Machine Learning, Investor Sentiment and Stock Market Portfolios

Iles Acly Ayati

Dissertation written under the supervision of Dr. João Brogueira de Sousa.

Dissertation submitted in partial fulfilment of requirements for M.Sc. in Economics, major in Banking & Finance, at Universidade Católica Portuguesa and for M.Sc. in Finance at BI Norwegian Business School, November 2019.

Abstract

This paper investigates linear and non-linear relationships between a selected set of investor sentiment measures and a range of stock portfolio returns from the Kenneth French Data Library. First, I test for contemporaneous linear relationships between the features and portfolio returns, both individually and jointly. Second, I test for causal relationships using the traditional Granger causality framework as well as an alternative machine learning adaptation of said framework. The findings show that some of these features do indeed have significant contemporaneous relationships with the stock portfolios. Moreover, the causality tests show that a subset of these features also seem to affect some portfolio returns, though with varying consistency.

Este artigo investiga as relações lineares e não-lineares entre medidas de sentimento dos investidores e carteiras de ações da Kenneth French Data Library. Primeiro, testo as relações lineares contemporâneas entre as medidas de sentimento e os retornos do portfólio, individualmente e em conjunto. Segundo, testo relações causais pelo o método tradicional de causalidade à Granger, bem como uma adaptação de machine learning desse método. Os resultados mostram que algumas dessas medidas têm, de fato, relações contemporâneas significativas com as carteiras de ações. Além disso, os testes de causalidade mostram que um subgrupo desses medidas também parece afetar alguns retornos, embora com consistência variável.

Contents

1 Introduction

This paper examines the relationships between a selected set of investor sentiment measures with a range of U.S. stock market portfolios, categorized by size and book-to-market ratio. The selected investor sentiment features are closed-end fund discounts, the change in the VIX Index, and a set of the CFTC Commitment of Traders Report data. While it has been documented that the VIX Index and closed-end fund discounts are good predictors of stock market returns (Lee et al., 1991; Bekaert and Hoerova, 2014), the literature analyzing the Commitment of Traders (COT) report and its use for stock market predictions is limited. Still, and especially by some foreign exchange speculators, this data is considered a market signalling tool. Furthermore, I could not find any literature analyzing the causality of investor sentiment measures on the stock market and vice versa. If there existed concrete evidence that changes in these features affect the following market movements, they could potentially be tools to be considered for further research and investment analysis. For this reason, this paper investigates the contemporaneous as well as causal relationships that might exist between a selected set of sentiment features and stock market portfolios. To capture more information about the relationships, both contemporaneous and causal, a wide set of models are deployed. Some of these are linear machine learning algorithms that are popular and frequently used in finance and economics. Lastly, a Random Forest Regression is deployed to capture non-linear relationships between features and portfolios. The results largely confirm the existing literature: Closed-end fund discounts and changes in the VIX Index are indeed significant predictors of stock market movements. Notably, changes in the VIX Index is a powerful feature to explain stock market movements, even when used as a lone feature. What is new here is the evidence that trader positioning on the S&P 500 futures market can indeed explain price movements, especially that of stocks with high book-to-market ratios. Moreover, the results show that the causal relationships are not as evident. There are cases where forecast errors are considerably improved, especially for high book-to-market ratio stocks, but due to some important violations of the Classical Linear Regression Model assumptions, any steadfast conclusions are at risk of being spurious.

2 Related Literature

Empirical studies on stock price predictions are among the most popular in financial market research. Economists are divided when it comes to modelling price movements. On one side, you have the proponents of randomness in the price movements - they don't follow any measurable patterns and are therefore impossible to fully predict. Samuelson (1965) set to prove that stock prices follow a random walk if rational competitive investors require a fixed rate of return. *Efficient capital markets: A review of theory and empirical work* (Malkiel and Fama, 1970) is a highly cited paper that demonstrates how stock prices indeed follow a process close to a random walk.

The opposing view that market movements are rooted in human behavior and our expectations - which are not always rational, has also been well documented, though usually on a more anecdotal basis. For instance, the early 60's were characterized by a high demand for small, newly issued growth stocks (Malkiel, 1999). In pages 55-57, Malkiel mentions a 'new-issue mania' that later led to a decline in growth stocks much stronger than the rest of the market. Investors seemed to have ignored the uncertainty of newly issued electronics stocks in favor of promised overnight returns. On the other hand, De Long et al. (1990) takes a more theoretical approach, and argues that changes in investor sentiment are not fully countered by arbitrageurs and therefore affect security returns. Arbitrageurs are "fully rational" investors who trade to exploit that the price of a security may sometimes deviate from that of its perfect substitute - a portfolio of other securities with the same risk and return profile. For example, if the price of a security is below the price of the substitute portfolio, arbitrageurs will sell the portfolio and buy the security until the prices are back to equal. When the substitute is indeed perfect, this arbitrage is completely riskless. Thus, arbitrageurs have perfectly elastic demand for the security at the price of its substitute portfolio. Such riskless arbitrage is very effective for derivative securities such as futures and options, but for individual stocks and bonds, perfect substitutes are usually not available. Instead, *close* substitutes are traded, which implies that the arbitrageurs' demand is no longer perfectly elastic as there will exist some degree of risk-return differences between the security and its close substitute. In that regard, De Long et al. (1990) argues that arbitrageurs will not be able to completely offset investor sentiment. This "noise trader approach" was further developed upon by other economists. Lee et al. (1991)

examined closed-end funds and proposed that their discount fluctuations are driven by investor sentiment. One thing that both schools of thought agree upon is the difficulty in predicting asset price movements. Many studies aim to provide a rational model directly proving that prices are indeed predictable (De Bondt and Thaler, 1985; Jegadeesh, 1990; Lo and MacKinlay, 1990). However, profiled economists seem to share the opinion that they don't really add more to what we already have in form of the Capital Asset Pricing Model (CAPM). [1]

The Latin saying; *'Cum hoc ergo propter hoc'*, is a reminder about the logical fallacy we are faced with when drawing conclusions that are exclusively based on observations of two events occurring simultaneously. Today, this is better known as; 'correlation is not causation'. Granger (1969) neatly provided the framework to test for causality between time series, and is widely recognized as *the* method for investigating causal relationships in econometrics. The basic premise of the paper relied heavily on the notion that cause cannot happen after effect. Consequently, it focused on 'lagged' causality in linear relationships. Granger's definition of causality is as follows: X_t is causing Y_t if we are better able to predict Y_t using all available information U than if the information apart from X_t had been used. Obviously, using 'all available information' is impossible, which in practice means that we have to select the relevant information for the time series Y_t instead. To test for instantaneous causality, we would also include the contemporaneous or new information in the universe U, while traditional or 'lagged' causality will only contain information up to and including $t-1$. Due to its simplicity it gained a lot of popularity as well as criticism by economists. The paper was later extended to include frameworks for instantaneous causality as well as tackling some issues and limitations regarding non-linear and deterministic relationships. In an efficient market, the best price forecast of tomorrow's prices are today's prices, since the probability of an increase equals that of a decrease (Malkiel and Fama, 1970). By that reasoning, the presence of causality could be interpreted as evidence of market inefficiencies of the weak form (Niarchos and Alexakis, 1998). However, not all economists agree on this point. Two cointegrated time series as defined in Engle and Granger (1987), implies Granger causality in at least one direction, but it has been shown that there is no general equivalence between the existence of arbitrage opportunities and cointegration (Dwyer Jr and Wallace, 1992). This is an important premise for this thesis that

[1] Thaler, R.H. and Fama, E. F. (2014, October 18). Chicago Booth Review Interview: https://review.chicagobooth.edu/economics/2016/video/are-markets-efficient

I will stress in the methodology. In finance, most studies on causality focus on the relationship between stock prices and their respective trading volume, where most test for instantaneous causality (Hiemstra and Jones, 1994). However, there are studies in this field relying exclusively on traditional, lagged Granger causality tests (Smirlock and Starks, 1988; Jain and Joh, 1988). Although such tests can be good in explaining linear causal relations, their usefulness in non-linear casual relations is debatable (Hiemstra and Jones, 1994). For this reason, traditional Granger causality tests might overlook significant non-linear patterns that ensemble algorithms such as Random Forest may capture.

In recent years, machine learning has practically become a buzzword in finance and economics. It is safe to say that machine learning tools have for many years changed the playing field and that they are continuously developed by analytic departments across the industries. However, there are limitations to all tools, and the issues associated with financial time series predictions still prevail - the time series are inherently noisy, complex and chaotic (Kumar and Thenmozhi, 2006; Kara et al., 2011; Patel et al., 2015). But that has not stopped researchers from adopting new and experimental techniques in their pursuit. An example of innovative adaptations of machine learning tools in economics is Natural Language Processing (NLP). NLP is used to analyze the effects of web texts for stock market predictions (Das and Chen, 2007; Tetlock, 2007; Tetlock et al., 2008). Words from web articles are collected and transformed into measures of investor sentiment by their positive or negative charges. They are then fed to Neural Networks along with stock returns, which in turn, are predicted. Tetlock (2007) finds that high media pessimism is able to predict downward pressure on prices, and that unusually high or low pessimism predicts high market trading volume. Another example is Patel et al. (2015), which compares four popular ensemble algorithms; Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest and Naive Bayes. The study focuses on data pre-processing where continuous data is transformed into discrete data to improve prediction performances. The results show that Random Forest produces the best classification predictions. Tsai et al. (2011) also applies classifier ensemble methods to analyze stock returns. It compares prediction performances using hybrid methods of majority voting and bagging, as well as compare combining two types of classifier ensembles with single classifiers. The findings are somewhat conflicting, but they do conclude that on average, combining multiple classifiers (e.g. an ensemble of ANNs, random forest classifiers, and logistic regressions) provide better prediction

accuracies than a homogenous ensemble of classifiers (e.g. ensemble of ANNs or random forest classifiers only).

The introduction mentioned a lack of literature on the use of the CFTC Commitment of Traders report. Masters and White (2008) is a three-act report conducted by hedge fund manager Michael W. Masters during the financial crisis in 2008 and 2009. It largely centers around the proposition that 'Index Speculators', a subset of institutional investors who speculate using index-tracker futures induce an artificial demand for commodities such that their market prices are pushed up to levels beyond the fundamentals. However, this report is not a traditional research paper and does not justify its rationale using anything other than ad hoc measures. While it could be an accurate analysis about market forces, its proposition would need to be re-formulated into a testable hypothesis in order to be considered as scientific literature. Chatrath et al. (1997) investigates whether or not hedgers and speculators exchange risk premia, using data on agricultural commodities from the COT report. Their findings show that although large speculators are profitable in commodity futures markets, they do not impose an instantaneous risk premia on hedgers. They argue that these findings support the efficient market hypothesis, in which the existence of a risk premium would suggest market inefficiencies.

3 Data and Feature Definitions

The collected data sets consist of monthly figures from January 1994 to December 2018. The number of observations varies between 279 to 299, depending on the model and the number of lags used. When no lags are used, all 299 observations are included. Response variables are prepared in the Kenneth French Data Library as monthly percentage changes, so no further manipulations are made here. The sentiment measures, however, are a mix of index prices and levels transformed into ratios. I use the log change on most of the features' time series, to get closer to a normal distribution on all variables. Now, let's conduct more comprehensive definitions of each variable.

Note that when I refer to a variable without any time subscript, e.g. $(x_j)_{j \in F}$, I am referring to the full time series - that is, all observations of the appropriate time series up to and including time t. When presenting the models, I include the time subscript to signal that observations

up to time t are used. For models testing causal relationships, only observations up to time $t-1$ are used for the right hand side variables, and hence, they will be denoted accordingly (e.g. $y_{i,t-1}$).

3.1 Response Variables

Following the reasoning of Lee et al. (1991) I expect that the closed-end fund discount is better at explaining the variance in small stock returns than big stock returns. On the other hand, one could hypothesize that due the aggregate nature of the COT features, they are better at explaining big stock returns. The same logic applies to the CBOE Volatility Index, which is directly derived from the total S&P 500 and not its subsets. Therefore, all regression routines will be performed on a range of stock portfolios to better capture their performance on different subsets of the New York Stock Exchange.

