Pontifícia Universidade Católica
do Rio de Janeiro

Raphael Alexander Rottgen

Institutional Ownership as a Predictor of Future Security Returns

DISSERTAÇÃO DE MESTRADO

Dissertation presented to the Programa de Pós-Graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Ciências - Informática

Advisor: Prof. Eduardo Sany Laber

Rio de Janeiro
June 2015

Pontifícia Universidade Católica
do Rio de Janeiro

Raphael Alexander Rottgen

Institutional Ownership as a Predictor Of Future Security Returns

Dissertation presented to the Programa de Pós-Graduação em Informática, of the Departamento de Informática do Centro Técnico Científico da PUC-Rio, as partial fulfillment of the requirements for the degree of Mestre.

Prof. Eduardo Sany Laber
Orientador
Departamento de Informática – PUC-Rio

Prof. Marco Antonio Casanova
Departamento de Informática – PUC-Rio

Prof. Ruy Luiz Milidiú
Departamento de Informática – PUC-Rio

Prof. José Eugenio Leal
Coordinator Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, June 16th, 2015

Raphael Alexander Rottgen

Raphael Alexander Rottgen obtained bachelor degrees in economics (specializations in Finance and Decision Sciences) and Psychology from the University of Pennsylvania (Philadelphia, United States). He has seventeen years of financial markets experience acting as a mergers and acquisitions banker for J.P. Morgan and Greenhill \& Co., a Director of Proprietary Trading for Deutsche Bank, an analyst for Gandhara Capital and a private investor, based in New York, London, Sydney, Tokyo, São Paulo and Rio de Janeiro.}

Bibliographic data

Rottgen, Raphael Alexander

Institutional ownership as a predictor of future security returns / Raphael Alexander Rottgen ; advisor: Eduardo Sany Laber. – 2015.
96 f. : il. (color.) ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2015.
Inclui bibliografia

1. Informática – Teses. 2. Big data. 3. Aprendizado de máquina. 4. Modelos de fatores para ações. 5. Investidores institucionais. I. Laber, Eduardo Sany. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

To my advisor Professor Eduardo Sany Laber, for the support, the everyday kindness and the incentive for the realization of this work.

To the people of the Computer Science department at PUC-Rio for the constant help, particularly to Marco Antonio Casanova and Ruy Luiz Milidiú.

To my partner in life Cynthia Pereira de Souza Leusin Rottgen.

Abstract

Rottgen, Raphael Alexander; Laber, Eduardo Sany (Advisor). **Institutional Ownership as a Predictor of Future Security Returns.** Rio de Janeiro, 2015. 96p. MSc Thesis - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Data on institutional ownership of securities is nowadays publicly available in a number of jurisdictions and can thus be used in models for the prediction of security returns. A number of recently launched investment products explicitly use such institutional ownership data in security selection. The purpose of the current study is to apply statistical learning algorithms to institutional ownership data from the United States, in order to evaluate the predictive validity of features based on such institutional ownership data with regard to future security returns. Our analysis identified that a support vector machine managed to classify securities, with regard to their four-quarter forward returns, into three bins with significantly higher accuracy than pure chance would predict. Even higher accuracy was achieved when "predicting" realized, i.e. past, four-quarter returns.

Keywords

Big Data; Machine Learning; Securities Factor Models; Institutional Ownership.

Resumo

Rottgen, Raphael Alexander; Laber, Eduardo Sany. **Uso de Dados das Carteiras de Investidores Institucionais na Predição de Retornos de Ações**. Rio de Janeiro, 2015. 96p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Texto Dados sobre as carteiras de investidores institucionais em ações agora estão disponíveis em vários países e portanto podem ser usados em modelos para prever os futuros retornos de ações. Recentemente, vários produtos comerciais de investimento foram lançados que explicitamente usam tal tipo de dados na construção da carteira de investimentos. O intuito deste estudo é aplicar algoritmos de aprendizado de máquina em cima de dados das carteiras de ações de investidores institucionais nos Estados Unidos, a fim de avaliar se tais dados podem ser usados para prever futuros retornos de ações. Nosso trabalho mostra que um modelo usando um support vector machine conseguiu separar ações em três classes de futuro retorno com acurácia acima da esperada se um modelo aleatório fosse usado.

Palavras-chave

Big Data; Aprendizado de Máquina; Modelos de Fatores para Ações; Investidores Institucionais.

Contents

List of Figures

List of Tables

By three methods we may learn wisdom: first, by reflection, which is noblest; second, by imitation, which is easiest; and third by experience, which is the bitterest.

Confucius, *551-479 BC.*

I
Introduction

This dissertation investigates a broad range of stock ownership features with regard to their predictive validity for future returns of such stocks. Institutional ownership of stocks must nowadays be disclosed in a number of jurisdictions worldwide, e.g. the United States and Brazil, and is hence available as an input factor for prediction models of stock returns. A number of commercial investment products have been launched over the last few years that explicitly purport to use ownership data as an input in their asset allocation methodologies.

The current study uses quarterly U.S. SEC data on institutional ownership between 1Q2004 and 2Q2014. We developed routines to scrape the relevant data from the SEC website, including extensive plausibility checks, and to insert it into quarterly ”investors x securities” matrices. Based on this preprocessed data, we derive a number of ownership-related features, including some that, to our knowledge, have not been investigated so far, e.g. entropy of holdings. We use the features to train a variety of models. Here, we also depart from the standard ordinary least squares multiple regression and employ a range of contemporary statistical learning techniques, including support vector machines and decision trees.

Our support vector machine classifier achieved an accuracy of almost 0.38 when classifying securities into three bins with regard to their expected return over the next four quarters. Given that the true classes were balanced, this performance is clearly significantly above pure chance.

The remainder of this dissertation is organized as follows: chapter II provides a brief review of key existing research with regard to security return models as well as a summary of the commercial investment products that purport to use ownership-related factors; chapter III explains our methodology, including dataset source, data pre-processing, predicted variable, features, and learning algorithms; chapter IV presents the results of our learned prediction models; in chapter V, we offer our conclusions as well as suggestions for future research.

II Background

II.1 Related Research

(a) Factor Models of Security Returns

Myriad features (or factors) can be proposed, calculated, and used in factor models of security returns, and this has indeed been and continues to be a rich field of academic research. The original Capital Asset Pricing Model (”CAPM”) of Sharpe [16], Lintner [13] and Black [2] explained expected security returns via a single factor, the famous β (the slope of the regression line of a security's return upon the market return). Mounting empirical evidence against CAPM over the years has led to the development of multi-factor models of security returns (see e.g. [9] for a brief review of such contradictory evidence). E.g., Fama and French [9] find that two factors, size (as measured by a security's market capitalization) and the ratio of book equity to market equity, explain cross-sectional variation in average U.S. stock returns from 1963-1990. Haugen & Baker [10] define five factor categories (risk, liquidity, price level, growth potential, stock price history) and investigate a total o 71 factors from these categories as predictors for stock price returns in five countries (U.S., Germany, France, United Kingdom, Japan).

We could imagine an even broader list of factors, along the lines of the following categories:

– macroeconomic data - e.g. GDP, inflation, unemployment, current account data;

– macro market data - e.g. interest rates, commodity prices, foreign exchange rates;

– market technicals - e.g. absolute and relative index levels, volumes, put-call ratios, mutual fund cash balances, number and size of offerings;

– industry sector data;

- company fundamentals - e.g. revenue growth, margins, return ratios, debt ratios, cash flow measures, operating statistics;
- stock fundamentals - e.g. valuation ratios;
- stock technicals - e.g. absolute and relative price and volume levels, short interest, put/call ratio;
- stock ownership data – which we shall explore in more detail in this study.

(b) Research focusing on Ownership Factors

The current study focuses on analyzing data regarding the holdings of institutional investors in stocks, which falls somewhere between fundamental and technical data. With regard to this ownership data, there are several classes of factors/features that can be derived, e.g. with regard to:

- institutional ownership magnitude;
- institutional ownership dispersion;
- institutional owners' characteristics; and
- ownership interaction with other variables (principally, security prices).

There are a number of studies investigating ownership factors in various contexts. Although only a few explicitly attempt to use ownership factors as predictors of stock returns, the various studies at a minimum provide us with ideas for factors to include in the present study.

Lakonishok, Shleifer and Vishny [12] investigate whether a sample of 769 U.S. funds (in 1985-1989) exhibits "herding" (simultaneously buying or selling the same stocks as other funds) and/or "positive-feedback trading" (buying past winners / selling past losers). They find little evidence for either, especially when considering large-cap stocks. However, the two phenomena still make for legitimate features to include in the present study in some form. They also point out that the two phenomena may exist after all, but in subgroups, e.g. subgroups of investors (rather than the broad universe of investors their study considers). It is worth noting that the features used in the present study effectively allow the learning algorithms to consider such subgroups. Finally, the authors do find statistically significant excess returns for stocks that were bought, on net, during a calendar quarter by the funds – however, they investigate the relation within the same quarter, invalidating the use of this result for prediction.

Several other studies pick up on the topic of herding, e.g. Choi and Sias [4] found evidence of institutional herding at the industry level (i.e. institutional investors tend to follow each other buying into / selling out of a given industry sector). However, this and similar studies do not investigate validity of herding as a predictor of future stock returns, nor do they introduce further ownership features that could be of use in the present study.

Dasgupta, Prat and Verardo's [5] 2011 study focuses on multi-quarter (say, 3-5 quarters), persistent institutional trading of U.S. portfolio managers from 1983-2004 and concludes that it negatively predicts long-term (about 2 years') stock returns. They also find that this effect is concentrated among smaller stocks (in line with Lakonishok, Shleifer and Vishny) and stronger for stocks with high institutional ownership.

Chen, Hong and Stein [3] analyze breadth (defined as the ratio of the number of funds having a long holding in a stock to the total number of funds) of mutual fund ownership in U.S. stocks from 1979-1998. They find "that those stocks whose change in breadth in the prior quarter places them in the lowest decile of the sample underperform those in the top change-in-breadth decile by 3.82% in the first six months after portfolio formation and by 6.38% in the first 12 months." They also analyze a metric representing the change in the fraction of total shares outstanding of a stock that is owned by mutual funds, but find that this metric becomes statistically insignificant once it is added to the change-in-breadth metric. The present study includes both of these ownership-related metrics as features.

Barabanov [1] finds that, for NASDAQ stocks in 1983-2000, future stock returns (1, 4, and 12 quarters forward) are positively related to levels of institutional ownership (percentage of total shares outstanding held by institutions) and negatively related to concentration of ownership (using the Herfindahl index as a metric). Barabanov also considers the first derivatives of these two variables. He finds that quarterly changes in the level of institutional ownership are, on average, positively related to returns for the following quarter and returns for the following year, but negatively related to three-year returns, and suggests that this effect may be related to the institutional investors' average holding periods (the present study includes a feature representing average holding period). In contrast, his study finds that changes in concentration of total institutional ownership are mostly negatively related to future quarterly and yearly returns and significantly positively related to three-year returns. It is worth noting that Barabanov also considers a classification of investors by type (e.g. banks, insurance companies, independent investment advisors). The present study does not replicate this type of feature, partly due to the fact that it is not possible to conduct such a classification automatically based on our main data source, the SEC 13-F filings.

It is worth pointing out that, even where studies have found effects, while coefficient estimates tend to be statistically significant given the sheer amount of data, the results are typically rather weak with regard to explaining the variance in forward stock returns. E.g. Chen, Hong and Stein [3] conducted univariate regressions of forward returns on their two ownership-related predictors (as described above) and found R^2 statistics of between 0.7-1.2%. Barabanov [1] conducted quarterly multiple regressions, using the institutional ownership and concentration metrics described above, as well as ten additional non-ownership-related predictors, to predict forward returns, and found R^2 statistics of between 3.7-5.8% (averaged over the quarters).

In the following section, we turn our attention from academic research to a practical implementation: investment products that explicitly attempt to use ownership features to achieve superior investment returns.

II.2 Practical Application: Replicator ETFs

"Liquid alternatives" are currently one of the fastest-growing types of financial products. They are typically investment products that seek to offer return and risk characteristics approaching those of hedge funds, but with lower fees, higher liquidity, and accessible to a wider investor universe. One specific type of liquid alternatives that has emerged in the recent past (the below examples are from 2012 or later) are indices and exchange-traded funds ("ETFs") that aim to replicate hedge fund equity positions [6]. As these replicator ETFs use security ownership-based features in order to construct portfolios to maximize returns, they are essentially a practical implementation of the subject of the current study. As we can construct our own (virtual, for now) replicator ETF from the trained models of the current study, these replicator ETFs could also serve as benchmarks for a practical evaluation of our models.
For this dissertation, we have reviewed four examples of replicator indices and ETFs that are already available for investing:

- AlphaClone Hedge Fund Long/Short Index (ALFA)
- Direxion iBillionnaire (IBLN)
- Solactive Guru Index (GURU)
- Solactive Hedge Fund Holdings US Index (HEDGEUS)

We provide background information on all four products in the appendix, in section A. It is worth noting that, as at the time of writing this dissertation, further such investments products are being developed, e.g. by Goldman Sachs.

For the purposes of the present study, one specific area of interest with regard to these products was to understand the ownership-related features and methodologies that the product sponsors use to arrive at their investment decisions. Upon review of the publicly available information on the products, we find that with regard to ownership-related features, the products consider such factors as:

- number of hedge funds (as defined by the product sponsor) holding the shares of a given company;
- hedge fund ownership fraction – the number of shares held by "hedge funds" (as defined by the product sponsor) over the number of total shares outstanding of a given company;
- ownership fraction change – essentially, the change in the variable mentioned immediately above;
- "clone score" – essentially, a score, for each stock, based on the average past returns achieved by investors holding the stock.

The present study includes variations of all of these features, as we will detail in section III.3. It is worth noting that, besides ownership-related features, all products also have minimum size (in terms of market capitalization) and liquidity (in terms of trading volume) requirements, and they exclusively invest in U.S. stocks.

With regard to "methodology", i.e. how the product sponsors use these features, the publicly available information does not always provide much detail, but it appears to be essentially a simple decision tree-type approach in all cases.

II.3 Statistical learning techniques in academic finance

All of the studies cited above use ordinary least squares regression, likely reflecting the assumption of linearity with regard to the relationships between financial variables, as well as a modus operandi of academic finance that started with an article by Fama and Macbeth in 1973 [8], when many of the modern statistic learning techniques described in section III.5 had not yet been formulated. There are exceptions to the widespread use of OLS regression, though, e.g. Eakins, Stansell and Buck, in a study investigating various financial characteristics of firms as predictors for institutional ownership [7], test various neural network models, pointing out the non-linear relationships between their predictors and predicted variable.

In the present study, we add several other statistical learning techniques to regression for two principal reasons. First, from a theoretical point of view, regression results are only valid if a number of assumptions hold with regard to the underlying data (such as a linear relationship) which may quite simply not be the case for our data.

Second, from a practical point of view, we are interested in the predictive validity of our models, as measured by their performance on a test set, and are therefore agnostic about the selection of the model as long as it performs well (notwithstanding this last comment, if we end up with two or more models with similar performance level, we shall, as customary, prefer the one that shows greater parsimony and ease of intuitive understanding).

Our inclusion of a broad range of modern statistical learning techniques (or "machine learning" techniques) is part of a recent trend of the application of these techniques to an ever broader range of problems. We are convinced that academic finance, too, will increasingly embrace these methodologies, given their flexibility and performance.

III
Methodology

III.1 Dataset

Our dataset is of tabular/matrix format with securities in the rows and features related to the securities in the columns. The majority of the features are based (either directly or derived) on institutional ownership data (e.g., as the simplest example, number of funds holding the security in a specific calendar quarter). Some further features, specifically security prices and the number of shares outstanding of the security, are not based on the institutional ownership data, but are needed to calculate the predicted variable as well as some of the features.

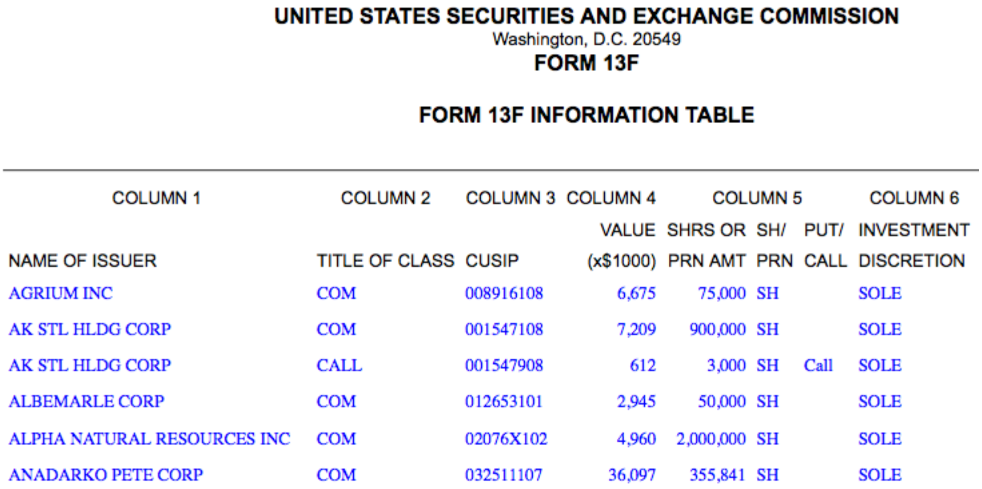
As institutional ownership data is available by calendar quarter, a full set of features for each security is also available on the same quarterly basis.

(a) Data Sources

Institutional Ownership Data

Our institutional ownership data is based on the Form 13F filings made publicly available by the United States Securities and Exchange Commission (”SEC”). Institutional investment managers that exercise investment discretion over US$100 million or more in Section 13(f) securities (generally, securities listed on a U.S. exchange) are required to make 13F filings within 45 days of the end of each calendar quarter. Further details with regard to the Form 13F are available on the site http://www.sec.gov/answers/form13f.htm. The 13F filings themselves are also available on the SEC website: http://www.sec.gov/cgi-bin/srch-edgar. The 13F filings contain descriptive information on the filing entity (i.e. institutional investor / fund), such as name and address, the total number of investment holdings, the total value of the holdings, as well as table providing the individual investment holdings, including number of shares held of each security.

An excerpt of such a table, based on the third quarter, 2014, 13F filing by investor 3G is reproduced below.



UNITED STATES SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549
FORM 13F

FORM 13F INFORMATION TABLE

COLUMN 1	COLUMN 2	COLUMN 3	COLUMN 4	COLUMN 5			COLUMN 6
			VALUE	SHRS OR	SH/	PUT/	INVESTMENT
NAME OF ISSUER	TITLE OF CLASS	CUSIP	(x$1000)	PRN AMT	PRN	CALL	DISCRETION
AGRIUM INC	COM	008916108	6,675	75,000	SH		SOLE
AK STL HLDG CORP	COM	001547108	7,209	900,000	SH		SOLE
AK STL HLDG CORP	CALL	001547908	612	3,000	SH	Call	SOLE
ALBEMARLE CORP	COM	012653101	2,945	50,000	SH		SOLE
ALPHA NATURAL RESOURCES INC	COM	02076X102	4,960	2,000,000	SH		SOLE
ANADARKO PETE CORP	COM	032511107	36,097	355,841	SH		SOLE

Figure III.1: Example of a 13F filing (excerpt)

Note that in this study, each security is identified via its "CUSIP" (Committee on Uniform Security Identification Procedures) code, as this is an unambiguous identifier, in contrast to e.g. security name or ticker symbol (ticker symbols may belong to different companies over time or change e.g. due to a company changing its name).

Other Data

In order to calculate the predicted variable as well as some of the predictors, we also need historical prices and numbers of shares outstanding for each security. This information was obtained from FactSet Research Systems Inc. ("FactSet"), a commercial provider of financial information.

(b) Data Processing

A number of steps were undertaken to obtain and "wrangle" the raw institutional ownership data into a suitable input format for machine learning. A certain amount of filtering and error-checking was undertaken along the way, as explained in what follows. No comparable processing was performed for the price and share number data from FactSet given its simpler structure and commercial nature (which, optimistically, implies that e.g. error-checking should already have taken place).

Ownership Data Collection

Routines in Python (including the Pandas and Numpy libraries) were developed in order to browse all 13F filings in a given calendar quarter. For each investment position of an institutional investor (e.g. fund A holds 100 shares of Apple, worth $500,000 as of the filing date, 1000 shares of Microsoft ... and so on), the routine extracted the number and value of shares held and inserted them into dataframe objects, with securities in the rows and investors in the columns, yielding such dataframes for each calendar quarter investigated.

The study used 13F filings starting from the quarter ending March 31st, 2004, through the quarter ending June 30th, 2014 (i.e. 42 quarters in total). Before 1Q04, the 13F security lists used as part of error-checking (see below) are not available in machine-friendly format. It is also worth pointing out that, as of 2Q13, 13F filings are made in XML rather than simple text format, substantially reducing the potential for some of the errors discussed below. While an XML parser (Beautiful Soup) was used for the filings in XML format, filings in text format were analyzed substantially with the help of regular expressions.

The routines took on average between 3-4 seconds per filing. Given the large number of filings in the analysis period (127,395), we ran the routines on several (up to eight) Amazon EC2 (elastic cloud) instances in parallel in order to save user (but obviously not computing) time.

Ownership Data Filtering

The following filters were implemented with regard to the dataset.

Security type We excluded investment positions in options, fixed income securities and convertibles from the analysis. This is substantially due to the fact that reliable historical price quotes are often not available for these types of instruments. Options and fixed income instruments can be easily identified via their special CUSIP format. Convertible identification was attempted via regular expressions. Over the entire period (1Q04 - 2Q14), approximately 2.2% and 0.3% of total initially scraped investment positions were excluded due to being options and convertibles, respectively. While our script did not calculate the isolated reduction of total investment positions due to the exclusion of fixed income positions, it did calculate that, over the entire period, approximately 7% of securities were excluded for this reason. As fixed income securities are typically not widely held, the percentage reduction in the number of investment positions maintained for analysis should be significantly smaller than 7%.

Minimum average position size Only 13F filings of institutional investors with a minimum average position size of 0.25% of total position value (equivalent to a maximum restriction of 400 investment positions) were included in the analysis. This filter is a domain expertise-based proxy to try to filter out e.g.

- large institutions that make one filing mixing a number of funds with different strategies and/or securities held for end customers in a non-discretionary function;
- funds running passive investment strategies; and
- quantitative high-frequency trading-focused funds who typically hold thousands of positions but for very short timeframes (making the quarterly filing data irrelevant).

Clearly this crude approach is inferior to e.g. a manual filtering of funds each quarter by a domain expert, but reflects the time constraints of the implementation of the study.

