Forecasting Stock Price Changes: Is it Possible?[1]

Pedro N. Rodriguez[2] and Simon Sosvilla-Rivero[3]

First Draft: February 1, 2006

This Draft: May 7, 2006

Abstract

We examine the relation between monthly stock returns and lagged publicly available information. Our primary objective is to determine whether the variables proposed in the literature to predict the equity premium contain incremental information to an investor. We find that certain variables do provide incremental information and may have some practical value. Although this not necessarily imply that return-forecasting models may be used to predict future stock returns, some model specifications may be used to predict future stock movements.

JEL Classification: G12, G14.

Keywords: Stock return predictability, stock movement predictability.

[1] Pedro N. Rodriguez thanks CONACYT (Mexico) for financial support (Fellowship: 170328).

[2] Universidad Complutense de Madrid (UCM), Facultad de Ciencias Económicas y Empresariales, Departamento de Estadística e Investigación Operativa II, Campus de Somosaguas, 28223 - POZUELO DE ALARCÓN (MADRID), Spain, phone: +34 696 58 40 04, e-mail: pedro_nahum@ucm.universia.es.

[3] UCM and Fundación de Estudios de Economía Aplicada (FEDEA), C/ Jorge Juan, 46, 28001 – Madrid, Spain, phone: +34 914359020, e-mail: simon.sosvilla@fedea.es (corresponding author).

1. - Introduction

Explaining either the variations or movements of future stock returns is one of the most elusive goals in academic finance. In their recent article on stock return predictability, Goyal and Welch (2005, p. 27) conclude

> …our paper suggests only that our profession has yet to find a variable that has a meaningful robust empirical equity premium forecasting power, both IS [in-sample] and OOS [out-of-sample], at least from the perspective of a real-world investor.

This paper shows when Goyal and Welch's (2005) conclusion, regarding out-of-sample evidence, is met and when it is not.

A widely recognized fact in finance is that model selection criteria fail to detect out-of-sample predictability in terms of level-based accuracy measures, such as the mean square prediction error (see, e.g., Bossaerts and Hillion (1999)). The evidence suggests that the "true" return generating model is either nonlinear or unstable. However, when one acknowledges the consequences of a potentially misspecified model, return-forecasting models can exhibit external validity via a weighted average model [see, e.g., Avramov (2002) and Cremers (2002)].

Another widely recognized fact in finance is that despite their inability to generate satisfactory level-based out-of-sample results, models may be estimated using publicly available information to predict future stock movements [see, e.g., Pesaran and Timmermann (1995, 2002) and Aiolfi and Favero (2005)]. The evidence suggests that the

variables that economists have suggested to predict the equity premium may have, after all, some practical value.

Although there is some evidence that links movement and return predictability (see, e.g., Christoffersen and Diebold (2005)), it cannot be reconciled with the (multivariate) empirical evidence in a way that offers a consistent description of the behavior of model selection criteria or weighted average models when predicting either the returns or signs of the returns. This paper proposes a simply explanation that bridges this gulf between return and movement predictability.

We investigate the extent to which the monthly stock price returns and movements of the Standard and Poors 500 (S&P 500) are predictable. Our approach differs from prior work in several ways. First, rather than assuming the natural class distribution of the test-set, we follow the intuition of a popular proverb in the Latino community "a good rooster can crow anywhere," that the return-forecasting models should generate desirable properties when predicting out-of-sample regardless of the test-set distribution. We find that return-forecasting models exhibit a disproportional large percentage of error when forecasting negative equity premiums, but lower than the one generated by the historical equity premium mean. In fact, if the test-set would have had a different distribution—i.e., higher proportions of negative equity premiums than positive premiums—model selection criteria and weighted averaging models would have statistically outperformed the historical equity premium mean in terms of mean square prediction error. Under the natural class distribution of the test-set, however, return-

forecasting models cannot outperform the simple historical equity premium mean, consistent with Goyal and Welch (2005).

Next, rather than limiting ourselves to Pesaran and Timmermann's (1992) market timing test statistic, we gauge the direction-of-change predictability of the models using several accuracy measures widely used in epidemiology, machine learning, and radiology and find that even though the Pesaran and Timmermann's (1992) test statistic rejects the null of no market timing skills, return-forecasting models are not properly predicting negative equity premiums.

Finally, we compare the predictive performance of return-forecasting models against coin-toss classifiers.[4] We find that neither model selection criteria, nor methodologies robust to model misspecification, nor predictors evaluated singly outperform coin-toss classifiers. Thus, the empirical evidence we document imply that movement predictability have been overstated in the existing literature when using either model selection criteria or weighted average models.

However, an ex-post analysis reveals that if an investor would have clung to some specifications with 4 or 5 predictors, he could have easily outperformed the historical equity premium mean, model selection criteria, all-inclusive model, and weighted model

[4] Coin-toss classifiers, also known as random classifiers, use a discrete distribution to resolve the dispute between our two alternatives: positive premiums (or up movements) and negative premiums (or down movements).

averaging models in terms of directional-based accuracy measures, but not in terms of level-based accuracy measures. Stock movement predictability is feasible, whereas stock return predictability is not.

The remainder of the paper proceeds as follows. In Section 2, we reexamine the existing sample evidence on return predictability. Section 3 implements several strategies to improve the predictive performance of the return-forecasting models. To check whether some specifications would have helped an investor, we evaluate the predictive performance of all competing regression specifications and report the results in Section 4. We conclude in Section 5.

2. – Another look at the sample evidence on return predictability: Are weighted committees increasing our ability to describe the time-series behavior of stock returns?

As Fama (1991, p. 1577) states, "[t]here is a resurgence of research on time-series predictability of stock returns…" Not only "traditional" variables are nowadays considered in empirical tests for return predictability, tests now also consider the predictive power of a large set of potential variables.[5] Moreover, in contradistinction to the pre-1991 research, which focused on evaluating the in-sample predictability, recent

[5] Avramov (2002), Cremers (2002), Goyal and Welch (2005), and Campbell and Thompson (2005), for instance, examine the information content of more than 13 conditioning variables.

tests assess in-sample and out-of-sample predictability in terms of either level-based or directional-based accuracy measures.[6]

Arguably, one of the most noticeable new results is that predictive regressions in which model uncertainty is assessed and propagated generate desirable properties when predicting out-of-sample. Avramov (2002) and Cremers (2002) find that the Bayesian model averaging's out-of-sample performance is superior to that of model selection criteria in terms of level-based accuracy measures. This finding, however, prompts several (and unanswered) questions: Are weighted committees increasing our ability to describe the time-series behavior of stock returns? Are model averaging techniques able to discriminate stock price movements? Are ensembles extracting information beyond that contained in an i.i.d. model?

We first consider the sample evidence on return predictability through level-based accuracy measures. Then we discuss the implications of evaluating model averaging and model selection criteria via directional-based accuracy measures.

A. - Equity premium predictability

[6] See, for example, Pesaran and Timmermann (1995, 2002), Bossaerts and Hillion (1999), Avramov (2002), Cremers (2002), Goyal and Welch (2003), Ang and Bekaert (2004), Lunde and Timmermann (2005), Aiolfi and Favero (2005), Goyal and Welch (2005), and Rapach and Wohar (2005).

Consider monthly excess returns on the S&P 500 index over the sample period 1953:04 through 2002:12 using the following $p = 11$ conditioning variables (taking one lag):

1. Dividend yield on the S&P 500 index (d/y).
2. Size Premium (SMB).
3. Value Premium (HML).
4. Earnings price ratio on the S&P 500 index (e/p).
5. Stock variance of the S&P 500 index (svar).
6. Cross-sectional premium (csp).
7. Book-to-market ratio (b/m).
8. Net equity expansion of NYSE stocks (ntis).
9. Term spread, defined as the difference between the long term yield on government bond and the 3-month T-bill (tms).
10. Default yield spread, defined as the difference between the BAA- and AAA-rated corporate yields (dfy).
11. Default return spread, defined as the difference between the returns on long-term corporate bonds and the returns on long-term government bonds (dfr).

The data set was kindly provided by Amit Goyal, Ivo Welch, and Kenneth R. French.

Goyal and Welch (2005) examine the usefulness of the aforesaid set of variables, excluding SMB and HML, to predict stock returns. They find that none of the variables evaluated singly or in an all-inclusive linear model or with a model selection criterion outperform the then-prevailing mean. Their conclusions, however, are solely based on a

level-based accuracy measure — i.e., mean squared prediction error— and as we will show below that it is *(a)* highly sensitive to the class distribution of the test-set and *(b)* not useful to gauge direction-of-change predictability.

As Avramov (2002), we perform a fixed-size rolling windows analysis, in which model parameters are first estimated with data from 1 to T (our T corresponds to 180 observations), next with data from 2 to $T + 1$,…, and finally with data from $N - T$ to $T - 1$. At each iteration, one forecasts one-step ahead. Table 1 reports several statistics examining the properties of out-of-sample monthly forecasts generated by several models and composite weighted ensembles.

