RELEVANCE-BASED IMPORTANCE:

A COMPREHENSIVE MEASURE OF VARIABLE IMPORTANCE IN PREDICTION

THIS VERSION: December 10, 2024

Megan Czasonis, Mark Kritzman, and David Turkington

Megan Czasonis is a managing director at State Street Associates in Cambridge, MA.
mczasonis@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Mark Kritzman is the chief executive officer of Windham Capital Management in Cambridge, MA, and a senior lecturer at the MIT Sloan School of Management in Cambridge, MA.
kritzman@mit.edu
100 Main Street, Cambridge MA, 02142

David Turkington is senior managing director and head of State Street Associates in Cambridge, MA.
dturkington@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Key Takeaways

A t-statistic measures the importance of variables to a prediction when the prediction is formed from linear regression analysis. However, it is difficult to interpret if the predictive variables are collinear, and it is impossible to interpret if the relationship shifts as conditions change.

A Shapley value measures the importance of variables used in machine learning models, but it only considers reliability on average across all predictions. It does not account for a variable's contribution to the reliability of individual predictions.

Relevance-based importance gives a measure of variable importance that is robust to collinearity and conditionality. It accounts for a variable's contribution to reliability on average across all predictions as well as its contribution to the reliability of individual predictions.

Abstract

The notion of variable importance is not uniquely defined. If a prediction is formed from a linear regression model, it is common to measure variable importance as a t-statistic, but a t-statistic is difficult to interpret if the predictive variables are collinear, and it is uninterpretable if the relationship between the predictive variables and the outcomes shifts as conditions change. A Shapley value measures variable importance when a prediction is formed from machine learning models. It is robust to collinearity and conditionality, but it does not account for a variable's contribution to the reliability of individual predictions. It only considers a variable's contribution to the reliability of predictions on average across all predictions. The authors introduce a new measure of variable importance, called relevance-based importance that, unlike a t-statistic, is robust to collinearity and conditionality and, unlike a Shapley value, accounts for a variable's contribution to the reliability of individual predictions. The authors show that in the special case in which the predictive variables are uncorrelated with one another and the relationship remains constant, relevance-based importance provides the same information as a t-statistic when averaged across all predictions. They also show that when relevance-based importance is averaged across all predictions, it converges to the Shapley value where the chosen value function is the R-squared of a linear regression model.

RELEVANCE-BASED IMPORTANCE:

A COMPREHENSIVE MEASURE OF VARIABLE IMPORTANCE IN PREDICTION

When we form a prediction from data, we must select predictive variables that we believe to be important to the prediction. We typically lean on theory, observation, and intuition to guide us initially. But we also seek to validate our assessment of a variable's importance by observing what the data reveals to us. Validation of a variable's importance can be challenging, though, because the notion of importance is not uniquely defined and because some definitions of importance are difficult or even impossible to interpret under certain conditions.

We introduce a new measure of variable importance, called relevance-based importance, hereafter referred to as RBI, that overcomes important limitations of widely used alternative measures. We demonstrate mathematically and by simulation that RBI compares favorably to a t-statistic, which is the conventional measure of variable importance for linear regression models, and it compares favorably to a Shapley value, which is used to measure variable importance in machine learning models.

We do not consider causality in our analysis of variable importance because our interest is with prediction, not how best to intervene in a relationship to change an outcome. Moreover, causality does not affect the issues we address in our analysis.

We proceed as follows. We first describe relevance-based prediction and show how RBI is a natural byproduct of this prediction method. We then introduce RBI. Next, we compare RBI to a t-statistic. We start by considering the special case in which the predictive variables are uncorrelated with each other and the relationship between the predictive variables and the

outcomes is constant across the full sample of observations. This special case allows us to connect RBI's prediction-specific measure of variable importance to a t-statistic's measure of average importance. We then relax the assumption that the predictive variables are uncorrelated and show how RBI accounts for collinearity more carefully than a t-statistic. We then consider relationships between predictive variables and outcomes that shift as conditions change and discuss how RBI, unlike a t-statistic, accounts for this conditionality. Next, we compare RBI to a Shapley value, which like RBI, accounts for collinearity and conditionality. We show that both measures account for a variable's contribution to reliability on average across all predictions, but that only RBI accounts for a variable's contribution to the reliability of individual predictions. We then present an empirical analysis of RBI for a set of variables used to predict stock market volatility. We conclude with a summary.

Relevance-Based Prediction

RBI is a natural byproduct of relevance-based prediction. Therefore, to understand how RBI is computed and to appreciate its useful properties, we first describe relevance-based prediction, which has three key components: relevance, fit, and grid prediction.[1]

Relevance

Relevance is a precise statistical measure of the importance of an observation to a prediction. It is composed of similarity and informativeness, which are both measured as Mahalanobis distances, as shown by Equations 1 through 4.[2]

$$r_{it} = sim(x_i, x_t) + \frac{1}{2}\big(info(x_i, \bar{x}) + info(x_t, \bar{x})\big) \quad (1)$$

$$sim(x_i, x_t) = -\frac{1}{2}(x_i - x_t)\Omega^{-1}(x_i - x_t)' \quad (2)$$

$$info(x_i, \bar{x}) = (x_i - \bar{x})\Omega^{-1}(x_i - \bar{x})' \quad (3)$$

$$info(x_t, \bar{x}) = (x_t - \bar{x})\Omega^{-1}(x_t - \bar{x})' \quad (4)$$

In Equations 1 through 4, x_i is a vector of the values of K predictive variables for a prior observation, x_t is a vector of the values of the predictive variables for a specific prediction task, $\bar{x} = 1_N 1_N' X N^{-1}$ is the average of the predictive variables across all observations, and Ω^{-1} is the inverse covariance matrix of all the observations of the variables. The vector $(x_i - x_t)$ measures how distant the observations are independently from the circumstances of the prediction task. By multiplying this vector by the inverse covariance matrix, we capture the interaction of the predictive variables, and at the same time we standardize the distances by dividing by variance. By multiplying this product by the transpose of the vector $(x_i - x_t)$ we consolidate the outcome into a single number. All else being equal, observations that are like current circumstances but different from average circumstances are more relevant than those that are not.

This definition of relevance is not arbitrary. We know from information theory that the information contained in an observation is the negative logarithm of its likelihood.[3] We also know from the Central Limit Theorem that the relative likelihood of an observation from a multivariate normal distribution is proportional to the exponential of a negative Mahalanobis distance. Therefore, the information contained in a point on a multivariate normal distribution is proportional to a Mahalanobis distance.

