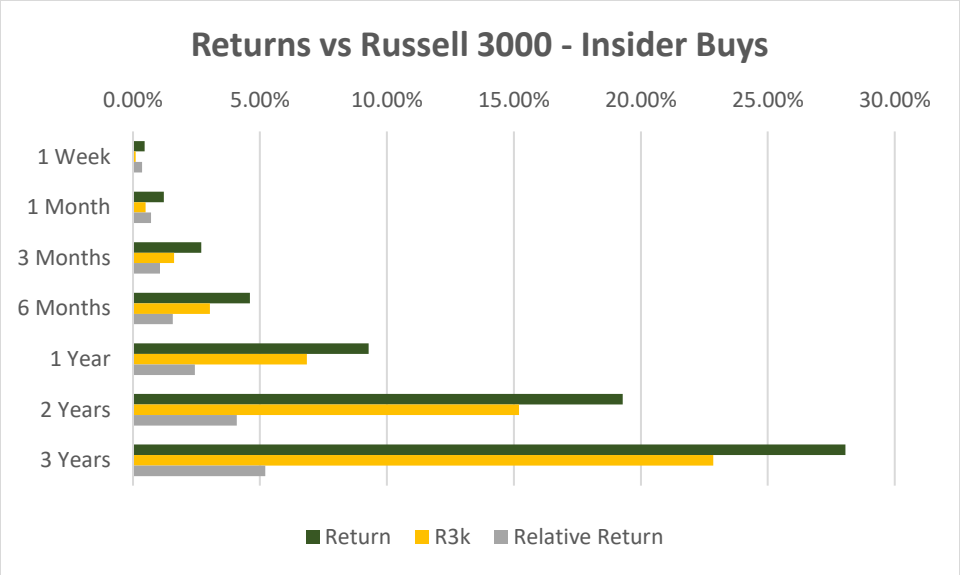


Exhibit 2.3 Returns Following Insider Sells

Period	Return	R3k	Relative Return
1 Week	-0.05%	0.14%	-0.19%
1 Month	0.55%	0.53%	0.02%
3 Months	1.70%	1.52%	0.18%
6 Months	3.31%	2.82%	0.49%
1 Year	7.64%	6.19%	1.45%
2 Years	16.39%	13.19%	3.21%
3 Years	25.83%	20.50%	5.33%

Exhibit 2.4 Returns Following Insider Buys

Period	Return	R3k	Relative Return
1 Week	0.46%	0.10%	0.36%

1 Month	1.21%	0.50%	0.71%
3 Months	2.68%	1.62%	1.06%
6 Months	4.60%	3.03%	1.57%
1 Year	9.28%	6.85%	2.43%
2 Years	19.29%	15.21%	4.09%
3 Years	28.06%	22.86%	5.21%

Exhibit 2.5 Period Returns for Sells, Buys and Baseline Average

Time Frame	Trade Type	Return	R3k	Relative Return
1 Week	Sells	-0.05%	0.14%	-0.19%
	Average	0.28%	0.12%	0.17%
	Buys	0.46%	0.10%	0.36%
1 Month	Sells	0.55%	0.53%	0.02%
	Average	0.98%	0.51%	0.47%
	Buys	1.21%	0.50%	0.71%
3 Months	Sells	1.70%	1.52%	0.18%
	Average	2.34%	1.59%	0.75%
	Buys	2.68%	1.62%	1.06%
6 Months	Sells	3.31%	2.82%	0.49%
	Average	4.15%	2.96%	1.19%
	Buys	4.60%	3.03%	1.57%
1 Year	Sells	7.64%	6.19%	1.45%
	Average	8.71%	6.62%	2.09%
	Buys	9.28%	6.85%	2.43%
2 Years	Sells	16.39%	13.19%	3.21%
	Average	18.28%	14.50%	3.78%
	Buys	19.29%	15.21%	4.09%
3 Years	Sells	25.83%	20.50%	5.33%

	Average	27.28%	22.03%	5.25%
	Buys	28.06%	22.86%	5.21%

It is important to note the reason that the Russell 3000 returns are different for Buys vs Sells. This difference exists because the returns are date-specific. They represent the average returns beginning on each transaction date. For example, if there was a trade on February 6th of 2005, its 1 year return and corresponding Russell 3000 return would reflect the returns over the following 252 trading days. Note in Exhibit 2.5 that the 3 Year Relative Return for Sells was better than for Buys, but from Exhibits 2.3 and 2.4 we can see that the benchmark Russell 3000 was 2.36% less during the weighted average Sell period, so we see the Buys-based absolute return for 3 years was better by 2.23 percent. Also note the Average is closer to the Buys because there were roughly two insider buys for every sell (653,299 vs 349,221).

In any event, returns relative to the Russell 3000 were better for all time-frames except the 1-week period for Sells. This could certainly be due to independent effects, as perhaps the Russell isn't the most representative index for this data set. It is a market-cap weighted, and many of the companies in this dataset are quite small. In any event, returns following insider buys outperform returns following insider sells on an absolute basis over any time period.

2.4 Specifically, What We Are Predicting

The prediction has been set up for purposes of a theoretical trading engine. This engine which would scan the SEC's EDGAR RSS feed intraday for the latest Form 4s. It would score each filing based on the winning model in this paper. If the filing's score was high enough, the engine

would issue a Buy order on the particular company's stock. The engine would look at attributes specific to the individual form 4, such as transaction code type ("TranCodePrimary"), the insider's position/title/role at the firm ("Role"), or whether the individual holds a significant portion of the shares ("IsTenPercentOwner.") These fields already exist on the existing form, but for this project I've engineered some additional variables which look back at other recent Form 4 Filings involving either the insider or her firm. This is captured in fields such as "NetOwnerSharesTraded30Days" and "NumIssuerTradesSameDirection30Days."

Also, we attempt to capture preplanned trades, otherwise known as 10b5-1 plans[8]. 10b5-1 plan allow insiders to transact at a future date on a scheduled basis. However, this information is not clearly identified on a Form 4. It is generally (but not necessarily) captured in the footnotes. I've parsed through this field looking for terms such as "10b5-1" or "10b51" or "preplanned". It is an imperfect indicator, as this text-based filter is certain to miss some pre-planned trades.

The table in Exhibit 2.6 presents the Predictor variables in rows 1 through 26, and the response variable, 3-month returns relative to the Russell 3000 index, labeled "R63_REL" (for 63 trading days), in row 27. 3-month returns were chosen as the response because it provides an adequate time for an insider-transaction affect to be incorporated in price, yet is a short enough time frame as to allow for a larger dataset to work with. The longer the return period, the less data we'll have results available to work with. In any event, as exhibits 2-1-2.5 show, the aggregate return numbers are representative of the periods in question. In other words, relative returns expand at a steady rate from 1-week to 1-month, 1-month to 3-months and onwards. I'd anticipate similar results regardless of the time frame. The market doesn't seem to adjust immediately to insider trading disclosures.