****
Table 1 about here
****

Following Avramov (2002), we make use of three regression-based tests of predictive accuracy. Namely, forecasts errors' mean equal to zero, zero correlation between forecasts errors and predictive returns (Efficiency), and of zero first-order serial correlation. In addition, we computed the Mean Square Error (MSE) in percent.

The MSE was decomposed in order to assess whether or not the forecasts errors are higher in either positive or negative premiums. The decomposition may be expressed as in the following equation:

$$MSE = x \cdot MSE\,|_{premium>0} + (1-x) \cdot MSE\,|_{premium\leq 0}, \qquad (1)$$

where x represents the proportion of positive premiums in the test sample. Approximately 60 percent of the test sample corresponds to positive premiums.

We use eight forecasting models: First, we consider four models selected by adjusted R-squared (R2a), Akaike Information Criterion (AIC), Schwartz Information Criterion (SIC), and Posterior Information Criterion (PIC), all of which are described by Bossaerts and Hillion (1999). Second, we examine the i.i.d model predicting the then-prevailing mean in stock returns. Finally, we generate three composite weighted ensembles by considering all linear data-generating processes in the presence of 11 conditioning variables (2^{11} models). In particular, the model denoted by Ave (Med) forecasts the average (median) of the 2^{11} models, whereas the model denoted by BMA computes posterior probabilities for the collection of all 2^{11} models. The posterior probability for each model was obtained via the BIC approximation (see Raftery (1995)). Models were estimated using Ordinary Least Squares (OLS).

The results in Table 1 indicate that model averaging techniques tend to outperform model selection criteria in terms of regression-based tests of predictive accuracy. Indeed, the prediction errors have zero mean and are essentially uncorrelated. In addition, the prediction errors are uncorrelated with predicted returns. Moreover, model averaging techniques produce mean square errors smaller than those corresponding to the i.i.d. model and to model selection criteria, consistent with Avramov (2002). It is worth noting that the three model averaging techniques exhibit very similar

predictive performance, consistent with the low discrimination power of model selection criteria (see, e.g., Dell'Aquila and Ronchetti (2006)).

By focusing on the complete picture of the mean square error criterion, however, we may be misguided enough to believe that model averaging techniques outperform either the model selection criteria or the then-prevailing equity mean. In fact, models that outperform in one movement consistently decrease their ability to describe the other movement in terms of the MSE.

To further assess how different test-set distributions affect the MSE criterion, we evaluate the following the test-set distributions (expressed as the percentages of down movements or negative equity premiums): 2%, 10%, 25%, 50%, 75%, 90%, and 95%. To ensure that all experiments have the same test-set size, no matter the class distribution, the test-set size is made equal to the total number of down movements. Each test set is then formed by random sampling from the original test-set data, without replacement, such that the desired class distribution is achieved. To enhance our ability to identify differences in predictive performance with respect to changes in test-set class distribution, the experiments are based on 1000 runs. The results are shown in Table 2.

****
Table 2 about here
****

Table 2 report the effect test-set class distribution on the MSE. The first column in Table 2 specifies the model (or weighted committee). The next seven columns present

the average MSE for the 7 fixed class distributions. The values reported in the main rows are the actual mean square errors, and the entries enclosed in parenthesis are the standard errors. As can be seen from Table 2, a general pattern persists: return-forecasting models' MSE increases as the proportion of down movements increases in the test-set.

The intuition behind of varying the test-set class distribution is that a good forecasting model should generate desirable properties when predicting out-of-sample regardless of the test-set distribution. Patently, this is not case. Models exhibit a disproportional large percentage of error when forecasting negative equity premiums. In fact, using the two-sided test of the null that the population mean difference is zero against the alternative that the population mean difference is not zero, we find that for higher proportions of down-movements, all the estimated models except for the R-adjusted criterion outperform the i.i.d. model. These results suggest that the only reason why in many tests for return predictability the then-prevailing mean cannot be outperformed by return-forecasting models is due to the high proportion of positive premiums in test-sets.

B. - Direction-of-change predictability

Thus far, the analysis exhibits evidence supporting an asymmetry in equity premium predictability: negative equity premiums are not as predictable as positive equity premiums in terms of a level-based accuracy measure. However, the analysis is solely based on the level and not on the direction of the change. Are the probabilities of correctly predicting the sign of change also asymmetrical?

To answer this question, we first evaluate the forecasts of each model using the 0/1 loss function. The 0/1 loss function is usually the main criterion for classification problems, and may be represented as in the following equation:

$$E_i = L(t_i, \hat{y}_i), \tag{2}$$

where at time i, t_i is the "output" or "response" variable $t \in C$, where C is a set of class labels. In this paper, $C \in \{0,1\}$, where C equals to 1 if the observed equity premium is higher than zero, 0 otherwise. $\hat{y}_i$ is the predicted movement. $\hat{y}_i$ equals to 1 if the predicted equity premium is higher than zero, 0 otherwise. E_i equals to 1 if $t_i \neq \hat{y}_i$, 0 otherwise. In other words, the 0/1 loss function represents one less the proportion of correctly predicted signs.

In the machine learning literature, the bias-variance decomposition is widely used as key tool for understating function approximation algorithms. Although the bias-variance decomposition was originally proposed for the square loss (see, e. g., Geman *et al*., 1992), this paper uses the bias-variance decomposition for the 0/1 loss function for one main reason: level accuracy is not as strongly correlated with profits with a trading strategy based on a set of predictions as directional accuracy [see, e.g., Leitch and Tanner (1991) and Pesaran and Timmermann (1995)].

Following Domingos (2000) and Valentini and Dietterich (2004), bias and variance in a noise-free setting can be defined in terms of the main prediction. The main prediction y_m can be defined as the movement that is predicted more often in the test

sample. Thus, the bias (systematic loss incurred by the function) at time *i* can be computed as,

$$B_i = \begin{cases} 1 \text{ if } y_m \neq t_i \\ 0 \text{ if } y_m = t_i \end{cases}. \tag{3}$$

To distinguish between the two different effects of the variance on the loss 0/1 loss function, Domingos defines the unbiased variance, V_u, to be the variance when $B_i = 0,$ and can be calculated as,

$$V_u^i = \left\| (y_m = t_i) \text{ and } (y_m \neq \hat{y}_i) \right\|, \tag{4}$$

where $\|s\| = 1$ if s is true, 0 otherwise. The unbiased variance evaluates the extent to which the estimated function deviates from the correct predictions. The biased variance, V_b^i, occurs when $B_i = 1,$ and evaluates the extent to which the estimated function deviates from the incorrect predictions. The biased variance can be estimated as,

$$V_b^i(\mathbf{x}) = \left\| (y_m \neq t_i) \text{ and } (y_m \neq \hat{y}_i) \right\|. \tag{5}$$

To obtain the loss associated with a given observation a time *i* [denoted by E_i], we simply compute the algebraic sum of bias, unbiased and biased variance as,

$$E_i = B_i + V_u^i - V_b^i. \tag{6}$$

In order to compute the aforementioned variables in a test set, we simply obtain the average for each variable. Clearly, if we want a good function that distinguishes

between up-and-down movements, we want the bias and the unbiased variance to be small. The results for the fixed-size rolling windows scheme are presented in Table 3.

****
Table 3 about here
****

Table 3 shows that model averaging techniques do not outperform model selection criteria in terms of direction-of-change predictability. In fact, model averaging techniques achieve identical error rates as model selection criteria in the test sample. The analysis, however, reveals that the bias, and not variance, plays a significant role in its contribution to the error rate. But why should a financial economist care about the role that the bias and variance play in the error rate?

****
Figure 1 about here
****

We depict the case analysis of error in Figure 1. Consider the leftmost braches of the "tree." We branch to the left if $B_i = 1$ or if $y_m \neq t_i$. In our test sample, for every model, the main prediction was 1— i.e., a positive equity premium. Thus, at time *i* there will be a bias if the models attempt to predict a down movement. The return-forecasting models, however, are not properly predicting down movements, since the estimated Biased Variance for each model is relatively small. In other words, good forecasting models should generate a high biased variance in order to neutralize the effects of the bias on the error rate.

To further illustrate the inability of the estimated models to predict down movements, we gauged the direction-of-change predictability of the models using several accuracy measures widely used in radiology and epidemiology.

