PREDICTION WITH TRANSPARENCY: OFFENSIVE VALUE IN BASEBALL

THIS VERSION NOVEMBER 5, 2025

Megan Czasonis is a Founding Partner of Cambridge Sports Analytics in Cambridge, MA.
245 Main Street, Cambridge MA, 02142
mczasonis@csanalytics.io

Miles Kee is a data scientist at Cambridge Sports Analytics in Cambridge, MA.
245 Main Street, Cambridge MA, 02142
mkee@csanalytics.io

Mark Kritzman is a Founding Partner of Cambridge Sports Analytics in Cambridge, MA, and a senior lecturer at the MIT Sloan School of Management in Cambridge, MA.
245 Main Street, Cambridge MA, 02142
mkritzman@csanalytics.io

David Turkington is a Founding Partner of Cambridge Sports Analytics in Cambridge, MA.
245 Main Street, Cambridge MA, 02142
dturkington@csanalytics.io

Abstract

Accurate and interpretable prediction of player performance requires analytic methods that account for contextual player-specific information and give visibility into the influence of observations and variables on the formation of the predictions. We describe a novel model-free prediction system called relevance-based prediction (RBP) that addresses these needs, and we show how it enables prediction-specific interpretability that is beyond the reach of model-based approaches such as linear regression analysis or machine learning models. RBP reveals the specific reliability of each prediction before the prediction is made, the importance of each prior player to each prediction, the contribution of each predictive variable to each prediction's value, and the contribution of each predictive variable to each prediction's reliability. We illustrate this new prediction system by applying it to predict wRC+ for major league baseball players. The prediction-specific information given by RBP stands in contrast to R-squared, beta, and t-statistics, which only give information about average effects, as we illustrate with specific player examples.

PREDICTION WITH TRANSPARENCY: OFFENSIVE VALUE IN BASEBALL

Baseball analytics has made substantial progress in quantifying offensive performance using composite statistics. For example, metrics such as wRC+ (weighted runs created plus) provide a holistic and league-normalized assessment of a player's hitting value. However, despite the availability of sophisticated performance metrics, the task of predicting future performance remains a significant challenge. Traditional linear models treat predictive variables as independent factors, thereby failing to capture their interaction with player traits and circumstances. Complex machine learning models capture nonlinear and conditional effects, but they lack transparency and often learn spurious patterns that undermine the quality of their predictions. To translate complex data into trustworthy predictions, it is necessary to account for conditional, context-dependent effects in a way that is fully transparent.

We describe a model-free prediction system called relevance-based prediction (RBP) which forms transparent predictions that adapt to the unique circumstances of each prediction task. RBP forms a prediction as a weighted average of observed outcomes in which the weights are based on a rigorously defined and theoretically justified statistic called relevance. Unlike predictive models such as linear regression analysis or machine learning models, which work by estimating model parameters and then applying those parameters to new tasks, RBP operates by evaluating patterns in the relationship between outcomes and predictive variables given the specific circumstances of each prediction task.

Like machine learning models, RBP captures nonlinear effects and conditionality, but unlike machine learning models, RBP's model-free approach to prediction gives remarkable visibility into the formation of each individual prediction:

- It reveals the reliability of each prediction before the prediction is made.
- It reveals precisely how each player informs each prediction.
- It shows how each predictive variable contributes to each prediction's value.
- It shows how each predictive variable contributes to each prediction's reliability.

The prediction-specific information given by RBP, as we illustrate in chosen examples, stands in stark contrast to the summary statistics given by a model. For example, linear regression analysis provides one R-squared, one set of beta coefficients, and one set of t-statistics for a calibrated model. None of these statistics distinguish among the circumstances of different prediction tasks. By contrast, RBP's assessment of reliability, a variable's contribution to a prediction's value, and a variable's contribution to a prediction's reliability, reflect the player-specific context of each prediction task.

We proceed as follows. We first describe the three key features of RBP: relevance, fit, and grid prediction.[1] We then describe how RBP measures the influence of the predictive variables on a prediction's value and on its reliability. Next, we apply RBP to predict wRC+ for MLB players. We describe our experiment setup, and we present results which give evidence of the transparency and efficacy of RBP. We conclude with a summary.

Relevance-Based Prediction

As described comprehensively by Czasonis, Kritzman, and Turkington (2022a, 2022b, 2023, and 2024), RBP is a model-free prediction system that forms a prediction as a relevance-weighted average of observed outcomes in which relevance has a precise statistical meaning. RBP also depends crucially on fit, which quantifies the extent to which there are useful patterns in a dataset. The final feature of RBP is grid prediction, which forms a composite prediction as a reliability-weighted average of many predictions given by different combinations of players and predictive variables.

Relevance

In the context of predicting offensive production for MLB players, relevance provides a statistical measure of the importance of a previous player to the prediction for a current player given a chosen set of predictive variables. It is composed of two components, similarity and informativeness, as shown in equation 1.

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

In equation 1, similarity and informativeness are computed as Mahalanobis distances (Mahalanobis 1936) rather than absolute distances or Euclidean distances.

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

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

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

In equations 1 through 4, x_i is a row vector of the values of the predictive variables for a previous player, x_t is a row vector of the values of the predictive variables for the current player, $\bar{x}$ is a vector of the average values of the predictive variables for all previous players in the sample, Ω^{-1} is the inverse covariance matrix of the values of the predictive variables for all previous players, and $'$ denotes matrix transpose.

The vector $(x_i - x_t)$ measures how different a previous player is from the current player, whereas the vector $(x_i - \bar{x})$ measures how different he is from average, and $(x_t - \bar{x})$ measures how different the current player is from average. By multiplying these vectors by the inverse covariance matrix, we capture the correlation of the attributes of the previous players. Also, this calculation implicitly standardizes the differences by dividing them by variance. By multiplying the product by the transpose of the vector differences we consolidate the outcome into a single number, which represents the covariance-adjusted distance between the two vectors.