Exhibit 2.6 Variables Used in Analysis

	Long Name	**Short Name**	Data Type	**Description**
1	DocId	DocId	INT	Unique Identifier
2	ReportDate	RptDate	DateTime	Date of form 4 submission.
3	NetSharesTraded	NetSharesTraded	NUMERIC	Total net shares traded on form 4. There can be more than 1 transaction on each filing.
4	TranCodePrimary	TranCode	Factor	Transaction code group (General, Rule 16b3, Derivative, Small (Gift/Trust/By Will), Other, Mix. Since there can be multiple transactions on each Form 4, if a form has less than 50% as one transaction code type, it is labeled 'Mix.
5	Role	Role	Factor	Categorized role of insider. Chief Other includes Chief of Marketing, Chief Operating Officer, etc, Other is any other insider such as Officer, Director.
6	IsTenPercentOwner	Is10Pct	Binary/Logical	A ten percent owner may or may not also be an owner of 10% or more of company stock.
7	Price	Price	Numeric	Price of the stock.
8	NetAmt	NetAmt	Numeric	Price multiplied by NetSharesTraded
9	DaysSinceLast8k	Last8k	Numeric	Days since last 8-K filing.
10	DaysSinceLast10k	Last10k	Numeric	Days since last 10-K filing.
11	DaysSinceLast10q	Last10q	Numeric	Days since last 10-Q filing.
12	Preplanned	Preplanned	Binary/Logical	Indicates if the security part of a 10b5-1 plan. This simply checks the footnotes data for presence of "10b5-1."
13	MarketCap	MarketCap	Numeric	The company's market capitalization.
14	TrdPctMktCap	TrdPctMktCap	Numeric/Percentage	NetAmt / MarketCap
15	NetOwnerSharesTraded30Days	NetOwnrShrTrd30	Numeric	Shares Purchased less Shares Sold by Insider over past 30 days.
16	NumAddtlOwnerTrades30Days	AddtlOwnrTrd30	Numeric	Number of trades by Insider in addition to existing trades over past 30 days.
17	NumOwnerTradesSameDirection30Days	OwnrTrdSameDir30	Numeric	Number if trades by insider in same direction as current filing.
18	NumOwnerTradesOppositeDirection30Days	OwnrTrdOppDir30	Numeric	Number if trades by insider in opposite direction as current filing.
19	NetDirectionOwnerNumTrades30Days	NetDirOwnrNumTrd30	Numeric	Net Number of Trades by Insider over 30 days.

20	NetIssuerSharesTraded30Days	NetCoShrTrd30	Numeric	Shares Purchased less Shares Sold By all insiders in company over past 30 days.
21	NumAddtlIssuerTrades30Days	AddtlCoTrd30	Numeric	Number of trades by all insiders in addition to existing trades over past 30 days.
22	NumIssuerTradesSameDirection30Days	NumCoTrdSameDir30	Numeric	Number if trades by all insiders in same direction as current filing.
23	NumIssuerTradesOppositeDirection30Days	NumCOTrdOppDir30	Numeric	Number if trades by all insiders in opposite direction as current filing.
24	NetDirectionIssuerNumTrades30Days	NetDirCoNumTrd30	Numeric	Net Number of Trades by all insiders over 30 days.
25	MixedDirectionSameDayByOwner	MixedDirOwnr	Numeric	Were there trades in opposite direction same day by insider.
26	MixedDirectionSameDayByIssuer	MixedDirCo	Numeric	Were there trades in opposite direction same day by any insider.
27	R63_REL	R63_REL	Numeric	Response variable. 3 month returns relative to Russell 3000.

3. Data Exploration

3.1 Initial Queries

For the total of 653,299 Insider buys, we see a 3-month relative return of 1.061%.

We'll begin the data analysis process by seeing how filtering the data affects returns. We'll take a look at the two primary factor variables – Role and TranCode.

We'll filter through the Role factor variable. This predictor attempts to categorize the role of insiders based on their corporate title. The Title field is freeform text. To separate the titles into categories, I searched for key words in SQL, labeling chief executives, presidents and chairman as "Chief", Chief Financial Officers (and similar titles) as "CFO", other senior executives with "chief" in their title as "Chief Other" and basically everyone else (VPs, Directors, Officers etc.) as "Other." It is difficult to completely eliminate the noise involved in corporate title, but

perhaps the effort will offer up some insights. Nejat Seyhun (1998) organized his insider groups as Top Executives, Executives, Directors, and Officers. My model mergers directors and officers, but separates CFOs from other Executives, while 'Top Executive' maps closely to 'Chief'.

When we break down this factor, we witness the following returns:

Exhibit 3.1 - Returns by Role

Role	Count	Relative Return
Chief	86,716	1.44%
CFO	44,480	1.31%
Chief Other	39,674	1.23%
Other	482,429	0.96%

These results may indicate a hierarchy of knowledge within firms, with CEO trades seeing subsequent 3 month relative returns of 1.44, and CFO's trailing slightly, at 1.31%. Perhaps with privileged access to financial information, these individuals can make more informed trades.

The other factor variable is TranCode. We see the following results:

Exhibit 3.2 Returns by Transaction Code grouping

TranCode	Count	Relative Return
Genrl	110,521	2.61%
Deriv	6,478	1.98%
Small	2,098	1.56%
16b3	505,517	0.77%
Mix	3,663	0.43%
Other	25,022	-0.08%

The General ("Genrl") category shows a stronger return then we see in alternative categories. Perhaps the most interesting difference is that between General and 16b-3("16b3") trades. General Trades include open market transactions. 16b-3 trades are transactions between the company and insiders. Intuitively, it makes sense that open market buys would outperform the intra-company trades such as compensation-plan based option exercises. The "Other" category literally represents Form 4 trades labeled "Other" as well as those involving tenders or swaps. This was the only category to underperform the Russell 3000.

Exhibit 3.3 groups by Role and Transaction Code group (limiting the results shown to those with counts over 5,000). We can see that CFO/General trades experiencing impressive 3.58% relative returns, and Chief/General at 2.83%. When the role is "Other" and transaction group is "16b-3", we see subsequent relative returns of a mere 0.69%.

Exhibit 3.3 Returns by Role, Transaction Code group

Role	TranCode	Count	Relative Return
CFO	Genrl	6,762.00	3.58%
Chief	Genrl	20,740.00	2.83%
Other	Genrl	78,808.00	2.46%
Chief	16b3	60,396.00	1.06%
Chief Other	16b3	33,195.00	1.03%
CFO	16b3	35,182.00	0.92%
Other	16b3	376,744.00	0.69%
Other	Other	18,781.00	-0.19%

The IsTenPercentOwner variable also provides insights. There is overlap between this category and with Role. An individual may be a company officer as well as a 10% Owner.

The results are as follows:

Exhibit 3.4 Returns for IsTenPercentOwner groups

IsTenPercentOwner	Count	Relative Return
True	22,745.00	2.41%
False	630,554.00	1.01%

While it appears that 10% Owner filings show greater subsequent returns, only 3% of trades are from such individuals. An infinite array of filters can be developed. Those above indicate some of the possibilities and help describe to what degree this data set makes intuitive sense.

3.2 Plots of Non-Normal Distributions

Several of the predictors appear to show non-normal, skewed distributions, some of which have negative values. We will utilize a Yeo-Johnson transformation using the R's preprocess function from the caret package.