Relevance-based prediction forms a prediction as a weighted average of prior outcomes for Y.

$$\hat{y}_t = \sum_{i=1}^{N} w_{it} y_i \tag{5}$$

If we define weights in terms of relevance as follows, which admits the relevance-weighted average of every prior outcome in the observed data sample, the result is precisely equivalent to the prediction that results from linear regression analysis.[4]

$$w_{it,linear} = \frac{1}{N} + \frac{1}{N-1} r_{it} \tag{6}$$

In most cases, however, we can produce a more reliable prediction by censoring the observations that are less relevant than a chosen threshold, which leads to the following definition of prediction weights.

$$w_{it,retained} = \frac{1}{N} + \frac{\lambda^2}{n-1} (\delta(r_{it}) r_{it} - \varphi \bar{r}_{sub}) \tag{7}$$

$$\delta(r_{it}) = \begin{cases} 1 & if\ r_{it} \geq r^* \\ 0 & if\ r_{it} < r^* \end{cases} \tag{8}$$

$$\lambda^2 = \frac{\sigma^2_{r,full}}{\sigma^2_{r,retained}} = \frac{\frac{1}{N-1}\sum_{i=1}^{N} r_{it}^2}{\frac{1}{n-1}\sum_{i=1}^{N} \delta(r_{it}) r_{it}^2} \quad (9)$$

In Equations 6 through 9, $n = \sum_{i=1}^{N} \delta(r_{it})$ is the number of observations that are fully retained, $\varphi = n/N$ is the fraction of observations in the retained sample, and $\bar{r}_{sub} = \frac{1}{n}\sum_{i=1}^{N} \delta(r_{it}) r_{it}$ is the average relevance value of the observations in the retained sample. It is important to note that $w_{it,retained}$ depends crucially on the prediction circumstances x_t. Relevance is reassessed for each prediction circumstance which further affects the identification of the retained subsample and introduces nonlinear conditional dependence of the prediction $\hat{y}_t$ on the prediction circumstances x_t. The scaling factor λ^2 compensates for a bias that would otherwise result from relying on a small subsample of highly relevant observations. In the case of linear regression analysis $n = N$ and $\lambda^2 = 1$. Lastly, note that the regression weights always sum to 1.[5]

Fit

Fit reveals how much confidence we should have in a specific prediction task, separately from the confidence we have in the overall prediction routine. In addition, fit provides a principled way to evaluate the relative merits of alternative calibrations for each prediction task.

Consider a pair of observations that are used to form a prediction. Each observation has a weight and an outcome. We are interested in the alignment of the weights of the two observations with their outcomes. We first standardize them by subtracting the average value and dividing this difference by standard deviation – in essence, converting them to z-scores.

We then measure their alignment by taking the product of these standardized values. If the product is positive, their relevance is aligned with their outcomes, and the larger the product, the stronger the alignment. We perform this calculation for every pair of observations in our sample. We should also note that all the formulas we have thus far considered for weights rely only on relevance, which in turn relies only on the x_is, the x_t, and the $\bar{x}$. They do not use any of the information from observed outcomes. To determine fit, however, we must consider outcomes (the y_is).

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j z_{w_{it}} z_{w_{jt}} z_{y_i} z_{y_j} \tag{10}$$

Equation 11 intuitively describes fit as the squared correlation of relevance weights and outcomes, which conceptually matches the notion of the conventional R-squared statistic. As we soon show, this connection of fit to R-squared is critically important.

$$fit_t = \rho(w_t, y)^2 \tag{11}$$

Although we compute fit from the full sample of observations, the weights that determine fit vary with the threshold we choose to define the relevant subsample. As we focus the subsample on observations that are more relevant, we should expect the fit of the subsample to increase, but we should also expect more noise as we shrink the number of observations. The fit across pairs of all observations in the full sample implicitly captures this tradeoff between subsample fit and noise by overweighting observations that are more relevant and underweighting observations that are less relevant.

Like relevance, fit is not arbitrary. In the case of linear regression analysis with $n = N$, the informativeness-weighted average fit across all prediction tasks in the observed sample equals R-squared.[6]

$$R^2 = \frac{1}{N-1}\sum_{t=1}^{N} info(x_t, \bar{x}) fit_t \quad (12)$$

Censoring observations that fall below a relevance threshold is more effective to the extent there is asymmetry between the fit of the weights formed from the retained subsample of observations and the fit of the weights formed from the complementary set of censored observations. We measure asymmetry between the fit of the retained and censored subsamples as shown by Equation 13. The $(+)$ superscript designates weights formed from the retained observations while the $(-)$ superscript designates weights formed from the censored observations. Asymmetry recognizes the benefit of censoring non-relevant observations that contradict the predictive relationships that exist among the relevant observations. This assessment also inherently considers the relative sample sizes of the two subsamples.

$$asymmetry_t = \frac{1}{2}\left(\rho\left(w_t^{(+)}, y\right) - \rho\left(w_t^{(-)}, y\right)\right)^2 \quad (13)$$

To calculate adjusted fit, we add asymmetry to fit and multiply this sum by K, the number of predictive variables included in the prediction, as shown by Equation 14. Multiplication by the number of predictive variables allows us to compare predictions based on different numbers of predictive variables. Adjusted fit recognizes that we are more likely to observe a spurious relationship from prediction weights based on just one or a few variables than we are based on a collection of many variables.

$$adjusted\ fit_t = K(fit_t + asymmetry_t) \quad (14)$$

Grid Prediction

Grid prediction employs a grid in which the columns represent different combinations of predictive variables, and the rows represent subsamples of observations determined by different relevance thresholds. Each cell contains a prediction and an associated adjusted fit. The assessment of reliability using adjusted fit occurs before the prediction is rendered and the subsequent outcome is known. Grid prediction forms a composite prediction as a reliability-weighted average of the predictions from all possible calibrations. Equation 15 defines reliability weights, ψ_θ, as the adjusted fit for a parameter calibration, θ, divided by the sum of all adjusted fits across all parameter calibrations.

$$\psi_\theta = \frac{adjusted\ fit_\theta}{\sum_{\tilde{\theta}} adjusted\ fit_{\tilde{\theta}}} \quad (15)$$

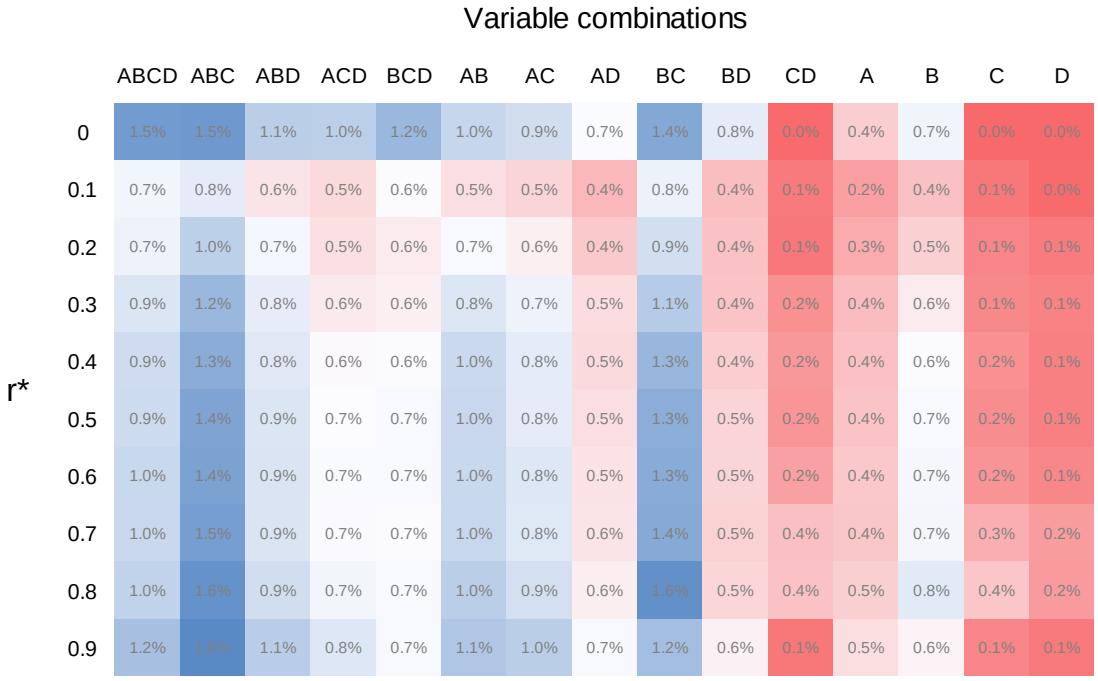
Equation 16 describes the composite prediction.

$$\hat{y}_{t,grid} = \sum_\theta \psi_\theta \hat{y}_{t,\theta} \quad (16)$$

Exhibit 1 gives a visual representation of grid prediction based on a contrived data set of four predictive variables and 400 randomly simulated observations. The columns represent different subsets of variables, and the rows represent different subsamples of observations as determined by different relevance thresholds. Each cell represents a calibration θ; that is, a unique combination of retained predictive variables and retained observations. The values shown in the cells in Exhibit 1 are the weights (ψ_θ) we apply to the calibration-specific

predictions to form the composite grid prediction. Blue cells are more important to forming the prediction than red cells. The values in the grid are specific to each prediction task.

