

The Value of Economic Constraints in Boosting Equity Premium Prediction

Kendro Vincent

October 24, 2024

Department of Money and Banking
National Chengchi University

- Goyal and Welch (2008) suggest that out-of-sample R^2 are fairly low, even negative, for most linear predictive regression models for equity premium.
- Campbell and Thompson (2008) argue that:
 1. Low R^2 is still useful for investors
 2. Only positive equity premium prediction is relevant (to compensate for the risk aversion)
 3. Investors only use the “correct sign” of the estimated coefficients.

For example,

$$r_{t+1} = \alpha + \beta \times \text{Div. Yield}_t + \varepsilon_{t+1},$$

investors only use \hat{r}_{t+1} if $0 < \hat{\beta} < 1$ and $\hat{\alpha} > 0$.

- The value of economic constraints
 1. Pettenuzzo et al. (2014) develop Bayesian model that restrict the sign and Sharpe ratio.
 2. Li and Tsiakas (2017) apply shrinkage and sign constraints.
 3. Zhang et al. (2019) truncate extreme prediction.
- Machine learning in equity premium prediction
 1. Kelly et al. (2022) use random Fourier features as predictors.
 2. Shen and Xiu (2024) study how to use machine learning in weak signals environment.

Incorporate economic restriction to train gradient boosted regression tree (GBRT) model for equity premium prediction:

- Train GBRT with upper bound on predictability implied by asset pricing theory
- Impose monotonic constraint on the predictor
- Modify loss function to encourage positive prediction

Sample period: 1957–2022; Predictors: Goyal and Welch (2008)'s 14 variables; rolling-window: 120 months.

- Baseline method (validation sample): $R_{\text{oos}}^2 = 0.2\%$, Sharpe ratio (SR) = 0.33.
- Use fixed upper bound on R_{in}^2 : $R_{\text{oos}}^2 = 0.77\%$, SR = 0.4.
- Asymmetric loss function + fixed upper bound:
 $R_{\text{oos}}^2 = 1.25\%$, SR = 0.44.
- Asymmetric loss function + upper bound + sign constraint:
 $R_{\text{oos}}^2 = 1.63\%$, SR = 0.51.

The benchmark for economic value is based on the historical mean forecasts, the SR is 0.32 in this sample period.

Campbell and Thompson (2008) assume the investor maximizes the mean-variance utility function:

$$E_t[w_t \times r_{t+1}] - \frac{\gamma}{2} \text{var}_t(w_t \times r_{t+1}),$$

and the excess return process follows

$$r_{t+1} = \mu + x_t + e_{t+1},$$

where w_t is the portfolio weight on the risky asset, μ is the unconditional mean, x_t is the predictor with zero mean and variance σ_x^2 , e_{t+1} is the unpredictable noise with mean zero and variance σ_e^2 .

The solution to the portfolio choice problem: $w_t = \frac{E_t[r_{t+1}]}{\gamma \text{var}_t(r_{t+1})}$.

The agnostic investor, who does not use the information x_t , employs a fixed optimal portfolio weight

$$w = \frac{\mu}{\gamma \times (\sigma_x^2 + \sigma_e^2)}$$

and would earn on average (in excess of risk-free rate)

$$\frac{1}{\gamma} \left(\frac{\mu^2}{\sigma_x^2 + \sigma_e^2} \right) = \frac{S^2}{\gamma},$$

where S^2 is the unconditional Sharpe ratio of the risky asset.

The market-timing investor, who utilizes the information of x_t , uses the optimal weight

$$w_t = \frac{1}{\gamma} \left(\frac{\mu + x_t}{\sigma_e^2} \right)$$

and would earn on average

$$\frac{1}{\gamma} \left(\frac{\mu^2 + \sigma_x^2}{\sigma_e^2} \right) = \frac{1}{\gamma} \left(\frac{S^2 + R^2}{1 - R^2} \right).$$

The value of R^2

The market-timing investor beats the agnostic investor in terms of mean returns by

$$\frac{1}{\gamma} \left(\frac{R^2}{1 - R^2} \right) (1 + S^2)$$

The proportional increase in expected excess return is

$$\left(\frac{R^2}{1 - R^2} \right) \left(\frac{1 + S^2}{S^2} \right)$$

Suppose the unconditional Sharpe ratio is 0.4, and the predictable variation (R^2) is 0.05. The expected excess return could be improved by

$$\left(\frac{0.05}{0.95} \right) \left(\frac{1 + 0.16}{0.16} \right) = 38.16\%$$

- Review of GBRT and the default choice of hyperparameters
- Upper bound of in-sample R^2
- Sign constraints on predictors
- Asymmetric adjustment toward positive prediction

The GBRT model can be written as

$$r_{t+1} = \sum_{j=1}^J \eta \mathcal{T}_j(\mathbf{x}_t; \Theta^{(j)}) + \varepsilon_{t+1},$$

where the j -th tree \mathcal{T}_j can be written as a piece-wise constant function

$$\mathcal{T}_j(\mathbf{x}_t; \Theta^{(j)}) = \sum_{l=1}^L c_{l,j} I(\mathbf{x}_t \in S_{l,j}).$$

The model parameters are $\Theta^{(j)} = \{c_{l,j}, S_{l,j}\}_{l=1}^L, j = 1, \dots, J$, while $\{\eta, L, J\}$ are the hyperparameters.