The Kenneth French Data Library (KFDL) contains portfolios formed on sets of characteristics such as size, operating income, investment levels etc. The chosen response variables in this paper are *6 Portfolios Formed on Size and Book-to-Market* from KFDL. These portfolios are constructed at the end of each June. They are the intersections of two portfolios formed on size of market equity (ME) and three portfolios formed on the ratio of book equity (BE) to market equity (BE/ME). The size breakpoint for year t is the median NYSE market size at the end of June of year t. BE/ME for June of year t is the book equity for the last fiscal year end in t-1 divided by ME for December of t-1. The BE/ME breakpoints are the 30th and 70th NYSE percentiles. For instance, a small sized stock portfolio with low book-to-market is denoted SMALLLoBM. To simplify the appearance of the regression models, the six portfolios' time series are instead denoted $(y_i)_{i \in I}$, where $I = \{1, ..., 6\}$, ranging from small size stocks with low BE/ME to big size with high BE/ME. The order corresponds to that of the descriptive statistics displayed in Table 1.

3.2 Sentiment Features

Closed-End Fund Discount: The difference between a publicly traded fund's price and the net asset value of this fund. The net asset value is calculated from the *fund's* balance sheets, which intuitively suggests that the discount should be close to zero. However, the funds may

have advantages that allow them to acquire shares at prices different from the common share trading prices. Moreover, closed-end funds are only traded secondhand - the funds do not buy back shares or issue new ones. Consequently, the supply of the fund's shares remains constant, while their demand may be subject to investor sentiment. Combining these effects result in price-to-nav deviations.

$$cefd_t = \frac{NAV_t - Trading\ Price_t}{NAV_t},$$

where

$$NAV_t = \frac{Assets_t - Liabilities_t}{Number\ of\ outstanding\ shares_t}.$$

CBOE Volatility Index: A market index that represents the market's expectation of 30-day forward-looking volatility. It is derived from the price inputs of the S&P 500 index options and provides a measure of market risk and investors' sentiment. Implied volatility is calculated by taking the market price of an option, entering it into the Black-Scholes formula, and back-solving for the value of the volatility.

$$vixret_t = \log\left(\frac{VIX_t}{VIX_{t-1}}\right).$$

Historical CFTC Commitment of Traders Report for Financial Futures:
The COT reports provide a breakdown of each Tuesday's open interest for futures and options on futures markets in which 20 or more traders hold positions equal to or above the reporting levels established by the Commodity Futures Trading Commission (CFTC).

The COT reports are based on position data supplied by reporting firms (e.g. FCMs, clearing members, foreign brokers and exchanges). While the position data is supplied by reporting firms, the actual trader category or classification is based on the predominant business purpose self-reported by traders on the CFTC Form 401 and is subject to review by CFTC staff for reasonableness. CFTC then aggregates the data and publishes it every Thursday. Thus, all the figures are of aggregate nature.

The selected features from this data are related to, and inspired by that of Sanders et al. (2004). The names are directly borrowed from that paper, as the intuition behind their construction

remains somewhat similar: I want to transform trader positioning into some measure of 'net pressure'. However, their actual definitions are not the same here, as I use the long-to-short ratio as a measure of pressure, whereas the cited paper uses spreads. This is partly done for experimental reasons, but also for the practical reason that the long-to-short ratio is non-negative, which simplifies calculating monthly changes.

Here are the COT feature definitions for this paper:

OI: Open Interest is the total of all S&P500 futures contracts entered into and not yet offset by a transaction, delivery, or by exercise. The aggregate of all long open interest is equal to the aggregate of all short open interest. Here, I use the log-change in Open Interest as an investor sentiment measure.

$$OI_t = \log\left(\frac{Open\ Interest_t}{Open\ Interest_{t-1}}\right).$$

CPNL: Commercials' Pressure Net Long. Here defined as the log-change in Commercials' long-to-short ratio (LS) on S&P500 futures. Commercial traders are defined by the CFTC as producers, merchants, processors and users of the physical commodity who manage their business risks with use of futures or option markets.

$$CPNL_t = \log\left(\frac{Commercial\ LS_t}{Commercial\ LS_{t-1}}\right).$$

NCPNL: Non-Commercials' Pressure Net Long. Here defined as the log-change in non-commercials' long-to-short ratio on S&P500 futures. Non-commercial traders are defined by the CFTC as professional money managers (e.g. CTAs, CPOs, and hedge funds) as well as a wide array of other non-commercial (speculative) traders.

$$NCPNL_t = \log\left(\frac{Non\text{-}commercial\ LS_t}{Non\text{-}commercial\ LS_{t-1}}\right).$$

TOTPNL: Total Reportables' Pressure Net Long. Here defined as the log-change in total reportables' long-to-short ratio on S&P500 futures: Total Reportables refers to the sum of futures contracts held by commercial and non-commercial traders registered by CFTC.

$$TOTPNL_t = \log\left(\frac{Total\ reportables\ LS_t}{Total\ reportables\ LS_{t-1}}\right).$$

NONPNL: Non-Reportables' Pressure Net Long. Here defined as the log-change in non-reportables' long-to-short ratio on S&P500 futures: The sum of non-reportable contracts should always offset the sum total reportables. Because of this duality, only NONPNL is used in the regressions.

$$NONPNL_t = \log\left(\frac{Non\text{-}reportables\ LS_t}{Non\text{-}reportables\ LS_{t-1}}\right).$$

Note again that to simplify the notation in the regression models, the sentiment features' full time series are denoted $(x_j)_{j\in F}$, where $F = \{1, 2, ..., 6\}$. The ordering of the features reflect the order in Table 1; feature x_1 refers to $cefd$, feature x_2 refers to $NONPNL$, and so on.

Figure 1 shows the full time series of each sentiment measure. Just by looking, the variance in the COT features seems to be considerably higher over the recent years. One may also suspect that stationarity in the closed-end fund discount is lacking and may pose an issue. This is addressed in the methodology.

Figure 1: Time series of all sentiment features

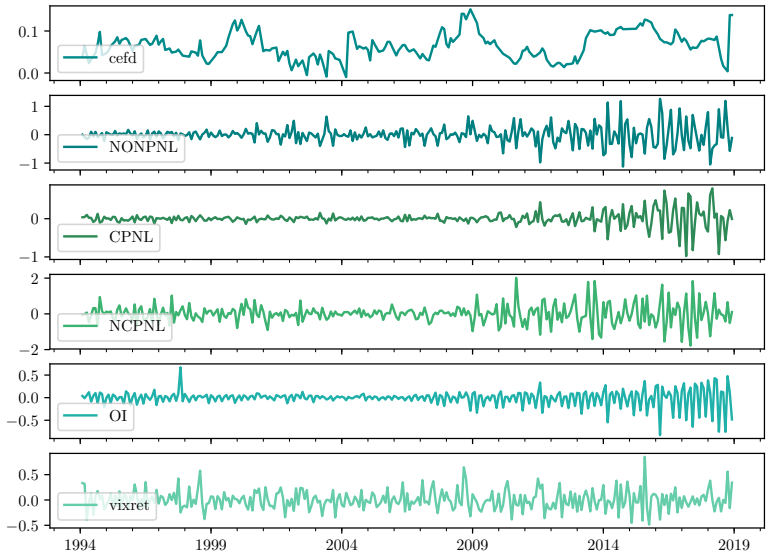


Table 1: Descriptive statistics: All variables

	SMALLLoBM	ME1BM2	SMALLHiBM	BIGLoBM	ME2BM2	BIGHiBM	mrp	cefd	NONPNL	CPNL	NCPNL	OI	vixret
obs	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.000
mean	0.007	0.009	0.010	0.008	0.006	0.006	0.006	0.064	0.000	0.004	0.006	-0.005	0.003
std	0.067	0.052	0.055	0.042	0.043	0.051	0.043	0.031	0.324	0.190	0.538	0.176	0.191
min	-0.245	-0.193	-0.206	-0.151	-0.181	-0.225	-0.172	-0.010	-1.124	-0.979	-1.776	-0.829	-0.486
25%	-0.034	-0.020	-0.018	-0.015	-0.018	-0.021	-0.020	0.042	-0.106	-0.050	-0.236	-0.058	-0.118
50%	0.011	0.012	0.010	0.012	0.011	0.013	0.012	0.061	0.009	0.001	-0.006	0.026	-0.004
75%	0.048	0.043	0.044	0.036	0.029	0.038	0.034	0.083	0.113	0.043	0.225	0.068	0.103
max	0.271	0.166	0.172	0.099	0.123	0.175	0.114	0.152	1.266	0.796	2.018	0.672	0.853
skew	-0.257	-0.477	-0.643	-0.603	-0.741	-0.772	-0.744	0.301	0.407	-0.546	0.274	-1.191	0.567
kurt	1.383	1.480	1.663	0.781	2.374	2.424	1.293	-0.233	3.140	8.375	2.179	5.080	1.398

The first six variables are the portfolio time series. The market risk premium, *mrp*, is not yet defined as a sentiment feature nor a response; it will be introduced when investigating causal relationships later. The variables' skewness and kurtosis indicate that they are not normally distributed. Most variables are left-skewed and have low kurtosis relative to the normal distribution.

4 Methodology

In this paper I want to investigate the relationship between a selected basket of investor sentiment measures and a small range of stock market portfolios. More specifically, I want to test the hypothesis that these features may have explanatory power as well as the hypothesis that there might exist causal relationships between the investor sentiment and portfolio returns. Moreover, I want to test *which* features best explain or cause changes in the portfolio returns.

The question regarding investors' rationality is, as previously mentioned, a hard one to answer. For instance, Warren Buffet states that the stock markets are rational most of the time, but he acknowledges that it has its occasional 'seizures' [2]. Investor sentiment is often confused with the irrationality that is usually associated with those periods of 'seizure'. This likely follows from the sense that sentiment is hard to measure and is deeply rooted in human behavior. Consequentially, one might think that sentiment has an insignificant role in explaining price movements over time, and that it should certainly not be able to forecast them. However, recall that an efficient market implies that prices reflect *all* current and past relevant information, which should include investor sentiment as well. Just because it is hard to measure or define does not mean that it no longer qualifies as information that prices should reflect. As for causality, recall from Dwyer Jr and Wallace (1992) that cointegration and arbitrage are not necessarily connected. For this reason, it is important to have a clear definition of what is actually being tested. I am not testing or trying to prove the inefficiency of the market. Instead, I simply look for evidence that these investor sentiment measures can both explain and cause price movements.

To do this, I first have to clear some definitions. The term 'prediction' is used a lot in statistical inference, but it does not have a universal definition. For this thesis, the term prediction will be a general reference to any predictions made from any model. Forecasts, on the other hand, will be used when referring to predictions made from lagged time series regressors exclusively. For example, a prediction from a vector autoregressive model would fall under the term 'forecast', since it uses lags of one or more time series to predict another, but a contemporaneous prediction from an ordinary least squares model would not. This distinction is important to avoid any confusion when the models are compared.

[2] Buffett, W. E. (2018, February 23). Letter to Shareholders of Berkshire Hathaway Inc.

The approach can be divided into 3 stages:
In Stage 1, Generalized Least Squares models are deployed to test the features' and their principal components' explanatory power. Stage 2 tests for causal relationships between the features/principal components and the responses using a vector autoregressive model. Stage 3 tests for causal relationships specifically in the feature $\rightarrow$ response direction, only this time using machine learning algorithms to test for causality - again using both features and their principal components as causing variables.

4.1 Univariate Generalized Least Squares

First, let's analyze how the features perform individually. To do this, I first perform a sequence of univariate Generalized Least Squares (GLS) regressions for each feature on each of the six portfolios. This is a two-step process where that starts with a classical Ordinary Least Squares (OLS) regression, to obtain information about the covariance structure of the residuals. Then the regression is transformed into a GLS. This is further explained below. Regressing all six features on all six portfolios implies a total of 36 regressions;

$$
\begin{aligned}
&y_{it} = \alpha_i + \beta_1 cefd_t + \varepsilon_{it} \quad , \\
&y_{it} = \alpha_i + \beta_1 NONPNL_t + \varepsilon_{it} \quad , \\
&\quad \vdots \\
&y_{it} = \alpha_i + \beta_1 vixret_t + \varepsilon_{it} \quad . \qquad (1)
\end{aligned}
$$

In traditional univariate OLS regressions we use the observed data $\{y_t, x_t\}$ and minimize the sum of squared residuals of $Y = X\beta + \varepsilon$. Then, so long as the Classical Linear Regression Model (CLRM) assumptions hold, we have that $\mathrm{E}[\varepsilon \mid X] = 0$, and $\mathrm{Cov}[\varepsilon \mid X] = \Omega = \sigma^2 I_n$. An important notice is that the constant variance assumption of CLRM implies that the covariance matrix Ω is diagonal and constant $\sigma^2 I$. The estimated coefficients are then unbiased, consistent and efficient. However, taking a quick glance at the features over time, some of them seem to have time-varying volatility, which can mean that the residuals too are heteroscedastic. While the coefficients will remain unbiased and consistent (i.e. the estimated coefficients reflect the true coefficients), they will not be efficient - their standard errors will likely be underestimated. Recall that the main reason for doing these univariate regressions is to paint a realistic picture

of the explanatory power of these features. A quick glance at Figure 1 gives reason to suspect heteroscedastic errors, and hence, I perform a White's test to confirm whether this is true or not.