Exhibit 1: Grid Prediction – Illustrative Example



Note that each cell's prediction is a linear function of observations, and the grid prediction is a linear function of each cell's prediction. Therefore, we can express the grid prediction in terms of composite weights applied to each observation, as shown by Equation 17. Composite weights are important because they preserve the transparency of each observation's contribution to the current prediction task, and they allow us to calculate fit from composite weights as a final gauge of the grid prediction's reliability.

$$w_{it,grid} = \sum_{\theta} \psi_{\theta} w_{it,\theta} \tag{17}$$

RBI

RBI measures a variable's total contribution to the reliability of a specific prediction. All the information needed to compute RBI is contained within the prediction grid.

As shown by Equation 18, RBI_{tk} for prediction t and variable k is computed as the weighted average adjusted fit for grid cells that contain k (for which the variable censoring indicator $\Delta_k(\theta) = 1$) minus the weighted average adjusted fit for cells that do not contain k (for which $\Delta_k(\theta) = 0$). We express RBI as a sum over all grid cells θ.

$$RBI_{tk} = \sum_\theta \alpha_\theta \frac{\Delta_k(\theta)(adjusted\ fit_{t\theta}) - (1-\Delta_k(\theta))(adjusted\ fit_{t\theta})}{\sum_{\tilde{\theta}} \Delta_k(\tilde{\theta})} \quad (18)$$

The term $\sum_{\tilde{\theta}} \Delta_k(\tilde{\theta})$ counts the number of cells that include variable k. For a grid that includes every variable combination, this number is nearly equal to the number of cells that do not include variable k, but the counts are not identical unless we include a column in the grid for predictions that do not use any of the X variables (for which adjusted fit is always zero). Thus, we divide by the number of cells that include variable k regardless of whether a given cell contains k or not.

Alternatively, we may interpret RBI as a weighted average of the differences in adjusted fit for pairs of cells that are otherwise identical in specification other than the inclusion of variable k in one case. We also include a scaling term, α_θ, which adjusts the otherwise equal weights defined by $\sum_{\tilde{\theta}} \Delta_k(\tilde{\theta})$. This scaling term accounts for the fraction of all cells that include the same number of non-k variables as the current cell, which for notational simplicity we denote simply as $fraction_\theta$. To the extent this fraction differs from $1/K$, the influence of the

cell is either increased or decreased. This scaling term boosts the influence of subsets of variables that are more sparsely represented in the grid, so that each size subset has equal influence on the result. As we will describe later, this adjustment enables RBI to converge to the Shapley value formula. However, we show that in practice it gives results that are nearly identical to results based on the much simpler weighting scheme of $\alpha_\theta = 1$.

$$\alpha_\theta = \frac{\frac{1}{K}}{fraction_\theta} \tag{19}$$

The value of RBI may be positive, zero, or negative. A positive value indicates that a variable adds value to the reliability of a prediction, and the higher the value the more substantial its contribution. A zero, or near-zero, value indicates that a variable contributes benign noise to a prediction. A negative value indicates that a variable contributes harmful noise to a prediction by generally obscuring the effects of otherwise compelling relationships.

Properties of RBI

The prediction grid reveals several useful properties of RBI.

- **RBI is prediction specific.** RBI is calculated from the adjusted fits of the predictions associated with the grid cells. The adjusted fits are calculated from the relevance values of the observations. These values are determined by the specific circumstances of each prediction task. Therefore, RBI explicitly considers the specific circumstances of each individual prediction task.

- **RBI measures a variable's total importance.** RBI is calculated from the grid columns, which collectively account for every configuration in which a variable is used to form a prediction irrespective of the usage of the other variables. A variable can be recognized for containing useful predictive information even if the same information overlaps with other variables. RBI, therefore, measures each variable's total contribution to a prediction's reliability.
- **RBI accounts for conditionalities.** The top row of the grid uses the full sample of observations. The predictions formed by these cells reflect the unconditional (linear) part of the relationship between the predictive variables and the outcomes across the full sample of observations. The remaining rows of the grid use subsamples of observations to form predictions from the premise that conditions change and cause the relationship to shift from its unconditional pattern. Although the relationship between the predictive variables and the outcomes may not be consistent across the full sample of observations, it is more consistent within these condition-specific subsamples. To the extent the adjusted fits of the predictions given by the cells below the first row are relatively high, RBI accounts for conditionalities that defy independent relationships across variables.
- **RBI accounts for a prediction's reliability**. RBI is calculated from adjusted fit which measures the alignment of relevance and outcomes across all pairs of observations that go into a prediction. This alignment gives a measure of a prediction's reliability, which means that RBI explicitly accounts for reliability as opposed to just the magnitude of a prediction.

We next explore the connection of RBI to a t-statistic.

RBI Compared to a t-statistic

A t-statistic measures the estimated average response of the outcome Y that is being predicted to a one-unit change in a predictive variable, divided by the standard error of the estimated average response, as shown by Equation 20.

$$t_k = \frac{\widehat{\beta}_k}{SE(\widehat{\beta}_k)} \tag{20}$$

In Equation 20, $\hat{\beta}_k$is the observed linear regression coefficient of the predictive variable, and $SE(\hat{\beta}_k)$ equals the standard error of $\hat{\beta}_k$.[7]

A t-statistic's main virtue is that it accounts for the reliability of the predictions because it is estimated from the predicted values as well as the outcomes. A t-statistic is limited in three critical ways, however. (1) It only measures average importance across all predictions; it does not measure the importance of a variable to an individual prediction. (2) It only measures a variable's marginal importance and not its total importance. (3) It fails to address conditionalities, which is to say, it is uninterpretable if a relationship shifts away from its typical average pattern when conditions change.

RBI shares a t-statistic's main virtue in that it too considers reliability. But it does not suffer from the limitations of a t-statistic, as is evident from the earlier discussion of RBI's useful properties.

A t-statistic's inability to measure total importance deserves further consideration. This limitation arises when the predictive variables are collinear, even if there are no conditionalities that complicate the overall relationship between the predictive variables and the outcomes. If the predictive variables are uncorrelated with each other, a t-statistic does measure the total importance of a predictive variable. This special case gives rise to an important near equivalence. An informativeness-weighted average of RBI across all prediction tasks nearly converges to a squared t-statistic in the absence of collinearity and conditionality.

$$\tau_k^2 = \frac{(N-K)}{1-R^2} \sum_t \omega_t RBI_{tk} \tag{21}$$

$$\omega_t = \frac{info(x_t)}{K(N-1)} \tag{22}$$

In Equation 21, τ_k^2 equals the average RBI approximation to the squared t-statistic t_k^2, and in Equation 22, $info(x_t)$ is from Equation 7 and is computed using all K predictive variables.

