Becker, Janis; Leschinski, Christian

Working Paper

Directional predictability of daily stock returns

Hannover Economic Papers (HEP), No. 624

Provided in Cooperation with:
School of Economics and Management, University of Hannover

Directional Predictability of Daily Stock Returns

Janis Becker, Christian Leschinski[1]

Leibniz University Hannover, Germany

This version: January 12, 2018

Abstract

The level of daily stock returns is generally regarded as unpredictable. Instead of the level, we focus on the signs of these returns and generate forecasts using various statistical classification techniques, such as logistic regression, generalized additive models, or neural networks.

The analysis is carried out using a data set consisting of all stocks that were part of the Dow Jones Industrial Average in 1996. After selecting the relevant explanatory variables in the subsample from 1996 to 2003, forecast evaluations are conducted in an out-of-sample environment for the period from 2004 to 2017. Since the model selection and the forecasting period are strictly separated, the procedure mimics the situation a forecaster would face in real time.

It is found that the sign of daily returns is predictable to an extent that is statistically significant. Moreover, trading strategies based on these forecasts generate positive alpha, even after accounting for transaction costs. This underlines the economic significance of the predictability and implies that there are periods during which markets are not fully efficient.

JEL-Numbers: G12, G14, G17, C38

Keywords: Asset Pricing · Market Efficiency · Directional Predictability · Statistical Classification

[1]Corresponding author:
Phone: +49-511-762-5383
Fax: +49-511-762-3923
E-Mail: leschinski@statistik.uni-hannover.de

1 Introduction

Asset return prediction is of vital importance for practitioners in the context of portfolio allocation as well as for academics due to its relation to market efficiency. Research along these lines has focused on long horizon predictability and typically utilizes monthly or quarterly returns. In this context, several studies find statistically significant predictability using simple predictive regressions (e.g. Fama and French (1988), Ang and Bekaert (2007)). Other contributions point out that this predictability is time inconsistent and negligible out-of-sample (Welch and Goyal (2008)), not economically meaningful (Bajgrowicz and Scaillet (2012)), and based on falsely specified statistical tests (Choi, Jacewitz, and Park (2016)). Recently, more sophisticated methods such as restricted regressions (Campbell and Thompson (2008)), forecast combination (Rapach, Strauss, and Zhou (2010)), and diffusion indices (Neely et al. (2014)) provide new evidence that there might be predictability in monthly stock returns after all.

While this debate about predictability on longer time horizons is still ongoing, there is a consensus that daily stock returns are unpredictable, which is theoretically implied by the efficient market hypothesis (EMH). In its semi-strong form, the EMH requires that asset prices fully reflect all publicly available information at all times (cf. Fama (1970)). Price changes can therefore only reflect the arrival of new information, which is unpredictable by definition. This gives rise to the random walk hypothesis for the level of (log-)prices. If we denote the continuously compounded return by r_t, then it can be decomposed as

$$r_t = \pi_t + \mu_t + \varepsilon_t, \tag{1}$$

where π_t denotes the risk-free interest rate, μ_t is the equity premium, and ε_t is a mean zero innovation term that is serially uncorrelated. If investors are risk-neutral, we have $\mu_t = 0$ and the EMH implies that excess returns should be unpredictable. On the other hand, if investors are risk averse, a certain degree of predictability in the equity premium μ_t is possible in efficient markets - as far as the predictability reflects time-varying aggregate risk (cf. Rapach and Zhou (2013)). On a daily frequency, however, both π_t and μ_t are close to zero since their scale is minuscule compared to the variation of ε_t. The daily return r_t should therefore be essentially unpredictable.

Predictability of stock returns on short horizons is also precluded by consumption based asset pricing models (cf. Cochrane (2009)). Based on these models, Ross (2009) and Zhou (2010) derive an upper bound of potential predictability which increases in the variance of the stochastic discount factor and in π_t. Since on a daily horizon the risk-free rate is essentially zero and the variance of the stochastic discount factor is small

compared to the monthly case, there is no theoretical basis to expect significant predictability of daily returns.

These theoretical arguments are supported by the empirical findings in the literature. Research along these lines culminated in the Nobel prize being awarded to Eugene Fama in 2013 for "showing that asset prices are extremely hard to predict in the short term". On a monthly horizon, some recent contributions suggest to forecast the sign of asset returns instead of the level. Directional predictability - predictability of the sign of r_t - is directly related to market timing, as considered for example in Henriksson and Merton (1981) and Pesaran and Timmermann (1995), and provides a natural basis for the development of trading strategies. Instead of the representation in (1), consider the decomposition

$$r_t = sign(r_t)|r_t|, \tag{2}$$

where $sign(r_t) = 1$ if $r_t > 0$, and $sign(r_t) = -1$ if $r_t \leq 0$. It is well known that the absolute value of the return $|r_t|$ is highly predictable and its behavior is well approximated by long memory processes (cf. for example Granger and Ding (1996), or Corsi (2009)). An explanation for predictability of $sign(r_t)$ is suggested in Christoffersen and Diebold (2006), who argue that directional predictability can be generated by persistence in $|r_t|$ if $\mu_t \neq 0$. Christoffersen et al. (2007) argue that variation in higher moments can generate directional predictability even if $\mu_t = 0$, as long as the distribution is asymmetric.

Empirically, Leung, Daouk, and Chen (2000) show that directional predictability is stronger than predictability of r_t itself. Evidence for significant directional predictability in US stock markets is also found by Nyberg (2011), who applies a dynamic binary probit model with recession indicators, Pönkä (2017), who uses the dynamic binary probit model and incorporates lagged industry returns, and Nyberg and Pönkä (2016), who extend the model of Pönkä (2017) to the bivariate case and include interaction effects with the United States for return predictions in other financial markets.

On a daily horizon, Linton and Whang (2007) find evidence for directional predictability of S&P 500 returns, Chung and Hong (2007) find directional predictability in daily changes of nominal exchange rates that appears to be rooted in the persistence of higher moments, and Han et al. (2016) show that realized variance is predictive for the sign of daily S&P 500 returns.

All of these findings are based on non-parametric tests that give little to no insight about the functional form of the dependence structure. Actual predictions for the sign of daily returns can only be found in the machine learning literature, for example in Kara, Boyacioglu, and Baykan (2011), or Qiu and Song (2016). These, however, report that they are able to correctly classify returns 70-80 percent of the time, which is theoretically

implausible and points to an issue with overfitting.

Motivated by these findings, we take a comprehensive look at the directional predictability of daily asset returns. For this purpose, we use a data set consisting of all stocks that were part of the *Dow Jones Industrial Average* (DJIA) in 1996 and various statistical classification methods.

The classification methods include logistic regression, generalized additive models, neural networks, support vector machines, random forests, and boosted classification trees.

For each method, the relevant explanatory variables are selected in the subsample from 1996 to 2003 based on a forward selection procedure that utilizes cross-validation techniques.

Subsequently, the predictive performance of the selected models is evaluated in an out-of-sample environment for the period from 2004 to 2017, where each model is re-estimated in a rolling window to generate one step-ahead forecasts. Since the model selection and the forecasting period are strictly separated, the procedure mimics the situation a forecaster would face in real time.

In addition to the statistical significance of the results, we also consider their economic significance. For this purpose, we propose trading strategies that are suitable to exploit directional predictability and evaluate the generated returns using Sharpe ratio, realized utility, and alpha. To account for trading costs, we consider actual bid-ask returns.

This procedure allows us to establish several key findings. First, directional predictability on a daily frequency exists, it is of a magnitude that is statistically significant, and it is consistent over time. Second, trading strategies based on these predictions generate abnormal returns even after accounting for known risk factors and transaction costs. Third, the parametric logistic regression model generates the best predictions, whereas non-parametric machine learning techniques are too flexible and too prone to overfitting.

As discussed above, equation (1) implies that the level as well as the sign of future returns should be unpredictable on a daily horizon where $\pi_t \approx \mu_t \approx 0$. Therefore, our findings are in clear contradiction to the random walk hypothesis. However, in its weakest form, the EMH allows for deviations from the random walk, as long as these cannot be exploited due to transaction costs (cf. Fama (1991)). Statistical significance is therefore not equivalent to an economically meaningful violation of the EMH. Nevertheless, even after accounting for transaction costs, we observe significant alphas. The majority of the trading returns appears to be concentrated in brief periods during which the predictability intensifies so that substantial trading returns can be generated.

The remainder of this paper is organized as follows. The next section introduces the data set and the explanatory variables. Afterwards, we discuss our model selection and forecasting procedure in Section 3. Section 4 then reports the empirical results showing the statistical, as well as economic significance of our forecasts. The direction of the

influence of the explanatory variables is explored in Section 5. Finally, Section 6 shows the robustness of our results, before Section 7 concludes.

2 Data

We consider 5-minute data for $N = 30$ stocks included in the DJIA on January 1, 1996 and several explanatory variables. The data set is obtained from the *Thomson Reuters Tick History* data base and ends on January 31, 2017.

Since high frequency data is often subject to minor recording mistakes, it is common practice to apply some form of data cleaning. Here, we adopt the approach of Barndorff-Nielsen et al. (2009), which comprises, among other things, the removal of observations with negative stock prices, negative bid-ask spreads, and abnormal high or low entries in comparison to other observations on the same day. The resulting cleaned data set is then used to calculate daily stock returns and more than 20 explanatory variables.[1]

For modeling and forecasting purposes, we calculate logarithmic returns that facilitate the computation of explanatory variables such as moving averages or realized volatilities. Even though the differences between discrete and logarithmic returns are small due to the short time horizon, the logarithmic returns are transformed to discrete returns when considering trading strategies in Section 4.5. This allows for the calculation of portfolio returns.

Our analysis is based on the companies that were components of the DJIA in the beginning of our sample in January, 1996. Over the course of time, several companies faced bankruptcy or were taken over so that the respective time series end before 2017.[2]

2.1 Explanatory Variables

Since predictability of daily stock returns has been largely ruled out in the literature, there is no established set of potential explanatory variables. Valuation ratios such as dividend-price or earnings-price ratios, for example, which some studies find to be predictive on a monthly or quarterly basis, cannot be used, as they do not change on a daily basis.

We therefore focus on variables that (i) exhibit meaningful variation on a daily frequency, (ii) are easily available, and (iii) for which a plausible economic argument can be made,

[1] Since bid-ask spreads tend to be higher at market closing time than they are five minutes before and we intend to use the resulting forecasts for trading purposes, daily returns are calculated using the closing prices at 3:55pm each day.

[2] This concerns *Bethlehem Steel, Eastman Kodak, Sears Roebuck, Texaco, Union Carbide*, and *Westinghouse Electric*. *General Motors* went bankrupt in 2009 but returned to the *New York Stock Exchange* only one year later.

or where predictability on a longer horizon has been found in previous studies.
This includes measures of moments of the return distribution, such as log-realized variances and realized skewness, since their potential predictive power can be motivated by the theoretical arguments of Christoffersen and Diebold (2006) and Christoffersen et al. (2007).
Market measures, such as S&P 500 returns and realized betas, are considered because the CAPM implies a strong relationship between market and stock returns. The log-realized variance of the S&P 500 and the VIX are included for the same reasons as the other risk measures mentioned above.
A number of recent studies such as Chernov (2007), Bollerslev, Tauchen, and Zhou (2009) and Bollerslev et al. (2014) introduce the variance premium that is defined as the difference between the implied variance under the assumption of risk neutrality and the conditional expectation of the variance. This measure is related to the aggregate risk aversion and it is shown to be a significant predictor for the level of monthly stock returns. Therefore, it is included as well.[3]
The yield curve is found to be predictive for future macroeconomic activity (cf. Ang, Piazzesi, and Wei (2006)), which should also influence stock prices. We therefore include the first three principal components (PC) of the yield curve (along with their changes) that can be interpreted as level, slope, and curvature of the curve. These are calculated using US bonds with over 40 different maturities.
Finally, some technical indicators implemented in the aforementioned machine learning studies are used as well. This results in the following list of 24 explanatory variables.

- **(Realized) Measures of Moments:** log-realized variance (cf. e.g. Amaya et al. (2015)), high-low variance (cf. Corrado and Truong (2007)), realized skewness (cf. e.g. Amaya et al. (2015)).
- **Financial Market Indicators:** S&P 500 return, realized betas calculated from S&P 500 5-minute returns, log-realized variance S&P 500, level VIX, VIX return, oil return.
- **Risk Aversion Indicators:** variance premium (cf. Bollerslev, Tauchen, and Zhou (2009)).
- **Yield Curve Measures:** level and change of first PC (level of the yield curve), second PC (slope of the yield curve), and third PC (curvature of the yield curve).
- **Technical Indicators:** stock return, 5-day moving average stock return, on-balance volume (cf. Neely et al. (2014)), 12-day moving average of binary stock

[3] We estimate the conditional expectation of the variance using the HAR model of Corsi (2009).