Minimum total holdings value Only institutional investors with a minimum of US$100 million of reportable holdings are required to make 13F filings. We decided to implement a higher minimum cut-off of US$200 million in order to attempt to exclude short-lived investors or investors that "drift in and out" of the mandatory reporting zone, as well as to reduce the total number of investors in order to reduce computational time. This filter excluded approximately 24% and 23% of total reporting investors in 2Q14 and 1Q14, respectively.

Ownership Data Errors

Errors in the dataset can occur due to: errors in the filings themselves or errors resulting from incorrect scraping of the filing data.

Some errors are detectable via e.g. cross-referencing of data or plausibility checks, while others are not, or at least not easily so, e.g. if a share number and/or value were incorrectly entered in the filing in the first place. Detectable errors may or may not be fixable, e.g. it may be detected that a filing has lines with the security identifier (CUSIP) missing, but the CUSIPs may not be easily retrievable in alternative ways.

The below table show some typical errors along with potential remedies.

ERROR	REMEDIES
Incorrect share number	Plausibility check on total share number Check share price implied by filing data against actual historical share price
Incorrect shareholding value	Plausibility check on total market cap. Check implied share price OBS.: not an actual problem as holding value can in any event be calculated using actual historical prices
Incorrect CUSIP identifier	Check against SEC CUSIP list If CUSIP not easily matched, use string similarity
Missing numbers in filing	Results in invalid data row that is flagged in error log
Concatenated numbers in filing (e.g. share no. & holding value)	Some concatenations can be resolved Otherwise, concatenation will result in invalid data row that is flagged in error log

The following paragraphs detail some of the routines that were used to attempt to eliminate errors.

Error log For every quarter of filings that is scraped, an error log is generated, showing the filing URL as well as further relevant details for the following errors:

- number of scraped securities different from total number of securities that is usually (but not always, and not always correctly) given at the beginning of the filings themselves
- implausibly high shareholding value
- implausibly high number of shares
- empty security table

The total number of errors detected, as a percentage of the total number of investment positions scraped per quarter, averages between 2-3%. In spot checks performed on some quarters, it was found that the majority of errors flagged are differences in the number of scraped securities vs. the number of securities indicated in the filing. A majority of these differences have been found to be due to:

- presence of foreign securities that do not have a valid CUSIP (or no CUSIP at all) and hence do not get scraped; and
- total number of securities given in the filing being incorrect and not corresponding to the actual number of securities listed.

CUSIP validator routine The CUSIP validator script attempts to validate CUSIPs scraped from filings by comparing them against the CUSIP list that is provided quarterly by the SEC. The script automatically fixes CUSIPs that are missing leading zeroes or a final checksum digit. If, post these automatic fixes, a CUSIP can still not be validated against the SEC list, a series of secondary checks are performed and the CUSIP may be automatically or manually accepted or rejected:

- if the security name associated with the CUSIP (from the scraped data) is sufficiently similar (as measured by Levenshtein string similarity) to a security name in the SEC CUSIP list, the corresponding CUSIP from the SEC list is used;
- if the CUSIP translates to a valid stock ticker, the CUSIP is assumed to be valid and kept;
- if more than three holders for the CUSIP exist and its length is valid, the CUSIP is assumed to be valid and kept;
- otherwise, a variety of information (similar security names, based on Levenshtein string similarity, and their CUSIPs; number of other holders of the security) is displayed for the user to make a decision on whether to keep, alter, or delete the CUSIP.

The CUSIP validator script also automatically deletes any fixed income securities, for the reasons explained above.

Implied share price check As part of the scraped information from the 13F filings, for each investment position that an investor holds in a security, the number of shares held and the Dollar value of the position is obtained and stored. This allows for the calculation of an implied share price as of the date of the filing. Clearly, such implied prices should be substantially identical for the same security on the same date, otherwise we can conclude that there is an error with a share number, a Dollar value number, or both. A script goes through every security, for every filing quarter, comparing the prices implied for the security by the investment positions. Investment positions whose implied price deviates from the median implied price by more than 5% are deleted.

Investment positions where there is only one holder in a security are also deleted.

Total share number check For each security, in every quarter, an automatic routine added up the number of shares held by the institutional investors and compared this total with the total number of shares outstanding as reported by FactSet. Where the total shares held by institutions exceeded the total number of shares outstanding, an error was included in an error log, for manual inspection. The number of security-quarters with such errors was reported by the routine to be 241, or less than 0.01% of total security-quarters in the dataset.

Securities with zero holdings In every quarter, securities may exist that have zero holders. Such securities, for obvious reasons, do not even show up in the quarter's 13F filings. In order to prevent the dataset from showing ”NAs” for the number of holders for such these securities, we conduct a cross-check to see whether a security had an existing market capitalization in a given quarter – if affirmative, the number of holders is set to zero for the quarter. If a missing market capitalization number indicates that the security indeed did not exist (at least as a publicly traded security) in the quarter, the number of holders is left as ”NA.” This cross-check also impacts other features whose calculation is based on the number of holders, e.g. fraction of institutional investors, as well as the ”first derivative” features based on these features (e.g. change in holders over one, two, and three quarters). Where a change calculation would result in infinity (e.g. change in holders for a security that does have holders in the current quarter, but did not have any in the previous quarter), the change is arbitrarily set to a high number (1000%).

(c) Data Summary

A total of 127,395 filings were analyzed, covering the period from March 31st, 2004 through June 30th, 2014, representing a total of approximately 6.2 million investment positions (post the application of the filters and error-correction mechanisms described above). The size of our dataset is determined by the number of ”security-quarters”, i.e. the sum, over all quarters, of the total number of securities in each quarter. Each such security quarter represents a row in our final dataset, with columns representing the features. Note that such rows may not be completely filled in as some of the features include backward-looking data (over up to three quarters) in their calculation, and not all securities appear in every quarter. The total number of security-quarters is 243,446. Clearly, many securities appear in several quarters (and some in every quarter). The total number of unique securities is 15,360.

We also present here some summary information on the number of investors, albeit it is not of direct relevance to the current study. Over the entire period, there are a total of 4,495 unique investors. The median number of quarters filed per investor are 9. The below histogram illustrates the frequency of occurrence of a fund filing a given number of quarters.

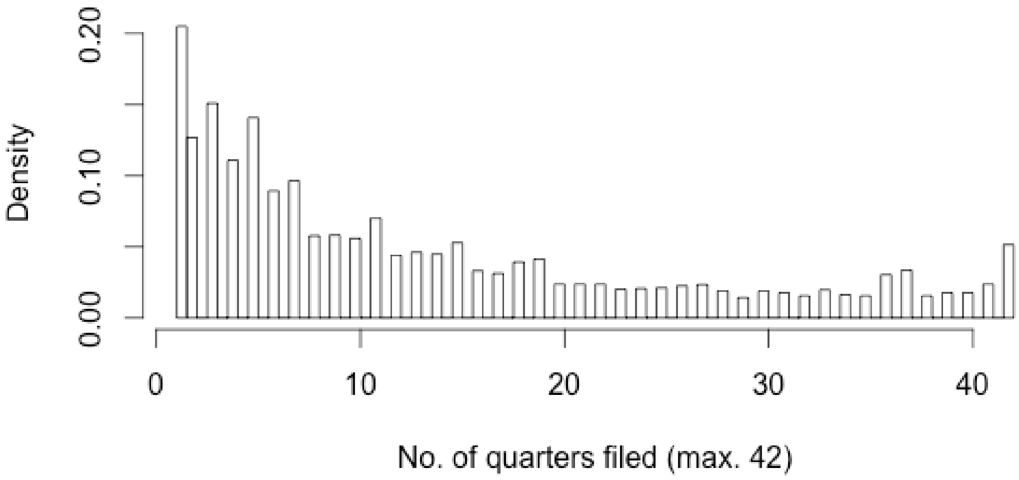
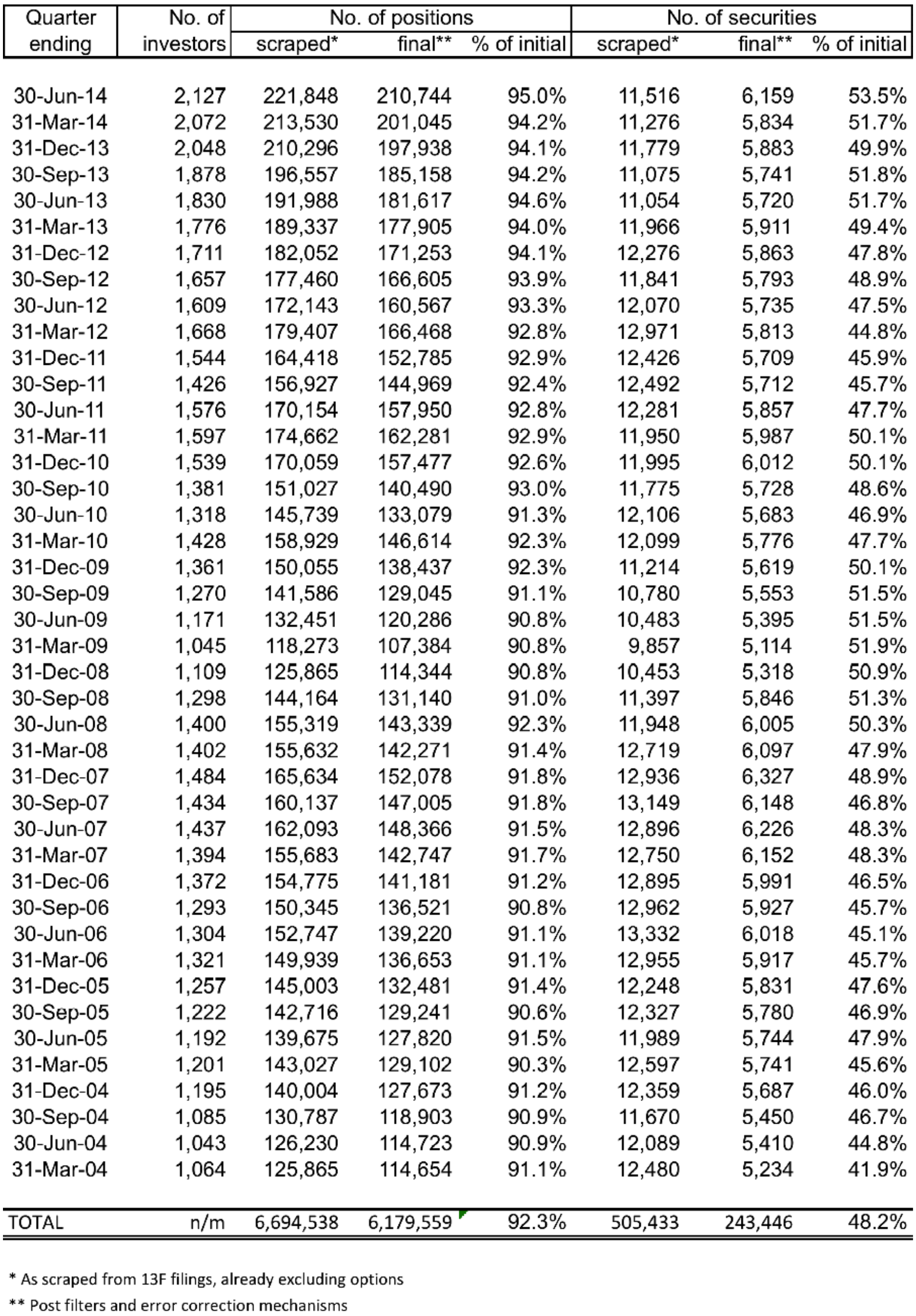


Figure III.2: Histogram: filing frequency (no. of quarters)



Quarter	No. of	No. of positions			No. of securities		
ending	investors	scraped*	final**	% of initial	scraped*	final**	% of initial
30-Jun-14	2,127	221,848	210,744	95.0%	11,516	6,159	53.5%
31-Mar-14	2,072	213,530	201,045	94.2%	11,276	5,834	51.7%
31-Dec-13	2,048	210,296	197,938	94.1%	11,779	5,883	49.9%
30-Sep-13	1,878	196,557	185,158	94.2%	11,075	5,741	51.8%
30-Jun-13	1,830	191,988	181,617	94.6%	11,054	5,720	51.7%
31-Mar-13	1,776	189,337	177,905	94.0%	11,966	5,911	49.4%
31-Dec-12	1,711	182,052	171,253	94.1%	12,276	5,863	47.8%
30-Sep-12	1,657	177,460	166,605	93.9%	11,841	5,793	48.9%
30-Jun-12	1,609	172,143	160,567	93.3%	12,070	5,735	47.5%
31-Mar-12	1,668	179,407	166,468	92.8%	12,971	5,813	44.8%
31-Dec-11	1,544	164,418	152,785	92.9%	12,426	5,709	45.9%
30-Sep-11	1,426	156,927	144,969	92.4%	12,492	5,712	45.7%
30-Jun-11	1,576	170,154	157,950	92.8%	12,281	5,857	47.7%
31-Mar-11	1,597	174,662	162,281	92.9%	11,950	5,987	50.1%
31-Dec-10	1,539	170,059	157,477	92.6%	11,995	6,012	50.1%
30-Sep-10	1,381	151,027	140,490	93.0%	11,775	5,728	48.6%
30-Jun-10	1,318	145,739	133,079	91.3%	12,106	5,683	46.9%
31-Mar-10	1,428	158,929	146,614	92.3%	12,099	5,776	47.7%
31-Dec-09	1,361	150,055	138,437	92.3%	11,214	5,619	50.1%
30-Sep-09	1,270	141,586	129,045	91.1%	10,780	5,553	51.5%
30-Jun-09	1,171	132,451	120,286	90.8%	10,483	5,395	51.5%
31-Mar-09	1,045	118,273	107,384	90.8%	9,857	5,114	51.9%
31-Dec-08	1,109	125,865	114,344	90.8%	10,453	5,318	50.9%
30-Sep-08	1,298	144,164	131,140	91.0%	11,397	5,846	51.3%
30-Jun-08	1,400	155,319	143,339	92.3%	11,948	6,005	50.3%
31-Mar-08	1,402	155,632	142,271	91.4%	12,719	6,097	47.9%
31-Dec-07	1,484	165,634	152,078	91.8%	12,936	6,327	48.9%
30-Sep-07	1,434	160,137	147,005	91.8%	13,149	6,148	46.8%
30-Jun-07	1,437	162,093	148,366	91.5%	12,896	6,226	48.3%
31-Mar-07	1,394	155,683	142,747	91.7%	12,750	6,152	48.3%
31-Dec-06	1,372	154,775	141,181	91.2%	12,895	5,991	46.5%
30-Sep-06	1,293	150,345	136,521	90.8%	12,962	5,927	45.7%
30-Jun-06	1,304	152,747	139,220	91.1%	13,332	6,018	45.1%
31-Mar-06	1,321	149,939	136,653	91.1%	12,955	5,917	45.7%
31-Dec-05	1,257	145,003	132,481	91.4%	12,248	5,831	47.6%
30-Sep-05	1,222	142,716	129,241	90.6%	12,327	5,780	46.9%
30-Jun-05	1,192	139,675	127,820	91.5%	11,989	5,744	47.9%
31-Mar-05	1,201	143,027	129,102	90.3%	12,597	5,741	45.6%
31-Dec-04	1,195	140,004	127,673	91.2%	12,359	5,687	46.0%
30-Sep-04	1,085	130,787	118,903	90.9%	11,670	5,450	46.7%
30-Jun-04	1,043	126,230	114,723	90.9%	12,089	5,410	44.8%
31-Mar-04	1,064	125,865	114,654	91.1%	12,480	5,234	41.9%
TOTAL	n/m	6,694,538	6,179,559	92.3%	505,433	243,446	48.2%

* As scraped from 13F filings, already excluding options

** Post filters and error correction mechanisms

Figure III.3: Summary information on dataset

III.2 Predicted Variable

Our predicted variable is the forward return of a security:

$$\begin{aligned} P_{Q_0} &= \text{price of a security at the end of the current quarter} \\ P_{Q_{0+t}} &= \text{price of security at the end of } t \text{ quarters from now} \end{aligned}$$

$$return = r = \frac{P_{Q_{0+t}} - P_{Q_0}}{P_{Q_0}}$$

We calculate returns for $t = [1, 2, 3, 4]$ quarters forward and denominate these variables Y1, Y2, Y3 and Y4, respectively. For the purposes of some the statistical learning methods employed, we use a discretized version of the returns, e.g. defined by the terciles of the distribution of forward returns:

let F be the cumulative distribution function of all returns r for all securities s in a given timeframe, and let Y_s be the return class of security s; then:

$$\begin{array}{lcl} F(r_s) \in (0, 0.3\overline{3}] & \rightarrow & YC_s = 1 \\ F(r_s) \in (0.3\overline{3}, 0.6\overline{6}] & \rightarrow & YC_s = 2 \\ F(r_s) \in (0.6\overline{6}, 1] & \rightarrow & YC_s = 3 \end{array}$$

Note that this results in balanced classes and, the higher the class number, the better the returns. For illustrative purposes, we show plots of the return distribution (for 1-year/4-quarter returns) in Figure III.4 as well as of the mean returns for each bin (Figure III.5), over the entire analysis period, when binning 4-quarter returns into 3 bins. As expected, the distribution of returns is bell-shaped with positive skew (of course, returns can never be lower than -100%, but can be infinite), and excess kurtosis (financial returns are well-known, and feared, for having fat tails).

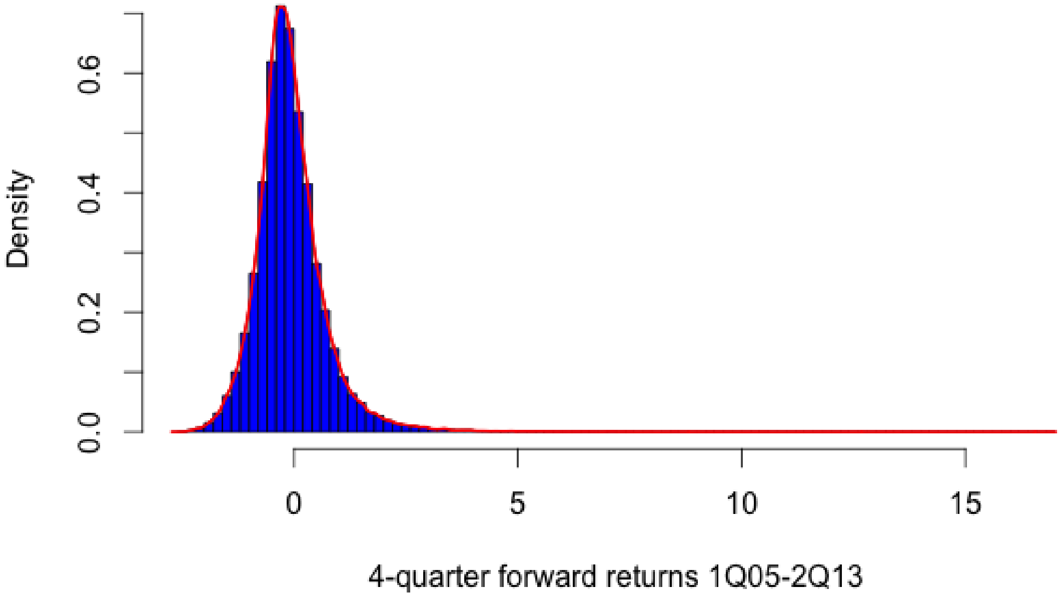


Figure III.4: Histogram of returns 1Q05-2Q13

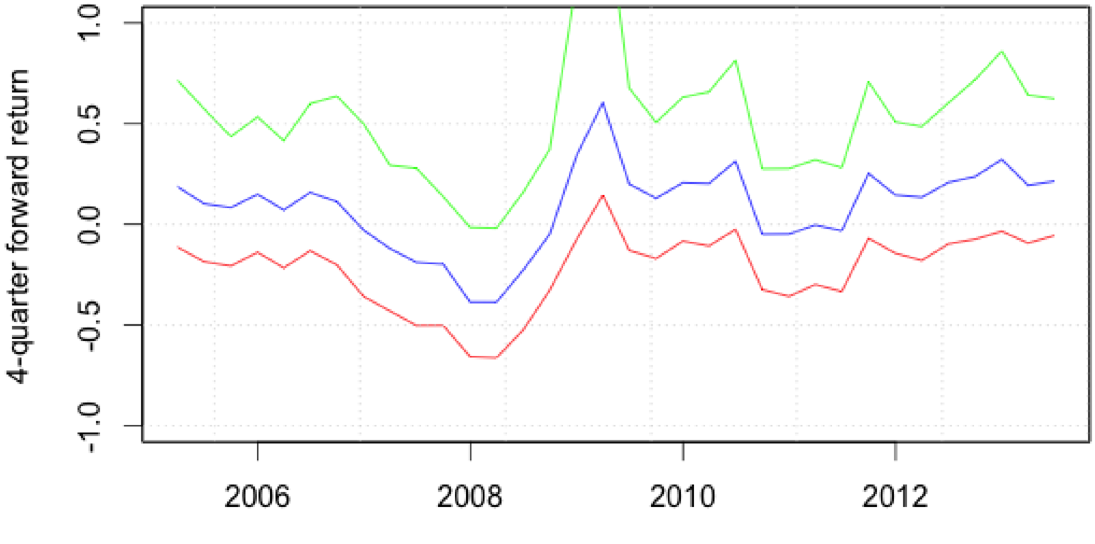


Figure III.5: Mean returns per bin (3 bins)

III.3 Features

As we noted in the Background section, the current study focuses substantially on institutional ownership-related features. In the following section, we present descriptions of the features used to train our prediction models. We classify these ownership-related features into sub-categories as presented in section II.1b:

- institutional ownership magnitude;
- institutional ownership dispersion;
- institutional owners' characteristics; and
- ownership interaction with other variables (principally, security prices).

We provide basic exploratory analysis on each feature in the appendix. In order not to "contaminate" our validation and test sets, we only conducted such exploratory analyses on quarters in our training set timeframe, as defined further below. We present the calculation of each feature as a continuous variable. These are easily transformed into categorical variables, where a classifier type so requires, via assigning the security to quantile bins. Besides the value of the features in a given quarter, we also consider the change in most features over time, specifically comparing the value of a feature in the current quarter to its value 1, 2, and 3 quarters ago. For simplicity, we do not explicitly show neither the quantile nor "first derivative" transformations of the basic features in what follows. Each feature is calculated with regard to a specific security and quarter and should hence be subscripted with such security and quarter – again, for simplicity's sake, we omit such subscripts in what follows.

