Feature Importance for Model Fit: Decomposing Mean Squared Error in Regression

January 18, 2026

Abstract

We use Euler's theorem to decompose predictive accuracy in regression models. The resulting contributions allocate realized predictive performance across additive components of a prediction, providing a natural, additive, and model-conditional measure of feature importance. We also derive standard errors for the contributions, allowing formal inference and facilitating model monitoring over time.

In this framework, a component contributes positively to predictive performance if it moves the prediction closer to the outcome, either by directly explaining the outcome or by correcting errors left by other components.

Under ordinary least squares estimated in sample, the resulting proportional attributions coincide with the Pratt decomposition of explained variance. Outside the estimation sample and beyond ordinary least squares, the Euler decomposition remains well defined for any prediction with an additive structure, yielding a unified notion of feature importance that does not rely on estimator-specific orthogonality conditions.

www.ludgerhentschel.com ludger@ludgerhentschel.com

Contents

Acknowledgements
For helpful comments, I am grateful to Nishant Gurnani, Ingemar Hentschel, and Shubham Jaiswal.

1 Introduction

In regression and machine learning models, we often wish to gauge the importance of individual features, especially when the number of included features is large.[1] This objective has given rise to a wide range of feature-importance measures.

In this paper, we focus on evaluating realized predictive performance for a given, fitted model. We measure model fit as the reduction in mean squared error relative to an intercept-only baseline predictor. This loss-based notion of predictive performance is well defined both in and out of sample and applies uniformly across linear and nonlinear prediction methods. Rather than allocating importance across a space of potential models or counterfactual feature inclusions, we attribute realized improvements in predictive accuracy to additive components of the deployed prediction itself. This perspective naturally supports model monitoring, diagnostics, and performance attribution for an existing model whose structure and fitted values are treated as fixed. The Euler decomposition provides a unified and computationally efficient way to attribute realized improvement in predictive accuracy across prediction components, without reference to the estimation procedure that produced them.

We show that the reduction in mean squared error admits an exact Euler decomposition when expressed in terms of its homogeneous components. The resulting attribution allocates predictive fit across additive prediction components based on their contribution to reducing realized forecast error. The decomposition is model-conditional: it attributes realized predictive performance for the prediction actually used and does not rely on counterfactual refitting, feature removal, or perturbation. This mirrors the marginal contribution framework used in portfolio risk attribution (Litterman, 1996; Tasche, 2008), where total risk is allocated across additive portfolio components, holding fixed the actual portfolio and without reference to the portfolio construction method.

A large literature proposes measures of "relative importance" based on partitioning predicted variation, including standardized coefficient measures following Pratt (1987), heuristic variance partitions proposed by Bring (1995), and Shapley-value or dominance-based decompositions developed by Lindeman, Merenda, and Gold (1980) and Kruskal (1987). These approaches differ in their axiomatic foundations and computational complexity, but they often share a common focus on decomposing predictions or predicted variance

[1] We use the term regression model in the broad sense of a model for predicting a continuous outcome, rather than to refer to a specific estimation method.

across features. Many of these measures prove useful when exploring alternative model specifications, comparing competing models, or interpreting individual predictions. Our approach differs in that it attributes realized predictive performance for a fixed, deployed model, rather than allocating importance across hypothetical model variations.

In the proposed decomposition, an Euler component contributes positively to predictive performance if it moves the prediction closer to the outcome, either by directly explaining the outcome or by correcting errors left by other components. Unlike variance- or correlation-based measures, the attribution depends on how components interact within the fitted prediction rather than just on their marginal associations with the outcome. The attribution requires only realized predictions and their additive components, without refitting, perturbation, or estimator-specific orthogonality conditions.

In linear regression models, predictions decompose naturally into regressor-specific components given by features multiplied by their fitted coefficients. The Euler decomposition applies equally to ordinary, weighted, or regularized linear models, including generalized least squares, Ridge, Lasso, and Elastic Net regressions. More generally, the same logic applies to any predictive model whose predictions decompose into additive components of interest.

Finally, we derive standard errors for the Euler contributions that reflect sampling variability in the data. This allows us to assess whether observed variation in feature contributions across samples or over time is plausibly attributable to noise or instead reflects changes in predictive relevance.

The remainder of the paper proceeds as follows. Section 2 introduces the loss-based framework and derives the Euler decomposition of predictive accuracy, together with standard errors. Section 3 examines the linear regression case and its connection to in-sample OLS and the Pratt decomposition. Section 4 presents Monte Carlo illustrations. Section 5 relates the approach to existing feature-importance measures. Section 6 discusses extensions, including grouped attributions and broader applicability. Section 7 concludes.

2 An Euler Decomposition of Explained Fit

After establishing notation and defining predictive accuracy, we can apply Euler's theorem to obtain an exact additive decomposition of model fit across prediction components and derive the corresponding standard errors.

2.1 Setup

Let $\tilde{y} \in \mathbb{R}^N$ denote a vector of observed outcomes with finite, nonzero variance and define the centered outcome

$$y = \tilde{y} - \mathbb{E}[\tilde{y}]. \tag{1}$$

We interpret the intercept-only predictor $\widehat{y}_0 = \mathbb{E}[\tilde{y}]$ as the baseline model and evaluate predictive performance relative to this baseline.

Throughout, expectations, variances, and covariances are sample averages. Since y is centered, $\mathbb{E}[y] = 0$ and $\mathrm{Var}(y) = \mathbb{E}[y^2]$.

Let $\widehat{y} \in \mathbb{R}^N$ denote the predictions of a regression or forecasting model, and assume that $\widehat{y}$ is centered so that $\mathbb{E}[\widehat{y}] = 0$.[2]

We measure predictive accuracy using mean squared error and define the improvement in fit relative to the baseline predictor as[3]

$$\Delta\mathcal{L} = \mathrm{Var}(y) - \mathrm{Var}(y - \widehat{y}). \tag{2}$$

With perfect predictions $\widehat{y} = y$, predictive accuracy attains a maximum value of $\mathrm{Var}(y)$. With poor predictions, predictive accuracy can be close to 0 or even negative.