	Mean	SD	1%	25%	50%	75%	99%
Log-Realized Variance	-8.592	0.911	-10.476	-9.344	-8.708	-8.036	-6.370
High-Low Variance in Percent	0.033	0.163	0.001	0.006	0.013	0.029	0.309
Realized Skewness	0.033	1.009	-2.670	-0.544	0.019	0.603	2.758
S&P 500 Return in Percent	0.025	1.210	-3.412	-0.522	0.056	0.612	3.386
Realized Beta	0.800	0.396	-0.049	0.654	0.914	1.151	1.868
Log-Realized Variance S&P 500	-9.715	1.015	-11.784	-10.443	-9.794	-9.109	-6.888
Level VIX	21.562	8.632	10.680	15.160	20.300	25.240	50.930
VIX Return in Percent	-0.001	6.690	-16.040	-3.818	-0.411	3.308	20.197
Variance Premium	6.868	4.217	0.000	4.092	6.318	9.257	19.507
Oil Return in Percent	0.028	2.380	-6.305	-1.260	0.036	1.379	6.029
Level First PC (Level)	-1.301	6.389	-12.334	-6.492	-0.684	5.907	8.162
Level Second PC (Slope)	-0.138	1.120	-2.800	-0.798	-0.057	0.761	1.750
Level Third PC (Curvature)	0.004	0.310	-0.807	-0.197	0.023	0.224	0.574
Change First PC (Level)	0.003	0.171	-0.480	-0.082	0.000	0.087	0.460
Change Second PC (Slope)	0.000	0.090	-0.240	-0.043	0.001	0.046	0.222
Change Third PC (Curvature)	0.000	0.049	-0.122	-0.019	0.000	0.018	0.151
Stock Return in Percent	0.013	2.167	-5.905	-0.926	0.000	0.966	5.811
5-Day Moving Average Return in Percent	0.014	0.944	-2.683	-0.407	0.040	0.465	2.439
On-Balance Volume ($\times 10^{-4}$)	-0.426	3.146	-10.268	-1.262	-0.098	0.606	12.443
12-Day Binary Moving Average	0.497	0.140	0.167	0.417	0.500	0.583	0.833
Momentum Indicator	0.047	1.544	-4.559	-0.539	0.053	0.665	4.312
A/O Oscillator	0.515	0.538	-0.680	0.167	0.500	0.861	1.753
Rate-of-Change Indicator	1.006	0.077	0.791	0.969	1.007	1.044	1.211

Table 1: Summary statistics of the explanatory variables. Mean corresponds to the mean of the variables, SD to the standard deviation and the remaining columns state the respective quantiles of the variables.

returns (cf. Qiu and Song (2016)), momentum indicator (cf. Qiu and Song (2016)), A/O oscillator (cf. Qiu and Song (2016)), and rate-of-change indicator (cf. Qiu and Song (2016)).[4]

Detailed explanations and discussions of these variables can be found in the referenced articles. Moreover, Table 1 reports summary statistics and Figure 1 shows the correlation matrix of the variables.

3 Methodology

In this section, we describe our model selection and forecasting framework and discuss the econometric reasoning behind the respective modeling choices.

[4]Let $C_{i,t}$ be the closing price, $L_{i,t}$ the lowest price, $H_{i,t}$ the highest price, and $V_{i,t}$ the volume of trade of stock i at day t. Moreover, let $\Theta_{i,t} = 1$ if $C_t \geq C_{i,t-1}$ and $\Theta_{i,t} = -1$ otherwise. Then the on-balance volume (OBV) is given by $OBV_{i,t-1} + \Theta_{i,t} * V_{i,t}$, the momentum indicator by $C_{i,t} - C_{i,t-4}$, the A/O oscillator by $\frac{H_{i,t}-C_{i,t-1}}{H_{i,t}-L_{i,t}}$, and the rate-of-change indicator by $C_{i,t}/C_{i,t-14} \times 100$.

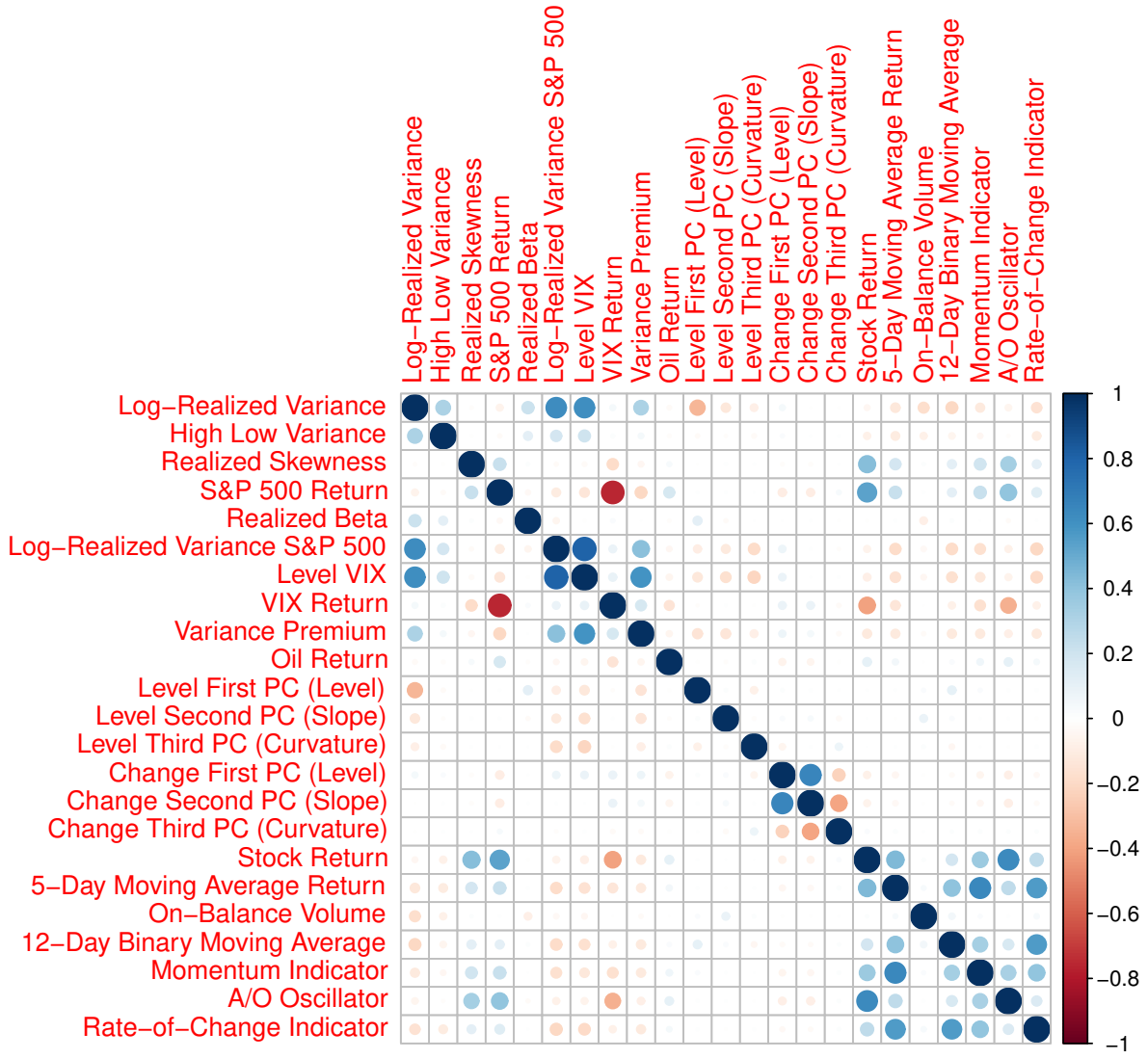

Figure 1: Correlation plot of the explanatory variables. Size and color of the circles correspond to the degree and direction of the correlation.

3.1 Model Framework

In contrast to regression problems, where the dependent variable can take values in $\mathbb{R}$, directional predictability is a classification problem, which means that the dependent variable $y_{i,t+1} = I(r_{i,t+1} > 0)$ takes the value one if the return of stock i at time t is positive, and zero otherwise. The classification model can be represented in the form

$$y_{i,t+1} = G^{(a)}(x_{i,t}, \theta_i) + \varepsilon_{i,t+1}, \tag{3}$$

where the function $G^{(a)}$ denotes the classifier, $\varepsilon_{i,t+1}$ is an error term, the index a specifies the classification method that is used, $i = 1, ..., N$, and $t = 1, ..., T$. For simplicity, the dependence of $x_{i,t}$, θ_i, and $\varepsilon_{i,t+1}$ on a is omitted from the notation. Each classifier

$G^{(a)}(x_{i,t}, \theta_i)$ is a function of the parameter vector θ_i and the vector of explanatory variables $x_{i,t}$. Note that the explanatory variables in $x_{i,t}$ are lagged by one period relative to the dependent variable $y_{i,t+1}$ as the aim is to obtain one-step ahead forecasts.

The list of classification methods considered includes logistic regression, generalized additive models (Hastie and Tibshirani (1990)), support vector machines (Cortes and Vapnik (1995)), neural networks (Ripley (1996)), random forests (Breiman (2001)), and boosted trees (Friedman (2001)). Detailed explanations of these methods can be found in the referenced books and articles. As mentioned in the introduction, previous studies on directional predictability of daily stock returns claim that some of these classifiers are able to correctly predict up to 80 percent of the returns (cf. Kara, Boyacioglu, and Baykan (2011) and Qiu and Song (2016)).

The classical stock return predictability literature typically employs simple bivariate predictive regressions. This is due to the low signal-to-noise ratio in stock returns and the low number of monthly or quarterly observations. Since we consider observations on a daily frequency, we have at least 22 times more observations for the same time span and therefore face a different bias-variance trade-off. This enables us to consider more complex multivariate models with several explanatory variables. These are more likely to successfully approximate the complex interrelation between stock returns and the explanatory variables.

The aforementioned low signal-to-noise ratio in stock returns is challenging for statistical analysis in several ways. First, all methods except for logistic regression are nonlinear in variables, and some are even non-parametric so that they are able to adapt extremely well to patterns in the data set and capture highly complex relations. Due to their flexibility, they are, however, also especially prone to overfitting, which is particularly problematic if the signal is low relative to the noise. Second, we consider a large number of potential explanatory variables so that it is likely that variables appear to be significant due to random covariation with the noise component.

To address the first issue, we evaluate the predictive performance in an out-of-sample environment. In the empirical analysis, the model selection and the forecasting period are strictly separated. While the model is chosen in the years from 1996 to 2003 ($T_M \approx 2{,}000$ observations per stock), out-of-sample forecasts take place from 2004 to 2017 ($T_F \approx 3{,}300$ predictions per stock).[5] Obviously, it holds that $T_M + T_F = T$. It needs to be emphasized that this separation entails that our out-of-sample analysis simulates the situation an investor starting to invest in 2004 would have faced in real time. Clearly, the length of these two periods is selected rather arbitrarily. However, in Section 6 it is

[5] Some companies went bankrupt which results in varying T_M and T_F per stock. Furthermore, days with recording errors (indicated by the procedure of Barndorff-Nielsen et al. (2009)) were excluded, which also implies slightly different sample sizes per stock.

shown that the results are robust with respect to the choice of these periods.
To deal with the second issue, we restrict the set of explanatory variables $\mathbb{M}^{(a)}$, which is included in the regressor vector $x_{i,t}$, to be the same across all stocks. This has the advantage that the number of observations for model selection increases drastically compared to the case where each stock is allowed to have a different set of regressors. Consequently, we obtain more stable results. This restriction neither means that the values of the explanatory variables are the same for all stocks, nor that the coefficients θ_i associated with the variables are identical.

3.2 Model Selection

With 24 possible explanatory variables and a low signal-to-noise ratio, we require a model selection procedure to obtain a more parsimonious model. In general, a model selection procedure consists of two components, a goodness-of-fit criterion to evaluate the performance of each candidate model, and a rule that defines which models are considered as candidate models.
With respect to the choice of the goodness-of-fit criterion, we face a number of statistical challenges in the setup considered here. First, as the statistical methodology and implementation of the models varies heavily, e.g. some parametric and some non-parametric methods, it is almost certain that the optimal variable combination varies across classification methods. This implies that it is sensible to choose a different model for each classifier. At the same time, the goodness-of-fit criterion should be the same for all classifiers so that the results are comparable across the different methods. Second, with regard to the flexibility of the non-parametric methods, in-sample analysis is most likely meaningless due to overfitting. Third, some of the explanatory variables exhibit strong temporal dependence, which may pose problems for procedures that assume a weaker form of persistence.
To summarize, we require a goodness-of-fit criterion for the performance evaluation of a classifier $G^{(a)}$ that is (i) applicable to all classifications methods so that the results are comparable, (ii) robust to overfitting, and (iii) robust to the strong serial dependence in some of the explanatory variables.
We therefore consider one-step-ahead forecasts for the latter part of the model selection period generated from the classifier estimated in an expanding window. We then calculate the proportion of correctly predicted $y_{i,t}$ (the so-called hitrate) in this pseudo out-of-sample experiment.
More formally, denote the classifier estimated using the observations from time 1 to time t by $G^{(a)}_{[1,t]}(x_{i,t}, \hat{\theta}_i)$. Then for a given set of explanatory variables $\mathbb{M}^{(a)}$ the average

out-of-sample hitrate (OOSH) is defined as

$$OOSH(\mathbb{M}^{(a)}) = N^{-1} \sum_{i=1}^{N} \frac{\#\left\{I\left(G^{(a)}_{[1,t]}(x_{i,t}, \hat{\theta}_i) > 0.5\right) = y_{i,t+1}\right\}_{t \in \{S,\ldots,T_M-1\}}}{(T_M - S)}. \tag{4}$$

In the machine learning literature, this approach is referred to as *last block cross-validation* (cf. Hjorth (1982)). This is a modified version of ordinary cross-validation that is applicable when observations are not *iid* but exhibit temporal dependence.
The OOSH is a direct estimate of the expected out-of-sample hitrate conditional on the model. Since we aim to maximize the latter, the former is an appropriate measure to evaluate the performance of a model. Further information on this relation can be found in Arlot and Celisse (2010).
With respect to the set of candidate models, it would be optimal to consider all possible combinations of variables. This approach is referred to as best subset selection. Unfortunately, this would require to evaluate 2^{24} models, which is computationally infeasible. This is especially true for computationally demanding models such as support vector machines or neural networks. Therefore, we use stepwise forward selection as a computational surrogate.
Recall that the set of explanatory variables used by the classifier $G^{(a)}$ is denoted by $\mathbb{M}^{(a)}$. If the set of all K possible explanatory variables is denoted by $\mathbb{P}$, and $\mathbb{M}^{(a)}_k$ is the set of explanatory variables that is already selected as part of the model in step k, then $\mathbb{P}^{(a)}_k = \mathbb{P} \setminus \mathbb{M}^{(a)}_{k-1}$ is the set of variables that could still be added to the model. Each of these variables is referred to as $\mathbb{P}^{(a)}_{k,j}$. The procedure then proceeds as follows:

0. Initialization:

Set $k = 1$ and $\mathbb{M}^{(a)}_0 = \emptyset$.

