Feature Importance for Model Fit: Decomposing Log Loss in Binary Classification

January 18, 2026

Abstract

We use a path-integral generalization of Euler's theorem to decompose the predictive accuracy of binary classification models, measured by improvement in expected log loss. Relative to a constant-probability baseline, this improvement admits an exact additive decomposition across components of the fitted score, yielding a principled notion of feature importance as contribution to global predictive fit.

Because log loss is nonlinear in the score, an exact decomposition requires averaging the canonical residual, the derivative of log loss with respect to the score, along the path from the baseline score to the fitted score. The resulting attribution is exact and does not rely on optimality conditions or local approximations. Specializing to logistic regression, the natural components are $\beta_j x_j$, directly mirroring the fitted-value decomposition in linear regression.

In this framework, a component contributes positively to predictive performance if it moves the fitted score closer to the outcome, either by directly explaining the outcome or by offsetting misalignment introduced by other components.

www.ludgerhentschel.com ludger@ludgerhentschel.com

Contents

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

1 Introduction

We develop an additive decomposition of model fit for binary classification models. We extend the Euler decomposition of explained signal strength in regression models proposed by Hentschel (2026b) to settings with binary outcomes $y \in \{0, 1\}$ and fitted probabilistic predictions $\widehat{p} \in (0, 1)$.

We define feature importance as a model-conditional attribution of predictive fit, measured by the improvement in expected log loss relative to a constant-probability baseline. This definition attributes performance within the fitted model itself, rather than to marginal associations, local sensitivities, or counterfactual feature removal.

In linear regression, the reduction in mean squared error admits an exact Euler decomposition because it is homogeneous in the fitted signal. In classification models, squared error loss does not provide a natural measure of fit. Probabilistic classification models instead evaluate predictions using log loss, a proper scoring rule. Because log loss is nonlinear in the fitted score, a direct Euler decomposition does not apply, and variance-based regression results do not extend mechanically. We address this nonlinearity by averaging the canonical residual along the path from the baseline score to the fitted score. This construction yields an exact, additive, and computationally tractable attribution.

Our analysis focuses on the fitted predictions themselves and does not depend on the estimation procedure that produced them. We derive the decomposition without invoking optimality conditions, likelihood equations, or refitting-based arguments. Under log loss, the decomposition admits a closed-form expression and retains the key advantages of the regression case: additivity, interpretability, and low computational cost.

We also derive standard errors for the Euler contributions that reflect sampling variability in the data. These standard errors allow us to assess whether observed variation in feature contributions across samples or over time plausibly reflects noise or instead indicates changes in predictive relevance.

A large literature proposes feature importance measures based on coefficients, perturbation or permutation schemes, and Shapley-value attributions. These methods address important but different questions, including sensitivity of predictions to inputs, robustness to feature removal, and explanation of individual predictions. In contrast, we focus on decomposing a fitted model's global predictive fit. Section 3 provides a detailed comparison with existing approaches.

The remainder of the paper proceeds as follows. Section 2 develops improvement in log loss relative to a baseline model as a measure of clas-

sification fit and derives the core Euler-style decomposition of this model fit. Section 3 compares the proposed decomposition to existing feature-importance measures. Section 4 concludes. The appendices discuss the path-integral generalization of EulerÕs theorem, standard errors for the feature contributions, the extension to multinomial classification, and the computational algorithm for the decomposition.

2 Decomposing Model Fit in Binary Classification

This section develops the core decomposition of predictive fit for binary classification models. We begin by formalizing model fit using log loss and defining explained fit relative to a constant-probability baseline. We then derive an exact, additive decomposition of this explained fit at the level of fitted score components, discuss extensions to weighted evaluation, and interpret the resulting contributions as measures of feature importance. The section concludes by specializing the general framework to logistic regression, where the decomposition takes a particularly transparent form.

2.1 Log Loss as a Measure of Classification Fit

We measure model fit using improvement in expected log loss rather than threshold-based classification accuracy. In probabilistic classification, log loss plays the same conceptual role as explained variance in regression: it evaluates how well a model explains outcomes relative to a baseline, rather than how often it produces correct classifications at an arbitrary threshold.

Log loss is a strictly proper scoring rule for binary outcomes: the expected loss is uniquely minimized when the predicted probability equals the true conditional probability Gneiting and Raftery (2007). As a result, log loss evaluates the quality of probabilistic predictions directly, rewards calibration, and admits well-defined population expectations.

These properties distinguish log loss from threshold-dependent metrics such as accuracy or F1, and from rank-based measures such as AUC. Such metrics do not evaluate probabilistic forecasts directly and do not support additive decompositions of model fit. By contrast, log loss aggregates additively across observations and is therefore well suited to the decomposition developed in this paper.

We evaluate model fit using the log loss

$$\ell(y, \widehat{p}) = -y \log \widehat{p} - (1 - y) \log(1 - \widehat{p}). \tag{1}$$

Although log loss coincides with the negative log likelihood for Bernoulli models, we use it here purely as an evaluation metric. The decomposition

applies to any classification model that produces probabilistic predictions, regardless of how those predictions were constructed.

2.2 Baseline Model and Explained Log Loss

As in regression, we define explained fit relative to a constant baseline model. We take this baseline to be the unconditional class probability $\overline{p} = \mathbb{E}[y]$, corresponding to an intercept-only model. This choice is directly analogous to centering the dependent variable in regression: if the baseline is misspecified, the notion of explained fit itself becomes ambiguous.

We define explained log loss as the reduction in expected log loss relative to this baseline,

$$\Delta\mathcal{L} = \mathbb{E}[\ell(y, \overline{p})] - \mathbb{E}[\ell(y, \widehat{p})]. \tag{2}$$

This quantity is the natural classification analogue of explained variance. It measures how much predictive information the fitted model captures beyond unconditional class frequencies.