I proceed to use a Generalized Least Squares (GLS) model to better capture the true significance of each feature. The optimization problem is;

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \, (Y - Xb)^{\intercal} \, \Omega^{-1} (Y - Xb)$$

The procedure begins by factoring the non-diagonal covariance matrix Ω such that $\Omega = CC^{\intercal}$, where $C = \Omega^{-\frac{1}{2}}$. Then the transformation $C^{-1}Y = C^{-1}X\beta + C^{-1}\varepsilon$ will assure constant variance in the residuals. If we let $\varepsilon^* = C^{-1}\varepsilon$ then it can be shown that $\operatorname{Var}[\varepsilon^* \mid X] = C^{-1}\Omega \left(C^{-1}\right)^{\intercal} = I$. The coefficients will remain the same, but will have slightly different standard errors. They are then estimated by $\hat{\beta} = \left(X^{\intercal}\Omega^{-1}X\right)^{-1} X^{\intercal}\Omega^{-1}Y$.
This shows that a GLS regression is equivalent to performing an OLS regression on a linearly transformed version of the data, which means that the interpretation of the results will remain the same.

4.2 Multivariate Generalized Least Squares

To examine how the features perform together, I proceed with multivariate GLS regressions. However, this immediately poses an issue: The features are correlated, which means that CLRM's linear independence assumption that $\Pr\big[\operatorname{rank}(X) = k_{\text{features}}\big] = 1$ is violated. Measuring joint significance is primarily for comparison with the machine learning approach later and the comparison will not make a lot of sense if the results are inconsistent. There are a few solutions to the multicollinearity issue, but most of them involve making changes to the data, i.e. either cutting out some features or changing the number of observations. Cutting out features would defeat the purpose of our feature selection goal, and increasing the number of observations is not feasible as collecting more data would require new channels and permissions that are not available. Thus, we can either ignore it, use a subset of the data, or transform the feature vectors so that they are orthogonal or linearly independent. This can be done by extracting their principal components and use them as features. Doing that will retain most of the information of the features and simultaneously assure linear independence. One reason

for conducting these multivariate regressions is to compare with the machine learning approach later, where I will use principal components anyway. For this reason, I will use the principal components here as well. However, I first ignore the collinearity issue and simply bundle all the features together to see how they perform jointly, and then repeat the multivariate regression using the first three principal components. This will also give an idea of how much the collinearity issue increases the estimated coefficients' standard errors.

This GLS regression routine is also performed on all $i \in I$ portfolios. Note that this routine consists of 12 GLS regressions;

$$\begin{aligned} y_{it} &= \alpha_i + \beta_1 cefd_t + \beta_2 NONPNL_t + \beta_3 CPNL_t + \beta_4 NCPNL_t + \beta_5 OI_t + \beta_6 vixret_t + \varepsilon_{it} \\ y_{it} &= \alpha_i + \beta_1 PC1_t + \beta_2 PC2_t + \beta_3 PC3_t + \varepsilon_{it} \end{aligned} \tag{2}$$

The significance of each coefficient is tested using F-tests. The null-hypothesis is that keeping all other parameters constant, adding the appropriate feature will, on average, not lead to a lower residual sum of squares. Formally,

$$F = \frac{\left(\frac{\mathrm{RSS}_R - \mathrm{RSS}_{UR}}{p_{UR} - p_R}\right)}{\left(\frac{\mathrm{RSS}_{UR}}{n - p_{UR}}\right)},$$

where $\mathrm{RSS}_i = (y_i - \hat{y}_i)^2$ is the residual sum of squares of the appropriate model obtained from the observed residuals. The degrees of freedom are determined by $p_{UR} - p_R$ and $n - p_{UR}$, where p is the number of parameters estimated in the model. The restricted model omits the feature to be tested, and hence, is nested within the unrestricted. Rejecting the null-hypothesis implies an F-statistic above the significance thresholds in the F-distribution.

4.2.1 Principal Components Analysis

Principal Component Analysis (PCA) is a method that is often used to reduce the dimensionality of large data sets. It transforms a set of variables into a smaller one such that it still contains most of the information of the large set. In python, both the Statsmodels module and the Sci-kit Learn (SKlearn) library can be used to perform PCA. For the multivariate

regressions, using the statsmodels.PCA function is straightforward and easy: I simply set it to standardize the variables and use the Singular Value Decomposition (SVD) method to obtain the principal components. The SKlearn approach is slightly different, though only because of the data-splitting. The formal procedure stays the same: First, the data is standardized such that all vectors have a zero mean and standard deviation equal to one. This is done to assure that the relative scaling of the vectors is uniform across all variables so that no variables weigh more than others when the principal components are obtained. Then, SVD is applied to transform the scaled data into principal components and their weights. The mathematical representation of the SVD procedure is as follows:

Let $\Sigma_{n\times p}$ be a diagonal matrix of positive numbers σ_k called the singular values of the feature matrix X. Singular values are equal to the square root of the eigenvalues λ_k of $X^{\intercal}X$. Now, let $U_{n\times n}$ and $W_{p\times p}$ be matrices where all columns are orthogonal vectors. The columns of U are called the left singular vectors of X, while the columns of W are the right singular vectors. U and W are obtained by computing the eigenvectors of $XX^{\intercal}$ and $X^{\intercal}X$, respectively. Then, it can be shown that $X = U\Sigma W^{\intercal}$. After SVD is performed, the principal components are obtained as the vectors of the score matrix $T = XW$.

Again, the only difference between Statsmodels' and Sklearn's approach is that in the SKlearn approach, X will be restricted to a subset of the entire sample of observations - the training set. Thus, the principal components obtained from the SVD will be slightly different from those of the Statsmodels approach, but in both cases, they will contain most of the variance in the features.

4.3 Granger Causality

The previous regressions test the explanatory power of the features where the time indexing is contemporary. Naturally, the data points in the previous regressions were denoted $\{y_i, x_j\}_{i\in I, j\in F}$, implying that I used all observations up to and including time t. Thus, I tested if and how the features *correlate* with the portfolios month by month, rather than test whether they *cause* the variation. Recall that one of the reasons for using investor sentiment measures as features is the hypothesis that they might also *affect* next month's portfolio returns. This is tested using Granger causality tests. The model used to test this is a vector autoregressive

model of order p, a VAR(p). The order refers to how many lag variables are used, and is selected by the Akaike Information Criterion (AIC). AIC is a model selection estimator that maximizes the model fit evaluated at the maximum likelihood estimates of the parameters, but penalizes for the number of lags to be used in the model (Akaike, 1974). Instead of selecting the number of lags to use in the model arbitrarily, it selects the number of lags subject to capturing most of the variation in the portfolio returns. The information criterion is defined as;

$$\begin{aligned} AIC &= 2k - 2\ln(\hat{L}), \\ \ln(L) &= -\left(\frac{n}{2}\right)(\ln|\Omega| - K\ln(2\pi) - K) \end{aligned}$$

where k is the number of estimated parameters, K the number of estimated equations, and Ω the average residual sum of squares RSS/n. $\hat{L}$ is the maximum value of the likelihood function for the model, obtained by replacing Ω with Ω_{MLE} - the average RSS with the maximum likelihood estimators.

In reality, a VAR process is atheoretical - all variables are treated as endogenous variables. The model does not distinguish between the 'dependent' responses and 'independent' features. Instead, it regards all variables, whether it be a response or a sentiment feature, as plain *variables*. For this reason, one can define a matrix X that includes *all* time series; that is responses $(y_i)_{i\in I} \subset X$ and features $(x_j)_{j\in F} \subset X$. This characteristic of VAR processes comes with a few caveats that I will discuss a little further. First, a VAR process is expressed as;

$$X_t = \sum_{q=1}^{p} A_q X_{t-q} + \varepsilon_t \tag{3}$$

where ε_t is a white Gaussian random vector, and A_q is a $n \times K$ matrix for every lag q. Feature $(x_j)_{j\in F}$, is said to Granger causes response $(y_i)_{i\in I}$ if the respective coefficients $a_{i,j} \in A_q$ for all lags $q = 1, \ldots, p$ are significantly different from zero. This is determined through F-tests as described in the multivariate least squares estimation. Recall from the definition of Granger causality that the universe of information should contain as much relevant information as possible. CAPM demonstrates that the market risk premium mrp contains most of the relevant information to explain stock returns, which justifies including it in our model. It is defined as

the return on a value-weight market portfolio minus the U.S. one month T-bill rate, commonly denoted $R_M - R_f$. The unrestricted model will include all variables in the universe of information, while the nested, restricted model will omit the feature to be tested. The null-hypothesis of the F-test is Granger non-causality; Adding the new feature does not significantly reduce the residual sum of squares.

Let x_0 be the *mrp* time series. For each $i \in I$, and each $j \in F$, a new VAR(p) is estimated;

$$\begin{bmatrix} y_{i,t} \\ x_{0,t} \\ x_{j,t} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} a^1_{1,1} & a^1_{1,2} & a^1_{1,3} \\ a^1_{2,1} & a^1_{2,2} & a^1_{2,3} \\ a^1_{3,1} & a^1_{3,2} & a^1_{3,3} \end{bmatrix} \begin{bmatrix} y_{i,t-1} \\ x_{0,t-1} \\ x_{j,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a^p_{1,1} & a^p_{1,2} & a^p_{1,3} \\ a^p_{2,1} & a^p_{2,2} & a^p_{2,3} \\ a^p_{3,1} & a^p_{3,2} & a^p_{3,3} \end{bmatrix} \begin{bmatrix} y_{i,t-p} \\ x_{0,t-p} \\ x_{j,t-p} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \\ e_{3,t} \end{bmatrix}.$$

The VAR(p) framework comes with a few preparation procedures. First, one has to make sure that all variables are stationary, or $I(0)$. If a variable is non-stationary, it would imply that the expected error term in the respective equation of the model is non-zero, which in turn means that the regression is spurious. I already tested and confirmed for stationarity in the differences prior to making any regressions using an Augmented Dickey-Fuller test. However, to determine whether it is more appropriate to use a Vector Error Correction Model (VECM) as opposed to a VAR model, I have to be certain that the variables are non-stationary in the levels. That is, if the variables were already stationary before differencing once, it would imply that there could exist long-run relationships as well as short-run relationships between the variables. Then, the process would resemble a VECM process more so than a VAR process, which only captures one of the two types of relationships. With this, I first transform the variables into levels and perform another Dicky-Fuller test of a unit root in each of the 'leveled' variables. After having confirmed that they are in fact non-stationary in the levels, I can proceed to deploy the VAR(p).

Executing this routine is done using the VAR model provided in the Statsmodels module. This model produces robust regressions; It is set up to test for heteroscedasticity and re-run the regressions so that the errors are $\sim N(0, \sigma_K^2)$. Looping through all features successively, then for each iteration, it will compute the optimal lag order according to AIC and run a new VAR(p). This module is also equipped with a Granger causality test submodule. I simply run the regressions and use their outputs as inputs for the causality test submodule, along with the instructions to test for the isolated effects of adding the sentiment features. It will then perform

an F-test to conclude whether the coefficients of the feature lags are jointly significant. Table 3 shows the p-values of these tests. Table 6 shows the corresponding results of the causality tests in the opposite direction. That is, whether the portfolio returns affect the sentiment features. In cases where $(x_j)_{j\in F} \rightarrow (y_i)_{i\in I}$ AND $(y_i)_{i\in I} \rightarrow (x_j)_{j\in F}$, there exists a feedback system, implying that both variables mutually affect each other. In theory, this would indicate a more complex relationship between feature and portfolio, and it would make it difficult to draw any steadfast conclusions.

5 Machine Learning Approach

In python, Statsmodels is generally the most frequently used module for traditional statistical analysis in finance, as it provides many outputs for in-depth analysis, such as summaries, T-stats, log-likelihood and so on. The GLS and VAR regressions earlier were all performed using Statsmodels classes, as the main goal was to illustrate the relationship between the features and portfolios. However, the downside of this approach is that results may be subject to sampling bias. In other words, the results were all obtained in retrospect, and one cannot be confident that the features would perform similarly if they had been analyzed in real-time and tested out-of-sample. Alternatively, one can use SKlearn, which contains a wide array of both supervised and unsupervised learning modules. All the models in the library are designed so that we fit the algorithms on a subset of the data and test them on a different, unseen subset. This partitioning is known as train-test splitting and is in the core design of SKlearn - it is better tailored for out-of-sample regressions than Statsmodels. Moreover, we can validate the models by cross-validation, which will re-split the data into new partitions and conduct the same regressions again, using different hyperparameters. This way, one can be more confident in how generalized the models really are (Bishop, 2006).