1. Forward Selection:

Set $\mathbb{M}^{(a)}_k = \arg\max_{j=1,\ldots,K-k+1} OOSH\left(\mathbb{M}^{(a)}_{k-1} \cup \mathbb{P}^{(a)}_{k,j}\right)$.

2. Model Selection:

If $k < K$, increase k by one and go back to Step 1.
Otherwise set $\mathbb{M}^{(a)} = \arg\max_{k=1,\ldots,K} OOSH\left(\mathbb{M}^{(a)}_k\right)$ and terminate the procedure.

In simpler terms, we generate a sequence of models by iterating through a procedure that starts out with the empty model in Step 0 and sequentially adds variables to the model, until the full set of regressors is used in the K-th model $\mathbb{M}^{(a)}_K$.
In each iteration k of Step 1, we sequentially add each of the remaining variables in $\mathbb{P}^{(a)}_k$ to the $k-1$-variable model from the previous iteration to generate candidate models

with k regressors. For each candidate model we calculate the OOSH and then select the model that generates the largest improvement. Since some of the classification methods require the selection of tuning parameters, these are also determined as part of Step 1 using a grid search procedure and the OOSH as the goodness-of-fit criterion. To reduce the computational effort, the expanding window estimate $G^{(a)}_{[1,t]}(x_{i,t}, \hat{\theta}_i)$ is updated only every 40th observation.
Finally in Step 2, we consider the sequence of models of increasing size and select the set of regressors $\mathbb{M}^{(a)}$ that achieves the lowest OOSH.

3.3 Forecasting Procedure

After selecting the set of explanatory variables $\mathsf{M}^{(a)}$ for each classifier based on the model selection period from 1996 to 2003, the actual tests of directional predictability are carried out for the out-of-sample period from 2004 to 2017.
Based on (3), we generate one-step-ahead forecasts using

$$G^{(a)}_{[T_M-W+t+1,T_M+t]}(x_{i,t}, \hat{\theta}_i) = \hat{P}(y_{i,t+1} = 1|x_{i,t}) \tag{5}$$

for $t = 0, ..., T_F - 1$, where $x_{i,t}$ contains the variables in $\mathbb{M}^{(a)}$. This means the model is re-estimated in each period using a rolling window of the previous $W = 1000$ observations. As indicated by the equation, each value of $G^{(a)}_{[T_M-W+t+1,T_M+t]}(x_{i,t}, \hat{\theta}_i)$ is an estimate of the probability that there is a positive return on the next day.[6] To convert these probabilities into actual forecasts for $y_{i,t+1}$, we need to define a threshold above which we predict the next days return to be positive.
The simplest threshold is the Bayes classifier that assigns

$$\hat{y}_{i,t+1} = I\Big(G^{(a)}_{[T_M-W+t+1,T_M+t]}(x_{i,t}, \hat{\theta}_i) > 0.5\Big). \tag{6}$$

This decision rule is optimal since the loss of falsely classifying positive and negative returns is equal. Note that this threshold has already been used in equation (4).

4 Empirical Analysis

In the following, we present the results of our empirical analysis. After the selected models are discussed in Section 4.1, we evaluate the forecasting performance of the models compared to several naive benchmarks in Sections 4.2, 4.3, and 4.4. Sections 4.5 and 4.6 then address the economic significance of the results.

[6] The only exception are support vector machines, where the sign of $G^{(a)}$ is directly indicative of the class.

	Logit	GAM	NN	SVM	RF	Boost
Log-Realized Variance		X		X	X	X
High-Low Variance						X
Realized Skewness					X	
S&P 500 Return	X	X	X	X	X	X
Realized Beta				X	X	
Log-Realized Variance S&P 500		X		X	X	
Level VIX				X		
VIX Return	X				X	
Variance Premium	X	X		X	X	
Oil Return		X		X		
Level Third PC (Curvature)	X	X	X		X	
Stock Return	X	X	X	X	X	X
5-Day Moving Average Return	X	X	X	X	X	
12-Day Moving Average of Binary Stock Returns			X	X	X	X
Momentum Indicator			X	X	X	
A/O Oscillator	X	X	X	X	X	
Rate-of-Change Indicator		X				X

Table 2: Selected models separated by classification method. An X indicates that the selection procedure included this variable in the respective model. Logit abbreviates Logistic Regression, GAM Generalized Additive Model, NN Neural Network, SVM Support Vector Machine, RF Random Forest, and Boost Boosted Trees.

4.1 Model Selection

Table 2 shows the results of the model selection procedure described in Section 3.2 for each of the classification methods. The size of the selected models varies between 7 variables for the logistic regression and 13 variables in the random forest. All of the selected models contain the lagged S&P 500 return as well as the lagged return of the respective stock itself. Furthermore, the technical indicators 5-day moving average return, and the A/O oscillator are included by all models, except for the boosted classification tree. Four out of six models include the log-realized variance, the variance premium, and the curvature of the yield curve. Therefore, all of these variables seem to be relevant for directional predictions.

The variables not included in Table 2 were not selected by any of the models. This is the case for the level and change of the first and second principal component of the yield curve, and the on-balance volume.

Analyzing the hitrates in the pseudo out-of-sample part of the model selection period is not sensible, since these observations have been used for the selection of explanatory

variables and tuning parameters. It can therefore be expected that the models are fitted to some extent to the random variation in these pseudo out-of-sample observations. This results in the fact that the hitrate on these observations will overestimate the true out-of sample hitrate - an effect referred to as *optimism bias*.
Of course it would be desirable to gain further insights into the form of the dependence between the selected explanatory variables and the sign of the next days return. Unfortunately, the majority of the classification methods used here are typically regarded as black box models, since the estimated models are to complex to be easily interpretable. Furthermore, the parameters of the models are allowed to vary across stocks, and the estimations are carried out in an expanding window so that the size and direction of effects may change over time.
Instead of considering the functional form here, we therefore conduct a full sample analysis for the best performing classifier later on in Section 5, once it is established that there actually is significant predictability.

4.2 Forecast Results

To evaluate the performance of the forecasts generated by the procedure described in Section 3.3, it is helpful to consider the performance of a naive benchmark forecast.
In the level predictability literature, the naive benchmark is typically the expanding mean of previous returns. This is because on monthly or longer horizons the average return should be equal to the equity premium. Since the size of the equity premium is unknown, the expanding mean is the best available estimate of the unconditional expectation.
For directional predictions there is no such established benchmark. We therefore consider three possible approaches.
A first naive approach could be to assume that positive and negative returns are equally likely. This would correspond to a random walk model without drift and with a symmetric innovation distribution. To avoid sampling variation induced by randomly assigning predictions for positive and negative returns, we alternate between predicting a positive and a negative return. Due to its relation to the random walk hypothesis, this benchmark is referred to as the *random walk* forecast.
The second benchmark allows the random walk model to have a non-zero drift and a negatively skewed innovation distribution. Although the scale of the equity premium and the risk-free rate is minuscule compared to the variation of the returns, theoretically both should be slightly positive. This introduces a positive drift. Furthermore, daily stock returns are typically found to have slightly negative skewness (cf. Christoffersen (2003)). Therefore, positive returns can occur more often than negative returns, even if the mean

				DM test		
	HR	SE	SP	RW	HM	Opt
Random Walk Forecast (RW)	50.13	50.13	50.13	0.00		
Historical Majority Forecast (HM)	50.03	35.35	65.23	-0.33	0.00	
Optimist Forecast (Opt)	50.86	100.00	0.00	1.44*	1.26*	0.00
Logistic Regression	51.99	55.99	47.86	5.66***	5.14***	2.14**
Generalized Additive Model	51.35	52.34	50.32	3.81***	3.60***	0.87
Neural Network	50.51	52.52	48.43	1.41*	1.50*	-0.73
Support Vector Machine	51.02	59.73	42.02	3.21***	3.11***	0.32
Random Forest	50.51	50.86	50.15	1.61*	1.44*	-0.68
Boosted Tree	50.92	59.62	41.92	2.55***	2.80***	0.11

Table 3: Aggregated forecasting results. All values represent the average over all forecasts for all stocks. HR, SE, and SP are given in percent and the right panel shows the values of the DM statistic. Here, under the null hypothesis, the forecasts of the model in the respective row is at best equally good to that in the respective column. The symbols ***, (**), and [*] then indicate that the null is rejected at the 1%, (5%), or [10%] level.

of the returns is zero. Taking these two arguments together, the hitrate obtained by predicting a positive return for each day should be higher than that of the random walk benchmark. We refer to this benchmark as the *optimist* forecast.

Finally, for comparability with the level predictability literature, we consider an analogue to the expanding mean as a third benchmark. This forecast, referred to as the *historical majority* forecast, is obtained by predicting a positive or a negative return depending on whether the majority of the previous returns of the stock in an expanding window was positive or negative.

The quality of the predictions will be judged based on the hitrate (HR), the sensitivity (SE), and the specificity (SP). For each stock these are defined as follows:

$$HR_i = \frac{\sum_{t=1}^{T_F} I(\hat{y}_{it} = y_{it})}{T_F},$$
$$SE_i = \frac{\sum_{t=1}^{T_F} I(\hat{y}_{it} = y_{it} = 1)}{\sum_{t=1}^{T_F} I(y_{it} = 1)},$$
$$\text{and} \quad SP_i = \frac{\sum_{t=1}^{T_F} I(\hat{y}_{it} = y_{it} = 0)}{\sum_{t=1}^{T_F} I(y_{it} = 0)}.$$

The hitrate is therefore simply the proportion of returns that are correctly classified, the sensitivity is defined as the proportion of positive returns that are correctly classified, and the specificity is the proportion of negative returns that are correctly classified.

Consequently, hitrate, sensitivity, and specificity lie between zero and one with higher values indicating better classification performance.

There is typically a trade-off between sensitivity and specificity. For example, the optimist forecast, which always predicts an up-movement, will classify all positive returns correctly but none of the negative ones, so that $SE_i = 100$ percent and $SP_i = 0$ percent for all $i = 1,...,N$. In contrast, the random walk forecast, which alternates between predicting an up-day and a down-day, will have $SE_i \approx 50$ percent and $SP_i \approx 50$ percent if the proportions of positive and negative returns in the sample are roughly equal.

Table 3 present the averages of these measures over all stocks. As can be seen, the benchmarks obtain hitrates between 50.03 percent for the historical majority forecast and 50.86 percent for the optimist forecast. Among the classifiers, the logistic regression achieves the highest hitrate with 51.99 percent, followed by the generalized additive model with 51.35 percent. The forecast with the lowest out-of-sample hitrate is the random forest with a hitrate of 50.51 percent. All models have higher sensitivity than specificity. Therefore, the proportion of positive returns that are correctly classified is higher than that of negative returns. This is due to the fact that the models on average predict an up movement in more cases than the true fraction of 50.86 percent. A similar observation is made by Nyberg (2011), who predicts monthly stock index movements using a dynamic probit model.

Overall, two key observations can be made. First and most importantly, the hitrate for logistic regression (51.99 percent) is clearly above that of all benchmarks. This is first evidence that the sign of daily returns is, to some extent, predictable. Second, the linear parametric logistic regression model outperforms the more flexible non-linear and non-parametric procedures.

Testing the statistical significance of these results is complicated by the panel structure of the forecasts. The tests conventionally applied in the directional forecasting literature such as those of Pesaran and Timmermann (1992) and DeLong, DeLong, and Clarke-Pearson (1988) are constructed for simple cross sections or time series and pooling of the forecasts for all stocks is not possible due to the strong cross-sectional dependence. It is, however, possible to construct a test in the spirit of Diebold and Mariano (1995). Calculating the average loss function over all units in the cross section for each day t circumvents the cross-sectional dependence. More formally, denote the forecasts of two competing models by $\hat{y}_{it}^{(1)}$ and $\hat{y}_{it}^{(2)}$. Then the loss differential between these forecasts at day t is given by

$$l_t = N^{-1} \sum_{i=1}^{N} \left(y_{it} - \hat{y}_{it}^{(1)}\right)^2 - \left(y_{it} - \hat{y}_{it}^{(2)}\right)^2. \tag{7}$$

	Logit	GAM	NN	SVM	RF	Boost
Logit	0.00	2.30**	4.73***	3.14***	5.03***	3.49***
GAM		0.00	2.92*	1.02	3.26**	1.39*
NN			0.00	-1.92	-0.06	-1.61
SVM				0.00	1.73**	0.50
RF					0.00	-1.49
Boost						0.00

Table 4: Aggregated model comparison results using the DM test. As before, under the null hypothesis the forecasts of the model stated in each row are at best as good as the forecasts of the model stated in the respective column. The symbols ***, (**), and [*] then indicate that the null is rejected at the 1%, (5%), or [10%] level. Logit abbreviates Logistic Regression, GAM Generalized Additive Model, NN Neural Network, SVM Support Vector Machine, RF Random Forest, and Boost Boosted Trees.

A simple Diebold-Mariano statistic that is asymptotically standard normal is then given by

$$DM = \sqrt{T_F}\frac{\sum_{t=1}^{T_F} l_t}{\sqrt{Var(l_t)}}.$$

This boils down to testing whether the hitrates of two forecasts are significantly different from each other. For the comparison with the benchmark forecasts, the results of the test are displayed on the right hand side of Table 3. It is found that the forecasts of the logistic regression model are significantly better than those of all possible benchmarks. All other classifiers produce forecasts that are significantly better than those of the random walk and the historical majority benchmark but not than those of the optimist benchmark. It can further be seen that the optimist forecast significantly outperforms the other two benchmarks.

Obviously, the DM statistic can also be used for pairwise comparisons between the classifiers. These results are shown in Table 4. It can be seen that the logistic regression forecasts are significantly better than those of all other classifiers. Furthermore, the forecasts generated by the generalized additive model significantly outperform those from the neural network, the random forest, and the boosted tree - at least at the 10 percent level.

4.3 Disaggregated Forecasting Results for Individual Stocks

After establishing the existence of predictability jointly for the whole cross section, we now consider the results for individual stocks. As in the previous section, we first investigate whether significant directional predictability is present and then we compare

the models among each other.