Notice that in the formula for similarity we multiply the Mahalanobis distance of a previous player from the current player by negative one half. The negative sign converts a measure of difference into a measure of similarity. We multiply by one half because the average squared distances between pairs of players is twice as large as the players' average squared differences from the average of all players. When we measure informativeness, we retain its positive sign, and we need not multiply by one half. By measuring informativeness as a difference from average, we are recognizing that unusual players contain more information than typical players. Intuitively, this occurs because the outcomes for an unusual player are likely to reveal genuine relationships, whereas outcomes for highly typical players are likely to

contain more noise. Finally, note that we measure the unusualness of the current player. We do so to center our measure of relevance on zero. All else being equal, previous players who are like the current player but different from the average of all previous players are more relevant to a prediction than those who are not.

This definition of relevance is not arbitrary. We know from information theory that the information contained in an observation is the negative logarithm of its likelihood (Shannon 1948). We also know from the Central Limit Theorem that the relative likelihood of an observation from a multivariate normal distribution is proportional to the exponential of a negative Mahalanobis distance. Therefore, the information contained in a point on a multivariate normal distribution is proportional to a Mahalanobis distance.

We can also justify the non-arbitrariness of relevance by considering a limiting case of the predictions it yields. RBP forms a prediction as a weighted average of prior player outcomes for Y.

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

If we define relevance weights as follows, which admits the relevance-weighted average of every previous player outcome in the observed data sample, the result is precisely equivalent to the prediction that results from linear regression analysis.[2]

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

Owing to this equivalence, the theoretical justification given by Gauss for linear regression analysis applies as well to RBP.[3] In most cases, however, we can produce a more

reliable prediction by taking a relevance-weighted average of a subset of relevant players, especially if the relationship between the predictive variables and the outcomes is not perfectly static and symmetric. RBP censors the influence of previous players who are less relevant than a chosen threshold, which leads to the following definition of prediction weights.

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

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

$$\lambda^2 = \frac{\sigma^2_{r,full}}{\sigma^2_{r,partial}} = \frac{\frac{1}{N-1}\sum_i r_{it}^2}{\frac{1}{n-1}\sum_i \delta(r_{it})r_{it}^2} \tag{9}$$

In equations 7 through 9, $n = \sum_{i=1}^{N}(r_{it})$ is the number of players who are fully retained, $\varphi = n/N$ is the fraction of players in the retained sample, and $\bar{r}_{sub} = \frac{1}{n}\sum_{i=1}^{N}\delta(r_{it})r_{it}$ is the average relevance value of the players in the retained sample. It is important to note that $w_{it,retained}$ depends crucially on the prediction circumstances x_t. Relevance is reassessed for each prediction circumstance which further affects the identification of the retained subsample and introduces nonlinear conditional dependence of the prediction $\hat{y}_t$ on the prediction circumstances x_t . The scaling factor λ^2 compensates for a bias that would otherwise result from relying on a small subsample of highly relevant players. In the case of linear regression analysis $n = N$ and $\lambda^2 = 1$. Lastly, note that the prediction weights always sum to 1.[4]

Fit

Fit is a critical component of RBP. It quantifies the prevalence of useful patterns in a dataset which reveals how much confidence we should have in a specific prediction task, separately

from the confidence we have in the overall prediction system. Fit provides a principled way to evaluate the relative merits of alternative calibrations for each prediction task.

Consider, for example, a pair of previous players who are used, in part, to form the prediction of an outcome for a current player. Each previous player has a relevance weight and an outcome. We are interested in the alignment of the relevance weights of the two previous players with their outcomes. But we must standardize them by subtracting the average value and dividing by standard deviation – in essence, converting them to z-scores. We then measure their alignment by taking the product of the standardized values. If this product is positive, their relevance is aligned with their outcomes, and the larger the product, the stronger the alignment. We perform this calculation for every pair of previous players in our sample. We should also note that all the formulas we have thus far considered for the relevance weights rely only on the $x_i s$, the $x_t s$, and the $\bar{x} s$. They do not make use of any of the information from previous player outcomes. To determine fit, however, we must consider outcomes (the y_is).

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

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

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

Although we compute fit from the full sample of players, the weights that determine fit vary with the threshold we choose to define the relevant subsample. As we focus the subsample on players who are more relevant, we should expect the fit of the subsample to

increase, but we should also expect more noise as we shrink the number of players. The fit across pairs of all players in the full sample implicitly captures this tradeoff between subsample fit and noise by overweighting players who are more relevant and underweighting players who are less relevant accordingly.

Like relevance, fit is not arbitrary. In the case of linear regression analysis with $n = N$, the informativeness-weighted average fit across all prediction tasks in the observed sample equals R-squared.[5]

$$R^2 = \frac{1}{T-1}\sum_t info(x_t)fit_t \quad (12)$$

This convergence of fit to R-squared reveals an intriguing insight. R-squared is the result of some good predictions, some average predictions, and some bad predictions; that is, some predictions with high fit, some with average fit, and some with low fit. R-squared reveals the average reliability of a prediction model. It reveals much less about the reliability of specific prediction tasks, which can vary substantially. Fit is much more nuanced. It gauges the reliability of a specific prediction task in a non-arbitrary way, as demonstrated by its convergence to R-squared. Fit is the fundamental building block of R-squared. To compute fit, we must know the weight of each player in a prediction. These weights are inherent to RBP, but they are not available in model-based prediction algorithms which rely exclusively on calibrated parameters rather than weighted players to form predictions.

This notion of prediction-specific fit warrants particular emphasis. Because it offers advance guidance about a specific prediction's reliability, it enables teams to discard or view with greater caution predictions that are foreseen to be unreliable.

Grid Prediction

We have thus far shown how to form a prediction as a relevance-weighted average of player outcomes. And we have shown how we can use fit to measure the reliability of a specific prediction task. But we have left unanswered the question of how to determine the threshold for the subsample of relevant players. We have only noted that a prediction depends on the choice of a parameter, r^*, which is the censoring threshold for relevance. In addition, we have implicitly assumed up to this point that the full menu of predictive variables is used to measure relevance and form a prediction. However, it is possible that a subset of the predictive variables will render a better prediction for a specific prediction task. The efficacy of previous players for a given prediction task depends on the predictive variables, and the efficacy of the predictive variables depends on the players. These choices are codependent on the traits and circumstances of the current player. We, therefore, turn to the last key feature of RBP, which is grid prediction. But before we show how to form predictions that consider a range of alternative calibrations, we must first describe an enhanced version of fit called adjusted fit.

