How to Predict the Performance of NBA Draft Prospects

Paper Track: Basketball
Paper ID: 193977

1. Introduction

We propose a new mathematical system for predicting outcomes of NBA draft prospects based on the statistical concept of relevance. This new approach to prediction identifies the combinations of previously drafted players and predictive variables that are most informative for each individual prediction task. Additionally, this method provides a measure of fit for each prediction, which enables teams to assess in advance the unique reliability of each individual prediction task, thereby offering guidance about how committed they should be to a draft prospect. Relevance-based prediction addresses hidden complexities that are beyond the reach of conventional prediction models, but in a way that is more transparent, more flexible, and more theoretically justified than widely used machine learning algorithms.

We proceed by first describing the key tenets of relevance-based prediction conceptually and mathematically. We then compare relevance-based prediction to linear regression analysis and machine learning. Next, we illustrate our relevance-based prediction system by showing how it would have predicted VORP (value over replacement player)[1] and total minutes played of NBA players during their rookie seasons, based on certain attributes of these players and their pre-NBA basketball performance, as well as attributes and performance of NBA players who came before them, and we compare our predictions to the draft prospects' actual outcomes in the NBA. We conclude with a summary.

2. Relevance-Based Prediction

Relevance-based prediction rests on three key tenets: relevance, which measures the importance of a previously drafted player to a prediction; fit, which measures the reliability of each individual prediction task; and codependence, which is the notion that the efficacy of previously drafted players for a given prediction task depends on the selected predictive variables, and the efficacy of predictive variables depends on the selected players.

2.1. Relevance

Relevance has three components: the similarity of a previously drafted player to the draft prospect, the informativeness of the previously drafted player, and the informativeness of the draft prospect, as shown in Equation 1.

[1] VORP (value over replacement player) is an estimate of the points per 100 team possessions a player scores over a replacement player during the entire season assuming his teammates perform in line with the average of all NBA players. Replacement players are bench players and have a VORP of -2.

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

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

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

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

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

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

The vector $(x_i - x_t)$ measures how different a previously drafted player is from the draft prospect, whereas the vector $(x_i - \bar{x})$ measures how different he is from average, and $(x_t - \bar{x})$ measures how different the draft prospect is from average. By multiplying these vectors by the inverse of the covariance matrix, we capture the correlation of the attributes of the previously drafted players. Also, this calculation implicitly standardizes the differences by dividing them by variance. By multiplying the product by the transpose of the vector 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 previously drafted player from the draft prospect 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 underlying relationships to his personal attributes and circumstances, whereas outcomes for highly typical players are likely to contain more noise and less information. Finally, note that we measure the unusualness of the draft prospect. We do so to center our measure of relevance on zero. All else being equal, previously drafted players who are like the draft prospect but different from the average of all previously drafted players are more relevant to a prediction than those who are not.

This definition of relevance is not arbitrary. We know from the Central Limit Theorem that aggregations of independent events generally tend toward a normal distribution, which establishes the fundamental importance of the normal distribution in statistics and explains its prevalence in empirical data. We also know that the relative likelihood of an observation from a multivariate normal distribution is proportional to the exponential of a negative Mahalanobis distance:

$$\text{likelihood}(x_i) \propto e^{-\frac{1}{2}(x_i-\bar{x})\Omega^{-1}(x_i-\bar{x})'} \tag{5}$$

Additionally, we know from foundational principles of information theory (Shannon 1948) that the information contained in an observation is the negative logarithm of its likelihood. Therefore, the information contained in a point on a multivariate normal distribution is proportional to a Mahalanobis distance.

$$\text{information}(x_i) \propto (x_i - \bar{x})\Omega^{-1}(x_i - \bar{x})' \tag{6}$$

We can also justify the non-arbitrariness of relevance in the following sense. As we show in the Appendix, a relevance weighted average of outcomes for the full sample of previously drafted players yields a prediction that is precisely equivalent to the prediction that would result from a linear regression equation. Therefore, the theoretical justification of linear regression analysis applies as well to relevance-based prediction. We also show in the Appendix that our definition of relevance aligns with the key breakthrough that enables large language models such as ChatGPT.

The equivalence of relevance and linear regression reveals an intriguing insight. A linear regression equation places as much emphasis on non-relevant previously drafted players as it does on relevant players; it just flips the sign of how a non-relevant player informs the prediction. Relevance-based prediction, by contrast, forms a prediction as a relevance-weighted average of a subsample of relevant players. This approach to prediction is called partial sample regression. The weights that are used to form the prediction in partial sample regression are given by Equation 7.

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

In Equation 7, $\delta(r_{it})$ is a censoring function that equals 1 if $r_{it} \geq r^*$ and 0 otherwise, in which r^* is the threshold for relevance. For notational concision we write the number of players for which $\delta(r_{it}) = 1$ as $n = \sum_i \delta(r_{it})$ and the proportion of all players for which $\delta(r_{it}) = 1$ as:

$$\varphi = \frac{n}{N} \tag{8}$$

In addition, we write the subsample average of relevance over the retained players as:

$$\bar{r}_{sub} = \frac{1}{n}\sum_i \delta(r_{it})r_{it} \tag{9}$$

Finally, we include a term:

$$\lambda^2 = \frac{\sigma^2_{r,full}}{\sigma^2_{r,partial}} = \frac{\frac{1}{N-1}\sum_i r_{it}^2}{\frac{1}{n-1}\sum_i \delta(r_{it}) r_{it}^2} \tag{10}$$

To the extent it differs from 1, λ^2 compensates for a bias that would otherwise arise from focusing on a small subsample of highly relevant players. The partial sample regression prediction is given by a weighted average of prior outcomes, y_i:

$$\hat{y}_{t,psr} = \sum_i w_{it,psr} y_i \tag{11}$$

We now turn to the second key tenet of relevance-based prediction, fit.