2.3 A Model-Level Decomposition

Define the log score

$$f(y, \widehat{\eta}) = y\widehat{\eta} - \log(1 + e^{\widehat{\eta}}), \tag{3}$$

where $\widehat{\eta}$ denotes the fitted score (log-odds) associated with the predicted probability $\widehat{p} = \sigma(\widehat{\eta})$. Let $\overline{\eta} = \log(\overline{p}/(1 - \overline{p}))$ denote the baseline score. The improvement in model fit can be written as

$$\Delta\mathcal{L} = \mathbb{E}[f(y, \widehat{\eta}) - f(y, \overline{\eta})]. \tag{4}$$

Although the logistic function $\sigma(\eta) = 1/(1 + e^{\eta})$ is commonly associated with logistic regression estimation, it enters our analysis solely through the choice of log loss as the evaluation metric. The decomposition therefore depends only on the form of the scoring rule and the fitted scores, and not on the estimation procedure or objective function that produced them.

In linear regression with squared error loss, homogeneity of the loss function allows the improvement in fit to be decomposed by direct application of Euler's theorem.[1] For log loss, this property no longer holds: the log score is not homogeneous of degree one in the fitted score $\widehat{\eta}$. As a result, an endpoint evaluation of the score derivative does not yield an exact decomposition of global model fit.

[1] See Silberberg (1978) or Tasche (2008).

Instead, we apply the fundamental theorem of calculus along the straight-line path from the baseline score $\overline{\eta}$ to the fitted score $\widehat{\eta}$ on the score scale. For each observation,

$$f(y,\widehat{\eta}) - f(y,\overline{\eta}) = \int_0^1 \frac{d}{dt} f\left(y,\ \overline{\eta} + t(\widehat{\eta}-\overline{\eta})\right) dt \tag{5}$$

$$= (\widehat{\eta}-\overline{\eta}) \int_0^1 \left(y - \sigma\left(\overline{\eta} + t(\widehat{\eta}-\overline{\eta})\right)\right) dt. \tag{6}$$

Appendix A shows that this decomposition is the natural generalization of Euler's theorem when homogeneity fails: endpoint derivatives are replaced by averages of the directional derivative along a path.

Now, assume the fitted score admits an additive representation

$$\widehat{\eta} = \overline{\eta} + \sum_{j=1}^{K} \widehat{\eta}_j. \tag{7}$$

Substituting this representation yields an additive decomposition of the log-loss improvement,

$$\Delta\mathcal{L} = \sum_{j=1}^{K} C_j, \tag{8}$$

with component contributions

$$C_j = \mathbb{E}\left[\widehat{\eta}_j \int_0^1 \left(y - \sigma\left(\overline{\eta} + t(\widehat{\eta}-\overline{\eta})\right)\right) dt\right]. \tag{9}$$

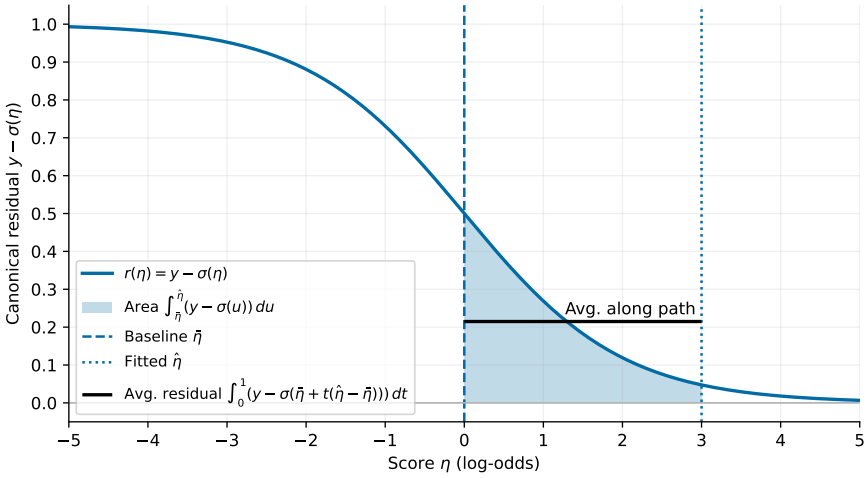
For log loss, the path integral admits a closed-form solution because the loss is generated by a convex log-partition function. We retain the integral representation to make explicit the nonlinear geometry underlying the additive decomposition. For log loss, the closed-form contributions are

$$C_j = \mathbb{E}\left[\widehat{\eta}_j \left(y - \frac{\log(1+e^{\widehat{\eta}}) - \log(1+e^{\overline{\eta}})}{\widehat{\eta}-\overline{\eta}}\right)\right], \tag{10}$$

with the ratio interpreted by continuity as $\sigma(\overline{\eta})$ when $\widehat{\eta} = \overline{\eta}$. This closed form preserves exact additivity while avoiding numerical integration.

The contribution C_j measures how much the score component $\widehat{\eta}_j$ shifts predicted probability mass in regions where the model is still uncertain about the outcome. Each component is weighted not by its raw magnitude, but by how strongly it aligns with residual predictive uncertainty, averaged along

Figure 1: Logistic Residual vs. Score and the Path Integral



The figure illustrates the path-integral calculation used to attribute model fit to a single score component. The plot shows the canonical residual $y - \sigma(\eta)$ as a function of the score η for a single observation with $y = 1$, a baseline score $\overline{\eta} = 0$, and a fitted score $\widehat{\eta} = 3$. Here, $\sigma(\cdot)$ is the logistic function. The shaded region corresponds to the path integral

$$\int_0^1 \left(y - \sigma\left(\overline{\eta} + t(\widehat{\eta} - \overline{\eta})\right)\right)\, dt,$$

taken along the straight-line path from the baseline score $\overline{\eta}$ to the fitted score $\widehat{\eta}$. The horizontal line indicates the average residual over this path. Multiplying this average residual by the score contribution $\widehat{\eta}_j$ of a component yields that componentâĂŹs contribution to the improvement in log loss for the observation.

the path from the baseline score to the fitted score.

This mirrors the regression case, where a component's importance is governed by its covariance with the fitted signal. In classification, the analogous object is the alignment between a score component and the canonical residual $y - \sigma(\cdot)$, integrated along the score path. Score movements that occur when predicted probabilities are near 0 or 1 contribute little to fit, because the model is already confident; score movements near the decision boundary contribute more, because they meaningfully reduce uncertainty.