A forecasting model with good market timing abilities produces out-of-sample predictions satisfying several important properties, including *(a)* high sensitivity and specificity. Sensitivity of a model (or ensemble) is defined as the proportion of truly up-movement cases that have a predicted equity premium higher than zero and the specificity the proportion of truly down-movements cases that have a predicted premium lower or equal to zero; *(b)* High (low) positive (negative) likelihood ratio. The positive likelihood ratio (LR+) represents the odds ratio that a predicted premium higher than zero will be observed in an up-movement population compared to the odds that the same result will be observed among a down-movement population. The negative likelihood ratio (LR-) represents the odds ratio that a predicted premium lower or equal to zero will be observed in a down-movement population compared to the odds that the same result will be observed among an up-movement population (see, e.g., Biggerstaff (2000)); and *(c)* high before-test rule-in/out potentials. A rule-in potential represents the number of times greater, on average, that an up-movement case will be rule in as an up-movement after performing the prediction in the estimated function, and a rule-out potential represents the number of times that a down movement is more likely to be a down movement (see, e.g., Lee (1999)).

In addition, we evaluated Pesaran and Timmermann's (1992) market timing test statistic, widely used in academic finance. Panel A of Table 4 shows the results for the estimated models. To evaluate the extent to which the results provided in Panel A can be explained by randomness, we simulated 5000 coin-toss classifiers. To obtain each random classifier, we generate random values from a discrete distribution in which two values where possible: 1's and 0's. Each value was assigned 50 per cent of probability. The distribution of the accuracy measures of the random classifiers is presented in Panel B of Table 4.

****
Table 4 about here
****

Table 4 shows that neither model averaging techniques nor model selection criteria have desirable properties for market timing, although positive premiums exhibit large predictable components. Model averaging techniques produce specificity levels worst than random and the specificity levels that model selection criteria generate are perfectively explainable by randomness, even though the rolling-window predictions get the sign of the equity premium right in at least 60 percent of all months over the 1968 to 2002 period! Thus, the percentage of correctly predicted signs of the excess of returns that does not necessarily convey important information to an investor in an unbalanced test-set, as it may clearly mislead an investor to think that his/her model outperform a random classifier.

The low before-test rule-in/out potentials are another way to see the lack of discriminatory power of the estimated models. As all rule-in/out levels can be explained by coin-toss classifiers. Low (high) positive (negative) likelihood ratios can also be confirmed by coin-toss classifiers. In sum, neither model selection criteria nor model averaging techniques detect out-of-sample predictability in terms of the direction of change simply because the return-forecasting models fail to understand negative premiums.

Cooper and Gulen (2006) find that random "inputs" can largely explain the literature's out-of-sample evidence. Our results are consistent with them in the sense that random classifiers can entirely explain the apparent predictability. Perhaps the most striking result in Table 4, consistent with Cooper and Gulen, is that coin-toss classifiers can reject the null hypothesis of no market timing against the alternative of market timing skills at usual significance levels.

3. – Learning from unbalanced data sets.

Weiss and Provost (2003, p. 323) indicate that "[p]ractitioners have noted that learning performance often is unsatisfactory when learning from data sets where the minority class is substantially underrepresented." To illustrate the extent to which negative premiums are underrepresented each time the return-forecasting models are trained, we plot in Figure 2 the proportion of positive premiums (or up-movements) to negative premiums (or down-movements) in each training-set of the rolling window analysis.

****
Figure 2 about here
****

Figure 2 illustrate that each time we estimate a return-forecasting model more than 60 percent of the observations correspond to up-movements. In this section, we implement two strategies to deal with the imbalanced data each time we train a return-forecasting model. First, we evaluate whether or not different training-set distributions improve the predictive performance. Then, we assess the effectiveness of cost sensitive learning.

A. - Varying the training-set distribution

We now analyze how the return-forecasting models perform under several training-set distributions. This strategy is gaining more ground in the machine learning community (see, e.g., Kubat and Matwin (1997) and Weiss and Provost (2003)). We evaluate the following the training-set distributions (expressed as the percentages of down movements or negative equity premiums): 25%, 40%, 50%, 60%, and 75%. To ensure that all experiments have the same training-set size, no matter the class distribution, the training-set size is made equal to the total number of down movements available at time i when attempting to predict $i + 1$. Each training set is then formed by random sampling from the training-set data, without replacement, such that the desired class distribution is achieved. Note that the training-set data contains the only the information accessible at each iteration from the rolling window analysis. Due to the high computational burden of the analysis, the experiments are only based on 5 runs. The results are shown in Table 5.

****
Table 5 about here
****

As can be seen from Table 5, all premium-forecasting models, with the exception of the i.i.d model, improve their ability to understand (and predict) negative premiums as more negative premiums are available in the training-set. However, this improvement is not cost-free. The proportion of down-movements has an inversely related effect on sensitivity levels of the estimated models. Note also that if the real training-set distribution would have been different—i.e., higher proportion of down-movements—our concern would be explaining the unpredictability of positive premiums.

B. - Cost sensitive learning

The trade-off between sensitivity and specificity levels as more negative premiums are included in the training-set can be explained by a loss in information, as a large part of the positive premiums population is not used for training. Another approach to make return-forecasting models more suitable for learning from imbalanced data sets follows the idea of cost sensitive learning. The idea is to assign a higher cost (or weight) to the error of the negative premiums in the training phrase. We re-estimated the fixed-size rolling windows analysis from Section 2, but in the learning process we included a vector of positive weights. Negative premiums were weighted 20 percent more than positive premiums. In other words, instead of minimizing the sum-of-squares we minimize the weighted sum-of-squares. The results are shown in Table 6.

****
Table 6 about here
****

Table 6 shows that the return-forecasting models improve, albeit slightly, their ability to predict negative premiums. However, cost sensitive learning is as infective as varying the training-set distribution. As a model improves its ability to describe negative equity premiums its ability to describe positive equity premiums deteriorates.[7]

4. – Are the "explanatory" variables irrelevant?

As we have shown above, return-forecasting models are not properly discriminating positive from negative premiums. Thus, the lack of predictive power of the return-forecasting variables could be explained by either the irrelevance of the "explanatory" variables or by the failure of model selection criteria to deliver the "best" specifications. One of the easiest ways to evaluate the relevance and redundancy of the "explanatory" variables, although very inefficient for high-dimensional data, is to search for a minimum subset of variables that maximizes predictability. Clearly, maximizing predictability with a set of ex-ante observable variables is more appropriate. However, our goal here is to test whether or not some specifications would have provided incremental information to an investor. The out-of-sample mean square error, in percent, for all the competing regression specifications is shown in Figure 3.

[7] We have also experimented with different weights, but achieved the same pattern. The results, nevertheless, are available upon request to the authors.

****
Figure 3 about here
****

Figure 3 shows that the lowest mean square error, 0.2002, is achieved using a subset of size 4. Note that the mean square error of the best return-forecasting model inside a subset increases as the size of the subset increases, after the subset of size 4. This suggests that some explanatory variables act as redundant.

To formally test whether or not the return-forecasting models contain information that it is not present in the then-prevailing mean (or i.i.d. model), we implement Diebold and Mariano's (1995) sign test statistic. Let p_m be the vector of predictions of the model *m*, *t* be the vector of observable equity premiums, and p_{iid} be the vector of predictions of the i.i.d. model. Then, $e_m = (t - p_m)$ and $e_{iid} = (t - p_{iid})$ denote the corresponding errors. The sign test statistic $\{S\}$ is defined for model *m* by,

$$S_m = \frac{2}{\sqrt{n}} \sum_{j=1}^{n} \left(I[d_{m,j} > 0] - \frac{1}{2} \right) \overset{a}{\sim} N(0,1), \tag{7}$$

where $d_{m,j}$ is the so-called loss differential at time *j*, $d_{m,j} = e_{iid,j}^2 - e_{m,j}^2$, and *I* is an indicator function. We compute the *S* statistic for all competing regression specifications and depict the results in Figure 4.

****
Figure 4 about here
****

Significant and positive (negative) values for *S* indicate a significant difference between the two forecasting errors, which imply a better accuracy of the *m* (i.i.d) model. Figure 4 shows that the predictions errors of all competing regression specifications cannot outperform the historical equity premium mean. However, the historical mean is able to generate better accuracy than many models. Another possibility to test the null hypothesis that there is no qualitative difference between forecasts from two models is to use re-sampling techniques.

Re-sampling techniques are computer-intensive statistical tools for estimating the distribution of a parameter that in other ways would be difficult to obtain.[8] The traditional re-sampling algorithm two compute the difference between two parameters, such as mean square prediction error, is: (1) draw a sample of size *n* with replacement from the observed sample of values of model 1 and calculate its mean square prediction error, (2) using the same random rows from step 1, for model 2 calculate the mean square prediction error, (3) compute the difference between the MSEs, (4) repeat steps 1 and 2 thousand times to obtain a set of bootstrap replications.

****
Figure 5 about here
****

[8] Re-sampling techniques are described in more technical detail in Hall (1992) and Davison and Hinkley (1997). Practical examples of confidence interval construction are given by Efron and Tibshirani (1993). Guide for choosing a bootstrap confidence method when using nonparametric or parametric simulation is given by Carpenter and Bithell (2000).