The selected models for this paper are OLS, Ridge, Lasso and Random Forest Regression (RFR). OLS, Ridge and Lasso are linear regression models, similar to GLS and VAR, while Random Forest is an ensemble learning algorithm that is non-parametric and also captures non-linear relationships. Unfortunately, there are no GLS nor VAR model algorithms in the SKlearn library (Pedregosa et al., 2011), and constructing them from scratch is beyond the aim of this thesis. For that reason, the existing OLS, Ridge and Lasso regression algorithms are

chosen to be compared with the VAR regressions. Recall that the main reasons for the machine learning approach here is; 1) to see how the features perform out-of-sample, and 2) investigate which features that contribute the most, using a wide array of models.

To isolate the effects of adding the features to our model, the strategy will make use of a nested model. That is, I will compare a partially restricted model with an unrestricted model. For both models, the design matrices will only contain lagged time series. Similar to VAR, I include as much relevant information as possible and then test whether adding the selected features to that universe will improve the predictions. Hence, the market risk premium is a predictor variable in both models. For the nested model, I restrict the universe of information to contain the lagged portfolio returns and the lagged market risk premium only. Then, the unrestricted model will include the investor sentiment features in addition to the lagged responses and market risk premium. Response matrix Y is defined such that it only contains the temporal series (not lagged) of the respective responses. Consequently, I am effectively performing two Autoregressive Distributed Lag models (ADL) of order p, where one is nested in the other. As in the VAR model, the lag order is computed by AIC. Lastly, I compare the prediction errors with and without the features. A more rigorous way to explain this procedure is this:

Let Y be the set containing the real-time portfolio returns so that the partitions $(Y_i)_{i\in I} \subset Y$ represent each portfolio time series. Let $\bar{U}$ be the universe of available and relevant information. The bar signals that the set only includes lagged information. Now let $\bar{X}'$ and $\bar{X}$ be subsets of that universe such that $\bar{X}' \subset \bar{X} \subseteq \bar{U}$. The partitions $(\bar{X}'_i)_{i\in I} \subseteq \bar{X}'$ only contain lags of the respective responses and lags of the market risk premium. More precisely, for any $i \in I$ and including all p' lags chosen by AIC, we have that $\bar{X}'_i = \{\bar{Y}_i, \overline{mrp}\}_{t-p'}^{t-1}$. Now, let $\bar{X}_{i,j} \subset \bar{X}$ be a superset of $\bar{X}'_i$ that also contains the lagged time series of the sentiment features $(\bar{X}_j)_{j\in F}$. Again, AIC will choose p lags to include in the model such that $\bar{X}_{i,j} = \{\bar{Y}_i, \overline{mrp}, \bar{X}_j\}_{t-p}^{t-1}$. Finally, I iterate through the portfolios and features as before and estimate the models;

$$Y_i = \bar{X}'_i\beta + \varepsilon_i \tag{4}$$

$$Y_i = \bar{X}_{i,j}\beta + \varepsilon_{i,j} \tag{5}$$

Comparing the errors in models (4) and (5) will provide an indication on causal effects in the feature $\rightarrow$ response direction. When testing the principal components, they are simply bundled

together such that $\bar{X}_{i,PC} = \{\bar{Y}_i, \overline{mrp}, \overline{PC1}, ..., \overline{PC3}\}_{t-1}$. Because they are linearly independent, there will be no collinearity issue between them. However, the components are correlated with the response lags and/or the market risk premium. This issue persists in all of the linear models conducted in the machine learning approach. Also, note that with the principal components, I only use one lag. This is to avoid overcomplicating the model - after some experimentation, including more than one lag did not improve the model.

Having explained the setup for all the linear models, a few more issues should be addressed: First, how do I compare the fit of the models? Traditionally, Granger causality tests use F-statistics to determine causality and R^2 to measure the fit of the model. However, this ADL model is not a direct translation of VAR, so I should consider a few important things: R^2 automatically increases with the number of regressors, and since I am comparing two models with different number of regressors, it will provide a poor measure for comparison. For this reason, adjusted R^2 would be better in measuring the causal effects. However, another issue with this approach has to do with the CLRM violations. When the expected prediction errors are not $IID(0, \sigma^2)$, the estimated coefficients will be biased (Keele and Kelly, 2006). This is unfortunately the case in some of the regressions. While the residuals are not significantly correlated, heteroscedasticity and collinearity between regressors are persisting issues. Thus, some coefficients will be biased and their standard errors overestimated. Since R^2 and F-tests are only accurate when the CLRM assumptions hold, AIC is instead used as the measure of how well the forecasts are improved. Similar to adjusted R^2, AIC measures the goodness of fit while also penalizing for the number of estimated parameters. Note that the actual value of AIC has no inherent meaning. Instead, it is a comparative measure of fit between models, where the model with the lowest AIC is always preferred. For model comparisons, AIC can be expressed as;

$$AIC = 2k - n\ln(RSS),$$

where k is the number of parameters estimated, n is the number of observations and RSS is the residual sum of squares of the forecast. The difference between AIC of the unrestricted model's forecast with that of the restricted model indicates whether the forecast improved or not.

5.1 OLS, Ridge, Lasso and Random Forest

I will now introduce the machine learning algorithms used for this approach, but first I will clear another definition. In this thesis, a 'learning' algorithm refers to one that does not use a closed-form solution to solve its optimization problem. For instance, OLS obtains a vector of regression coefficients exclusively by matrix inversion and multiplication - it does not need to 'learn' patterns within the data set. If, on the other hand, the algorithm uses an iterative optimization algorithm to minimize a cost function, then it is a learning algorithm. In Ridge and Lasso, the optimization problems are specified to be solved through gradient and coordinate descent, respectively. In fact, the SKlearn modules provide an array of solvers to choose from, depending on the model. For example, Ridge can use least squares, singular value decomposition or variations of stochastic and conjugate gradient descent. Since least squares and singular value decomposition are closed-form and hence not 'learning' algorithms, both are avoided here. When faced with multiple choices of this kind, grid-search cross-validation will choose the optimization algorithm that best minimizes the cost function. An explanation of grid-search cross-validation is found in the subsection below. There are a few variations of optimization through gradient descent, and since Ridge, Lasso and Random Forest all use some form of this, I will briefly explain gradient descent;

The goal is to choose the coefficient vector $\mathbf{w} = \{w_1, w_2, \ldots, w_k\}$ that minimizes a cost function. In this case, the cost function is the residual sum of squares $RSS(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} RSS_i(\mathbf{w})$. It first picks a random starting point for the coefficient vector, $\mathbf{w}^{(0)}$, and computes $RSS_{\mathbf{w}^{(0)}}$ accordingly. The iterative process will choose new vectors $\mathbf{w}^{(m)}$ so that $RSS(\mathbf{w})$ decreases. It does so by computing the gradient - that is, the partial derivatives of $RSS(\mathbf{w})$ with respect to each coefficient, and replacing these 'old' coefficients with new ones such that;

$$w_i^{\text{new}} \rightarrow w_i^{\text{old}} - \gamma \frac{\partial}{\partial w_i} RSS(\boldsymbol{w}). \tag{6}$$

Learning rate γ is a constant that helps determine the step size for the next guess of the coefficient, w_i^{new}. It is a hyperparameter that is set to some arbitrary small positive number, along with the maximum number of iterations to allow - naturally there will exist a trade-off between the learning rate and maximum iterations. When the step size is lower than the learning rate, the gradient is close enough to zero and the optimization algorithm stops - the optimized

coefficients are then used as the coefficient vector $\mathbf{w}$. This iterative learning procedure is one reason why I categorize Ridge, Lasso, and Random Forest as machine learning algorithms. Another reason is cross-validation, which I will come back to in the subsection below. Since all four models use either cross-validation and/or iterative learning, they all fall under the machine learning section.

SKLearn's *LinearRegression* algorithm corresponds to the traditional **OLS** regression. It takes no hyperparameters and simply fits a standard linear regression with coefficients $\mathbf{w} = (w_1, ..., w_p)$ to minimize the residual sum of squares between the observed targets in the data set and the targets predicted by the linear approximation. The mathematical representation of the optimizing problem is;

$$\min_{\mathbf{w}} \quad ||X\mathbf{w} - Y||^2. \tag{7}$$

This method is identical to the methodology in (1), except that here β is substituted for $\mathbf{w}$ to denote the coefficients. It is worth to mention that the algorithm does not provide any summary statistics except the residual sum of squares. This makes it difficult to perform a Cholesky decomposition of the covariance matrix, and consequentially I have to accept that some CLRM assumptions will be violated. This applies to the Ridge and Lasso regressions too.

Another recurring issue related to machine learning models is that fitting the regression to the training set and then predicting on unseen data runs the risk of *overfitting*. That is, the algorithm may not generalize the relationships between variables very well. Ridge and Lasso regressions address this overfitting issue by imposing a penalty on the size of the coefficients. By doing this, the estimated coefficients will have a slightly 'worse' fit on the training set, but at the advantage of generalizing better so that predictions on the test set may improve. This is otherwise known as regularization. The coefficients are obtained by minimizing a *penalized* residual sum of squares. **Ridge** regression uses L2 regularization, and takes on the following optimization problem;

$$\min_{\mathbf{w}} \quad ||X\mathbf{w} - Y||_2^2 + \alpha||\mathbf{w}||_2^2. \tag{8}$$

It resembles the traditional OLS optimization problem, but adds a regularization term. Here, the hyperparameter α is a non-negative number that reflects the strength of the regularization.

If α is zero, the coefficients are identical to that of a traditional OLS solution. It can be shown that as α increases, the coefficients asymptotically converge towards zero, resulting in a near-fully restricted model with horizontal predictions (Hastie et al., 2005, p. 73).

Lasso regression, on the other hand, uses L1 regularization;

$$\min_{\mathbf{w}} \quad \frac{1}{2n}||X\mathbf{w} - Y||_2^2 + \alpha||\mathbf{w}||_1. \tag{9}$$

Similar to Ridge, α is a hyperparameter, a non-negative number that pushes the optimal coefficients towards zero as it increases towards infinity. An important distinction from Ridge is that Lasso allows the coefficients to be zero. As a result, Lasso has feature selection embedded in its algorithm, whereas Ridge will never completely neglect any features, no matter how irrelevant they are. This follows from the type of norm they use. In (8) and (9) the subscripts denote what type of norm the respective regularization methods use. L2 regularization solves the optimization problem with respect to a *Euclidean distance*, while L1 regularization minimizes with respect to a *Manhattan distance*. The difference between these norms is mathematically expressed;

$$\begin{aligned} ||\mathbf{w}||_1 = &\quad ||\mathbf{w}_1|| + ||\mathbf{w}_2|| + \cdots + ||\mathbf{w}_N||, \\ ||\mathbf{w}||_2 = &\quad \sqrt{||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + \cdots + ||\mathbf{w}_N||^2}. \end{aligned}$$

It can be shown that it is due to this difference in the norms that makes L1 regularization a tool for feature selection, whereas L2 regularization will only provide an indication for which features are less important (Hastie et al., 2005, p. 73).

Until now, all of the machine learning have been linear models. However, there could exist non-linear relationships between these investor sentiment features and portfolio returns as well. **Random forest** regression captures non-linearities through decision trees and bagging. The following is an explanation for the algorithm as provided in (Hastie et al., 2005, p. 588).