At step k , we have:

- $F_{k-1}(\mathbf{x}_i) = \sum_{j=1}^{k-1} \eta \mathcal{T}_j(\mathbf{x}_i)$, the aggregated trees after $k - 1$ steps.
- $y_i^{(k)} = r_i - F_{k-1}(\mathbf{x}_i)$, the pseudo-residuals.

For a given loss function $\ell(y, f)$, we want to find

$$\mathcal{T}_k^* = \arg \min_{\mathcal{T}_k} \sum_{i=1}^N \ell(y_i^{(k)}, F_{k-1}(\mathbf{x}_i) + \eta \mathcal{T}_k(\mathbf{x}_i)),$$

such that the updated GBRT model is given by

$$F_k(\mathbf{x}_i) = F_{k-1}(\mathbf{x}_i) + \eta \mathcal{T}_k^*(\mathbf{x}_i).$$

1. $S_{l,j}$ is determined by the split points found by an approximate greedy algorithm to minimize the given objective function. For example, the histogram-based split finding algorithm in Ke et al. (2017).
2. $c_{l,j}$ is computed as follows,

$$\hat{c}_{l,k} = -\frac{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) g_{ik}}{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) h_{ik}},$$

where g_{ik} and h_{ik} are, respectively, the first- and second-order derivatives of $\ell(r, f)$ with respect to f and evaluated at F_{k-1} .

If we use the squared-error loss function, $\hat{c}_{l,k}$ would be

$$\frac{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) y_i^{(k)}}{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k})},$$

where $y_i^{(k)} = r_i - F_{k-1}(\mathbf{x}_i)$.

I use Ke et al. (2017)'s LightGBM and a fixed set of hyperparameters. The value of each hyperparameter is picked to encourage slow learning and obtain a high-bias tree.

- Maximum number of leaves $L = 4$
- Learning rate $\eta = 0.01$. The maximum possible number of trees $J_{\max} = 1/\eta$. The actual J is determined by the validation method or upper bound of return predictability.

I also use a version of stochastic gradient boosting proposed by Friedman (2002). The stochastic GBRT draws a random subsample *without replacement* as the training data to construct the tree in each step. The purpose of random subsampling:

- Decrease the correlation among the trees constructed in each iteration so that the total variance of the combined trees is reduced.
- GBRT model might be highly sensitive to the first few trees because the subsequent learning process depends heavily on the pseudo-residuals of the former trees.

- Random subsample (row-wise): 50%.
- Random predictor (column-wise): $\lfloor \sqrt{p} \rfloor$, where p is the total number of predictors. For example, if $p = 14$, then 3 predictors are randomly selected at each step.

Let \hat{f}_J denote the estimated GBRT with J trees, and its corresponding in-sample predictive power is measured by

$$R_{\text{in}}^2(J) = 1 - \frac{\sum_{\tau=t-N}^{t-1} (r_{\tau+1} - \hat{f}_J(\mathbf{x}_\tau))^2}{\sum_{\tau=t-N}^{t-1} (r_{\tau+1} - \bar{r})^2},$$

where $\bar{r} = \frac{1}{N} \sum_{\tau=t-N}^{t-1} r_{\tau+1}$.

The “optimal” number of trees is:

$$J^* = \max\{J : R_{\text{in}}^2(J) \leq \mathbf{Upper\ Bound}, J = 1, \dots, J_{\max}\},$$

where the **Upper Bound** is guided by asset pricing theory.

The cornerstone of asset pricing theory:

$$E[mr_i] = 0 \quad \text{for all } i,$$

where r_i is the excess return on a risky asset i and m is the **stochastic discount factor** (SDF). The existence of an SDF is equivalent to the law of one price.

Hansen–Jagannathan bound theorem states that the maximum attainable SR is bounded by the volatility of SDF. That is,

$$\max_i \frac{E(r_i)}{\sigma_i} \leq \frac{\sigma_m}{E(m)}.$$

Intuitively, the greater predictability would give us a higher SR.

Ross (2005) shows that

$$R_{in}^2 \leq (1 + r_f)^2 \sigma_m^2,$$

where r_f is the risk-free rate and σ_m^2 is the maximum bound for the variance of the stochastic discount factor (SDF). Ross (2005) assumes that σ_m is bounded by

$$5 \times \frac{\text{Volatility of S\&P 500 returns}}{1 + r_f},$$

so $R_{in}^2 \leq 25 \times \text{variance of S\&P 500 returns}$.

Suppose we use the above approximation and assume that the annual volatility of S&P 500 returns is 20%. If we use daily data to run the predictive regression, we can expect

$$R_{in}^2 \leq 25 \times 0.2^2 / 365 \approx 0.27\%.$$

With weekly data,

$$R_{in}^2 \leq 25 \times 0.2^2 / 52 \approx 1.92\%.$$

With monthly data,

$$R_{in}^2 \leq 25 \times 0.2^2 / 12 \approx 8.33\%.$$

1. Fixed upper bound:

$$J^* = \max\{J : R_{in}^2(J) \leq 8.33\%, J = 1, \dots, J_{max}\}.$$

2. Variance targeting:

$J^* = \max\{J : R_{in}^2(J) \leq 25 \times VOL_t^2, J = 1, \dots, J_{max}\}$, where VOL_t is the square root of the sum of squared daily returns on the S&P 500 index at the end of month t .