RBP is more effective to the extent there is strong alignment between relevance and outcomes, as measured by fit. It is also more effective to the extent there is asymmetry between the fit of the retained subsample of previous players and the fit of the censored players. In the presence of asymmetry, we trust the more relevant sample on principle. In the absence of asymmetry, the full sample relationship is linear, and linear regression analysis will suffice. Therefore, to compare properly the efficacy of two predictions formed from different values of r^*, we need a way to measure not only fit but asymmetry.

We measure asymmetry between the fit of the retained and censored subsamples as shown by equation 13.

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

The $(+)$ superscript designates weights formed from the retained subsample of players while the $(-)$ superscript designates weights formed from the complementary sample of censored players. Asymmetry recognizes the benefit of censoring non-relevant players that contradict the predictive relationships that exist among the relevant observations. This assessment also inherently considers the relative sample sizes of the complementary groups

To calculate adjusted fit, we add asymmetry to fit and multiply this sum by K, the number of predictive variables, as shown by equation 14.

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

Multiplication by the number of predictive variables allows us to compare predictions based on different numbers of predictive variables. It corrects a bias that would otherwise occur, whereby adding a pure noise variable decreases fit in proportion to the increase in the number of variables, even if the predictions themselves do not change (consider, for example, the case of a full sample linear regression analysis with a large sample of players). Another way to view the intuition for K is that we are more likely to observe a spurious relationship from weights based on any one variable in isolation than we are based on a collection of many variables.

We now return to the question of how to form a prediction given uncertainty in the calibration of r^* and variable selection, which are codependent choices. To address this issue, we could consider every possible calibration that combines a choice of r^* with a choice of a subset of variables and select the prediction with the greatest reliability as measured by adjusted fit. It is critical to remember that the assessment of reliability using adjusted fit is made before the prediction is rendered and the subsequent outcome is known. And it is also critical to remember that the assessment of reliability is specific to the prediction task.

Instead of selecting one optimal calibration for a given prediction task, it may be more prudent to compute a composite prediction as a reliability-weighted average of the predictions from all possible calibrations. Equation 15 defines reliability weights, ψ_θ, as the adjusted fit for a parameter calibration, θ, divided by the sum of all adjusted fits across all parameter calibrations.

$$\psi_\theta = \frac{adjusted\ fit_\theta}{\sum_{\tilde{\theta}} adjusted\ fit_{\tilde{\theta}}} \tag{15}$$

Equation 16 describes the composite prediction.

$$\hat{y}_{t,grid} = \sum_\theta \psi_\theta \hat{y}_{t,\theta} \tag{16}$$

Exhibit 1 gives a visual representation of grid prediction based on hypothetical values. The columns represent different combinations of predictive variables and the rows represent different subsamples of previous players as determined by different relevance thresholds. Each cell represents a calibration θ; that is, a unique combination of predictive variables and previous players. In practice, we would consider all 63 combinations of six variables, but for

illustrative purposes we show only seven columns in exhibit 1. The first values shown in the cells are the calibration-specific predictions $\hat{y}_t$ for a given prediction task t. The second values are the weights ψ_θ we apply to the calibration-specific predictions to form the composite prediction. The values in the grid are specific to each prediction task. This illustration gives a composite prediction of 16.30 (15.7 x 1.72% + 15.7 x 1.15% + 10.1 x 0.24% + . . . + 9.3 x 0.04%).

Exhibit 1: Grid Prediction – Illustrative Example

Variable Combinations

Player Censoring Thresholds	$X_1 X_2 X_3 X_4 X_5 X_6$		$X_1 X_2 X_3 X_4$		$X_1 X_3 X_4$		$X_2 X_5 X_6$		$X_3 X_6$		X_2		X_6	
0.0	15.7	1.72	15.7	1.15	10.1	0.24	15.3	1.37	10.9	0.54	15.3	0.47	7.4	0.06
0.1	16.4	2.02	16.7	1.39	10.4	0.23	15.4	1.88	12.5	0.73	15.5	0.50	7.7	0.04
0.2	17.5	2.20	17.4	1.43	10.3	0.18	15.4	1.91	12.6	0.64	15.5	0.44	7.9	0.05
0.3	17.8	2.17	17.7	1.43	10.5	0.20	15.5	2.24	12.6	0.62	15.5	0.42	7.9	0.05
0.4	18.2	2.29	18.0	1.50	10.6	0.22	15.4	2.18	12.7	0.65	15.5	0.41	8.1	0.07
0.5	18.6	2.50	18.2	1.58	10.7	0.25	14.3	2.50	12.8	0.70	15.3	0.41	8.1	0.06
0.6	18.7	2.47	18.4	1.61	10.7	0.23	15.4	1.21	13.1	0.73	15.4	0.42	8.8	0.10
0.7	19.0	2.47	18.8	1.63	10.7	0.19	15.4	2.20	12.9	0.62	15.4	0.41	8.7	0.07
0.8	19.4	2.32	19.1	1.50	11.5	0.20	15.3	2.04	13.7	0.57	15.5	0.37	8.6	0.04
0.9	19.5	1.26	18.8	0.81	12.9	0.22	15.5	1.73	14.0	0.32	15.3	0.25	9.3	0.04

Composite prediction: 16.30

Note that each cell's prediction is a linear function of player observations, and the grid prediction is a linear function of each cell's prediction. Therefore, we can express the grid prediction in terms of composite weights applied to each player, as shown in equation 17. Composite weights are important because they preserve the transparency of how each

previous player informs the prediction, and they allow us to calculate fit from composite weights as a final gauge of the grid prediction's reliability.