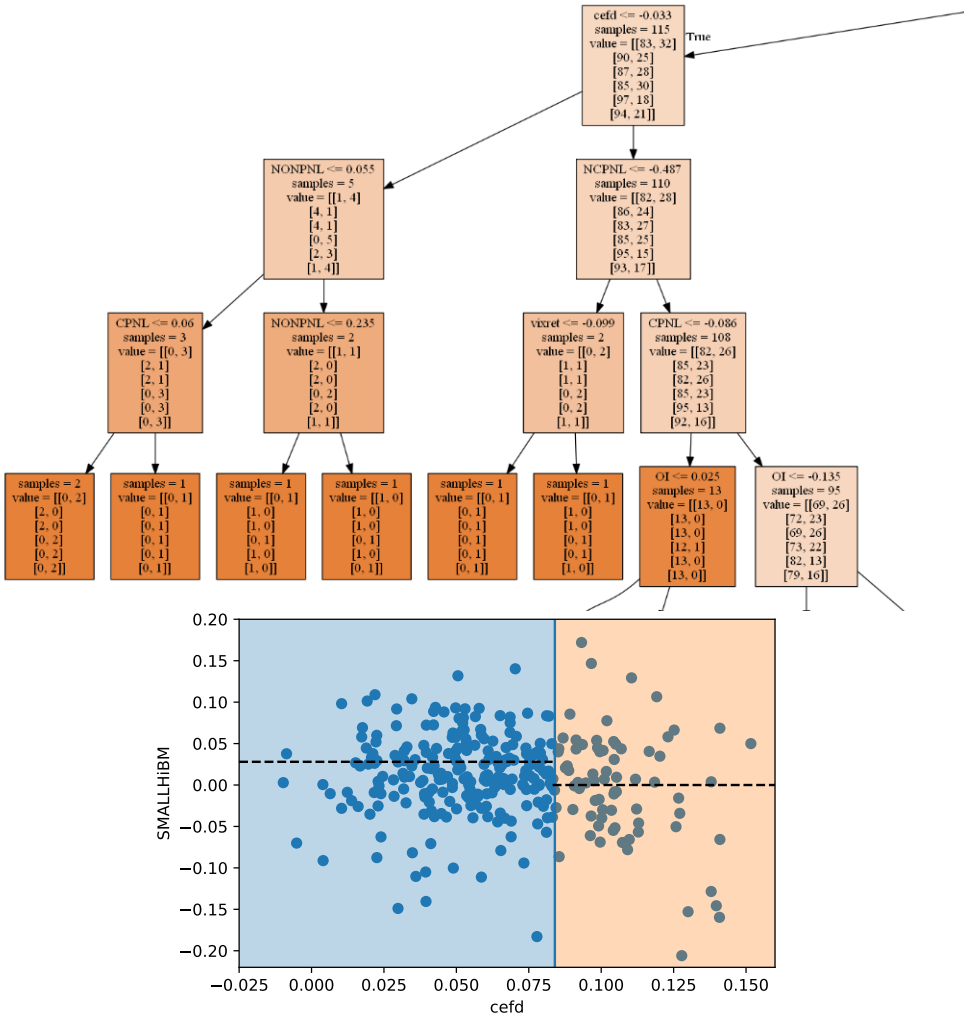
The algorithm first divides the predictor space - the set of possible values for all time series, into k distinct and non-overlapping regions $\{R_1, R_2, ..., R_k\}$. The regions are split using a top-down, greedy approach known as *recursive binary splitting*. Figure 2 below is a graphical example of

this method. In order to perform recursive binary splitting, it first selects a feature X_j and a cutpoint s, and splits the predictor space into the regions $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$ so to achieve the greatest possible reduction in residual sum of squares. That is, for each feature $(X_j)_{j \in F}$ and all possible values for the cutpoint s, it chooses the feature and cutpoint such that the resulting tree's predictions have the lowest residual sum of squares. In greater detail, for any j and s, it defines the pair of half-planes $R_1(j,s) = \{X|X_j < s\}$ and $R_2(j,s) = \{X|X_j \geq s\}$, and seek the values of j and s that minimize the sum;

$$\sum_{i \in R_1} (Y_i - \hat{Y}_{iR_1})^2 + \sum_{i \in R_2} (Y_i - \hat{Y}_{iR_2})^2,$$

where $\hat{Y}_{iR_k}$ is the mean of the response in region $R_k(j,s)$. Next, the process repeats, picking the best feature and best cutpoint in order to split the data further, always seeking to minimize the RSS within each of the resulting regions. However now, instead of splitting the entire predictor space, it splits one of the two previously identified regions. Now there are three regions. The process continues until a stopping criterion is reached; For example, it may continue until no region contains more than five observations. Then we have a fully grown decision tree $T(X;\theta_b)$, where θ_b is the set of hyperparameters for decision tree b. This way of splitting the observations in the predictor space, by seeking 'clusters' of observations, is exactly what captures non-linearities - Instead of fitting a straight line through the predictor space, it partitions it and fits a new straight line within each cluster it finds.

Figure 2: Graphical Illustration of Recursive Binary Splitting



The first illustration is an example of a decision tree $T(cefd; \theta_b)$. Here, the first optimal cutpoint is $cefd \leq -0.033$. Then, the tree is split further into two new branches using features $NONPNL$ and $NCPNL$. The second illustration is an example of the first step in recursive binary splitting. Here, the regions are split such that $R_1(j,s) = \{X|cefd \leq 0.084\}$ and $R_2(j,s) = \{X|cefd > 0.084\}$. The corresponding regression forecast $\widehat{\text{SMALLHiBM}}_{R_k}$ is the average of observations within each region, displayed by the two stapled lines. This figure is just an illustration and does not reflect an actual decision tree or region-split made in the forest.

The forest is set up to produce $B = 1000$ trees. This is the only hyperparameter that stays constant throughout the cross-validation. All other hyperparameters, e.g. minimum samples per split, mininum number of samples per leaf, or maximum number of features, are fed to grid-search cross-validation in ranges of alternatives. It will then 'tune' the estimators and pick the combination that minimizes the squared errors. Again, this is explained in greater detail

in the next section.

Since random forest regression is non-parametric, the feature selection mechanism is not determined by whether or not coefficients are zero. Instead, it outputs feature importances: At each split in each tree, the improvement in the split criterion, the decrease in RSS is saved. It is a measure of importance, that is rewarded to the current feature. For each feature, the improvements are then accumulated over all trees in the forest, and transformed into weights called *relative feature importance*. For this reason, instead of using the nested model approach as described in Equations (4) and (5), it makes more sense to simply feed all the features to the model. One advantage of random forest is that collinearity does not violate any assumptions. Thus, bundling the features together is feasible and will not incur any issues.

For the implementation, I want to stay consistent with the methodology of comparing the forecast errors of two models where one is nested within the other. Therefore, I re-define the models slightly:

The setup will remain similar except that only one lag is used for both the restricted and unrestricted model matrices. This mean that for the restricted model, the model matrix $\bar{X'}_i = \{\bar{Y}_i, \overline{mrp}\}_{t-1}$. When testing the sentiment features, the unrestricted model matrix is $\bar{X}_i = \{\bar{Y}_i, \overline{mrp}, \overline{cefd}, \ldots, \overline{vixret}\}_{t-1}$. For the principal components, the unrestricted model matrix is $\bar{X}_{i,PC} = \{\bar{Y}_i, \overline{mrp}, \overline{PC1}, \ldots, \overline{PC3}\}_{t-1}$. Then, I perform the random forest regressions for both predictor spaces and compare the forecasts;

$$\hat{f}(\bar{X'}_i) = \frac{1}{B}\sum_{b=1}^{B} T(\bar{X'}_i; \theta_b^{'}), \tag{10}$$

$$\hat{f}(\bar{X}_i) = \frac{1}{B}\sum_{b=1}^{B} T(\bar{X}_i; \theta_b), \tag{11}$$

$$\hat{f}(\bar{X}_{i,PC}) = \frac{1}{B}\sum_{b=1}^{B} T(\bar{X}_{i,PC}; \theta_b^{PC}). \tag{12}$$

Since random forest regression is non-parametric, there are no coefficients to be obtained and penalized for in AIC. For this reason, the forecasts are measured by the mean squared error. However, this does not really capture how well each feature 'incrementally' contributes to improving the forecasts. I therefore made an attempt at constructing a pseudo-measure for

AIC by penalizing for the size of the forest. In theory, this would allow for better comparison with the other models, but the lack of literature on the matter makes this measure hard to justify. Thus, the feature selection is evaluated based on feature importance. Figure 6 compares the fifth split model forecasts (10) and (12) relative to the full unrestricted model forecast (11). Figure 10 and Table 7 shows the fifth split relative feature importances.

5.2 Training, Cross-validation and Testing

Now that the models are presented, how do I deploy them? How should I split the data and implement the models?

To my understanding, it seems that most of the machine learning literature deals with classification rather than time series regression. Classification problems and cross-sectional analysis allow for splitting methods where the data is randomly shuffled before train and test indices are split. However, for time series regressions, it wouldn't make a lot of sense to train the model on sets without respect to sequentiality. For example, suppose we have a small training set containing only observations on February 1994, December 2001, and March 2016. Now, let our test set contain observations on March 2005 and October 2007. If we fit the model to the training set, we would train it using observations many years ahead of the observations in the test set. This is not feasible in real life, so I should therefore use a different approach. The TimeSeriesSplit function provides train/test indices to split time series data samples that are observed at fixed time intervals. In each split, test indices must be higher (i.e. more recent) than before. It also uses an expanding window, implying that the last split contains most observations. After experimenting with different numbers of splits and strategies, the best results were obtained using grid-search cross-validation for hyperparameter tuning and then test on a different subset. I will explain this a little further:

First, I set TimeSeriesSplit to make five splits on the entire data set. Each split contains a train subset and a complementary test subset. Suppose a train set contains observations 50 to 150, and that it is complemented with a test set containing observations 151 to 180. Let's denote these observations T^1_{train} and T^1_{test}, where the superscript denotes that this is one of five splits. The remaining observations are not included in this particular split. To pick the best combination of hyperparameters, grid-search cross-validation will then further split the train

set T^1_{train} into three new train/test pairs $T^1_{train,k}, T^1_{test,k} \subset T^1_{train}$, where $k \in \{1,2,3\}$. It will then pick a combination of hyperparameters, perform the regressions on these new train/test pairs, and store the combination that lead to the lowest prediction errors across the three train/test pairs. This combination is then used to make forecast on the unseen test set T^1_{test}. That is, a forecast of observations 150 to 180. This procedure is repeated for all five T^m_{train}, T^m_{test}, $m \in \{1,2,...,5\}$.

Cross-validation is computationally challenging; five splits, six responses and six features amounts to 180 regressions when testing the individual features. The joint tests using bundled principal components or features amount to 30 regressions for each model.

6 Results

Due to the large number of regressions, interpreting the results one by one will not be done here. Instead, summaries of the results are shown. Some tables and figures are shown here and discussed more in detail, while others are located in the appendices. For this reason, be well aware of the table references.

Table 2: Univariate GLS

	cefd	NONPNL	CPNL	NCPNL	OI	vixret
SMALLLoBM	-0.1999	0.0148	-0.0393*	0.0119*	-0.0367*	-0.1897***
ME1BM2	-0.1772*	0.0124	-0.0347**	0.0124**	-0.0260	-0.1560***
SMALLHiBM	-0.2300**	0.0112	-0.0337**	0.0129***	-0.0246	-0.1466***
BIGLoBM	-0.1357*	0.0106	-0.0234*	0.0074*	-0.0001	-0.1414***
ME2BM2	-0.2189***	0.0137*	-0.0310***	0.0106***	-0.0078	-0.1358***
BIGHiBM	-0.1798*	0.0162*	-0.0393***	0.0137***	-0.0128	-0.1381***
* $\Rightarrow \alpha = 10\%$	** $\Rightarrow \alpha = 5\%$	*** $\Rightarrow \alpha = 1\%$				

The table is a consolidated table of coefficients from the 36 univariate regressions. Each element is the estimated predictor coefficient of the associated regression. For convenience, the intercepts are not shown here. The asterisks show the significance level of the respective T-tests, with null-hypothesis $\beta_k = 0$.

Starting with the univariate GLS in Table 2: The coefficients of each regression tell us about

the directional relationship between each feature and the portfolios. The asterisks denote the level of significance each estimated coefficient has, on average. A positive coefficient indicates that the feature varies *with* the portfolio returns, while a negative one indicates that the feature varies *against* the portfolio returns. For instance, the negative coefficients for $cefd$ indicates that an *increase* in the closed-end fund discounts, the portfolio returns *decrease* on average. This is consistent with Lee et al. (1991), and strengthens the hypothesis that closed-end fund discounts widen when the stock market is performing poorly and vice versa. However, note that the coefficients for $cefd$ seem to be slightly more significant for large stocks than for small stocks, which is not what the paper shows. One possible explanation for that has to do with the definition of $cefd$. Here, it is an equal-weight average of all closed-end fund discounts in the database, as opposed to using market sizes as weights like they do in the paper. Thus, the discount as defined here may not capture the proper size-corresponding relationships. Another take-away from this figure is the obvious power of *vixret* - the change in 'the fear guage' seems to be the best in explaining the variation of relatively risky equity-only portfolios. It suggests that when the VIX Index is increasing, risk aversion turns investors away from the stock market and position themselves in other markets - hence the strong, negative coefficients. As for the COT features, the commercial and non-commercial traders' aggregate positioning both seem to be significant in explaining the variation in the portfolio returns. As expected, the commercials, or 'hedgers', increase their net long position when the stock market is falling, whereas the non-commercials, e.g. leveraged mutual funds, increase their net long position when the market is rising. This makes sense as hedgers seek to minimize their exposure, thereby entering a contrarian position in the futures market. "Long-only" managers, on the other hand, seek profits and do so by allocating funds pro-cyclically. Moreover, the results show that trader positioning better explains the change in stocks with high book-to-market ratios than low book-to-market ratios. High book-to-market stocks seem to be more sensitive to market sentiment. A possible reason for this is that these stocks, also known as value stocks, are companies that tend to have low earnings on book equity relative to low book-to-market growth stocks. This is consistent with the findings in Fama and French (1995). High book-to-market stocks are typically more distressed relative to growth stocks. Although uncertainty towards future earnings is commonly associated with growth stocks, that may be a result of idiosyncratic factors rather than systematic, market-wide uncertainties. Combining this with

the fact that the COT features are based on aggregate data on traders' futures positioning in the S&P 500, it makes intuitive sense that they better predict high book-to-market stocks than low book-to-market stocks.