Figure 5 shows the *p*-value for each competing regression specification. The *p*-value represents the proportion of bootstrap estimates in which the difference between the MSEs is greater than zero. In other words, the *p*-value denotes the fraction of times where the simulated MSE of model *m* is lower than the one from the i.i.d. model. Therefore, low *p*-values indicate that the MSE of model *m* is statistically lower than the MSE of the i.i.d. model. As can be seen from Figure 5, none of all the competing regression specifications outperform the i.i.d. model, consistent with Goyal and Welch (2005).

The results presented in Figure 4 and 5 indicate that model selection criteria does not fail to deliver specifications that are the best predictors with respect to a set of ex-ante observable economic variables simply because there is not a single specification that outperform the historical equity premium mean in terms of level-based accuracy measures. As a result, an investor needs to exercise caution in transforming a level forecast from a return-forecasting model into a trading strategy.

We also computed the average between sensitivity and specificity for all the competing regression specifications to gauge the overall discriminatory accuracy. We depict the results in Figure 6.

****
Figure 6 about here
****

Figure 6 shows that the highest overall discriminatory accuracy, 0.5941, is attained using a subset of size 5. If we compare the discriminatory accuracy of all

competing regression specifications against the coin-toss classifiers simulated in Section 2.B, 21 specifications would have outperformed random predictions at 1 percent of significance and 287 at 5 percent of significance.

However, Cooper and Gulen (2006) show that much of the literature's out-of-sample time-series-based predictability is consistent with data snooping. Can random "inputs" explain the results depicted in Figure 6? Following Cooper and Gulen, we use non-repeating seeds to generate eleven random $N(0,1)$ predictive variables. We compute the average between sensitivity and specificity for all competing regression specifications in the presence of these random variables. In the analysis, we followed the same instructions as in Section 2.A and run the simulation 10 times since Cooper and Gulen obtain with 10 iterations 100% or greater of the real-data predictability. The results are shown in Figure 7.

****
Figure 7 about here
****

Figure 7 illustrates the discriminatory accuracy of 20480 models. The highest discriminatory accuracy achieved by random inputs is 0.5525. Thus, the sample evidence does not allow us to conclude that the "explanatory" variables proposed in the literature are irrelevant, as they can be use to estimate a return-forecasting model that is able to discriminate financial movements.

5. – Concluding remarks

In this paper, we have examined the relation between stock returns and lagged publicly available information. Our primary objective was to determine whether return-forecasting models contain incremental information to an investor. We find that certain variables proposed in the literature to predict the equity premium do provide incremental information and may have some practical value. Although this not necessarily imply that return-forecasting models may be used to predict future stock returns, some model specifications may be used to predict future stock movements.

Our results suggest that predictive regressions can be improved by data-intensive techniques. In particular, it may be possible to determine "optimal specifications" using out-of-bag estimates or to search for nonlinear relationships using General Additive Models. Such considerations may lead to an entirely new paradigm of what the important conditioning variables are for predicting stock returns. We hope to explore those issues more fully in future research.

References

Aiolfi, M., and C. A. Favero. "Model uncertainty, thick modeling and the predictability of stock returns." *Journal of Forecasting* 24 (2005), 233-254.

Ang, A., and G. Bekaert. "Stock return predictability: Is it there?." Unpublished manuscript, Graduate School of Business, Columbia University (2001)..

Avramov, D. "Stock return predictability and model uncertainty." *Journal of Financial Economic* 64 (2002), 423-458.

Avramov, D., and T. Chordia. "Predicting stock returns." *Journal of Financial Economics*, forthcoming, (2005).

Baker. M. and J. Wurgler. "The equity share in new issues and aggregate stock returns." *Journal of Finance* 55 (2000), 2219-2257.

Bauer, E., and R. Kohavi. "An empirical comparison of voting classification algorithms: Bagging, boosting and variants." *Machine Learning* 36 (1999), 525-536.

Biggerstaff, B. D. "Comparing diagnostic tests: a simple graphic using likelihood ratios." *Statistics in Medicine* 19 (2000), 649-663.

Boothe, P. and D. Glassman. "Comparing exchange rate forecasting models: Accuracy versus profitability." *International Journal of Forecasting* 3 (1987), 65-79.

Bossaerts, P., and P. Hillion. "Implementing statistical criteria to select return forecasting models: What do we learn?." *Review of Financial Studies* 12 (1999), 405-428.

Bradley, A. P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognition* 30 (1998), 1145-1159.

Breiman, L. "Using iterative bagging to debias regressions." Machine Learning 45 (2001), 261-277.

Campbell, J. Y. "Stock returns and the term structure." *Journal of Financial Economics* 18 (1987), 373-399.

Campbell, J. Y. and R. J. Shiller. "Stock prices, earnings, and expected dividends." *Journal of Finance* 43 (1998), 661-676

Campbell, J. Y. and S. B. Thompson. "Predicting the equity premium out-of-sample: Can anything beat the historical average?." Working paper 11468, Bureau of Economic Research, July, (2005).

Carpenter, J., and J. Bithell. "Bootstrap confidence intervals: when?, which?, what? A practical guide for medical statisticians." *Statistics in Medicine* 19 (2000), 1141-1164.

Chen, N.F., R. Ross, and S. Ross. "Economic forces and the stock market." *Journal of Business* 59 (1986), 383-404.

Christoffersen, P.F., and F.X. Diebold. "Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics." *Management Science*, forthcoming, (2005).

Cooper, M., and H. Gulen. "Is time-series based predictability evident in real time." Unpublished manuscript, Krannert Graduate School of Management, Purdue University (2004).

Cremers, K. J. M. "Stock return predictability: A Bayesian model selection perspective." *Review of Financial Studies* 15 (2002), 1223-1249.

Davison, A. C., and D. V. Hinkley. *Bootstrap Methods and their Applications.* Cambridge University Press (1997).

Dell'Aquila, R., and E. Ronchetti. "Stock and bond return predictability: The discrimination power of model selection criteria." *Computational Statistics and Data Analysis* 50 (2006), 1478-1495.

Diebold, F. X., and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13 (1995), 253-265.

Domingos, P. A unified bias-variance decomposition for zero-one and squared loss, in *Proceedings of the Seventeenth National Conference on Artificial Intelligence* Austin: AAAI Press (2000), 564-569.

Efron, B., and R. Tibshirani. *An Introduction to the Bootstrap.* London: Chapman and Hall (1993).

Fama, E, "Efficient capital markets: II," *Journal of Finance* 46 (1991), 1575-1617.

Fama, E., and K. R. French. "Dividend yields and expected stock returns," *Journal of Financial Economics* 22 (1988), 3-25.

Fama, E., and K. R. French. "The cross-section of expected stock returns," *Journal of Finance* 47 (1992), 427-465.

Geman, S., E. Bienenstock, and R. Doursat. "Neural networks and the bias-variance dilemma," *Neural Computation* 4 (1992), 1-58.

Goyal, A., and I. Welch. "Predicting the equity premium with dividend ratios," *Management Scienc*e 49 (2003), 639-654.

Goyal, A., and I. Welch. "A comprehensive look at the empirical performance of equity premium predictions," *Review of Financial Studies* (2005), forthcoming.

Hall, P., 1992, *The Bootstrap and Edgeworth Expansion.* London: Springer-Verlag (1992).

Kothari, S. P. and J. Shanken. "Book-to-market time series analysis," *Journal of Financial Economics* 44 (1997), 169-203.

Kubat, M., and S. Matwin. Addressing the curse of imbalanced data sets: One-sided sampling, in D. H. Fisher, ed.: *Proceedings of the 14th International Conference on Machine Learning*. Nashville: Morgan Kaufmann (1997), 179-186.

Lee, W., 1999, Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance, *International Journal of Epidemiology* 28, 521-525.

Leitch, G, and J. E. Tanner. "Economic forecast evaluation: Profits versus the conventional error measures," *American Economic Review* 81 (1991), 580-590.

Lunde, A., and A. Timmermann. "Completion time structures of stock price movements," *Annals of Finance* 2 (2005), 193-226.

Pesaran, M. H., and A. Timmermann. A simple nonparametric test of predictive performance, *Journal of Business and Economic Statistics* 10 (1992), 461-465.

Pesaran, M. H., and A. Timmermann. "Predictability of stock returns: Robustness and economic significance," *Journal of Finance* 50 (1995), 1201-1228.

Pesaran, M. H., and A. Timmermann. "Market timing and return prediction under model instability," *Journal of Empirical Finance* 9 (2002), 495-510.

Polk, C., S. Thompson, and T. Vuolteenaho. "Cross-section forecasts of the equity premium," *Journal of Financial Economics* (2006), forthcoming.

Pontiff, J., and L. D. Schall. "Book-to-market ratios as predictors of market returns," *Journal of Financial Economics* 49 (1998), 141-160.

Rangvid, J. "Output and expected returns," *Journal of Financial Economics* (2005), forthcoming.

Raftery, A. E. "Bayesian model selection in social research (with Discussion)," *Sociological Methodology* 25 (1995), 111-196.