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

The prediction grid also yields a comprehensive measure of how each variable contributes to the value of a prediction. This measure is called contribution to prediction (CTP). As shown by equation 18, CTP_{tk} for prediction t and variable k is computed as the weighted average prediction value for grid cells that contain k (for which the variable censoring indicator $\Delta_k(\theta) = 1$) minus the weighted average prediction value for cells that do not contain k (for which $\Delta_k(\theta) = 0$). We express CTP_{tk} as a sum over all grid cells θ.

$$CTP_{tk} = \sum_{\theta} \alpha_{\theta} \frac{\Delta_k(\theta)(\hat{y}_{t\theta}) - (1-\Delta_k(\theta))(\hat{y}_{t\theta})}{\sum_{\tilde{\theta}} \Delta_k(\tilde{\theta})} \tag{18}$$

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

We can use the same computation method to measure a variable's contribution to the reliability of a prediction, which we call relevance-based importance (RBI).[6] We simply replace each grid cell's prediction value with its adjusted fit.

$$RBI_{tk} = \sum_{\theta} \alpha_{\theta} \frac{\Delta_k(\theta)(adjusted\ fit_{t\theta}) - (1-\Delta_k(\theta))(adjusted\ fit_{t\theta})}{\sum_{\tilde{\theta}} \Delta_k(\tilde{\theta})} \tag{19}$$

RBI has several advantages over alternative measures of variable importance. Linear regression analysis relies on t-statistics and their corresponding p-values, which only measure a variable's marginal importance. RBI, by contrast, captures a variable's total importance. RBI also captures conditional relationships which t-statistics fail to address. And unlike the Shapley value, which is the accepted standard for assessing variable importance in machine learning models, RBI accounts for the reliability of individual predictions.

A final note on grid prediction. For some prediction tasks, it may be preferable to select the subsample of players and predictive variables based on similarity rather than relevance. We need not worry whether we should use similarity or relevance to identify the optimal combination of players and variables. We simply include both censoring rules as candidates in the grid. However, even when we censor players based on similarity, we should still form the predictions as a relevance-weighted average of the retained players.

Experiment Setup

To illustrate how RBP is used to predict player outcomes, we apply it to predict wRC+ for MLB players. The performance metric wRC+ stands for weighted runs created plus. It combines a player's batting outcomes into a single measure of runs created, adjusting for league and ballpark effects to facilitate comparison across players.

Our training sample comprises 1,601 players with 200 or more plate appearances per season from 2015 through 2023. Our prediction sample, which we use to compare player results cross sectionally, comprises 265 players with 200 or more plate appearances in the 2024

season. Our prediction task is to predict each player's wRC+ for the 2025 season.

We form our predictions based on the following set of predictive variables. Appendix B gives definitions of these variables.

Exhibit 2: Predictive Variables

Contextual Factors	Hitting Statistics	Plate Discipline	Batted Ball Profile	Statcast Metrics
Batter handedness	wRC+	BB %	Pull %	xwOBA
Age	K %	K %	Oppo %	Exit velocity
Plate appearences	HR/PA	Out of zone swing %	Soft contact %	Launch angle
	WAR/162	In zone swing %	Hard contact %	
	BABIP	Out of zone contact %	Ground ball rate	
	wOBA	In zone contact %	Fly ball rate	
		Called strike + whiff rate	HR/FB	

Results

Transparency

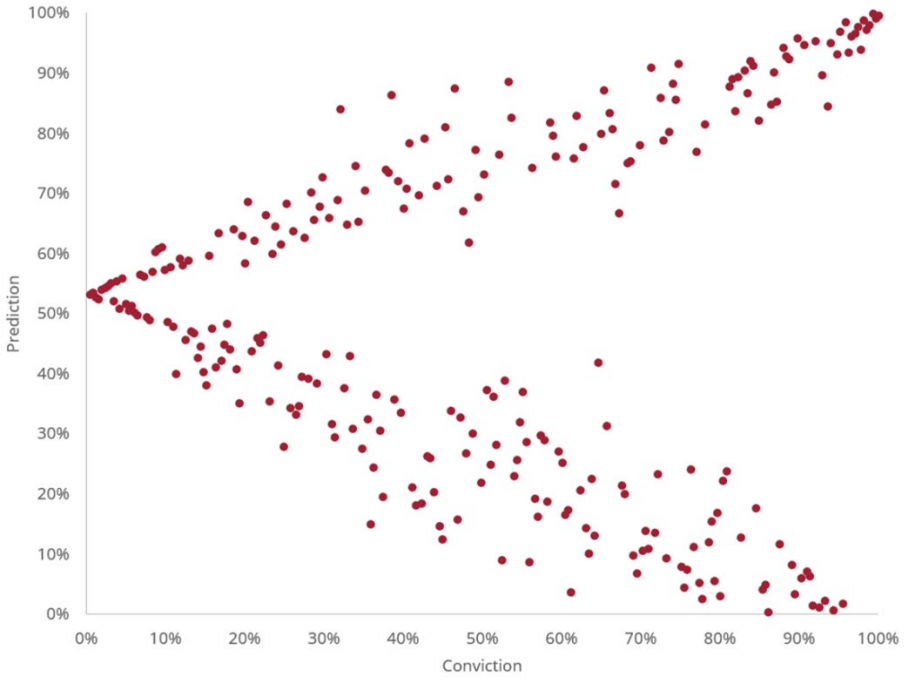
We first present several exhibits that reveal the transparency of RBP as we apply it to predict wRC+.

Exhibit 3 shows a scatter plot of predicted wRC+ for players with 200 or more plate appearances in the 2024 season, reported as cross-sectional percentile ranks on the vertical axis, and their corresponding conviction levels based on fit, also reported as cross-sectional percentile ranks, on the horizontal axis. These results yield two key insights. First, conviction varies dramatically from one prediction to the next, even for predictions at similar levels. This underscores the value of fit, which reveals the reliability of each prediction before it is made, thereby enabling teams to view more cautiously predictions that are likely to be less trustworthy. Second, larger magnitude predictions tend be based on stronger patterns leading

to higher conviction, while lower conviction predictions based on weaker patterns tend to revert to the mean.

It is important to contrast this prediction-specific detail with the assessment of conviction given by a model. A linear regression model, for example, assigns the same R-squared to all predictions, ignoring the vast differences in reliability across predictions as shown by exhibit 3.

Exhibit 3: wRC+ Predictions and Convictions for the 2025 Season



In addition to customizing each prediction task to account for a player's specific circumstances, RBP reveals precisely how each previous player informs the prediction. For example, exhibit 4 shows the 10 most important players for forming the wRC+ prediction for Juan Soto, which include Soto himself from previous seasons. The product of the relevance weights of these players, shown in the third column, and the outcomes shown in the right-most

column, along with the weights and outcomes of all the other relevant players, gives the prediction of wRC+ for Soto. Soto's predicted wRC+ of 157 ranked at the 99.6th percentile among all players for the 2025 season. At the 99.9th percentile, it also had extraordinarily high conviction. It is reaffirming to note that Soto's realized wRC+ for 2025 was 156. Not surprisingly, the players who were most relevant for forming Soto's prediction had similarly strong wRC+ outcomes for the seasons in which they were most relevant.