To reduce the effect of extreme values of realized volatility, set

$$\frac{0.15}{\sqrt{12}} \leq VOL_t \leq \frac{0.35}{\sqrt{12}}.$$

The variance target method implicitly assumes that we expect the extent of predictability to be greater when the market is in bad times, which is consistent with the model proposed by Cujean and Hasler (2017)

- Cochrane (1999) shows that the attainable unconditional squared Sharpe ratio (S_*^2) of a market timing strategy based on the model with R_{in}^2 is related to the unconditional buy-and-hold Sharpe ratio (S_0^2) by the following equation:

$$S_*^2 = \frac{S_0^2 + R_{in}^2}{1 - R_{in}^2}.$$

- If the unconditional average of monthly Sharpe ratio of 0.1 ($\approx 0.32/\sqrt{12}$), the upper bound of 8.33% for R_{in}^2 is equivalent to assigning 0.32 (1.1) as the upper bound for the monthly (annualized) SR of any market timing strategies.
- Pettenuzzo et al. (2014) also use the upper bound of 1 for the prior of SR in their Bayesian model.

Choosing the number of trees based on SR

- Assume that the greatest squared Sharpe ratio improvement is δ times the buy-and-hold squared Sharpe ratio.
- Calculate **Upper Bound** = $(\delta - 1)\tilde{S}_{0t}^2 / (1 + \delta\tilde{S}_{0t}^2)$, where \tilde{S}_{0t}^2 is the estimated buy-and-hold SR with the same sample size for estimating the GBRT model.
- In the empirical study, $\delta = 2$.
- We truncate the \tilde{S}_{0t}^2 on the left at 0.0025.

I call this SR targeting in the empirical analysis. In contrast to the variance targeting, it implicitly assumes that we expect more predictability when the past market condition is good.

Remark on the max SR improvement

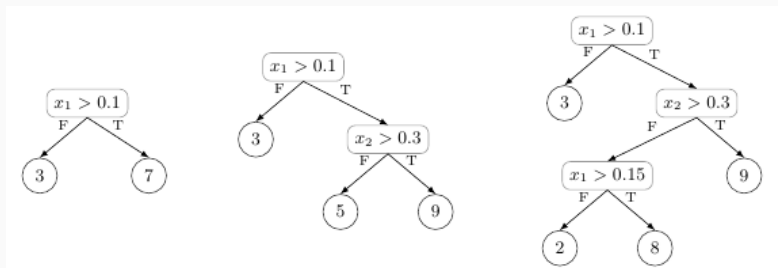
- The choice of δ is based on the consideration that the monthly time-varying S_{0t} could be as high as 0.2.
- The expected improvement with $\delta = 2$ implies the annualized conditional Sharpe ratio is $(12 \times 2)^{1/2} \times 0.2 = 0.98$.
Therefore, any multiplier greater than 2 may result in an unrealistic Sharpe ratio improvement.
- In the factor asset pricing theory, Ross (1976), Haddad et al. (2020), and Kozak et al. (2018) also rule out the investment opportunity which squared Sharpe ratio greater than twice the market portfolio's.

The GBRT enforces the sign constraint on a tree in two ways:

1. In the construction of the first split, it discards any split that violates the sign constraints.
2. Record the midpoint of the left and right nodes' values whenever they are used as optimal splitting variables. In the subsequent binary splitting, the algorithm would restrict the estimated parameters of the constrained variables within a boundary value based on the midpoint so that the sign constraints are satisfied at all nodes.

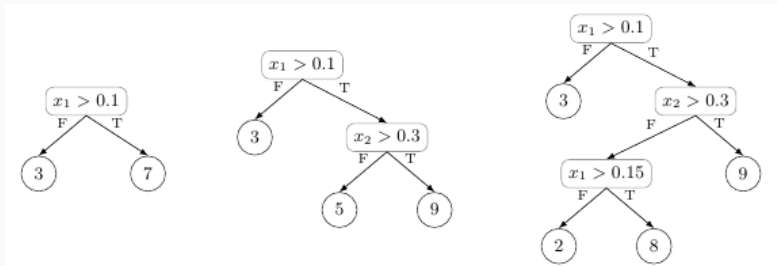
Enforcing sign constraints on predictors

To explain the second step, consider the following (unconstrained) tree example



Suppose we want the tree function to be increasing in x_1 and leave the relation between f and x_2 unconstrained.

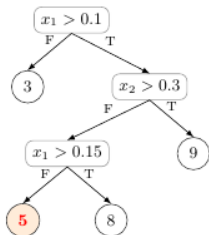
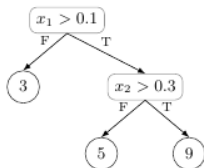
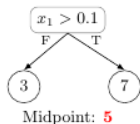
Enforcing sign constraints on predictors



The predicted value f is not increasing in x_1 for $x_2 \leq 0.3$.

- $f(\{x_1 \leq 0.1, x_2 \leq 0.3\}) = 3$.
- $f(\{0.1 \leq x_1 \leq 0.15, x_2 \leq 0.3\}) = 2$.

Enforcing sign constraints on predictors



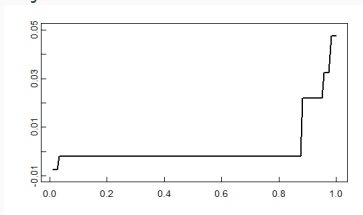
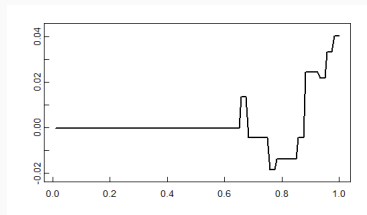
Lower bound is the midpoint recorded in the previous split

The predicted value f is now increasing in x_1 everywhere.