Sample plots showing benefits of the transform include the following before and after histograms:

Exhibit 3.5 Last8k

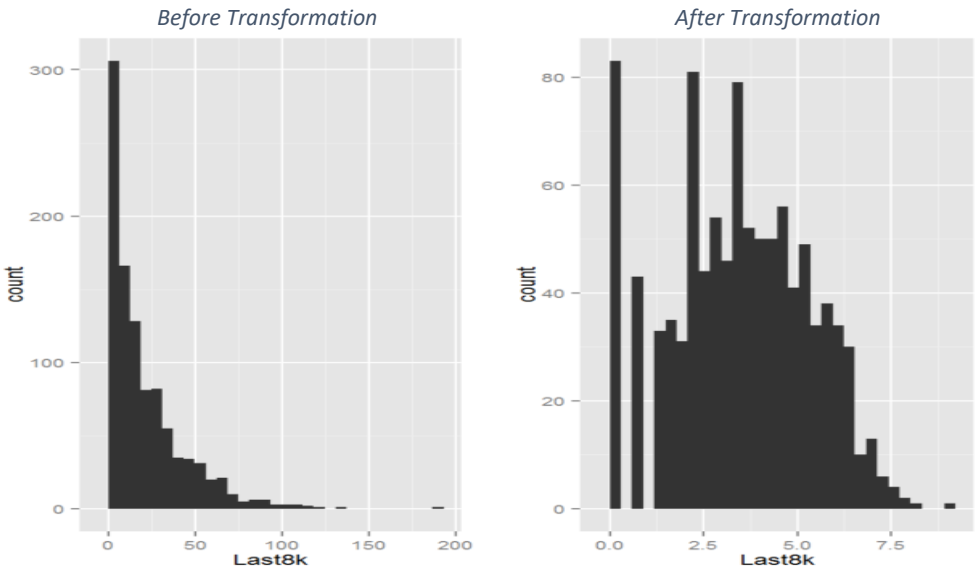
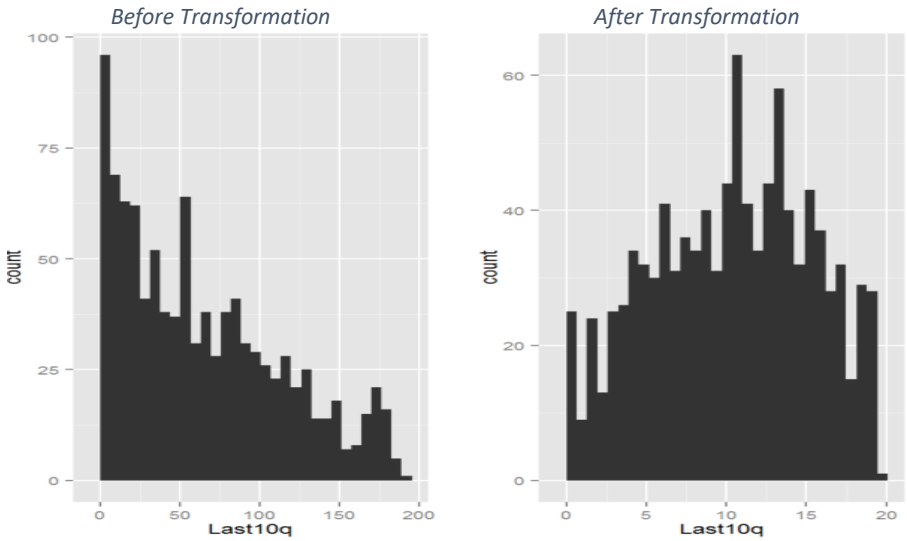


Exhibit 3.6 Last10k

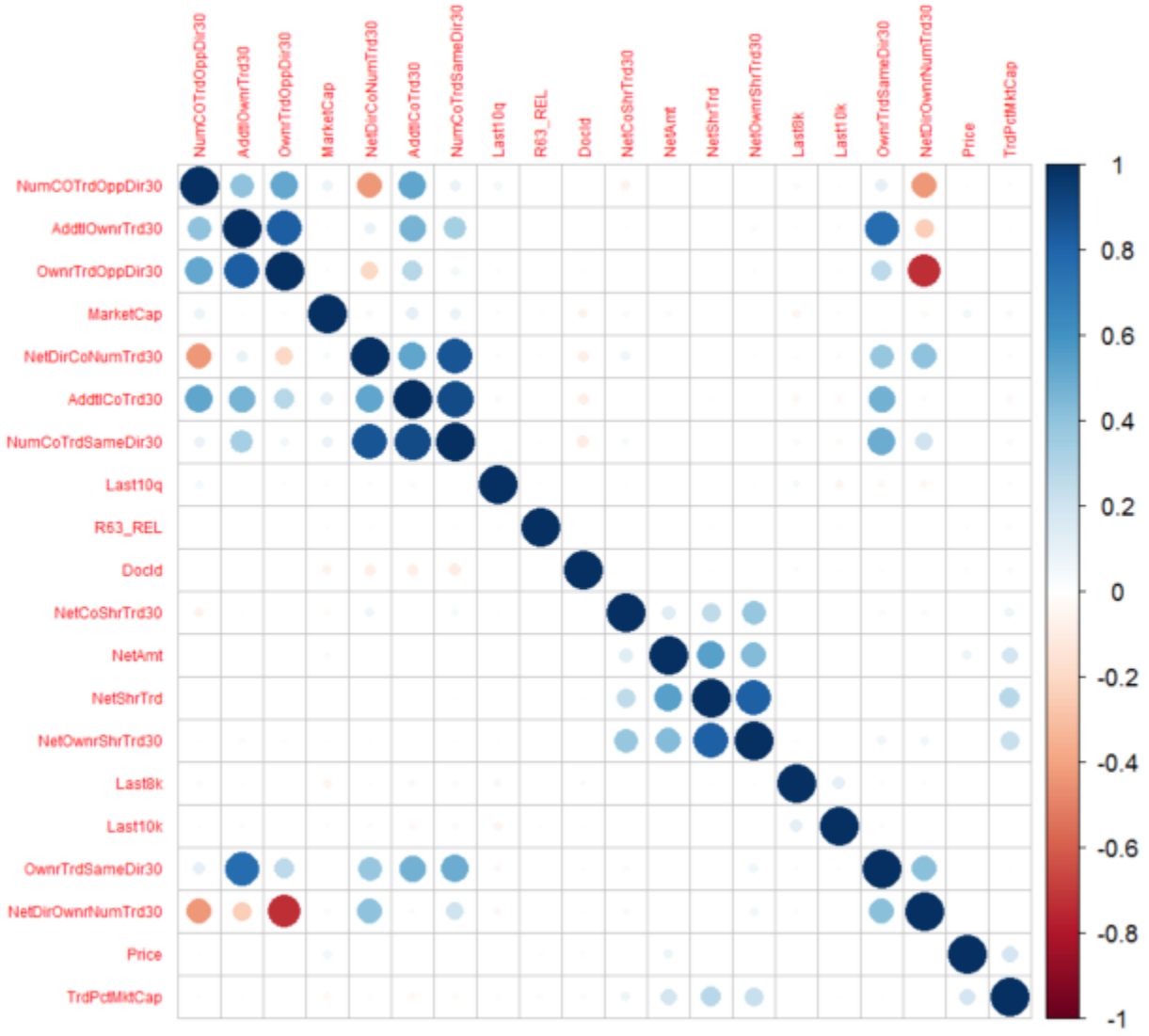


Similar results exist for other predictors, though some do experience fairly normal distributions without transformation.

3.3 Correlation

The following correlation plot summarizes the relationships between the numeric variables:

Exhibit 3.7 Predictor Correlations



Fortunately, we're not seeing too much excessive correlations, although a few predictors, such as NumCoTrdSameDir30 and NetDirCoNumTrd30 are very close to 1.0. It is noteworthy that we are seeing minimal correlation between the predictors and with R63_REL, our response variable. Each of these are very close to zero.