Because RBP gives the identity of the players who are most relevant to each prediction, it enables us to observe the characteristics of those players that explain to their relevance. Exhibit 4, for example, shows which variables best explain the relevant players' similarity to Soto. It is important to keep in mind, though, that a player's similarity also accounts for covariation across the predictive variables which is not shown here. Nonetheless, this basic information about the players who are most relevant to the formation of Soto's prediction, and which is unobtainable from a model, goes a long way in facilitating dialogue between analytics professionals, coaches, and scouts.

Exhibit 4: Most Relevant Players for Prediction of Juan Soto's wRC+

Juan Soto Mets, RF						
Prediction Percentile: 99.6%			Conviction Perentile: 99.9%			
10 Most Relevant Players	Season	Weight	Most Similar Characteristics			wRC+
Juan Soto	2022	1.791%	FB%	C+SwStr%	Z-Swing%	154
Juan Soto	2021	1.746%	C+SwStr%	wOBA	HR/FB	146
Yordan Alvarez	2022	1.731%	HR/FB	xwOBA	wOBA	170
Juan Soto	2023	1.667%	Z-Swing%	O-Contact%	BABIP	180
Yasmani Grandal	2021	1.658%	Hard%	Pull%	HR/PA	68
Aaron Judge	2022	1.546%	xwOBA	Oppo%	Z-Contact%	172
Ronald Acuna, Jr.	2023	1.482%	xwOBA	HR/PA	HR/FB	105
Mike Trout	2016	1.309%	Z-Swing%	wOBA	FB%	180
Aaron Judge	2023	1.288%	wOBA	Pull%	xOBA	218
Mike Trout	2018	1.267%	C+SwStr%	HR/FB	O-Swing%	177

Exhibit 5 presents the same information for Corbin Carroll. RBP predicted that Carroll's wRC+ would rank at the 77.7th percentile with conviction at the 62.6th percentile. RBP predicted that Carroll's wRC+ would improve from 107 to 114, but it underestimated the extent of improvement, as Carroll recorded a wRC+ of 139 in 2025. This directionally correct but less accurate prediction is not surprising, as the conviction percentile was much lower for Carroll than the conviction assigned to Soto's much more accurate prediction. Together, exhibits 4 and 5 highlight how RBP conditions each prediction on the specific attributes of the player for whom the prediction is made, which explains why each prediction is informed by different players. To emphasize this point, it is worth noting that the weights of the most relevant players for Soto's prediction are about twice as large as the 10 most relevant players for Carroll's prediction. This difference occurs because there are fewer players who are like Soto than Carroll; hence the relatively few relevant players for Soto are weighted more heavily on average than the greater number of relevant players for Carroll.

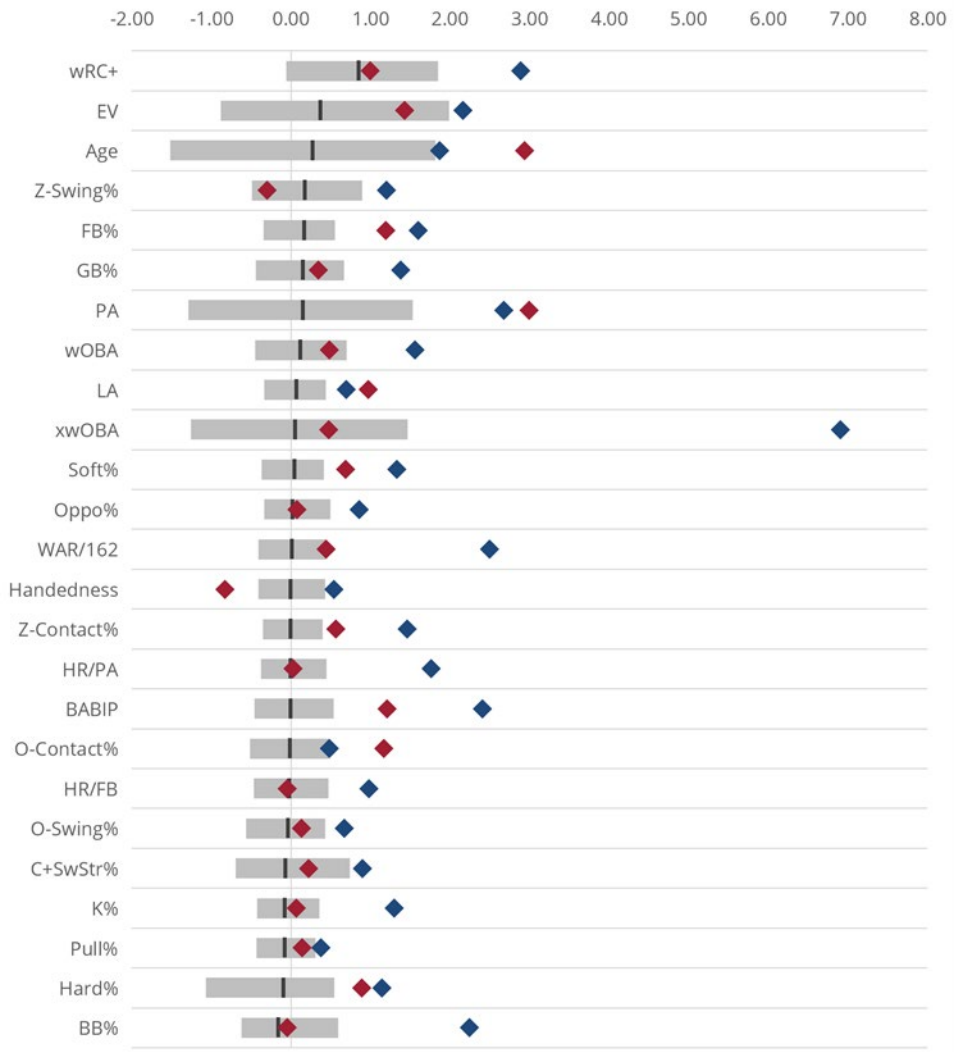
Exhibit 5: Most Relevant Players for Prediction of Corbin Carroll's wRC+