In terms of practical implementation, the features were calculated automatically in fully-vectorized environments, either in Python (using the Numpy and Pandas libraries) or R.

(a) Institutional ownership magnitude

$X1$ Number of investors holding the security

This is the simplest of all features. Let H be the set of all investors h holding a given security in a given calendar quarter, then:

$X1 = |H|$

$X2$ **Fraction of institutional investors**

A more refined ownership size metric than X1 is to consider the sum of all shares of a security that are held by 13F investors vs. the total number of shares outstanding of the security. Let s_h be the number of shares in security s held by investor $h \in H$; and S be the total number of shares outstanding of security s:

$$X2 = \frac{\sum\limits_{h \in H} s_h}{S}$$

$X3$ **Adjusted number of investors**

Our domain experience suggests that the size of institutional ownership in a security should be impacted by the market capitalization and trading volume of the security. Intuitively, an institutional investor (unless the investment process is substantially automated) simply cannot afford to spend human analysis time on small (relatively to total amount invested) investment positions. However, institutional investors are typically also wary of liquidity risk and often want to be able to exit any investment position within a few days without adversely impacting its price. Looking at our dataset e.g. in 2Q14, the median total amount invested per institutional investor in our sample was approximately US\$600 million and the median number of investment positions per investor was 76, implying an average investment size of approximately of (600/76) almost US\$8 million. In order to exit this type of position without negative impact on the price, the security either has to trade in sufficient size on-exchange (say, e.g. if the investor wanted to exit over the course of four days and could not represent more than 1/3 of trading volume in order not to impact price, then the security would have to trade US\$ 6 million per day) or the investor would at least have to be able to effect some type of block trade(s) off-exchange. The latter would typically only be possible without impacting the price negatively if the stake sold by the investor represented a fairly small fraction of the overall shares issued by the company, say no more than 1% (which in our example would imply that the company needed to have a market capitalization of at least US\$800 million).

Our intuition is confirmed by our data, e.g. in 2Q14, as shown by regressions of the number of investors in a security vs. the market capitalization and/or trading volume of the security. The regression of number of holders vs. trading volume (which included the use of a spline) yields an R^2 of 0.7116 (refer to Figure III.7). The relationship is also visible in the scatter plot in

Figure III.6. We show the regression and scatter plot for the relationship between number of holders and market capitalization ($R^2 = 0.5023$) in the appendix, in section B.3. A multiple linear regression of number of holders as a function of market capitalization and trading volume yields an R^2 of 0.7653 and is also shown in section B.3 of the appendix. Albeit market capitalization and trading volume are correlated (r=0.66) and therefore there is collinearity in the regression, this is not at an excessive level (adjusted generalized variance-inflation factors of 1.44 and 1.13 for market capitalization and volume, respectively, in 2Q14).

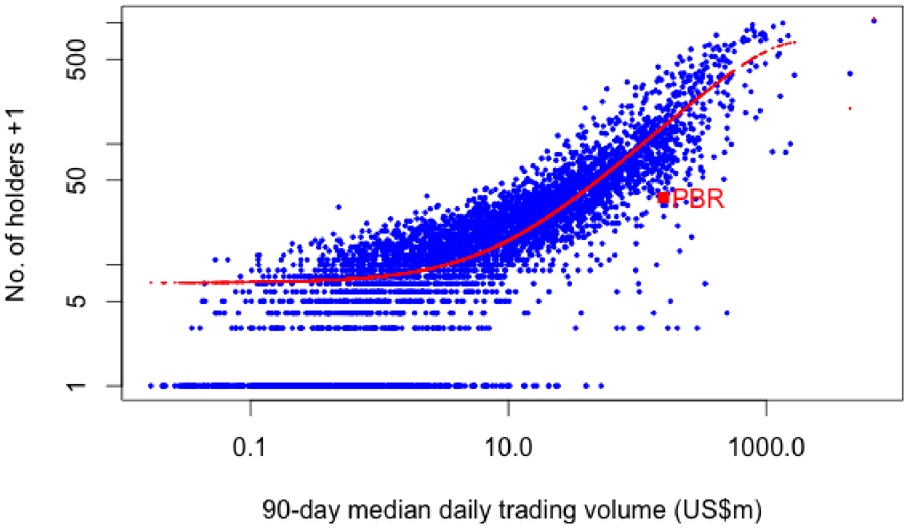
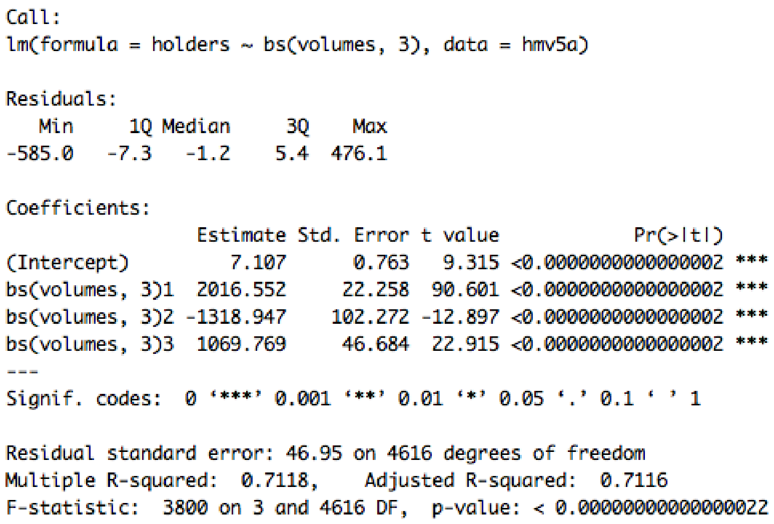


Figure III.6: No. of holders vs. trading volume (2Q2014, log-log)



```
Call:
lm(formula = holders ~ bs(volumes, 3), data = hmv5a)

Residuals:
   Min     1Q Median     3Q    Max
-585.0   -7.3   -1.2    5.4  476.1

Coefficients:
                  Estimate Std. Error t value            Pr(>|t|)
(Intercept)          7.107      0.763   9.315 <0.0000000000000002 ***
bs(volumes, 3)1   2016.552     22.258  90.601 <0.0000000000000002 ***
bs(volumes, 3)2  -1318.947    102.272 -12.897 <0.0000000000000002 ***
bs(volumes, 3)3   1069.769     46.684  22.915 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.95 on 4616 degrees of freedom
Multiple R-squared:  0.7118,    Adjusted R-squared:  0.7116
F-statistic:  3800 on 3 and 4616 DF,  p-value: < 0.00000000000000022
```

Figure III.7: Regression of no. of holders vs. trading volume (2Q2014)

We can hence construct a feature that eliminates the impact of market capitalization and trading volume on the number of holders of the security. In other words, a feature representing the residual. One final issue here to consider is the heteroskedasticity of the residuals. We address this issue by constructing a metric that divides the residual by the corresponding predicted value. Let $\widehat{|H|}$ be the predicted number of holders of a security.

$$X3 = \frac{(|H| - \widehat{|H|})}{\widehat{|H|}}$$

Plot of the unadjusted and adjusted residuals are provided in Figures B.6 and B.7 in the appendix.

Adjusted fraction of institutional investors

Analogously, we could calculate an adjusted metric based on the fraction of institutional investors $X2$. However, as we can see from the scatter plot (Figure III.8) and regression (Figure III.9) of institutional ownership fraction vs. trading volume, the relationship appears to be rather weak. It actually seems to be the case that the institutional ownership fraction is relatively constant, except for securities with very large market capitalizations (holding 1% of Apple would be equivalent to a holding worth more than US$5 billion as of June 30, 2014). The same holds true for the relationship between institutional ownership fraction and market capitalization, as shown in Figures B.8 and B.9 in the appendix.

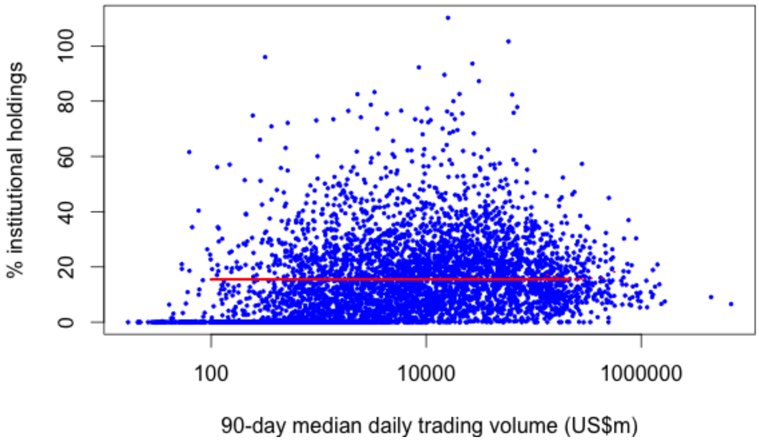
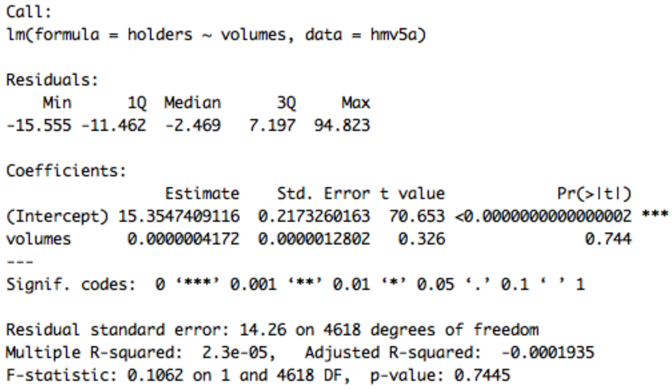


Figure III.8: Institutional holdings% vs. trading volume (2Q2014, log scale on x-axis)



```
Call:
lm(formula = holders ~ volumes, data = hmv5a)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.555 -11.462  -2.469   7.197  94.823 

Coefficients:
                Estimate    Std. Error t value            Pr(>|t|)    
(Intercept) 15.3547409116  0.2173260163  70.653 <0.0000000000000002 ***
volumes      0.0000004172  0.0000012802   0.326               0.744    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.26 on 4618 degrees of freedom
Multiple R-squared:  2.3e-05,	Adjusted R-squared:  -0.0001935 
F-statistic: 0.1062 on 1 and 4618 DF,  p-value: 0.7445
```

Figure III.9: Regression of institutional holdings% vs. trading volume (2Q2014)

(b) Institutional ownership dispersion

$X4$ Entropy 1

We consider a feature representing the level of "consensus" with regard to being overweight or underweight a certain security. For this purpose, we calculate Shannon's entropy with regard to the probabilities of investors' position weight of a security falling into a certain quantile.

We calculate the investor's "position weight" of a security in two ways. First, we consider the Dollar value of the position in the security over the total Dollar value of the investor's entire investment portfolio (in a given quarter) which effectively represents an absolute position weight. Let p and P be the Dollar value of an individual investment position in a security and the Dollar value of the entire portfolio of a specific investor h, respectively.

$$pw1_h = \frac{p_h}{P_h}$$

For every security, we calculate this absolute position weight $pw1$ for every investor h in the security, in a given quarter. We then assign those weights to one hundred equal-sized bins in order to derive probabilities of the absolute weight falling into a certain bin i. Lastly, using these probabilities (denoted $P(pw1_i)$ below) we calculate Shannon's entropy (using the *entropy* package in R):

$$X4 = -\sum_i P(pw1_i) \log_2 P(pw1_i)$$

$X5$ Entropy 2

Second, we calculate a type of "relative" position weight for each security, by deriving the absolute position weight in the same way as above, but dividing it by the average position weight given an investor's portfolio (e.g. an investor with ten investment positions has an average absolute position weight of 10%). We then assign those weights to ten equal-sized bins in order to derive probabilities of this relative weight falling into a certain bin i. The calculation of entropy, of course, remains the same:

$$X5 = -\sum_i P(pw2_i) \log_2 P(pw2_i)$$

It is worth noting that we do not hold any ex ante intuitions or hypotheses with regard to what type of impact entropy should have on forward security returns.

$X6$ **Gini**

We also consider it interesting to calculate a feature representing the dispersion of absolute (Dollar) holding sizes of a security among institutional investors. Consider that e.g. a security may have high institutional ownership due to either all investors being overweight the security in equal measure or to a subset of investors being overweight to an even higher degree. From a domain expertise point of view, the former scenario (all investors overweight) often represents an unattractive situation as there may not be any incremental institutional buyers left to help drive the security price higher. In contrast, the latter scenario (a subset of investors very overweight) may be more attractive as it may indicate that a few savvy investors have managed to identify an interesting investment situation, with scope for further institutional investors to still catch on and buy the security. Let $|H|$ be the number of holders of a security and s_i be the number of shares held by the i-th investor in the security. We assume that the shareholdings in a security are indexed/sorted in non-decreasing order. We calculate the Gini coefficient (using the *ineq* package in R) for each security:

$$X6 = \frac{2\sum_{i=1}^{|H|} i s_i}{|H|\sum_{i=1}^{|H|} s_i} - \frac{|H|+1}{|H|}$$

The higher the Gini coefficient, the more concentrated are the holdings in the security. The coefficient can lie between 0 (perfect equality) and 1 (perfect inequality).

(c) Institutional owners' characteristics

The next group of features is related to characteristics of the institutional investors holding a specific security. The chosen features are, as usual, explained in further detail below. It is, however, worth noting that we considered, but ultimately decided against, introducing a further features based on whether institutional holders in a security include investors that we, based on our domain knowledge, consider to be "smart" investors. In the end, we found it impossible to establish such a list of smart investors without forward-looking bias (i.e. e.g. we simply do not remember with confidence who we would have considered to be a smart investor as of, say, 1Q2004, without this judgment being impacted by our knowledge of events post 1Q2004).

$X7$ Investor performance score

The ALFA and IBLN replicator indices/ETFs presented in section II take into account the past returns of institutional investors holding a security. We therefore adopt this feature in the current study, too. The general question of whether past returns can be a predictor of future returns is a rich and controversial field. We decided not to procure published investment returns of all institutional investors in the study, partly due to time constraints, partly due to the fact that such published returns may suffer e.g. from

- inconsistencies resulting from differing return calculation methodologies (between investors, or even within investors over time);
- selective reporting;
- returns reflecting investments outside of the "13F scope", e.g. foreign stocks, derivatives, and short positions.

Rather, we calculated a return measure based solely on the published 13F holdings. Unfortunately we have no way of knowing the exact purchase and sales prices of the investment positions. We therefore value positions at the prices at the quarter end dates and, where there are changes in position size, we assume that the incremental shares were bought or sold at the volume-weighted average price prevailing during the quarter (this also therefore fails to capture any "trading around" that an investor may have done during the quarter - e.g. an investor may start and end the quarter with 500 shares of a certain security, but have traded in and out of the security multiple times during the quarter).

At the top level, we construct our metric for every security as the weighted average of the investors holding the security. Let PI_h be the performance index of an investor and w be a weighting function (both defined in the following paragraphs). As usual, we omit the subscript for the security. Hence, for an individual security:

$$X7 = \sum_{h \in H} w_h PI_h$$

Note that, in cases where a metric is not available for a fund in a quarter, we substitute in the median metric of the funds holding the same security during the quarter.

For the weighting function w, which determines how much weight a particular investor's performance score carries in the calculation of the metric of a given security, we choose to consider the percentage weight that the security has in an investor's portfolio (in terms of Dollar value) in the context of such weights for all investors holding the security. Let p and P be the Dollar value of an individual investment position in a security and the Dollar value of the entire portfolio of a specific investor, respectively. The weight w_{hi} assigned to a specific investor i, with regard to his holding in a specific security (security subscript omitted as usual) is calculated as:

$$w_{hi} = \frac{\frac{p_{hi}}{P_{hi}}}{\sum\limits_{h \in H} \frac{p_h}{P_h}}$$

For the construction of the performance index PI of an investor, we were guided by the desire for this metric to reflect both the ranking of a specific investor versus all of its peers in a specific quarter as well as the size of its "track record" (i.e. how many quarters of investment performance do we have available for the investor). We consider the "track record" element as important as, without it, we may give undue importance to investors that may have achieved good rankings by pure chance. This is also the reason why we do not simply calculate an average of past available rankings as our metric. Rather, we decided to design a metric that, in a way, is meant to reflect the probability of achieving a particular investor's performance track record:

- first, for every quarter q, we rank all investors by their performance in the quarter and assign the highest rank to the top performer (i.e. if e.g. there are 100 funds, the top performer will be assigned rank 100);
- second, we divide 1 by the rank of each investor in the quarter, in order to derive a probability-type metric for each investor; let pr be this metric (we omit a subscript to denote individual investors);
- third, we multiply these probabilities for every quarter where an investor has performance data (and hence a rank) in order to come up with a type of aggregate probability pr_{cum} of achieving the investor's track record; let Q and q be the set of all quarters for which performance ranks for a fund are available and a specific quarter from within this set, respectively:

$$pr_{cum} = \prod_{q \in Q} pr_q$$

- fourth, in each quarter we assign each investor to a centile (1-100) bin, based on its pr_{cum}, with 100 representing the highest/best score.

We note that, due to the specific way the metric is constructed, while it achieves our objective to reflect both individual quarter performances and track records of investors, it is slightly biased to overweighting the latter. In a future refinement of the study, we will likely separate out length of track record of a security's investors as a separate feature.

We should point out that the automatic calculation of this metric suffered from the shortcoming that a handful of funds changed names during the observation period, resulting in separate track records for funds that should really be given credit for a single, longer-term track record. In most cases, this issue could be fixed by automatic detection (via string similarity). However this has not been undertaken due to the time constraints of the study.

$X8$ **Investor performance score - recent**

The calculation of feature $X8$ is identical to the calculation of $X7$ save that, for the calculation of the performance index PI of an investor, we only consider performance shown during the most recent four quarters rather than the full performance track record (which, in most cases, will have more than four calendar quarters).

$X9$ **Investor portfolio turnover score**

Another quantifiable characteristic of investors is the speed with which they change their investment portfolio composition over time – e.g. a conservative long-term fundamental investors may hold the same set of stocks for years, while a trading-oriented fund may substantially change its portfolio from quarter to quarter (or even minute to minute in the case of some high-frequency funds). We start by deriving a Jaccard similarity measure for each investor, for each two consecutive quarters. Let p and P be the Dollar value of an individual investment position in a security i and the Dollar value of the entire portfolio of a specific investor, respectively. For simplicity, we omit a subscript to denote the investor.

$$pw_i = \frac{p_i}{P}$$

Let $pw_{i,q0}$ and $pw_{i,q1}$ be the Dollar position weights of security i in the consecutive quarters $q0$ and $q1$, respectively. Jaccard similarity, as always, is defined as the intersection of the sets over their union. As we are looking at percentage holdings, the denominator (union) is the maximum of the two percentage holdings. For the intersection, we have to consider the minimum of the two percentage holdings (in $q0$ and $q1$) of each security the investor holds in either $q0$ and/or $q1$ (e.g. if an investor has 3% of his portfolio invested in a specific security in $q0$ but only 1% in $q1$, then the overlap, for the purpose of the similarity calculation, is 1%).

$$Portfolio_{q0} \cap Portfolio_{q1} = \sum_i \min\left(pw_{i,q0}, pw_{i,q1}\right)$$

$$Jaccard(q0, q1) = \frac{Portfolio_{q0} \cap Portfolio_{q1}}{Portfolio_{q0} \cup Portfolio_{q1}}$$

The Jaccard metric has a minimum of 0 for totally different portfolios and a maximum of 1 for identical portfolios. We transform the Jaccard metric into a type of turnover metric in order to make it more intuitively understandable from a practical financial markets point of view. The turnover TO_h of a specific investors h in a specific quarter (subscript omitted) hence becomes:

$$TO_h = 4(1 - Jaccard)$$

Note that our calculated turnover measure can never exceed four, resulting from the fact that we look at quarterly data (in reality, many of the funds in our universe may turn over their portfolio more frequently, but we have no systematic way of knowing this).

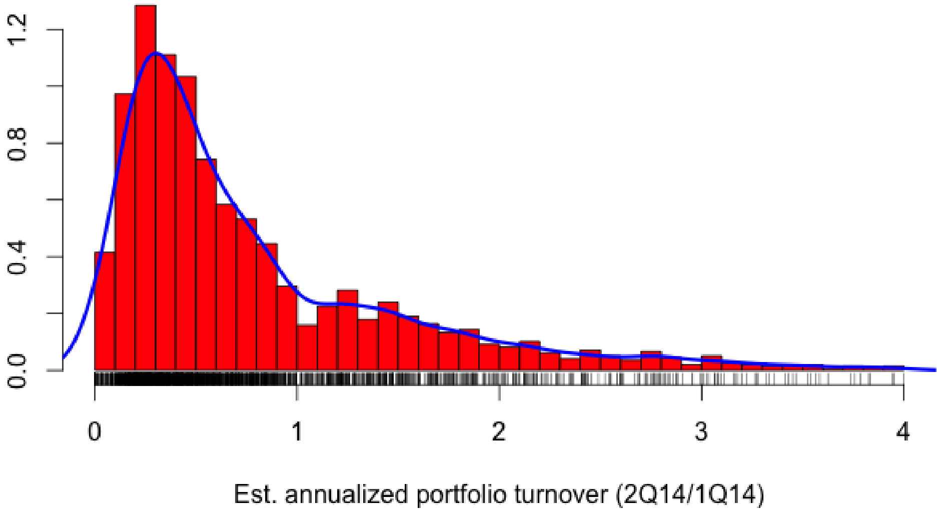


Figure III.10: Histogram of est. annualized portfolio turnover (2Q14/1Q14)

We now calculate a weighted turnover score for each security given the individual turnover metrics of the security's holders in a given quarter. We again use the same weighting function as we defined above in the context of investor performance. Let TO_h be an investor's turnover:

$$X9 = \sum_{h \in H} w_h TO_h$$

$X10$ Investor diversity score 1

We now introduce metrics that are meant to measure the breadth of the universe of securities that an investor typically holds. This is meant to help differentiate between "specialist" investors (say, a mid-cap technology fund) and investors that may cover the entire market. It is worth pointing out that we hold no ex ante hypotheses with regard to the significance of such diversity scores as potential predictors for future security performance. We do, however, consider it a measurable characteristic of a security's investor base and hence valid to include as a feature.

We calculate these "diversity" measures in two ways. In the first way, we consider the cumulative number of unique securities that a fund has held over the cumulative number of unique securities existent in the market (in both cases up to a determined quarter in time q).