- $f(\{x_1 \leq 0.1, x_2 \leq 0.3\}) = 3$.
- $f(\{0.1 \leq x_1 \leq 0.15, x_2 \leq 0.3\}) = 5$.

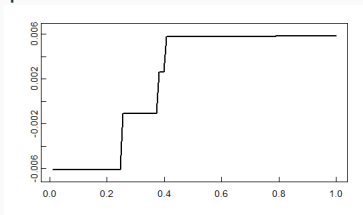
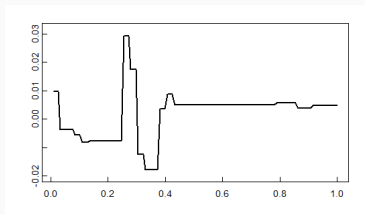
Illustration of enforcing sign constraints

Dividend yield



The target variable is the S&P500 index excess returns from February 1970 to January 1980.

Term spread



The target variable is the S&P500 index excess returns from February 1970 to January 1980.

I use the linear-exponential (linex) loss function, introduced by Varian (1975). The linex loss function of a predictor f and its actual value y can be expressed as

$$\ell(f, y; a) = e^{a(f-y)} - a(f-y) - 1,$$

where $a \neq 0$ is the shape parameter that determines the asymmetric cost of over- or underestimation.

- $a > 0$, the loss rises exponentially faster on the overestimation region. $a < 0$, underestimation would be more costly than overestimation.
- Campbell and Thompson (2008) suggest that we should shift the prediction toward positive values, i.e. pick $a < 0$ (penalizes underestimation more severely).

Two reasons for not completely ruling out negative equity premium forecasts.

- The truncated method implicitly discards the variation of equity premium and does not efficiently utilize the information from the given data.
- A rare but large negative shock could cause the expected equity premium to be negative. Such implied forecasts might still be useful for investors who want to insure against possible market crashes; see, e.g. Beason and Schreindorfer (2022).

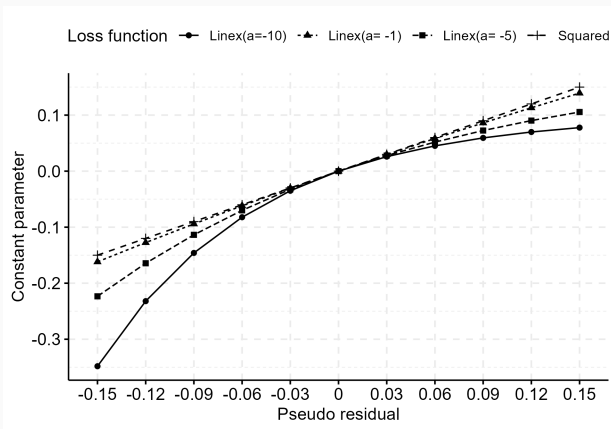
Recall that the constant parameter in a tree is:

$$\hat{c}_{l,k} = -\frac{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) g_{ik}}{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) h_{ik}}.$$

Under the linex loss,

$$\begin{aligned}\hat{c}_{l,k}^{\text{linex}} &= -\frac{1}{a} \left(\frac{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) \left(\exp \{a(F_{k-1}(\mathbf{x}_i) - r_i)\} - 1 \right)}{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) \exp \{a(F_{k-1}(\mathbf{x}_i) - r_i)\}} \right) \\ &= -\frac{1}{a} \left(1 - \left(\frac{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k}) e^{-a \times y_i^{(k)}}}{\sum_{i=1}^N I(\mathbf{x}_i \in S_{l,k})} \right)^{-1} \right),\end{aligned}$$

Linex loss parameter



For the typical values of the pseudo-residuals of monthly returns, the linex loss with a small value of a could be fairly similar to a symmetric loss. I set $a = -10$ in the empirical analysis.

Out-of-sample R^2 :

$$R_{\text{OOS}}^2 = 1 - \frac{\sum_{t=N}^{T-1} (r_{t+1} - \hat{r}_{t+1|t})^2}{\sum_{t=N}^{T-1} (r_{t+1} - \bar{r}_{t+1|t})^2},$$

where $\hat{r}_{t+1|t}$ and $\bar{r}_{t+1|t}$ are the monthly predictions of r_{t+1} at time t based on an estimated model and the historical mean benchmark, respectively.

To test $H_0 : R_{\text{os}}^2 \leq 0$, we first construct the time series

$$cw_t = (r_{t+1} - \bar{r}_{t+1|t})^2 - ((r_{t+1} - \hat{r}_{t+1|t})^2 - (\hat{r}_{t+1} - \bar{r}_{t+1|t})^2),$$

and then compute the *CW* test statistic as the mean of cw_t divided by its Newey-West standard error.

Clark and West (2007) show with simulation that the term $(\hat{r}_{t+1|t} - \bar{r}_{t+1|t})^2$ could be a good correction for adjusting the accumulated estimation uncertainty in $\hat{r}_{t+1|t}$ due to its greater number of parameters. With this adjustment, the *CW* statistic is **approximately standard normal distribution**.

To assess the economic value of the prediction $\hat{r}_{t+1|t}$, we adopt the same market timing experiment as in Farmer et al. (2023) and Kelly et al. (2022). In this experiment,

- An investor allocates $c \times \hat{r}_{t+1|t}$ of her capital into the equity and the rest into the risk-free asset.
- The monthly excess return on the portfolio is then given by $c \times \hat{r}_{t+1|t} \times r_{t+1}$.
- The normalizing constant c is assigned so the portfolio has the desired level of unconditional volatility. In the empirical analysis, we report the Sharpe ratio of the portfolio, so the economic performance measure is invariant to the normalizing constant.