2.2. Fit

Fit is a critical component of relevance-based prediction. It reveals how much confidence we should have in a specific prediction task, separately from the confidence we have in the overall prediction system. In addition, it enables us to increase our prediction's reliance on information from the combinations of predictive variables and previously drafted players that are most informative for each prediction task.

Consider, for example, a pair of previously drafted players who are used, in part, to form the prediction of an outcome for a draft prospect. Each previously drafted player has a relevance weight and an outcome. We are interested in the alignment of the relevance weights of the two previously drafted 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 previously drafted 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 observed player outcomes. To determine fit, however, we must consider outcomes (the y_is). We express fit as a pairwise sum that involves the relevance of weights and outcomes for both players in all pairs.

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j r(w_{it}, w_{jt}) r(y_i, y_j) \tag{12}$$

In Equation 12:

$$r(w_{it}, w_{jt}) = \frac{(w_{it} - \bar{w})(w_{jt} - \bar{w})}{\sigma_w^2} \tag{13}$$

$$r(y_i, y_j) = \frac{(y_i - \bar{y})(y_j - \bar{y})}{\sigma_y^2} \tag{14}$$

From Equations, 12, 13, and 14, we can restate fit in terms of normalized z-scores as shown in Equation 15 or as a squared correlation as in Equation 16:

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j z_{w_{it}} z_{w_{jt}} z_{y_i} z_{y_j} \qquad (15)$$

$$fit_t = \rho(w_t, y)^2 \qquad (16)$$

Equation 16 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.

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. The informativeness-weighted average fit across all prediction tasks in a sample equals the classical R-squared statistic in the case of full sample linear regression (Czasonis, Kritzman, and Turkington 2022):

$$R^2 = \frac{1}{T-1} \sum_t info(x_t) fit_t \qquad (17)$$

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 a specific prediction task, 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 observation in a prediction. These weights are inherent to relevance-based prediction, but they are not available in model-based prediction algorithms which rely exclusively on calibrated parameters rather than weighted observations to form predictions.

This notion of prediction-specific fit warrants particular emphasis. Because it offers advance guidance of a specific prediction's reliability, it enables analysts to discard or view with greater caution predictions that are foreseen to be unreliable. As we show later, this feature further enables analysts to form predictions that are significantly more reliable than those generated by linear regression analysis.

2.3. Codependence

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 partial sample regression 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 partial sample 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 previously drafted 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. We, therefore, turn to the third key feature of relevance-based prediction, which is codependence. 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.

Partial sample prediction using relevance 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 previously drafted 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 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 18. The plus superscript designates weights formed from the retained subsample of players while the negative superscript designates weights formed from the complementary sample of censored players. Asymmetry recognizes the benefit of censoring non-relevant observations that contradict the predictive relationships that exist among the relevant observations.

$$asymmetry_t = \frac{1}{2}\left(\rho\left(w_t^{(+)}, y\right) - \rho\left(w_t^{(-)}, y\right)\right)^2 \tag{18}$$

To calculate adjusted fit, we add asymmetry to fit and multiply this sum by K, the number of predictive variables, as shown by Equation 19. 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 observations). 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.

$$adjusted\ fit_t = K(fit_t + asymmetry_t) \tag{19}$$

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, CKT regression (Czasonis, Kritzman, and Turkington 2023) considers every possible calibration that combines a choice of r^* with a choice of a subset of variables, and it selects the prediction with the greatest expected reliability. 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.

In this paper, we introduce a further extension to CKT regression. Instead of selecting one optimal calibration for a given prediction task, we compute a composite prediction as a reliability-weighted average of the predictions from all possible calibrations. Equation 20 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.

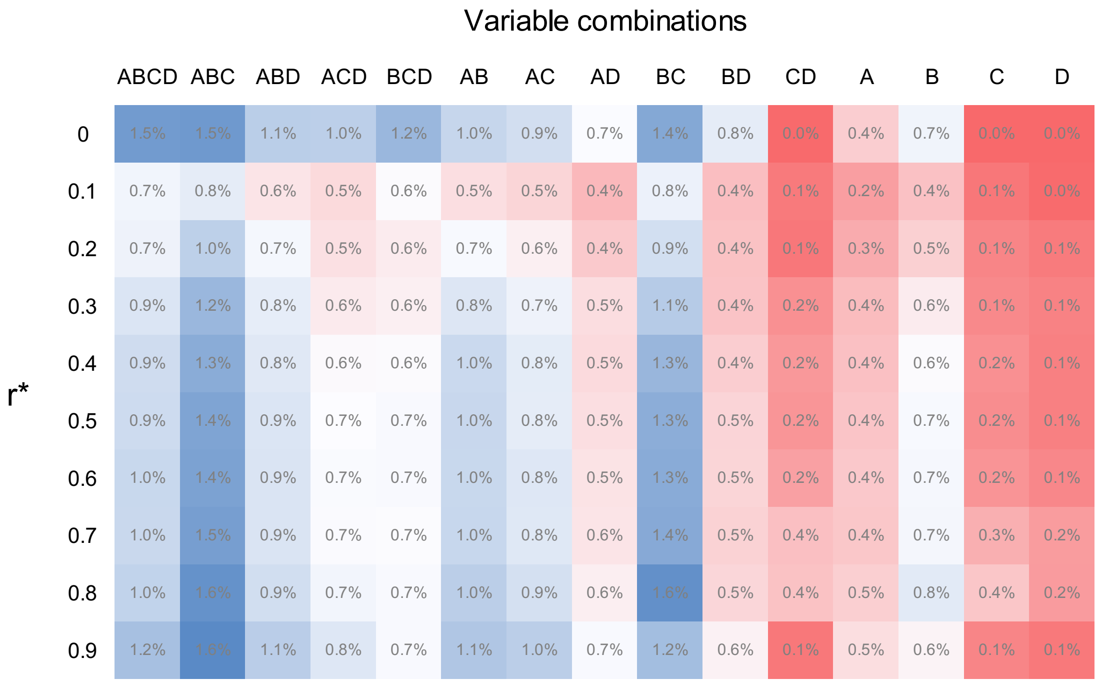
$$\psi_\theta = \frac{adjusted\ fit_\theta}{\sum_{\tilde{\theta}} adjusted\ fit_{\tilde{\theta}}} \tag{20}$$