Exhibit 3.8 Correlations with Response Variable R63_REL (3 Month Relative Returns)

Variable	Correlation
DocId	0.0066
NetShrTrd	0.0082
Price	(0.0119)
NetAmt	(0.0009)
Last8k	0.0070
Last10k	0.0126
Last10q	0.0130
MarketCap	(0.0102)
TrdPctMktCap	(0.0165)
NetOwnrShrTrd30	0.0087
AddtlOwnrTrd30	(0.0012)
OwnrTrdSameDir30	0.0027
OwnrTrdOppDir30	(0.0042)
NetDirOwnrNumTrd30	0.0038
NetCoShrTrd30	0.0051
AddtlCoTrd30	0.0069
NumCoTrdSameDir30	0.0104
NumCOTrdOppDir30	(0.0043)
NetDirCoNumTrd30	0.0117

3.4 Highly Correlated Variables

To eliminate issues pertaining to multicollinearity, we will utilize a variable elimination function created by Marcus W. Beck from the "R is My Friend" blog[9]. This will help eliminate unnecessary variables based on their variance inflation factors, or VIFs. With a threshold VIF of 5, 3 variables were eliminated:

AddtlOwnrTrd30
AddtlCoTrd30
NetDirOwnrNumTrd30

3.5 Stepwise Selection for Parameter Elimination

I ran the step function from R's stats package, utilizing both forward and backward selection.

This removed the following variables:

OwnrTrdOppDir30

NetCoShrTrd30

MixedDirOwnr

The final predictors are the following:

NetShrTrd
TranCode
Role
Is10Pct
Price
NetAmt
Last8k
Last10k
Last10q
Preplanned
MarketCap
TrdPctMktCap
NetOwnrShrTrd30
OwnrTrdSameDir30
NumCoTrdSameDir30
NumCOTrdOppDir30
NetDirCoNumTrd30
MixedDirCo

4. Model Construction

4.1 Splitting the Data

The data set consists of Form 4 Filings between 2005-2015. Years 2005-2011 represent the training Set (385,624 rows), while 2012-2013 will be used as validation (126,915 rows) and we'll hold out 2014-2015 for the test set (140,760). We'll start out with the linear model.

4.2 Linear Regression

By utilizing the 18 predictors mentioned above, we'll run a multiple regression using R's lm function. The summary output is as follows:

Exhibit 4.1 Linear Regression Model summary results

```
Call:
lm(formula = R63_REL ~ . - DocId - RptDate, data = trainF4)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1913 -0.1084 -0.0180  0.0795  7.4832

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.541e-03  1.005e-03   2.527 0.011499 *
NetShrTrd          6.123e-09  1.483e-09   4.127 3.67e-05 ***
TranCodeGenrl      2.316e-02  9.809e-04  23.613  < 2e-16 ***
TranCodeSmall     -6.867e-03  6.327e-03  -1.085 0.277779
TranCodeMix       -6.285e-03  5.027e-03  -1.250 0.211186
TranCodeOther     -1.101e-02  1.730e-03  -6.367 1.93e-10 ***
TranCodeDeriv     -2.631e-03  3.630e-03  -0.725 0.468558
RoleChief          4.708e-03  1.062e-03   4.432 9.32e-06 ***
RoleCFO            6.301e-03  1.451e-03   4.342 1.41e-05 ***
RoleChief Other    1.492e-03  1.600e-03   0.932 0.351253
Is10PctTRUE        2.061e-02  2.046e-03  10.073  < 2e-16 ***
Price             -4.988e-06  1.336e-06  -3.734 0.000189 ***
NetAmt            -7.193e-11  2.615e-11  -2.751 0.005939 **
Last8k             5.866e-05  1.548e-05   3.789 0.000151 ***
Last10k            2.375e-05  3.060e-06   7.761 8.45e-15 ***
Last10q            7.586e-05  7.087e-06  10.703  < 2e-16 ***
PreplannedTRUE     1.286e-02  1.872e-03   6.868 6.54e-12 ***
MarketCap         -6.213e-14  1.393e-14  -4.460 8.21e-06 ***
TrdPctMktCap      -7.596e-01  5.924e-02 -12.821  < 2e-16 ***
NetOwnrShrTrd30    1.610e-09  1.089e-09   1.478 0.139292
OwnrTrdSameDir30  -3.973e-03  3.639e-04 -10.916  < 2e-16 ***
```

```
NumCoTrdSameDir30 -1.696e-03  8.111e-04  -2.091 0.036525 *
NumCOTrdOppDir30   2.636e-03  8.702e-04   3.030 0.002448 **
NetDirCoNumTrd30   2.469e-03  8.327e-04   2.965 0.003026 **
MixedDirCoTRUE    -9.996e-03  1.322e-03  -7.564 3.93e-14 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.2203 on 385599 degrees of freedom
Multiple R-squared:  0.003846,    Adjusted R-squared:  0.003784
F-statistic: 62.03 on 24 and 385599 DF,  p-value: < 2.2e-16
```

We can see above that most of the predictors are showing significance at least at the 5% level. For the Role factor, we see that CFO and Chief so significance, while 'Other' does not. This fits well with our intuition. Perhaps the highest level leaders can make more informed decisions when buying in the open market. It's also noteworthy that MarketCap is inversely correlated with 3 month returns. This reflects the view of Seyhun,1999 that insider buys are more informative amongst smaller capitalization firms.

It is also noteworthy that the Adjusted R-Squared is so close to zero, at 0.003784. The residual standard error is 0.2203. While the low R-Squared may be cause for concern, we'll examine the validation set to see how well it performs on unseen data. We see a Root-Mean-Square Error (RMSE) of 0.1787. To further gain clarity, we can break out these scored results into quintiles based on their predicted values, join these predicted returns with the actual returns, and then calculate the actual mean returns, based on their precited rank. We see these results below:

Exhibit 4.2 Mean Actual Returns Per Prediction Quintile, Linear Regression Model, Validation Set

Quintile, Predicted Relative Return	Actual Relative Return
1	0.61%
2	0.37%
3	1.02%
4	0.93%
5	2.09%

The model actually did pretty well at predicting returns. Quintiles 1 and 2 flipped position in actual returns, as did quintiles 3 and 4. Generally, however, predicted return groupings coincided roughly with their actual returns.