Figure 1 illustrates the geometry of the decomposition. The canonical residual $y - \sigma(\eta)$ varies nonlinearly with the score η, reflecting the diminishing marginal value of increasingly confident predictions. The contribution of a score component is proportional to the area under this residual curve along the path from the baseline score to the fitted score. The integral therefore weights score movements by how much predictive uncertainty they resolve, rather than treating all movements equally.

In logistic regression, where $\widehat{\eta}_j = \widehat{\beta}_j x_j$, this interpretation is especially transparent: a feature is important if it systematically pushes the score in

directions that resolve uncertainty across the sample. Large coefficients attached to rarely informative features contribute little, while modest coefficients acting where predictions are uncertain can contribute substantially to model fit.

The decomposition $\Delta\mathcal{L} = \sum_j C_j$ is complete in the sense that the contributions sum exactly to the total improvement in expected log loss. It attributes model fit at the level of fitted score components and applies to any classification model that produces an additive score representation, regardless of how the fitted probabilities were obtained.

Additivity is assumed on the score scale rather than on the probability scale. While the logistic link is nonlinear, the fitted score $\widehat{\eta}$ is linear and homogeneous of degree one in its components. The nonlinearity of the link affects the form of the residual term through the path integral but does not alter the additive structure of the decomposition.

Algorithm 1 in appendix D outlines how to compute these contributions in practice, with the addition of weighted cases.

2.4 Weighted Evaluation

The same decomposition applies when predictive fit is evaluated under a weighted empirical measure.

Let $w_i \geq 0$ denote observation weights defining the empirical measure under which predictive fit is evaluated, and define $\mathbb{E}_w[g] = \sum_i \tilde{w}_i g_i$ with $\tilde{w}_i = w_i / \sum_m w_m$. The constant-probability baseline becomes $\bar{p}_w = \mathbb{E}_w[y]$, with corresponding score $\bar{\eta}_w = \log(\bar{p}_w/(1-\bar{p}_w))$, and explained fit is measured by the improvement in weighted expected log loss. Replacing $\mathbb{E}[\cdot]$ by $\mathbb{E}_w[\cdot]$ throughout yields an exact additive decomposition of weighted log-loss improvement.[2]

Throughout, we interpret weights as defining the metric in which predictive fit is evaluated, not as ad hoc adjustments for class imbalance or cost sensitivity. This interpretation parallels the role of weights in WLS regression and ensures that the resulting attributions remain model-conditional and properly calibrated.

[2] In contrast to linear regression, there is no direct analogue of generalized least squares for Bernoulli outcomes, because binary responses do not admit additive noise with an unrestricted covariance structure in the same sense as Gaussian regression. As a result, weighting in classification is naturally interpreted as defining the evaluation metric for predictive fit, rather than as correcting for correlated errors. Accordingly, all extensions beyond the unweighted case enter solely through weighted expectations of log loss, in direct analogy with weighted least squares evaluation in regression.

2.5 Interpretation and Feature Importance

The contribution C_j depends on the magnitude of the score component $\widehat{\eta}_j$ and on its alignment with an integrated residual term. The integral averages the canonical residual $y - \sigma(\cdot)$ along the straight-line path on the score scale from the baseline score to the fitted score. For squared error loss, this path integral collapses to an endpoint expression because the loss is quadratic. For log loss, the integral remains, making explicit how nonlinearity affects attribution while preserving additivity across fitted components.

As in the regression setting, this defines feature importance as a property of the fitted model rather than of marginal relationships in the data. Measures based on correlations, mutual information, or univariate classification accuracy describe predictive structure present in the data but need not align with the contribution of a component to the fitted model once other components are included.

2.6 Specialization to Logistic Regression

In a logistic regression model, the predicted probabilities are obtained from a fitted score (log odds)

$$\widehat{p} = \sigma(\widehat{\eta}), \qquad \widehat{\eta} = \widehat{\beta}_0 + \sum_{j=1}^{K} \widehat{\beta}_j x_j. \tag{11}$$

We evaluate model fit relative to the constant-probability baseline $\overline{p} = \mathbb{E}[y]$, with baseline score

$$\overline{\eta} = \log\left(\frac{\overline{p}}{1-\overline{p}}\right). \tag{12}$$

To apply the decomposition, we write the fitted score as a baseline plus additive components,

$$\widehat{\eta} = \overline{\eta} + \sum_{j=1}^{K} \widehat{\eta}_j \tag{13}$$

$$\widehat{\eta}_j = \widehat{\beta}_j x_j, \tag{14}$$

so that the intercept is absorbed into the baseline. Substituting these components into the general decomposition in equation (10) yields the feature-level attributions C_j. By construction, the contributions satisfy $\Delta\mathcal{L} = \sum_{j=1}^{K} C_j$.

The baseline score $\overline{\eta}$ corresponds to the intercept-only model defined by the unconditional class probability and should be distinguished from the fitted intercept $\widehat{\beta}_0$ of the full logistic regression. When regressors are not

Table 1: Conceptual classification of feature-importance measures

Method class	Object	Model fixed	Mechanism	Scope
Coefficients	Pred/score	Yes	Local sens	Global
Perm. / Pert.	Acc/pred	No	Discrt rem	Global
Shapley	Pred	No	Discrt avg	Local (agg)
Int Grads	Pred	Yes	Cont path	Local
Euler	Acc	Yes	Cont path	Global

The table summarizes high-level characteristics of feature importance measures.

centered, $\widehat{\beta}_0 \neq \overline{\eta}$, but this distinction has no effect on the decomposition: Any constant shift in the score cancels in the baseline comparison and therefore contributes nothing to $\Delta\mathcal{L}$. Accordingly, the attribution is carried entirely by the components $\widehat{\eta}_j = \widehat{\beta}_j x_j$, regardless of whether regressors are centered.[3]

In linear regression, a fitted component contributes to explained variance in proportion to its covariance with the overall fitted signal. In logistic regression, the analogous object is the covariance between a score component and unresolved classification uncertainty, as captured by the canonical residual. The contribution C_j is large when the component $\widehat{\eta}_j$ varies substantially across observations and tends to move in directions that reduce uncertainty in the outcome, averaged over the entire range from the baseline prediction to the fitted prediction. Components that primarily reinforce already confident predictions or that vary in directions orthogonal to residual uncertainty contribute little to overall model fit.