Equation 21 describes the composite prediction.

$$\hat{y}_{t,grid} = \sum_\theta \psi_\theta \hat{y}_{t,\theta} \tag{21}$$

Figure 1 gives a visual representation of CKT grid prediction, based on a contrived data set of four predictive variables and 400 randomly simulated observations. The column labels represent alternative variable subsets, and the row labels represent alternative previously drafted player subsets. Each cell represents a codependent calibration θ; that is, a unique combination of predictive variables and previously drafted players. The values in the cells are the weights (ψ_θ) we apply to the calibration-specific predictions to form the overall CKT grid prediction. Cells that are shades of red are less important to forming the prediction while blue shaded cells are more important. This grid prediction method is not computationally trivial. In our subsequent empirical illustrations, we consider four thresholds for player sample sizes, 14 predictive variables, and two censoring criteria (which we soon describe). Because we consider all combinations of predictive variables, we evaluate 131,064 separate calibrations for each player prediction.

Figure 1: CKT Grid Prediction – Toy Example



r*	ABCD	ABC	ABD	ACD	BCD	AB	AC	AD	BC	BD	CD	A	B	C	D
0	1.5%	1.5%	1.1%	1.0%	1.2%	1.0%	0.9%	0.7%	1.4%	0.8%	0.0%	0.4%	0.7%	0.0%	0.0%
0.1	0.7%	0.8%	0.6%	0.5%	0.6%	0.5%	0.5%	0.4%	0.8%	0.4%	0.1%	0.2%	0.4%	0.1%	0.0%
0.2	0.7%	1.0%	0.7%	0.5%	0.6%	0.7%	0.6%	0.4%	0.9%	0.4%	0.1%	0.3%	0.5%	0.1%	0.1%
0.3	0.9%	1.2%	0.8%	0.6%	0.6%	0.8%	0.7%	0.5%	1.1%	0.4%	0.2%	0.4%	0.6%	0.1%	0.1%
0.4	0.9%	1.3%	0.8%	0.6%	0.6%	1.0%	0.8%	0.5%	1.3%	0.4%	0.2%	0.4%	0.6%	0.2%	0.1%
0.5	0.9%	1.4%	0.9%	0.7%	0.7%	1.0%	0.8%	0.5%	1.3%	0.5%	0.2%	0.4%	0.7%	0.2%	0.1%
0.6	1.0%	1.4%	0.9%	0.7%	0.7%	1.0%	0.8%	0.5%	1.3%	0.5%	0.2%	0.4%	0.7%	0.2%	0.1%
0.7	1.0%	1.5%	0.9%	0.7%	0.7%	1.0%	0.8%	0.6%	1.4%	0.5%	0.4%	0.4%	0.7%	0.3%	0.2%
0.8	1.0%	1.6%	0.9%	0.7%	0.7%	1.0%	0.9%	0.6%	1.6%	0.5%	0.4%	0.5%	0.8%	0.4%	0.2%
0.9	1.2%	1.6%	1.1%	0.8%	0.7%	1.1%	1.0%	0.7%	1.2%	0.6%	0.1%	0.5%	0.6%	0.1%	0.1%

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 observation, as shown in Equation 22. Composite weights are important because they preserve the transparency of how each previously drafted player contributes to the current prediction task, and they allow us to calculate fit from composite weights as a final gauge of the grid prediction's reliability.

$$w_{it,grid} = \sum_{\theta} \psi_{\theta} w_{it,\theta} \tag{22}$$

One final point is worth noting about CKT 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 these observation censoring rules as candidates in the grid. However, even when we censor based on similarity, we should still form the predictions as a relevance-weighted average of the retained observations.

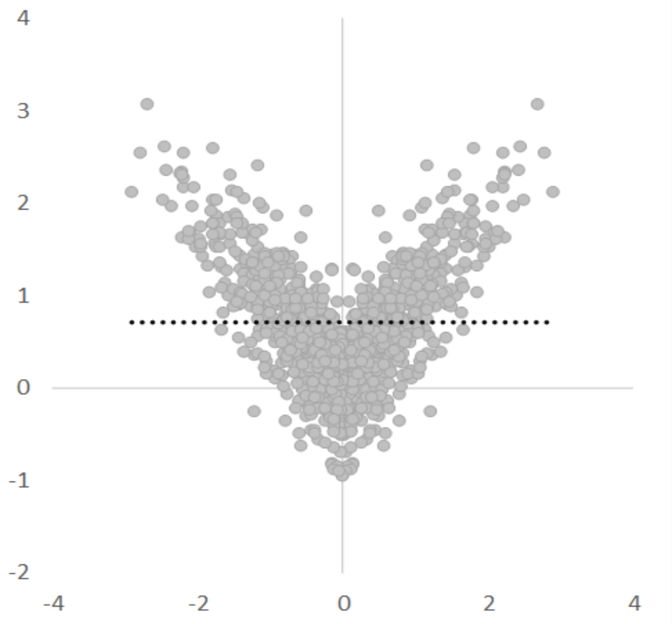
3. Relevance-Based Prediction and Linear Regression Analysis