For this purpose, we again consider the Diebold-Mariano test as described above but without cross-sectional averaging. Note, however, that this test may be ill-suited to test the null hypothesis of no predictive ability due to the dependence of the test results on the loss function. As an example, consider the case where instead of the loss differential in (7) we use $l_{it} = \left(y_{it} - \hat{p}_{it}^{(1)}\right)^2 - \left(y_{it} - \hat{p}_{it}^{(2)}\right)^2$, where $\hat{p}_{it}^{(1)}$ and $\hat{p}_{it}^{(2)}$ are the probabilities for a positive return implied by the two competing models. For the optimist benchmark the implied probability is $p_{it}^{(Opt)} = 1$ and for the random walk forecast the probability is given by $p_{it}^{(RW)} = 0.5$, for every period t and all stocks i.

Even though this also appears to be a sensible choice for the loss differential, the results are counter intuitive. The logistic regression forecasts still significantly outperform the optimist forecasts, but they tend to do worse than the random walk forecasts. Moreover, under this loss function the random walk forecasts are significantly better than the optimist forecasts for all stocks. Both of these findings are in contradiction to the results discussed above.

Therefore, we also apply the aforementioned test of Pesaran and Timmermann (1992) that is now applicable, since we consider the univariate time series for each stock separately. The test is specifically designed to test the null of no directional predictability. Under the null hypothesis, the forecasts $\hat{y}_{it}$ and the realizations y_{it} are independent, implying that the realizations are not predictable using the model under consideration. Under the alternative hypothesis, there is a positive relationship between $\hat{y}_{it}$ and y_{it}. This would imply that the direction of the returns is, to some extent, predictable.

Following Pesaran and Timmermann (1992), the test statistic is given by

$$PT_i = \frac{\sqrt{T_F}(SE_i + SP_i - 1)}{\left(\frac{\bar{P}_{y_i}(1-\bar{P}_{y_i})}{\bar{y}_i(1-\bar{y}_i)}\right)^{1/2}},$$

where sensitivity and specificity are defined as above, $\bar{y}_i$ is the average class of the stock, and $\bar{P}_{y_i}$ evolves as $\bar{P}_{y_i} = \bar{y}_i \cdot SE_i + (1-\bar{y}_i) \cdot (1-SP_i)$.

The test is consistent and asymptotically standard normal distributed under the null hypothesis. In case of no predictability, the values of sensitivity and specificity will sum up to one, meaning the numerator and consequently the test statistic will be zero. If a positive relationship between the forecasts and the realizations is present, it holds that $SE_i + SP_i > 1$ and hence $PT_i \to \infty$ for $T_F \to \infty$.

The disaggregated results of the forecasting procedure using the logistic regression model for each stock are shown in Table 5. It can be seen that the hitrate varies between 50.06 percent for Industrial Paper (IP) and 53.27 percent for General Electric (GE). With regard to the sensitivity and the specificity, it can be observed that the sensitivity typi-

cally outweighs the specificity, except for those stocks where the proportion of positive returns in the sample (the hitrate of the optimist forecast) is below 50 percent. The model therefore overpredicts the majority class. In comparison to the benchmark forecasts, the hitrate achieved with the logistic model is higher for the majority of stocks.
With regard to the Diebold-Mariano test, the null of equal predictive accuracy is rejected at the 10 percent level in 12 cases for the random walk benchmark, in 19 cases for the historical majority benchmark, and in 9 cases for the optimist benchmark.
When considering the results of the PT test, we find that the statistic is positive for all but one stock, indicating that there is some extent of predictability. This predictability is reported to be significant at the 10 percent level for 19 out of 26 cases. Overall, it can therefore be concluded that there is strong evidence for the statistical significance of directional predictability - even if individual stocks are considered.
In the last section we found that on average logistic regression produces better forecasts than the other classification methods considered. For pairwise comparisons among the different classifiers for each stock, we again consider the Diebold-Mariano test as well as an alternative procedure that is specifically designed for classification models. This is the well known test of DeLong, DeLong, and Clarke-Pearson (1988) that compares the AUCs of two competing classification models using the test statistic

$$W = \frac{AUC^{(1)} - AUC^{(2)}}{\sqrt{Var(AUC^{(1)} - AUC^{(2)})}}. \tag{8}$$

The AUC is the area under the *receiver operator characteristics* curve and a commonly used measure for classification models. It can be interpreted as the average sensitivity for all possible values of the specificity obtained by varying the values of the threshold in (6). The test statistic is standard normally distributed under the null hypothesis of equal predictive accuracy. For further details confer DeLong, DeLong, and Clarke-Pearson (1988).
Table 6 states the number of stocks for which the forecasts of the models stated in the respective row are significantly better than those of the model stated in the respective column at the 10 percent level.
The results show that the forecasts of the logistic regression model are not outperformed once by the forecasts generated by any other classification method. On the contrary, the logistic model forecasts significantly better than generalized additive models, neural networks, support vector machines, random forests, and boosted trees for almost half of the stocks.
As stated in the introduction, the general accepted opinion of academic financial economists is that daily stock returns are unpredictable, as implied by the EMH and the random walk hypothesis. In contrast to that, the predictability found here is in clear

RIC	*HR*	*SE*	*SP*	DM test RW	DM test HM	DM test Opt	PT test
AA	51.95	43.17	60.46	51.06	50.76	49.24**	2.11**
AXP	52.68	56.90	48.29	49.00***	50.27**	51.00*	2.98***
BA	52.14	55.44	48.62	49.26**	49.77**	51.66	2.33**
CAT	51.69	57.45	45.64	50.81	48.28***	51.23	1.78**
CVX	52.66	75.70	27.50	51.90	50.32**	52.21	2.09**
DD	50.75	57.75	43.27	50.38	48.37**	51.63	0.59
DIS	51.75	58.33	44.82	50.23*	50.11*	51.29	1.82**
EK	53.17	23.75	78.87	51.83	53.36	46.64***	1.41*
FL	52.18	48.52	55.90	49.28***	49.59**	50.41*	2.54***
GE	53.27	45.04	61.39	49.29***	50.35***	49.65***	3.73***
GM	52.41	27.59	74.92	50.03**	52.44	47.56***	1.53*
GT	50.78	44.24	57.45	49.71	49.47	50.53	0.98
HON	52.49	62.07	42.26	50.18**	50.40**	51.64	2.54***
IBM	52.33	51.94	52.73	50.84	50.05**	50.90	2.68***
IP	50.06	52.54	47.50	49.42	49.18	50.82	0.02
JPM	52.39	49.76	55.05	51.57	49.74**	50.26**	2.76***
KO	53.09	58.93	47.07	48.68***	48.92***	50.71**	3.47***
MCD	51.81	77.90	22.43	50.59	50.08*	52.96	0.23
MMM	52.23	64.78	38.51	49.29***	49.89**	52.23	1.96**
MO	53.27	78.60	24.00	49.80***	53.09	53.61	1.78**
MRK	51.41	47.58	55.28	50.02	48.80***	50.23	1.65**
PG	52.20	60.66	43.53	50.81	50.62*	50.62*	2.44***
S	50.49	71.14	30.77	49.84	51.15	48.85	0.36
T	50.85	51.16	50.53	49.30*	48.69**	51.31	0.97
UTX	51.06	61.74	40.22	51.13	49.30*	50.40	1.15
XOM	51.87	63.80	39.45	49.59**	50.44*	50.99	1.93**

Table 5: Disaggregated forecasting results separated by stock for the logistic regression model. RIC states the ticker code used by *Thomson Reuters* and all values, except the PT statistics, are given in percent. The panel in the middle shows the hitrate of the three benchmarks and the asterisks indicate whether the logistic forecasts are significantly better than the respective benchmark according to a one-sided DM test. Again, the symbols ***, (**), and [*] indicate that the null is rejected at the 1%, (5%), or [10%] level.

contradiction to the random walk hypothesis. Whether these findings also contradict the EMH cannot be determined without considering the effect of transaction costs. This will be investigated in Section 4.5, where trading strategies based on the directional forecasts are implemented.

	Logit	GAM	NN	SVM	RF	Boost	Σ
				DM test			
Logit	0	8	13	11	14	9	55
GAM	0	0	7	5	8	5	25
NN	0	1	0	2	3	1	7
SVM	0	1	4	0	5	4	14
RF	0	2	4	2	0	1	9
Boost	0	1	4	2	5	0	12
Σ	0	13	32	22	35	20	122
				W-Test			
Logit	0	7	10	14	13	12	56
GAM	0	0	5	7	8	10	31
NN	0	1	0	4	3	2	10
SVM	0	0	2	0	3	5	10
RF	0	2	3	4	0	4	13
Boost	0	1	1	2	2	0	6
Σ	0	11	21	31	29	33	126

Table 6: Disaggregated model comparison results separated by stock. Each field shows the number of stocks for which the forecasts of the model stated in the respective row are significantly better than the forecasts of the model stated in the respective column. Logit abbreviates Logistic Regression, GAM Generalized Additive Model, NN Neural Network, SVM Support Vector Machine, RF Random Forest, and Boost Boosted Trees.

With regard to the relative performance of the logistic regression and the machine learning methods, it can be concluded that the low signal-to-noise ratio leads to overfitting of the more complex models. This is in contrast to the findings of Leung, Daouk, and Chen (2000), who observe that neural networks outperform logistic regression for directional predictions. However, this finding is based on monthly returns where the signal-to-noise ratio is higher.

Apparently, on a daily horizon a linear parametric model contains sufficient useful information for predicting future returns. This linear relationship can be captured by all classification methods, but more flexible methods also adjust to non-linear, noise-driven relations. Since these relations do not resurface in the out-of-sample period, these methods produce worse forecasting results than logistic regression. In statistical terms, the increased bias of the linear model is outweighed by the decreased variance.

Overfitting problems are very common in the field of machine learning, even in cases where the signal-to-noise ratio is larger than in the situation considered here. Therefore, most of the classification methods considered already include penalty terms that are meant to prevent this issue.

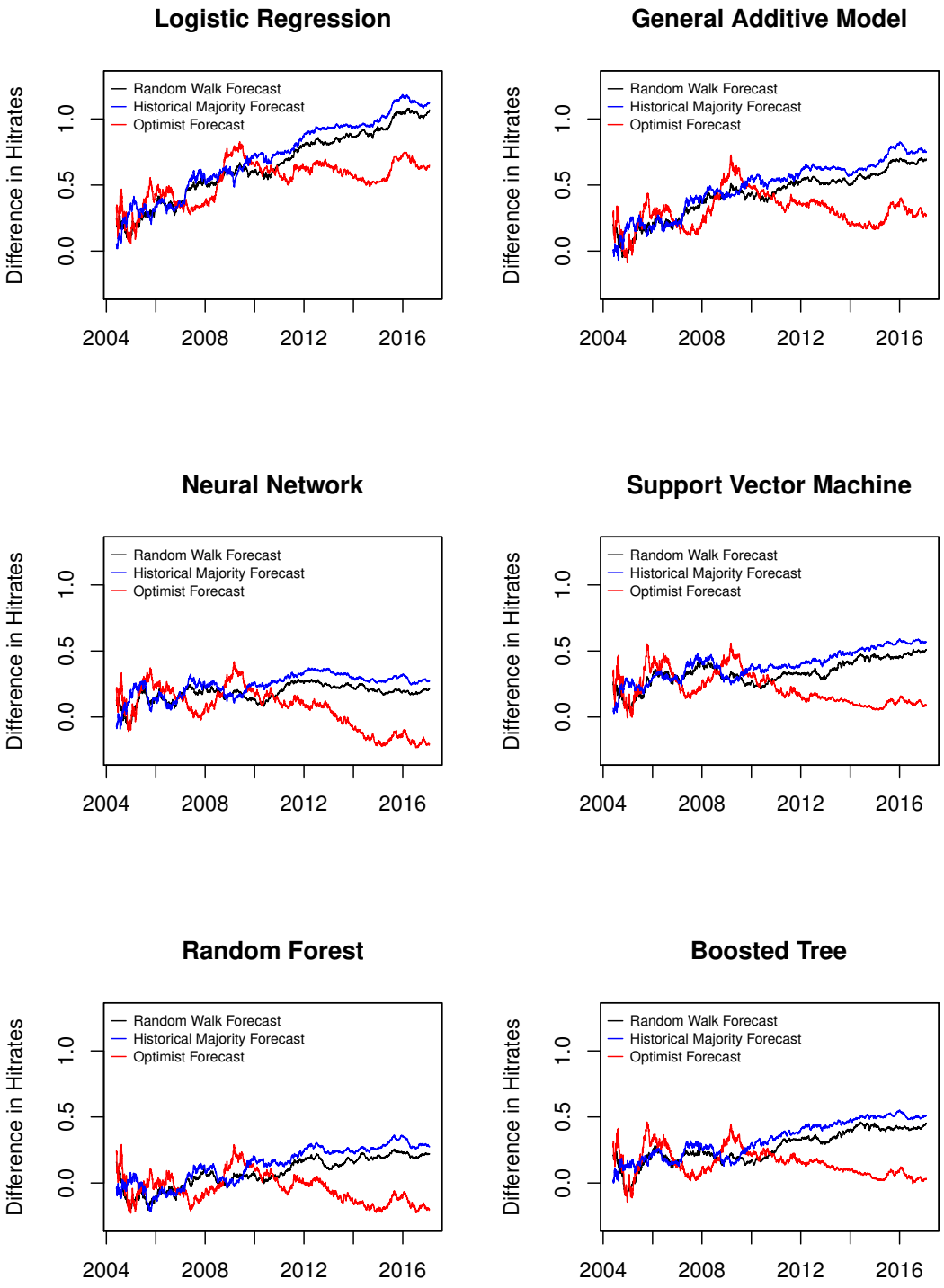

Figure 2: Rescaled cumulative difference between the hitrate of the classification model and the respective benchmark.

4.4 Stability of the Forecast Performance over Time

In the level predictability literature, it is often found that predictability is concentrated in recession periods. Welch and Goyal (2008), for example, find a short period of predictability after the oil price shock in 1973. We therefore investigate the stability of the forecasting performance of the classification models over time.