3 Relation to Other Feature Importance Measures

A large literature proposes measures of feature importance for classification models. These measures differ in the object they attribute, predictions, model performance, or counterfactual behavior, and in whether importance is defined locally or at the level of the fitted model. Table 1 summarizes the main characteristics for several of these feature importance measures.

3.1 Coefficient-Based Measures

In parametric classification models such as logistic regression, feature importance is often assessed using coefficient magnitudes, standardized coefficients, odds ratios, or associated test statistics. These quantities describe the local sensitivity of the fitted score or log odds to changes in individual inputs, holding other inputs fixed.

[3] Centering regressors simplifies interpretation by aligning the fitted intercept with the baseline score and levels the playing field in the presence of regularization, but it is not required for the validity of the decomposition.

Coefficient-based measures do not, however, account for the empirical distribution of the inputs or for how often a feature contributes meaningfully to prediction in the data. As a result, features with large coefficients but little variation may have limited impact on predictive fit, while features with smaller coefficients but substantial variation may be more important in practice. The decomposition proposed here incorporates both effect size and realized variation by attributing improvement in log loss using the fitted score components themselves. Even in correctly specified logistic regression models, coefficient magnitudes do not decompose global log-loss improvement and therefore cannot be interpreted as contributions to predictive fit.

3.2 Perturbation and Permutation Importance

Model-agnostic approaches such as permutation importance and feature masking assess importance by measuring the deterioration in predictive performance when a feature is randomly permuted, corrupted, or removed. These methods are widely used in applied machine learning and provide a notion of model reliance or robustness rather than contribution within the fitted model itself Breiman (2001); Fisher, Rudin, and Dominici (2019).

Perturbation-based measures depend on the perturbation scheme, feature correlations, and the choice of evaluation metric, and they do not yield an additive decomposition of total model fit.

Interpreted this way, perturbation importance measures how much a model depends on a feature for producing its predictions, not how that feature contributes to realized predictive accuracy. In contrast, the present approach decomposes the fitted model's log-loss improvement exactly, without perturbing the data or refitting the model.

3.3 Shapley-Based Attribution Methods

Shapley-value-based methods provide local, observation-level explanations by averaging marginal contributions over all possible subsets of features. In the context of machine learning models, these methods are most prominently associated with SHAP Lundberg and Lee (2017), which adapts the Shapley value framework Shapley (1953); Lindeman, Merenda, and Gold (1980); Kruskal (1987) to prediction models.

Shapley-based attributions answer a different question from the one considered here. They explain how a specific prediction is constructed relative to a baseline, rather than how global predictive performance is generated. Aggregating local Shapley values to obtain global feature importance is possible but requires additional averaging choices and does not generally

produce an exact decomposition of a population-level fit measure such as expected log loss.

Shapley values are additive with respect to the chosen value function $V(S)$, but that value function typically corresponds to counterfactual refit or feature-removal objectives rather than to realized model fit for a fixed fitted model.

3.4 Gradient-Based and Local Explanation Methods

Gradient-based attribution methods, including saliency maps and integrated gradients, assess how predictions change under infinitesimal or path-based perturbations of the inputs. Integrated gradients Sundararajan, Taly, and Yan (2017) are among the most widely used approaches in this class and provide axiomatic guarantees for local attribution of individual predictions.

While these methods also rely on derivatives, their objective differs fundamentally from the present one. Gradient-based methods typically quantify local sensitivity of individual predictions, whereas the decomposition developed here provides a global, additive attribution of predictive fit.

When applied to individual predictions, integrated gradients explain sensitivity. Whey applied to model fit, they can be viewed as a numerical approximation to Euler-style attribution, but only after redefining the value function to be population-level log loss rather than individual predictions.

3.5 Threshold-Dependent Accuracy Metrics

Feature importance is sometimes assessed using changes in classification accuracy, AUC, or related threshold-dependent metrics. These measures depend on arbitrary classification thresholds or ranking behavior and do not evaluate the quality of predicted probabilities.

Because such metrics are non-additive and do not correspond to proper scoring rules, they do not support a principled decomposition of model fit. By working directly with log loss, the present framework avoids threshold dependence and evaluates probabilistic predictions on the scale at which classification models are typically estimated and compared.

3.6 Summary

Existing feature importance measures emphasize local sensitivity, robustness to feature removal, or explanation of individual predictions. The approach developed here addresses a complementary objective: decomposing a model's global predictive fit into additive contributions from components of the fitted score. By operating directly on log loss and conditioning on the fitted model,

the decomposition provides a principled and computationally efficient notion of feature importance aligned with probabilistic classification.

Although alternative approaches could, in principle, be applied to decompose model fit, doing so requires applying them to the log loss of a fixed fitted model without refitting. Shapley methods are maximally general in that they require only evaluations of a value function defined on subsets of features. This generality comes at two costs: substantial computational complexity and the inability to exploit known structure of the value function, such as smoothness or derivative information.

Gradient-based methods attempt to incorporate additional structure by using derivatives to learn about the local shape of the value function. However, when applied to individual predictions, these methods quantify local sensitivity rather than contribution to global predictive performance. Integrated gradients applied to model fit can be viewed as a numerical approximation to Euler-style attribution, trading analytic clarity for generality at increased computational cost.

From the integral formulation in equation (9), we see that the general Euler attribution for differentiable models reduces to a one-dimensional line integral between baseline and fitted models. When analytic solutions are available, as in regression and classification under log loss, this yields exact, additive attributions at minimal cost. More generally, numerical evaluation of this line integral provides a unified approach to attributing predictive fit across a wide class of models Hentschel (2026a).

4 Conclusion

We have developed a model-level decomposition of predictive fit for binary classification models that parallels the Euler decomposition of explained variance in linear regression. By evaluating predictive performance using log loss and working on the additive score scale, the improvement in expected log loss relative to a constant-probability baseline admits an exact additive decomposition across components of the fitted score.