The most common approach to statistical prediction is linear regression analysis. Linear regression analysis focuses on the use of preselected predictive variables that are weighted based on an assumed linear relationship between the values for the predictive variables and the outcomes, to give a prediction of a new outcome given a new set of values for the predictive variables. The weights that are applied to the predictive variables are derived by fitting a line through a scatter plot of values for the predictive variables and outcomes such that the sum of the squared distances of the observations from the line is minimized. Carl Friedrich Gauss, who originated this method of least squares circa 1795, proved that it gives a prediction whose expected variance from the truth is lower than any other linear and unbiased prediction.[2]

However, linear regression analysis is limited in a significant way. It assumes that the relationship between the predictive variables and the outcomes is static across all observations, or put differently, linear. Consider, for example, the scatter plot shown in Figure 2 in which there is an asymmetric relationship between the predictive variable and the outcome. Because linear regression analysis uses all the observations, the subset of observations that are positively correlated offset the subset of negatively correlated observations. Linear regression analysis cannot detect these subsample relationships and therefore gives as its prediction the average outcome for the dependent variable, at least in this contrived example. We contrived this extreme example to illustrate our point. Nonetheless, even though asymmetry is typically less extreme, it may be more subtle and more pervasive when we include more predictive variables.

[2] Gauss's original proof that the least squares line gives the best estimate of the true relationship rests on the assumption that the errors around the least squares line are normally distributed. It is now known from the Gauss-Markov Theorem that the optimality of least squares requires less strict assumptions, specifically that the errors are spherically distributed (which means they are centered on zero), are independent, and have finite variance.

Figure 2: Asymmetric Relationship between Predictive Variable (horizontal axis) and Outcomes (vertical axis)



Relevance-based prediction overcomes the inability of linear regression analysis to address asymmetry by favoring observations from the positively correlated subsample when they are relevant to the prediction task and observations from the negatively correlated subsample when they are relevant. It also favors subsamples for predictions that rely on variables with more complex asymmetries. For example, if the relationship between certain predictive variables and outcomes differs depending on whether the relationship is measured from regular season games or playoff games, relevance-based prediction will automatically capture this distinction whereas linear regression analysis will fail to detect it.

4. Relevance-Based Prediction and Machine Learning

Prior to relevance-based prediction, the standard tool for addressing asymmetry has been machine learning. It is convenient to separate machine learning algorithms into two broad categories: model-based algorithms such as lasso regression and neural networks, and model-free algorithms such as nearest neighbor and Gaussian kernels.

Model-based algorithms are essentially extensions of regression analysis, albeit very complex extensions, as some neural networks include billions of parameters. A distinguishing feature of model-based algorithms is that their parameters are determined in advance of their application and fixed for all prediction tasks. To carry out a new prediction task that depends on features of the data not originally considered by the model's parameters, one must reconstruct the model to address the specific circumstances of this new prediction task.[3] Relevance-based prediction, by contrast, automatically adapts to new prediction tasks. Rather than rely on pre-trained parameters, relevance-based prediction retains all prior observations and prioritizes the appropriate subsamples for each new prediction task. Essentially, it is a dynamic alternative to a pre-trained fixed algorithm. Relevance-based prediction also compares favorably to model-based algorithms by its transparency. It reveals precisely how each previously drafted player informs the prediction, whereas most model-based algorithms are highly opaque. Relevance-based prediction is therefore less susceptible to overfitting than model-based algorithms. And relevance-based prediction, as we have shown earlier, is theoretically grounded, whereas model-based algorithms rely on rules that are determined by trial and error.

Model-free algorithms form predictions as weighted averages of past values of the outcomes. In this sense, they serve as a bridge to relevance-based prediction. In fact, we can think of our relevance-based system as a theoretically grounded refinement to kernel regression. For example, a Gaussian kernel regression forms a prediction as a weighted average of local observations, by applying a Gaussian decay to normalized Euclidean distances to compute the weight of each player. Our relevance-based approach, by contrast, uses the Mahalanobis distance instead of the Euclidean distance to measure nearness, and, critically, it adds the component of informativeness to determine relevance. Furthermore, it combines these components in precisely the correct way, by which we mean the only way that gives the same answer as linear regression analysis when applied across the full sample.

5. Illustration of Relevance-Based Prediction for NBA Outcomes

To illustrate how relevance-based prediction is used to predict player outcomes, we apply it to predict VORP and the total number of minutes played during their NBA rookie year for players drafted in 2018, 2021, and 2022 subject to data availability. We chose to predict VORP because it reflects a variety of ways in which a player affects scoring and therefore has the potential to incorporate hidden complexities. We chose to predict total minutes played because it implicitly summarizes all the factors, as perceived by the coach, that bear upon a player's potential to impact the outcome of a game. It also allows teams to standardize how they compensate players. But mainly, we chose to predict these player outcomes because we think they are reasonable and uncontroversial. Having said that, we wish to emphasize that relevance-based prediction can be applied to predict any player outcome.

Our full data sample comprises 359 players who were drafted in the years 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2021, and 2022 from Division I U.S. colleges who played at least one season in the NBA. We excluded players from 2019 and 2020 to avoid distortions that might have occurred from the effect of COVID on both the collegiate and NBA player statistics. For each player in the 2018, 2021, and 2022 drafts and for each previously drafted player, we collect data in four categories: physical attributes, individual college performance, team performance in college, and

[3] Online learning algorithms adapt automatically to new circumstances, but only in a limited way. They adapt by applying a correction algorithm based on observed errors, but they do not re-estimate the algorithm from the full sample of previously used data. Relevance-based prediction, by contrast, retrieves all the previously used data for each new set of prediction circumstances.

team performance in the NBA. We chose these predictive variables merely to illustrate our relevance-based prediction system. We do not have expertise in determining the most effective predictive variables. For each prediction, we use training data from previously drafted players that would have been available at the time of that year's draft.