Rapach, D. E., and M. E. Wohar. "In-sample vs. out-of-sample tests of stock return predictability in the context of data mining," *Journal of Empirical Finance* (2005), forthcoming.

Schwert, G. M. Anomalies and market efficiency, in G. M. Constantinides, M. Harris, and R. Stulz, eds.: *Handbook of the Economics of Finance*, Vol. 1, Part. 2 Amsterdam: North-Holland (2003) 937-972.

Valentini, G., and T. G. Dietterich. "Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods," *Journal of Machine Learning Research* 5 (2004), 725-775.

Weiss, G. M., and F. Provost. "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of Artificial Intelligence Research* 19 (2003), 315-354.

Table 1: Out-of-sample results

The table displays several statistics examining the properties of out-of-sample forecast errors generated by several models and weighted committees or ensembles. The former set includes the i.i.d model and five models selected by adjusted R-squared (r2a), AIC, SIC, FIC, and PIC. We examine three ensembles: Ave, Median, and BMA. Ave represents the collection of all 2^p models (where p denotes the number of explanatory variables in the study) in which each model is equally-weighted. Median forecasts the median of all 2^p models. BMA stands for Bayesian Model Averaging. BMA computes posterior probabilities for the collection of all 2^p models. The posterior probability for each model was obtained via the BIC approximation. The forecasts of each model or weighted committees were evaluated with several regression-based test of prediction accuracy, such as MPE, Efficiency, and Serial correlation, all of which are described by Avramov (2002) and West and McCracken (1998). In addition, we computed the Mean Squared Error (MSE) for each model. The MSE was decomposed in order to evaluate whether or not the forecast errors are higher in either positive or negative movements as $MSE = x \cdot MSE(y=1) + (1-x) \cdot MSE(y=0)$, where x denotes the proportion of positive premiums in the test-set.

	Ave	Med	BMA	AIC	SIC	PIC	i.i.d	R2a
MPE	0.002	0.002	0.002	0.005	0.004	0.006	-0.001	0.000
t-statistic	*0.920*	*0.720*	*0.925*	*2.009*	*1.813*	*2.421*	*-0.285*	*-0.069*
Serial correlation	-0.066	-0.071	-0.066	-0.041	-0.023	-0.016	-0.002	0.006
t-statistic	*-1.358*	*-1.448*	*-1.356*	*-0.847*	*-0.459*	*-0.327*	*-0.049*	*0.126*
Efficiency	-0.319	-0.285	-0.319	-0.652	-0.546	-0.678	-5.387	-1.411
t-statistic	*-1.463*	*-1.265*	*-1.470*	*-5.933*	*-4.942*	*-6.605*	*-2.060*	*-4.385*
MSE	0.203	0.202	0.203	0.221	0.211	0.226	0.207	0.215
MSE (y=1)	0.178	0.174	0.178	0.217	0.210	0.224	0.159	0.166
MSE (y=0)	0.240	0.244	0.240	0.226	0.214	0.229	0.278	0.287

Table 2: Effect of test-set class distribution on the MSE.

The table displays the MSE criterion (generated by several models and weighted committees described in Table 1) according to several the test-set distributions. We evaluate the following test-set distributions (expressed as the percentages of down movements): 2%, 10%, 25%, 50%, 75%, 90%, and 95%. The values reported in the main rows are the actual mean square errors, and the entries enclosed in parenthesis are the standard errors.

	Out-of-sample MSE when using specified test-set distributions (training distribution expressed as % of down-movements)						
	2	10	25	50	75	90	95
2k (Ave)	0.1785	0.1831	0.1935	0.2099	0.2241	0.2333	0.2375
	(0.0156)	(0.0183)	(0.0219)	(0.0229)	(0.0188)	(0.0129)	(0.0094)
2k (Med)	0.1759	0.1816	0.1911	0.2089	0.2268	0.2370	0.2402
	(0.0156)	(0.0184)	(0.0209)	(0.0212)	(0.0183)	(0.0122)	(0.0099)
BMA	0.1788	0.1841	0.1939	0.2106	0.2246	0.2341	0.2368
	(0.0159)	(0.0181)	(0.0212)	(0.0227)	(0.0187)	(0.0128)	(0.0090)
AIC	0.2164	0.2165	0.2197	0.2216	0.2228	0.2248	0.2259
	(0.0167)	(0.0187)	(0.0221)	(0.0217)	(0.0176)	(0.0118)	(0.0092)
SIC	0.2097	0.2111	0.2100	0.2108	0.2133	0.2134	0.2138
	(0.0157)	(0.0177)	(0.0197)	(0.0199)	(0.0165)	(0.0119)	(0.0082)
PIC	0.2228	0.2243	0.2235	0.2263	0.2282	0.2282	0.2287
	(0.0181)	(0.0189)	(0.0221)	(0.0219)	(0.0179)	(0.0127)	(0.0094)
i.i.d	0.1670	0.1702	0.1884	0.2176	0.2478	0.2650	0.2710
	(0.0133)	(0.0169)	(0.0214)	(0.0244)	(0.0196)	(0.0133)	(0.0104)
R2a	0.1682	0.1769	0.1966	0.2256	0.2580	0.2743	0.2804
	(0.0148)	(0.0185)	(0.0227)	(0.0252)	(0.0202)	(0.0140)	(0.0104)

Table 3: Bias-variance decomposition.

The table displays several statistics examining the properties of out-of-sample forecast errors generated by several models and weighted committees described in Table 1. The forecasts of each model or composite weighted ensembles were evaluated with 0/1 loss function, which evaluates the usefulness of the estimated model to distinguish up from down movements. 0/1 Loss, Bias, Net Variance, Unbiased Variance, and Biased Variance are all described by Domingos (2000) and Valentini and Dietterich (2004).

	0-1 Loss	Bias	Net Variance	Unbiased Variance	Biased Variance
2k (Ave)	0.4038	0.4038	0.0000	0.1010	0.1010
2k (Med)	0.3918	0.4038	-0.0120	0.0913	0.1034
BMA (Ave)	0.4038	0.4038	0.0000	0.1010	0.1010
AIC	0.4207	0.4038	0.0168	0.1971	0.1803
SIC	0.4255	0.4038	0.0216	0.1851	0.1635
PIC	0.4038	0.4038	0.0000	0.1851	0.1851
i.i.d	0.4038	0.4038	0.0000	0.0000	0.0000
R2a	0.4038	0.4038	0.0000	0.0168	0.0168

Table 4: Direction-of-change predictability.

Table 3A-B displays several statistics examining the properties of out-of-sample forecast errors generated by several models and weighted committees described in Table 1. Sensitivity of a model (or ensemble) is defined as the proportion of truly up-movement cases that have a predicted return higher than zero and the specificity the proportion of truly down-movements cases that have a predicted return lower or equal to zero. The positive likelihood ratio (LR+) represents the odds ratio that a predicted return higher than zero will be observed in an up-movement population compared to the odds that the same result will be observed among a down-movement population. The negative likelihood ratio (LR-) represents the odds ratio that a predicted return lower or equal to zero will be observed in a down-movement population compared to the odds that the same result will be observed among an up-movement population. The next two columns are Kullback-Leibler distances. exp[D(f||g)] represents the number of times greater, on average, that an up-movement case will be rule in as an up-movement after performing the prediction, whereas a down-movement case, on average, will become exp[D(g||f)] times more likely to be down-movement. The column labeled Sign represents the probability of correctly predicting the sign of change, while the column labeled PT represents the Pesaran and Timmermann (1992) test statistic. The distribution of the accuracy measures of the coin-toss classifiers is shown in Panel B.

A. Discriminatory power of the estimated models

	Sensitivity	Specificity	LR+	LR-	exp(D(f || g))	exp(D(g || f))	Sign	PT
2k (Ave)	0.8306	0.2500	1.1075	0.6774	1.0082	1.0091	0.5962	2.0129
2k (Med)	0.8468	0.2560	1.1381	0.5986	1.0135	1.0154	0.6082	2.5994
BMA	0.8306	0.2500	1.1075	0.6774	1.0082	1.0091	0.5962	2.0129
AIC	0.6694	0.4464	1.2092	0.7406	1.0122	1.0126	0.5793	2.3933
SIC	0.6895	0.4048	1.1584	0.7671	1.0083	1.0086	0.5745	1.9824
PIC	0.6895	0.4583	1.2730	0.6774	1.0200	1.0210	0.5962	3.0680
i.i.d	1	0	1.0000	*1/0*	*1/0*	*1/0*	0.5962	0.0000
R2a	0.9718	0.0417	1.0140	0.6774	1.0011	1.0013	0.5962	0.7468

B. Random predictability (Up-movements=50%)

Percentiles	Sensitivity	Specificity	LR+	LR-	exp(D(f || g))	exp(D(g || f))	Sign	PT
0.01	0.4274	0.4107	0.7952	0.79374	1.0000	1.0000	0.4423	-2.3590
0.05	0.4474	0.4345	0.8487	0.84872	1.0000	1.0000	0.4591	-1.6860
0.25	0.4758	0.4762	0.9315	0.93548	1.0005	1.0005	0.4832	-0.6897
0.50	0.5000	0.5000	0.9992	1.00073	1.0023	1.0023	0.5000	-0.0038
0.75	0.5202	0.5238	1.0696	1.07331	1.0068	1.0068	0.5168	0.6852
0.95	0.5524	0.5595	1.1833	1.18344	1.0196	1.0197	0.5409	1.6780
0.99	0.5726	0.5893	1.2752	1.26779	1.0336	1.0333	0.5577	2.3979

Table 5: Effect of training-set class distribution on the out-of-sample accuracy measures.
Table 5 displays several statistics examining the properties of out-of-sample forecast errors generated by several models and weighted committees described in Table 1. For each model, we report the out-of-sample accuracy measures when using specified training distributions (training distribution expressed as % of down-movements).