This near equivalence of average RBI and a t-statistic in this special case reveals a t-statistic to be a special case of average RBI, which reconciles their seemingly different interpretations. We defined RBI as a predictive variable's contribution to the reliability of a prediction, while we defined a t-statistic as the sensitivity of the outcome being predicted to a one-unit change in the predictive variable divided by the standard error of this sensitivity. In the absence of collinearity, these definitions converge. However, if the predictive variables are collinear, a t-statistic's definition of importance holds only as a measure of marginal importance, whereas RBI always measures total importance.

This comparison of total importance and marginal importance merits further explanation. As we show in Appendix B, we can express a squared t-statistic equivalently as a function of R-squared.

$$t_k^2 = \frac{(N-K)}{1-R^2}\left(R^2 - R_{\setminus k}^2\right) \tag{23}$$

In Equation 23, R^2 applies to a prediction that uses all K predictive variables, whereas $R_{\setminus k}^2$ apples to a prediction that uses all the variables except variable k. As we mentioned earlier, RBI measures total importance because it is calculated from every combination of predictive variables. As Equation 23 shows, a squared t-statistic expressed as a function of R-squared only considers R-squared with all the variables and R-squared with all the variables but variable k; hence, it only measures marginal importance. However, we can expand Equation 23 to consider all marginal differences in R-squared where variable k is added to each cohort of variables including the null set, rather than only the cohort that already includes all the other variables, as shown by Equation 24.

$$\tau_{k,Rsquared}^2 = \frac{(N-K)\sum_\theta \alpha_\theta\left(\frac{\Delta_k(\theta)R_\theta^2 - \left(1-\Delta_k(\theta)\right)R_\theta^2}{\sum_{\tilde{\theta}}\Delta_k(\tilde{\theta})}\right)}{1-R^2} \tag{24}$$

The sum in the numerator mirrors that of RBI from Equation 18, except that the expression pertaining to R-squared values applies to a summary assessment of predictions across all tasks rather than the adjusted fit of a specific prediction task. Note that this concept, which naturally derives from grid prediction, equates to the Shapley method of assigning value which we will next discuss.

Empirically, we observe that the value τ_k^2 which is aggregated from individual prediction variable importance is nearly equal to $\tau_{k,Rsquared}^2$ which is computed at the level of an entire linear regression model. Therefore, RBI_{tk} offers a prediction-level decomposition of $\tau_{k,Rsquared}$ analogous to how fit offers a prediction-level decomposition of R-squared for a single linear regression.[8,9]

We can generalize the prediction from a single linear regression cell to a grid prediction that includes a row in the grid: one linear regression cell for each variable combination. Though the grid predictions formed from a one-row grid are not identical to those from a one-cell linear regression that includes every variable, they are highly similar.

Simulation of Variable Importance with Uncorrelated Variables

To support our claim of the convergence of RBI to a t-statistic when the predictive variables are uncorrelated with each other, we simulate 1,000 observations from prespecified distributions and then calculate variable importance as measured by RBI and by a t-statistic. In this simulation, as well as the ones that follow, we assume that the relationship between the predictive variables and the outcomes is constant across the full sample of observations. We do not allow for conditionalities that would alter the relationship.

We begin by considering three X variables that follow uncorrelated standard normal distributions. We compute Y outcomes using linear betas of 1 for each X variable and then add independent random noise with a standard deviation of 5.

The following exhibit shows the t-statistics of the three X variables. The simulation contains randomness, so the variables are not all equally important to explaining the outcomes. Nevertheless, all are highly statistically significant. The τ_k values are similar to the t-statistics; however, they are slightly higher. Note that the τ_k and $\tau_{k,Rsquared}$ values are nearly identical. The R-squared decomposition at the model level consists of three values that sum to precisely the R-squared of a one-cell linear regression that includes all the predictive variables.

Exhibit 2: Aggregate Variable Importance
Correlations Equal 0

	X1	X2	X3
R-squared decomposition (model level)	0.0461	0.0497	0.0430
Informativeness-weighted RBI	0.0469	0.0491	0.0430
Linear regression t-statistic	7.00	7.51	6.94
RBI τ-statistic	7.36	7.53	7.04
RBI τ-statistic (simplified weighting)	7.37	7.53	7.05

Simulation of Variable Importance with Correlated Variables

Exhibit 3 shows how variable importance measured by RBI and a t-statistic diverge when the predictive variables are correlated with one another. In this simulation, we draw X values from a multivariate normal distribution where $X1$ and $X2$ are 90 percent correlated. $X3$ remains uncorrelated to the other variables. The t-statistics are much lower for $X1$ and $X2$, because each one largely substitutes for the other one, so neither one registers as important. The τ_k measure, however, registers $X1$ and $X2$ as both highly important, and each more so than $X3$. Intuitively, this may reflect the fact that $X1$ and $X2$ serve to validate each other's predictive value, whereas $X3$ can only vouch for itself.

Exhibit 3: Aggregate Variable Importance
$X1$ and $X2$ are 90 Percent Correlated

	X1	X2	X3
R-squared decomposition (model level)	0.0525	0.0495	0.0216
Informativeness-weighted RBI	0.0522	0.0500	0.0214
Linear regression t-statistic	2.84	2.18	5.05
RBI τ-statistic	7.67	7.51	4.91
RBI τ-statistic (simplified weighting)	7.69	7.53	4.95

Simulation of Variable Importance with Uncorrelated Variables and Different Betas

Exhibit 4 shows variable importance again assuming all three X variables are uncorrelated but this time with different assumed betas. $X1$ has a beta of 2 to Y, $X2$ has a beta of 1, and $X3$ has a beta of 0. Once again, the relative value assessment of the variables is sensible and intuitive.

Exhibit 4: Aggregate Variable Importance
Betas Equal 2, 1, and 0; Correlations Equal 0

	X1	X2	X3
R-squared decomposition (model level)	0.1066	0.0397	0.0003
Informativeness-weighted RBI	0.1065	0.0397	0.0005
Linear regression t-statistic	11.08	6.68	0.49
RBI τ-statistic	11.13	6.80	0.74
RBI τ-statistic (simplified weighting)	11.14	6.81	0.82

Simulation of Variable Importance with Correlated Variables and Different Betas

Exhibit 5 preserves the same assumptions of differing betas, but now we assume that all three X variables have correlations of 0.5 with each other. The relative rankings of the variables are consistent across the alternative measures. But again, the t-statistics assign much lower

importance to variables for which much of their useful predictive information is captured by the other variables. It remains the case, however, that even $X3$, which has no direct influence on Y by assumption, contains useful information about Y through its correlation with the other variables. The information given by $X3$ is useful not only in isolation (if $X1$ and $X2$ were omitted) but also as a confirmation mechanism for the information in the other variables.