Prediction tasks:

- Rookie year VORP for 2018, 2021, and 2022 draft cohorts
- Rookie year total minutes played in the season for 2018, 2021, and 2022 draft cohorts

Training sample:

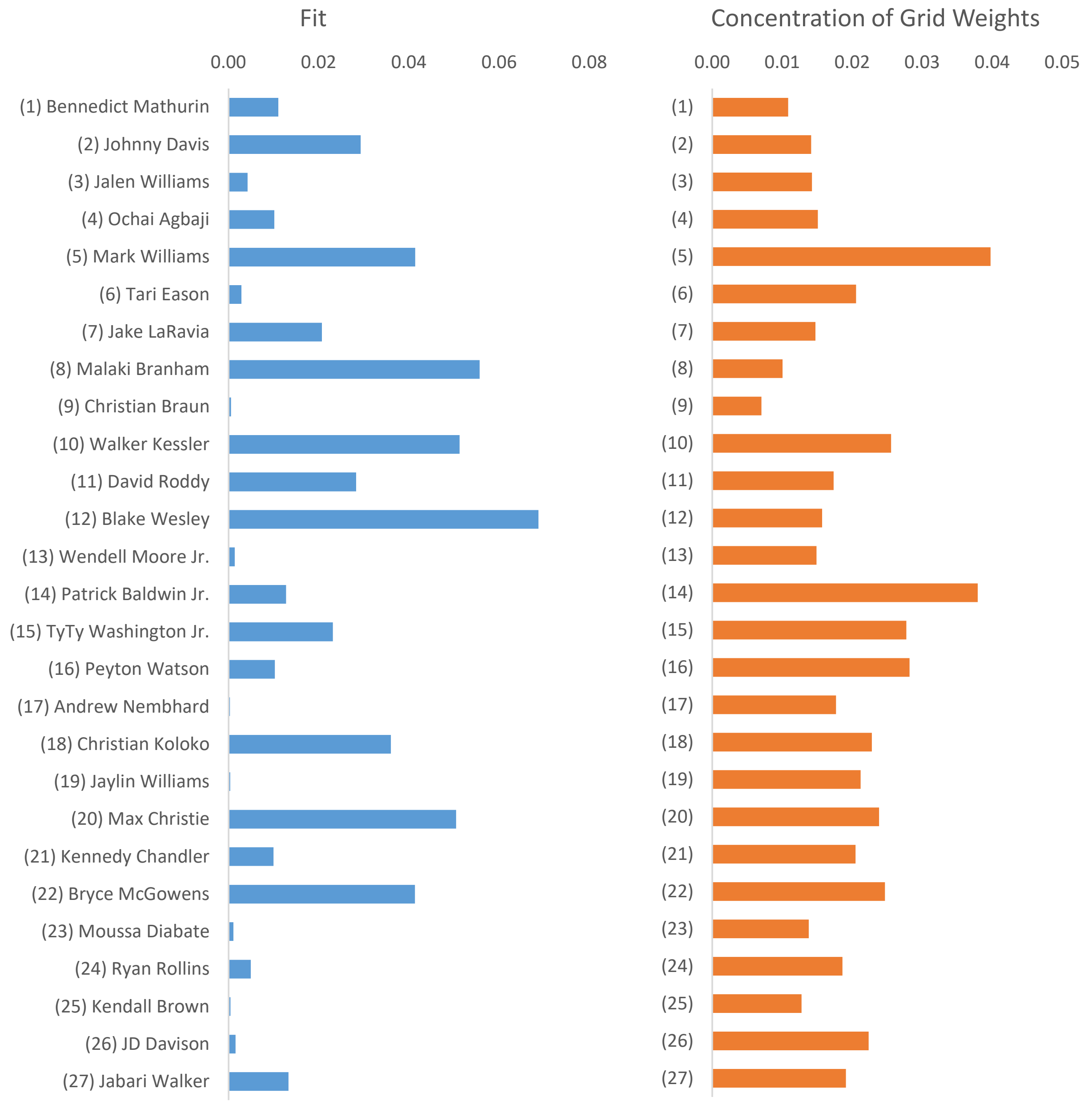
- Players drafted from 2011 through 2021 who were drafted prior to the draft class that is currently being predicted, excluding 2019 and 2020, from Division I colleges with at least one NBA season

Predictive variables:

- Physical attributes
  - Height
  - Weight
- College performance (final college season)
  - True shooting percentage
  - Free throws/minute
  - 3-point shots/minute
  - 2-point shots/minute
  - Offensive rebounds/minute
  - Defensive rebounds/minute
  - Assists/minute
  - Average player game score
- Non-player factors – College:
  - School's conference winning percentage (final college season)
  - Number of players from school drafted into the NBA (prior 10 years)
- Non-player factors – NBA team (prior season):
  - Win percentage
  - Average point spread