For this purpose, Figure 2 presents cumulated scaled difference plots as a graphical illustration. These are the classification equivalents to the cumulative plots given in Welch and Goyal (2008). The plots show the cumulated difference between the hitrate of the classification model and the benchmark. To ensure that the variance of the curve is stable over time, this difference is scaled by the square root of the time index.[7]

As argued in Welch and Goyal (2008), the level of these curves cannot be interpreted - but the slope. For periods with positive slopes, it holds that the directional forecasts outperform the benchmark and for periods with negative slopes the opposite holds. Therefore, the plots help to identify whether a model is superior compared to its benchmark for any chosen period by simply comparing the height of the curve at the beginning of the period with the height of the curve at the end of the period.

We observe that the blue and black curves comparing random walk forecasts and historical majority forecasts with the classification model forecasts are almost identical. For these two benchmarks, Figure 2 reveals a predominantly positive slope for all classification methods. In line with the findings in Section 4.2 and 4.3, this slope is especially strong for the logistic regression model and the generalized additive model and moderate for all other methods. Consequently, when comparing forecasts with these benchmarks, there is only one conclusion that can be drawn: the directional predictability of daily stock returns is consistent over time and not restricted to recessions (as it seems to be the case for level predictability at lower frequencies).

When considering the optimist benchmark, the picture changes slightly. For the first part of the forecast period from 2004 until 2010 the curves still exhibit a predominantly positive slope that is especially strong for the logistic regression model and the generalized additive model, as it was the case for the other benchmarks. However, when considering the time period after the subprime mortgage crisis from 2010 until 2017, we do not observe predominantly positive slopes anymore. For the logistic regression model slopes are alternating and for all other models we observe moderate negative slopes.

The reason for this behavior is rooted in the nature of the optimist forecast. While the hitrates of the forecasting models remain relatively stable over time, the hitrate of the optimist forecast is extremely low during the subprime mortgage crisis and relatively

[7] We also excluded the first 100 observations, as these were too variable leading to scales that distort the true relationship. Therefore, most curves do not start at zero.

high during the subsequent bull market period. The dynamic of the cumulated scaled difference plot is therefore dominated by that of the stock market.
Overall, we conclude that the performance of the directional forecasts is stable over time and not limited to specific time windows. It should further be noted that the forecast evaluations based on the hitrate do not discriminate between directional predictions that were assigned a high probability by the forecasting model and those where the assigned probability is close to 50 percent. This will be addressed in detail in the next section, where we consider the economic significance of the observed predictability.

4.5 Economic Significance

The previous sections show that signs of daily stock returns are, to some extent, predictable. The obvious question raised by this finding is whether this predictability is of a magnitude that is economically meaningful so that it can be exploited to generate abnormal returns. This section therefore addresses the design and performance of trading strategies based on sign predictability.
Directional forecasting is as an attempt to time the market. It is therefore obvious to buy stocks which are expected to have positive returns and to sell stocks which are expected to have negative returns. This is the basis for a long-short equity strategy that trades a value neutral portfolio with zero net-investment and where the market risk is hedged. There are, however, several a priori considerations that have to be addressed in the design of the trading strategy.
As mentioned before, the classification methods do not only produce a binary forecast $\hat{y}_{it}$ that specifies whether the stock is expected to have a positive or negative return. Instead, equation (5) gives an estimate of the probabilities $\hat{P}(y_{i,t+1} = 1|x_{i,t})$ that the stock return will be positive.
The plot on the left hand side of Figure 3 shows the density of these predicted probabilities for the logistic model in the pseudo out-of-sample part of the model selection period. It can be seen that the majority of the probabilities are in the neighborhood of 50 percent. Obviously, the hitrate that can be expected from directional forecasts based on these probabilities will also be close to 50 percent. Conversely, if we only consider predictions for which the probability to be positive is further away from 50 percent, then the hitrate will be higher. This can be seen on the right hand side of Figure 3 that shows the hitrate in the pseudo out-of-sample part of the model selection period for those stocks where the probability to be positive was at least v percent higher or lower than 50 percent. We refer to this distance v of the predicted probabilities from the 50 percent threshold as the *confidence* of the prediction. It is clear to see that the hitrate for those stocks that are predicted to have a probability to move up of at least

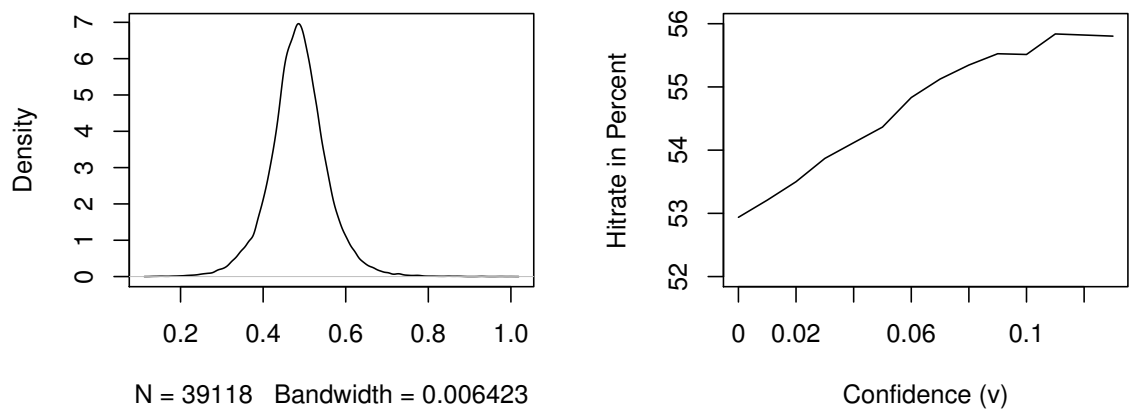

Figure 3: Model confidence and density plots for the logistic regression model: the left graph shows the density of the forecasted probabilities in the model selection period across all stocks, the right graph shows the hitrate in percent conditional on the difference of the forecasts from the threshold 0.5.

60 percent or at most 40 percent is nearly 56 percent.

Now, with regard to the design of the trading strategy, there is a trade-off between trading few stocks, for which the model generates a strong signal so that the hitrate and therefore the average return are high, and trading many stocks to reduce the variability of the portfolio return.

Furthermore, it should be noted that a strategy that requires daily trading will generate much higher transaction costs than a strategy with monthly portfolio re-balancing. It is therefore crucial to keep the portfolio turnover low and to carefully consider the effect of transaction costs on the performance of the portfolio.

It would be possible to consider a strategy where all stocks for which the probability of a positive return is higher than $50+v$ percent are bought and where all stocks for which the probability is lower than $50-v$ percent are sold. This could, however, cause situations where trading is suspended, because all stocks are predicted to have positive returns with a probability of more than $50-v$ percent so that the short portfolio is empty. The same would be possible for the long portfolio. These effects can arise even though the dispersion among the predicted probabilities is high so that we would have a strong signal to trade on. To avoid this effect, the strategy is formulated in terms of the difference w in the probability of up-movements between pairs of stocks.

Assume, without loss of generality, that there is an even number of stocks N. The strategy is then implemented as follows:

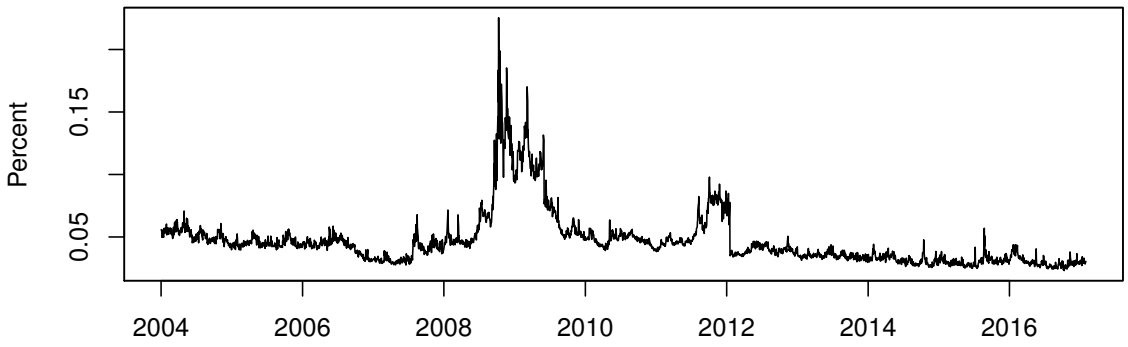

Figure 4: Plot of the average daily bid-ask spread in percent.

For each day $t = 1, ..., T_F$:

1. Sort all stocks in ascending order according to the predicted probabilities $\hat{P}(y_{i,t+1} = 1|x_{i,t})$ from (5) and denote the rank by $r = 1, ..., N$.

2. Form pairs of stocks so that the stock with the lowest probability to have a positive return ($r = 1$) and the stock with the highest probability ($r = N$) are matched together, the stock with the second lowest and highest probability ($r = 2$ and $r = N - 1$) are matched together, and so on.

3. For all m pairs where the difference between the probabilities to have a positive return is at least w percent, buy the stock that is more likely to go up and sell the stock that is less likely to go up.

This strategy allows us to trade on the cross-sectional dispersion in the predicted probabilities for positive returns. If the probabilities are close to 50 percent for all stocks, trading is suspended. If there is a strong signal for a low number of stocks, the number of pairs that is traded is kept low, and if the signal is strong for a large number of stocks, the traded portfolio contains a large number of stocks.
As mentioned above, when evaluating the performance of our trading strategy, it is critical to account for the effect of transaction costs. It is common practice to assume that transaction costs are constant over time and to deduct a constant fee from each return, as for example in Pesaran and Timmermann (1995) or Bajgrowicz and Scaillet (2012). However, if for example predictability is concentrated in more volatile markets where transaction costs tend to be higher, it is likely that this leads to an overestimation of the trading returns.

We therefore adopt a different approach and consider actual bid-ask spreads that account for the largest proportion of transaction costs. It should, however, be mentioned that there are still some limitations. Our strategy requires shortselling and additional costs associated with this are ignored. Furthermore, we do not consider the impact that trading the strategy with a larger portfolio itself would have on the market, which traders refer to as *slippage*. On the other hand, the bid-ask spread might overestimate the actual level of transaction costs, since it basically implies that we always buy at market, while in practice a more sophisticated order management might avoid crossing the spread for every trade.

Figure 4 shows the evolution of the average bid-ask spread for the stocks in our sample over the trading period from 2004 to 2017. To guarantee comparability across stocks, this measure is calculated by dividing the bid-ask spread through the price of the stock before averaging the percentages across stocks. The figure reveals that the average level of transaction costs is decreasing from around 0.05 percent in the beginning of the period to about 0.03 percent in recent years. However, the bid-ask spreads were much higher during the subprime mortgage crisis in 2008 and 2009 and again in 2011. In 2009 the average bid-ask spread was as high as 0.2 percent and in 2011 as high as 0.1 percent. We therefore conclude that this approach allows us to obtain much more realistic results on the impact of transaction costs.

Table 7 reports the trading returns of the strategy outlined above with $w = (0.13,\ 0.15,\ 0.17,\ 0.19)$. All results are based on the forecasts of the logistic regression model. In addition to mean and standard deviation of the trading returns, Table 7 presents several commonly used performance measures.

For mean-variance investors, we state the Sharpe ratio and the realized utility gain. For the latter, we follow Rapach, Strauss, and Zhou (2010) and state the measure in excess to holding the S&P 500 and assuming a risk aversion parameter that equals three. The table further reports CAPM-alphas as well as alphas for the five-factor model considered by Fama and French (2015).[8]

As a benchmark, we report performance measures for buying and holding the S&P 500. This benchmark doubled between 2004 and 2017 and is closely related to the optimist forecast, since an investor that predicts a positive return for every stock and every trading day could simply buy and hold the index portfolio.

The table reveals that even after accounting for bid-ask spreads, for all considered values of w, all performance measures indicate superior performance of the trading strategy compared to the benchmark.

For $w = 0.15$, for example, the average return of the trading strategy after accounting

[8]Returns of the factors are obtained from Kenneth French's website at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html.

	Mean	SD	SR	Δ Utility	α CAPM	α (5-factors)
Benchmark (S&P 500)	0.029	1.178	0.024	0.000	-	-
				$w = 0.13$		
Daily	0.056	0.967	0.058	0.253	0.060***	0.057***
Daily (Bid-Ask Costs)	0.025	0.966	0.026	0.224	0.029**	0.026*
				$w = 0.15$		
Daily	0.073	1.128	0.065	0.102	0.078***	0.074***
Daily (Bid-Ask Costs)	0.045	1.118	0.041	0.085	0.050***	0.046***
				$w = 0.17$		
Daily	0.069	1.118	0.062	0.109	0.072***	0.070***
Daily (Bid-Ask Costs)	0.045	1.107	0.041	0.097	0.048***	0.046***
				$w = 0.19$		
Daily	0.057	1.115	0.051	0.100	0.060***	0.056***
Daily (Bid-Ask Costs)	0.038	1.107	0.034	0.090	0.041**	0.037**

Table 7: Performance measures for the trading strategy based on the logistic forecasts. All measures are calculated using the daily percentage return of the strategy. SD states the standard deviation and SR the Sharpe ratio, which is calculated without considering the risk-free rate since this is marginal on a daily basis. Moreover, Δ Utility states the realized utility gain of a mean-variance investor with a risk aversion parameter of three in excess to holding the S&P 500. The symbols ***, (**), and [*] indicate that alpha is significantly larger zero at the 1%, (5%), or [10%] level.

for transaction costs is 0.045 percent per trading day, whereas that of the benchmark is only 0.029 percent. At the same time, the standard deviation of 1.118 percent is lower than that of the benchmark, which is 1.178 percent. Consequently, the Sharpe ratio is larger and the utility gain is positive. Cumulated over the trading period, the trading strategy yielded a return of 367 percent after transaction costs, whereas the benchmark only increased by 205 percent during the same period.