The results are not as clear to interpret when it comes to the joint regressions (Table 4 in Appendix A). As mentioned, the features are correlated, as shown in the heatmap, Figure 7. An important issue with multicollinearity is that the standard errors of the estimated coefficients will likely be overestimated, which exposes the model to make Type 2 errors (Brooks, 2019, p. 170-175). Thus, the coefficients may come out as insignificant when in fact they are not. This is likely what is seen when comparing the significance levels in Tables 2 and 4. A consequence of this issue is that the portfolio-average R^2_{adj} of 34.6% measure is most likely misleading. Using the features' principal components instead fully removes the collinearity issue and nearly doubles the portfolio-average R^2_{adj} to 68.1%, despite having used only three predictor variables. Note that here, unlike Table 2, the significance of a feature is based on an F-test as described in the methodology. F-tests of this kind are also used to determine Granger causality between a feature and a portfolio, which is discussed below.

Table 3: Granger Causality Test Results: Feature → Response

	SMALLLoBM	ME1BM2	SMALLHiBM	BIGLoBM	ME2BM2	BIGHiBM
cefd	0.3514	0.2841	0.218	0.1911	0.101	0.1961
NONPNL	0.4709	0.3731	0.0138**	0.1506	0.4355	0.1705
CPNL	0.3873	0.1705	0.0812*	0.682	0.2853	0.0953*
NCPNL	0.5308	0.5911	0.2585	0.1852	0.3298	0.2232
OI	0.7423	0.5092	0.0326**	0.266	0.1238	0.7043
vixret	0.4707	0.244	0.2753	0.5519	0.638	0.9487

Rows and columns represent causing and caused variables, respectively. Each element is the p-value from the corresponding F-test, where the null-hypothesis is Granger non-causality.

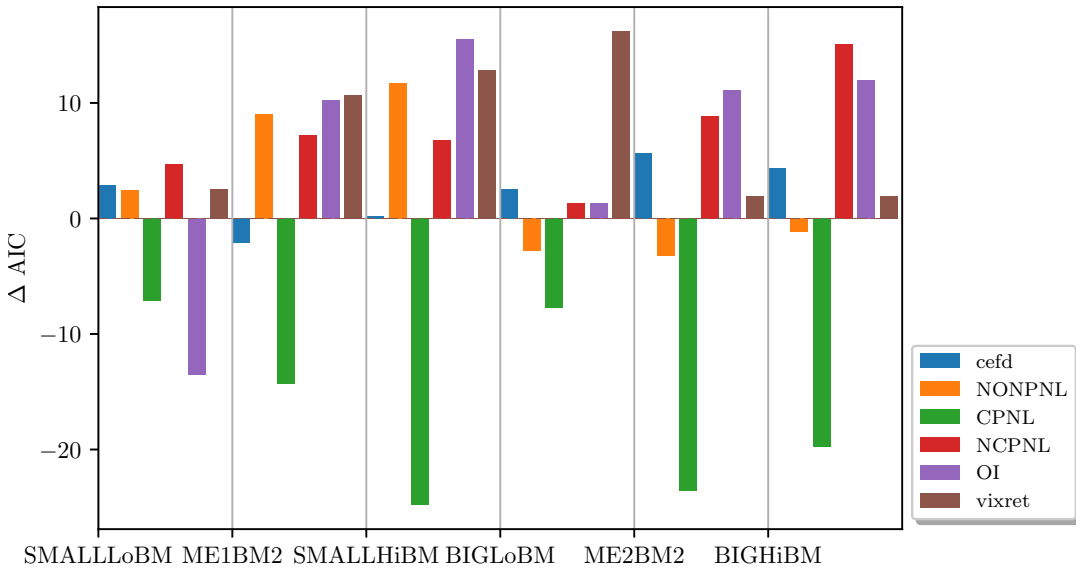
Interpreting the coefficients in a VAR model is difficult, since the model is dynamic and all variables are endogenous (Johansen, 2005). To avoid any spurious claims about these relationships, I will only conclude on whether a causal relationship is found or not. The VAR results in Table 3 show that causal relationships between the sentiment features and portfolio returns do

exist, but the evidence is underwhelming. The table shows p-values of the causality tests in the feature $\rightarrow$ response direction. Of the 36 causality tests, only four tests reject the null-hypothesis of Granger non-causality, where only two of those are significant on a 5% level. However, it is worth to mention that all four are COT features, and that they seem to only affect stocks with a high book-to-market ratio. This suggests that these features affect high book-to-market stocks in particular. Again, this could be because lower earning, distressed stocks are more exposed to market-wide uncertainties and investor sentiment relative to low book-to-market stocks. As for causality in the response $\rightarrow$ feature direction, Table 6 shows evidence of five causal relationships. None of which suggest a feedback system. That is, there does not seem to exist a bi-directional causal relationship between any of these features and portfolios.

For the machine learning approach, we will not have any 'yes' or 'no' answers on whether the results show evidence of causality. Even though I do perform F-tests on the coefficients, the violations of important CLRM assumptions make any strong claims questionable. Instead, comparing AICs of the forecasts with and without the sentiment features will provide an indication of causality: Recall from the definition of Granger causality that if one can *better* forecast the portfolio returns using the lagged features rather than omitting them, then they Granger cause the portfolio returns. Since AIC rewards goodness of fit while penalizing for the number of parameters, it will serve as a good indication of whether there exists a causal relationship.

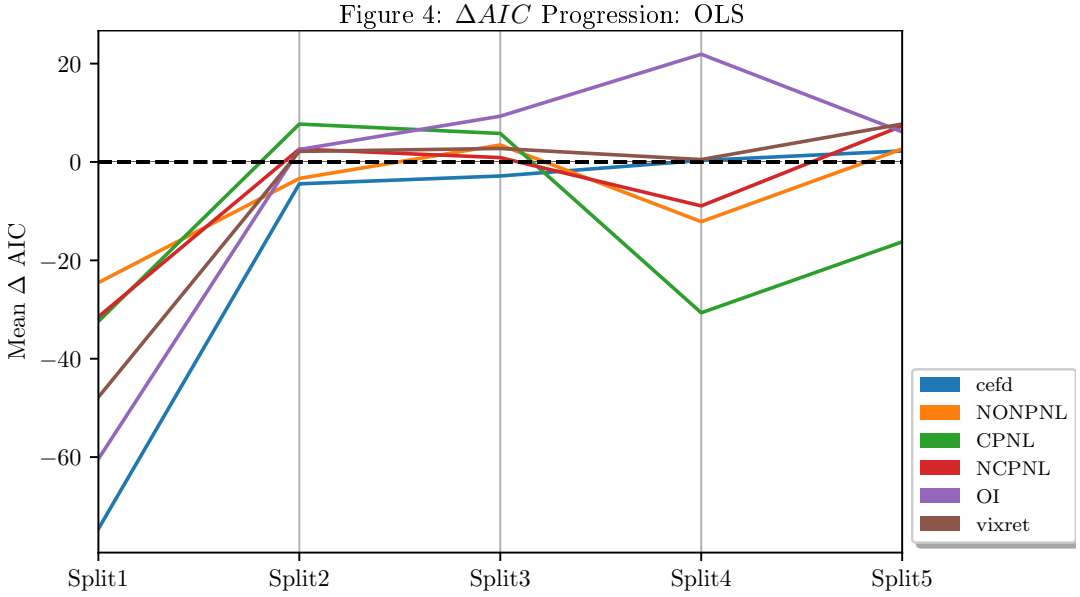
Figures 3 and 5 display the difference in AIC values for the fifth split's forecasts. Each bar represents the improvement or deterioration of the forecast as a result of including the respective feature as a predictor variable. Mathematically, each bar is $\Delta AIC = AIC_{UR} - AIC_R$. Recall that a lower AIC value indicates a better forecast, so a negative value here indicates that the forecast improved when the given feature was included. Comparing AICs from the traditional VAR model in section 4.3 with those in section 5, the ladder results are slightly worse. There could be several reasons for this, but the first obvious one is that the AICs in section 5 are computed on unseen test sets, whereas the AIC values from the VAR are computed on the 'training' data itself. Another reason for this discrepancy could be due to the lack of robustness in the residuals. In some of the regressions, they are not normally distributed with mean zero and constant variance.

Figure 3: ΔAIC: OLS



Each bar represents ΔAIC of the fifth-split forecasts. A negative value indicates an improvement of the forecast.

Figure 4: ΔAIC Progression: OLS



Progression of ΔAIC over five splits, on average across portfolios. A negative difference indicates that the feature improved the forecast of that split, on average across portfolios. Note, this shows the progression in the OLS model, but similar patterns are found in all models.

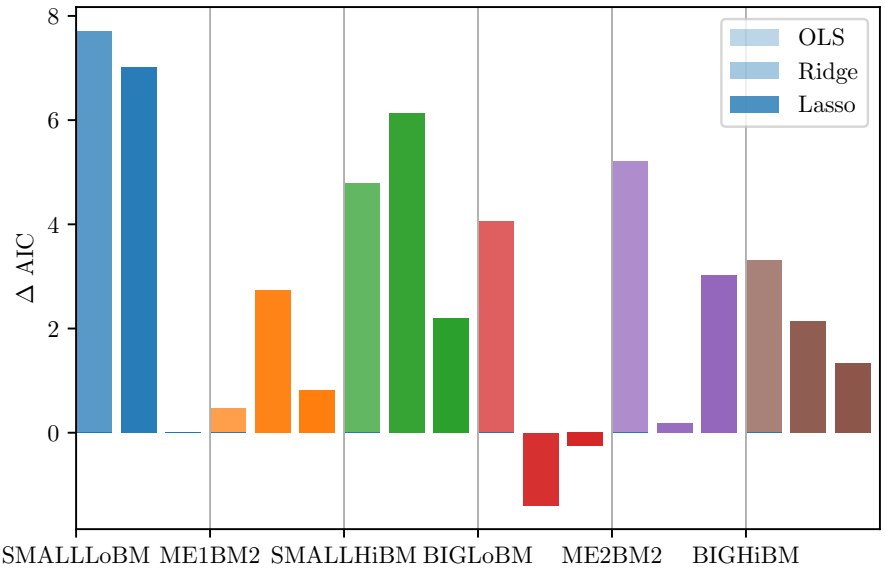
Figure 4 is only intended to serve as an illustration of how cross-validation works. It displays the forecast improvements over all five splits in the OLS model. To simplify the interpretation,

each line is the average of ΔAIC across the portfolios. This shows that in the former splits, which have fewer observations, the model always prefers to include the additional information embedded in the features, relative to omitting it. However, as the number of observations increases, the conclusion is not as unilateral. The last split is thus what the cross-validation results converge to. Hence, I will discuss those results more in detail; that is, the results of the last split.

There is an emerging pattern in Figure 3. The third feature, $CPNL$, seems to improve AIC considerably on all forecasts of portfolio returns. For the OLS model, AIC improves by 38.82% on average across portfolios. The fifth feature, OI, improves the forecast on small stocks with a low book-to-market ratio by 21.64%, but does not improve the forecasts on any other portfolio. The first feature, $NONPNL$, slightly improves the forecasts of all the large stocks portfolios, though only by an average of 4.89%. These results are consistent with the results from the traditional Granger causality tests, where $NONPNL$, OI and $CPNL$ were the only features whose test results rejected the null-hypothesis of Granger non-causality. Comparing with Figures 8 and 9, one can observe some model-dependent variations. The biggest difference is when including OI. In Lasso, it seems to improve more portfolio forecasts than for OLS and Ridge. This is because the algorithm selects fewer lags, which in turn means fewer parameters to penalize for. Using OLS, the feature only seemed to affect the small, low book-to-market portfolio. Using Lasso, the feature improved four out of six portfolio forecasts. On the other hand, $CPNL$, which improved all of the forecasts considerably in the OLS and Ridge, does not seem to perform as well in the Lasso regressions. Note that the Lasso algorithm does not take into account *which* variables to drop: It simply picks those that minimize L1 errors; In this case, it most likely picked the most significant lags, and not all of the lags. This could complicate the interpretation of causality; Granger (1969) does not mention 'picking and choosing' lags of this kind - the Lasso method did not exist at the time [3].