If we scale the predictive accuracy by $\mathrm{Var}(y)$, we obtain the standard regression coefficient of determination

$$R^2 = \frac{\Delta\mathcal{L}}{\mathrm{Var}(y)} = 1 - \frac{\mathrm{Var}(y - \widehat{y})}{\mathrm{Var}(y)}. \tag{3}$$

Although the expression $R^2 = \mathrm{Var}(\widehat{y})/\mathrm{Var}(y)$ is often treated as equivalent, this is only true when predictions $\widehat{y}$ are orthogonal to prediction errors $y - \widehat{y}$. Although this condition holds in-sample for ordinary least squares regressions, this is a special case. Without this orthogonality, $\Delta\mathcal{L}$ remains a measure of predictive accuracy while $\mathrm{Var}(\widehat{y})/\mathrm{Var}(y)$ compares the scale of predictions and outcomes without any notion of alignment between them.

Expanding the squared error yields

$$\Delta\mathcal{L}(\widehat{y}) = 2\ \mathrm{Cov}(y, \widehat{y}) - \mathrm{Var}(\widehat{y}). \tag{4}$$

[2] If the fitted model includes an intercept and is evaluated on centered regressors, this condition holds automatically. In regularized regressions, the intercept is typically excluded from regularization to preserve this property.

[3] Defining explained fit relative to an intercept-only baseline model is partly conceptual and partly a matter of convenience. The intercept is not a feature, but a normalization that centers predictions and defines the baseline level of predictive accuracy. The Euler decomposition allocates only the incremental explained fit arising from additive prediction components beyond the mean. When the contribution of the unconditional mean is of interest, we can track it separately as a baseline component of total fit.

This quantity measures realized predictive accuracy of the model relative to the intercept-only baseline. It is well defined both in and out of sample and does not rely on orthogonality or optimality conditions specific to any estimation procedure.

2.2 Euler Decomposition

The improvement in predictive accuracy is the difference of two homogeneous functions,

$$\Delta\mathcal{L}(\widehat{y}) = g_1(\widehat{y}) - g_2(\widehat{y}), \tag{5}$$

where $g_1(\widehat{y}) = 2\,\mathrm{Cov}(y, \widehat{y})$ and $g_2(\widehat{y}) = \mathrm{Var}(\widehat{y})$. The first term g_1 is homogeneous of degree one in $\widehat{y}$, while the second term g_2 is homogeneous of degree two in $\widehat{y}$.

For a function homogeneous of degree k, Euler's theorem states $g(x) = \frac{1}{k} x^\top \nabla g(x)$.

We can apply Euler's theorem separately to each term. Euler's theorem implies

$$g_1(\widehat{y}) = 2\ \mathrm{Cov}(\widehat{y}, y) = \widehat{y}^\top \frac{\partial g_1}{\partial \widehat{y}} = \frac{2}{N}\,\widehat{y}^\top y, \tag{6}$$

and

$$g_2(\widehat{y}) = \mathrm{Var}(\widehat{y}) = \frac{1}{2}\,\widehat{y}^\top \frac{\partial g_2}{\partial \widehat{y}} = \frac{1}{N}\,\widehat{y}^\top \widehat{y}. \tag{7}$$

Although Euler's theorem resembles a local, gradient-based expansion, it is an exact identity for homogeneous functions and holds globally for all admissible inputs.

In linear regression and quadratic risk models, feature or asset contributions follow immediately from covariance algebra. These decompositions can also be viewed as special cases of Euler decompositions of homogeneous fit or risk measures. We adopt the Euler perspective because it extends unchanged to nonlinear models and alternative loss functions.

In many regression models, the fitted signal admits an additive decomposition

$$\widehat{y} = \sum\nolimits_j \widehat{y}_j, \tag{8}$$

where $\widehat{y}_j$ denotes a component of the prediction associated with feature, regressor, or model component j.

In linear regression models with an intercept and centered regressors, each regressor-specific fitted component is automatically mean zero. In more general additive decompositions, component means are not uniquely determined: constants can be shifted arbitrarily across components without affecting the overall fitted prediction. To obtain a well-defined attribution, we therefore impose the normalization

$$\mathbb{E}[\widehat{y}_j] = 0 \qquad \text{for all } j, \tag{9}$$

which assigns all level effects to the intercept-only baseline. This normalization does not affect the fitted predictions or the model's predictive accuracy, but it is essential for identifying component-level contributions.

This gives an exact additive decomposition of predictive accuracy in terms of $\widehat{y}_j$,

$$\Delta\mathcal{L}(\widehat{y}) = \sum\nolimits_j C_j \tag{10}$$

$$C_j = 2\,\mathrm{Cov}(y, \widehat{y}_j) - \mathrm{Cov}(\widehat{y}, \widehat{y}_j) \tag{11}$$

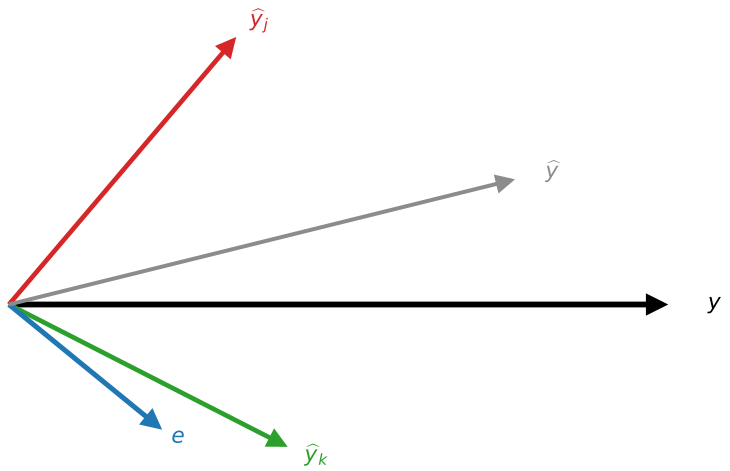
$$= \mathrm{Cov}(y, \widehat{y}_j) + \mathrm{Cov}(e, \widehat{y}_j). \tag{12}$$

The structure of the Euler contribution has a geometric interpretation, illustrated in figure 1. View y, $\widehat{y}$, and the additive components $\widehat{y}_j$ as vectors in $\mathbb{R}^N$ equipped with the inner product $\langle a, b\rangle = \mathrm{Cov}(a, b)$. Because the model has already been fitted, all of these objects in figure 1, including the realized prediction error $e = y - \widehat{y}$, are fixed.