The magnitude of these results shines a different light on the extent of directional predictability. We discussed above that the hitrate can be expected to be significantly higher for those days where the model predicts a probability for a positive return that is further away from 50 percent. In fact, the hitrate for all stocks traded by the strategy with $w = 0.15$ is as high as 54.09 percent, which explains the magnitude of the differences in the mean returns shown in Table 7.

Most importantly, however, a positive alpha is indicated by the CAPM as well as the five-factor model for all considered values of w and it is found to be statistically significant. This implies that the trading returns cannot be explained by any of the established risk factors.

To gain further insights into the performance of the trading strategy, Table 8 shows the

w		Long		Short	
		Mean	SD	Mean	SD
0.13	Daily	0.093	1.700	0.020	1.997
	Daily (Bid-Ask Costs)	0.067	1.691	-0.018	2.000
0.15	Daily	0.102	1.995	0.045	1.998
	Daily (Bid-Ask Costs)	0.079	1.967	0.012	2.001
0.17	Daily	0.092	1.973	0.045	1.916
	Daily (Bid-Ask Costs)	0.072	1.942	0.018	1.920
0.19	Daily	0.071	1.907	0.043	1.848
	Daily (Bid-Ask Costs)	0.055	1.879	0.021	1.858

Table 8: Trading returns separated for the long and short side of the trading portfolio. The "Long" column shows the daily percentage return and its standard deviation for long investments only, and conversely, the "Short" column shows the respective values for short investments.

	Return				Proportion			
m/w	0.13	0.15	0.17	0.19	0.13	0.15	0.17	0.19
0	0.00	0.00	0.00	0.00	0.21	0.36	0.52	0.64
2	0.03	0.10	0.12	0.13	0.34	0.35	0.30	0.23
4	0.05	0.12	0.17	0.23	0.21	0.14	0.10	0.06
6	0.10	0.06	0.17	0.19	0.11	0.06	0.04	0.03
8	0.14	0.14	0.06	0.25	0.05	0.03	0.02	0.01

Table 9: Average percentage return for all days where m stocks are traded and the proportion of days for which this is the case.

average return and the standard deviation for the stocks that are bought and those that are sold. While the variation of the return of the long and short side is similar (except for $w = 0.13$), the mean return achieved on the stocks that are bought is considerably higher than that on the stocks that are sold.

These findings indicate that the model is more successful in predicting which stock prices are likely to increase than those who are likely to decrease, which is in line with the findings in Section 4.4. Nevertheless, with one exception, the return after accounting for transaction costs is positive so that the short side still contributes to the portfolio return and does not only work as a hedge against market risk.

Table 9 gives further details on the effect of changing the dispersion between the predicted probabilities for a positive return that is required to initiate a trade. As can be expected, the higher the required dispersion w, the higher the proportion of days where a low number of pairs is traded. Furthermore, if we consider all days where some number m of stocks is traded, the average return on these days tends to be higher, the higher

the dispersion parameter w. This again underlines that the trading strategy works as intended.

As discussed in the introduction, the efficient market hypothesis in its weakest form allows for some extent of predictability, as long as this is of a magnitude that cannot be exploited after accounting for transaction costs.

The findings presented in this section clearly show that the opposite is the case. Even after accounting for transaction costs, we obtain higher Sharpe ratios than the S&P 500 and significant alphas as measured by the 5-factor model of Fama and French (2015). The positive alphas suggest that no risk-based explanation for the generated returns exists. It should be emphasized that the results presented here are based on a relatively small cross section of 30 stocks or less. A trading strategy based on a larger asset universe can be expected to produce even more dispersion among the predicted probabilities and therefore higher trading returns.

Overall, we find that there is some market inefficiency in form of an overlooked non-linear dependence that is economically meaningful.

4.6 Performance of the Trading Strategy over Time

As mentioned in Section 4.4, previous studies find that predictability is concentrated in certain periods such as recessions. This is also backed by theoretical arguments, for example those of Timmermann and Granger (2004) and Lo (2004), who argue that rational investors will pick up on emerging patterns and exploit them so that it cannot be expected that any forecasting patterns will persist for long periods of time. Therefore, this section investigates the performance of the trading strategy over time focusing on the case with $w = 0.15$. The results for other values of w are qualitatively similar.

Figure 5 presents cumulative plots of the trading returns of our strategy on the left hand side and the difference between the cumulated return of the trading strategy and the buy-and-hold strategy on the right hand side. In contrast to the plots in Figure 2, the curves in Figure 5 are not scaled since they do not represent means and therefore exhibit equal variance over time. Therefore, we cannot only interpret the slope of these curves but also the level.

Both curves in the left plot exhibit a predominantly positive slope, which indicates that the model picks up on a signal that is consistently present over time. However, there are brief periods with extremely high slopes, which means that the magnitude of the signal must be increased during these periods. This concerns mostly the period from 2008 to 2009 and that from 2014 onwards.

The jump on October 03, 2011 is due to a very large return of 57.17 percent for *Eastman Kodak* that was correctly predicted by the model. *Eastman Kodak* went through sub-

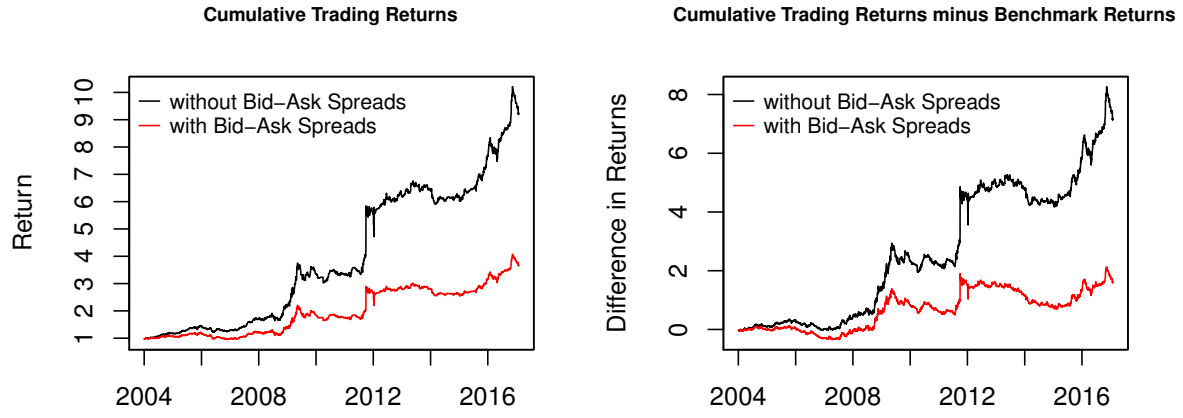

Figure 5: Cumulative trading return of the directional strategy (left) and cumulative difference between trading returns and the buy-and-hold strategy (right) for $w = 0.15$.

	Level	p-Value
Intercept	0.157	0.000
Probability of Bull Market	-0.202	0.000

Table 10: Regression results for the daily trading return explained by the probability of a bull market estimated by a Markov switching mean-variance model.

stantial turbulences in the autumn of 2011, before eventually going bankrupt on January 19, 2012. This is the cause for some of the large movements that can be seen in the plots around this time and also explains the increasing magnitude of the average bid-ask spread in 2011 that can be seen in Figure 4. A robustness check presented in Section 6 shows that the predictability persists even if *Eastman Kodak* is completely excluded from the sample.

The plot on the right hand side of Figure 5 looks quite similar to that on the left hand side, although the comparison to the benchmark leads to short periods of negative slopes, especially when considering transaction costs.

Since the slopes of the curves are more pronounced in the period from 2008 to 2009, one might suspect that the strategy performs better in times of crises. This would be in line with the finding that level predictability appears to be concentrated in recessions, as discussed above. However, since our analysis relies on high frequency data, there are not enough recession periods in the sample to test this hypothesis. Instead, Table 10 presents the result of a regression of the daily trading return of our strategy on the probability of a bull market obtained through a Markov switching mean-variance model with two regimes corresponding to a bull and a bear market. Since the coefficient of

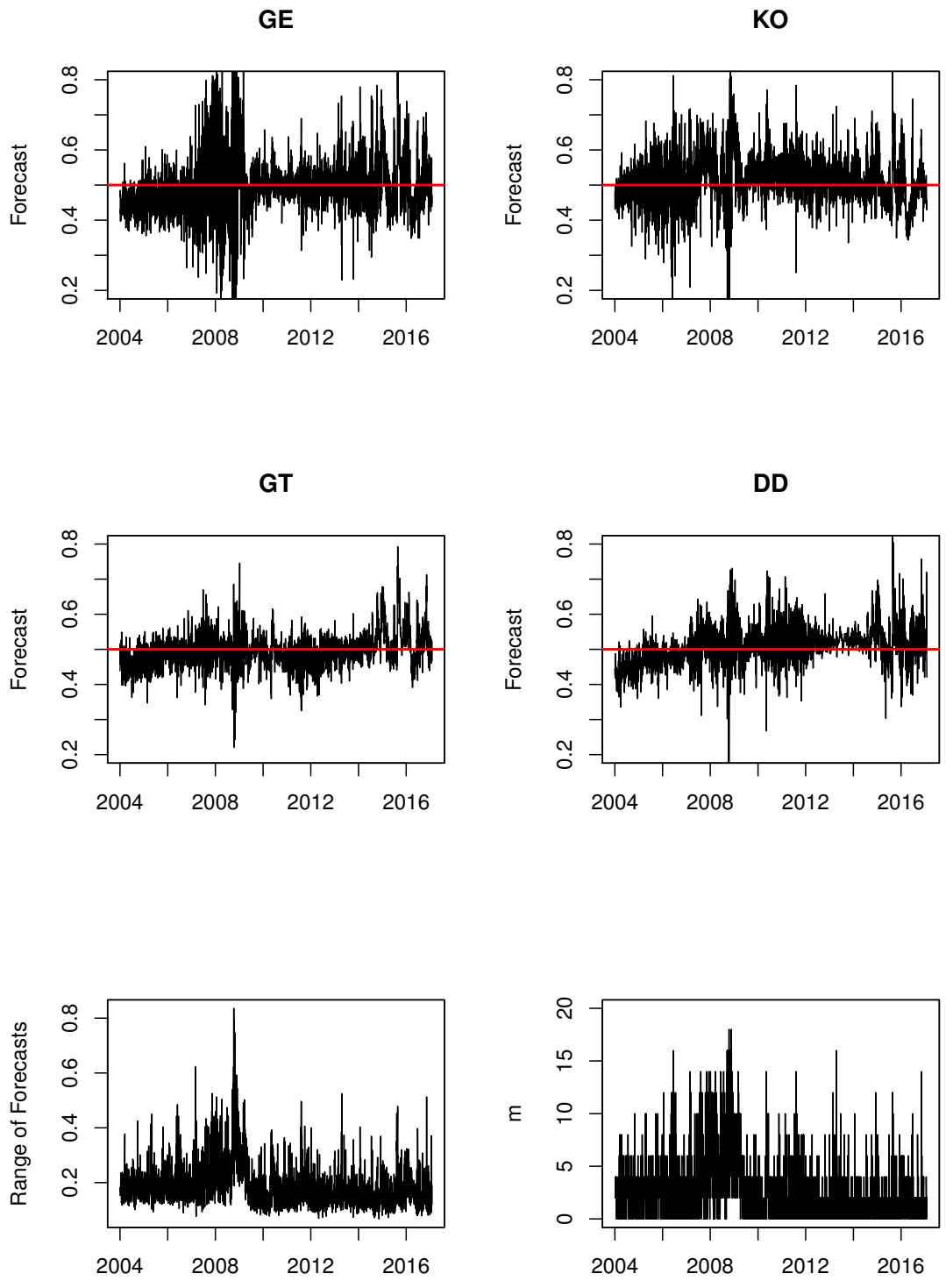

Figure 6: Forecasts of the logistic regression model for four stocks (top), range of the predicted probabilities for positive returns across all stocks (bottom left), and number of stocks m traded by the strategy with $w = 0.15$ (bottom right).

the bull market probability is negative, the trading returns are indeed higher in bear markets. It should be noted, however, that the intercept is still significantly larger than zero so that the market sentiment alone is not able to explain the average returns generated by the trading strategy.

Further results on the behavior of the trading strategy and the forecasts are shown in Figure 6. The first two rows of the figure show the predicted probabilities from equation (5) exemplarily for four stocks. It can be seen that the model tends to predict a probability for a positive return that is close to 50 percent for most of the time. However, the variation in the predicted probabilities changes significantly over time and there is some degree of persistence in the deviations from 50 percent. It is interesting to note that the variation in the predicted probabilities seems consistently higher for *General Electric* (GE) and *Coca-Cola* (KO) than for *Goodyear Tire* (GT) and *Du Pont* (DD). When considering the results on the significance of the observed predictability in Table 5, it is found that GE and KO, for which the model generates stronger variation in the predicted probabilities, are indicated to be significantly predictable by all tests, whereas GT and DD are not.

The overall dispersion of the forecasts across all stocks is considered in the bottom left plot. The figure reveals that the model differentiates especially strong between potential winner- and loser stocks during the subprime mortgage crisis. An effect that can also be seen, although less obvious, in the four plots above. Consequently, the trading strategy invests in a larger number of stocks during this period as the dispersion across stocks is larger. This is confirmed in the bottom right graph of Figure 6, where the number of stocks m in which the strategy invests is plotted over time. This behavior corresponds to the more pronounced slopes in the period from 2008 to 2009 in Figure 5.

Overall, the results suggest that the trading strategy produces positive returns relatively consistently. These are especially high in times when the market performs bad. During these periods the inefficiency becomes stronger so that the model is able to differentiate better between potential winner and loser stocks. In consequence, a larger portfolio is traded.

With regard to the overall performance of the forecasts and the trading strategy, we conclude that the logistic model picks up on an inefficiency that is moderate for most of the time, and for several periods it is so small that it cannot be exploited due to transaction costs. However, the predicted probabilities by the model itself account for this effect and indicate when the signal becomes stronger so that it can be successfully exploited.