[3] A Lasso Granger method has been introduced in recent years (Arnold et al., 2007), but the lack of supporting literature makes it difficult to conclude whether this actually suggests Granger causality

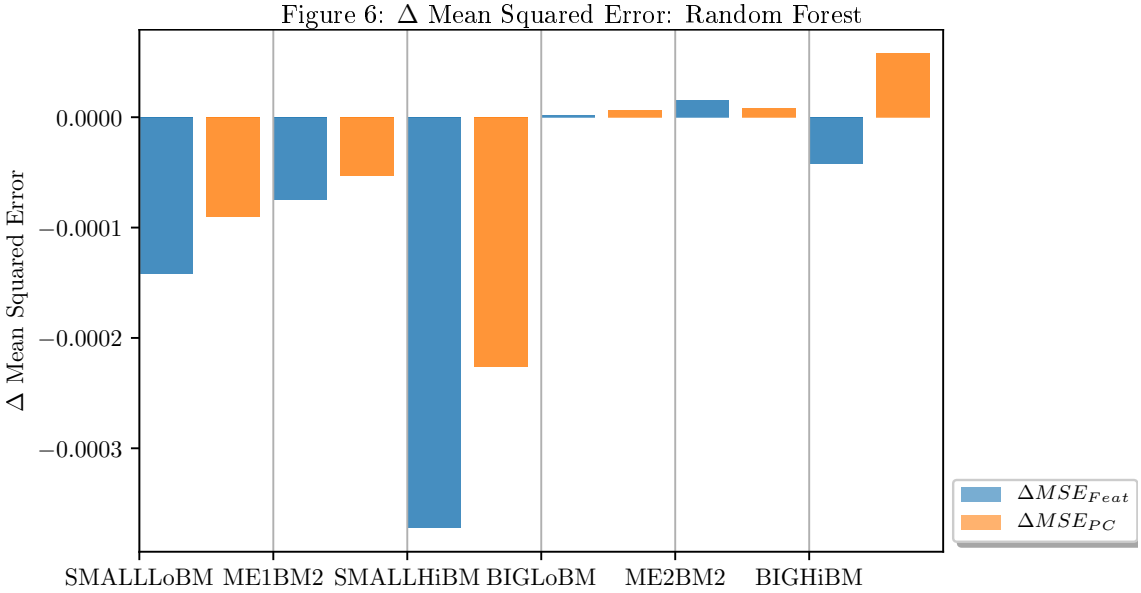
Figure 5: ΔAIC with PCs: OLS, Ridge and Lasso



For each portfolio, the light shade represents ΔAIC for the OLS model, and the darkest shade is for the Lasso model. Each bar represents ΔAIC of the fifth-split forecasts. A negative value indicates an improvement of the forecast.

Improvements using the features' principal components also seems to be dependent on the model used. Overall, when considered jointly, the components do not seem to improve the forecasts. Figure 5 shows that the components only improves the big, low book-to-market portfolio forecast when using the Ridge and Lasso models. However, the magnitude of this improvement is very low, at 2.89%. In the random forest, the PCs additionally improve forecasts on the small, high book-to-market portfolio. The differences between the models make it difficult to draw a conclusion on whether they affect the portfolio returns. However, this could also be evidence of the random forest capturing non-linearities that the linear models cannot capture. Comparing Figures 5 and 6, the PCs did not improve the forecast of the small high book-to-market portfolio in any of the linear models, but considerably did so in the random forest.

Figure 6: Δ Mean Squared Error: Random Forest



Each bar represents ΔMSE of the fifth-split forecasts. A negative value indicates an improvement of the forecast.

Figure 6 shows that for the small size portfolios, adding the features decreases the mean squared error of the forecasts. On average across all portfolios, the features contribute to a 6.92% improvement. In particular, for the small stocks high book-to-market portfolio, the improvement is a notable 18.3%. The mean feature importances across portfolios are displayed in Figure 10. The forest was not able to distinguish the relative importance of each feature very well. This interpretation comes from the small differences in relative importance, which are close to the mean of 12.49% relative importance. Regardless, the forest still seems to attribute more importance to $CPNL$ and $PC2$ than it does the responses' own lags, indicating that including the sentiment features into the universe of information increases the performance of the forest regression. Table 7 shows the relative importances on each portfolios. Again, the responses' own lags, here indicated by $response.L1$, seems to be less prioritized consistently across portfolios. In fact it only comes up as the most important feature on the small, low book-to-market portfolio. From this, we can also see that the market risk premium is consistently a strong contributor to the forest. This could have a 'damping' effect on the features' additional information. However, including as much information as possible is an key assumption to Granger causality and cannot be ignored.

6.1 Conclusion

This thesis investigated both linear and non-linear relationships between a selected set of investor sentiment features and a range of stock portfolio returns. While some of the chosen models are fairly straightforward to implement, the models using machine learning algorithms had to be partly constructed. This discrepancy in design may have suppressed the goal of achieving one-to-one comparisons between the traditional models and the machine learning models. Regardless, the results from the machine learning approach do stay relatively consistent with their traditional counterparts.

The results show that when it comes to explaining the variation of stock portfolio returns, four features are particularly significant: $vixret$, $NCPNL$, $CPNL$ and $cefd$. For a 95% confidence interval or higher, all of their coefficients are significant in explaining changes in small stocks with high book-to-market ratios. The first three are also significant on a 99% confidence interval for large stocks with high book-to-market ratio. When considered jointly, the collinearity issue seems to be too strong to draw any good conclusion. However, what *can* be deduced from the multivariate regressions is that the features' first three principal components are highly significant in explaining the variation in all of the chosen portfolios. This should provide sufficient evidence to state that there does indeed exist a contemporaneous linear relationship between some of these investor sentiment features and stock portfolios.

As for causal relationships, the evidence indicates that some COT features indeed affect the following portfolio returns. From the causality test conclusions in the traditional VAR(p) model $NONPNL$, $CPNL$ and OI show significant F-statistics. This suggests that these features do indeed Granger cause some returns, particularly on stocks with high book-to-market equity. In the machine learning approach, the same features show some, or even substantial improvements in forecasting portfolio returns, relative to omitting them. In particular, commercial traders' positioning on the futures market (read: hedgers' positioning) consistently seems to affect stock portfolio returns. As for the features' three principal components, when considered jointly, the results highly depend on the model used. In the linear models, overall, they do not seem to improve the portfolio forecasts - except for a slight improvement in the large stocks, low book-to-market portfolio. The random forest conducted multivariate regressions only, and possibly picked up a non-linear relationship between the features and one portfolio. In particular,

comparing the forecast improvements between models, the forest seems to indicate a possible non-linear causal relationship between the principal components and the small, high book-to-market portfolio. The linear models did not show an improvement in the forecast of this portfolio, while the random forest did. In the random forest, the features jointly improved the forecasts of all except one portfolio.

For future research on the topic, the first obstacle would be to deal with the issue regarding the robustness of the causal regressions: This is a known issue with dynamic models, which are highly dependent on the quality of the data - one violation of any assumption quickly turns into a spurious regression. One suggestion to deal with this issue is data preprocessing. One might avoid this issue all together if one can turn these investor sentiment measures into linearly independent features with time-invariant volatility. For example, one could rather focus on using the principal components as features to be tested individually as well as jointly. The downside to this is of course the interpretation of the results - they might lose economic intuition. Another suggestion is to use a whole new set of investor sentiment measures all together, although the approach should provide reliable results regardless. As long as the data quality is good, this approach should paint a good picture of both contemporaneous and causal relationships.

Appendices

A

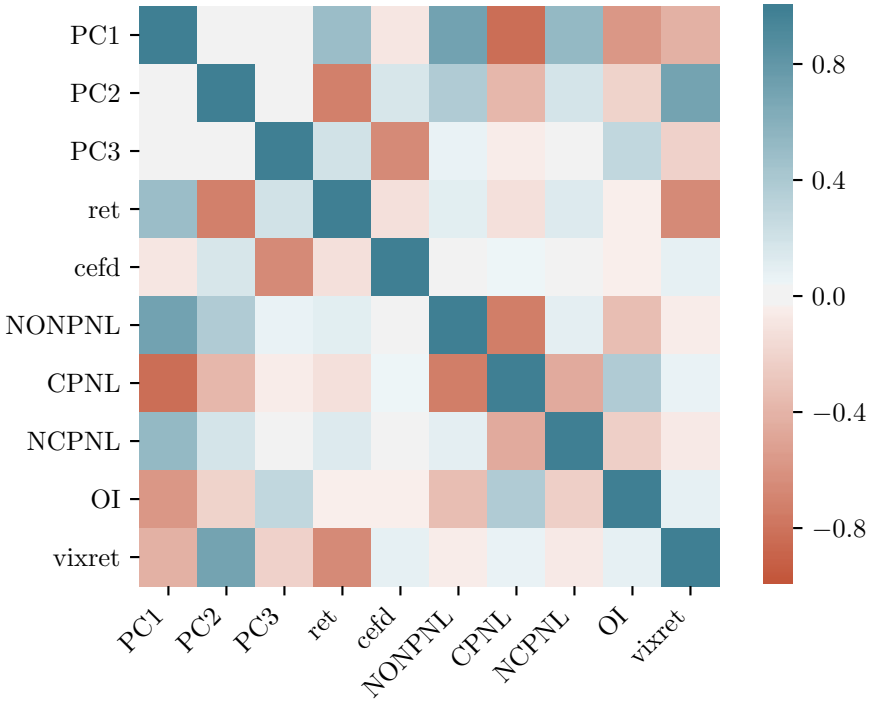
Contemporeneous Relationships

Table 4: Multivariate GLS: Features and Principal Components, Concatenated

	cefd	NONPNL	CPNL	NCPNL	OI	vixret	PC_1	PC_2	PC_3
SMALLLoBM	-0.1018	-0.0011	-0.0203	0.0032	-0.0122	-0.1858***	0.8571***	-0.6255***	-0.0836***
ME1BM2	-0.0981	0.0009	-0.0146	0.0061	-0.0035	-0.1526***	0.6329***	-0.3807***	0.1049***
SMALLHiBM	-0.1546	0.0018	-0.0125	0.0073	-0.0032	-0.1429***	0.6174***	-0.3406***	0.1964***
BIGLoBM	-0.0518	0.008	-0.0057	0.0043	0.0224**	-0.1406***	0.4862***	-0.3696***	0.1652***
ME2BM2	-0.1422**	0.0099	-0.0059	0.0071	0.0156	-0.1332***	0.4400***	-0.2179***	0.3652***
BIGHiBM	-0.0951	0.0108	-0.0105	0.0094*	0.015	-0.1356***	0.4922***	-0.1874***	0.4837***
Mean R^2_{adj}	0.3458						0.6806		

The table displays the coefficients from the two multivariate regressions, one with all six sentiment features and one with all three PCs. Each element is the estimated predictor coefficient of the associated regression. For convenience, the intercepts are not shown here. Each coefficient's significance is calculated based on an F-test.

Figure 7: Heatmap of predictor variables



Correlation matrix between all predictor variables: Features, principal components, and the market risk premium. *Note that $ret = mrp$ here.*

B

Causal Relationships: Traditional

Table 5: Causality test p-values.

	SMALLLoBM	ME1BM2	SMALLHiBM	BIGLoBM	ME2BM2	BIGHiBM
cefd	0.3514	0.2841	0.218	0.1911	0.101	0.1961
NONPNL	0.4709	0.3731	0.0138**	0.1506	0.4355	0.1705
CPNL	0.3873	0.1705	0.0812*	0.682	0.2853	0.0953*
NCPNL	0.5308	0.5911	0.2585	0.1852	0.3298	0.2232
OI	0.7423	0.5092	0.0326**	0.266	0.1238	0.7043
vixret	0.4707	0.244	0.2753	0.5519	0.638	0.9487

Rows,columns represent causing and caused variables, respectively.