- The data for the S&P 500 returns, risk-free rates, and economic predictors are downloaded from Prof. Amit Goyal's website.
- The sample covers the period from April 1957 to December 2022.
- Except for the inflation rate (INFL), the one-month lag of economic variables is used to construct the forecasts. The variable INFL is lagged by two months to account for the one-month announcement delay in the consumer price index report.

- The rolling windows of size 60 and 120 months are used to construct the one-month ahead returns.
- Benchmark method is early stopping with validation sample:
 - The first two-thirds is the training sample for estimation.
 - The last one-third is the validation sample for evaluating the objective function.
 - Early stopping round = 3.

Predictors

Variable	Definition	Sign	Reference
DP	Dividends to price ratio of the S&P 500 index	1	Campbell and Shiller (1988)
EP	Earnings to price ratio of the S&P 500 index	1	Campbell and Shiller (1988)
DE	Dividends to earnings ratio of the S&P 500 index	0	Campbell and Thompson (2008)
BM	Book-to-market ratio of the Dow Jones Industrial Average Index	1	Pontiff and Schall (1998)
VOL	Square root of the sum of squared daily returns on the S&P 500 index	-1	French et al. (1987)
NTIS	Aggregate net equity issues by the NYSE listed stocks	-1	Baker and Wurgler (2000)
TBL	Three-month Treasury-bill rate	-1	Campbell (1987)
LTY	Long-term government bond yield	0	Campbell (1987)
LTR	Long-term government bond returns	0	Campbell (1987)
TMS	Difference between long-term government bond yield and three-month Treasury-bill rate	1	Campbell (1987)
DFY	Difference between BAA and AAA-rated corporate bond yields	-1	Fama and French (1989)
DFR	Difference between corporate bond and long-term government bond returns	0	Campbell (1987)
INFL	Monthly rate of change of consumer price index	-1	Fama and Schwert (1977)
MLAG	One-month lag of the S&P 500 index excess returns	0	Kelly et al. (2022)

Main Empirical Results

	Out-of-sample R^2			
	Asymmetric		Symmetric	
	Unsigned	Signed	Unsigned	Signed
Panel A. 60-month rolling window.				
$R_{in}^2 \leq 7\%$	1.04***	1.17***	0.1	0.3
$R_{in}^2 \leq 8.33\%$	1.04***	1.23***	0.18	0.32
$R_{in}^2 \leq 10\%$	1.07***	1.27***	0.06	0.34
SR targeting	0.2**	0.25**	-0.24	-0.05
Var. targeting	0.97***	1.01***	-0.21	-0.23
Validation set	0.32**	0.55***	-0.21	-0.07
Panel B. 120-month rolling window.				
$R_{in}^2 \leq 7\%$	1.06***	1.52***	0.69**	0.7**
$R_{in}^2 \leq 8.33\%$	1.25***	1.63***	0.77**	0.73*
$R_{in}^2 \leq 10\%$	1.43***	1.64***	0.91**	0.69*
SR targeting	-0.29	-0.08	0.26*	0.43**
Var. targeting	1.61***	1.91***	0.68**	0.33
Validation set	0.49**	0.1	0.2	0.15

Note: The asterisks *, **, and *** indicate the hypothesis $H_0 : R_{oos}^2 \leq 0$ is rejected by CW statistic at 10%, 5%, and 1% significance levels.

Main Empirical Results

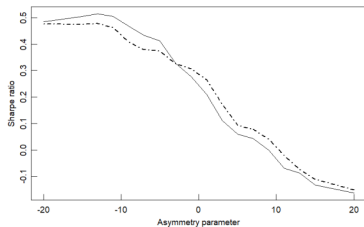
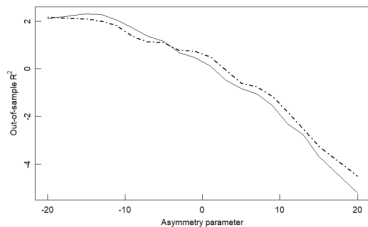
	Sharpe ratio			
	Asymmetric		Symmetric	
	Unsigned	Signed	Unsigned	Signed
Panel A. 60-month rolling window.				
$R_{in}^2 \leq 7\%$	0.3	0.32	0.27	0.28
$R_{in}^2 \leq 8.33\%$	0.29	0.32	0.27	0.28
$R_{in}^2 \leq 10\%$	0.29	0.33	0.26	0.28
SR targeting	0.1	0.11	0.23	0.24
Var. targeting	0.23	0.24	0.24	0.23
Validation set	0.13	0.21	0.24	0.26
Panel B. 120-month rolling window.				
$R_{in}^2 \leq 7\%$	0.42	0.51	0.39	0.4
$R_{in}^2 \leq 8.33\%$	0.44	0.51	0.4	0.41
$R_{in}^2 \leq 10\%$	0.46	0.51	0.41	0.4
SR targeting	0.16	0.22	0.34	0.37
Var. targeting	0.44	0.49	0.36	0.33
Validation set	0.34	0.28	0.33	0.33

Note: The Sharpe ratio using historical mean forecasts are 0.27 and 0.32 for the 60-month and 120-month rolling window, respectively.