Exhibit 5: Variable Importance
Betas Equal 2, 1, and 0; Correlations Equal 0.5

	X1	X2	X3
R-squared decomposition (model level)	0.1057	0.0624	0.0216
Informativeness-weighted RBI	0.1060	0.0621	0.0215
Linear regression t-statistic	9.04	5.51	-0.30
RBI τ-statistic	11.40	8.72	5.14
RBI τ-statistic (simplified weighting)	11.19	8.45	4.65

RBI Compared to a Shapley Value

A Shapley value is used to measure variable importance for predictions that are performed by machine learning models. It was conceived from game theory to measure the contribution of participants in collaborative games.[10] Like RBI, the Shapley value measures variable importance for individual predictions, it measures total importance, and it is robust to conditionality. However, a Shapley value is significantly limited compared to RBI because it does not account for an individual prediction's reliability. It is computed only from the observations' predictions without consideration of how they compare to the outcomes for Y. Both RBI and a t-statistic account for reliability by considering both the predictions and the outcomes of the observations.

The formula for a Shapley value is given by Equation 25.

$$Shapley_k = \frac{1}{K}\sum_{S\subseteq K\backslash\{k\}}\binom{K-1}{|S|}^{-1}\left(v(S\cup\{k\})-v(S)\right) \quad (25)$$

If we consider Equation 25 within the context of variable importance in prediction rather than a cooperative game, K is the full set of variables, S is a subset of variables, k is a specific variable, v is a function that maps a subset of variables onto a scalar value, $S \subseteq K\backslash\{k\}$ represents iteration over all of the possible sets of variables composed of any number of the K total variables excluding the variable of interest, $S \cup \{k\}$ is the set of variables that combines S and the variable of interest, and $\binom{K-1}{|S|}$ represents how many combinations of $K-1$ variables have the same size as S.

When RBI is averaged across all predictions, it converges to a Shapley value where the chosen value function is the R-squared of a linear regression model. When RBI is applied to individual predictions it represents the contribution of each variable to the prediction's reliability (adjusted fit), which differs from Shapley values applied to predictions from machine learning models wherein the Shapley value is applied to the prediction value ($\hat{y}_t$) without regard to its reliability.

We thus have the following comparisons. A t-statistic considers average reliability but not total importance or conditionality. A Shapley value measures total importance and is robust to conditionality for individual predictions, but it does not consider reliability except when applied to aggregate model outputs. Only RBI measures total importance, is robust to

conditionality, and considers reliability both for individual predictions and on average across all predictions.

Empirical Example

We now compare variable importance as measured by RBI and a t-statistic for an empirical application to predicting market volatility. This builds upon the simulation-based results by allowing for conditional relationships between predictive variables and outcomes. The outcome we aim to predict is the subsequent one quarter (63-day) volatility of daily total returns of the S&P 500 index. We use 14 predictive variables as described in Exhibit 6, which are observed at the time of each prediction.

Exhibit 6: Predictive Variables

Predictive Variable	Proxy	Source
Market Conditions		
Trailing 1-month volatility	Trailing 21-day volatility of daily S&P 500 returns	Bloomberg
Trailing 3-month volatility	Trailing 63-day volatility of daily S&P 500 returns	Bloomberg
Implied volatility	CBOE VIX (with proxy based on options prices before 1990)	CBOE
Trailing 1-month market return	Trailing 21-day return of the S&P 500	Bloomberg
Trailing 3-month market return	Trailing 63-day return of the S&P 500	Bloomberg
Financial Conditions		
Short-term interest rate (level)	Fed funds rate: 12-m average	FRED
Short-term interest rate (change)	1-year change in short term interest rate	FRED
Long-term interest rate (change)	1-year change in 10y constant maturity rate	FRED
Credit spread (change)	1-year change in Baa corp. bond yield - 10y const. maturity rate	FRED
Economic Conditions		
Growth	1-year % change in industrial production	FRED
Payrolls	1-year % change in non-farm payrolls	FRED
Inflation	1-year % change in Consumer Price Index (CPI)	FRED
Money supply	3-year % change in M2 money supply	FRED
Debt-to-GDP	3-year change in public debt / GDP	FRED

Our historical sample consists of non-overlapping quarterly observations from Q1 1986 through Q4 2023. We use the first half of our sample (Q1 1986 through Q4 2004) as training

data for relevance-based prediction and linear regression, and we reserve the second half (Q1 2005 through Q4 2023) for out-of-sample testing.[11]

For relevance-based prediction, we consider a grid consisting of every possible variable combination (grid columns) and observation censoring percentile thresholds of 0, 0.2, 0.5 and 0.8 (grid rows). We consider censoring based on relevance as well as censoring based on similarity, which essentially multiplies the number of grid cells by two. Note that the cells that use censoring thresholds of zero are equivalent to linear regression predictions for a given set of predictive variables. There are more than 16,000 grid cells; however, we use a sparse sampling method whereby for each prediction we consider the full sample linear cell, each of the 14 single variable linear cells, and 100 randomly selected cells from the rest of the grid, for a total of 115 cells for each prediction task.

In-Sample Results

We first apply relevance-based prediction and linear regression to the training sample (Q1 1986 through Q4 2004) to measure the aggregate importance of the 14 predictive variables. Exhibit 7 shows the variables' linear t-statistics (reported as absolute values to facilitate comparison), RBI τ-statistics, as well as components of their respective calculations. As shown in Exhibit 7, the relative rankings of the variables are similar between the linear and relevance-based measures, though there are notable divergences. For example, trailing three-month volatility is one of the most important variables according to RBI, but it ranks among the least important variables based on its t-statistic. The opposite is true for trailing one-month market returns. In

part, these divergences arise from non-zero correlations between the variables (see Appendix C for their historical correlations), which we also observed in the simulation-based results. Additionally, in this empirical application we allow for conditional relationships between predictive variables and outcomes, which RBI captures, and t-statistics do not.

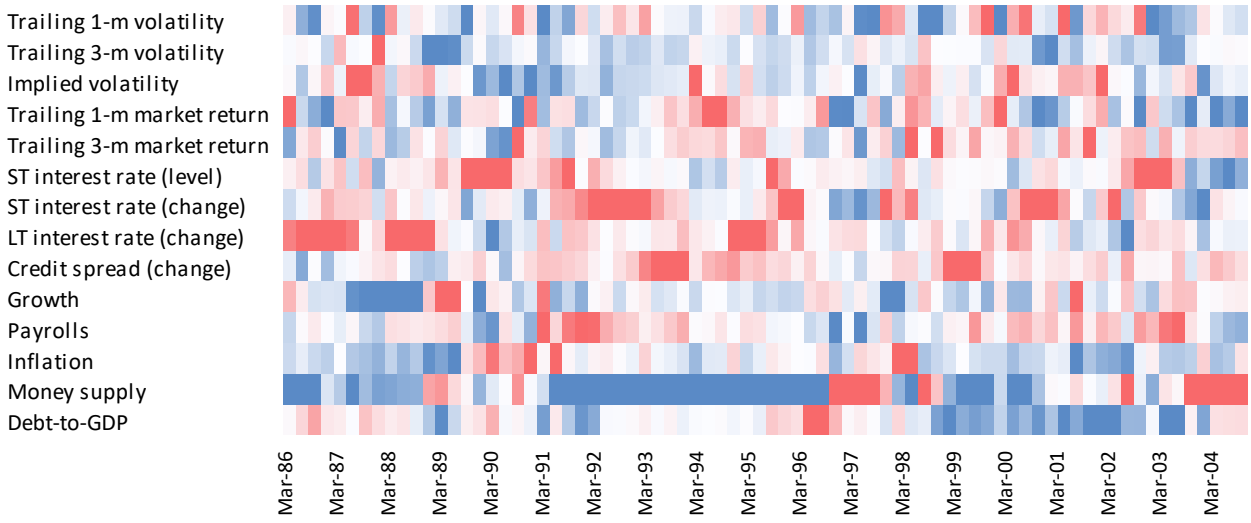
Exhibit 7: Aggregate Variable Importance
In-Sample Q1 1986-Q4 2004