Perhaps transforming the model using the Yeo-Johnson method can help improve the RMSE and refine the quintile returns. After performing the transform, the summary results are below:

Exhibit 4.3 Linear Regression Model with Yeo-Johnson transformation, summary results

```
Call:
lm(formula = R63_REL ~ . - DocId - RptDate, data = trainF4_t)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14101 -0.09084  0.00441  0.09629  1.60703

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -4.497e-01  7.039e-03 -63.893  < 2e-16 ***
NetShrTrd          8.269e-03  1.609e-03   5.140 2.74e-07 ***
TranCodeGenrl      2.237e-02  8.688e-04  25.750  < 2e-16 ***
TranCodeSmall      1.116e-02  5.328e-03   2.094 0.036272 *
TranCodeMix        4.917e-03  4.232e-03   1.162 0.245269
TranCodeOther      3.489e-03  1.485e-03   2.349 0.018801 *
TranCodeDeriv      4.676e-03  3.053e-03   1.531 0.125667
RoleChief          5.839e-03  9.207e-04   6.342 2.27e-10 ***
RoleCFO            4.542e-03  1.226e-03   3.704 0.000213 ***
RoleChief Other   -2.062e-03  1.352e-03  -1.525 0.127180
Is10PctTRUE        1.516e-02  1.700e-03   8.916  < 2e-16 ***
Price             -4.675e-02  3.528e-03 -13.249  < 2e-16 ***
NetAmt            -6.569e-03  1.442e-03  -4.554 5.26e-06 ***
Last8k             1.927e-03  1.739e-04  11.083  < 2e-16 ***
Last10k            4.471e-05  2.975e-05   1.503 0.132851
Last10q            1.727e-04  5.381e-05   3.209 0.001333 **
PreplannedTRUE     3.165e-03  1.571e-03   2.015 0.043930 *
MarketCap          3.857e-02  5.619e-04  68.647  < 2e-16 ***
TrdPctMktCap      -3.857e+01  1.979e+00 -19.488  < 2e-16 ***
NetOwnrShrTrd30    4.712e-10  7.601e-10   0.620 0.535310
OwnrTrdSameDir30  -1.640e-02  4.134e-03  -3.967 7.28e-05 ***
NumCoTrdSameDir30  2.818e-03  8.526e-04   3.305 0.000951 ***
NumCOTrdOppDir30  -1.375e-02  1.649e-03  -8.341  < 2e-16 ***
NetDirCoNumTrd30  -2.104e-04  6.950e-05  -3.027 0.002470 **
MixedDirCoTRUE    -5.288e-03  1.114e-03  -4.747 2.07e-06 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.1854 on 385599 degrees of freedom
Multiple R-squared:  0.03901,     Adjusted R-squared:  0.03895
F-statistic: 652.1 on 24 and 385599 DF,  p-value: < 2.2e-16
```

On the surface, not much has changed. There is a slight improvement in adjusted r-squared, to 0.03895 from 0.03784. Residual standard error dropped to 0.1854 from 0.2203.

When scored against the validation set, we see a drop in RMSE to 0.1613 from 0.1787.

As for return groupings, we see the following in exhibit 4.4:

Exhibit 4.4 Mean Actual Returns Per Prediction Quintile, Lin Reg, Transformed, Validation Set

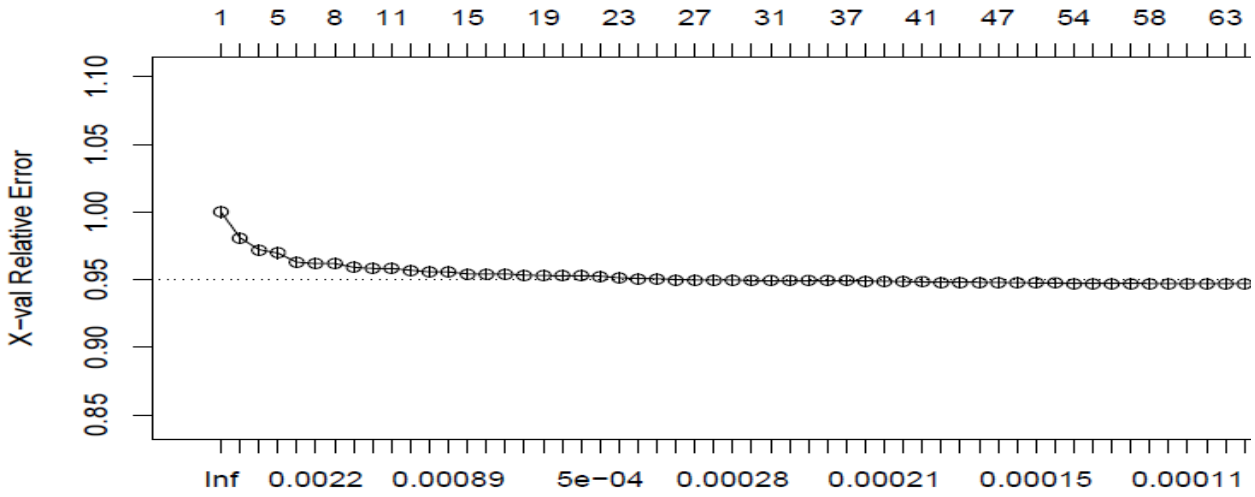
Quintile, Predicted Relative Return	Actual Relative Return
1	-2.21%
2	0.19%
3	1.24%
4	2.21%
5	3.60%

The predicted and actual groupings now line up perfectly. Also, the signal appears stronger, with the spread between bottom and top 20% at 5.81%, vs 2.7% in the original linear model.

4.3 Recursive Partitioning

We'll next utilize a decision-tree based recursive partitioning model. In this case, I will make use of Microsoft's rxDTree algorithm, which works similar to the RPART package in R, but scales much more effectively. It is part of the RevoScaleR package[10]. After an initial fitting with a maximum depth of 15 and maximum buckets (used to determine minimum records for split) of 2000, the following complexity parameter plot is generated:

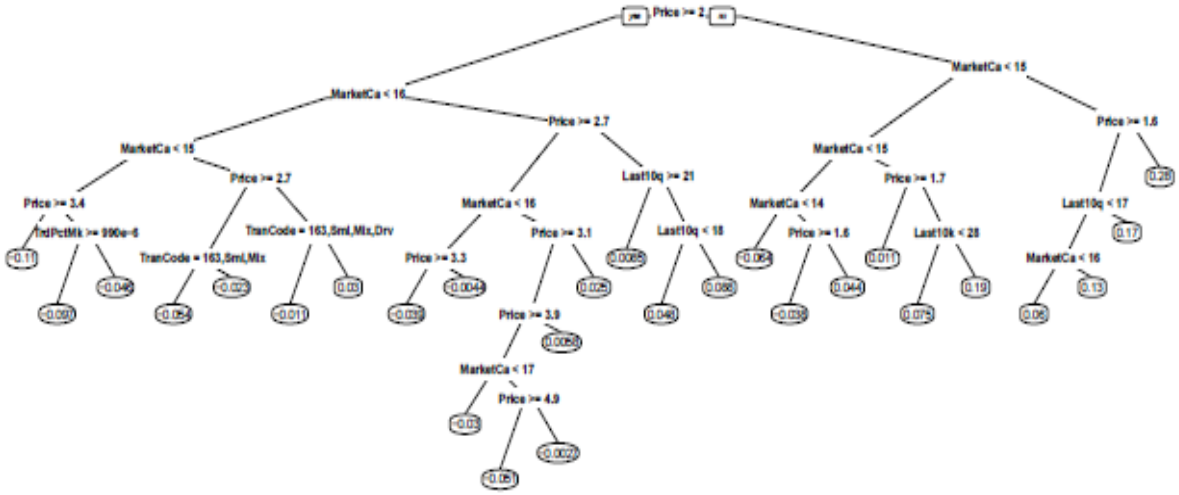
Exhibit 4.5 Recursive Partitioning Relative Error per Tree Size, Complexity Parameter



This helps us tune the parameter and provides a good pruning level for the tree. We can see via the top axis that a tree size of approximately 25, with a complexity parameter of `3.289676e-04` is a good point to set for pruning.