Because log loss is nonlinear in the score, an exact decomposition of global model fit cannot be obtained from endpoint derivatives alone. Instead, it requires averaging the canonical residual along the path from the baseline score to the fitted score. This path-integral formulation provides a natural generalization of Euler's theorem beyond homogeneous objectives and yields additive attributions that depend only on the evaluation metric and the fitted model, not on the estimation procedure that produced it.

Specializing the framework to logistic regression, the natural attribution components are $\widehat{\beta}_j x_j$, directly mirroring fitted-value decompositions in linear regression, with the intercept absorbed into the baseline. In this setting, a feature's contribution to predictive fit reflects both the magnitude of its score component and the extent to which it resolves residual classification uncertainty across the sample.

The resulting notion of feature importance is global and model-conditional: it attributes realized predictive fit within the fitted model actually used. This makes it well suited for monitoring, comparison, and diagnostic analysis of deployed models, where the object of interest is not sensitivity or counterfactual performance, but how predictive accuracy is generated by the model in operation.

We also derive standard errors for the Euler contributions, allowing formal assessment of sampling variability. This makes it possible to distinguish meaningful changes in feature contributions, across samples or over time, from fluctuations attributable to noise, an essential requirement for practical monitoring and inference.

More broadly, the integral formulation clarifies that attributing predictive fit for differentiable models reduces to a one-dimensional path integral between baseline and fitted predictions. When analytic solutions are available, as in regression and classification under log loss, this yields closed-form Euler-style attributions. When they are not, the same framework suggests a unified numerical approach. This perspective provides a principled and scalable foundation for feature attribution across a wide class of predictive models.

5 References

Breiman, Leo, 2001, Random forests, *Machine Learning* 45, 5–32.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici, 2019, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, *Journal of Machine Learning Research* 20 (1), 1–81.

Gneiting, Tilmann, and Adrian E. Raftery, 2007, Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* 102 (477), 359–378.

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

Hentschel, Ludger, 2026b, Feature importance: Euler attribution of regression fit, Working paper, Versor Investments, New York, NY.

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).

Lundberg, Scott M., and Su-In Lee, 2017, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems*.

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

Silberberg, Eugene, 1978, *The Structure of Economics: A Mathematical Analysis* (McGraw–Hill, New York, NY).

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan, 2017, Axiomatic attribution for deep networks, in Doina Precup, and Yee Whye Teh, eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328 (PMLR).

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).

A Euler Decompositions and Path Integrals

This appendix clarifies the relationship between Euler's theorem for homogeneous functions and the path-integral decomposition used in the main text. The purpose is to show that the latter is the natural generalization of the former when homogeneity does not hold.

Although the decomposition is expressed in terms of gradients and integrals, it is not a local or linear approximation. In particular, it does not rely on a Taylor expansion or on evaluating derivatives at a single point. The path-integral identity is an exact restatement of the function difference implied by the fundamental theorem of calculus.

A.1 A Path-Integral Identity for Differentiable Functions

Let $f : \mathbb{R}^K \to \mathbb{R}$ be continuously differentiable, and let $x, x_0 \in \mathbb{R}^K$ be two points. Consider the straight-line path connecting x_0 to x,

$$x(t) = x_0 + t(x - x_0), \qquad t \in [0, 1]. \tag{15}$$

By the chain rule,

$$\frac{d}{dt} f(x(t)) = \nabla f(x(t))^\top (x - x_0). \tag{16}$$

Applying the fundamental theorem of calculus yields the exact identity

$$f(x) - f(x_0) = \int_0^1 \nabla f(x(t))^\top (x - x_0)\, dt. \tag{17}$$

Rewriting the inner product gives an additive decomposition across coordinates,

$$f(x) - f(x_0) = \sum_{j=1}^{K} (x_j - x_{0j}) \int_0^1 \frac{\partial f}{\partial x_j}(x(t))\, dt. \tag{18}$$

This decomposition is additive because the directional derivative is linear in the coordinate increments $x - x_0$. The identity requires only differentiability of f and expresses the change in the function as the sum, over coordinates, of each coordinate increment multiplied by its average marginal effect along the path from x_0 to x.

A.2 Euler's Theorem as a Special Case

Now suppose that f is positively homogeneous of degree one, so that

$$f(\lambda x) = \lambda f(x) \qquad \text{for all } \lambda > 0. \tag{19}$$

Euler's theorem states that

$$f(x) = \sum_{j=1}^{K} x_j \frac{\partial f}{\partial x_j}(x). \tag{20}$$

This result follows directly from the path-integral identity above. Take $x_0 = 0$ and $x(t) = tx$. Then

$$f(x) - f(0) = \int_0^1 \nabla f(tx)^\top x \, dt. \tag{21}$$

For a function homogeneous of degree one, the gradient $\nabla f(tx)$ is homogeneous of degree zero and therefore constant along rays. As a result, $\nabla f(tx) = \nabla f(x)$ for all $t > 0$, and the integrand does not depend on t. Homogeneity therefore eliminates the need to average derivatives along the path. The integral therefore collapses to

$$\int_0^1 \nabla f(x)^\top x \, dt = \nabla f(x)^\top x, \tag{22}$$

which is Euler's theorem. See Silberberg (1978) or Tasche (2008), for example.

A.3 Interpretation

Euler's theorem is an endpoint identity for scalar-valued functions that exploits homogeneity to recover the function value from derivatives evaluated at a single point. Although the statement involves gradients, it is not a local or linear approximation and is fundamentally not a multidimensional gradient expansion. It is an exact identity that reconstructs the function value itself.

When homogeneity does not hold, the appropriate generalization is a path-integral identity: the change in the function equals the integral of its directional derivative along a path connecting a baseline to an evaluation point. The straight-line path we use here treats all components symmetrically by construction, preserves additivity aligned with the fitted model, and reduces exactly to Euler's theorem when homogeneity holds. Alternative paths introduce ordering or counterfactual structure that is not implied by the fitted model itself.