	Linear Regression			Relevance-Based Prediction	
	Partial R-squared	Absolute t-statistic	R-squared Decomposition	Info-Weighted RBI	RBI τ-statistic
Trailing 1-m volatility	0.0292	1.87	0.11	0.13	4.02
Trailing 3-m volatility	0.0002	0.16	0.09	0.11	3.71
Implied volatility	0.0041	0.70	0.08	0.08	3.20
Trailing 1-m market return	0.0305	1.91	0.05	0.09	3.24
Trailing 3-m market return	0.0096	1.07	0.04	0.09	3.28
ST interest rate (level)	0.0262	1.77	0.07	0.09	3.30
ST interest rate (change)	0.0272	1.80	0.07	0.08	3.09
LT interest rate (change)	0.0000	0.07	0.06	0.06	2.81
Credit spread (change)	0.0014	0.41	0.02	0.06	2.75
Growth	0.0380	2.13	0.09	0.15	4.25
Payrolls	0.0003	0.17	0.07	0.08	3.15
Inflation	0.0090	1.04	0.08	0.15	4.23
Money supply	0.0645	2.78	0.15	0.18	4.70
Debt-to-GDP	0.0298	1.89	0.07	0.13	4.01

Thus far we have focused on the aggregate importance of variables across many predictions. However, a key advantage of RBI is that it is prediction specific. Exhibit 8 shows a heat map (red indicates lower values, blue indicates higher values) of prediction-level RBIs underlying the summary τ-statistics in the previous table. Because RBI explicitly considers the specific circumstances of each prediction task, it can vary meaningfully across predictions even when they are formed from the same training data. For example, the previous table shows that in aggregate, money supply was the most important predictive variable based on its τ-statistic;

however, we observe in Exhibit 8 that there were several points in time, such as 1997 and 2004, where it was the least important variable. Conversely, in aggregate, the credit spread was the least important variable; however, in the mid- and late-1980s, it was one of the most important variables.

Exhibit 8: Variable Importance (RBI) by Prediction
In-Sample Q1 1986-Q4 2004



Out-of-Sample Results

We next compare linear regression analysis and relevance-based prediction in terms of their out-of-sample efficacy. Specifically, using the same training data as in the previous section (Q1 1986 through Q4 2004), we predict subsequent volatility outcomes for every quarter in the out-of-sample period Q1 2005 through Q4 2023.

Exhibit 9 shows correlations between predictions and actual outcomes, as well as actual volatility for low and high predictions. We observe that the relevance-based predictions (RBP in the table) are 52% correlated with actual outcomes, compared to only 25% for linear regression

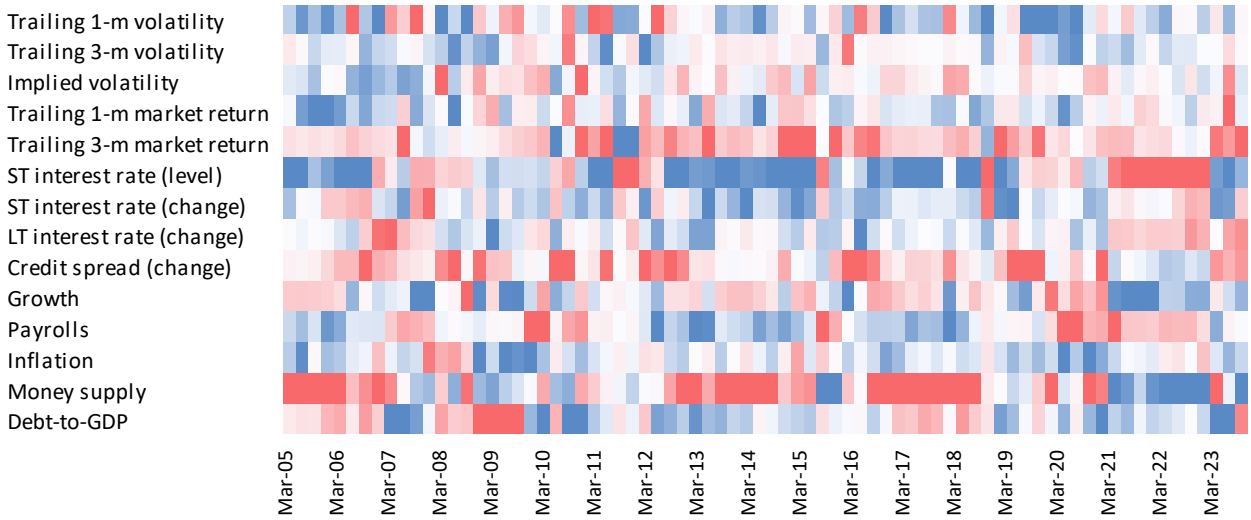
analysis. Moreover, when we focus on a subset of the most reliable relevance-based predictions as indicated by fit, we see an even stronger correlation with actual outcomes (60%). Similarly, the spread in actual volatility for low and high predictions is greater for relevance-based prediction (1.63) than linear regression (1.04), and even more so for a subset of the highest fit predictions (2.31).[12]

Exhibit 9: Prediction Performance
Out-of-Sample Q1 2005-Q4 2023

	Correlation with Actual	Average Out-of-Sample Outcomes		
		Low Predictions	High Predictions	High / Low
Linear regression	0.25	1.02%	1.06%	1.04
RBP	0.52	0.79%	1.29%	1.63
RBP - High fit	0.60	0.71%	1.63%	2.31

To further underscore the prediction-specific nature of RBI, Exhibit 10 shows a heat map of RBIs for every prediction in the out-of-sample period. Consistent with the in-sample results, we observe considerable variation in RBI across prediction tasks. Moreover, in comparing Exhibit 10 with Exhibit 8, both of which use the same training sample (Q1 1986-Q4 2004) and differ only in terms of their prediction tasks, we observe general shifts in variable importance. For example, for the in-sample predictions (Exhibit 8), money supply was the most important variable in aggregate; however, for the out-of-sample predictions (Exhibit 10), it is often the least important variable. Conversely, the level of short term interest rates appears to be more important in the out-of-sample period.

Exhibit 10: Variable Importance (RBI) by Prediction
Out-of-Sample Q1 2005-Q4 2023



Summary

We considered three measures of variable importance: RBI, a t-statistic, and a Shapley value. RBI measures importance as a variable's contribution to the reliability of a prediction for individual predictions and on average across all predictions. It follows naturally from the prediction grid, which is a key feature of relevance-based prediction. A t-statistic measures importance as the amount of variation in the variable that is being predicted that can explained by a predictive variable. It is used to measure variable importance in linear regression analysis. A Shapley value measures a variable's contribution to explaining variation across predictions. It is used to measure variable importance in machine learning models.

We described RBI in detail and listed its useful properties:

- RBI measures variable importance for individual predictions as well as on average across all predictions.
- RBI measures total variable importance.
- RBI is robust to conditionalities.
- RBI accounts for a prediction's reliability.