With this metric in hand, we calculate a weighted diversity score for each security given the individual diversity scores of the security's holder in a given quarter. We use the same weighting function that we introduced for the investor performance score above. Let $D1_h$ be an investor's diversity score:

$$X10 = \sum_{h \in H} w_h D1_h$$

$X11$ Investor diversity score 2

The second way to calculate investors' diversity scores takes into account the fact that investors may have constraints on the maximum number of investment positions. E.g. a fundamental, stock-picking, hedge fund may have a limit of, say, thirty investment positions (given constraints on human analysis time) and, under diversity measure 1 above, may therefore struggle to ever catch up with the diversity of a large index fund that virtually invests in the entire market. In order not to penalize the hedge fund vis-a-vis the index fund, we calculate the second diversity metric as the cumulative number of unique securities that a fund has held over the cumulative number of individual investment positions that the fund has had (in both cases up to a determined quarter in time q) – e.g. if the hedge fund was constrained to ten positions per quarter, then, after four quarters, the denominator of our second diversity metric would be forty.

As above, we calculate a weighted diversity score for each security given the individual diversity scores of the security's holder in a given quarter. We again use the same weighting function as above. Let $D2_h$ be an investor's diversity score:

$$X11 = \sum_{h \in H} w_h D2_h$$

(d) Ownership interaction with other variables

$X12$ Holding - price interaction

There are nine possible scenarios for the co-movement of the price and the institutional holdings of a given security during a given quarter. The table below show what proportion of total securities fell into each category in 2Q14.

	Share price change		
Holdings change	Down	Unchanged	Up
Down	17.1	0.1	31.4
Unchanged	0.2	0.0	0.3
Up	17.1	0.0	33.7

We can therefore easily construct a categorical feature with the nine possible values above. We could, of course, also attempt to construct a continuous feature that somehow would take into account the magnitudes of the price and ownership changes, but, for the time being, omit doing so for simplicity's sake. The construction of the feature is straightforward, given that the two required inputs, quarterly price change and institutional ownership percentage change, underlie our predicted variable and feature $X2$, respectively.

(e) Non-ownership-related features

$X13$ Security type

This is a binary variable indicating whether the security is a ”regular” common/preferred stock as opposed to e.g. a fund, convertible security, etc. The below table shows the breakdown of all unique securities in the study by their type, using FactSet data.

Security type	Number	%
Common stock	11,198	72.9
Mutual fund	1,558	10.1
Exchange-traded fund	1,224	8.0
Preferred stock	540	3.5
Warrant or right	315	2.1
(Unlabeled by FactSet)	219	1.4
Unit	181	1.2
Convertible preferrred	115	0.8
Money market mutual fund	5	0.0
Convertible bond	4	0.0
Corporate or Government security	1	0.0
TOTAL	15,360	100.0

As notes above, we only included common and preferred shares in the present stuy.

$X14$ **Security market capitalization**

We decided to include security market capitalization, a proxy for size, as a feature as various of the studies on securities returns reviewed above have identified size as an important factor. As noted above, the present study only includes securities that have in excess of US$ 100 million market capitalization in the "current" quarter. However, no such limit value has been used with regard to calculating forward performance, in order to avoid survivorship bias. E.g., a security that has $\geq$ US$ 100 million market capitalization in, say, 1Q2008, but in the following quarters falls to zero (Lehman Brothers is a specific example) would be included in our dataset in the 1Q2008 row, with its forward returns (the predicted value in the present study) showing highly negative values.

III.4 Dataset partioning

Following established practice, we partition our dataset into a training, validation and test set. Bearing in mind that our dataset represents time series data, we avoid any look-ahead bias (i.e. using data unknown at the time of prediction) by executing this three-way partition chronologically. As the exhibit below illustrates, our training set, validation set and test set comprise the quarters 1Q2005 through 2Q2007, 3Q2008 through 2Q2010 and 3Q11 through 2Q13, respectively (resulting in an approximate 40%/30%/30% split). It is important to remember that some of our features look up to four quarters backward in time (e.g. all the 'first derivative' features representing the change of a base feature vs. its value one, two, and three quarters ago) while our predicted variable is up to four quarters in the future. This also explains what appear to be the chronological gaps between our sets as listed above, e.g. the training set begins in 1Q2005 and ends in 2Q2007, while the cross-validation set only commences in 3Q2008: imagine that today is the last day of 2Q2008 – then we can train a model effectively only up to the 2Q2007, as we need to include our predicted variable, 1/2/3/4-quarter forward returns, and we obviously only know those up to today (i.e. 2Q2008); furthermore, due to the backward-looking features, the training set can only commence in 1Q2005, even though our dataset as such starts in 1Q2004. In the below exhibit, periods that contain purely backward-looking data are shaded in light blue, whereas periods that contain only forward-looking data are shaded in light orange. The "core" sets, which effectively contain quarters for which we have full data (incl. all backward and forward-looking data) available, are shaded in dark blue.

Note that the size of the forward-looking portion (light orange) obviously depends on our choice of predicted variable – i.e. whether we choose to predict returns that are 1, 2, 3 or 4 quarters into the future. Figure III.11 depicts the dataset portioning as it would need to look if we predict returns four quarters, i.e. the maximum period in this study, forward. If instead we chose the 1-quarter forward return as our predicted variable, then the required gap between sets (as depicted by the forward-looking part in light orange) would only comprise one, rather than four, quarters. This introduces a complication with regard to comparing the performance of classifiers on the different versions of the predicted variable: we can either compare the classifiers on always exactly the same training and validation periods (and in this case have to use the "worst case" four-quarter gap), or we can adjust the periods depending on what is "allowable" given the predicted variable, but then the training/validation periods will differ depending on the predicted variable chosen. For example, if our predicted variable is the 4-quarter forward return, then the partitioning of data into training and validation set is exactly as depicted in Figure III.11. If, instead, we used the 1-quarter forward return as our predicted variable, then we could either maintain the same training set but start and end the validation set three periods earlier, or we could start and end the training set three periods later, while maintaining the same validation set (we could also simply extend the training or validation set by three quarters, but we will ignore these options). In the below, when comparing classifier performance for the different options of predicted variable, we will simply use the same training and validation periods, independent of predicted variable choice (which means that we must use the maximum, four-quarter gap between training and validation set).

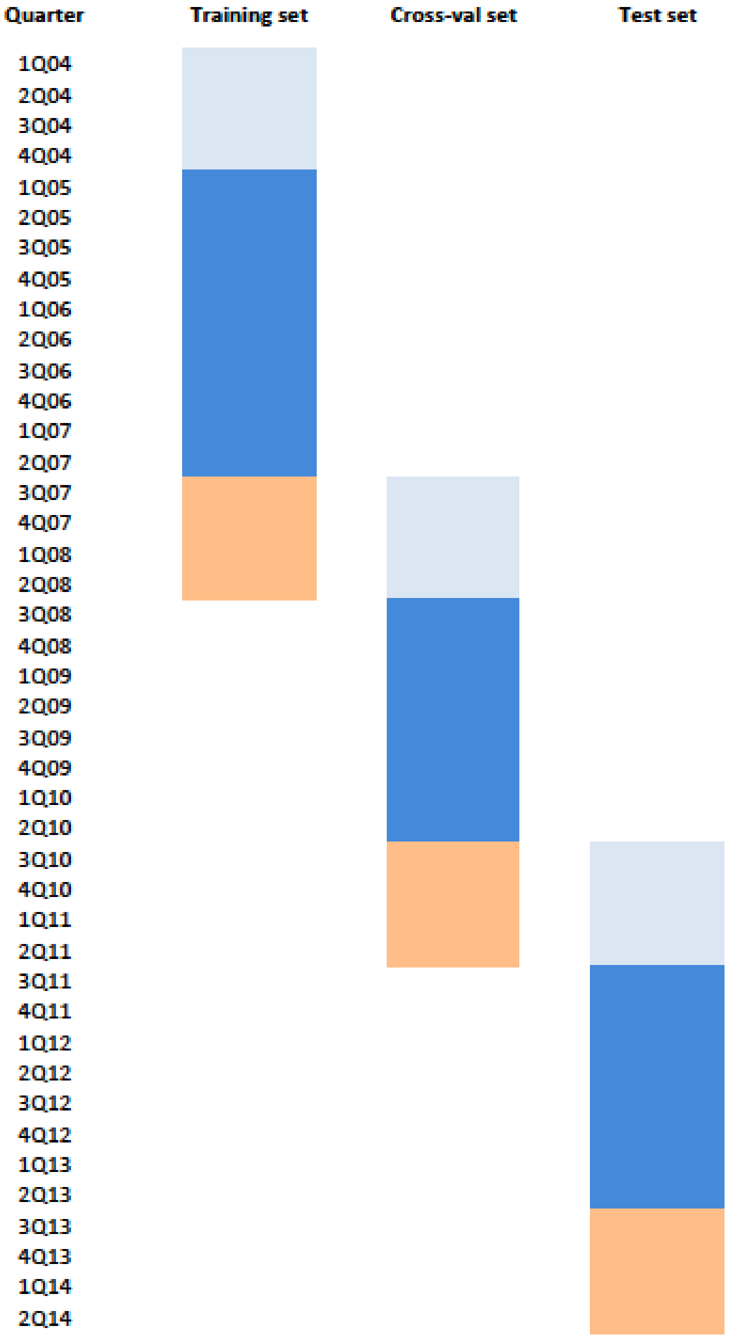


Figure III.11: Dataset split into training/cross-validation/test set

III.5 Learning techniques and their implementation

(a) Classifiers

We employed the following learning techniques, representing a broad spectrum of the available types of classifiers:

- Naive Bayes
- Logistic regression
- Support vector machine (SVM)
- k-nearest neighbor (KNN)
- Decision tree
- Random forest

Below, we give a succinct explanation of these methods. For an in-depth review, we suggest e.g. [14] and [11].

Naive Bayes

The Naive Bayes classifier simply assigns an observation with feature vector X_i to the class for which the probability of belonging to that class, given X_i, is maximized. Critical assumptions are that the features are independent of each other (which is not realistic) and, in the case of continuous input features, normally distributed. A key advantage of this classifier is its speed.

Logistic regression

Logistic regression fits a linear model to the log odds of an observation with feature vector X:

$$log\left(\frac{Pr(X)}{1-Pr(X)}\right) = w'X$$

Membership to a binary class depends on the sign of the log odds, i.e. zero being the decision boundary. The weights w are estimated by maximizing the log-likelihood.

Support vector machine (SVM)

The idea underlying SVMs is the use of a hyperplane to separate observations of different classes. The chosen plane is the one providing the maximum margin between training observations of differing classes, which is effectively defined by its support vectors. In order to accommodate non-linear decision boundaries, kernels can be used, such as a polynomial or radial kernel. In the end, the classification of an observation x is given by sign of a function of the form

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)$$

where S is the set of support vectors and K is a kernel function. SVMs can be applied to classification problems with more than two classes by e.g. using a one-vs.-one approach, whereby, for k classes, $\binom{k}{2}$ classifiers are constructed and an observation is assigned to the class into which it was most frequently voted. Note that SVMs and logistic regression often give very similar results as one can show that their loss functions are quite similar.

k-nearest neighbor (KNN)

The KNN classifier chooses the class j that maximizes the probability of an observation i belonging to that class given the classes to which i's K nearest neighbors belong:

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

The neighborhood of i is defined via its position in the n-dimensional feature space. The closest neighbors are chosen based on a distance metric, e.g. Euclidian distance. A key advantage of KNN is that its decision boundary is highly flexible. A key disadvantage is that its output often does not yield any intuitive understanding as to the role of the individual features.

Decision tree

Decision trees, or, more precisely for the purposes of this study, *classification trees*, split up the feature space into regions and assign an observation to the class that is most common in its region. Trees are typically greedily grown by always splitting on the feature with most "impurity", as indicated by an impurity measure such as entropy or the Gini index. In order to avoid over-fitting, trees are usually subjected to pruning. Key advantages of trees include speed and interpretability, as well as the fact that they do not require a linear decision boundary.

Random forest

Random forests consist of a pre-defined number of decision trees. The trees are forced to be "de-correlated" by, for each split, choosing the splitting feature from a subset of all available features (often approximately the square root of the number of all features). This effectively avoids that trees may always be dominated by the same strong features. The output class is defined by majority vote.

(b) Implementation

We implement the learning algorithms above using the Weka (Waikato Environment for Knowledge Analysis) workbench as well as R (using e.g. the *e1071* library for SVMs).

Most exploratory data analysis, including regressions, has been performed in R using, inter alia, the *car*, *glmnet* and *caret* packages.

III.6 Evaluation

(a) Summary evaluation metrics

For our main classification-based learning techniques, we evaluate the success of our learned models by considering the following metrics:

Overall success rate

The overall prediction success rate across all classes is given by the simple equation below, where TP, TN, FP, FN equal number of true positives, true negatives, false positives, false negatives, respectively.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Kappa statistic

The Kappa statistic puts the overall success rate into context by comparing it to the success rate that a random predictor would obtain on the given classes. This is achieved by subtracting the number of expected correct guesses of the random predictor from the model's correct predictions and dividing the result by the maximum possible number of correct predictions in excess of expected random predictor correct predictions. The resulting Kappa statistic ranges from a minimum of 0 (model no better than the random predictor) to 1 (model is a perfect predictor).

Other metrics

Some of the learning model ouput presented in the following sections also includes other metrics, such as precision, recall, F-measure and ROC area. We suggest e.g. [17] for a review of these and other evaluation metrics.

(b) Practical evaluation

Given the applicability of the present study to investing, we can also evaluate our models with regard to their success in an investing environment.

Long-short strategy

Each quarter, we could use our learned classification model to construct investment portfolios, whereby we buy the securities (”go long”) that the model classified into the highest expected return class and sell those (”go short”) classified into the lowest expected return class, holding the positions for the investment period that the model was trained on (i.e. either 1, 2, 3, or 4 quarters). We could then conduct t-tests to evaluate whether the differences in returns between the top and lowest class are significantly significant. Note that for a full evaluation of commercial feasibility of such a strategy we would need to include all transactions costs (incl. brokerage commission and stock lending fees). The present study does not include a test of such a strategy.

Benchmarking vs. existing replicator products

Each quarter, we could use our learned classification model to construct a long-only investment portfolio from the model's top-rated securities. We could then compare the performance of our portfolio against the performance of the replicator products described above. The present study does not include such a test.

IV Findings

In the following sections, we present and comment the performance of our learning models, first on the validation set, then on the test set.

IV.1 Training and validation

The following tables show the performance of our various learning models on the validation set, as measured by our chosen metrics (percentage of correct predictions and Kappa statistic), for classification into either five (first table) or three (second table) output classes, as outlined in section III.2. We provide our comments on the results immediately below the tables.

(a) 5 output classes

Note that we did not run a Random Forest classifier for the 5-output classes task due to the computational intensity.

CLASSIFICATION INTO 5 OUTPUT CLASSES
Mar-05 - Jun-07 training / Sep-08 - Jun-10 validation

	% correct predictions				**Kappa statistic**			
	Forecast timeframe				Forecast timeframe			
Classifier	1q	2q	3q	4q	1q	2q	3q	4q
Naïve Bayes	22.7	22.6	23.0	23.2	0.0341	0.0328	0.0377	0.0395
Logistic regression	22.7	22.9	23.1	23.0	0.0342	0.0360	0.0390	0.0369
SVM (linear)	23.0	22.7	23.2	23.2	0.0375	0.0336	0.0396	0.0396
KNN (10)	21.9	21.2	21.2	20.8	0.0186	0.0144	0.0154	0.0100
Tree (J48)	21.0	20.6	20.8	20.2	0.0131	0.0074	0.0104	0.0019

Table IV.1: Classification results - 5 output classes

(b) 3 output classes

CLASSIFICATION INTO 3 OUTPUT CLASSES
Mar-05 - Jun-07 training / Sep-08 - Jun-10 validation

	% correct predictions				**Kappa statistic**			
	Forecast timeframe				Forecast timeframe			
Classifier	1q	2q	3q	4q	1q	2q	3q	4q
Naïve Bayes	36.6	35.9	36.4	36.9	0.0496	0.0381	0.0463	0.0541
Logistic regression	37.1	36.7	36.7	36.9	0.0559	0.0498	0.0505	0.0539
SVM (linear)	37.4	37.1	36.9	37.2	0.0604	0.0571	0.0530	0.0579
KNN (10)	34.9	34.6	34.4	34.7	0.0236	0.0185	0.0157	0.0200
Tree (J48)	33.5	33.6	33.6	32.6	0.0032	0.0035	0.0038	0.0000
Random forest (100)	36.6	35.1	34.9	35.0	0.0483	0.0262	0.0232	0.0256

Table IV.2: Classification results - 3 output classes

(c) Conclusions from training and validation

We draw the following conclusions from the results, as shown in tables IV.1 and IV.2 above, of applying our various trained models to the specified validation set:

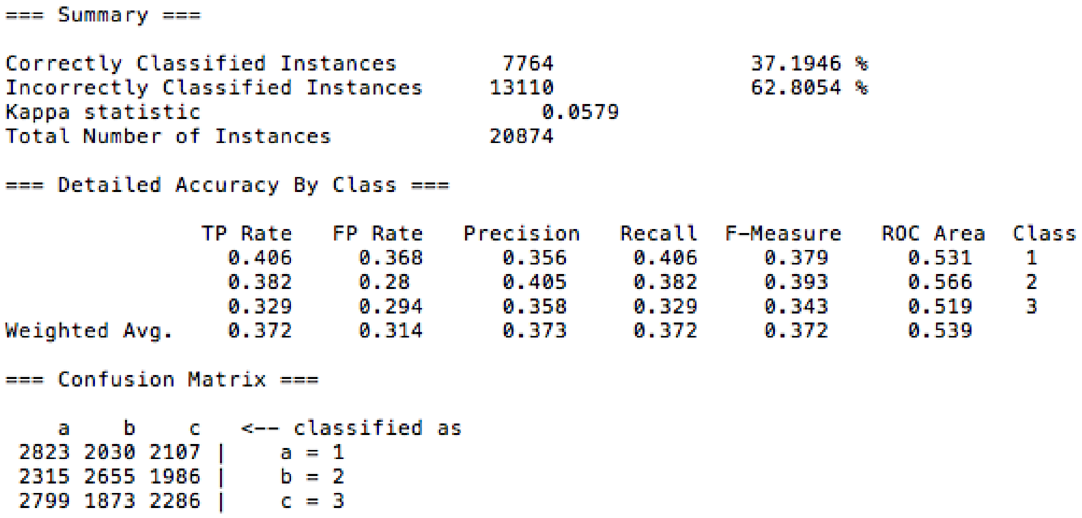
- logistic regression and support vector machine (SVM) showed the best performance among the tested classifiers (as noted above, this similarity is not surprising);
- classification into three rather than five bins yields slightly better results, as evidenced by the Kappa statistic; and
- the best performance is achieved when using either 1-quarter forward or 4-quarter forward returns as the predicted variable.

We also attempted a number of variations of the underlying data and employed classifiers, none of which resulted in any meaningful performance improvement, e.g.:

- using discretized versions of features that are continuous in their ”raw” form; and
- running a boosting algorithm (Ada Boost – see e.g. [15] for a review).

Given these results, for the purposes of the remainder of this study, we will focus entirely on classification of four-quarter forward returns into three bins, employing a support vector machine as the classifier algorithm. Why did we choose to not also look at the one-quarter forward return predictions? This is substantially based on a practical reason: SEC regulations allow 13F filings to be made up to 45 days after quarter end. Adding on a further days for processing/analysis and trade execution, this means that more than half of quarter is over before we can even put any investment positions into place. Note that, in a real world application, even for the case of using four-quarter forward returns, we would need to analyze how much of the performance of the securities occurs during these periods between quarter ends and actual filing submissions.

Figure IV.1 presents the results of the SVM for classifying four-quarter forward returns into three bins for the validation period (Sep-08 – Jun-10).



```
=== Summary ===

Correctly Classified Instances        7764               37.1946 %
Incorrectly Classified Instances     13110               62.8054 %
Kappa statistic                          0.0579
Total Number of Instances            20874

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.406     0.368      0.356     0.406     0.379      0.531    1
                 0.382     0.28       0.405     0.382     0.393      0.566    2
                 0.329     0.294      0.358     0.329     0.343      0.519    3
Weighted Avg.    0.372     0.314      0.373     0.372     0.372      0.539

=== Confusion Matrix ===

    a    b    c   <-- classified as
 2823 2030 2107 |    a = 1
 2315 2655 1986 |    b = 2
 2799 1873 2286 |    c = 3
```

Figure IV.1: Results: SVM classifier (4q-fwd returns, 3 bins) on validation period (Sep-08 – Jun-10)

(d) Optimization of training and testing periods

Given the time series nature of our data and the undeniable fact that our features capture but a minute fraction of all possible features that may explain our predicted variable we suspect that any model we build will probably grow ”stale” over time, i.e. its forecasting performance should deteriorate the longer we keep using it without updating the model. A related question is how far back the training period should reach – the tradeoff here being between quantity and ”freshness” of training data. In other words, we would like to know what the optimal training and testing periods should be if we want to optimize the performance of our prediction model. In order to obtain actual data to guide us in answering these questions, we programmed a routine in R that

- varies the training period start from 1Q05 up to 2Q07, while always training up to 2Q07, resulting in a total of ten training sets, with lengths ranging from one quarter (just 2Q07) up to ten quarters (1Q05 - 2Q07);
- varies the test period end from 3Q08 up to 2Q10, resulting in a total of eight one-quarter test sets;
- runs the support vector machine classifier on all of these eighty combinations of training and test sets; and
- outputs the accuracy for every variation tested.

Figure IV.2 shows the result graphically. The highest accuracy is obtained by using the biggest training set possible (starting from 1Q05), but only using this model on the nearest forecast quarter (3Q08), after which the performance already deteriorates.

(e) Optimization of SVM

We experimented with a number of settings for the SVM (using the e0171 package in R), e.g. varying the type of kernel. The highest accuracy was achieved when using a radial kernel.