Table 6: Causality test p-values

	cefd	NONPNL	CPNL	NCPNL	OI	vixret
SMALLLoBM	0.4071	0.2635	0.202	0.4381	0.0258**	0.1719
ME1BM2	0.353	0.32	0.0849*	0.4831	0.6176	0.7566
SMALLHiBM	0.6641	0.2209	0.1881	0.1764	0.8114	0.8686
BIGLoBM	0.4508	0.6123	0.1592	0.0151**	0.7866	0.9696
ME2BM2	0.6381	0.0985*	0.07*	0.2011	0.839	0.16
BIGHiBM	0.8997	0.8359	0.62	0.653	0.3822	0.7426

Rows,columns represent causing and caused variables, respectively.

C

Causal Relationships: Machine Learning

Figure 8: ΔAIC: Ridge

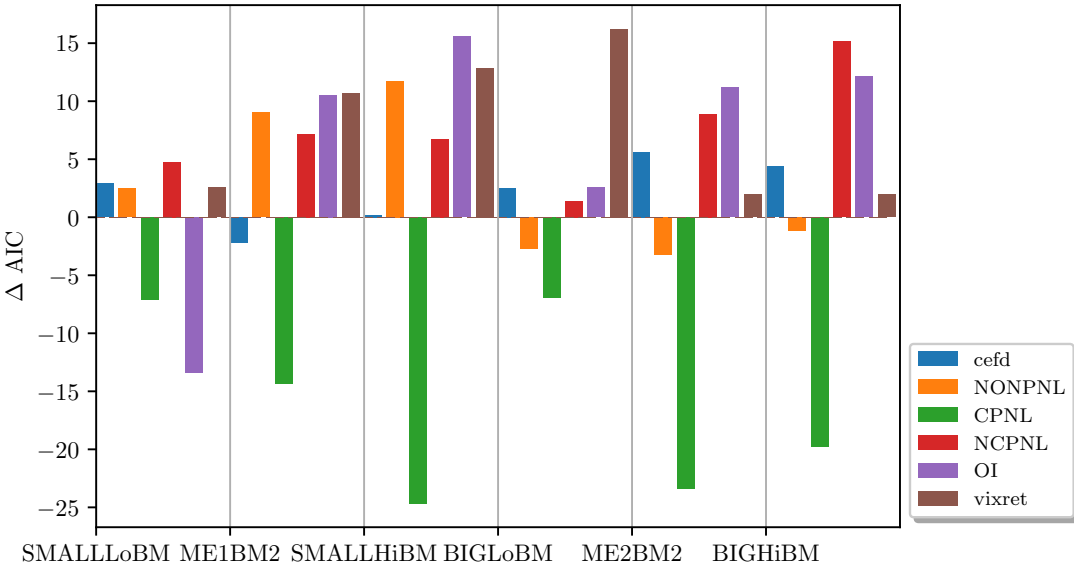


Figure 9: ΔAIC: Lasso

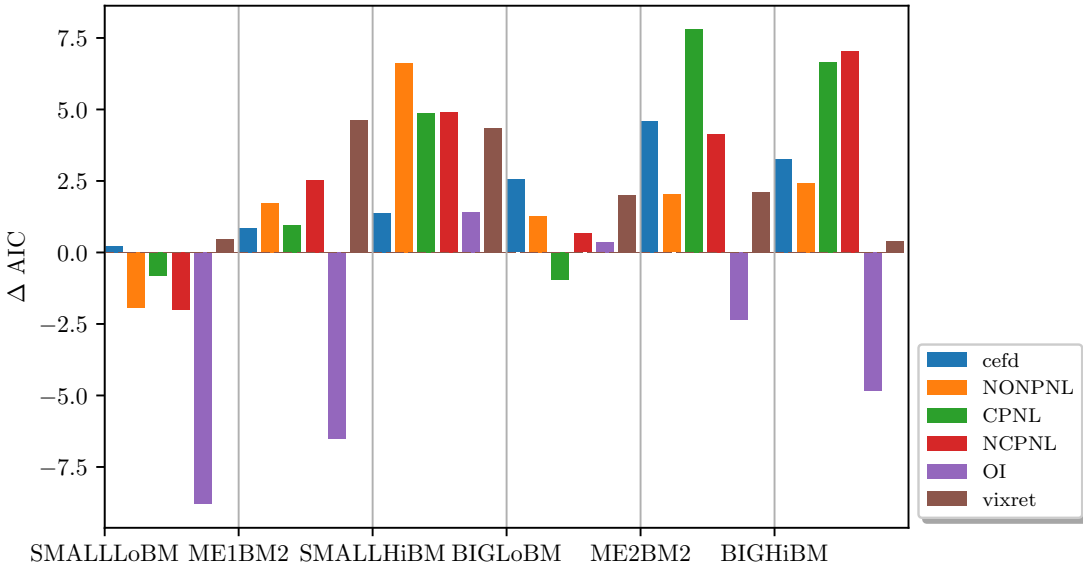
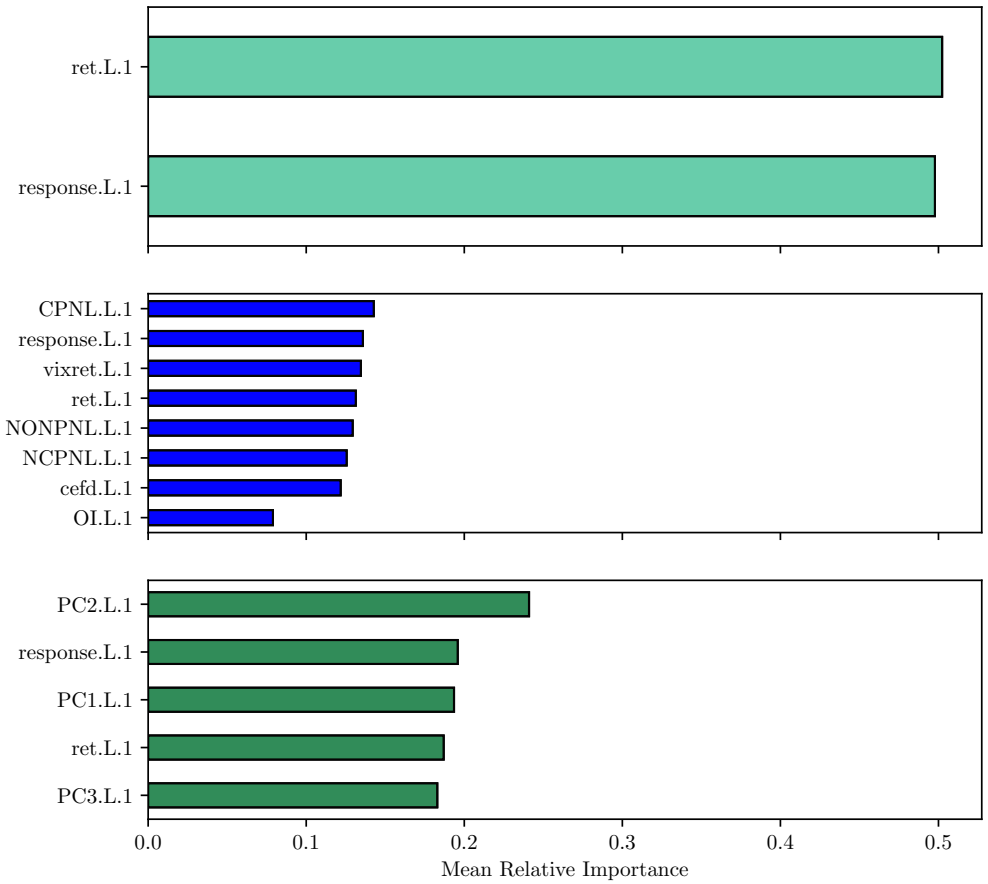


Figure 10: Feature Importances: Random Forest



Relative importances for the forest without the sentiment features, with the features, and with their principal components. Note that this displays the mean relative importance across portfolios, and hence, may not sum up to 1. *Note that $ret = mrp$ here.*

Table 7: Feature Importances; All Portfolios

	SMALLLoBM	ME1BM2	SMALLHiBM	BIGLoBM	ME2BM2	BIGHiBM
response.L.1	0.493	0.537	0.502	0.506	0.525	0.520
mrp.L.1	0.507	0.463	0.498	0.494	0.475	0.480
response.L.1	0.206	0.124	0.153	0.101	0.126	0.119
mrp.L.1	0.106	0.165	0.164	0.117	0.102	0.167
cefd.L.1	0.121	0.110	0.116	0.100	0.144	0.147
NONPNL.L.1	0.082	0.092	0.079	0.177	0.167	0.154
CPNL.L.1	0.090	0.119	0.132	0.182	0.145	0.112
NCPNL.L.1	0.166	0.186	0.140	0.100	0.065	0.096
OI.L.1	0.077	0.050	0.070	0.110	0.106	0.065
vixret.L.1	0.154	0.154	0.146	0.114	0.145	0.141
response.L.1	0.240	0.145	0.234	0.242	0.205	0.175
mrp.L.1	0.215	0.236	0.202	0.213	0.152	0.180
PC1.L.1	0.175	0.191	0.198	0.169	0.191	0.211
PC2.L.1	0.113	0.153	0.132	0.208	0.272	0.237
PC3.L.1	0.257	0.275	0.233	0.168	0.180	0.197

The table shows the relative feature importances on all portfolios. For each block, the sum of rows equals 1. None of the features are consistently more important than others. However, the sum importances of the six sentiment features is consistently higher than the sum of the responses and the market risk premium. Also, note that $PC2$, $NONPNL$ and $CPNL$ are considered more important as the stock size increases, while $NCPNL$ and $PC3$ seem to be more important in forecasting the smaller stocks.

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Arnold, A., Liu, Y., and Abe, N. (2007). Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75. ACM.

Bekaert, G. and Hoerova, M. (2014). The vix, the variance premium and stock market volatility. *Journal of Econometrics*, 183(2):1–5.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Brooks, C. (2019). *Introductory econometrics for finance*. Cambridge university press.

Chatrath, A., Liang, Y., and Song, F. (1997). Commitment of traders, basis behavior, and the issue of risk premia in futures markets. *The Journal of Futures Markets (1986-1998)*, 17(6).

Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9):1–17.

De Bondt, W. F. and Thaler, R. (1985). Does the stock market overreact? *The Journal of finance*, 40(3):793–805.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of political Economy*, 98(4):703–738.

Dwyer Jr, G. P. and Wallace, M. S. (1992). Cointegration and market efficiency. *Journal of International Money and Finance*, 11(4):318–327.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276.

Fama, E. F. and French, K. R. (1995). Size and book-to-market factors in earnings and returns. *The journal of finance*, 50(1):131–155.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 5–30, 110–200.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):61–73.

Hiemstra, C. and Jones, J. D. (1994). Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664.

Jain, P. C. and Joh, G.-H. (1988). The dependence between hourly prices and trading volume. *Journal of Financial and Quantitative Analysis*, 23(3):269–283.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of finance*, 45(3):881–898.

Johansen, S. (2005). Interpretation of cointegrating coefficients in the cointegrated vector autoregressive model. *Oxford Bulletin of Economics and Statistics*, 67(1):1–10.

Kara, Y., Boyacioglu, M. A., and Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, 38(5):5311–5319.

Keele, L. and Kelly, N. J. (2006). Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political analysis*, 14(2):186–205.

Kumar, M. and Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. In *Indian institute of capital markets 9th capital markets conference paper*.

Lee, C. M., Shleifer, A., and Thaler, R. H. (1991). Investor sentiment and the closed-end fund puzzle. *The journal of finance*, 46(1):4–75.

Lo, A. W. and MacKinlay, A. C. (1990). When are contrarian profits due to stock market overreaction? *The review of financial studies*, 3(2):175–205.

Malkiel, B. G. (1999). *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company.

Malkiel, B. G. and Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2).

Masters, M. W. and White, A. K. (2008). How institutional investors are driving up food and energy prices. *Special Report, The Accidental Hunt Brothers*.

Niarchos, N. A. and Alexakis, C. A. (1998). Stock market prices,'causality'and efficiency: evidence from the athens stock exchange. *Applied Financial Economics*, 8(2):167–174.

Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1):259–268.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Management Review*, 6(2).

Sanders, D. R., Boris, K., and Manfredo, M. (2004). Hedgers, funds, and small speculators in the energy futures markets: an analysis of the cftc's commitments of traders reports. *Energy Economics*, 26(3).

Smirlock, M. and Starks, L. (1988). An empirical analysis of the stock price-volume relationship. *Journal of Banking & Finance*, 12(1):31–41.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.

Tsai, C.-F., Lin, Y.-C., Yen, D. C., and Chen, Y.-M. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing*, 11(2):2452–2459.