The path-integral decomposition relies not only on differentiability, but also on the existence of a meaningful baseline and evaluation point. In predictive modeling, these are provided naturally by the baseline and fitted models that define explained fit. In the absence of such structure, the decomposition remains algebraically valid but loses its interpretation as an attribution of model performance.

The decomposition used in the main text follows this general principle. In regression, explained signal strength is homogeneous in the fitted signal, and Euler's theorem yields an exact endpoint decomposition. In classification, log loss is not homogeneous in the fitted score, and an exact decomposition requires averaging derivatives along the score path. Both constructions are exact, rely only on properties of the evaluation metric, and do not depend on the estimation procedure that produced the fitted model. The distinction between endpoint and path identities reflects differences in functional structure rather than differences in approximation quality.

B Standard Errors

This appendix derives standard errors for the component contributions C_j to explained log loss in binary classification. Throughout, we condition on the fitted prediction function $\widehat{\eta}(\cdot)$ and treat the model as fixed. Under this conditioning, inference reflects sampling variability in the evaluation sample we use to compute $\Delta\mathcal{L}$ and its decomposition, not uncertainty due to re-estimation of the model.

The standard errors are useful for assessing whether observed variation in contributions C_j across samples or over time reflects sampling variability in the evaluation data or meaningful changes in the relevance of individual prediction components.

B.1 Observation-Level Representation

Recall the log score

$$f(y, \eta) = y\eta - \log(1 + e^{\eta}), \tag{23}$$

the baseline score $\overline{\eta} = \log(\overline{p}/(1 - \overline{p}))$ implied by $\overline{p} = \mathbb{E}[y]$, and the additive score representation

$$\widehat{\eta} = \overline{\eta} + \sum_{j=1}^{K} \widehat{\eta}_j. \tag{24}$$

The improvement in fit is

$$\Delta\mathcal{L} = \mathbb{E}\big[f(y, \widehat{\eta}) - f(y, \overline{\eta})\big] = \sum_{j=1}^{K} C_j, \tag{25}$$

with component contributions (see equation (9))

$$C_j = \mathbb{E}\left[\widehat{\eta}_j \int_0^1 \big(y - \sigma\big(\overline{\eta} + t(\widehat{\eta} - \overline{\eta})\big)\big)\, dt\right]. \tag{26}$$

For log loss, the integral admits the closed form in equation (10). Define the scalar weight

$$w := y - \frac{\log(1 + e^{\widehat{\eta}}) - \log(1 + e^{\overline{\eta}})}{\widehat{\eta} - \overline{\eta}}, \tag{27}$$

interpreting the ratio by continuity as $\sigma(\overline{\eta})$ when $\widehat{\eta} = \overline{\eta}$. Then the contributions satisfy

$$C_j = \mathbb{E}[c_{ij}], \qquad c_{ij} := \widehat{\eta}_{ij}\, w_i, \tag{28}$$

where i indexes observations in the evaluation sample and $\widehat{\eta}_{ij}$ denotes the jth additive score component for observation i.

B.2 Standard Errors for IID Data

Let N denote the size of the evaluation sample. Conditional on the fitted model, C_j is the sample mean of $\{c_{ij}\}_{i=1}^{N}$. Under i.i.d. sampling, the standard error of C_j is the standard error of a sample mean,

$$SE(C_j) = \sqrt{\frac{1}{N}\,\mathbb{E}\big[(c_{ij} - C_j)^2\big]}. \tag{29}$$

A $100(1 - \alpha)\%$ confidence interval can be reported as $C_j \pm z_{1-\alpha/2}\, SE(C_j)$, conditional on the fitted model.

B.3 Grouped contributions.

For any group $G \subseteq \{1, \ldots, K\}$ define the grouped score component $\widehat{\eta}_{iG} = \sum_{j \in G} \widehat{\eta}_{ij}$ and grouped contribution $C_G = \sum_{j \in G} C_j$. Then

$$C_G = \mathbb{E}[c_{iG}], \qquad c_{iG} := \widehat{\eta}_{iG}\, w_i, \tag{30}$$

and the corresponding standard error is

$$SE(C_G) = \sqrt{\frac{1}{N}\,\mathbb{E}[(c_{iG} - C_G)^2]}. \tag{31}$$

Computing $SE(C_j)$ and $SE(C_G)$ requires only scalar variances of $\{c_{ij}\}$ and $\{c_{iG}\}$ and does not require estimating a full $K \times K$ covariance matrix.

B.4 Notes and Extensions

Because we decompose the fit of a given model and condition on the fitted score function $\widehat{\eta}(\cdot)$, the same formulas apply in-sample and out-of-sample. The only difference is the evaluation sample used to form $\{c_{ij}\}$ and its size N.

In logistic regression, the fitted score components take the parametric form $\widehat{\eta}_{ij} = \widehat{\beta}_j x_{ij}$ (with the intercept absorbed into $\overline{\eta}$ as in the main text). Substituting these components into equation (28) yields

$$c_{ij} = (\widehat{\beta}_j x_{ij})\, w_i, \qquad C_j = \mathbb{E}[(\widehat{\beta}_j x_{ij})\, w_i], \tag{32}$$

and the standard errors remain equations (29) and (31). No additional adjustment is required: the derivation depends only on the evaluation metric and the fitted scores, not on the estimation procedure used to obtain $\widehat{\beta}$.

If the evaluation sample exhibits heteroskedasticity or serial dependence, the i.i.d. variance estimators in equations (29) and (31) can be replaced by heteroskedasticity- or autocorrelation-consistent (HAC) estimators applied to the sequences $\{c_{ij}\}$ or $\{c_{iG}\}$. The observation-level representation equation (28) remains valid.

C Multinomial Classification

This appendix shows that the attribution framework developed in the main text extends mechanically from binary to multinomial classification models. The extension introduces no new attribution principle. Instead, it makes explicit how the same path-integral and Euler logic applies to softmax models once the geometry of the multinomial log loss is written out carefully.

The main text follows convention and defines explained fit as the improvement in expected log loss relative to a baseline. However, the improvement in log loss also has an exact quadratic representation along the score path. That representation exposes the geometry underlying the Euler decomposition and allows natural generalization to the multinomial case.