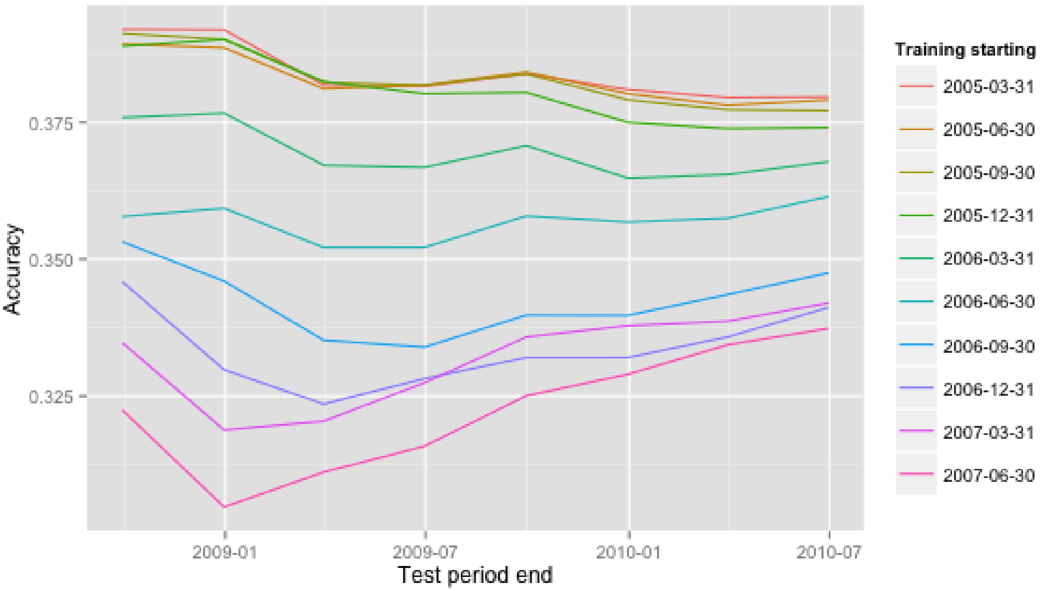


Figure IV.2: Training & test period optimization - 60pc training / 40pc testing

(f) Importance of features in prediction

The following tables (IV.3, IV.4, IV.5) highlight the importance of our features in the classification task, using weights assigned by the SVM (when classifying the validation set, i.e. Sep-08 - Jun-10, into three classes) as a proxy. We show the top ten features for each of the three class pair distinctions.

SVM attribute weights

Class 1 vs. Class 2

Feature		Weight
X2	Fraction of institutional investors	-4.0
X5	Entropy 2	2.6
X9	Investor portfolio turnover score	-2.4
X11	Investor diversity score 2	-2.3
X10	Investor diversity score 1	-2.2
X3	Adjusted number of investors	1.7
X13	Security type	1.5
X2	Fraction of inst. investors - chg. over 3q	1.4
X11	Investor diversity score 2 - change	-1.3
X9	Investor portfolio turnover score - chg. over 3q	1.3

Table IV.3: SVM attribute weights - class 1 vs. class 2

SVM attribute weights

Class 2 vs. Class 3

Feature		Weight
X11	Investor diversity score 2	4.1
X5	Entropy 2	-2.5
X2	Fraction of institutional investors	2.2
X9	Investor portfolio turnover score	2.1
X3	Adjusted number of investors	-2.0
X13	Security type	-1.9
X7	Investor performance score	1.7
X11	Investor diversity score 2 - chg. over 1q	-1.6
X11	Investor diversity score 2 - chg. over 3q	-1.3
X5	Entropy 2 - chg. over 3q	-1.3

Table IV.4: SVM attribute weights - class 2 vs. class 3

SVM attribute weights

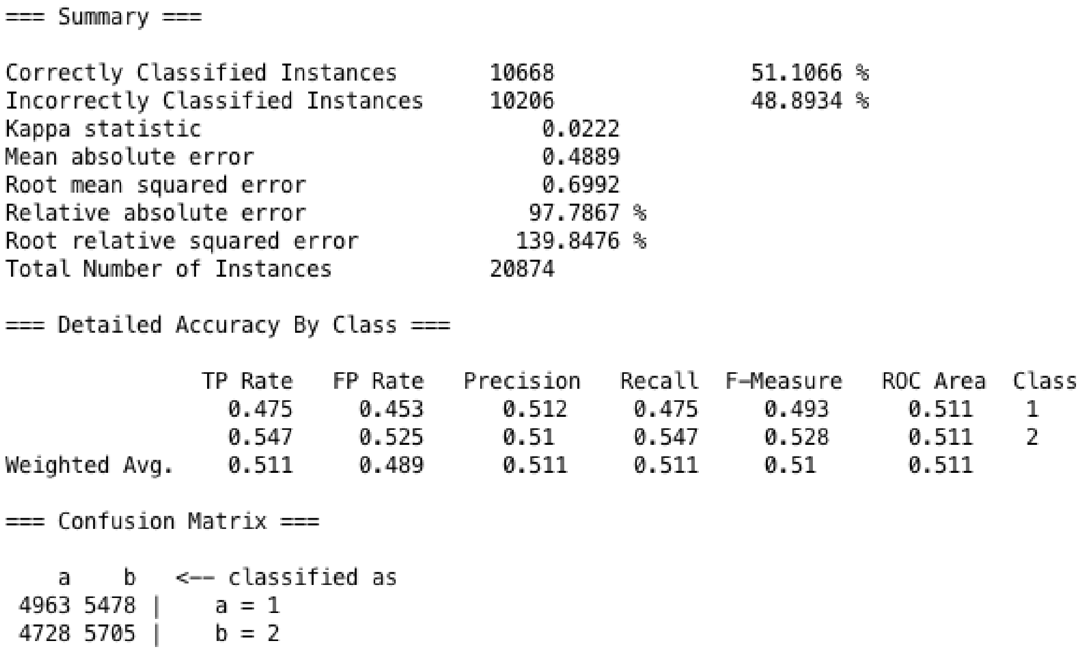
Class 1 vs. Class 3

Feature		Weight
X3	Adjusted number of investors - chg. over 2q	3.1
X10	Investor diversity score 1	-2.6
X2	Fraction of institutional investors	-2.3
X1	Number of investors - chg. over 2q	-2.3
X7	Investor performance score	2.2
X11	Investor diversity score 2	1.7
X8	Investor performance score - recent	-1.7
X3	Adjusted number of investors - chg. over 3q	1.4
X11	Entropy 2 - chg. over 2q	-1.2
X5	Gini - chg. over 3q	-1.1

Table IV.5: SVM attribute weights - class 1 vs. class 3

(g) Validation with two outcome classes

We note that our SVM classifier, when tasked with classification into three outcome classes, performed best with regard to isolating the middle outcome class (class 2). As this may indicate that the classifier may simply identify volatility (securities that moved significantly either up or down, as those in classes 1 and 3), we also ran the classifier on the same validation set, but with only two outcome classes. We present the results below in Figure IV.3. The classifier maintains performance that is statistically significantly different from pure chance.

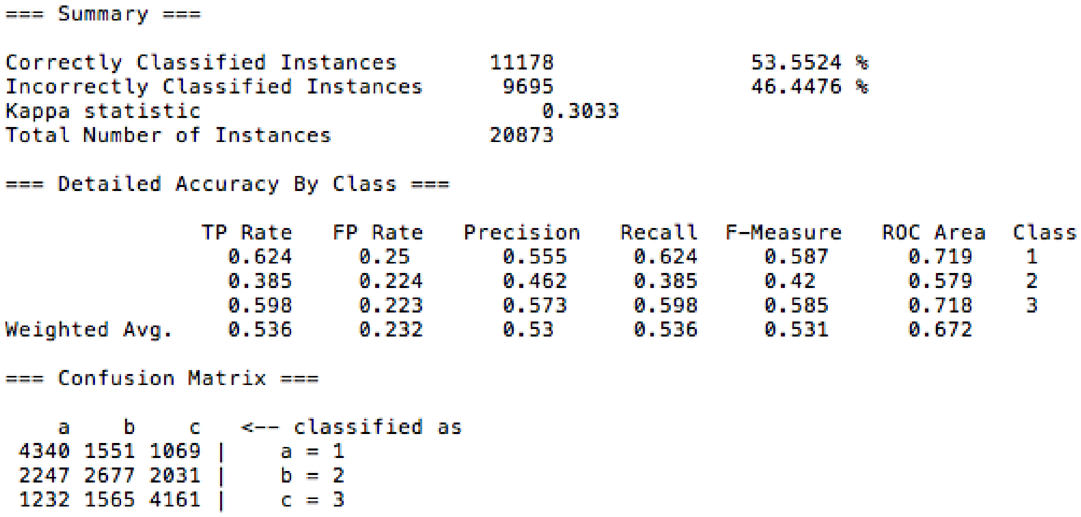


```
=== Summary ===

Correctly Classified Instances       10668               51.1066 %
Incorrectly Classified Instances     10206               48.8934 %
Kappa statistic                          0.0222
Mean absolute error                      0.4889
Root mean squared error                  0.6992
Relative absolute error                 97.7867 %
Root relative squared error            139.8476 %
Total Number of Instances            20874

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.475     0.453      0.512     0.475     0.493      0.511    1
                 0.547     0.525      0.51      0.547     0.528      0.511    2
Weighted Avg.    0.511     0.489      0.511     0.511     0.51       0.511

=== Confusion Matrix ===

    a    b   <-- classified as
 4963 5478 |    a = 1
 4728 5705 |    b = 2
```

Figure IV.3: Results: SVM classifier (forward 4q returns, 2 bins) on validation set (Sep-08 - Jun-10)

(h) Backwards test - ”explaining the past”

While the principal objective of the present study is to predict future returns, we were also curious to see whether our selection of features may at least be useful to explain the past. We therefore also ran our SVM classifier to classify the realized, past four-quarter return (rather than the forward four-quarter return) into three bins, using the same training (Mar-05 - Jun-07) and validation (Sep-08 - Jun-10) sets. The results are shown in Figure IV.4. Clearly, the performance is much improved relative to the performance shown when predicting the future.



```
=== Summary ===

Correctly Classified Instances       11178               53.5524 %
Incorrectly Classified Instances      9695               46.4476 %
Kappa statistic                          0.3033
Total Number of Instances            20873

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.624     0.25      0.555     0.624     0.587      0.719    1
                 0.385     0.224     0.462     0.385     0.42       0.579    2
                 0.598     0.223     0.573     0.598     0.585      0.718    3
Weighted Avg.    0.536     0.232     0.53      0.536     0.531      0.672

=== Confusion Matrix ===

    a    b    c   <-- classified as
 4340 1551 1069 |    a = 1
 2247 2677 2031 |    b = 2
 1232 1565 4161 |    c = 3
```

Figure IV.4: Results: SVM classifier (realized 4q returns, 3 bins) on validation set (Sep-08 - Jun-10)

IV.2 Testing

We now proceed to apply the chosen classifier, SVM with a radial kernel, to classify four-quarter forward returns into 3 bins for our test set, which comprises the quarters of 3Q10 through 2Q13. In accordance with the results from the previous section, we retrain the model every quarter in order to use the maximum training set available. For example, in order to predict the four-quarter forward returns in 3Q10, we train our model using the periods of 1Q05 through 2Q09.

(a) Test results

We present the overall results for the entire test period, as well as, further below, the results for every quarter in the test period. As is evident from the tables, the classifier maintained the level of performance that it had achieved on the validation set. Note that the "return" shown in the rightmost column of the quarterly tables is the mean four-quarter forward return for the specified class, as of the test quarter.

Results for entire test period

Training	Test			**Kappa**	**Accuracy**
1Q05 - 3Q09	3Q10 - 2Q13			**0.067**	**0.378**
True / Predicted ->		Class 1	Class 2	Class 3	Total
Class 1		4491	3455	3452	11398
Class 2		3362	5039	2987	11388
Class 3		4099	3908	3385	11392

Results by individual quarter in test period

Training	Test			**Kappa**	**Accuracy**
1Q05 - 2Q09	3Q10			**0.048**	**0.365**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	369	213	340	922	-32.3%
Class 2	294	393	234	921	-4.9%
Class 3	325	348	248	921	27.6%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 3Q09	4Q10			**0.047**	**0.364**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	443	223	276	932	-35.7%
Class 2	297	394	241	932	-4.8%
Class 3	357	383	192	932	27.7%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 4Q09	1Q11			**0.073**	**0.382**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	433	246	275	954	-29.8%
Class 2	307	411	235	953	-0.3%
Class 3	301	403	249	953	32.1%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 1Q10	2Q11			**0.071**	**0.381**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	395	238	306	939	-33.3%
Class 2	270	399	269	938	-3.1%
Class 3	283	378	278	939	28.2%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 2Q10	3Q11			**0.110**	**0.407**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	346	283	322	951	-6.9%
Class 2	249	435	266	950	25.3%
Class 3	295	277	379	951	70.8%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 3Q10	4Q11			**0.051**	**0.367**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	336	306	309	951	-14.3%
Class 2	283	401	266	950	14.7%
Class 3	357	284	310	951	50.9%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 4Q10	1Q12			**0.070**	**0.380**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	355	282	307	944	-17.9%
Class 2	263	417	263	943	13.5%
Class 3	323	317	303	943	48.6%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 1Q11	2Q12			**0.060**	**0.373**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	340	299	287	926	-9.7%
Class 2	272	411	242	925	20.8%
Class 3	339	302	285	926	60.3%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 2Q11	3Q12			**0.085**	**0.392**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	358	330	261	949	-7.3%
Class 2	264	447	237	948	23.8%
Class 3	349	294	305	948	71.8%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 3Q11	4Q12			**0.059**	**0.373**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	327	384	244	955	-3.4%
Class 2	276	434	244	954	32.2%
Class 3	343	305	306	954	86.1%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 4Q11	1Q13			**0.051**	**0.368**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	373	362	249	984	-9.3%
Class 2	278	441	265	984	19.4%
Class 3	398	315	271	984	64.3%

Training	Test			**Kappa**	**Accuracy**
1Q05 - 1Q12	2Q13			**0.076**	**0.384**
True / Predicted ->	Class 1	Class 2	Class 3	Total	Return
Class 1	426	289	276	991	-5.5%
Class 2	309	456	225	990	21.4%
Class 3	429	302	259	990	62.4%

(b) Significance of test results

We evaluate that the accuracy our prediction model obtained is significantly different from that one would expect by chance (1/3). Over the entire test period, our model made 12,915 successful predictions out of a total of 34,178. Given that we have three possible (and balanced) outcome classes, we can think of our model as rolling a three-sided dice, with a prediction success being equivalent to the dice falling onto a specific side. The null hypothesis is that our dice is fair and the probability of falling onto this specific side is 1/3. We calculate the probability of obtaining our outcome, using a binomial distribution (of course, given the large n, we could have also used a normal approximation):

$$Pr(X \geq 12915) = \sum_{k=12915}^{34178} \binom{34178}{k} (\frac{1}{3})^k (\frac{2}{3})^{34178-k}$$

The calculated probability is basically zero.

V
Conclusions and Discussion

Our task was to investigate whether we can use ownership-related features to predict forward stock returns. Given the results we obtained, as outlined above, we declare a limited victory on a purely theoretical front: the accuracy obtained for classification into three classes over the entire test period (0.378) is significantly different from the accuracy one would expect by chance (1/3). We are also pleased that a number of features that we especially developed (or at least adapted) for the present study appeared as having high weights in our SVM classifier, e.g. turnover scores, diversity scores and entropy.

However, we cannot yet declare victory from a practical perspective. Our classifier was best at identifying the middle return class, while confusing the extreme return classes relatively more often (see e.g. Figure IV.1). This makes an intuitive "go short the low return, go long the high return class" investment strategy less attractive. We may still be able to implement a strategy of e.g. going long the high return class, while dollar-for-dollar hedging via an index future short. As noted in another section above, one would have to be careful to backtest a trading strategy taking into account the fact that our underlying information, the 13-F filings, are only available up to 45 days post quarter ends. A rigorous backtest like that is a natural extension of the present study. Eventually, one could also test factor models that add our ownership factors to other types of factors (e.g. fundamental, economic, or technical ones).

We humbly think that, at a minimum, the present study has shown that ownership-related factors may be worth considering for inclusion in factor models for predicting security returns. This is also reinforced by our finding that the ownership factors did an even better job at explaining realized (past four-quarter) returns. As an aside, this of course also underlines the value of having ownership information in realtime (as e.g. brokers, employees, some investors of the funds may have) rather than quarterly, with a 45-day delay.

A related practical question is whether the commercially available "replicator products" are credible. Based on our conclusions, our answer, for the time being, has to to be "maybe" (which, in fairness, it probably always would have been – even if our study had found no effect at all, we could not have excluded the possibility that the sponsors of the replicator products simply conducted better research than ourselves!). In the end, as always, time will tell: as the replicator products extend their track records, it will be ever easier to evaluate whether they achieve statistically significant excess returns.

Lastly, we are excited about the richness of the 13-F ownership information dataset. The present study represents only one example of myriad analyses that could be run on this dataset. One specific suggestion for future research would be to conduct a clustering analysis on the investment funds.

Bibliography

[1] S. Barabanov. **The relationship between institutional ownership, concentration of ownership, bid-ask spread, and returns in nasdaq stocks**. *PhD thesis, Washington State University*, 2002. II.1(b)

[2] F. Black. **Capital market equilibrium with restricted borrowing**. *Journal of Business*, 45(3):444–455, July 1972. II.1(a)

[3] J. Chen, H. Hong and J. Stein. **Breadth of ownership and stock returns**. *Journal of Financial Economics*, 66:171–205, 2002. II.1(b)

[4] N. Choi and R. Sias. **Institutional industry herding**. *Journal of Financial Economics*, 94:469–491, Dec. 2009. II.1(b)

[5] A. Dasgupta, A. Prat and M. Verardo. **Institutional trade persistence and long-term equity returns**. *The Journal of Finance*, 66(2):635–653, Apr. 2011. II.1(b)

[6] M. Duffy. **More etfs play hedge fund copycats**. *The Institutional Investor*, page 243, Oct. 2012. II.2

[7] S. G. Eakins, S. R. Stansell and J. F. Buck. **Analyzing the nature of institutional demand for common stocks**. *Quarterly Journal of Business and Economics*, 37(2):33–48, Mar. 1999. II.3

[8] E. F. Fama and J. D. MacBeth. **Risk, return, and equilibrium: Empirical tests**. *Journal of Political Economy*, 81(3):607–636, 1973. II.3

[9] E. F. Fama and K. R. French. **The cross-section of expected returns**. *Journal of Finance*, 47(2):427–465, June 1992. II.1(a)

[10] R. A. Haugen and N. L. Baker. **Commonality in the determinants of expected stock returns**. *Journal of Financial Economics*, 41(3):401–439, July 1996. II.1(a)

[11] G. James, D. Witten, T. Hastie and R. Tibshirani. **An introduction to statistical learning**. Springer, 2013. III.5(a)

[12] J. Lakonishok, A. Shleifer and R. Vishny. **The impact of institutional trading on stock prices**. *Journal of Financial Economics*, 32:23–43, Aug. 1992. II.1(b)

[13] J. Lintner. **The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets**. *Review of Economics and Statistics*, 47(1):13–37, Feb. 1965. II.1(a)

[14] T. M. Mitchell. **Machine learning**. WCB/McGraw-Hill, 1997. III.5(a)

[15] R. E. Schapire. **A brief introduction to boosting**. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999. IV.1(c)

[16] W. F. Sharpe. **Capital asset prices: a theory of market equilibrium under conditions of risk**. *Journal of Finance*, 19(3):425–442, Sept. 1964. II.1(a)

[17] I. H. Witten, F. Eibe and M. A. Hall. **Data mining**. Morgan Kaufmann, 2011. III.6(a)

A
Descriptions of replicator products

A.1 AlphaClone Hedge Fund Long/Short Index (ALFA)

(a) ALFA Description

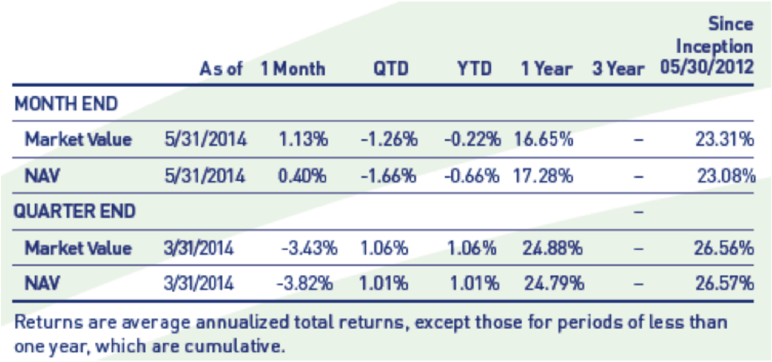
The index (inception date: 30-May-12) is calculated as follows:

- AlphaClone's proprietary Clone Score methodology ranks hedge funds and institutional investors based on the efficacy of replicating their publicly disclosed positions and selects equities from those managers with the highest ranking.
- Clone Scores are recalculated bi-annually and incorporate factors such as the persistence in excess returns over time when following different combinations of a manager?s disclosed positions.
- Constituents are equal weighted but with an overlap bias that would, for example, give a constituent held by twice the number of managers twice the weight.
- 330 hedge funds are chosen for size, length of 13F filing history and investment approach.

(b) ALFA Historical Performance

(c) ALFA Machine Learning Characteristics

AlphaClone does not provide enough specific information publicly in order to e.g. understand what exact features it uses to construct its index. From the general information above, the main feature appears likely to be the "Clone Score" of investors holding a security, which in turn is based partly on the investors' excess returns over time.



	As of	1 Month	QTD	YTD	1 Year	3 Year	Since Inception 05/30/2012
MONTH END							
Market Value	5/31/2014	1.13%	-1.26%	-0.22%	16.65%	–	23.31%
NAV	5/31/2014	0.40%	-1.66%	-0.66%	17.28%	–	23.08%
QUARTER END						–	
Market Value	3/31/2014	-3.43%	1.06%	1.06%	24.88%	–	26.56%
NAV	3/31/2014	-3.82%	1.01%	1.01%	24.79%	–	26.57%

Returns are average annualized total returns, except those for periods of less than one year, which are cumulative.

Figure A.1: ALFA historical statistics

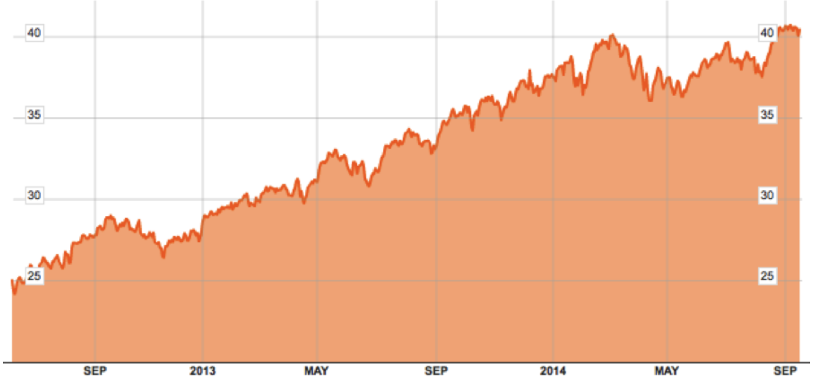


Figure A.2: ALFA historical price graph

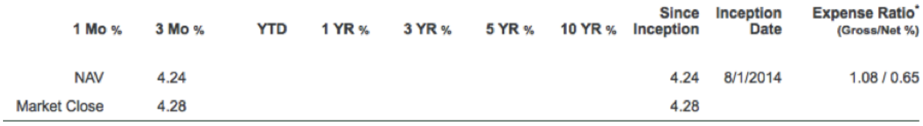
A.2 Direxion iBillionaire Index (IBLN)

(a) IBLN Description

The index (inception date: 25-May-12) is constructed as follows.