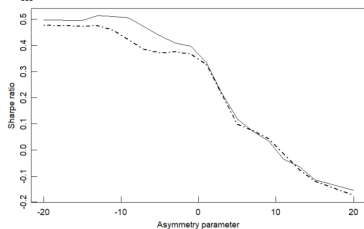
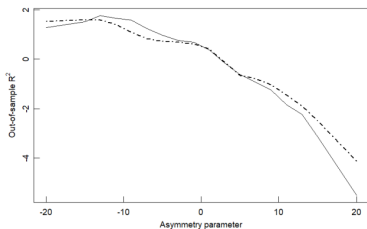
Sensitivity to linex parameter

Rolling window: 120 months. Solid line: sign constraints are imposed.

Time-varying bound: Variance targeting.



Fixed bound: $R_{in}^2 \leq 8.33\%$.



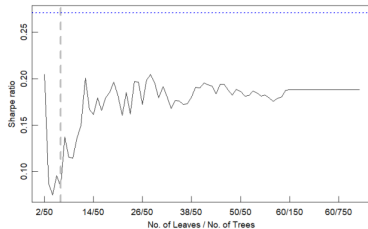
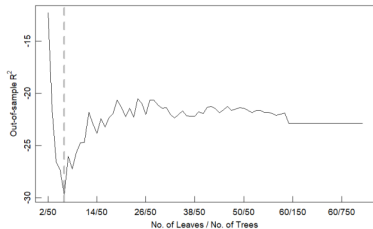
- Belkin et al. (2019) and Kelly et al. (2022) suggest that most ML models perform the worst surrounding interpolation regime.
- Once we move past further into the high complexity regime, then we might obtain increasing model accuracy that might even be better than the one below the interpolation threshold.

- GBRT needs different notion of high complexity because
 1. it is not possible to surpass the interpolation threshold with a single tree, where the maximum possible number of estimated constant parameters is $L = N$.
 2. the aggregation of multiple trees might have the effect of regularization.
- The interpolation threshold is defined as $R_{in}^2 = 90\%$.
- I follow similar approach by Curth et al. (2024):
 1. Set $\eta = 0.1$ and $J = 50$.
 2. Start from $L = 2$.
 3. After $L = N$, increase J gradually to 1,000.

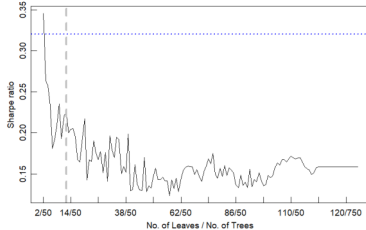
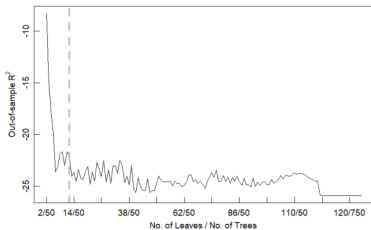
Performance of complex GBRT

Left: R_{OOS}^2 , right: SR (blue line is the benchmark).

Rolling window size = 60 months.

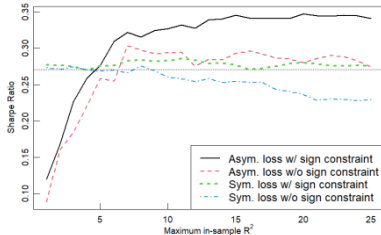
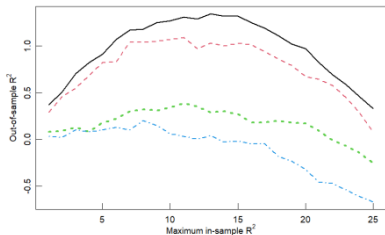


Rolling window size = 120 months.

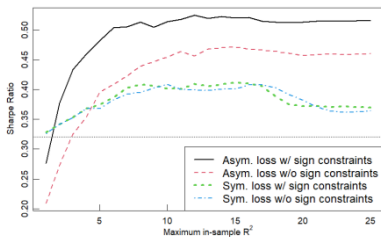
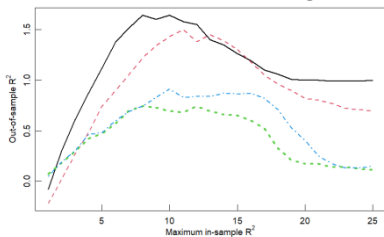


Sensitivity to upper bound (fixed case)

Rolling window size = 60 months.



Rolling window size = 120 months.



- Compute SHAP value decomposition

$$\hat{f}_t \equiv \hat{r}_{t+1|t} = \phi_0 + \sum_{m=1}^p \phi_m,$$

where ϕ_m is the SHAP value of predictor m and ϕ_0 is the baseline prediction.

- Define the drop- m prediction as

$$\hat{f}_t^{-m} = \phi_0 + \sum_{m' \neq m} \phi_{m'}.$$

- To evaluate the importance of predictor m :

$$\Delta R_{\text{oos}}^2(m) = R_{\text{oos}}^2 - \left(1 - \frac{\sum_{t=N}^{T-1} (r_{t+1} - \hat{f}_t^{-m})^2}{\sum_{t=N}^{T-1} (r_{t+1} - \bar{r}_{t+1|t})^2} \right),$$

Predictor importance

	$R_{in}^2 \leq 8.33\%$		Var. target	
	Signed	Unsigned	Signed	Unsigned
DP	0.438	0.299	0.582	0.405
EP	-0.195	-0.075	-0.157	-0.009
DE	0.055	0.038	-0.008	-0.025
BM	0.103	0.118	0.242	0.279
VOL	-0.012	-0.265	-0.043	-0.178
NTIS	-0.035	0.014	-0.065	-0.120
TBL	0.222	0.106	0.185	0.065
LTY	0.163	0.215	0.212	0.200
LTR	-0.021	-0.046	-0.021	-0.068
TMS	0.072	0.046	0.058	0.044
DFY	0.113	0.161	0.131	0.117
DFR	0.035	0.042	0.021	-0.001
INFL	-0.070	-0.057	-0.008	0.127
MLAG	0.012	-0.053	-0.047	-0.074

Note: Rolling window size is 120 months, loss function is linex.