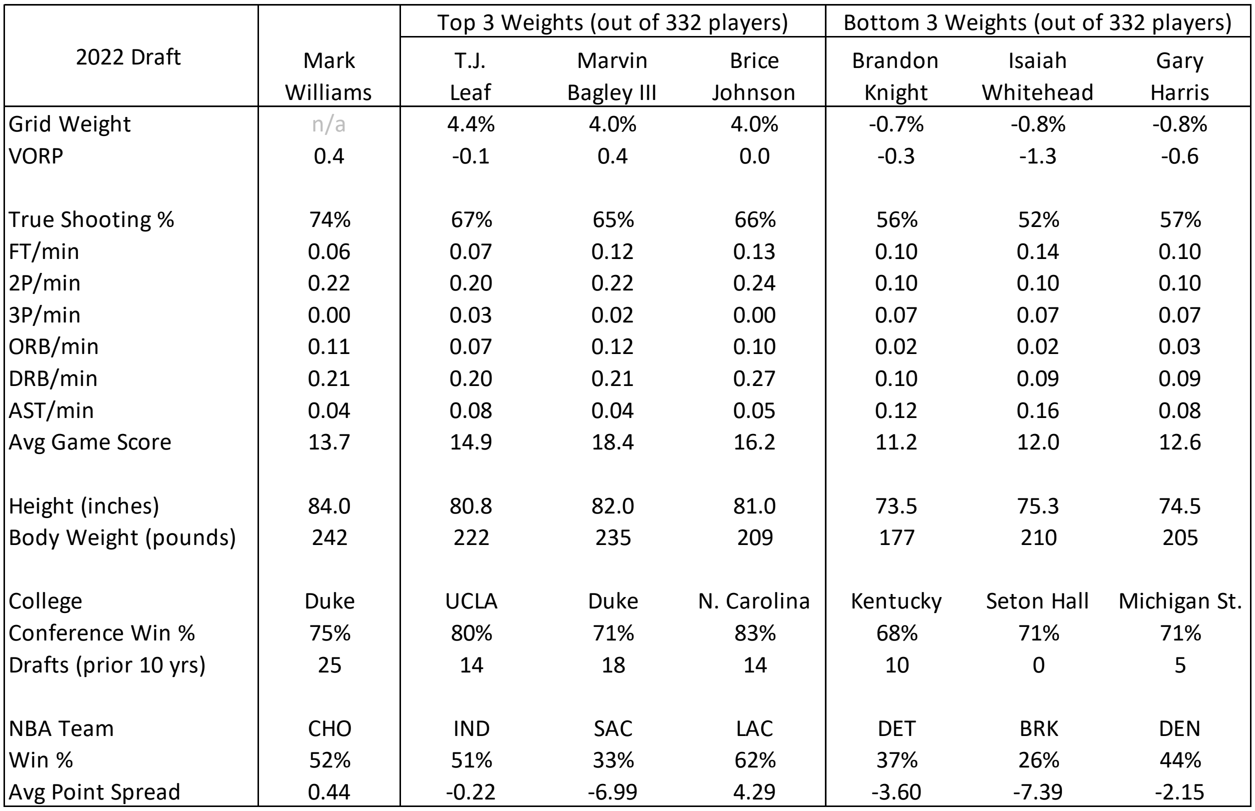
As we discussed previously, CKT grid prediction considers many subsets of previously drafted players and predictive variables for each individual prediction task. The information from every cell in the grid is aggregated to form one composite vector of weights across all previously drafted players. These weights directly determine the prediction: it equals the weighted average of player outcomes. The weights also contain other important information. For illustrative purposes, let us consider the VORP predictions for the 2022 draft cohort. The left-hand side of Figure 2 shows the fit of each prediction, which equals the squared correlation of weights and outcomes. These results reveal that expected reliability varies dramatically from one prediction to the next. The right-hand side of Figure 2 shows the degree of concentration in the weights vector, which we measure as the sum of squared weights. Equal weights to all prior players would have the lowest possible concentration $(1/N)$, and reliance on a single player would have the highest possible concentration (1). The prediction for Mark Williams is the most concentrated, and it also has a high fit. This means that players who are relevant to Mark Williams tend to have consistent outcomes that warrant large weights, whereas non-relevant players do not. The consistency in the retained sample is strong enough that it outweighs the noise of a smaller sample, and the overall fit is high. The draft prospect with the highest fit, Blake Wesley, has less concentrated weights and thus benefits from a broader set of relevant players. Meanwhile, the prediction for Moussa Diabate, with a similar concentration, has one of the lowest fits.

Figure 3: Fit and Grid Weight Concentration (VORP Predictions for 2022 Draft Cohort)



One of the most powerful features of relevance-based prediction is that it reveals precisely how each previously drafted player informs the prediction. For example, Figure 4 shows the three most relevant and three least relevant players for forming the VORP prediction for Mark Williams who was drafted 15th by the Charlotte Hornets in the 2022 draft. It is affirming to note that relevance-based prediction identified three players of similar height as most relevant to forming Mark Williams' VORP prediction, and three significantly shorter players as least relevant to forming his prediction. And it successfully identified a set of most relevant players with similar rookie season VORPs as what subsequently occurred for Williams, and a set of least relevant players with comparatively dissimilar VORPs, without foreknowledge of the outcome for Williams. It also identified a fellow Duke alumnus as one of the most relevant players. The main takeaway from Figure 4, though, is the extraordinary level of transparency relevance-based prediction affords, which is critical for facilitating dialogue between analytics professionals, coaches, and scouts.

Figure 4: Most and Least Relevant Players for Mark Williams



2022 Draft		Top 3 Weights (out of 332 players)			Bottom 3 Weights (out of 332 players)		
	Mark Williams	T.J. Leaf	Marvin Bagley III	Brice Johnson	Brandon Knight	Isaiah Whitehead	Gary Harris
Grid Weight	n/a	4.4%	4.0%	4.0%	-0.7%	-0.8%	-0.8%
VORP	0.4	-0.1	0.4	0.0	-0.3	-1.3	-0.6
True Shooting %	74%	67%	65%	66%	56%	52%	57%
FT/min	0.06	0.07	0.12	0.13	0.10	0.14	0.10
2P/min	0.22	0.20	0.22	0.24	0.10	0.10	0.10
3P/min	0.00	0.03	0.02	0.00	0.07	0.07	0.07
ORB/min	0.11	0.07	0.12	0.10	0.02	0.02	0.03
DRB/min	0.21	0.20	0.21	0.27	0.10	0.09	0.09
AST/min	0.04	0.08	0.04	0.05	0.12	0.16	0.08
Avg Game Score	13.7	14.9	18.4	16.2	11.2	12.0	12.6
Height (inches)	84.0	80.8	82.0	81.0	73.5	75.3	74.5
Body Weight (pounds)	242	222	235	209	177	210	205
College	Duke	UCLA	Duke	N. Carolina	Kentucky	Seton Hall	Michigan St.
Conference Win %	75%	80%	71%	83%	68%	71%	71%
Drafts (prior 10 yrs)	25	14	18	14	10	0	5
NBA Team	CHO	IND	SAC	LAC	DET	BRK	DEN
Win %	52%	51%	33%	62%	37%	26%	44%
Avg Point Spread	0.44	-0.22	-6.99	4.29	-3.60	-7.39	-2.15