- The Index Provider starts with a list of billionaire investors and institutional money managers (Managers) in the financial services industry derived from publicly available information.
- The billionaire list is then narrowed down utilizing the following criteria: i) the Manager has a personal net worth of at least $1 billion that is calculated and verified by industry publications; ii) financial markets and investments are the Managerś primary source of wealth; iii) the public portfolio of the Manager is valued at $1 billion or higher; iv) the Managerś portfolio has at least 10 securities; v) the portfolio turnover is less than 50%; vi) the Managerś equity allocation has a three year return that places the Manager in the top 15 financial billionaires; and vii) the Manager files Form 13F and has investments in the United States.
- The Index Provider limits the number of Managers to ten, however, the number of Managers could be less than ten if there are not ten Managers that meet the above criteria.
- Thirty stocks are selected based on highest allocations by 5-10 billionaires.
- Such stocks must be listed on the NYSE, or NASDAQ; and have at least $1 billion in market capitalization.
- Each of the thirty stocks in the index is allocated a fixed equal weight.

(b) IBLN Historical Performance



1 Mo %	3 Mo %	YTD	1 YR %	3 YR %	5 YR %	10 YR %	Since Inception	Inception Date	Expense Ratio* (Gross/Net %)
NAV	4.24						4.24	8/1/2014	1.08 / 0.65
Market Close	4.28						4.28		

Figure A.3: IBLN historical statistics

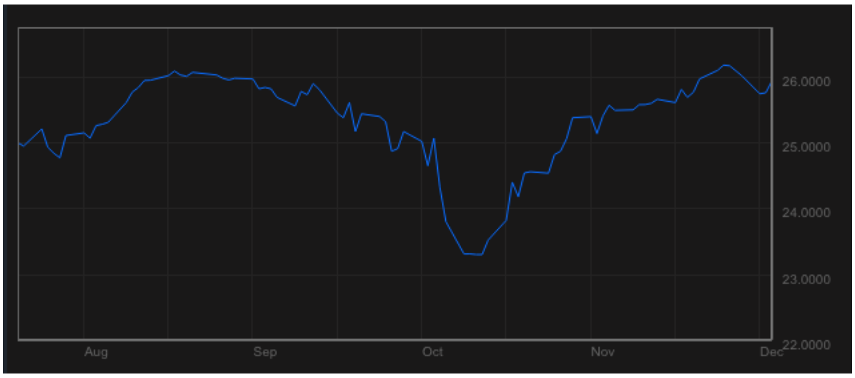


Figure A.4: IBLN historical price graph

(c) IBLN Machine Learning Characteristics

If we interpreted Direxion's index construction methodology for IBLN as a machine learning task, we could say that Direxion is essentially employing a decision tree classifier, using the features listed below, in order to predict a binary variable (will a security be included in the index or not?).

FEATURE	TYPE
Shares listed in the U.S.?	Discrete - Binary
Market capitalization *geq* US$1 billion?	Discrete - Binary
Top holding of a qualified billionaire?	Discrete - Binary

A.3 Solactive Guru Index (GURU)

(a) GURU Description

The Index (inception date: 25-May-12) is constructed as follows.

- The Selection Pool consists of each Top Holding of each hedge fund out of the Hedge Fund Pool according to the latest quarterly regulatory filings reported to the SEC in 13F filings.
- The Hedge Fund Pool is selected from the Hedge Fund Universe once a year on the last Business Day of January on the Hedge Fund Selection Day by applying certain rules and remains unaltered until the next Hedge Fund Selection Day.
- To determine the Selection Pool, the holdings of each hedge fund in the Hedge Fund Pool are ranked by market value as published in the most recent 13F Filing. For each hedge fund, the holding with the largest market value that meets certain requirements s assigned rank number 1 (Top Holding) and enters the Selection Pool. Holdings that do not meet the requirements are not considered for the Selection Pool.
- All holdings in the Selection Pool are then chosen as Index Components.
- If a Hedge Fund holding was an Index Component during the last quarter but is not part of the new Selection Pool, it remains in the index even if it is not a Top Holding any more but ranks second or third by market value and accounts for at least 4.8% of the total hedge fund market value.

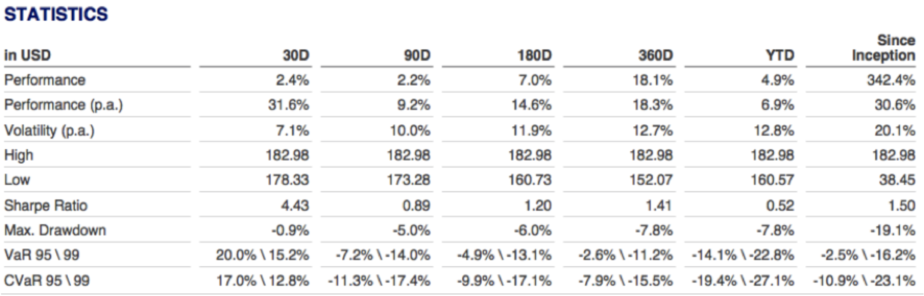
where some of the terms used above are defined as follows by Solactive:

Hedge Fund Pool in respect of a Hedge Fund Selection Day are those Hedge Funds out of the Hedge Fund Universe that fulfil the following conditions according to the most recent 13F Filing: (a) total market value of portfolio of at least US$500 million; (b) market value of top hedge fund portfolio holding accounts for at least 4.8% of total hedge fund portfolio market value; (c) year-over-year hedge fund portfolio turnover (Y-Y Fund Portfolio Turnover) of less than or equal to 50%. In case the market value of the top holding of a hedge fund is significantly larger than 4.8% of the total hedge fund portfolio market value, the Index Committee may decide to include a hedge fund, even if the turnover exceeds 50%.

Hedge Fund Universe is a list of hedge funds compiled from sources including Morningstar list of hedge funds, Bloomberg and Bloomberg magazine, and Barron?s Top 100 Hedge Funds.

Selection Pool , in respect of a Selection Day, is constituted by those holdings that fulfil the following conditions: (a) listed on a regulated stock exchange in the United States; (b) free float market capitalization of at least US$100 million; (c) average daily trading volume in the last three months of at least US$10 million; (d) average monthly trading volume of at least 75,000 shares in each of the last six months; (e) the holding not being an exchange traded product or a fund; (f) the Top holding that fulfils the above criteria and has a market value of at least 4.8% of the total hedge fund portfolio value.

(b) GURU Historical Performance



STATISTICS

in USD	30D	90D	180D	360D	YTD	Since Inception
Performance	2.4%	2.2%	7.0%	18.1%	4.9%	342.4%
Performance (p.a.)	31.6%	9.2%	14.6%	18.3%	6.9%	30.6%
Volatility (p.a.)	7.1%	10.0%	11.9%	12.7%	12.8%	20.1%
High	182.98	182.98	182.98	182.98	182.98	182.98
Low	178.33	173.28	160.73	152.07	160.57	38.45
Sharpe Ratio	4.43	0.89	1.20	1.41	0.52	1.50
Max. Drawdown	-0.9%	-5.0%	-6.0%	-7.8%	-7.8%	-19.1%
VaR 95 \ 99	20.0% \ 15.2%	-7.2% \ -14.0%	-4.9% \ -13.1%	-2.6% \ -11.2%	-14.1% \ -22.8%	-2.5% \ -16.2%
CVaR 95 \ 99	17.0% \ 12.8%	-11.3% \ -17.4%	-9.9% \ -17.1%	-7.9% \ -15.5%	-19.4% \ -27.1%	-10.9% \ -23.1%

Figure A.5: GURU historical statistics

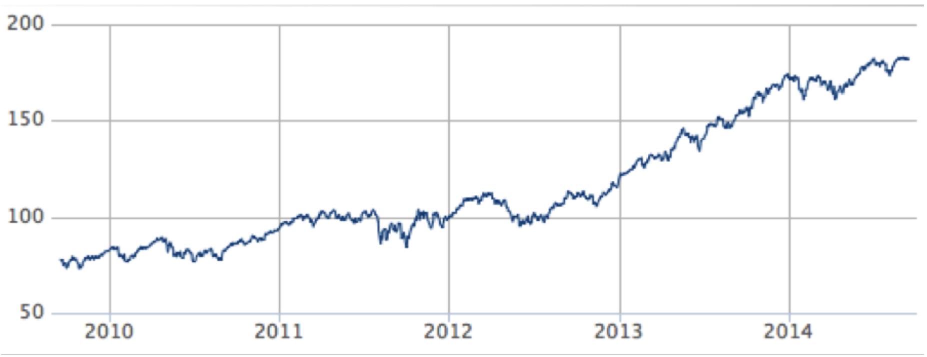


Figure A.6: GURU historical price graph

(c) GURU Machine Learning Characteristics

If we interpreted Solactive's index construction methodology for GURU as a machine learning task, we could say that Solactive is essentially employing a decision tree classifier, using the features listed below, in order to predict a binary variable (will a security be included in the index or not?).

FEATURE	TYPE
Total market cap of all Hedge Fund holdings of stock in top 500?	Discrete - Binary
Shares listed in the U.S.?	Discrete - Binary
Market capitalization *geq* US$100 million?	Discrete - Binary
Average 3-month trading volume $\geq$ US$ 10 million?	Discrete - Binary
Average monthly trading volume $\geq$ 75000 shares over last 6 months?	Discrete - Binary
Top holding of a qualified hedge fund?	Discrete - Binary

A.4 Solactive Hedge Fund Holdings US Index (HEDGEUS)

(a) HEDGEUS Description

The index (inception date: 28-Aug-14) is composed of a basket of shares, which are listed on a regulated exchange of the United States of America and which are invested in by hedge funds, as indicated in 13F Filings. These shares are selected from a Selection Pool which is determined by the index sponsor according to the following methodology (note that capitalized expressions represent terms defined by the index provider, Solactive, and are explained further below):

- *Aggregated Concentration Filter:* select the top five hundred shares, exchange traded funds and similar securities, as ranked in descending order according to their respective Aggregate Market Value Across 13F;
- *Share and Liquidity Filter:* after applying the Aggregated Concentration Filter, retain shares, which belong to the Eligible Universe. The Eligible Universe consists of Shares which: are listed on an exchange of the United States of America; which are not Shares issued by funds; and have a three-month Average Daily Volume exceeding US$3 million;
- *Stock Concentration Filter:* after applying the Share and Liquidity Filter, select the top one hundred shares, as ranked in descending order according to their Hedge Fund Concentration
- *Position Change Filter:* after applying the Stock Concentration Filter, select the top thirty shares, ranked in descending order according to their respective Position Change
- An equally-weighted basked is constructed from the selected shares

where some of the terms used above are defined as follows by Solactive:

Aggregate Market Value Across 13F is, for each Company, the combined market value of shares issued by such company across all hedge funds holdings.

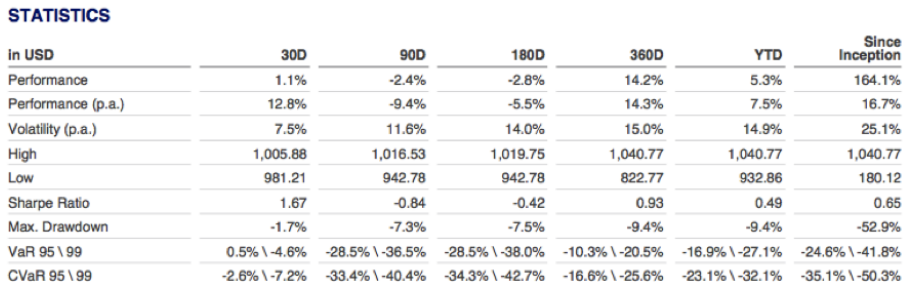
Hedge Fund means any institutional investor classified as a Hedge Fund by Bloomberg.

Hedge Fund Pool is the universe of all hedge funds obliged to file a 13F.

Hedge Fund Concentration means, in respect of a company, the ratio of (i) the number of shares issued by such Company and held by Hedge Funds with in the Hedge Fund Pool according to their 13F Filings to (ii) the total number outstanding shares issued by such company.

Position Change means, in respect of a company, the ratio of (i) the net number of shares issued by such company which have been bought by Hedge Funds within the Hedge Fund Pool according to their 13F Filings between the penultimate 13F Filing Date and the latest 13F Filing date to (ii) the total number outstanding shares issued by such company.

(b) HEDGEUS Historical Performance



STATISTICS

in USD	30D	90D	180D	360D	YTD	Since Inception
Performance	1.1%	-2.4%	-2.8%	14.2%	5.3%	164.1%
Performance (p.a.)	12.8%	-9.4%	-5.5%	14.3%	7.5%	16.7%
Volatility (p.a.)	7.5%	11.6%	14.0%	15.0%	14.9%	25.1%
High	1,005.88	1,016.53	1,019.75	1,040.77	1,040.77	1,040.77
Low	981.21	942.78	942.78	822.77	932.86	180.12
Sharpe Ratio	1.67	-0.84	-0.42	0.93	0.49	0.65
Max. Drawdown	-1.7%	-7.3%	-7.5%	-9.4%	-9.4%	-52.9%
VaR 95 \ 99	0.5% \ -4.6%	-28.5% \ -36.5%	-28.5% \ -38.0%	-10.3% \ -20.5%	-16.9% \ -27.1%	-24.6% \ -41.8%
CVaR 95 \ 99	-2.6% \ -7.2%	-33.4% \ -40.4%	-34.3% \ -42.7%	-16.6% \ -25.6%	-23.1% \ -32.1%	-35.1% \ -50.3%

Figure A.7: HEDGEUS historical statistics

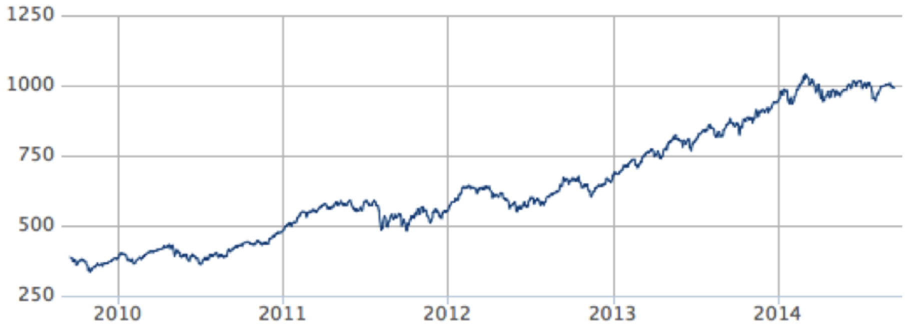


Figure A.8: HEDGEUS historical price graph

(c) HEDGEUS Machine Learning Characteristics

If we interpreted Solactive's index construction methodology for HEDGEUS as a machine learning task, we could say that Solactive is essentially employing a decision tree classifier, using the features listed below, in order to predict a binary variable (will a security be included in the index or not?).

FEATURE	TYPE
Total market cap of all Hedge Fund holdings of stock in top 500?	Discrete - Binary
Shares listed in the U.S.?	Discrete - Binary
Average 3-month trading volume $\geq$ US$ 3 million?	Discrete - Binary
Hedge Fund Concentration in top 100?	Discrete - Binary
Position Change in top 30?	Discrete - Binary

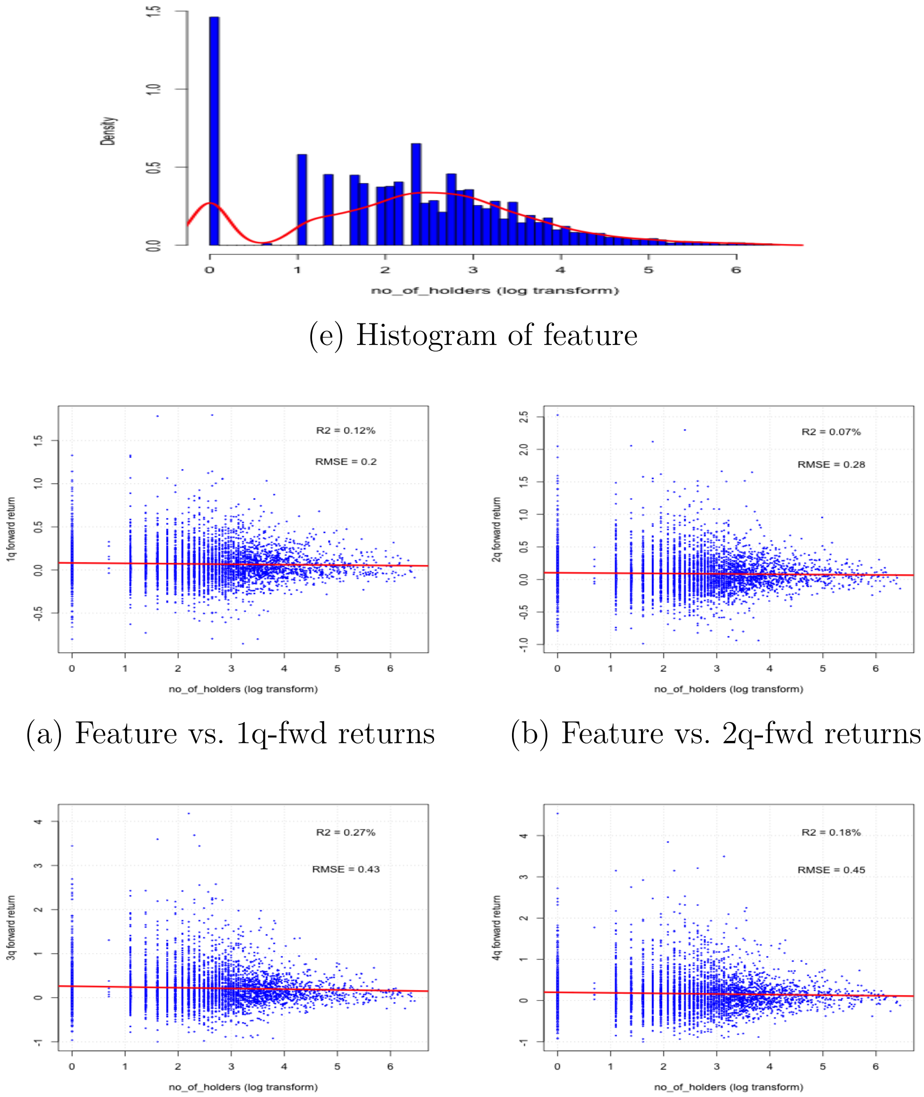
B
Exploratory analyses of features

In the following, we present basic exploratory analyses on our features:

- histogram of the feature
- scatter plot of the feature against the predicted variable – 1/2/3/4-quarter forward return

B.1 X1 - Number of holders

(a) X1 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

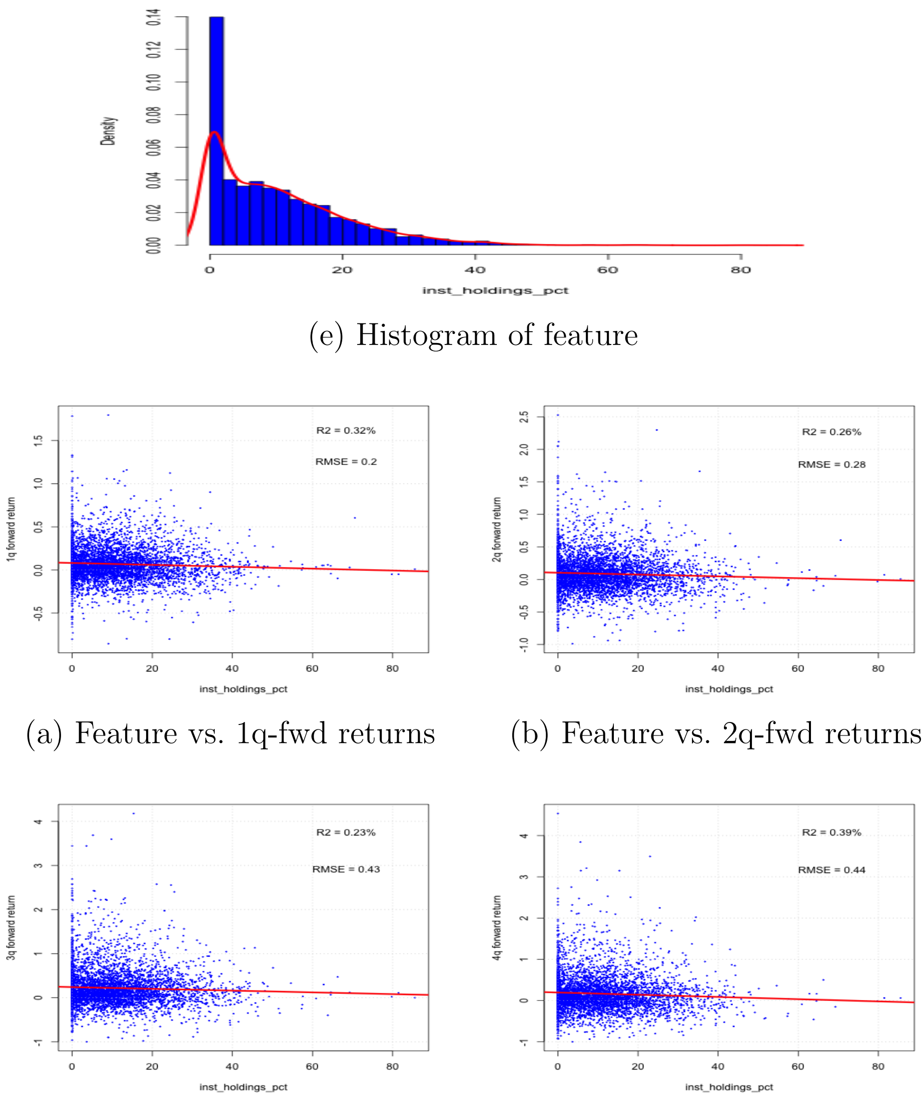
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.1: Feature X1 exploratory analyses (2Q2005)

B.2 X2 - Fraction of institutional investors

(a) X2 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.2: Feature X2 exploratory analyses (2Q2005)

B.3 X3 - Adjusted no. of investors

(a) Relationship between number of holders, market capitalization and trading volume

In the following, we provide:

- scatter plot of number of holders vs. market capitalization (Figure B.3);
- regression of number of holders vs. market capitalization;(Figure B.4);
- multiple regression of number of holders as a function of market capitalization and trading volume (Figure B.5);
- residual plot of the multiple regression (Figure B.6); and
- plot of adjusted residuals (Figure B.7).