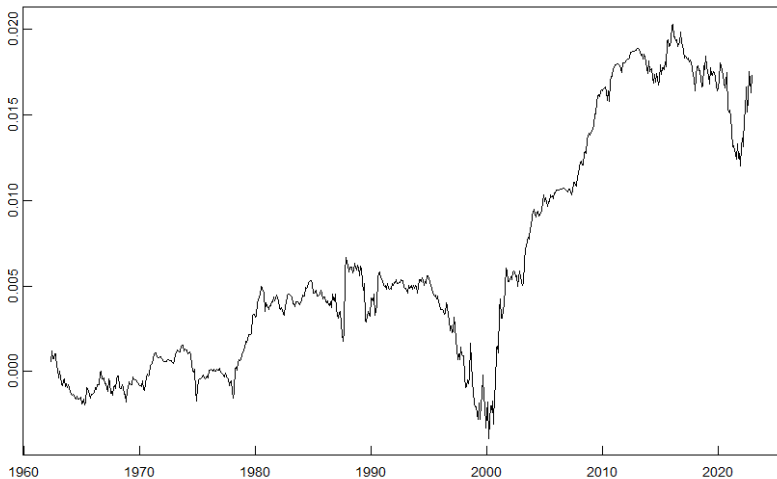
The cumulative difference of squared prediction error (CDSPE)

$$\text{CDSPE}_t = \sum_{\tau=N}^t (r_{\tau+1} - \bar{r}_{\tau+1|\tau})^2 - (r_{\tau+1} - \hat{r}_{\tau+1|\tau})^2.$$

The trending-upward CDSPE indicates that the proposed model predictions consistently outperform historical mean forecasts.

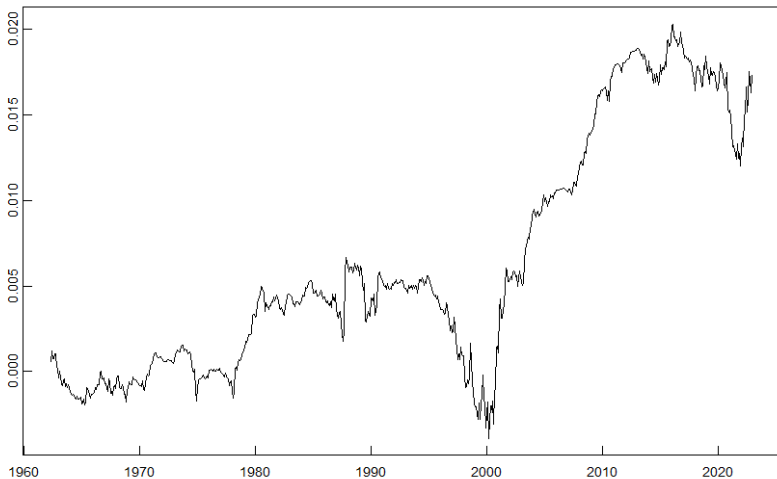
Time evolution of predictability

Rolling window size = 60 months. The GBRT model is estimated with asymmetric loss function and sign constraints on predictors. The optimal number of trees is selected by $R_{in}^2 \leq 8.33\%$ criterion.



Time evolution of predictability

Rolling window size = 120 months. The GBRT model is estimated with asymmetric loss function and sign constraints on predictors. The optimal number of trees is selected by $R_{in}^2 \leq 8.33\%$ criterion.



Similar CDSPE plot but with the historical mean forecast benchmark now replaced by *drop-m* prediction

$$cdspe_t(m) = \sum_{\tau=N}^t (r_{\tau+1} - \hat{f}_t^{-m})^2 - (r_{\tau+1} - \hat{f}_t)^2.$$

The trending $cdspe_t(m)$ would suggest predictor m becomes increasingly important.

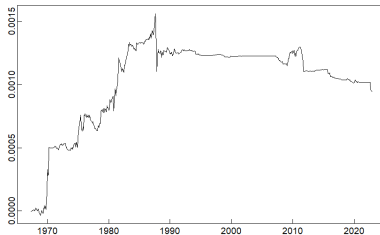
Time evolution of predictor importance

Rolling window size = 120 months.