After trimming we have the following tree:

Exhibit 4.6 Tree Generated via Recursive Partitioning



Using R's RPART varImp function, we see the rankings of most important variables:

Exhibit 4.7 Recursive Partitioning Model Variable Importance

Variable	Importance
Last10k	0.0297
Last10q	0.0228
MarketCap	0.1205
Price	0.1614
TranCode	0.0133
TrdPctMktCap	0.0057

It is interesting that the time-distance of company filings such as 10-K's and 10-Q's appear to be important influences on subsequent returns. In other words, informed trading is less likely to occur shortly after 10-K or 10-Q filings.

We see that this recursive partitioning model has a RMSE of 0.1714. As for returns, we see the following quintiles:

Exhibit 4.7 Mean Actual Returns Per Prediction Quintile, Recursive Partitioning, Validation Set

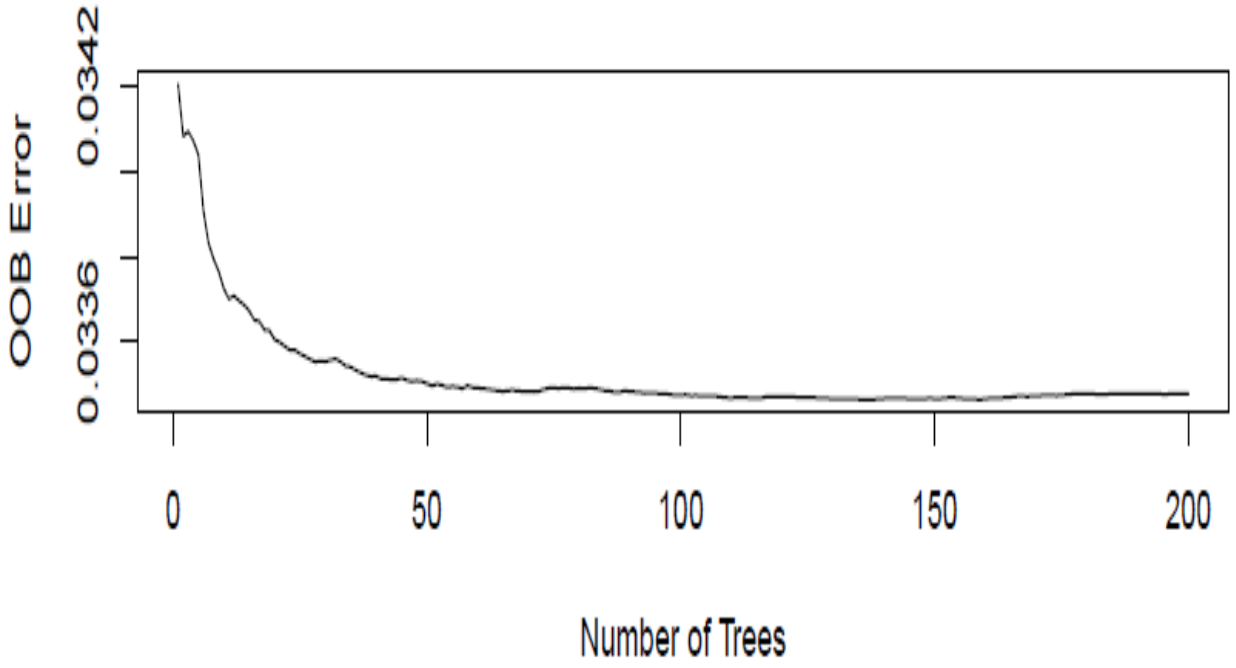
Quintile, Predicted Relative Return	Actual Relative Return
1	-1.52%
2	0.03%
3	0.34%
4	1.57%
5	3.48%

As with the transformed regression model, we see a consistency in the prediction vs actual returns in the validation set. The mean actual return spread between 1st and 5th quintile is 5.00%.

4.4 Random Forest

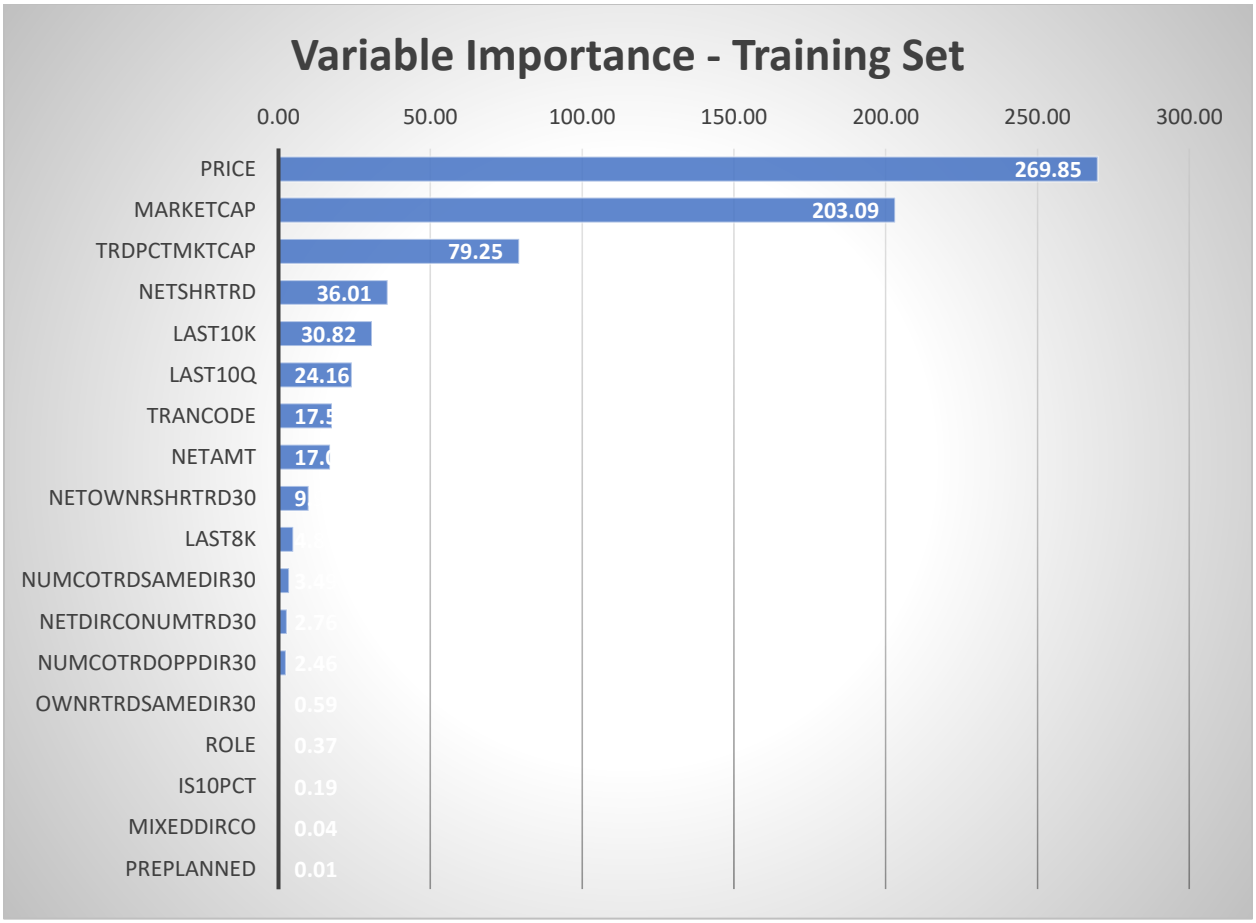
We'll utilize another tree-based method for regression, the random forest. In this case, we'll again use a RevoScaleR algorithm, rxDForest[12], which provides for a highly scalable random forest implementation. Using similar parameters as we used with recursive partitioning (max depth = 12, max bucket = 2000), we generated 200 trees. We can see the reduction in out-of-bag error as the tree grows below:

Exhibit 4.8 Random Forest Model Out of Bag Error, Training Set



As seen in exhibit 4.8, the OOB drops very sharply until around the 50-tree level and improves at a much slower rate. The variable importance output is displayed in Exhibit 4.9 below:

Exhibit 4.9 Random Forest Variable Importance



This model performed quite well, with an RMSE of 0.1608. It also broke out returns effectively be prediction quintile:

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%

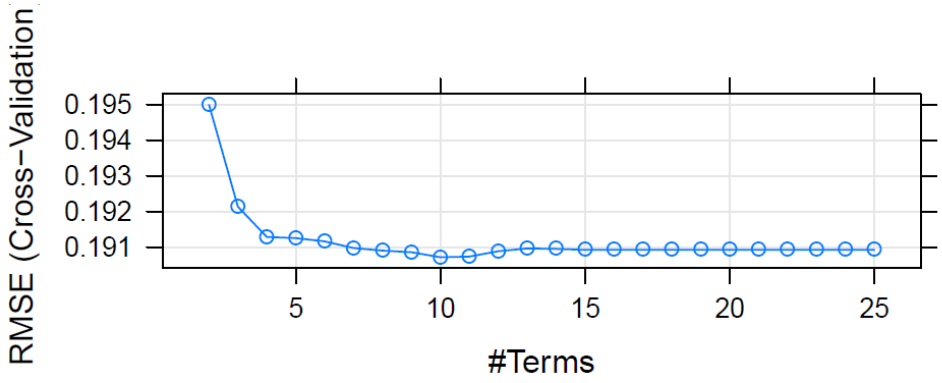
The validation set actual return spread between 1st and 5th quantile was 5.25%.

4.5 Multiplicative Adaptive Regression Splines (MARS)

We'll test one more model using multiplicative adaptive regression splines. This is a non-parametric model known to deal effectively with non-linearity and predictor interactions[12].

We tuned the model using the R caret package's train function, using 5-fold cross-validation to find the best hyperparameters.

Exhibit 4.11 MARS RMSE per # of terms



I then ran the model with degree of 1 and prune set to 10. This model had a RMSE of 0.1713, competitive with recursive partitioning. It's subsequent actual validation returns were as follows:

Exhibit 4.12 Mean Actual Returns Per Prediction Quintile, MARS Model, Validation Set

Quintile, Predicted Relative Return	Actual Relative Return
1	-1.83%
2	0.30%
3	0.70%
4	1.75%
5	4.10%

Interestingly enough, MARS showed the highest disparity in returns between the 1st and 5th quintile, at 5.93%

The results of each of these models was intriguing. We'll now compare the models.

5. Model Comparisons

5.1 Evaluations

The following grid helps to summarize the performance of each of the models:

Exhibit 5.1 Model Comparisons, Validation Set

	Lin Regression	Lin Reg with Y-J Transform	Recursive Partitioning	Random Forest	MARS
RMSE	0.1787	0.1613	0.1714	0.1608	0.1713
Quintile 1	0.61%	-2.21%	-1.52%	-1.29%	-1.83%
Quintile 2	0.37%	0.19%	0.03%	0.02%	0.30%
Quintile 3	1.02%	1.24%	0.34%	0.42%	0.70%
Quintile 4	0.93%	2.21%	1.57%	1.91%	1.75%
Quintile 5	2.09%	3.60%	3.48%	3.96%	4.10%
Q5 - Q1	1.48%	5.81%	5.00%	5.25%	5.93%

Each of these models have their merits. The random forest provided the best fit, but just barely better than the linear model with Yeo-Johnson transformation. Abiding by the principle of simplicity, it is tempting to go with the Linear model with Y-J transformation. Random forests are incredibly complex models which require extensive computation. The highly efficient, rxDForest implementation in this case, with 200 trees built upon nearly 400,00 rows took nearly 2 hours to run. A standard r implementation, randomForest, would have taken much longer. The linear model finished in less than 20 seconds. However, with the random forest model having the

lowest RMSE amongst the validation set, and its tendency to minimize overfitting, I'll choose the random forest model as the winner. Again, each of these models has merits, and further investigation could lead to both better-tuned existing models and/or better alternative models which could improve upon this performance.

5.2 The Winning Model

The winning random forest model was run against the test set (2014-2015). It's RMSE was 0.1694. To provide additional color, I broke out the return results into deciles, again based on the predicted score. The performance this time was more of a mixed bag:

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

It is worth noting that mean insider relative subsequent returns were -1.18% during this period. On a positive note, the top 5 deciles had a mean return of -0.94%, while the lowest 5 deciles has a return of -1.41%. Also, the model did place the worst performing returns in the appropriate bucket, the first decile, but deciles 8 and 10 were barely above the mean, and decile 9 was well below it. I've included Market Cap in this assessment, as perhaps this influential predictor had an effect on predicted returns. The lowest 3 market caps were in deciles in 8,9,10. As there was a negative relationship between returns and market cap, poor performance for small cap stocks would tend to skew returns lower. Overall, the test set results were far less promising than what was seen in the validation set.

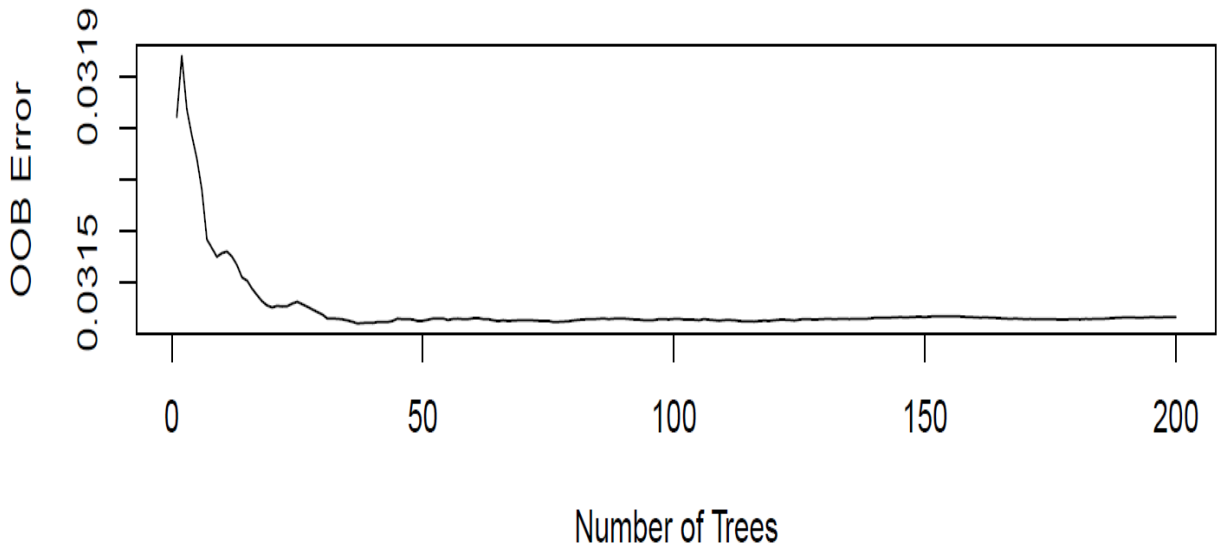
5.3 Combining Training and Validation Sets

It is disappointing to see a drop off in performance between the validation and test sets. Perhaps the strong results in the validation set was due purely to chance. However, it is possible there is a time dependency in the data, whereby the data immediately preceding the test set will yield better results. I will combine the training and validation sets, and retrain the model.