We then compared RBI to a t-statistic and showed that a t-statistic has only one of the useful properties of RBI. It accounts for reliability. We also showed that average RBI converges almost exactly to a t-statistic in the special case in which the predictive variables are uncorrelated with one another and the relationship between the predictive variables and the outcomes is constant across all observations. This convergence reveals a t-statistic to be a special case of RBI. We conducted several simulations to support our claims about RBI and t-statistics.

Next, we compared RBI to a Shapley value. We showed that a Shapley value has three of the four useful properties of RBI. A Shapley value gives a measure of variable importance for individual predictions, it measures total importance, and it is robust to conditionality. However, a Shapley value fails to account for reliability when applied to individual predictions. Setting aside this distinction, though important, we showed that average RBI across prediction tasks closely matches a Shapley value of model-level reliability and derives from the same calculation principles.

We then provided an empirical analysis of RBI. We used 14 variables to predict stock market volatility using both linear regression analysis and relevance-based prediction. Given

the same observations for training and testing both approaches, we showed that variable importance as measured by RBI changed significantly across prediction tasks both in sample and out of sample and that relevance-based predictions produced significantly more reliable predictions than linear regression analysis, especially for those predictions known in advance to be more reliable.

Given this analysis, we propose RBI as a comprehensive alternative to a t-statistic and a Shapley value for measuring variable importance. Of course, it is important to note that if one chooses to adopt RBI to measure variable importance, one must also choose relevance-based prediction to form predictions. But this choice should not be difficult for many prediction circumstances, when one considers the many virtues of relevance-based prediction compared to linear regression analysis and machine learning models.[13]

Appendix A: RBI and Grid Weights

Assuming we include every combination of variables in the grid of possibilities, including the null set of variables which always has adjusted fit of zero, there are $G/2 = \sum_{\tilde{\theta}} \Delta_k(\tilde{\theta})$ cells that include variable k, where G is the total number of cells in the grid. Therefore, we have the following expression.

$$RBI_{tk} = \sum_{\theta} \alpha_{\theta} \frac{\Delta_k(\theta)(adjusted\ fit_{t\theta}) - (1-\Delta_k(\theta))(adjusted\ fit_{t\theta})}{G/2} \quad \text{(A1)}$$

We can express RBI concisely as a weighted average of the adjusted fit of each cell.

$$RBI_{tk} = \frac{1}{G}\sum_{\theta} \alpha_{\theta}(4\Delta_k(\theta) - 2)(adjusted\ fit_{t\theta}) \quad \text{(A2)}$$

Further, we can express variable importance equivalently in terms of ψ_{θ}, the normalized adjusted fit weights for the grid from Equation 18 which are used to generate grid observation weights and grid predictions, and the average adjusted fit for the entire grid.

$$RBI_{tk} = \left(\frac{1}{G}\sum_{\tilde{\theta}} adjusted\ fit_{t\tilde{\theta}}\right)\sum_{\theta} \psi_{\theta}[\alpha_{\theta}(4\Delta_k(\theta) - 2)] \quad \text{(A3)}$$

Equation A3 highlights two interesting points. First, it shows that RBI is fundamentally measured in units of adjusted fit and that it reflects both the overall reliability of the prediction as determined by the entire grid and the relative contribution of a predictive variable across the grid cells. Second, it shows that all the information about relative variable importance for a given prediction is captured by a weighted average of the grid cell attribute $[\alpha_{\theta}(4\Delta_k(\theta) - 2)]$ where the weights ψ_{θ} are the same as those used to compute the grid prediction and the grid

observation weights. This fact underscores the essential nature of RBI within the paradigm of relevance-based prediction.

Appendix B: Squared t-statistic and R-squared

Without loss of generality, assume that the X and Y variables are centered to have zero averages, $1_N' X = 0$ and $1_N' Y = 0$, as additive shifts have no impact on t-statistics or R-squared for a linear regression model $Y = X\beta + \epsilon$ where β is a column vector of coefficients. The Ordinary Least Squares (OLS) estimate for the beta coefficients is $\hat{\beta} = (X'X)^{-1}X'Y$. We express the kth coefficient as $\hat{\beta}_k = d_k'(X'X)^{-1}X'Y$ where d_k is a column vector with 1 in the kth component and 0s elsewhere.

Let us start by establishing some expressions for the total sum of squares (TSS), the explained sum of squares (ESS), and the residual sum of squares (RSS).

$$TSS = Y'Y \tag{A4}$$

$$ESS = \hat{\beta}'X'X\hat{\beta} \tag{A5}$$

$$ESS = Y'X(X'X)^{-1}X'X(X'X)^{-1}X'Y \tag{A6}$$

$$ESS = Y'X(X'X)^{-1}X'Y \tag{A7}$$

$$RSS = TSS - ESS \tag{A8}$$

$$RSS = Y'Y - Y'X(X'X)^{-1}X'Y \tag{A9}$$

The t-statistic for a coefficient is defined as the coefficient divided by the standard error of the coefficient:

$$t_k = \frac{\widehat{\beta}_k}{SE(\widehat{\beta}_k)} \tag{A10}$$

We are interested in the squared t-statistic which maps directly to its magnitude but does not preserve its sign:

$$t_k^2 = \frac{\widehat{\beta}_k^2}{Var(\widehat{\beta}_k)} \tag{A11}$$

The covariance matrix of coefficient estimates is known to be expressed in terms of the unbiased estimate of the variance of the residuals $\hat{\sigma}^2$:

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1} \tag{A12}$$

$$Var(\hat{\beta}) = \frac{RSS}{N-K} (X'X)^{-1} \tag{A13}$$

$$Var(\hat{\beta}) = \frac{Y'Y - Y'X(X'X)^{-1}X'Y}{N-K} (X'X)^{-1} \tag{A14}$$

The variance of coefficient k is the corresponding diagonal element of the covariance matrix:

$$Var(\hat{\beta}_k) = \frac{Y'Y - Y'X(X'X)^{-1}X'Y}{N-K} (X'X)^{-1}_{kk} \tag{A15}$$

Plugging the necessary definitions into the formula for the squared t-statistic gives:

$$t_k^2 = \frac{Y'X(X'X)^{-1} d_k d_k' (X'X)^{-1} X'Y (N-K)}{(Y'Y - Y'X(X'X)^{-1}X'Y)(X'X)^{-1}_{kk}} \tag{A16}$$

We would like to relate this expression to one involving R-squared statistics. The first R-squared statistic we consider is that of the full regression using all variables.

$$R^2 = \frac{ESS}{TSS} = \frac{Y'X(X'X)^{-1}X'Y}{Y'Y} \tag{A17}$$

Next, we consider the R-squared of a reduced model that removes variable k. The Sherman-Morrison formula holds that the inverse of the reduced matrix without variable k can be computed from the original full inverse with an adjustment. Note here that the new inverse is still expressed as a K-by-K matrix but the kth row and column are neutralized to zero.

$$\left(X'_{\backslash k}X_{\backslash k}\right)^{-1} = (X'X)^{-1} - \frac{(X'X)^{-1}d_k d'_k (X'X)^{-1}}{(X'X)^{-1}_{kk}} \tag{A18}$$

The R-squared of the reduced variable model is:

$$R^2_{\backslash k} = \frac{Y'X\left((X'X)^{-1} - \frac{(X'X)^{-1}d_k d'_k (X'X)^{-1}}{(X'X)^{-1}_{kk}}\right)X'Y}{Y'Y} \tag{A19}$$

We take the difference between the full R-squared and the reduced variable model R-squared.

$$R^2 - R^2_{\backslash k} = \frac{Y'X(X'X)^{-1}d_k d'_k (X'X)^{-1}X'Y}{Y'Y(X'X)^{-1}_{kk}} \tag{A20}$$

Thus, we have:

$$t_k^2 = \frac{\left(R^2 - R^2_{\backslash k}\right)(N-K)}{(1-R^2)} \tag{A21}$$

Appendix C: Correlation of Predictive Variables

Exhibit A1 shows the empirical correlations across the 14 predictive variables used to predict market volatility.