	25	40	50	60	75
2k (Ave)					
mse	0.2200	0.2090	0.2188	0.2322	
Error rate	0.4019	0.4302	0.5254	0.5408	
sensitivity	0.9629	0.7596	0.4572	0.2370	
specificity	0.0595	0.2892	0.5000	0.7869	
2k (Med)					
mse	0.2182	0.2081	0.2189	0.2311	
Error rate	0.3995	0.4177	0.5221	0.5365	
sensitivity	0.9661	0.7846	0.4741	0.2435	
specificity	0.0607	0.2833	0.4833	0.7880	
BMA					
mse	0.2200	0.2090	0.2188	0.2322	
Error rate	0.4019	0.4302	0.5269	0.5413	
sensitivity	0.9629	0.7596	0.4556	0.2362	
specificity	0.0595	0.2892	0.4988	0.7869	
AIC					
mse	0.2431	0.2310	0.2411	0.2519	
Error rate	0.4312	0.4634	0.5288	0.5278	
sensitivity	0.8451	0.6467	0.4338	0.3024	
specificity	0.1607	0.3738	0.5261	0.7226	
SIC					
mse	0.2262	0.2268	0.2367	0.2453	
Error rate	0.4125	0.4596	0.5158	0.5408	
sensitivity	0.9000	0.6774	0.4774	0.2725	
specificity	0.1261	0.3380	0.4940	0.7345	
PIC					
mse	0.2548	0.2412	0.2454	0.2582	
Error rate	0.4384	0.4567	0.5091	0.5235	
sensitivity	0.8266	0.6403	0.4677	0.3233	
specificity	0.1702	0.4000	0.5250	0.7023	
R2a					
mse	0.2190	0.2079	0.2124	0.2283	
Error rate	0.4048	0.4043	0.4625	0.5451	
sensitivity	0.9927	0.9346	0.6532	0.2145	
specificity	0.0083	0.0952	0.3667	0.8095	

Table 6: Weighted ordinary least square regression results.

Table 6 displays several statistics examining the properties of out-of-sample forecast errors generated by several models and weighted committees, described in Table 1, trained using a weighted least square regression. Negative premiums were weighted 20 percent more than positive premiums.

	Ave	Med	BMA	AIC	SIC	PIC	R2a
MPE	0.0052	0.0047	0.0052	0.0076	0.0069	0.0086	0.0027
t-statistic	*2.3420*	*2.1330*	*2.3460*	*3.3050*	*3.0470*	*3.6910*	*1.1860*
Serial correlation	-0.0701	-0.0727	-0.0701	-0.0420	-0.0210	-0.0185	-0.0005
t-statistic	*-1.4335*	*-1.4867*	*-1.4335*	*-0.8589*	*-0.4286*	*-0.3776*	*-0.0102*
Efficiency	-0.2852	-0.2731	-0.2863	-0.6442	-0.5450	-0.6727	-1.8205
t-statistic	*-1.3440*	*-1.2370*	*-1.3510*	*-5.9510*	*-5.0360*	*-6.8180*	*-4.9510*
MSE	0.2043	0.2039	0.2043	0.2239	0.2146	0.2308	0.2166
MSE (y=1)	0.1974	0.1938	0.1974	0.2371	0.2282	0.2440	0.1843
MSE (y=0)	0.2145	0.2188	0.2144	0.2044	0.1946	0.2113	0.2643
Error	0.4159	0.4135	0.4159	0.4399	0.4351	0.4327	0.4183
Bias	0.4038	0.4038	0.4038	0.4038	0.4038	0.4038	0.4038
Unbiased Var	0.1490	0.1418	0.1490	0.2308	0.2163	0.2260	0.0505
Biased Var	0.1370	0.1322	0.1370	0.1947	0.1851	0.1971	0.0361
Sensitivity	0.7500	0.7621	0.7500	0.6129	0.6371	0.6210	0.9153
Specificity	0.3393	0.3274	0.3393	0.4821	0.4583	0.4881	0.0893

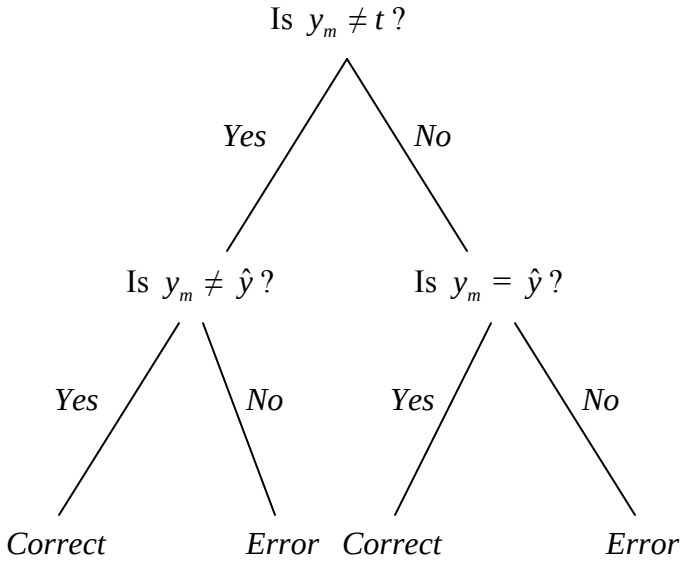

Figure 1: Case analysis of error in a noise-free setting.