Corbin Carroll Diamondbacks, RF						
Prediction Percentile: 77.7%			Conviction Perentile: 62.6%			
10 Most Relevant Players	Season	Weight	Most Similar Characteristics			wRC+
Juan Soto	2022	0.923%	O-Contact%	WAR/162	BABIP	154
Trent Grisham	2022	0.806%	HR/PA	BB%	GB%	90
Juan Soto	2023	0.769%	K%	Z-Swing%	C+SwStr%	180
Max Kepler	2018	0.766%	HR/PA	EV	xwOBA	122
Lars Nootbaar	2022	0.732%	GB%	FB%	WAR/162	118
Mookie Betts	2017	0.691%	wRC+	BB%	O-Contact%	185
Max Kepler	2016	0.674%	Soft%	Hard%	Z-Swing%	94
Jason Heyward	2016	0.672%	Pull%	Z-Swing%	LA	89
Rowdy Tellez	2022	0.665%	wOBA	BB%	Z-Swing%	78
Alex Bregman	2018	0.637%	EV	Z-Swing%	Hr/FB	167

Thus far, we have shown that RBP reveals the specific reliability of each prediction, unlike R-squared which treats the reliability of all predictions the same. And we have shown that RBP precisely quantifies how each previous player informs a current player's prediction and why these previous players are informative. This information is unknowable for predictions that come from models. We now turn to RBP's evaluation of predictive variables.

Exhibit 6 shows the contribution of each predictive variable to the wRC+ prediction (CTP) for the 2025 season. The gray bars show the 20th to 80th percentile range of the variables' contributions across all players for whom we formed predictions. The lines within the gray bars represent contributions to the wRC+ prediction for the median player. The blue diamonds show the contributions of each predictive variable for Juan Soto's wRC+ prediction, while the red diamonds show this measure for Corbin Carroll's wRC+ prediction.

Exhibit 6: Contribution of Predictive Variables to wRC+ Predictions for Soto and Carroll Relative to All Players



The key takeaway from exhibit 6 is the variation in the contribution of the variables overall, and especially between Soto and Carroll. For example, most variables had a larger impact on Soto's wRC+ prediction than on Carroll's prediction. This difference reflects the fact that there are more useful patterns to predict Soto's outcome than there are for Carroll. In other words, the impact of each predictive variable for Carroll's prediction is mitigated by noise and is therefore less apparent. A linear regression model's beta, by contrast, would judge the

contribution of each predictive variable to be proportional to the same beta coefficient for Soto and Carroll and all other players.

Exhibit 7 shows how each predictive variable contributed to the reliability of each player's prediction of wRC+ for the 2025 season, as measured by RBI. Again, the gray bars show the 20th to 80th percentile range, the lines within the gray bars represent the median player, and the red and blue diamonds show RBI for Soto and Carroll, respectively.

Exhibit 7: Relevance-Based Importance for wRC+ Predictions for Soto and Carrol Relative to All Players

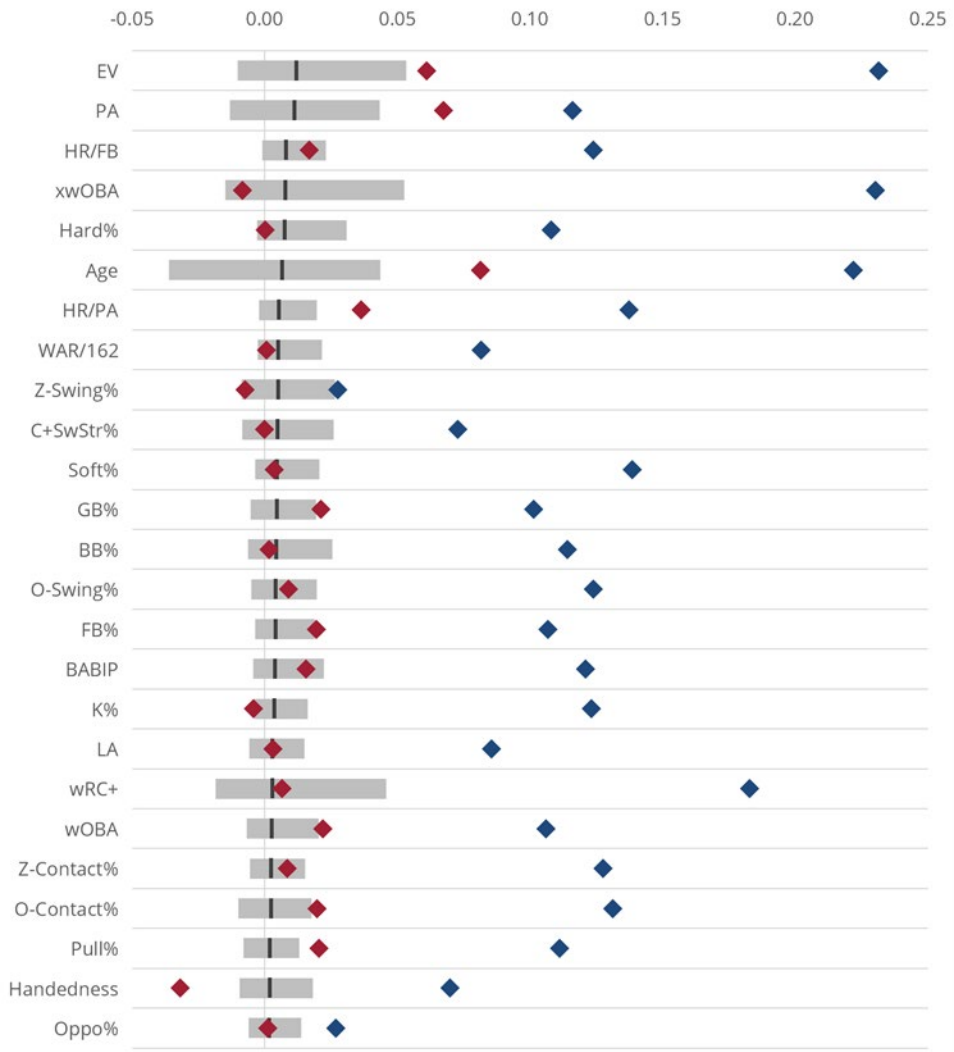


Exhibit 7 reinforces the importance of obtaining prediction-specific information. Whereas a t-statistic would assign the same average importance of a variable to all players, exhibit 7 shows that the influence of the predictive variables differs substantially across players. For example, all the predictive variables enhanced the reliability of Soto's prediction, which is unsurprising given the conviction (99.9th percentile) assigned to his prediction. But in Carroll's case, most of the predictive variables contributed very little to the reliability of Carroll's

prediction, which is consistent with the lower conviction level associated with Carroll's predicted wRC+.