The Euler contribution C_j summarizes how component $\widehat{y}_j$ aligns with these other realized objects. The first term measures how strongly the component aligns with the outcome y. The second term rewards alignment with the realized error. A component improves predictive accuracy to the extent that it aligns with the outcome and, conditional on the full fitted prediction, aligns with the error vector e. Such alignment indicates that the component contributes to reducing the distance between the fitted prediction $\widehat{y}$ and the outcome y, and therefore to lowering squared error. Conversely, a component that points away from the error contributes to a larger distance between $\widehat{y}$ and y and therefore worsens predictive accuracy. Components that primarily reinforce other fitted components without aligning with the error receive smaller or even negative contributions.

Additive prediction components $\widehat{y}_j$ necessarily distribute themselves around the fitted value $\widehat{y}$. This distribution creates dispersion in alignment with e and thereby produces winners and losers in the Euler attribution, even

Figure 1: Geometry of Euler Contributions to Model Fit



The figure illustrates the Euler decomposition of realized regression fit for an outcome vector y (black) and a fitted value $\widehat{y} = \widehat{y}_j + \widehat{y}_k$ (gray). Two additive prediction components, $\widehat{y}_j$ (red) and $\widehat{y}_k$ (green), appear with different alignment relative to the prediction error e (blue).

In the Euler decomposition, the contribution of component $\widehat{y}_\ell$ depends on both its covariance with y and its covariance with $e = y - \widehat{y}$. These covariances determine whether the component contributes new explanatory direction or primarily overlaps with the existing fitted value.

In the diagram, both $\widehat{y}_j$ and $\widehat{y}_k$ align positively with y, but they exhibit different alignment with the realized residual e. Here, $\widehat{y}_j$ makes a smaller Euler contribution to model fit because its negative alignment with e contributes to a larger distance between $\widehat{y}$ and y. Conversely, $\widehat{y}_k$ makes a larger Euler contribution because its positive alignment with e contributes to a smaller distance between $\widehat{y}$ and y.

The Euler decomposition and the figure do not rely on orthogonality assumptions; angles represent empirical covariances.

when components exhibit similar alignment with $\widehat{y}$ itself.

A component receives a negative contribution when it primarily offsets other components and thereby reduces predictive accuracy. Such negative contributions arise naturally from the variance identity $\text{Var}(a + b) = \text{Var}(a) + \text{Var}(b) + 2\,\text{Cov}(a, b)$ and reflect the fact that predictive accuracy depends on how components interact, not on their magnitudes in isolation.

When monitoring model performance across samples, a persistently negative Euler contribution need not be problematic; it may simply reflect stable redundancy or offsetting interactions among prediction components. By contrast, a contribution that changes sign indicates a change in how the component interacts with other parts of the fitted prediction. This suggests a shift in the structure of the predictive signal, especially when the change is statistically significant.

The Euler decomposition depends only on realized fitted values and their additive components. It does not differentiate through the estimation procedure that produced $\widehat{y}$ and does not require refitting, counterfactual

Algorithm 1: Euler Feature Importance

```
# Inputs:
# y      : (N,) vector of realized outcomes,
#            centered internally by the algorithm
# Y_hat  : (N, K) matrix of fitted signal components
#            with y_hat = sum_j Y_hat[:, j]
#
# Baseline:
# The intercept-only baseline is E[tilde y].
# All computations are performed relative to this baseline.
#
# Notes:
# For linear models with an intercept and centered regressors,
# Y_hat[:, j] = X[:, j] * beta[j] is already mean zero.
# Component centering ensures a unique attribution by assigning
# all level effects to the intercept-only baseline.
# Components are centered internally by the algorithm.
#
# For WLS or GLS, supply y and Y_hat already transformed
# by the appropriate weighting or whitening matrix.

# Center outcome (defines intercept-only baseline)
y_c = y - mean(y)

# Aggregate fitted signal and center
y_hat = Y_hat.sum(axis=1)
y_hat_c = y_hat - mean(y_hat)

# Center fitted components (normalization for attribution)
Yc = Y_hat - mean(Y_hat, axis=0)

# Euler contributions to improvement in MSE
for j in range(K):       # Can be vectorized
    C[j] = ( 2 * mean(y_c * Yc[:, j]) - mean(y_hat_c * Yc[:, j]) )

DeltaL = sum(C)  # Reduction in MSE relative to baseline

# Plain (i.i.d.) standard errors for contributions
N = len(y)
a = 2 * y_c - y_hat_c    # Observation-level term shared across
    features
for j in range(K):       # Can be vectorized
    c_ij = a * Yc[:, j]  # Observation-level contributions
    SE[j] = sqrt( mean((c_ij - C[j])**2) / N )

# Outputs:
# C          : Contributions to model fit
# SE         : Standard errors for C
# DeltaL     : Reduction in MSE relative to intercept-only baseline
# C / DeltaL : Proportional contributions
#                May be unstable if abs(DeltaL) is near 0
```

feature removal, or orthogonality between fitted values and residuals.

Algorithm 1 summarizes the computation for a general predictive model. Compared to many competing approaches, these computations are cheap, so they accommodate a large number of attribution components and frequent evaluation.

2.3 Standard Errors

The Euler decomposition expresses feature importance as an estimate derived from the evaluation data. Each contribution C_j is a sample average of observation-level contributions c_{ij} and therefore inherits sampling variability, even when the fitted model is treated as fixed. In this sense, feature importance is not merely a descriptive label attached to a model, but an estimated quantity whose precision depends on the amount and structure of the evaluation data.

Appendix B shows that standard errors for the Euler contributions $C_j = (1/N)\ \sum_i c_{ij}$ are

$$SE(C_j) = \sqrt{\frac{1}{N}\,\mathbb{E}\big[(c_{ij} - C_j)^2\big]}. \tag{13}$$

These standard errors quantify uncertainty due to sampling variability in the data used to evaluate model fit. They do not reflect uncertainty arising from re-estimation of the model, which we condition on throughout. As a result, they apply both in-sample and out-of-sample and can be used to assess whether observed variation in feature contributions across samples or over time plausibly reflects noise or instead indicates changes in predictive relevance.