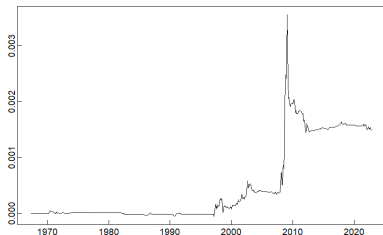
DP



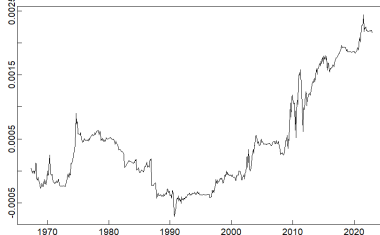
TMS



DFY



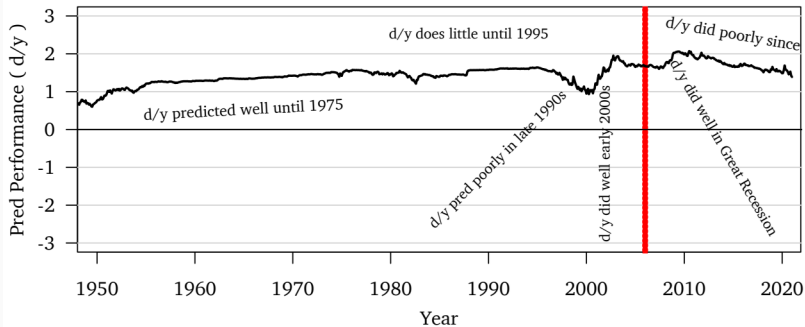
LTY



- The three economic constraints collectively help improve the GBRT for equity premium prediction.
 1. Asymmetric loss fn. plays a crucial role
 2. Var. target and a fixed upper bound perform equally well
- Empirical findings:
 1. Severe performance deterioration in the mid-1990 to early 2000.
 2. Dividend yield is the most important predictor in the multivariate GBRT model, even though it does not produce the best model as a lone predictor.

Outro

Goyal, Welch, and Zafirov (2024,RFS) "A comprehensive 2022 look at the empirical performance of equity premium prediction."



References

- Baker, M., & Wurgler, J. (2000). **The equity share in new issues and aggregate stock returns.** *Journal of Finance*, 55(5), 2219–2257.
- Beason, T., & Schreindorfer, D. (2022). **Dissecting the equity premium.** *Journal of Political Economy*, 130(8), 2203–2222.
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). **Reconciling modern machine-learning practice and the classical bias–variance trade-off.** *Proceedings of the National Academy of Sciences*, 116(32), 15849–15854.
- Campbell, J. Y. (1987). **Stock returns and the term structure.** *Journal of Financial Economics*, 18(2), 373–399.
- Campbell, J. Y., & Shiller, R. J. (1988). **Stock prices, earnings, and expected dividends.** *Journal of Finance*, 43(3), 661–676.
- Campbell, J. Y., & Thompson, S. B. (2008). **Predicting excess stock returns out of sample: Can anything beat the historical average?** *Review of Financial Studies*, 21(4), 1509–1531.
- Clark, T. E., & West, K. D. (2007). **Approximately normal tests for equal predictive accuracy in nested models.** *Journal of Econometrics*, 138(1), 291–311.
- Cochrane, J. H. (1999). **Portfolio advice for a multifactor world.** *NBER working paper*.
- Cujean, J., & Hasler, M. (2017). **Why does return predictability concentrate in bad times?** *Journal of Finance*, 72(6), 2717–2758.
- Curth, A., Jeffares, A., & van der Schaar, M. (2024). **A U-turn on double descent: Rethinking parameter counting in statistical learning.** *Advances in Neural Information Processing Systems*, 36.
- Fama, E. F., & French, K. R. (1989). **Business conditions and expected returns on stocks and bonds.** *Journal of Financial Economics*, 25(1), 23–49.
- Fama, E. F., & Schwert, G. W. (1977). **Asset returns and inflation.** *Journal of Financial Economics*, 5(2), 115–146.
- Farmer, L. E., Schmidt, L., & Timmermann, A. (2023). **Pockets of predictability.** *Journal of Finance*, 78(3), 1279–1341.
- French, K. R., Schwert, G. W., & Stambaugh, R. F. (1987). **Expected stock returns and volatility.** *Journal of Financial Economics*, 19(1), 3–29.
- Friedman, J. H. (2002). **Stochastic gradient boosting.** *Computational Statistics & Data Analysis*, 38(4), 367–378.
- Goyal, A., & Welch, I. (2008). **A comprehensive look at the empirical performance of equity premium prediction.** *Review of Financial Studies*, 21(4), 1455–1508.
- Haddad, V., Kozak, S., & Santosh, S. (2020). **Factor timing.** *Review of Financial Studies*, 33(5), 1980–2018.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). **LightGBM: A highly efficient gradient boosting decision tree.** *Advances in neural information processing systems*, 30.
- Kelly, B., Malamud, S., & Zhou, K. (2022). **The virtue of complexity everywhere.** *Working paper*.
- Kozak, S., Nagel, S., & Santosh, S. (2018). **Interpreting factor models.** *Journal of Finance*, 73(3), 1183–1223.
- Li, J., & Tsiakas, I. (2017). **Equity premium prediction: The role of economic and statistical constraints.** *Journal of Financial Markets*, 36, 56–75.
- Pettenuzzo, D., Timmermann, A., & Valkanov, R. (2014). **Forecasting stock returns under economic constraints.** *Journal of Financial Economics*, 114(3), 517–553.
- Pontiff, J., & Schall, L. D. (1998). **Book-to-market ratios as predictors of market returns.** *Journal of Financial Economics*, 49(2), 141–160.
- Ross, S. A. (1976). **The arbitrage theory of capital asset pricing.** *Journal of Economic Theory*, 13(3), 341–360.
- Ross, S. A. (2005). **Neoclassical finance.** Princeton University Press.
- Shen, Z., & Xiu, D. (2024). **Can machines learn weak signals?** *Working paper*.
- Varian, H. R. (1975). **A Bayesian approach to real estate assessment.** In S. E. Fienberg & A. Zellner (Eds.), *Studies in bayesian econometrics and statistics in honor of leonard j. savage* (pp. 195–208). North-Holland.
- Zhang, Y., Wei, Y., Ma, F., & Yi, Y. (2019). **Economic constraints and stock return predictability: A new approach.** *International Review of Financial Analysis*, 63, 1–9.