We now see a RMSE of 0.1683, a small improvement from the 0.1694 when run with the training set only.

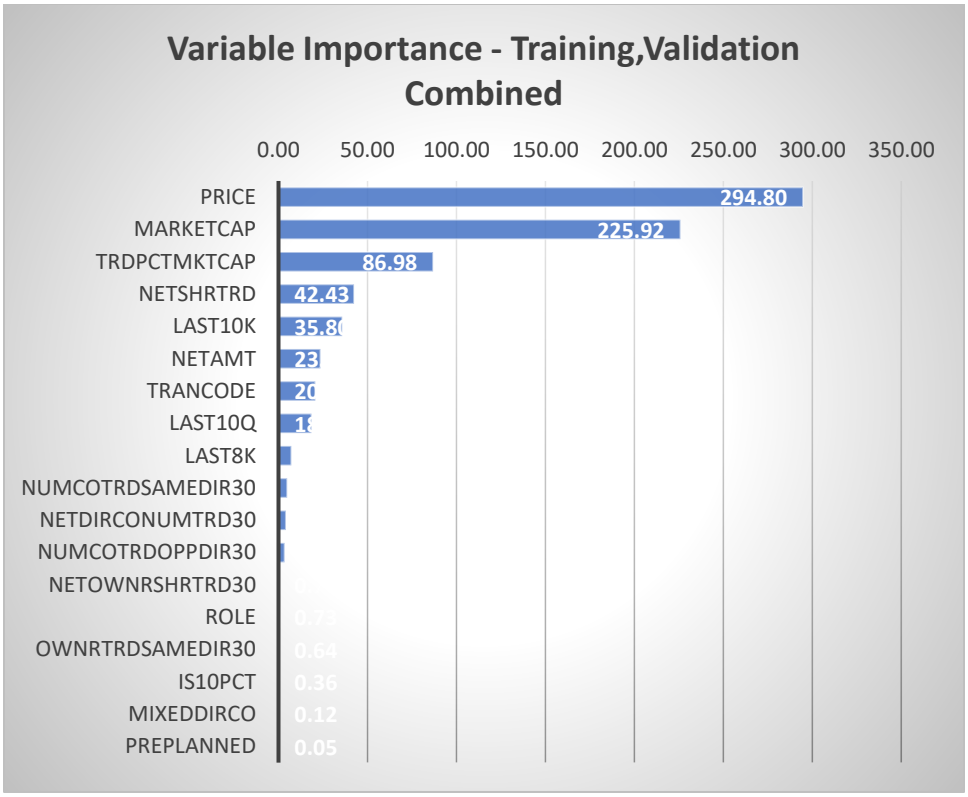
The out-of-bag error improves slightly, with a lower error achieved with less trees:

Exhibit 5.3 Random Forest Model Out of Bag Error, Combined Training and Validation Sets



Variable importance has changed only slightly, as seen in exhibit 5.4. Price, Market Cap, and Trade Percentage of Market Cap continue to be the 3 most important predictors. In fact, we don't see a change in ranking until Net Amount replaces Last10q as the 6th top ranking.

Exhibit 5.4 Random Forest Variable Importance – Combined Training and Validation Sets



When we run this model against the test set, we now see the following returns by decile:

Exhibit 5.5 Mean Actual Returns, Mkt Cap Per Prediction Decile, Random Forest, Combined Train, Val, scored on Test Set

Decile	Relative Return	Market Cap
1	-5.56%	774,064,885
2	-1.91%	6,347,980,804
3	-0.48%	15,026,063,719
4	-0.54%	17,018,007,332
5	-0.43%	20,403,812,722
6	0.03%	14,080,491,110
7	0.10%	10,551,000,000

8	-0.98%	4,738,295,164
9	-1.21%	2,288,693,959
10	-0.32%	777,824,963

This represents an improvement. The model was very effective in identifying the lowest two deciles. 4 of the top 5 deciles were above the median return of 1.18%. Though, the 9th decile was below the median, and deciles 3,4 and 5 were above it. The top five deciles return (-0.48) was 1.31% better than the bottom 5 deciles (-1.79%), a significant improvement over the 0.46% difference we saw when using the training set only. Overall, however, this performance pales in comparison to what we experienced when the training set model was scored against the validation set. Perhaps time-dependency was a factor, but also some degree of luck.

Conclusion

This study explored the viability of utilizing insider trading as a predictor of future stock returns. A database was built directly from the SEC's EDGAR database, and new parameters were derived from these filings to make such predictions. Five machine learning models were evaluated and scored against a validation set with indications of promise in predictability. However, when the winning random forest model was run against the test set, performance appeared much less clear. Perhaps this was due to other factor exposures, or perhaps the viability of an insider-based investment strategy simply is not as clear as it was in prior years.

Further study into the proper application of machine learning models in this domain could shed light on the uncertainties mentioned above. Feature engineering could be refined to model the

behaviors of insiders and future returns more effectively. Clever analysis into the interactions between form 4s and other regulatory filings may prove meaningful in the future.

Overall, there appears to be a relationship between insider filings and future returns, but this judgment is somewhat cloudier than it has been in the past.

References

1. https://www.sec.gov/fast-answers/answersform345htm.html
2. Lee, D. and Verbrugge, J. 1996, "The Efficient Market Theory Thrives on Criticism". Journal of Applied Corporate Finance. 1996
3. Maubossin, M. and Callahan, D. "Alpha and the Paradox of Skill". Credit Suisse. 2013. Retrieved from https://doc.research-and-analytics.csfb.com/docView?language=ENG&format=PDF&source_id=em&document_id=805456950&serialid=LsvBuE4wt3XNGE0V%2B3ec251NK9soTQqcMVQ9q2QuF2I%3D
4. Marks, H. (2014, January 16). Getting Lucky. Oaktree Capital Management. Retrieved from https://www.oaktreecapital.com/docs/default-source/memos/2014-01-16-getting-lucky.pdf?sfvrsn=2
5. Seyhun, N. Investment Intelligence From Insider Trading. The MIT Press. 1998
6. Roberts, E. When Is Insider Selling a Bad Sign? InvestorPlace. Retrieved from http://investorplace.com/2013/09/when-is-insider-selling-a-bad-sign/#.WP8_IYjys2x
7. Lakonishok, Josef, and Inmoo Lee, 2001, Are insider trades informative? Review of Financial Studies 14, 79-111.
8. What is the Rule 10b5-1? Investopedia. Retrieved from http://www.investopedia.com/terms/r/rule-10b5-1.asp
9. Marcus W Beck. Collinearity and stepwise VIF selection. [Web log post]. Retrieved from :https://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/
10. Calaway, R. Estimating Decision Tree Models. Microsoft R Services Guide. Retrieved from https://msdn.microsoft.com/en-us/microsoft-r/scaler-user-guide-decision-tree
11. Calaway, R.Estimating Decision Forest Models.Microsoft R Services Guide. Retrieved from https://msdn.microsoft.com/en-us/microsoft-r/scaler-user-guide-decision-forest
12. Kuhn, Max and Johnson, Kjell. Applied Predictive Modeling. Springer. 2013