3 Linear Regressions

Since linear regressions are obvious models with additive prediction components, they are an interesting special case. However, it is important to emphasize that the Euler decomposition does not depend on how the fitted signal $\widehat{y}$ and its components $\widehat{y}_j$ are obtained.

3.1 Features as Prediction Components

For linear models of the form $\widehat{y} = X\widehat{\beta}$ with centered regressors X, the fitted signal decomposes naturally into regressor-specific components

$$\widehat{y}_j = X_j\widehat{\beta}_j. \tag{14}$$

Substituting into the Euler contribution yields

$$C_j = 2,\widehat{\beta}_j\,\mathrm{Cov}(y, X_j) - \widehat{\beta}_j\,\mathrm{Cov}(\widehat{y}, X_j) \tag{15}$$

$$= \widehat{\beta}_j\,\mathrm{Cov}(y, X_j) + \widehat{\beta}_j\,\mathrm{Cov}(e, X_j), \qquad e = y - \widehat{y}. \tag{16}$$

The first term reflects the regressor's marginal association with the outcome, while the second captures how the regressor-specific fitted component aligns with the realized prediction error after accounting for the full model.[4]

This decomposition makes clear that marginal association alone does not determine predictive contribution. A regressor may have a large marginal correlation with y yet contribute little to predictive accuracy if its fitted component primarily reinforces other components without reducing residual error. Conversely, a regressor with modest marginal explanatory power may materially improve predictive accuracy by correcting systematic prediction errors.

3.2 In-sample OLS and the Connection to R-Squared

Ordinary least squares constitutes a special case in which predictive accuracy, explained variance, and correlation-based measures coincide in sample. Under OLS with centered variables, fitted values $\widehat{y}$ are orthogonal to residuals $e = y - \widehat{y}$, implying

$$\text{Cov}(e, X_j) = 0 \qquad \text{for all } j, \tag{17}$$

and therefore

$$C_j = \widehat{\beta}_j \, \text{Cov}(y, X_j). \tag{18}$$

Summing across features gives

$$\Delta\mathcal{L} = \sum\nolimits_j \widehat{\beta}_j \, \text{Cov}(y, X_j) = \text{Cov}(y, \widehat{y}) = \text{Var}(\widehat{y}), \tag{19}$$

where the final equality again follows from OLS orthogonality. Normalizing by $\text{Var}(y)$ yields

$$R^2 = \frac{\Delta\mathcal{L}}{\text{Var}(y)} = \frac{\text{Var}(\widehat{y})}{\text{Var}(y)}. \tag{20}$$

In this knife-edge setting, decomposing predictive accuracy, predicted variance, and R^2 are equivalent up to scale.

3.3 Pratt decomposition

Pratt (1987) proposes a decomposition of explained variance for linear regression based on marginal correlations. For standardized regressors estimated

[4] Because the fitted signal is linear in the coefficients, Euler contributions can equivalently be computed by differentiating with respect to $\widehat{\beta}_j$ rather than $\widehat{y}_j$, treating X as fixed.

by ordinary least squares,

$$R^2 = \sum_j \widehat{\beta}_j \operatorname{Corr}(y, X_j), \tag{21}$$

and the terms

$$P_j = \widehat{\beta}_j \operatorname{Corr}(y, X_j) \tag{22}$$

are interpreted as measures of variable importance.

The Pratt decomposition therefore allocates *explained variance* in the estimation sample and implicitly relies on the absence of correlation between fitted components and residuals.[5] By contrast, the Euler decomposition targets realized *predictive accuracy* of a fixed fitted model, measured as the reduction in mean squared error relative to a baseline predictor.

For in-sample OLS with standardized regressors, proportional Pratt and Euler attributions coincide exactly. This equivalence relies entirely on the orthogonality conditions imposed by ordinary least squares. Outside this setting, including out-of-sample evaluation, weighted or generalized least squares, and regularized linear models, residuals generally correlate with fitted components, so predicted variance no longer coincides with predictive accuracy. In these cases, the ratio $\operatorname{Var}(\widehat{y})/\operatorname{Var}(y)$ measures the scale of predictions but contains no information about their alignment with the outcome.

The Euler decomposition continues to apply without modification in these cases, providing an exact and additive attribution of realized predictive accuracy when variance-based decompositions break down.

4 Monte Carlo Illustration

Table 1 provides a numerical illustration of the Euler decomposition. The simulations use samples of 500 observations for fitting (when applicable) and 500 observations for evaluation, with results aggregated over 100,000 Monte Carlo replications. The true data-generating process is a linear regression with five features and coefficients $\{1.0, 0.6, 0.0, -0.4, 0.2\}$. All features are normally distributed with mean zero and unit variance. Features i and j have pairwise correlation $\rho_{ij} = \rho^{|i-j|}$. In Panels A and B, $\rho = 0.7$. The table computes sample statistics like $\Pr(\cdot)$, $\mathbb{E}[\cdot]$ and $\operatorname{Med}(\cdot)$ across Monte Carlo replications.

[5] Thomas, Hughes, and Zumbo (1998) provide a geometric interpretation of the Pratt decomposition.

Panel A verifies basic accounting identities and numerical stability. The simulations achieve the intended average out-of-sample R^2 values of 0.60, 0.30, 0.10, and 0.02. By construction, the Euler contributions sum exactly to the total mean-square improvement over the constant-only baseline on each evaluation sample, and this identity therefore also holds on average across simulations. Values reported as 0 are within machine precision. The final column shows that even correctly specified models can fail to improve upon the mean-only baseline in finite samples, particularly when true R^2 is low. As expected, this probability decreases with sample size, although the table does not explore that dimension.

Panel B compares Euler and Pratt attributions. For these direct comparisons, we use the covariance-form for the Pratt components, $P_j = \mathrm{Cov}(y, \widehat{y}_j)$, which are equal to Pratt's correlation-based formulation up to variable scale. The first column confirms the analytical result that, for ordinary least squares, Pratt and Euler attributions coincide exactly in the training sample. The second column shows that this equivalence breaks down out of sample: proportional Pratt and Euler shares can differ materially, with discrepancies becoming large when R^2 is low. The third column reports the same comparison for Elastic Net regressions. While the qualitative pattern mirrors the OLS case, regularization dampens estimation error and reduces the divergence between proportional Pratt and Euler attributions.