Exhibit 8 compares the player-specific information given by RBP with analogous information that would come from a linear regression model. We should note here that machine learning models have even less transparency than a linear regression model.

Exhibit 8: Transparency of RBP versus Linear Regression Analysis

	RBP	Linear Regression Analysis
Observations	RBP precisely quantifies how each player informs the prediction.	Linear regression analysis offes no visibiltiy into the influence of individual players on the formation of the prediction.
Conviction	Fit measures the unique reliability of each prediction, which across all predictions, aggregates to R-squared.	R-squared only measures a model's reliability. It assigns the same level of conviction to all predictions.
Contribution to prediction value	CTP measures the contribution of each predictive variable to the value of each prediction.	Linear regression analysis judges the contribution of each predictive variable to be proportional to the same beta coefficient for all players.
Contribution to prediction reliability	RBI measures how each preditive variable uniquely contributes to the reliability of each prediction.	A t-statistic assumes that each variable is equally important to all predictions.

Efficacy

We have so far given compelling evidence that RBP offers remarkable visibility into the formation of individual predictions. Additionally, RBP measures the unique reliability of each prediction, the contribution of each predictive variable to the reliability of each prediction, and the contribution of each predictive variable to the magnitude of each prediction. RBP's ability to yield this nuanced information extends far beyond the capabilities of linear regression

analysis and other model-based approaches to prediction. We now turn our attention to RBP's predictive efficacy.

Exhibit 9 shows the realized wRC+ for players with 200 or more plate appearances in the 2024 season. The first row shows the outcomes for players who were predicted to be in the top half of all players, while the middle row shows the outcomes for players predicted to be in the bottom half of all players. The third row gives the spread between the outcomes of players who were predicted to be in the top half set against those predicted to be in the bottom half. It, therefore, serves as a measure of prediction efficacy.

The first column of results shows the efficacy of linear regression analysis. These results reveal that linear regression analysis effectively anticipated the average difference between players who delivered more favorable results from those who performed less well. The second column reports the results given by RBP across all predictions, including those predictions known in advance to be less reliable. It reveals that RBP predicted the difference between above average outcomes and below average outcomes equally as reliably as linear regression analysis. For most datasets RBP would produce a larger spread than linear regression analysis, because it would capture nonlinearities that linear regression analysis would fail to detect. The equivalence in the spreads of both approaches reveals that, in this dataset, the relationship between the predictive variables and the wRC+ outcomes is linear, which removes the opportunity for RBP to extract additional information. Nevertheless, RBP offers a critical advantage over linear regression analysis, which is highlighted in the next two columns.

The third column shows the results given by RBP for those predictions judged in advance to be the 50% most reliable predictions based on fit. And the final column shows the results for those predictions anticipated by fit to be the 50% least reliable. These two columns demonstrate that RBP can effectively separate trustworthy predictions from those that are less reliable. By contrast, linear regression analysis gives no visibility into which predictions to trust and which to view with skepticism. Given a spread of 34 for the trustworthy predictions versus only 2 for the doubtful predictions, RBP's ability to anticipate each prediction's specific reliability constitutes a huge advantage over linear regression analysis.

Exhibit 9: Realized wRC+ for Players Predicted to Outperform
Versus Players Predicted to Underperform

	Linear Regression Analysis	RBP		
		All Predictions	High Conviction Predictions	Low Conviction Predictions
High prediction	112	112	119	105
Low prediction	94	94	85	103
Spread	18	18	34	2

Summary

We described a new approach for predicting performance outcomes for MLB players called relevance-based prediction. RBP forms predictions as weighted averages of past outcomes in which the weights are based on the relevance of previous players, measured in a mathematically rigorous and theoretically justified way.

Then we described fit, which quantifies the prevalence of useful patterns in a dataset and which indicates the specific reliability of each individual prediction. R-squared, by

comparison, measures only the average reliability of a prediction model. We showed that fit converges to R-squared in the case of linear regression analysis when aggregated properly across all prediction tasks.

Next, we introduced grid prediction, which uses fit to precisely blend the predictions that result from different combinations of players and predictive variables. Crucially, the blend places greater emphasis on players and variables that are most useful for an individual prediction task.

We then illustrated RBP by predicting wRC+ (weighted runs created plus) for MLB players who had 200 or more plate appearances in the 2024 season. Our analysis highlighted the extraordinary transparency of RBP. We reported how specific players contribute to the formation of individual predictions, which is almost always unobservable for predictions generated by models. We reported the specific reliability of individual predictions in contrast to R-squared which only gives a model's average reliability. We showed the unique contribution of each predictive variable to the value of each prediction in contrast to a linear regression equation's beta which assumes a predictive variable's contribution to the value of a prediction is proportional to the same parameter value across all predictions. And we reported the contribution of each predictive variable to the reliability of individual predictions in contrast to a t-statistic which only measures a variable's average importance.

Finally, we showed that RBP successively distinguished in advance players who produced more favorable outcomes from those who produced less favorable outcomes, and we

demonstrated that, unlike linear regression analysis, RBP could distinguish in advance which predictions to trust and which to discard or treat with caution.