Next, we show how well relevance-based prediction performed. To do so we compare the correlations of predictions with outcomes for the predictions foreseen to be most reliable, the predictions foreseen to be least reliable, and predictions based on linear regression analysis. We define most reliable as those predictions with the top 50% fits and the least reliable as those predictions with the bottom 50% fits. One could choose any threshold to delineate reliability. We chose 50% as a reasonable threshold for the purpose of illustration. Keep in mind that the fit values we use to discard less reliable predictions are known in advance of the predictions and are unknowable with other prediction systems. In addition to correlations of prediction values and outcome values, we show correlations of ranks in case the value correlations are unduly influenced by unusual values.

Figure 5 presents the correlations of VORP predictions and outcomes for the 2018, 2021, and 2022 draft cohorts as well as the average across these cohorts. Figure 6 presents the same comparisons based on ranks. They reveal that the relevance-based predictions foreseen to be reliable are substantially superior to the predictions that come from linear regression analysis and superior by an even greater margin, on average, than the predictions foreseen to be unreliable, as judged by fit.

Figure 5: Correlations of VORP Predictions with VORP Outcomes based on Value

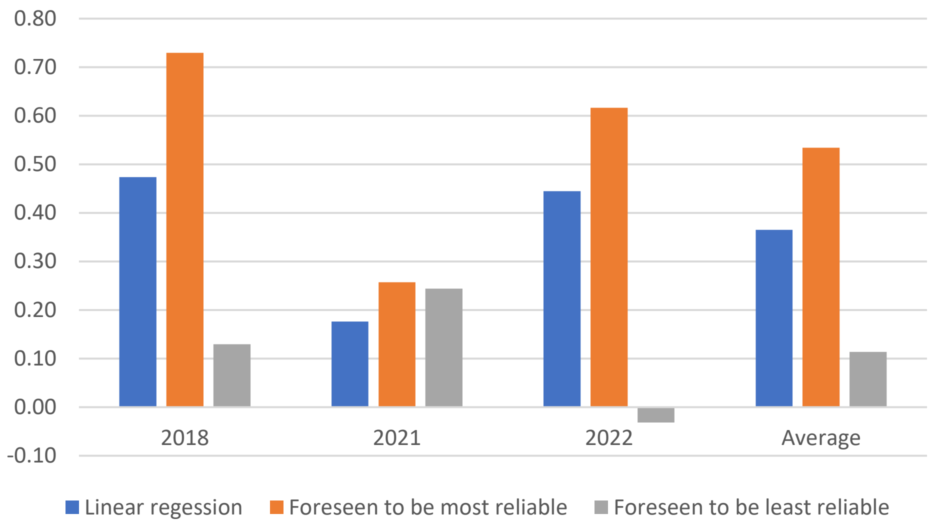
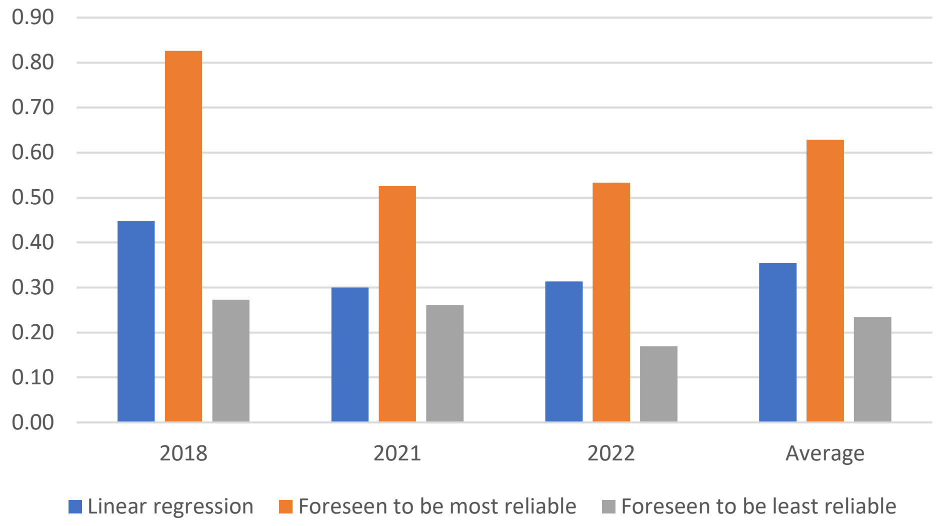


Figure 6: Correlations of VORP Predictions with VORP Outcomes based on Rank



Figures 7 and 8 present the value and rank correlations, respectively, of minutes played predictions and minutes played outcomes for the same draft cohorts. They reveal that relevance-based prediction is similarly superior to linear regression analysis in predicting minutes played and similarly, if not more effective, in rendering advance notice of a prediction's reliability.