Panel C examines the role of feature correlation by varying ρ. The first column shows that negative Euler contributions occur in finite samples but are rare unless features are strongly correlated. The second column indicates that most negative contributions are small in magnitude: the total share of negative contributions remains modest except at very high correlations. The final column confirms that the standard errors derived in appendix B deliver accurate coverage in this setting.[6]

For brevity, the table suppresses empirical Monte Carlo standard errors. For nearly all entries, these are below 0.001. The main exceptions occur for out-of-sample proportional differences between Euler and Pratt attributions at very low R^2. In these cases, the denominators, $\Delta\mathcal{L}$ or $\sum_j P_j$, can occasionally

[6] In the simulation setting, we can derive population values for the Euler contribution components and evaluate standard error coverage relative to these values. The simulations draw features $X \sim \mathcal{N}(0, \Sigma_X)$ and outcomes $y = X\beta + \varepsilon$, with ε orthogonal to X in the population. The fitted prediction function is $\widehat{y} = X\widehat{\beta}$. In this case, the population Euler contribution for feature j is

$$C_j = 2\,\mathrm{Cov}(y, \widehat{y}_j) - \mathrm{Cov}(\widehat{y}, \widehat{y}_j) = \widehat{\beta}_j \left[2(\Sigma_X \beta)_j - (\Sigma_X \widehat{\beta})_j \right].$$

Table 1: Monte Carlo Simulations

Panel A. Simulation characteristics

R^2	$\mathbb{E}[R^2_{oos}]$	$\mathbb{E}[\lvert\sum_j C_j - \Delta\mathcal{L}\rvert]$	$\Pr(\Delta\mathcal{L} < 0)$
0.60	0.599	0	0
0.30	0.299	0	0
0.10	0.099	0	0.000
0.02	0.020	0	0.056

Panel B. Comparison to Pratt

	In Sample OLS	Out of Sample OLS	Out of Sample Elastic Net
R^2	$\mathbb{E}[\max_j \lvert C_j - P_j\rvert]$	Med $\left(\max_j \lvert C_j/\Delta\mathcal{L} - P_j/\sum_j P_j\rvert\right)$	
0.60	0	0.015	0.006
0.30	0	0.029	0.010
0.10	0	0.075	0.022
0.02	0	0.591	0.124

Panel C. Variability of Contributions (Target $R^2 = 0.30$)

		Negative mass	Standard Errors
ρ	$\Pr(C_j < 0)$	$\mathbb{E}[\sum_{C_j<0} \lvert C_j\rvert/\Delta\mathcal{L}]$	95% coverage
0.00	0.029	0.002	0.949
0.30	0.078	0.006	0.949
0.60	0.183	0.038	0.949
0.90	0.200	0.203	0.948

The table reports Monte Carlo simulations with sample size 500 in the separate training and evaluation samples across 100,000 replications. The data-generating process is a linear model with $K = 5$ features and coefficients $\{1.0, 0.6, 0.0, -0.4, 0.2\}$. Regressors are jointly normal with mean zero, unit variance, and pairwise correlation $\rho_{ij} = \rho^{|i-j|}$. In Panels A and B, $\rho = 0.7$; panel C varies ρ. We set noise variance to target the listed population R^2 values. Entries reported as 0 are numerically zero within machine precision.

Euler contributions are $C_j = 2\,\mathrm{Cov}(y, \widehat{y}_j) - \mathrm{Cov}(\widehat{y}, \widehat{y}_j)$ and sum to $\Delta\mathcal{L} = \mathrm{Var}(y) - MSE(y - \widehat{y})$ on each evaluation sample, up to floating-point error. Pratt contributions are $P_j = \mathrm{Cov}(y, \widehat{y}_j)$. (We use the covariance-form for Pratt components, which coincide with Pratt's correlation-based formulation for standardized variables.) The corresponding proportional attributions are $C_j/\Delta\mathcal{L}$ and $P_j/\sum_j P_j$.

Panel A reports average out-of-sample R^2, the absolute add-up error $|\sum_j C_j - \Delta\mathcal{L}|$, and the frequency with which $\Delta\mathcal{L} < 0$. Panel B reports medians of the indicated discrepancies between Euler and Pratt attributions under OLS and Elastic Net estimation. Panel C reports the frequency of negative Euler contributions, the share of total contribution mass attributable to negative C_j, and empirical coverage of analytical 95% confidence intervals.

be close to zero, leading to numerically unstable proportional allocations even when the underlying level contributions remain well behaved.

5 Relation to Existing Measures

Feature-importance measures differ fundamentally in the object they seek to explain. The framework developed here attributes *predictive accuracy*, measured as the reduction in loss relative to a baseline predictor, to the components of a *fixed fitted model*. The resulting attribution is explicitly model-conditional and answers the question: *which features actually generated the predictive performance of the model that was used?*

Many existing measures instead assess association with the outcome, sensitivity of predictions to inputs, or counterfactual performance under feature removal or refitting. These quantities are useful for other purposes, but they generally do not decompose the realized predictive accuracy of a given fitted model.

Measures that refit models under alternative feature combinations are clearly different from our model-conditional contributions. They can assess the potential usefulness of features within a modeling approach, but they cannot attribute feature contributions to model fit for a specific, fixed model. In particular, refitting-based measures are poorly suited to identifying models that include the right features but assign them incorrect parameters.

5.1 Shapley and Perturbation Methods

Decompositions based on Shapley (1953), feature perturbations, and feature permutations are general attribution tools that can be applied to a wide range of model outputs or performance metrics. In practice, they are most commonly used to explain *individual predictions* by attributing deviations of a prediction function $\widehat{y} = f(x)$ from a baseline such as the unconditional mean; see Lundberg and Lee (2017), for example. These prediction-level explanations can be useful for interpreting individual predictions, but they address a fundamentally different question than attribution of predictive accuracy for the model overall.

In principle, Shapley values can be applied to measures of model fit, such as mean squared error or R^2. Doing so requires counterfactual evaluation under feature removal or refitting. Without refitting, Shapley and perturbation methods measure the sensitivity of a fixed model to input disruption; with refitting, they measure feature substitutability across alternative models. In either case, the resulting attributions do not decompose the realized performance of the fitted model actually used.