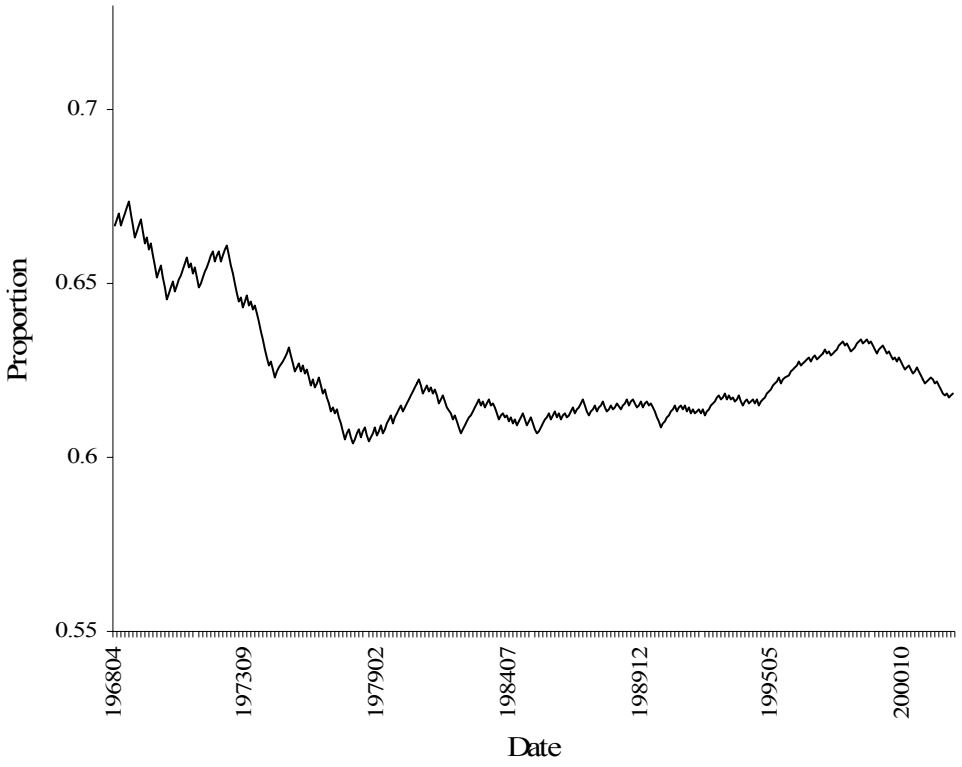

Figure 2: Training-set distribution expressed as % of up-movements

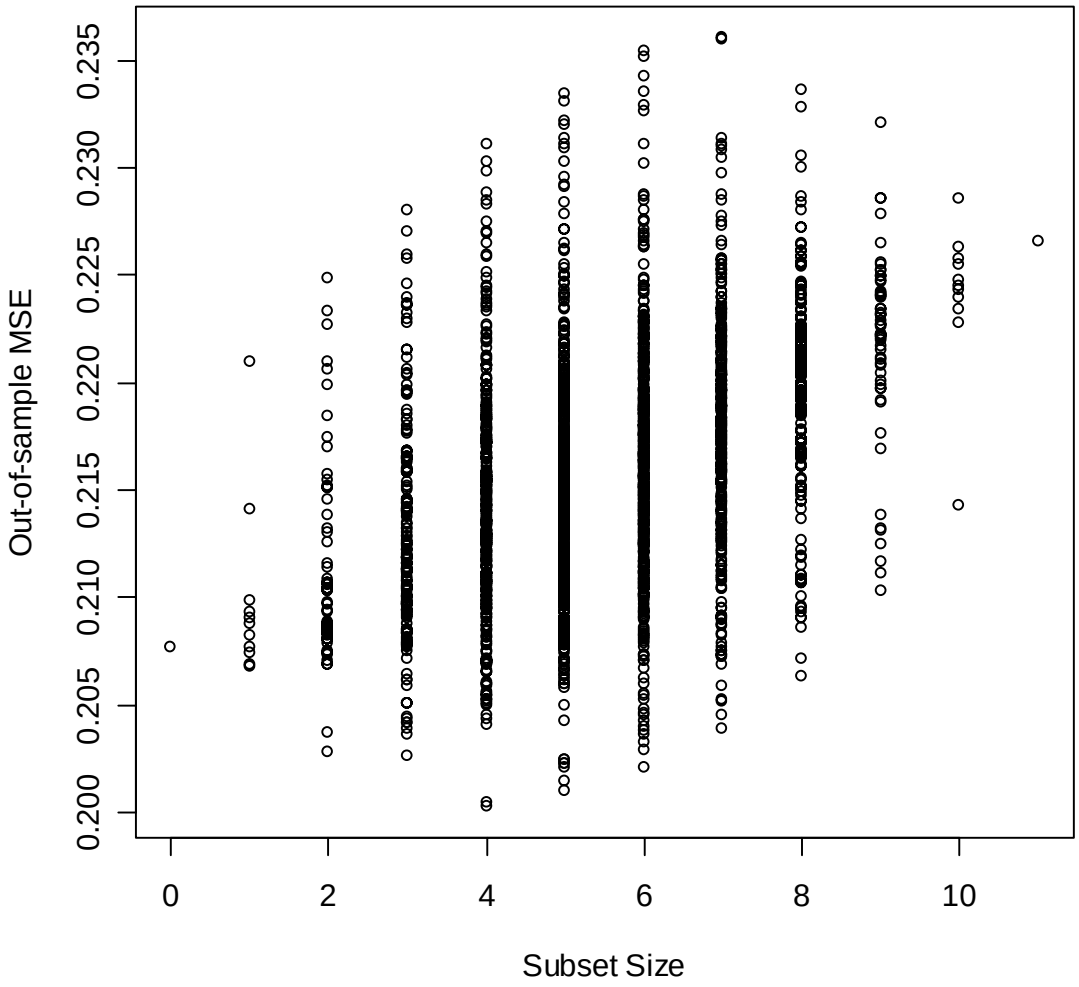

Figure 3: All possible subset models. At each subset size is shown the out-of-sample mean square error, in percent, for each model of that size.

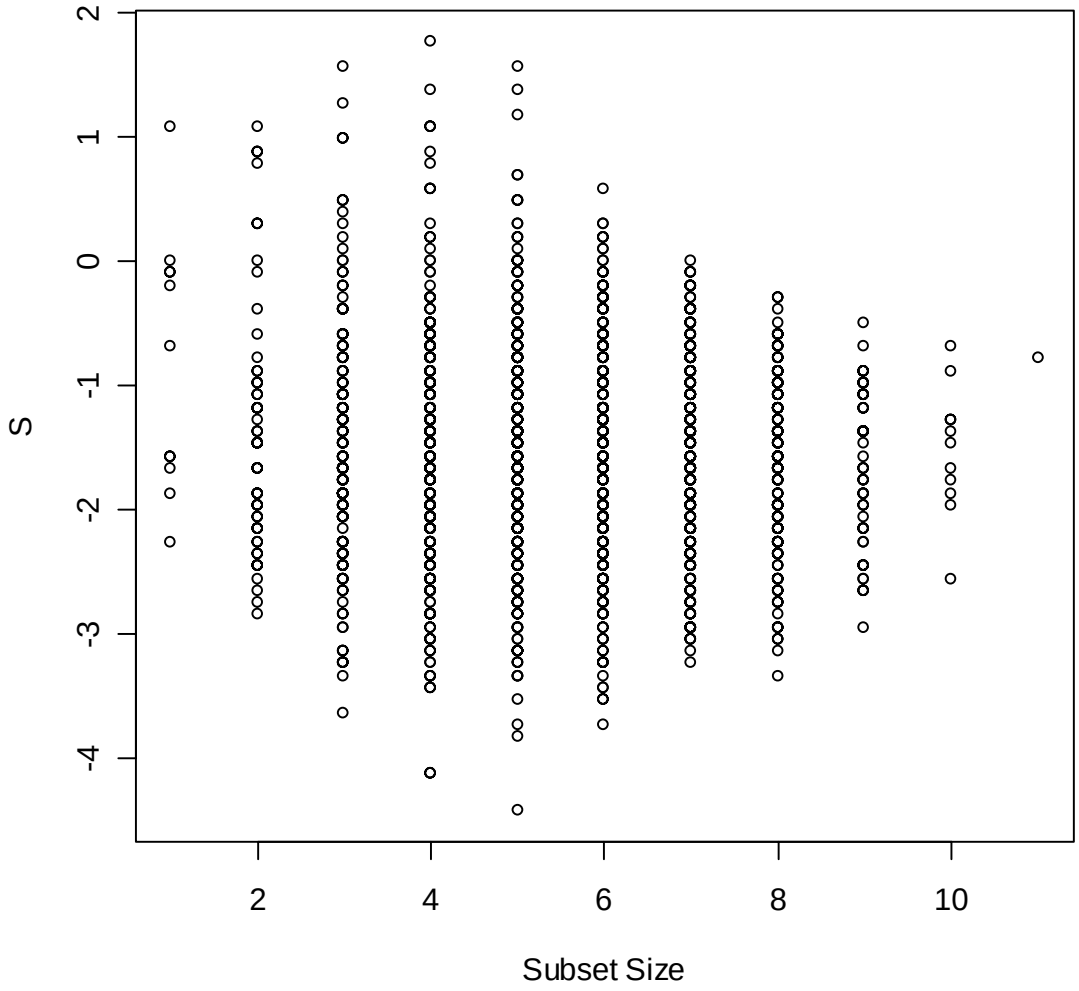

Figure 4: All possible subset models. At each subset size is shown the Diebold and Mariano's (1995) S test statistic on Eq. (7) for each model of that size.