A scatter plot and regression for the relationship between number of holders and volume have already been shown in the main body text, in section III.3(a). Note that market capitalizations are as of the quarter-end, while trading volumes are calculated as the median daily US$ volume since the start of the quarter. We show some of the scatter plots on logarithmic scales in order to facilitate visualization. Where this is the case, we increased the number of holders data by 1 throughout, in order to be able to show securities with 0 holders. In keeping with the practical objectives of the study, we limited the security universe to companies with a market capitalization $\geq$ US$100 million and security types to common and preferred shares.

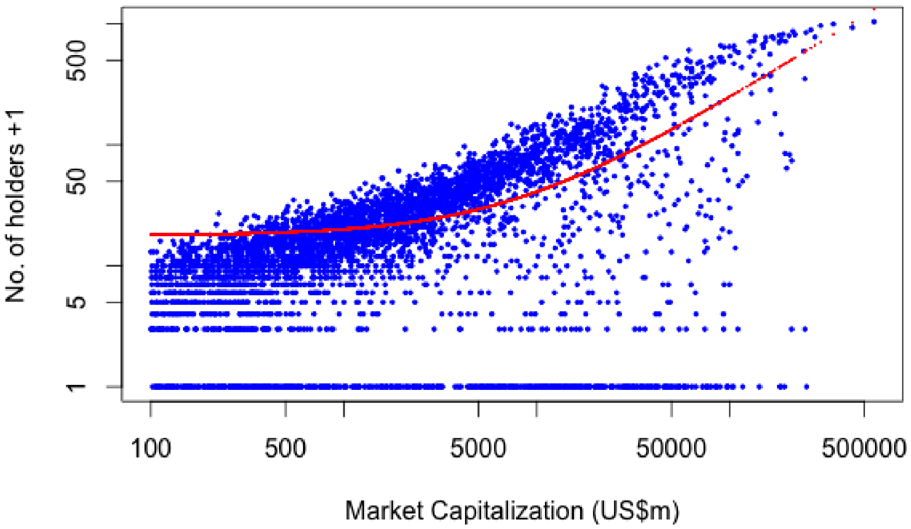
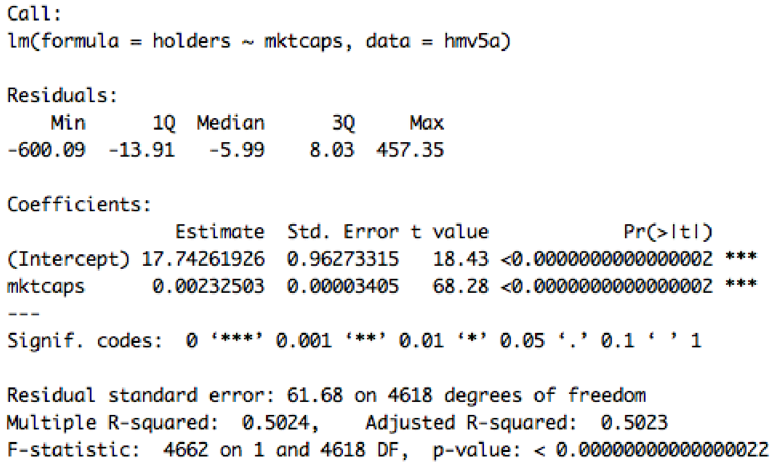
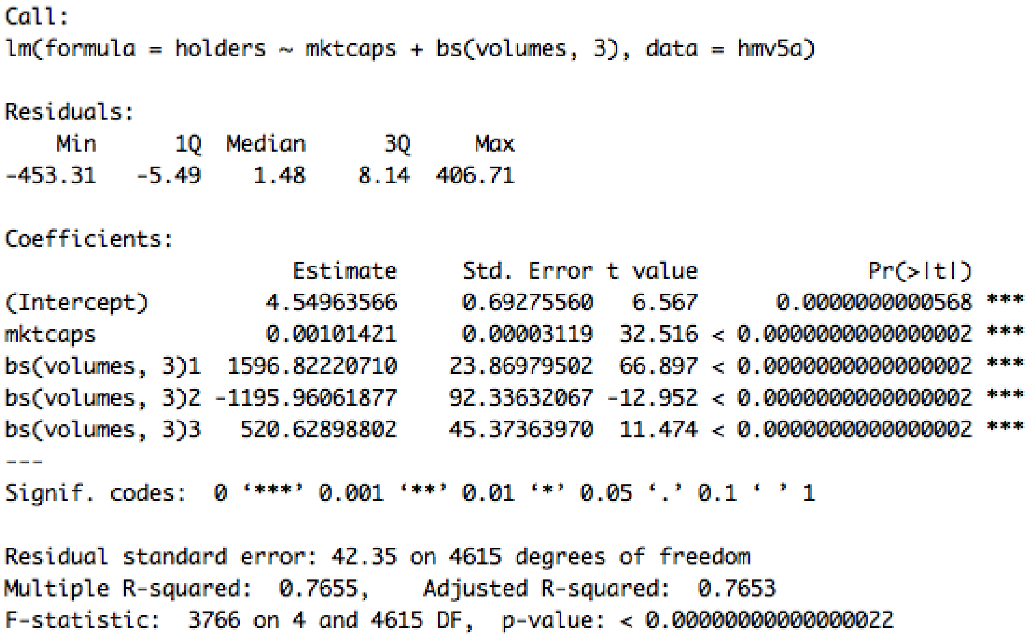


Figure B.3: No. of holders vs. mkt. caps. (2Q2014, log-log)



```
Call:
lm(formula = holders ~ mktcaps, data = hmv5a)

Residuals:
    Min      1Q  Median      3Q     Max 
-600.09  -13.91   -5.99    8.03  457.35 

Coefficients:
               Estimate  Std. Error t value            Pr(>|t|)    
(Intercept) 17.74261926  0.96273315   18.43 <0.0000000000000002 ***
mktcaps      0.00232503  0.00003405   68.28 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.68 on 4618 degrees of freedom
Multiple R-squared:  0.5024,	Adjusted R-squared:  0.5023 
F-statistic:  4662 on 1 and 4618 DF,  p-value: < 0.00000000000000022
```

Figure B.4: Regression of no. of holders vs. mkt. caps. (2Q2014)



```
Call:
lm(formula = holders ~ mktcaps + bs(volumes, 3), data = hmv5a)

Residuals:
    Min      1Q  Median      3Q     Max 
-453.31   -5.49    1.48    8.14  406.71 

Coefficients:
                     Estimate     Std. Error t value             Pr(>|t|)    
(Intercept)        4.54963566     0.69275560   6.567      0.0000000000568 ***
mktcaps            0.00101421     0.00003119  32.516 < 0.0000000000000002 ***
bs(volumes, 3)1  1596.82220710    23.86979502  66.897 < 0.0000000000000002 ***
bs(volumes, 3)2 -1195.96061877    92.33632067 -12.952 < 0.0000000000000002 ***
bs(volumes, 3)3   520.62898802    45.37363970  11.474 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.35 on 4615 degrees of freedom
Multiple R-squared:  0.7655,	Adjusted R-squared:  0.7653 
F-statistic:  3766 on 4 and 4615 DF,  p-value: < 0.00000000000000022
```

Figure B.5: Regression of no. of holders vs. mkt. caps. and trading volume (2Q2014)

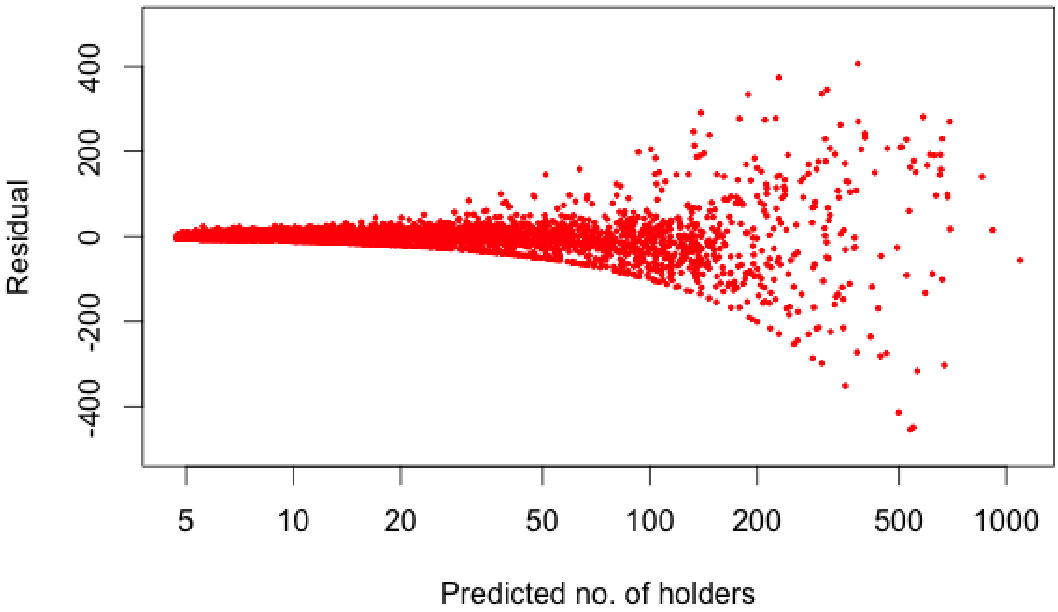


Figure B.6: Residuals vs. predicted no. holders (2Q2014)

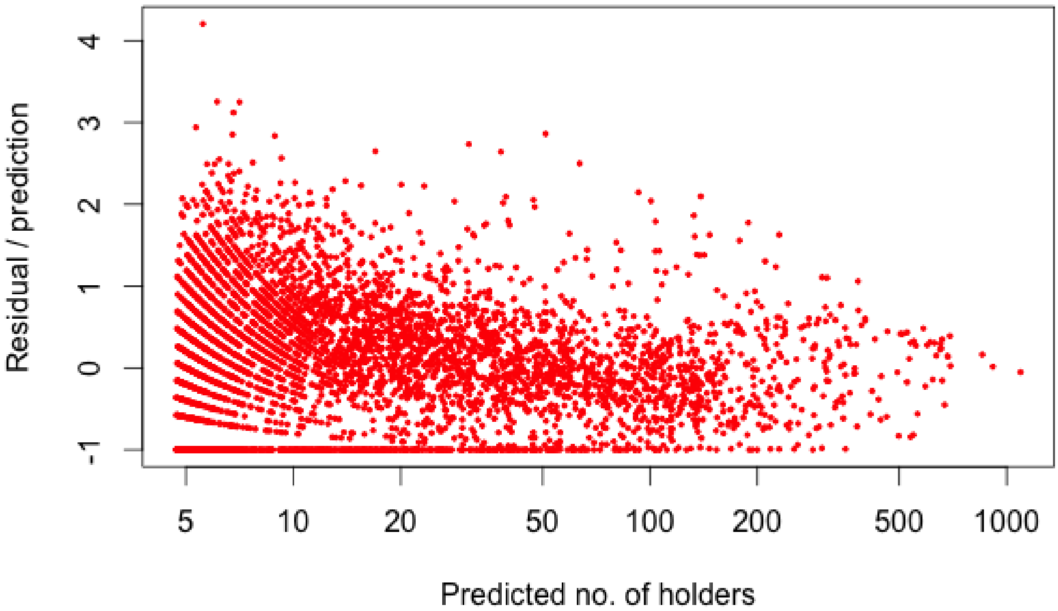


Figure B.7: Residuals/prediction vs. predicted no. holders (2Q2014)

(b) Relationship between institutional ownership fraction and market capitalization

Figure B.8 and Figure B.9 show a scatter plot and regression for the relationship between institutional ownership fraction and market capitalization, respectively. A scatter plot and regression for the relationship between institutional ownership fraction and volume have already been shown in the main body text, in section III.3(a).

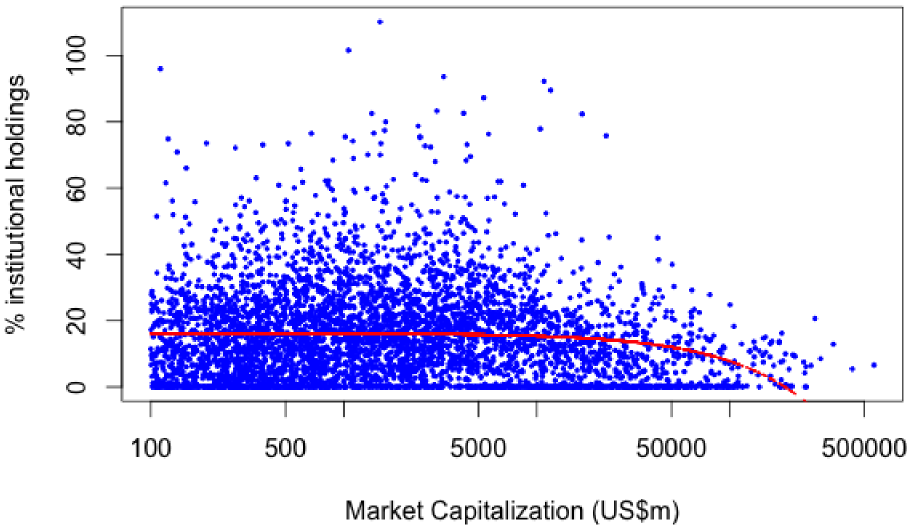
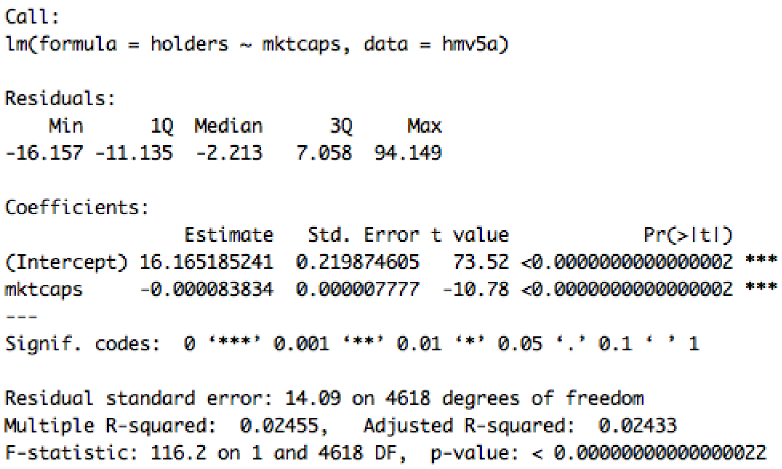


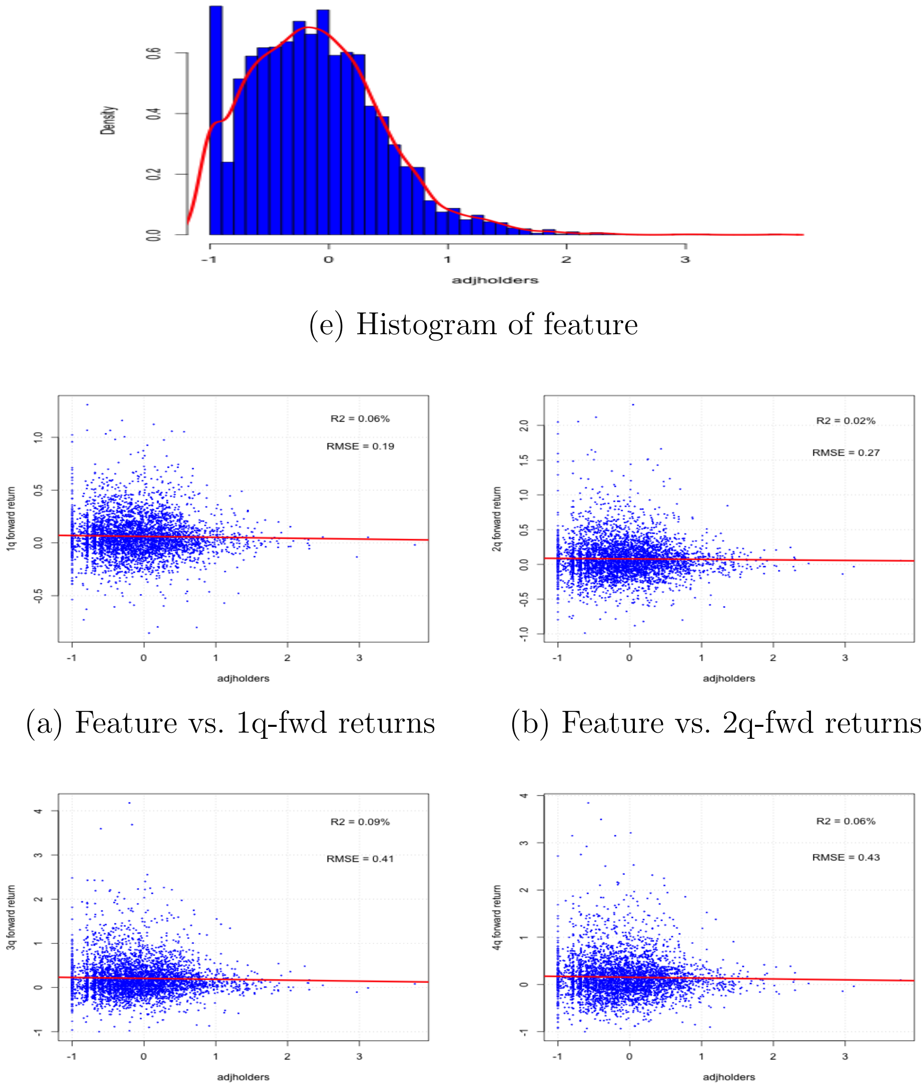
Figure B.8: Institutional holdings% vs. trading volume (2Q2014, log scale on x-axis)



```
Call:
lm(formula = holders ~ mktcaps, data = hmv5a)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.157 -11.135  -2.213   7.058  94.149 

Coefficients:
                Estimate   Std. Error t value            Pr(>|t|)    
(Intercept) 16.165185241  0.219874605   73.52 <0.0000000000000002 ***
mktcaps     -0.000083834  0.000007777  -10.78 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.09 on 4618 degrees of freedom
Multiple R-squared:  0.02455,	Adjusted R-squared:  0.02433 
F-statistic: 116.2 on 1 and 4618 DF,  p-value: < 0.00000000000000022
```

Figure B.9: Regression of institutional holdings% vs. trading volume (2Q2014)

(c) X3 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

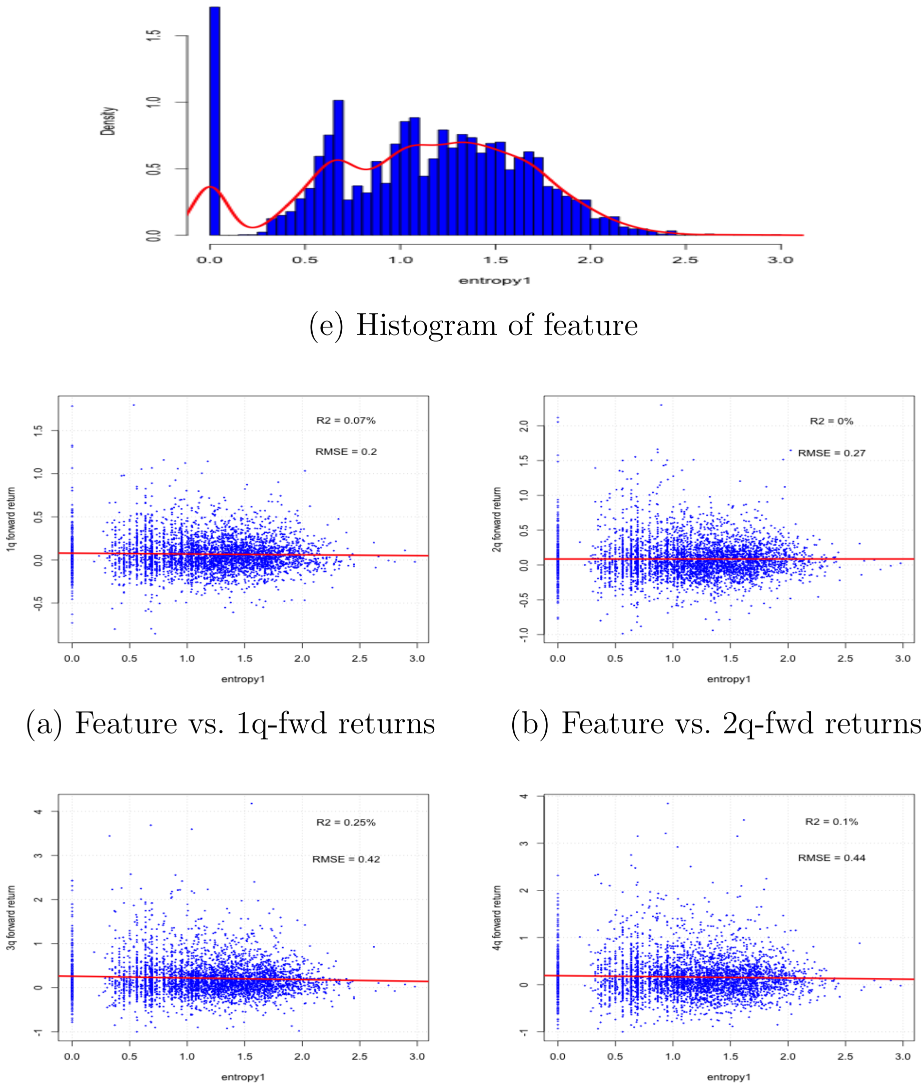
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.10: Feature X3 exploratory analyses (2Q2005)