This distinction is particularly stark in sparse or regularized models. A feature excluded from the fitted model contributes nothing to realized predictive accuracy and therefore receives zero Euler attribution. With

refitting, however, the same feature may receive positive importance because it can substitute for other features in counterfactual re-optimizations.

Ordinary least squares again constitutes a special case. Because explained variance is a homogeneous quadratic function of the fitted coefficients, Shapley values for explained variance computed without refitting coincide with Euler attributions. This equivalence reflects the quadratic structure of least squares and does not extend to loss-based accuracy measures, regularized models, or out-of-sample evaluation.

Shapley and perturbation methods are therefore best interpreted as tools for assessing feature reliance, robustness, or substitutability under information removal. They are also computationally expensive due to the large number of feature subsets they evaluate. By contrast, the Euler decomposition provides an exact, additive, model-conditional attribution of explained predictive accuracy at negligible computational cost once fitted values are available.

Unlike the Euler contributions, related measures of feature importance generally do not provide standard errors, even though they are clearly subject to sampling variation. This may be because they are often treated as indicators, not statistics, or because the standard errors are challenging to derive, which is certainly true under re-fitting of the model.

5.2 Informal Measures

A number of *ad hoc* feature-importance measures are widely used in practice, including standardized coefficients, squared standardized coefficients, absolute t-statistics, and marginal correlations with the dependent variable. These measures are appealing for their simplicity but do not have a principled interpretation as contributions to predictive accuracy.

Standardized coefficients adjust for regressor scale but ignore interactions among regressors in producing the fitted signal. Absolute t-statistics and p-values measure statistical significance rather than contribution to model performance. Marginal correlations reflect association with the outcome rather than contribution to the fitted model.

While these quantities can be useful for exploratory analysis or hypothesis testing, they address questions distinct from the decomposition of realized predictive accuracy considered here.

6 Extensions

We briefly discuss that we can easily group Euler contributions and that Euler contributions apply to a surprisingly broad class of prediction models.

6.1 Grouped Euler Decomposition

Because Euler contributions sum to total explained predictive accuracy, we can aggregate them naturally across groups of components to assess group-level importance.

The Euler decomposition applies to any number of additive components, including settings in which the number of components exceeds the number of observations. When components are numerous or highly collinear, individual contributions may be small or noisy, reflecting redundancy or cancellation within the fitted model. In such cases, aggregating related components yields more stable and interpretable attributions of predictive performance.

Suppose the prediction components are partitioned *ex ante* into disjoint groups. Define the contribution of group G as

$$C_G = \sum_{j \in G} C_j. \tag{23}$$

By additivity of Euler contributions,

$$\Delta\mathcal{L} = \sum_G C_G, \tag{24}$$

so the decomposition allocates total predictive accuracy exactly across groups.

Grouped Euler attribution reconciles diffuse importance across interchangeable individual components with concentrated attribution at the level of shared information sources. This logic parallels Owen values (Owen, 1977), which provide group-wise Shapley allocations, but avoids the combinatorial cost of counterfactual evaluation. Once the fitted model is available, we obtain grouped Euler contributions by direct aggregation at essentially no additional computational cost.

6.2 Scope of Euler Attribution

The Euler decomposition applies to any prediction method whose fitted signal admits a meaningful additive decomposition. The attribution is model-conditional: it allocates realized predictive performance to the fitted prediction components, regardless of how the prediction was constructed.

Any fitted signal that can be written as

$$\widehat{y} = \sum_j \widehat{y}_j \tag{25}$$

therefore supports Euler attribution, which assigns importance directly to the additive prediction components $\widehat{y}_j$. This logic does not rely on least squares or on orthogonality conditions.

Linear regression provides the canonical example, since the fitted signal decomposes naturally into regressor-specific components $\widehat{y}_j = X_j\widehat{\beta}_j$. The same additive structure extends immediately to weighted and generalized least squares, where Euler attribution follows after applying the implied weighting or whitening transformation. Penalized linear models, including Ridge (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996; Zou, 2006), and Elastic Net (Zou and Hastie, 2005), likewise admit Euler attribution because their fitted signals remain linear combinations of regressors, even though these models violate ordinary least squares orthogonality conditions.

Generalized linear models also admit Euler attribution when we take the fitted signal to be the linear predictor $\widehat{\eta} = X\widehat{\beta}$, rather than the conditional mean $g^{-1}(\widehat{\eta})$. On this scale, which often provides the most natural object for interpretation, the fitted signal decomposes additively into regressor-specific components $\widehat{\eta}_j = X_j\widehat{\beta}_j$.[7]

More broadly, any model with an explicitly additive predictor supports Euler attribution on that scale. Generalized additive models provide such a decomposition by construction. Many machine learning methods likewise produce fitted signals that are additive in meaningful internal components; see Hastie, Tibshirani, and Friedman (2009). Ensemble methods, including boosting and random forests, express predictions as sums of weak learners, while kernel methods admit additive representations in terms of training examples or kernel components.

We treat prediction components as primitive and do not require them to correspond to original input features. In polynomial models, neural networks, and other nonlinear architectures, predictions are linear in large collections of constructed features or internal activations, and Euler attribution assigns realized predictive performance directly to these components.

When the object of interest is attribution to the original input variables themselves, a direct additive decomposition of the fitted signal is generally unavailable outside linear models. In such settings, attribution requires aggregating marginal effects across nonlinear and interaction terms. The path-integral attribution developed in Hentschel (2026) extends the loss-based logic to input space, even when prediction components are not linear in the original features.

[7] Attribution on the mean scale generally requires nonlinear transformations and does not admit a simple additive decomposition.

7 Conclusion

This paper develops an Euler decomposition of realized predictive accuracy for regression models with additive prediction components. Measuring model performance as the reduction in mean squared error relative to an intercept-only baseline yields an exact, additive, and model-conditional attribution of explained fit across the components of a fitted prediction.

The resulting attribution answers a practical question: How much has each component of a deployed model contributed to its realized predictive accuracy? Unlike refitting-, perturbation-, or permutation-based approaches, the decomposition conditions on the prediction actually used and avoids counterfactual feature removal or re-optimization. This makes it well suited for model monitoring, diagnostics, and performance attribution.

The framework applies to any predictive system whose fitted signal admits an additive decomposition into components of interest. Because Euler attribution operates solely on realized predictions and their components, it is computationally cheap relative to model estimation, making frequent model evaluation feasible.