	Mean	Min	Median	Max	Positive
S&P 500 Return	-0.03	-0.13	-0.04	0.13	0.35
VIX Return	-0.01	-0.08	-0.01	0.07	0.46
Variance Premium	0.03	-0.01	0.05	0.10	0.96
Level Third PC (Curvature)	-0.03	-0.09	-0.03	0.01	0.08
Stock Return	-0.06	-0.21	-0.06	0.08	0.27
5-Day Moving Average Return	-0.05	-0.15	-0.04	0.04	0.08
A/O Oscillator	0.08	-0.04	0.08	0.22	0.77

Table 11: Summary statistics of the coefficients in the logistic regression model for all 26 stocks estimated for the full sample. The column "Positive" states the percentage of stocks for which the variable is estimated to have a positive impact on the probability of a positive return.

5 Interpretation of the Logistic Regression Model

So far, the forecasting models were treated as a black box since it is not necessary to consider the actual form of the model to establish the presence of predictability and its economic significance in an out-of-sample experiment. However, from an economic perspective, it is interesting to gain insights into the sources of predictability. In the following, we therefore analyze the influence of the selected variables on the predicted probability of a positive return on the next trading day. This can only be done for the logistic regression model. The other classification methods considered lack this kind of interpretability, which is often named as one of their main disadvantages.

The logistic regression model, however, can be interpreted as a linear model for the log-odds and the sign of the estimated coefficients indicates whether an increase of the respective explanatory variable leads to an increase or decrease in the probability to observe a positive return on the next day.

Due to the rolling window procedure and the fact that the estimated coefficients are allowed to vary across stocks, it is difficult to give a comprehensive overview of the direction of the effects. We therefore report the results of a full sample estimation of the selected model over the period from 1996 to 2017. This gives 26 different models - one for each stock. Due to issues with the non-stationarity of some of the regressors, standard inference is not applicable so that we cannot test whether the estimated effect for a specific regressor is significantly different from zero. Instead, we focus on the interpretation of the signs.

Table 11 shows summary statistics for the estimated coefficients. Since the variables are standardized to have zero mean and unit variance prior to the estimation of the models, the magnitude of the effects is comparable across variables. The last column on

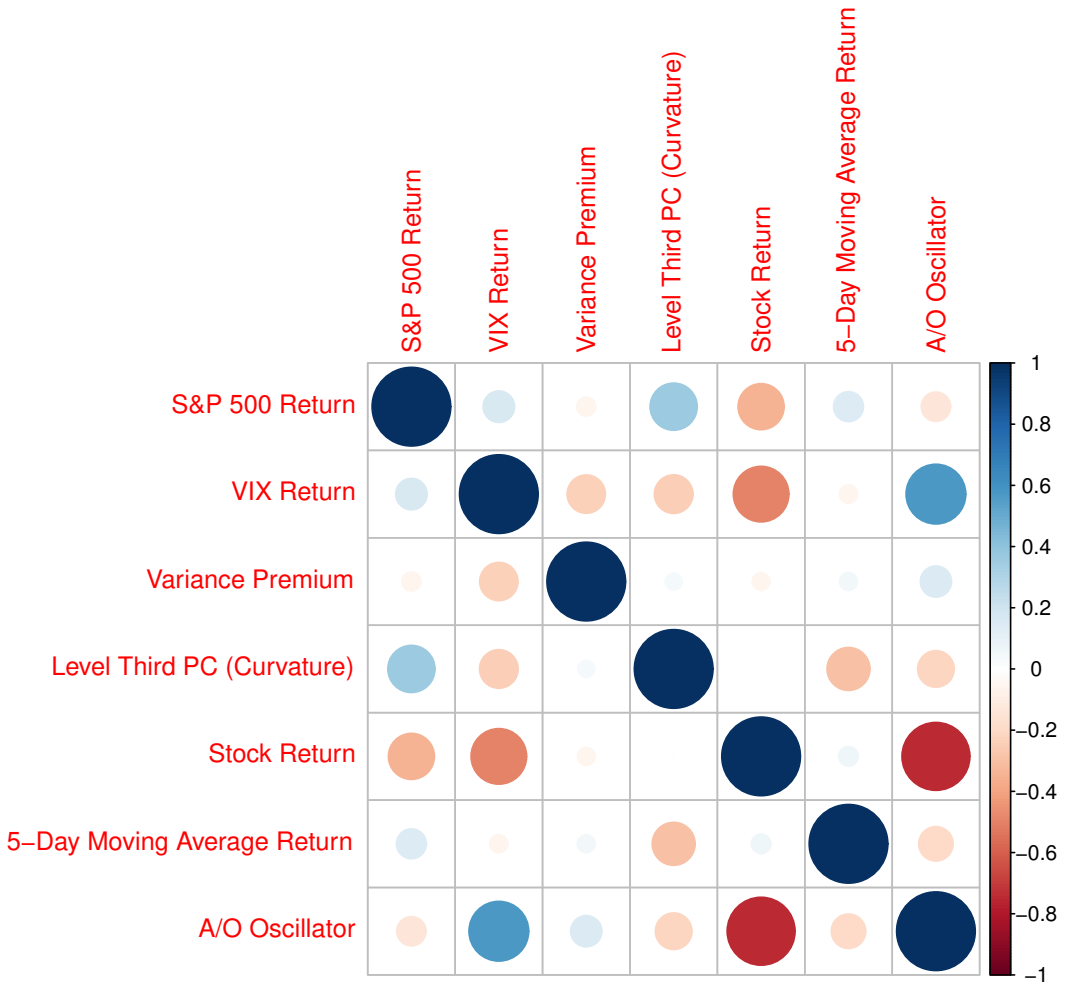

Figure 7: Correlation plot for the estimated coefficients from the logistic regression model for the different stocks. Size and color of the circles correspond to degree and direction of the correlation.

the right shows the proportion of the estimated coefficients for the respective variable that is positive. It is found that the effects of the A/O oscillator and the variance premium are predominantly positive. Conversely, both the 5-day moving average return and the curvature of the yield curve have a negative influence for nearly all stocks. Both the lagged return of the S&P 500 and the lagged return of the stock itself have a positive impact on the probability to observe a positive return on the following day for approximately a third of the stocks and a negative impact for two thirds of them. The percentage change of the VIX has a positive impact for approximately half of the stocks.

We therefore find that on average positive returns are more likely if the trading range on the previous day and the variance premium are high, and if the previous returns and the curvature of the yield curve are low. If one restricts the analysis to the model selection

				DM test		
	HR	SE	SP	RW	HM	Opt
Modelselection until 2002	51.77	53.61	49.85	50.27***	49.91***	50.88**
Modelselection until 2004	52.00	54.07	49.86	49.91***	50.10***	50.88**
Initial Model Selection Window 250	50.98	52.89	49.02	50.13***	50.03***	50.86
Initial Model Selection Window 750	51.93	56.41	47.29	50.13***	50.03***	50.86**
Forecasting Window 750	51.76	56.34	47.02	50.13***	50.03***	50.86**
Forecasting Window 1,250	51.70	53.05	50.30	50.13***	50.03***	50.86*
Excess Return Forecasting	51.85	54.85	48.76	50.14***	50.00***	50.82**

Table 12: Aggregated forecasting results for a number of alternative specifications that are considered as robustness checks. As before, all values are given in percent and represent the average over all forecasts and the symbols ***, (**), and [*] indicate that the null is rejected at the 1%, (5%), or [10%] level.

period, the results remain qualitatively similar.

The correlation of the estimated coefficients for the different stocks is given in Figure 7. It can be seen that there is a negative correlation among the coefficient on the lagged stock return and that of the A/O oscillator and the change of the VIX. This means that stocks that have a higher probability of a positive return after a negative return also tend to have higher returns if the previous day's trading range and the change of expected risk are high. In line with this, there is a positive relationship between the coefficient of the A/O oscillator and that of the return of the VIX.

6 Robustification

The previous sections establish that there is statistically as well as economically significant directional predictability in daily stock returns. However, the methodology established in Section 3 involves a number of ad-hoc choices, most notably the relative size of the model selection and the forecasting period, the size S of the initial window in the model selection period in (4), and the size W of the rolling estimation window for the forecasting model in (5).

In the following, we therefore consider the robustness of the forecasting performance with respect to these modeling choices. The results of this exercise are shown in Table 12, that is analogous to Table 3, and in Figure 8, that repeats the analysis in Figure 2. First, we consider changing the length of the model selection period so that the window ends in 2002 or 2004 instead of 2003. In theory, a longer model selection period should be beneficial for the model selection procedure as more observations are available and consequently the results get more stable. However, a longer model selection period

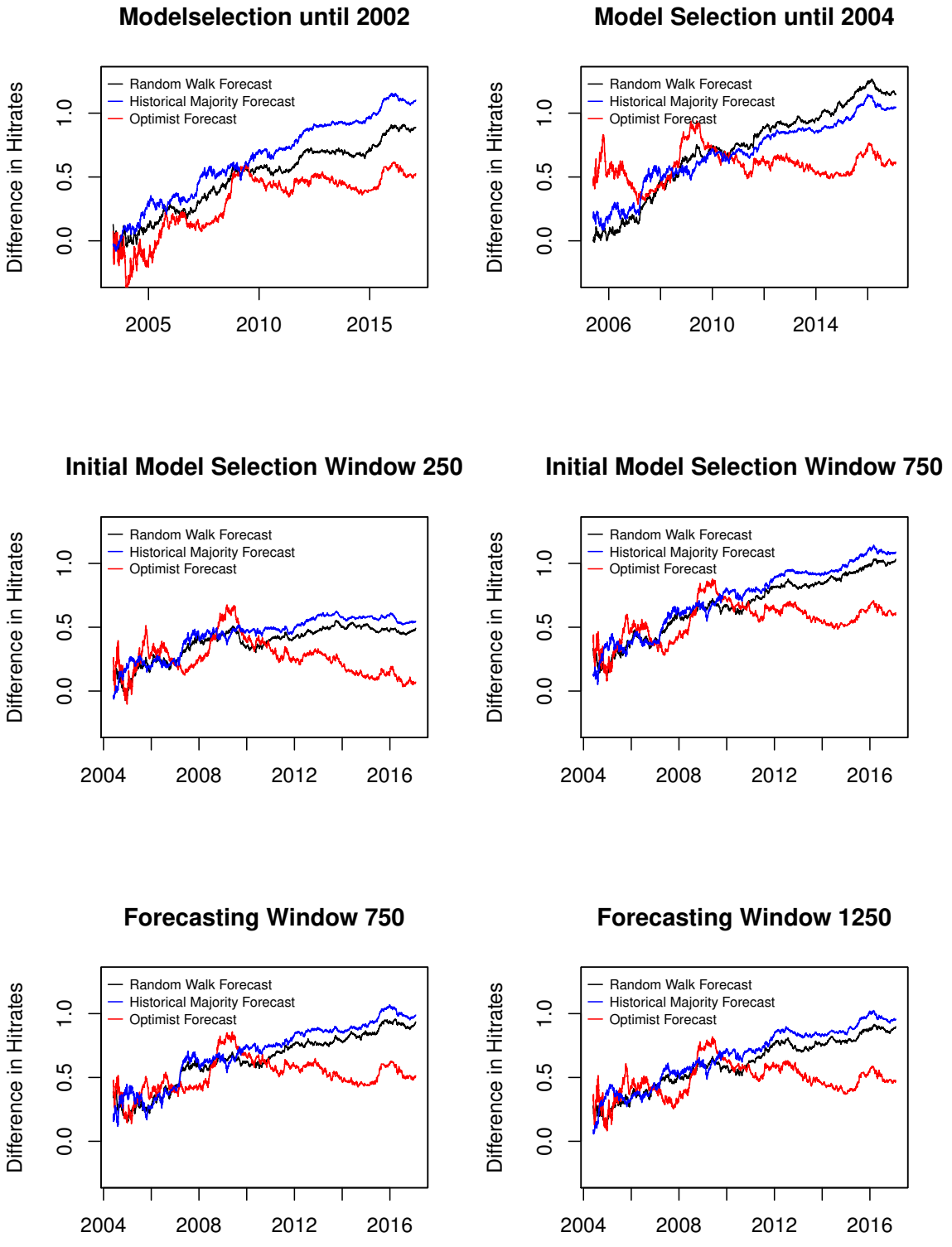

Figure 8: Cumulative hitrate plots in analogy to Figure 2.

means a shorter forecasting period so that less observations are available to evaluate the actual out-of-sample performance of the models.

For a model selection period from 1996 until the end of 2002 the procedure selected exactly the same model as reported in Section 4.1. Consequently, all results presented in Section 4.2 also hold if the model selection period is one year shorter. The hitrate changes slightly, however, due to the additional year in the forecasting period.

Adding another year, i.e. performing model selection from 1996 until the end of 2004, results in a slightly different model. The variables oil return and the 12-day moving average of the y_{it} are now included in the model with all other variables being the same. As the top right graph in Figure 8 shows, this has only a slight impact on the forecasting performance. The curves are still predominantly upward sloped indicating superior performance of the forecasts. Furthermore, the hitrate of 52.00 percent reported in Table 12 is almost exactly the same as in the original setup.

Second, we change the length S of the initial window in the model selection procedure from 500 to 250, respectively 750. In general, smaller values of S lead to less stability of the initial estimates, but there is also a larger number of pseudo out-of-sample observations available to select the variables. The right graph in the second row of Figure 8 reveals that increasing this number to 750 has a moderate effect on the forecasts. The selected model stays the same with the exception that the variable VIX return is replaced by the change of the third PC (curvature). Consequently, graph and hitrate (51.93 percent) are almost identical to the ones in the initial setup. This, however, does not hold when decreasing the size of the initial window S to 250. The results for this specification remain slightly better than the optimist benchmark, but between 2010 and 2017 a predominantly negative slope is observed when comparing with the optimist forecast.

In the third row of Figure 8, we changed the length W of the estimation window for out-of-sample forecasting in (5) from 1,000 to 750, respectively 1,250. As before, the graphs remain qualitatively similar, despite these changes.