B.4 X4 - Entropy 1

(a) X4 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

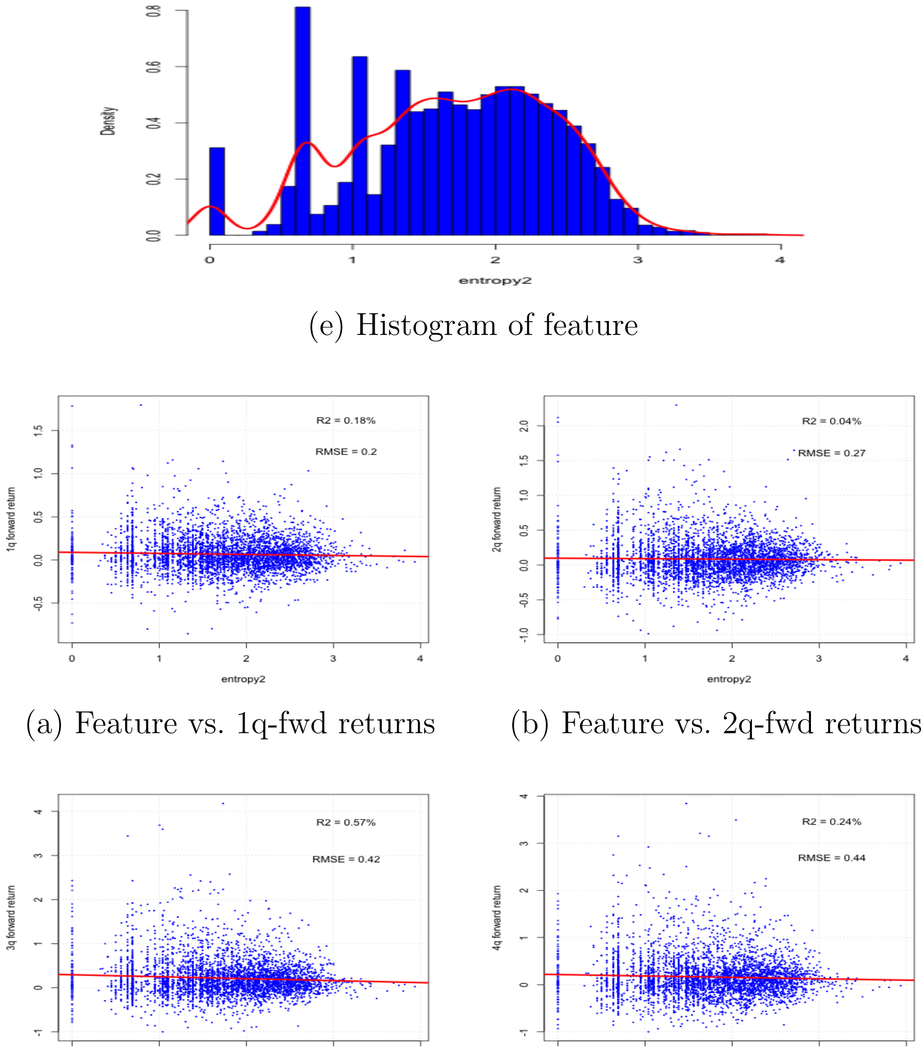
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.11: Feature X4 exploratory analyses (2Q2005)

B.5 X5 - Entropy 2

(a) X5 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

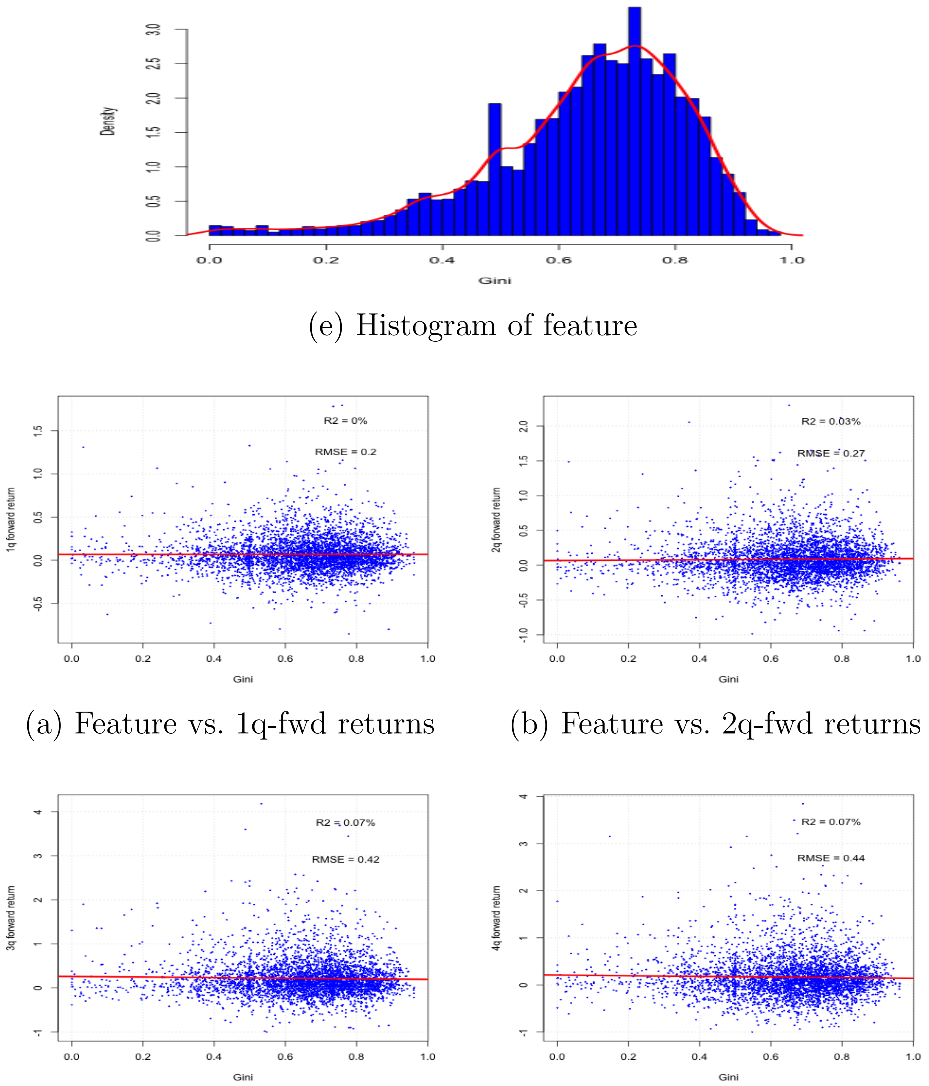
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.12: Feature X5 exploratory analyses (2Q2005)

B.6 X6 - Gini

(a) X6 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

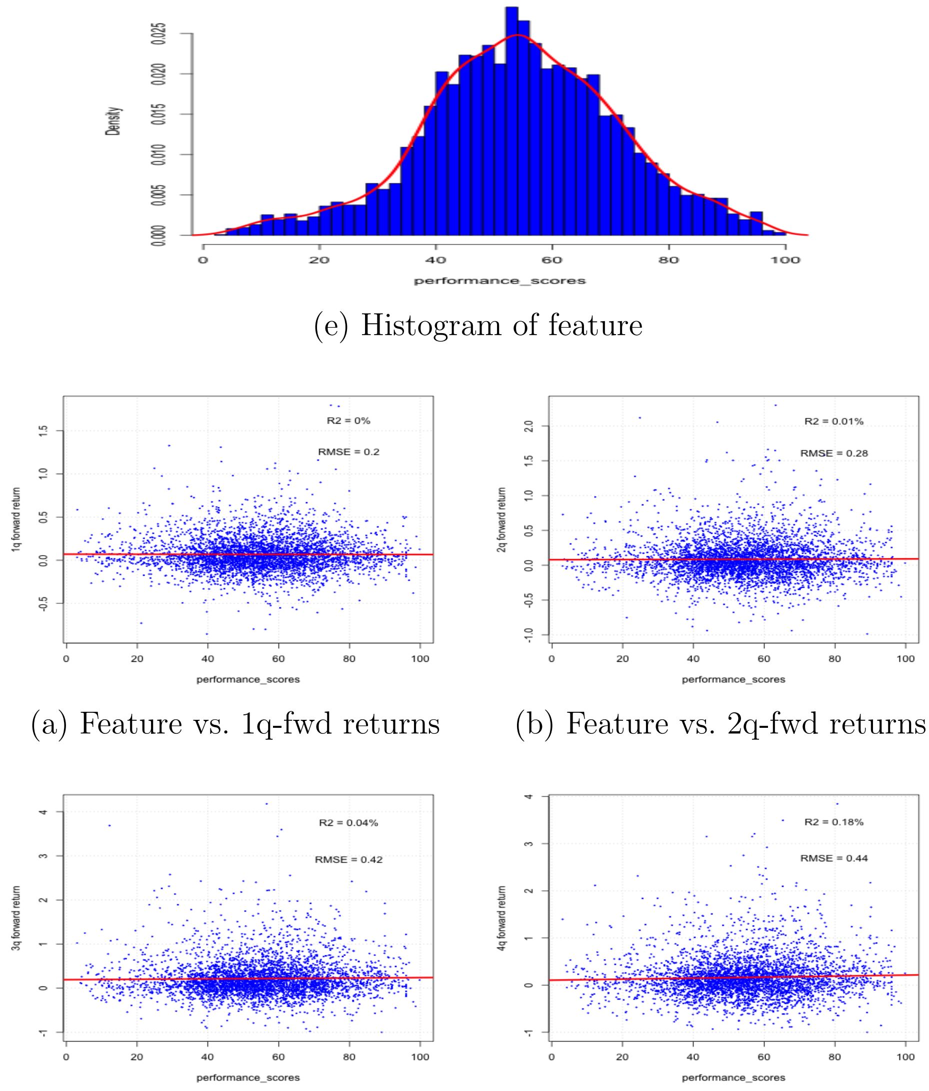
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.13: Feature X6 exploratory analyses (2Q2005)

B.7 X7 - Performance scores

(a) X7 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

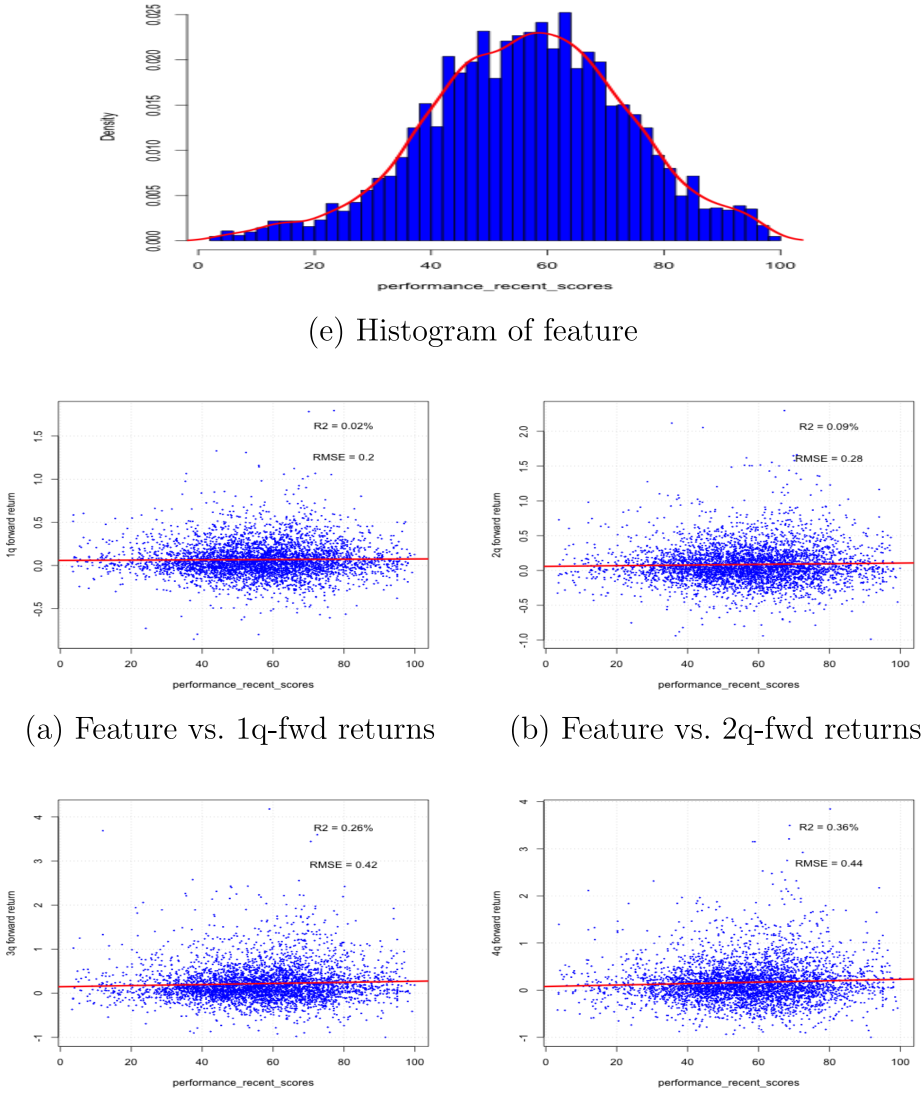
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.14: Feature X7 exploratory analyses (2Q2005)

B.8 X8 - Performance scores (recent)

(a) X8 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

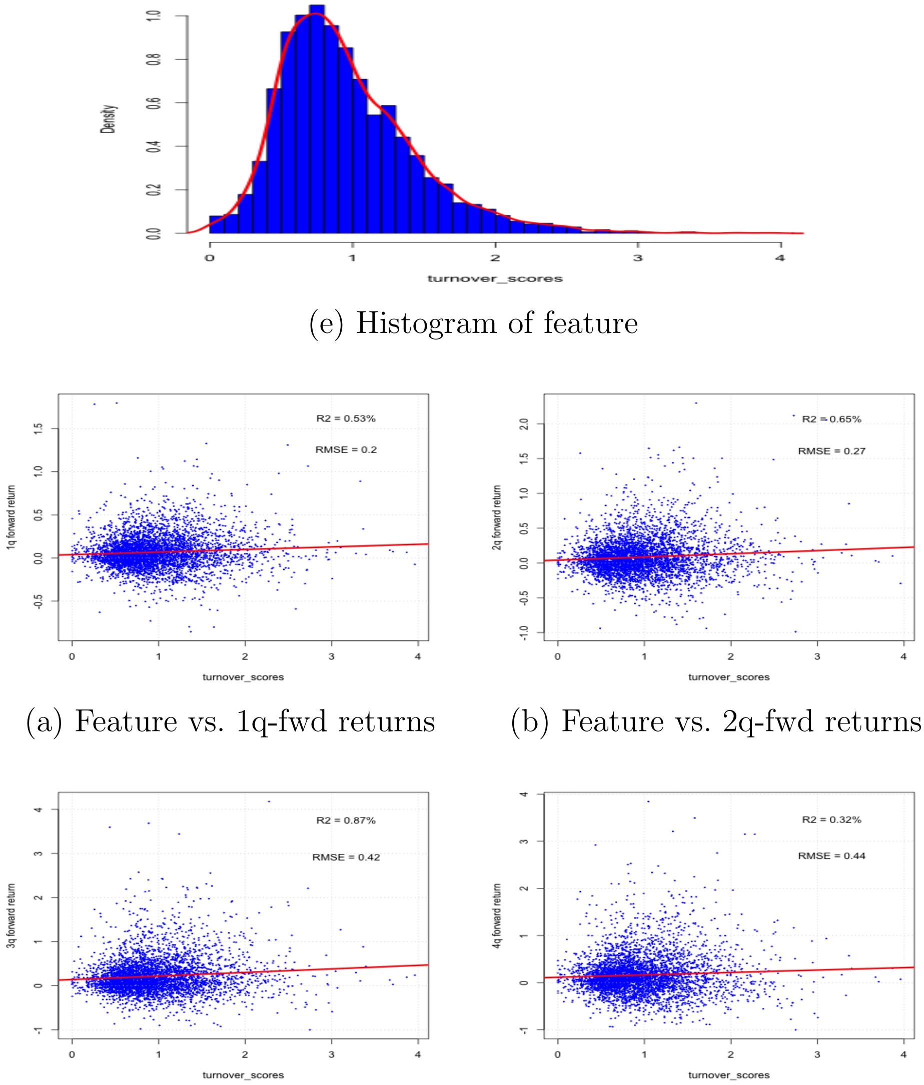
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.15: Feature X8 exploratory analyses (2Q2005)

B.9 X9 - Turnover scores (recent)

(a) X9 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

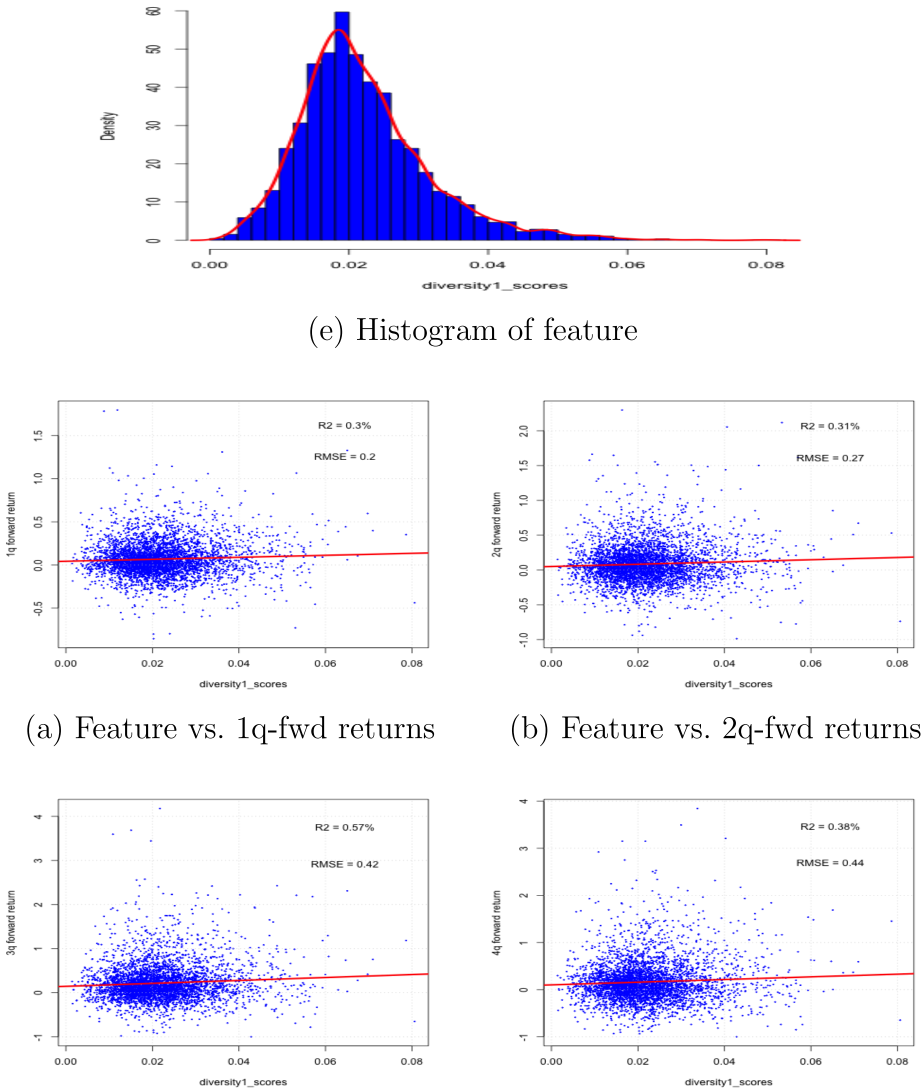
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.16: Feature X9 exploratory analyses (2Q2005)

B.10 X10 - Diversity 1 scores

(a) X10 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

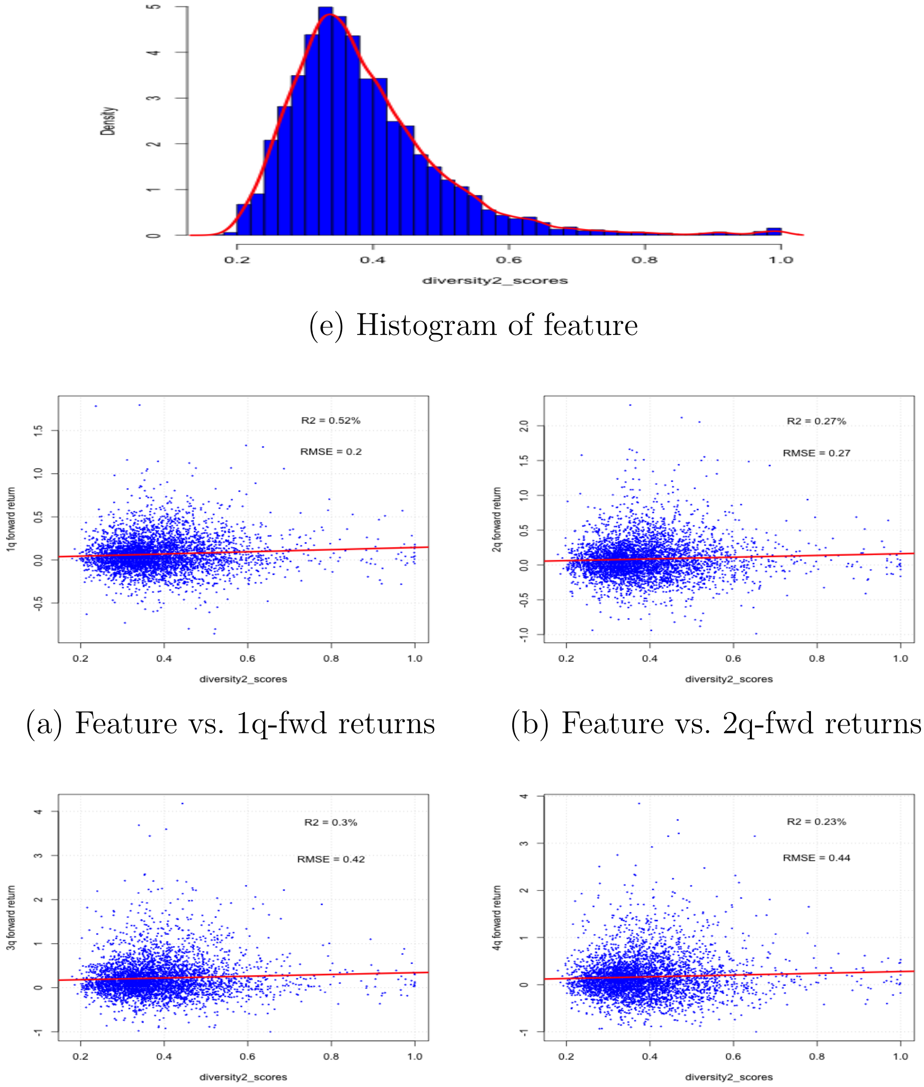
(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.17: Feature X10 exploratory analyses (2Q2005)

B.11 X11 - Diversity 2 scores

(a) X11 - Histogram and scatter plots



(e) Histogram of feature

(a) Feature vs. 1q-fwd returns

(b) Feature vs. 2q-fwd returns

(c) Feature vs. 3q-fwd returns

(d) Feature vs. 4q-fwd returns

Figure B.18: Feature X11 exploratory analyses (2Q2005)