We also derive standard errors for the Euler contributions that reflect sampling variability in the evaluation data while conditioning on the fitted model rather than the estimation process. These standard errors enable formal inference on feature importance and facilitate monitoring of contribution stability over time.

Under ordinary least squares evaluated in sample, orthogonality conditions cause explained predictive accuracy to coincide with explained variance, and proportional Euler attributions are equal to familiar variance-based decompositions such as the Pratt allocation. Outside this special case, including out-of-sample evaluation and regularized or weighted estimation, explained variance no longer coincides with predictive accuracy. The Euler decomposition, by contrast, remains well defined and additive.

8 References

Bring, Johan, 1995, Variable importance by partitioning R^2, *Statistical Papers* 36 (1), 1–16.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, second edition (Springer, New York).

Hentschel, Ludger, 2026, Feature importance: Decomposing model fit in nonlinear regressions, Working paper, Versor Investments, New York, NY.

Hoerl, Arthur E., and Robert W. Kennard, 1970, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12 (1), 55–67.

Kruskal, William, 1987, Relative importance by averaging over orderings, *The American Statistician* 41 (1), 6–10.

Lindeman, Richard H., Peter F. Merenda, and Ruth Z. Gold, 1980, *Introduction to Bivariate and Multivariate Analysis* (Scott, Foresman, Glenview, IL).

Litterman, Robert, 1996, Hot spots and hedges, *Journal of Portfolio Management* 23 (5), 52–75.

Lundberg, Scott M., and Su-In Lee, 2017, A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777 (Curran Associates Inc., Red Hook, NY).

Owen, Guillermo, 1977, Values of games with a priori unions, in Rudolf Henn, and Otto Moeschlin, eds., *Mathematical Economics and Game Theory*, volume 141 of *Lecture Notes in Economics and Mathematical Systems*, 76–88 (Springer, Berlin, Heidelberg).

Pratt, John W., 1987, Dividing the indivisible: Using simple symmetry to partition variance explained, in Timo Pukkila, and Simo Puntanen, eds., *Proceedings of the Second International Conference in Statistics*, 245–260 (University of Tampere, Tampere, Finland).

Shapley, Lloyd S., 1953, A value for n-person games, *Contributions to the Theory of Games* 2, 307–317.

Tasche, Dirk, 2008, Capital allocation to business units and sub-portfolios: The Euler principle, in Andrea Resti, ed., *Pillar II in the New Basel Accord: The Challenge of Economic Capital*, 423–453 (Risk Books, London).

Thomas, D. Roland, Edward Hughes, and Bruno D. Zumbo, 1998, On variable importance in linear regression, *Social Indicators Research* 45, 253–275.

Tibshirani, Robert, 1996, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B* 58 (1), 267–288.

Zou, Hui, 2006, The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association* 101 (476), 1418–1429.

Zou, Hui, and Trevor Hastie, 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2), 301–320.

A Bivariate Regression Illustration

This appendix provides a simple bivariate illustration of the Euler decomposition of improvement in mean squared error and contrasts it with marginal-correlation-based allocations such as the Pratt (1987) decomposition. The example is purely analytic and clarifies how Euler contributions behave in the presence of correlated regressors.

A.1 Setup

Let X_1 and X_2 be centered regressors with unit variance and correlation ρ, so that their covariance matrix is

$$\Sigma_X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{26}$$

Consider a linear prediction of the form

$$\widehat{y} = \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2, \tag{27}$$

where $\widehat{\beta}_1$ and $\widehat{\beta}_2$ need not be ordinary least squares estimates.

We assume that the outcome y is centered, so that the intercept-only baseline prediction is zero. We measure predictive performance by the reduction in mean squared error relative to this baseline,

$$\Delta\mathcal{L} = \mathrm{Var}(y) - \mathrm{Var}(y - \widehat{y}) = 2\,\mathrm{Cov}(y, \widehat{y}) - \mathrm{Var}(\widehat{y}). \tag{28}$$

A.2 Euler Contributions

The fitted prediction admits the additive decomposition

$$\widehat{y} = \widehat{y}_1 + \widehat{y}_2, \qquad \widehat{y}_j = \widehat{\beta}_j X_j. \tag{29}$$

The Euler decomposition yields feature-level contributions

$$C_j = 2\,\mathrm{Cov}(y, \widehat{y}_j) - \mathrm{Cov}(\widehat{y}, \widehat{y}_j). \tag{30}$$

In the bivariate case, these take the explicit form

$$C_1 = 2\widehat{\beta}_1\,\mathrm{Cov}(y, X_1) - \widehat{\beta}_1\big(\widehat{\beta}_1 + \rho\,\widehat{\beta}_2\big), \tag{31}$$

$$C_2 = 2\widehat{\beta}_2\,\mathrm{Cov}(y, X_2) - \widehat{\beta}_2\big(\widehat{\beta}_2 + \rho\,\widehat{\beta}_1\big). \tag{32}$$

A.3 Comparison with Pratt Allocation

The Pratt (1987) decomposition assigns importance proportional to

$$P_j = \widehat{\beta}_j \operatorname{Cov}(y, X_j), \tag{33}$$

where $\widehat{\beta}_j$ are ordinary least squares estimates. These attribution components reflects marginal association between each regressor and the outcome.

Under ordinary least squares estimation, in the estimation sample, the orthogonality conditions imply

$$\operatorname{Cov}(y - \widehat{y}, X_j) = 0, \tag{34}$$

so that

$$\operatorname{Cov}(y, X_j) = \operatorname{Cov}(\widehat{y}, X_j) = (\Sigma_X \widehat{\beta})_j. \tag{35}$$

Substituting into the Euler contribution yields

$$C_j = \widehat{\beta}_j \operatorname{Cov}(y, X_j) = P_j, \tag{36}$$

up to a common scaling factor. Thus, in-sample under ordinary least squares estimation for $\widehat{\beta}$, the Euler and Pratt decompositions induce identical proportional attributions, despite decomposing different objects.

Outside this special case, for example out of sample, under regularization, or for misspecified models, the orthogonality conditions fail. Then $\operatorname{Cov}(y, X_j) \neq \operatorname{Cov}(\widehat{y}, X_j)$, and the two attributions diverge.