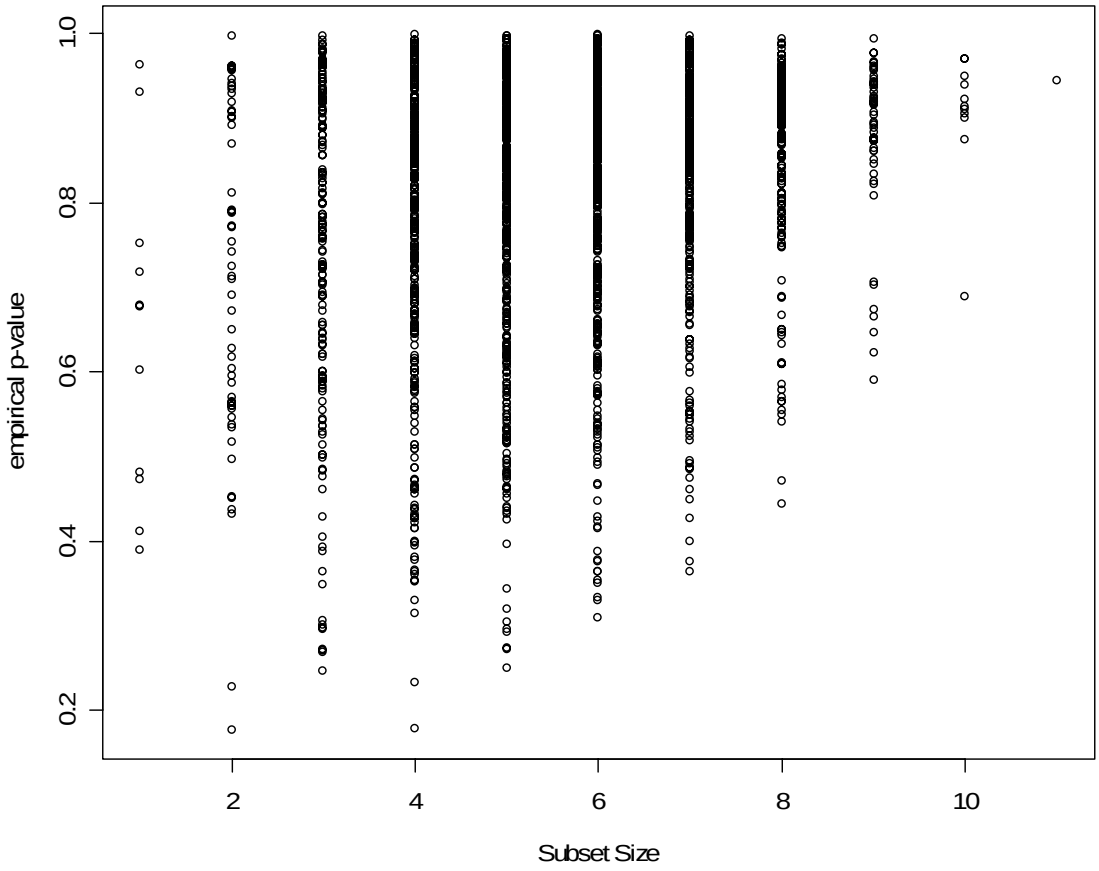

Figure 5: All possible subset models. At each subset size is shown the empirical p-value (the fraction of times where the simulated MSE of model m is lower than the one from the i.i.d. model) for each model of that size.

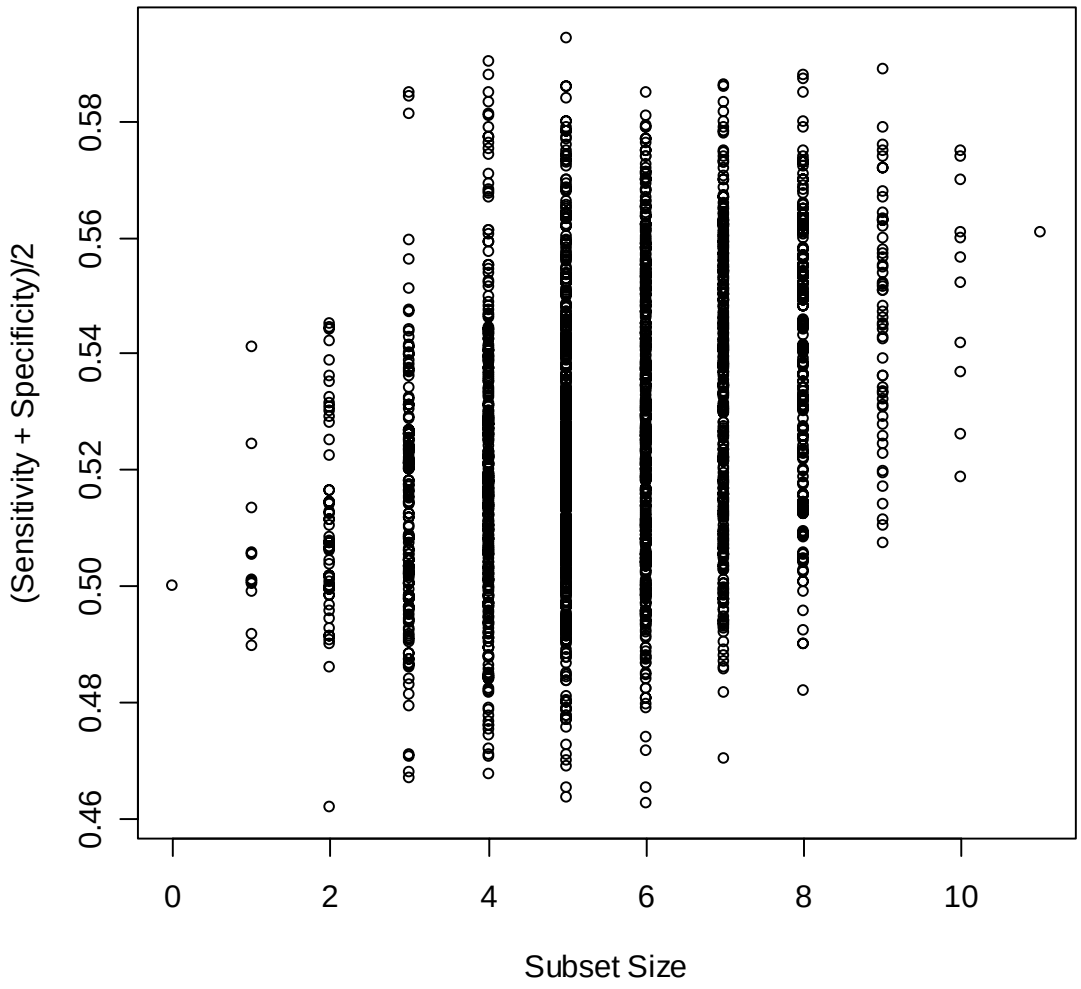

Figure 6: All possible subset models. At each subset size is shown the overall discriminatory accuracy [(sensitivity + specificity)/2] for each model of that size.

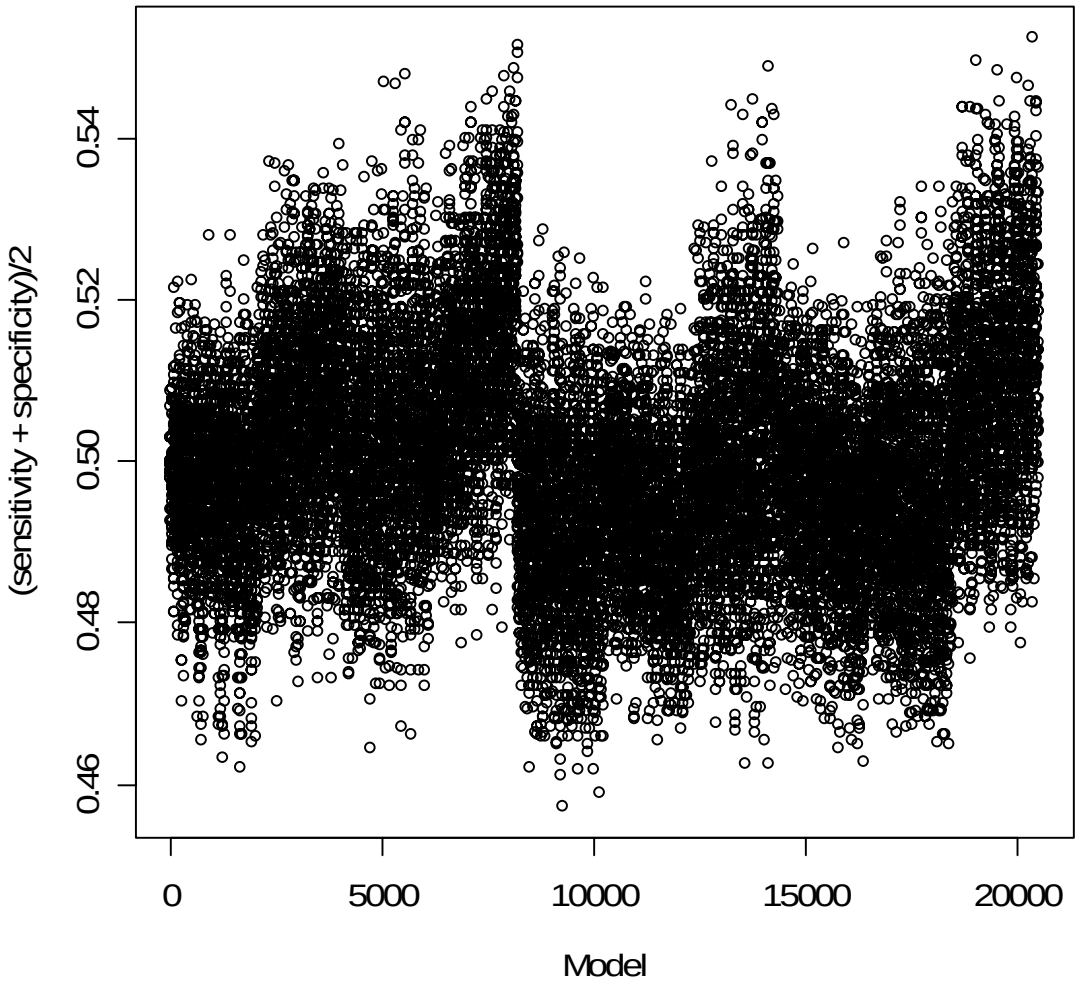

Figure 7: Random input predictability. For each model is shown the overall discriminatory accuracy [(sensitivity + specificity)/2].