Exhibit A1: Variable Correlations
Q1 1986-Q4 2004

		A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	Trailing 1-m volatility	1.00	0.78	0.85	-0.18	-0.50	0.00	-0.12	-0.05	0.34	-0.04	-0.13	-0.17	0.50	-0.47
B	Trailing 3-m volatility	0.78	1.00	0.84	-0.03	-0.50	0.07	-0.12	0.10	0.17	0.06	-0.08	0.01	0.50	-0.28
C	Implied volatility	0.85	0.84	1.00	-0.15	-0.44	0.10	-0.17	-0.03	0.31	0.02	-0.09	-0.06	0.54	-0.35
D	Trailing 1-m market return	-0.18	-0.03	-0.15	1.00	0.47	0.14	0.18	0.09	-0.13	0.28	0.29	0.02	-0.07	0.11
E	Trailing 3-m market return	-0.50	-0.50	-0.44	0.47	1.00	0.09	0.11	-0.09	-0.02	0.06	0.15	-0.08	-0.09	0.11
F	ST interest rate (level)	0.00	0.07	0.10	0.14	0.09	1.00	0.35	0.00	0.11	-0.06	0.36	0.58	0.04	0.23
G	ST interest rate (change)	-0.12	-0.12	-0.17	0.18	0.11	0.35	1.00	0.51	-0.47	0.36	0.68	0.28	-0.28	-0.10
H	LT interest rate (change)	-0.05	0.10	-0.03	0.09	-0.09	0.00	0.51	1.00	-0.65	0.36	0.25	0.37	-0.26	-0.20
I	Credit spread (change)	0.34	0.17	0.31	-0.13	-0.02	0.11	-0.47	-0.65	1.00	-0.40	-0.32	-0.29	0.26	-0.29
J	Growth	-0.04	0.06	0.02	0.28	0.06	-0.06	0.36	0.36	-0.40	1.00	0.74	-0.15	-0.31	0.13
K	Payrolls	-0.13	-0.08	-0.09	0.29	0.15	0.36	0.68	0.25	-0.32	0.74	1.00	0.09	-0.25	0.23
L	Inflation	-0.17	0.01	-0.06	0.02	-0.08	0.58	0.28	0.37	-0.29	-0.15	0.09	1.00	-0.11	0.33
M	Money supply	0.50	0.50	0.54	-0.07	-0.09	0.04	-0.28	-0.26	0.26	-0.31	-0.25	-0.11	1.00	-0.17
N	Debt-to-GDP	-0.47	-0.28	-0.35	0.11	0.11	0.23	-0.10	-0.20	-0.29	0.13	0.23	0.33	-0.17	1.00

Notes

This material is for informational purposes only. The views expressed in this material are the views of the authors, are provided "as-is" at the time of first publication, are not intended for distribution to any person or entity in any jurisdiction where such distribution or use would be contrary to applicable law and are not an offer or solicitation to buy or sell securities or any product. The views expressed do not necessarily represent the views of Windham Capital Management, State Street Global Markets®, or State Street Corporation® and its affiliates.

References

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022a. "Relevance." *The Journal of Investment Management*, 20 (1).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022b. *Prediction Revisited: The Importance of Observation*. Hoboken, New Jersey: John S. Wiley & Sons.

Czasonis, Megan, Mark Kritzman, and David Turkington. 2023. "Relevance-Based Prediction: A Transparent and Adaptive Alternative to Machine Learning." *The Journal of Financial Data Science*, 5 (1).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2024a. "The Virtue of Transparency: How to Maximize the Utility of Data Without Overfitting." *MIT Working Paper* (July).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2024b. "A Transparent Alternative to Neural Networks with an Application to Predicting Volatility." *MIT Working Paper* (September).

Mahalanobis, Prasanta Chandra. 1936. "On the Generalised Distance in Statistics." *Proceedings of the National Institute of Sciences of India*, 2 (1): 49–55.

Shannon, Claude. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal*, 27 (July, October): 379–423, 623–656.

Shapley, L. S. (1953), "A Value for n-Person Games," in *Contributions to the Theory of Games* (Vol. II), eds. H. W. Kuhn and A. W. Tucker, Princeton, NJ: Princeton University Press, pp. 307–318.

[1] The descriptions of these concepts follow closely from Czasonis, Kritzman, and Turkington (2022a), Czasonis, Kritzman, and Turkington (2022b), Czasonis, Kritzman, and Turkington (2023), and Czasonis, Kritzman, and Turkington (2024a), and Czasonis, Kritzman, and Turkington (2024b), but they are modified to fit the context of the current discussion.

[2] This measure was first introduced by Mahalanobis (1936).

[3] Shannon showed that information is an inverse logarithmic function of probability, which is a key insight from his comprehensive theory of communication. See Shannon (1948).

[4] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[5] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[6] See Czasonis, Kritzman, and Turkington (2022b) for proof of this result.

[7] See Appendix B for further discussion of this expression.

[8] Recall from Equation 12 that for the subset of variables that corresponds to any grid cell in the first row, the R-squared of a linear regression with that subset of variables is precisely equivalent to the informativeness-weighted average of the task-specific fits of the corresponding predictions. The same result holds for adjusted fit in the first row of the grid because asymmetry is zero for linear predictions. These cell-specific equivalences hold exactly when informativeness is computed using the subset of variables for each cell. Our result relating informativeness-weighted average RBI for predictions from the first row of the grid to the R-squared equivalent across the first row of the grid is approximate because task informativeness measured across all predictive variables and applied to the adjusted fit of each cell's variable subset does not recover the same task aggregation as taking a weighted average with the informativeness of each variable subset. Nevertheless, it is astonishing how close of a near equivalence we observe empirically for the informativeness-weighted RBI.

[9] This near equivalence to strictly-positive increases in R-squared for linear regressions with larger numbers of predictive variables also implies that the informativeness-weighted average RBI is rarely negative for a variable and if so only because of approximation error. This conceptual lower bound of zero for the aggregate importance of a variable across the training tasks occurs despite the perfectly reasonable occurrence of negative values for the RBI values of individual prediction tasks, as discussed earlier.

[10] The Shapley value has several desirable properties including efficiency, symmetry, linearity, and null player. See Shapley (1953).

[11] The empirical example follows closely from Czasonis, Kritzman, and Turkington (2024b), however it differs in two key ways: (1) We use non-overlapping quarterly observations of variables and outcomes whereas the cited paper uses overlapping monthly observations (2) We use a constant training sample from 1986 to 2004 to predict out-of-sample outcomes beginning in 2005 whereas the cited paper uses a growing training sample from 1986 to form out-of-sample predictions beginning in 2000.

[12] We define high and low predictions as those in the top and bottom half of all predictions, respectively. We define high fit predictions as those the 50% highest fits.

[13] To appreciate the many virtues of relevance-based prediction, see Czasonis, Kritzman, and Turkington (2023), Czasonis, Kritzman, and Turkington (2024a), and Czasonis, Kritzman, and Turkington (2024b).