A.4 Interpretation

The bivariate example highlights the central distinction emphasized in this paper. The Euler decomposition allocates realized predictive performance of a fixed prediction by measuring how each component contributes to reducing mean squared error. The attribution depends on both alignment with the outcome and interaction with other fitted components.

By contrast, the Pratt (1987) decomposition attributes marginal association with the outcome and coincides with Euler attribution only under the orthogonality conditions imposed by in-sample ordinary least squares. Outside that special case, only the Euler decomposition continues to provide a coherent, additive allocation of realized predictive accuracy.

B Standard Errors

This appendix derives standard errors for the Euler contributions to regression fit

$$C_j = 2\ \mathrm{Cov}(y, \widehat{y}_j) - \mathrm{Cov}(\widehat{y}, \widehat{y}_j), \tag{37}$$

for a fixed model $\widehat{y} = \sum_{k=1}^{K} \widehat{y}_k$ with additive structure. We evaluate these contributions in a given sample. The standard error calculation is the same for training and test samples. This is true because we condition on the fitted prediction function $\widehat{y}(\cdot)$ and treat the model as fixed. Under a given model, inference reflects sampling variability in the empirical covariances used to evaluate model fit, not uncertainty due to re-estimation of the model.

The standard errors are useful for assessing whether observed variations in contributions C_j reflect sampling variability in the evaluation data or meaningful changes in the relevance of individual prediction components.

B.1 Euler Contributions as Covariances

In deriving the standard errors for the contributions, it is helpful to express the contributions as a single covariance. Let

$$\widetilde{a}_i = 2y_i - \widehat{y}_i \qquad \text{and} \qquad \widetilde{b}_{ij} = \widehat{y}_{ij}. \tag{38}$$

Here, subscript i refers to observation i and subscript j to prediction component j. Define the centered versions $a_i = \widetilde{a}_i - \mathbb{E}[\,\widetilde{a}\,]$ and $b_{ij} = \widetilde{b}_{ij} - \mathbb{E}[\,\widetilde{b}\,]$, where expectations are sample means over N observations. To simplify notation, we drop tildes after centering. The sample means $\mathbb{E}[\,\widetilde{a}\,]$ and $\mathbb{E}[\,\widetilde{b}\,]$ correspond to the mean-only baseline. Then, each contribution can be written as a single covariance,

$$C_j = \mathrm{Cov}(a, b_j) = \mathbb{E}[a_i\, b_{ij}]. \tag{39}$$

We can collect c_{ij} into a K-vector

$$c_i = (c_{i1}, \ldots, c_{iK})^\top, \qquad c_{ij} = a_i\, b_{ij}. \tag{40}$$

Now, we can write the vector of all Euler contributions as

$$C = (C_1, \ldots, C_K)^\top = \mathbb{E}[c_i]. \tag{41}$$

B.2 Covariance Estimate

Under i.i.d. sampling on the evaluation sample, define the per-observation covariance matrix

$$\Sigma = \mathrm{Cov}(c_i) = \mathbb{E}\big[(c_i - \mathbb{E}[c_i])(c_i - \mathbb{E}[c_i])^\top\big]. \tag{42}$$

We estimate Σ by the usual sample covariance of $\{c_i\}$,[8]

$$\widehat{\Sigma} = \mathbb{E}\big[(c_i - C)(c_i - C)^\top\big], \tag{43}$$

Because $C = \mathbb{E}[c_i]$ is the sample average of c_i across N observations, we have $\mathrm{Cov}(C) = \Sigma/N$. Finally, the standard error for C_j is

$$\mathrm{SE}(C_j) = \sqrt{\frac{1}{N}\widehat{\Sigma}_{jj}} = \sqrt{\frac{1}{N}\mathbb{E}\big[(c_{ij} - C_j)^2\big]}. \tag{44}$$

For a group of features $G \subseteq \{1, \ldots, K\}$ we can define the 0/1 indicator vector $\mathbf{1}_G \in \mathbb{R}^K$ and define the grouped contribution $C_G = \mathbf{1}_G^\top C$. Then, the variance of the grouped contributions is

$$\widehat{\sigma}^2(C_G) = \frac{1}{N}\mathbf{1}_G^\top \widehat{\Sigma}\,\mathbf{1}_G \tag{45}$$

and

$$\mathrm{SE}(C_G) = \sqrt{\frac{1}{N}\mathbf{1}_G^\top \widehat{\Sigma}\,\mathbf{1}_G} = \sqrt{\frac{1}{N}\mathbb{E}[(c_{iG} - C_G)^2]}. \tag{46}$$

Although we have derived an analytical $K \times K$ covariance matrix, we do not need to estimate the full covariance matrix $\widehat{\Sigma}$. The standard error calculations in equation (44) require only one variance per feature; the standard error calculations in equation (46) require only one variance per grouped feature. Even for large K, we can compute these standard errors without any need to estimate or regularize a full covariance matrix.

In-sample versus out-of-sample evaluation

Because we are decomposing the fit of a given model, the derivation conditions on the fitted model $\widehat{y}(\cdot)$. As a result, the same formulas apply in-sample and out-of-sample. The only difference is the sample used to form c_i and its size N. The calculations are the same.

[8] If the evaluation sample exhibits heteroskedasticity or serial dependence, we can replace the i.i.d. estimator (43) with a HAC estimator applied to the time series $\{c_{it}\}$.

OLS as a special case

If we estimate the model by ordinary least squares on the same sample, the normal equations imply $\text{Cov}(e, \widehat{y}_j) = 0$ identically, where $e = y - \widehat{y}$. In this case, the contribution reduces algebraically to

$$C_j = \text{Cov}(y, \widehat{y}_j). \tag{47}$$

Of course, the vanished term is not estimated with error; it is exactly zero for every sample under in-sample OLS. The covariance estimator (43) applies directly to the reduced estimator, which coincides algebraically with (37). Once again, the calculations are the same.

Auxiliary regression interpretation

Equation (41) implies that we can view each contribution C_j as the intercept in a constant-only auxiliary regression

$$c_{ij} = \alpha_j + \varepsilon_{ij}, \tag{48}$$

with $c_{ij} = a_i b_{ij}$, as before. The OLS estimator of the intercept satisfies $\widehat{\alpha}_j = \overline{c}_j = C_j$, and the corresponding OLS standard error is equation (44).