As argued in the introduction, returns and excess returns are essentially the same on a daily horizon since the magnitude of the risk-free rate is marginal. The analysis in this paper is therefore conducted directly for the log-returns. To judge the impact of this modeling choice, we repeat the analysis using excess returns. As can be seen from Table 12, this delivers results that are very similar to those presented in Section 4.2. The overall hitrate changes only slightly to 51.84 percent.

Finally, we need to address two issues related to the economic significance of directional predictability. First, we calculated all performance measures assuming that stocks can be shorted at any time. In reality this was not the case, as there was a short time period of one month in 2008 where short selling was prohibited for most stocks by the

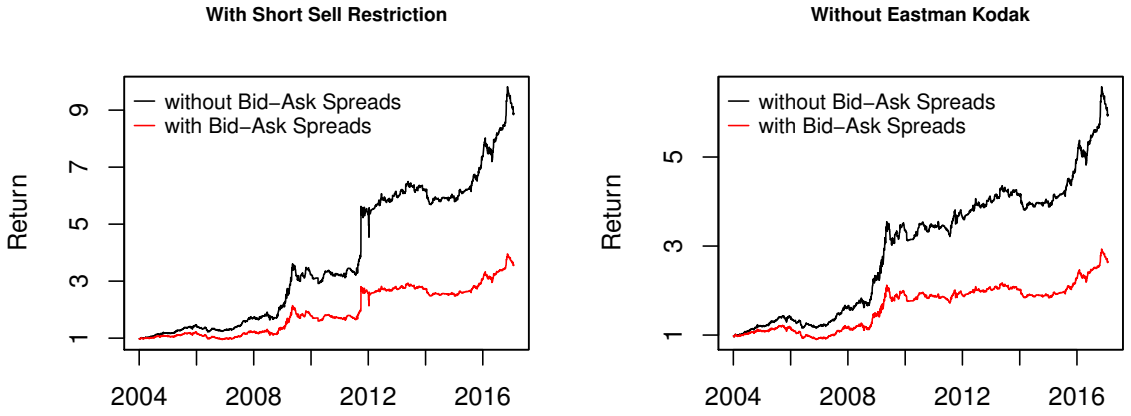

Figure 9: Cumulative trading return plots in analogy to Figure 5.

	Mean	S.D.	SR	Δ Utility	α CAPM	α (5-factors)
Benchmark (S&P 500)	0.029	1.178	0.024	0.000	-	-
	complying to the Short Sell Restriction in 2008					
Daily	0.073	1.128	0.064	0.102	0.078***	0.074***
Daily (Bid-Ask Costs)	0.045	1.117	0.040	0.085	0.050***	0.046***
	excluding *Eastman Kodak*					
Daily	0.057	0.779	0.074	0.419	0.060***	0.058***
Daily (Bid-Ask Costs)	0.033	0.777	0.042	0.396	0.036***	0.033***

Table 13: Performance measures for the trading strategy with $w = 0.15$ when complying with the short-sell restriction, or when excluding *Eastman Kodak* from the sample. The performance measures are calculated in analogy to those in Table 7.

Securities and Exchange Commission (SEC). Since our trading returns are particularly high during the subprime mortgage crisis, we repeat the analysis of trading returns for $w = 0.15$ under omission of this time period. The results are shown in the left graph of Figure 2 that shows the cumulative return of the trading strategy. It is clear to see that the strategy still consistently generates positive returns. The good performance of the strategy during the subprime mortgage crisis can therefore not be attributed to inefficiencies induced by the short-sale ban. This is also supported by Table 13, which reports only very small changes in the performance measures with an α that is still significant after excluding this period.

Second, we mentioned in Section 4.5 that the turbulences of *Eastman Kodak* in 2011 and 2012 severely influenced the performance of the trading strategy during this time. Therefore, we also report results for $w = 0.15$ when *Eastman Kodak* is completely ex-

cluded from the sample. The spike in 2011 is now absent from the time series plot of the cumulated trading returns shown in Figure 9. Instead, the figure shows a smooth positive trend, highlighting the stability of the trading strategy.
Moreover, the performance measures stated in Table 13 indicate that it might even be beneficial to exclude distressed stocks with large idiosyncratic risk from the pool of stocks traded by the strategy. Excluding *Eastman Kodak* leads to a lower average return, the decrease in standard deviation is, however, more pronounced. Therefore, the Sharpe ratio and utility gain reported are larger than those in the original setup.
Overall, it can be concluded that the results gained in the original setup are robust to changes in the selected parameters. This is also supported by the results of the Diebold-Mariano test in Table 12 that (with one exception) still rejects the null of equal predictive ability for all benchmarks.

7 Conclusion

Daily stock returns are generally regarded as unpredictable. This is theoretically implied by the efficient market hypothesis or consumption based asset pricing models and empirically well established for the level of daily stock returns. However, the results presented here show that the direction of daily stock returns is, to some extent, predictable. This predictability is shown to be statistically significant in an out-of-sample environment and of a magnitude that is economically meaningful so that it can be exploited by suitable trading strategies. These findings are in contradiction to the random walk hypothesis and the EMH implying that there is some degree of market inefficiency. While for most of the time the predictability seems to be small enough so that it cannot be exploited after accounting for transaction costs, there are periods, such as the subprime mortgage crisis or the recent years from 2014 onwards, during which the predictability intensifies and seems to constitute an actual market inefficiency.
Among the classification methods considered here, the parametric logistic regression model performs better than more flexible non-linear or even non-parametric models. However, the logistic regression model is not designed to address the properties of stock returns in an efficient manner, and it is likely that a more suitable statistical approach can generate even better forecast results. Furthermore, the trading strategies proposed in Section 4.5 trade on the dispersion between the predicted probabilities for expected returns, but the cross section of the DJIA data set considered is relatively small. It is therefore likely that the trading performance can be further improved by considering a larger asset universe.
While the degree of predictability is much lower than that found in previous studies in the machine learning literature, it is still clear that there is some form of non-linear

dependence in daily stock returns. These results may be unexpected from a theoretical perspective, but they are clearly in line with those of Linton and Whang (2007) and Han et al. (2016), who find evidence for directional predictability of daily returns based on non-parametric tests. We therefore conclude that directional forecasts are a promising field for future research on the predictability of stock returns - not only on a monthly horizon, as found by previous contributions such as Leung, Daouk, and Chen (2000), Nyberg (2011), Pönkä (2017), or Nyberg and Pönkä (2016), but also on a daily horizon.

Acknowledgements

We want to thank the finance group at the University of Hannover and the participants of the Statistical Week 2017 and the CFE-CMStatistics 2017 for their helpful comments. Furthermore, we are grateful for financial support from the University of Hannover through the program "Wege in die Forschung".

References

Amaya, D., P. Christoffersen, K. Jacobs, and A. Vasquez (2015). "Does realized skewness predict the cross-section of equity returns?" In: *Journal of Financial Economics* 118(1), pp. 135–167.

Ang, A. and G. Bekaert (2007). "Stock return predictability: Is it there?" In: *The Review of Financial studies* 20(3), pp. 651–707.

Ang, A., M. Piazzesi, and M. Wei (2006). "What does the yield curve tell us about GDP growth?" In: *Journal of Econometrics* 131(1), pp. 359–403.

Arlot, S., A. Celisse, et al. (2010). "A survey of cross-validation procedures for model selection". In: *Statistics surveys* 4, pp. 40–79.

Bajgrowicz, P. and O. Scaillet (2012). "Technical trading revisited: False discoveries, persistence tests, and transaction costs". In: *Journal of Financial Economics* 106(3), pp. 473–491.

Barndorff-Nielsen, O., P. Hansen, A. Lunde, and N. Shephard (2009). "Realized kernels in practice: Trades and quotes". In: *The Econometrics Journal* 12(3), pp. 1–32.

Bollerslev, T., G. Tauchen, and H. Zhou (2009). "Expected stock returns and variance risk premia". In: *The Review of Financial Studies* 22(11), pp. 4463–4492.

Bollerslev, T., J. Marrone, L. Xu, and H. Zhou (2014). "Stock return predictability and variance risk premia: statistical inference and international evidence". In: *Journal of Financial and Quantitative Analysis* 49(3), pp. 633–661.

Breiman, L. (2001). "Random forests". In: *Machine learning* 45(1), pp. 5–32.

Campbell, J. and S. Thompson (2008). "Predicting excess stock returns out of sample: Can anything beat the historical average?" In: *The Review of Financial Studies* 21(4), pp. 1509–1531.

Chernov, M. (2007). "On the role of risk premia in volatility forecasting". In: *Journal of Business & Economic Statistics* 25(4), pp. 411–426.

Choi, Y., S. Jacewitz, and J. Park (2016). "A reexamination of stock return predictability". In: *Journal of Econometrics* 192(1), pp. 168–189.

Christoffersen, P. (2003). *Elements of financial risk management*. Academic Press.

Christoffersen, P. and F. Diebold (2006). "Financial asset returns, direction-of-change forecasting, and volatility dynamics". In: *Management Science* 52(8), pp. 1273–1287.

Christoffersen, P., F. Diebold, R. Mariano, A. Tay, and Y. Tse (2007). "Direction-of-change forecasts based on conditional variance, skewness and kurtosis dynamics: international evidence". In: *Journal of Financial Forecasting* 1(2), pp. 1–22.

Chung, J. and Y. Hong (2007). "Model-free evaluation of directional predictability in foreign exchange markets". In: *Journal of Applied Econometrics* 22(5), pp. 855–889.

Cochrane, John H (2009). *Asset Pricing*. Princeton university press.

Corrado, C. and C. Truong (2007). "Forecasting stock index volatility: comparing implied Volatility and the intraday high–low price range". In: *Journal of Financial Research* 30(2), pp. 201–215.

Corsi, F. (2009). "A simple approximate long-memory model of realized volatility". In: *Journal of Financial Econometrics* 7(2), pp. 174–196.

Cortes, C. and V. Vapnik (1995). "Support-vector networks". In: *Machine learning* 20(3), pp. 273–297.

DeLong, E., D. DeLong, and D. Clarke-Pearson (1988). "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach". In: *Biometrics* 44, pp. 837–845.

Diebold, F. and R. Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13(3), pp. 253–263.

Fama, E. (1970). "Efficient capital markets: A review of theory and empirical work". In: *The Journal of Finance* 25(2), pp. 383–417.

Fama, E. (1991). "Efficient capital markets: II". In: *The Journal of Finance* 46(5), pp. 1575–1617.

Fama, E. and K. French (1988). "Dividend yields and expected stock returns". In: *Journal of Financial Economics* 22(1), pp. 3–25.

Fama, E. and K. French (2015). "A five-factor asset pricing model". In: *Journal of Financial Economics* 116(1), pp. 1–22.

Friedman, J. (2001). "Greedy function approximation: a gradient boosting machine". In: *The Annals of Statistics* 29, pp. 1189–1232.

Granger, C. and Z. Ding (1996). "Varieties of long memory models". In: *Journal of Econometrics* 73(1), pp. 61–77.

Han, H., O. Linton, T. Oka, and Y. Whang (2016). "The cross-quantilogram: measuring quantile dependence and testing directional predictability between time series". In: *Journal of Econometrics* 193(1), pp. 251–270.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Wiley Online Library.

Henriksson, R. and R. Merton (1981). "On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills". In: *Journal of Business*, pp. 513–533.

Hjorth, U. (1982). "Model selection and forward validation". In: *Scandinavian Journal of Statistics* 9, pp. 95–105.

Kara, Y., M. Boyacioglu, and Ö. Baykan (2011). "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The

sample of the Istanbul Stock Exchange". In: *Expert Systems with Applications* 38(5), pp. 5311–5319.

Leung, M., H. Daouk, and A. Chen (2000). "Forecasting stock indices: a comparison of classification and level estimation models". In: *International Journal of Forecasting* 16(2), pp. 173–190.

Linton, O. and Y. Whang (2007). "The quantilogram: With an application to evaluating directional predictability". In: *Journal of Econometrics* 141(1), pp. 250–282.

Lo, A. (2004). "The adaptive markets hypothesis". In: *The Journal of Portfolio Management* 30(5), pp. 15–29.

Neely, C., D. Rapach, J. Tu, and G. Zhou (2014). "Forecasting the equity risk premium: the role of technical indicators". In: *Management Science* 60(7), pp. 1772–1791.

Nyberg, H. (2011). "Forecasting the direction of the US stock market with dynamic binary probit models". In: *International Journal of Forecasting* 27(2), pp. 561–578.

Nyberg, H. and H. Pönkä (2016). "International sign predictability of stock returns: The role of the United States". In: *Economic Modelling* 58, pp. 323–338.

Pesaran, H. and A. Timmermann (1992). "A simple nonparametric test of predictive performance". In: *Journal of Business & Economic Statistics* 10(4), pp. 461–465.

Pesaran, H. and A. Timmermann (1995). "Predictability of stock returns: Robustness and economic significance". In: *The Journal of Finance* 50(4), pp. 1201–1228.

Pönkä, H. (2017). "Predicting the direction of US stock markets using industry returns". In: *Empirical Economics* 52(4), pp. 1451–1480.

Qiu, M. and Y. Song (2016). "Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model". In: *PloS one* 11(5).

Rapach, D., J. Strauss, and G. Zhou (2010). "Out-of-sample equity premium prediction: Combination forecasts and links to the real economy". In: *The Review of Financial Studies* 23(2), pp. 821–862.

Rapach, D., G. Zhou, et al. (2013). "Forecasting stock returns". In: *Handbook of Economic Forecasting* 2(Part A), pp. 328–383.

Ripley, B. (1996). "Neural networks and pattern recognition". In: *Cambridge University*.

Ross, S. (2009). *Neoclassical finance*. Princeton University Press.

Timmermann, A. and C. Granger (2004). "Efficient market hypothesis and forecasting". In: *International Journal of forecasting* 20(1), pp. 15–27.

Welch, I. and A. Goyal (2008). "A comprehensive look at the empirical performance of equity premium prediction". In: *The Review of Financial Studies* 21(4), pp. 1455–1508.

Zhou, G. (2010). "How much stock return predictability can we expect from an asset pricing model?" In: *Economics Letters* 108(2), pp. 184–186.