C.1 Binary Classification: Exact Quadratic Representation

We begin with binary outcomes $y \in \{0, 1\}$ and probabilistic predictions $\widehat{p} = \sigma(s)$, where $s \in \mathbb{R}$ denotes the fitted score (log odds). Define the log score

$$f(y, s) = ys - \psi(s), \qquad \psi(s) = \log(1 + e^s), \tag{33}$$

so that the log loss satisfies $\ell(y, \widehat{p}) = -f(y, s)$.

Let $\overline{p} = \mathbb{E}[y]$ denote the unconditional class probability and define the corresponding baseline score

$$s_0 = \log\left(\frac{\overline{p}}{1 - \overline{p}}\right). \tag{34}$$

By construction,

$$\psi'(s_0) = \sigma(s_0) = \overline{p} = \mathbb{E}[y]. \tag{35}$$

The improvement in expected log loss relative to the baseline is

$$\begin{aligned} \Delta\mathcal{L} &= \mathbb{E}\big[f(y, s) - f(y, s_0)\big] && (36)\\ &= \mathbb{E}\big[y(s - s_0)\big] - \big(\psi(s) - \psi(s_0)\big). && (37) \end{aligned}$$

Using $\mathbb{E}[y] = \psi'(s_0)$, this can be written exactly as

$$\Delta\mathcal{L} = \psi(s) - \psi(s_0) - \psi'(s_0)(s - s_0). \tag{38}$$

To express this quantity in quadratic form, define the straight-line path in score space

$$s(t) = s_0 + t(s - s_0), \qquad t \in [0, 1]. \tag{39}$$

Taylor's theorem with integral remainder yields the exact identity

$$\psi(s) - \psi(s_0) - \psi'(s_0)(s - s_0) = \int_0^1 (1 - t)\,(s - s_0)^2\,\psi''\big(s(t)\big)\,dt. \tag{40}$$

For the Bernoulli log-partition function,

$$\psi''(s) = \sigma(s)\big(1 - \sigma(s)\big). \tag{41}$$

Define the path-integrated curvature weight

$$W_{path} = 2 \int_0^1 (1-t)\, \sigma\big(s(t)\big)\big(1-\sigma\big(s(t)\big)\big)\, dt. \tag{42}$$

With this definition, the improvement in expected log loss admits the exact quadratic representation

$$\Delta\mathcal{L} = \frac{1}{2}(s-s_0)^2 W_{path}. \tag{43}$$

This representation shows that explained fit in binary classification is a homogeneous quadratic function of the score displacement $s - s_0$, with curvature determined by a path-averaged second derivative of the log-partition function.

C.2 Multinomial Logit Model

Now consider a classification problem with $K > 2$ classes. Let $\widehat{p}_k(x)$ denote the predicted probability of class k, and choose class K as a reference category. The quadratic representation allows a direct extension from one dimension to $K-1$ dimensions using standard vector notation.

Define the $(K-1)$-dimensional logit vector

$$s(x) = \begin{pmatrix} s_1(x) \\ \vdots \\ s_{K-1}(x) \end{pmatrix}, \qquad s_k(x) = \log \frac{\widehat{p}_k(x)}{\widehat{p}_K(x)}. \tag{44}$$

The binary case corresponds to $K = 2$, in which case $s(x)$ is scalar.

Let s_0 denote the baseline logit vector corresponding to the baseline probability vector $\overline{p}$.

C.3 Multinomial Log Loss and Its Derivatives

For a single observation $y \in \{1, \ldots, K\}$, the multinomial log score is

$$f(y,s) = s_y - \psi(s), \qquad \psi(s) = \log\left(1 + \sum_{k=1}^{K-1} e^{s_k}\right), \tag{45}$$

where class K is absorbed into the normalization.

The improvement in expected log loss relative to the baseline is

$$\Delta\mathcal{L} = \mathbb{E}\big[f(y,s) - f(y,s_0)\big] = \psi(s) - \psi(s_0) - \nabla\psi(s_0)^\top (s - s_0), \tag{46}$$

where $\nabla\psi(s_0)$ equals the baseline class-probability vector $\overline{p}$ (excluding the reference category).

The straight-line path in logit space is

$$s(t) = s_0 + t(s - s_0), \qquad t \in [0, 1]. \tag{47}$$

Applying Taylor's theorem with integral remainder in $\mathbb{R}^{K-1}$ yields the exact identity

$$\Delta\mathcal{L} = \int_0^1 (1 - t)\,(s - s_0)^\top \nabla^2\psi(s(t))\,(s - s_0)\,dt. \tag{48}$$

C.4 Fisher Information Interpretation

The Hessian of the multinomial log-partition function is

$$\nabla^2\psi(s) = \mathrm{diag}(p(s)) - p(s)p(s)^\top, \tag{49}$$

where $p(s) \in \mathbb{R}^{K-1}$ denotes the vector of predicted probabilities for the non-reference classes implied by s. This matrix equals the Fisher information matrix of the multinomial model evaluated at $p(s)$.

Define the path-integrated Fisher metric

$$W_{path} = 2\int_0^1 (1 - t)\,\nabla^2\psi(s(t))\;dt. \tag{50}$$

With this definition,

$$\Delta\mathcal{L} = \frac{1}{2}(s - s_0)^\top W_{path}(s - s_0), \qquad \|s - s_0\|_{W_{path}} = \sqrt{2\Delta\mathcal{L}}. \tag{51}$$

Thus, as in the binary case, explained fit admits an exact quadratic representation in the displacement of the fitted signal from the baseline, with curvature given by a path-integrated Fisher metric.

C.5 Additive Signal Representation

Suppose the fitted logit vector admits an additive decomposition

$$s(x) - s_0 = \sum_{j=1}^{p} s^{(j)}(x), \tag{52}$$

where $s^{(j)}(x)$ denotes the contribution of feature j to the fitted signal.

For linear multinomial logit models this decomposition is literal. For nonlinear models, it may be obtained via a path-integral representation of the Jacobian of $s(x)$, as discussed in the main text.