Appendix A: Convergence of Relevance to Other Prediction Methods

Convergence to Linear Regression Analysis

The prediction equation corresponding to full sample linear regression equals:

$$\hat{y}_t = \bar{y} + \frac{1}{N-1}\sum_{i=1}^{N} r_{it}(y_i - \bar{y}) \quad \text{(A1)}$$

Expanding the expression for relevance gives:

$$\hat{y}_t = \bar{y} + (x_t - \bar{x})\frac{1}{N-1}\sum_{i=1}^{N}\Omega^{-1}(x_i - \bar{x})'(y_i - \bar{y}) \quad \text{(A2)}$$

To streamline the arithmetic, we recast this expression using matrix notation:

$$X_d = (X - 1_N\bar{x}) \quad \text{(A3)}$$

$$\hat{y}_t = \bar{y} - \bar{x}\beta + x_t\beta - (x_t - \bar{x})(X_d'X_d)^{-1}X_d'1_N\bar{y} \quad \text{(A4)}$$

Where:

$$\beta = (X_d'X_d)^{-1}X_d'Y \quad \text{(A5)}$$

Noting that $X_d'1_N$ equals a vector of zeros, because X_d represents attribute deviations from their own respective averages, we get the familiar linear regression prediction formula:

$$\hat{y}_t = (\bar{y} - \bar{x}\beta) + x_t\beta \quad \text{(A6)}$$

$$\alpha = (\bar{y} - \bar{x}\beta) \quad \text{(A7)}$$

$$\hat{y}_t = \alpha + x_t\beta \quad \text{(A8)}$$

Relationship to Large Language Models

The key innovation that led to the success of large language models (LLMs) is the transformer, which is an information processing architecture based on attention mechanisms. Relevance is conceptually similar to attention and offers a novel interpretation of these models.

In the context of language processing, consider a sequence of words (or tokens) which is encoded as a vector, x_i. The goal is to transform each word into an enriched vector, z_i, with new dimensions, which represents a refined contextual meaning of the word within the passage.

As noted in Vaswani et al. (2017), attention in a transformer model is determined by a set of three transformation matrices: W^Q, W^K, and W^V, which compute what are commonly referred to as query, key, and value vectors from each word x_i. To highlight the link with RBP, we characterize this as follows:

$$q_t = x_t W^Q \tag{A9}$$

$$k_i = x_i W^K \tag{A10}$$

$$v_i = x_i W^V \tag{A11}$$

$$z_i = \sum_i softmax\left(\frac{q_t k_i'}{\sqrt{params}}\right) v_i \tag{A12}$$

We may intuitively think of v_i as representing the learned unconditional meaning of each word in the passage. These values represent the dependent variable, and we want to predict the contextual meaning as a weighted average of v_i for all words in the passage based on their relevance to x_i. We may express:

$$q_t k_i' = x_t W^Q W^K x_i' \quad \text{(A13)}$$

Equation A13 matches the definition of relevance in Equation 1 from earlier, if we assume $\bar{x} = 0$ and we have $W^Q W^K$ rather than the inverse covariance matrix to relate circumstances to each other. In other words, the learned matrices $W^Q W^K$ amount to a square matrix that is used to evaluate relevance. The letters used to characterize words are mostly arbitrary (compared to meaning), so learned mappings are necessary for language interpretation, whereas for meaningfully oriented data the inverse covariance matrix is well-motivated.

The softmax function serves as a censoring function that normalizes weights to sum to one, while also requiring them to be strictly positive. Thus, the use of softmax effectively censors observations to focus on the most relevant subset, similar to partial sample regression. There are many other complexities to transformers. We do not aim to provide a thorough accounting of how these models work. We merely wish to point out the striking similarity between the essence of the attention mechanism used in these models and the principles of RBP described in this article.

Appendix B: Variable Definitions

Batter Handedness	Indicates if batter is left handed or not.
Age	Age in years of the hitter.
Plate Appearances	Plate appearances.
wRC+	Weighted runs-created plus. All-in-one hitting stat that measures total offensive production, then normalizes so league average is 100. (120=20% above league average hitter)
HR/PA	Home runs per plate appearance.
WAR/162	Wins above replacement, prorated to 162 games played.
BABIP	Batting average on balls in play.
wOBA	Weighted on-base average, a rescaled version of OPS that better indicates the relative values of each batting outcome.
BB%	Percentage of at bats ending in a walk.
K%	Percentage of at bats ending in a strikeout.
Out of zone swing %	Percentage of pitches outside the zone that are swung at.
In zone swing %	Percentage of pitches in the zone that are swung at.
Out of zone contact %	Contact rate on pitches out of the strike zone.
In zone contact %	Contact rate on pitches in the strike zone.
Called strike + whiff rate	Percentage of pitches that are called strikes or swings and misses.
Pull %	Percentage of batted balls pulled.
Oppo %	Percentage of batted balls hit to the opposite field.
Soft contact %	Percentage of batted balls with low exit velocities.
Hard contact %	Percentage of batted balls with high exit velocities.
Ground ball rate	Percentage of batted balls that are ground balls.
Fly ball rate	Percentage of batted balls that are fly balls.
HR/FB	Home runs per fly ball.
xwOBA	Expected wOBA based on exit velocity and launch angle.
Exit velocity	Average exit velocity of batted balls.
Launch angle	Average launch angle of batted balls.

References

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

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

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

[4] Czasonis, Megan, Mark Kritzman, and David Turkington. 2024. “The Virtue of Transparency: How to Maximize the Utility of Data Without Overfitting.” *The Journal of Financial Data Science*, 7 (2).

[5] Czasonis, Megan, Mark Kritzman, and David Turkington. 2025a. “Prediction with Incomplete Information.” *The Journal of Financial Data Science*, 7 (4).

[6] Czasonis, Megan, Mark Kritzman, and David Turkington. 2025b. “Relevance-Based Importance: A Comprehensive Measure of Variable Importance in Prediction.” *The Journal of Portfolio Management*, 51 (9).

[7] Czasonis, Megan, Mark Kritzman, and David Turkington. 2025c. “A Transparent Alternative to Neural Networks with an Application to Predicting Volatility.” *The Journal of Investment Management*, 23 (3).

[8] Mahalanobis, Prasanta Chandra. 1936. “On the Generalised Distance in Statistics.” *Proceedings of the National Institute of Sciences of India* 2, no. 1: 49–55.

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

[10] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention is All you Need.” In *Advances in Neural Information Processing Systems* (vol. 30).

[1] The descriptions of the features of RBP and variable importance follow closely language used from Czasonis, Kritzman, and Turkington (2022a, 2022b, 2023, 2024, 2025a, 2025b, and 2025c), but they are modified to fit the context of the current discussion.

[2] See Appendix A for proof of this result.

[3] We also show in the Appendix that our definition of relevance aligns with the key breakthrough that enables large language models such as ChatGPT.

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

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

[6] See Czasonis, Kritzman, and Turkington (2025a) for a thorough description of relevance-based importance.