No assumption beyond additivity of the fitted signal is required; the decomposition concerns the fitted object itself and does not depend on the estimation procedure.

C.6 Euler Decomposition of Multinomial Signal Strength

Because $\|s - s_0\|_{W_{path}}$ is positively homogeneous of degree one in $s - s_0$, Euler's theorem yields the exact decomposition

$$\|s - s_0\|_{W_{path}} = \sum_{j=1}^{p} \frac{\langle s - s_0, s^{(j)} \rangle_{W_{path}}}{\|s - s_0\|_{W_{path}}}, \qquad \langle a, b \rangle_W = a^\top W b. \tag{53}$$

Each term represents the contribution of feature j to overall signal strength, measured in the geometry induced by the multinomial log-likelihood.

Thus, multinomial classification introduces no new attribution logic beyond that already present in the binary case. It replaces scalar quantities with their vector-valued analogues under the same path-integrated quadratic signal geometry.

Implementation of the multinomial case follows the same structure as the algorithm in appendix D, replacing scalar quantities with their vector-valued analogues and evaluating the path-integrated Fisher metric numerically when a closed form is unavailable.

C.7 Standard Errors

The standard-error calculations for multinomial classification follow the same logic as in the binary case. Throughout, we condition on the fitted score function $s(\cdot)$ and treat it as fixed. Inference therefore reflects sampling variability in the evaluation sample, not uncertainty due to re-estimation of the model.

For each observation i, define the observation-level contribution

$$c_{ij} = \frac{\langle s_i - s_0,\ s_i^{(j)} \rangle_{W_{path,i}}}{\|s_i - s_0\|_{W_{path,i}}}, \tag{54}$$

where $W_{path,i}$ denotes the path-integrated Fisher metric evaluated at observation i. The population contribution satisfies

$$C_j = \mathbb{E}[c_{ij}], \tag{55}$$

where expectations are taken over the evaluation sample.

Under i.i.d. sampling, the standard error of C_j is the standard error of a sample mean,

$$SE(C_j) = \sqrt{\frac{1}{N}\,\mathbb{E}[(c_{ij} - C_j)^2]}. \tag{56}$$

As in regression and binary classification, computing these standard errors requires only scalar variances of the observation-level contributions and does not require estimation of a full covariance matrix across features.

The same logic applies to grouped contributions obtained by summing c_{ij} across features.

D Decomposition Algorithm

This appendix contains the algorithm for computing exact Euler-like contributions to classification signal strength.

Algorithm 1: Feature importance for binary classification under log loss

```
# This algorithm computes exact Euler-style contributions under log
    loss,
# along with vanilla (i.i.d.) standard errors under the same
    evaluation weights.
#
# Inputs:
# y      : (N,) vector with y in {0,1}
# Eta_hat : (N, K) matrix of fitted score components
#          with eta = eta_bar + sum_j Eta_hat[:, j]
# w      : optional (N,) nonnegative observation weights defining the
#          evaluation metric (default: uniform weights)
#
# Notes:
# - For logistic regression, Eta_hat[:,j] = X[:,j] * beta[j].
# - Weights define the empirical measure used to evaluate fit
    (analogous to
#  WLS/GLS metrics), not ad hoc class rebalancing. All expectations
    below are
#  computed under these weights.
# - The baseline p_bar is the best constant predictor under the same
    weights.
# - For non-uniform weights, standard errors use the effective sample
    size
#  N_eff = 1 / sum_i wtil_i^2, where wtil are normalized weights.
```

```
# Helper functions
sigmoid(z) = 1 / (1 + exp(-z))
softplus(z) = log(1 + exp(z)) # implement stably if needed
logloss(y,p)= - y*log(p) - (1-y)*log(1-p)

# Dimensions
N, K = Eta_hat.shape

# Weighted mean helper (normalizes weights)
if w is None:
    w = ones(N)
w_sum = sum(w)
wtil = w / w_sum
wmean(a) = sum(wtil * a)

# Effective sample size under normalized weights
N_eff = 1.0 / sum(wtil**2)

# Baseline probability and baseline score (intercept-only, weighted)
p_bar  = wmean(y)
eta_bar = log(p_bar / (1 - p_bar))

# Aggregate fitted score and fitted probabilities
eta = eta_bar + Eta_hat.sum(axis=1)
p   = sigmoid(eta)

# Baseline and fitted log loss (fit improvement, weighted)
L_bar = wmean(logloss(y, p_bar))
L_hat = wmean(logloss(y, p))
DeltaL = L_bar - L_hat

# Closed-form path weight under log loss:
#   w_i^{path} = y_i - ( softplus(eta_i) - softplus(eta_bar) ) /
    delta_i,
# with the limit w_i^{path} -> y_i - sigmoid(eta_bar) as delta_i -> 0.
delta     = eta - eta_bar
sp_eta    = softplus(eta)     # (N,)
sp_eta_bar = softplus(eta_bar) # scalar

eps = 1e-12
w_path = zeros(N)
mask = abs(delta) > eps
```

```
w_path[mask] = y[mask] - (sp_eta[mask] - sp_eta_bar) / delta[mask]
w_path[~mask]= y[~mask] - sigmoid(eta_bar)

# Observation-level contributions and component contributions
    (weighted)
# c_ij = Eta_hat[i,j] * w_path[i], C[j] = E_w[c_ij]
c = Eta_hat * w_path[:,None]                 # (N, K)
C = sum(wtil[:,None] * c, axis=0)            # (K,)

# Plain (i.i.d.) standard errors for contributions under the same
    weights
# Var_w(c_j) = E_w[(c_ij - C[j])^2], SE(C_j) = sqrt( Var_w(c_j) /
    N_eff )
var_w = sum(wtil[:,None] * (c - C[None,:])**2, axis=0) # (K,)
SE   = sqrt(var_w / N_eff)                       # (K,)

# Output:
# C satisfies sum_j C[j] = DeltaL exactly (up to floating-point error)
# SE are vanilla standard errors under the evaluation weights
# Proportional importance: C / sum(C) = C / DeltaL
```