Figure 7: Correlations of Minutes Played Predictions with Minutes Played Outcomes based on Value

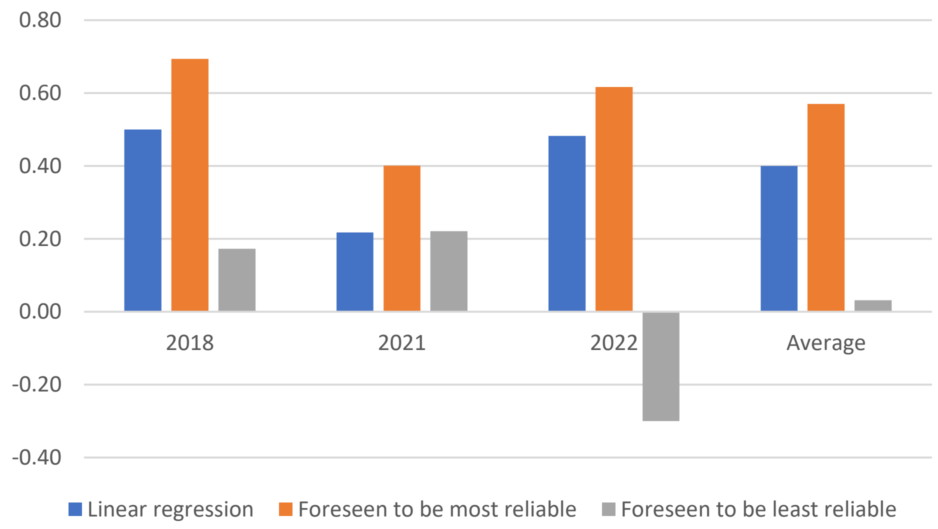
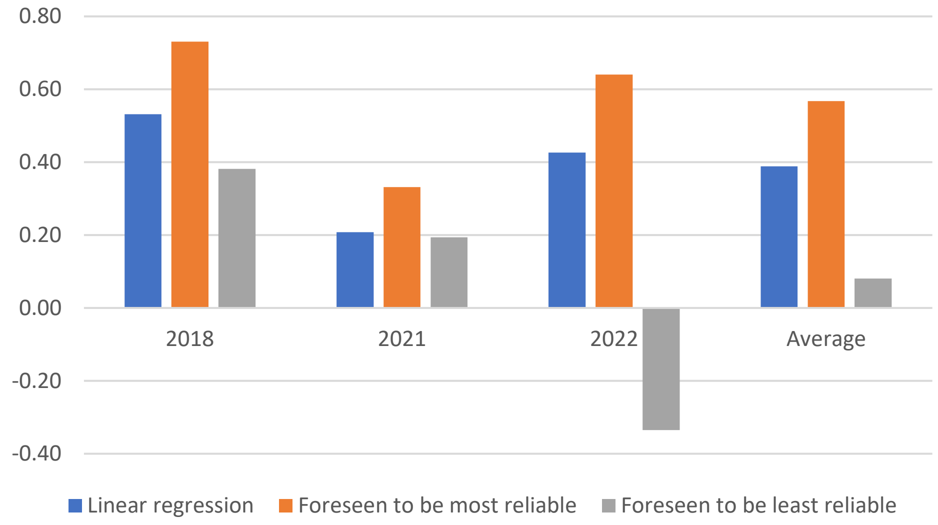


Figure 8: Correlations of Minutes Played Predictions with Minutes Played Outcomes based on Rank



We acknowledge that a skilled sports analytics professional might be able to produce more reliable predictions simply by using more effective predictive variables and more informative data in a conventional way. However, our results offer persuasive evidence that relevance-based prediction should produce superior results, given the same quality of variables and data, which we attribute to several key advantages.

- It considers complex asymmetries that are beyond the capacity of linear regression analysis.
- It dynamically customizes the use of previously drafted players and predictive variables for each new prediction task.
- It reveals precisely how each previously drafted player informs the prediction for the draft prospect.
- It reveals how much confidence one should assign to each specific prediction task in advance of the prediction, thereby enabling one to discard predictions that are foreseen to be less reliable.
- It is theoretically grounded.

Moreover, it is generally applicable. It can be applied to any set of data for the purpose of predicting any outcome, even beyond sports.

6. Summary

We described a new approach for predicting outcomes for NBA draft prospects called relevance-based prediction. This approach forms predictions as weighted averages of past outcomes in which the weights are based on the relevance of previously drafted NBA players, measured in a mathematically precise and theoretically justified way.

Then we described a measure of prediction-specific fit, which indicates the specific reliability of each individual prediction task. 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. And of critical importance, we showed that fit enables us to discard, or consider more cautiously, predictions that are foreseen to be less reliable.

Next, we introduced the concept of codependence, which holds that the efficacy of previously drafted players depends on the chosen predictive variables, and the efficacy of predictive variables depends on the chosen previously drafted players. We also showed how to blend the information gained from each unique combination of previously drafted players and predictive variables based on its relative reliability according to adjusted fit.

We then illustrated our new relevance-based system by predicting VORP (value over replacement player) and total minutes played during the season for the rookie seasons for players who were drafted in 2018, 2021, and 2022. Our analysis revealed that relevance-based prediction consistently produced results that were significantly superior to those generated by linear regression analysis, given the same sample of players, and set of predictive variables, and that it was able to distinguish in stark relief predictions foreseen to be reliable from those foreseen to be unreliable.

To conclude, we acknowledge that scouting information provides insights that would be unobtainable from any analytical system. However, we wish to emphasize that relevance-based prediction produces valuable information that is otherwise unknowable. We therefore recommend our relevance-based system as a complement to scouting and not as an alternative.

References

[1] Czasonis, Megan, Mark Kritzman, and David Turkington. 2020. "Addition by Subtraction: A Better Way to Forecast Factor Returns (and Everything Else)." *The Journal of Portfolio Management* 46, no. 8: 98–107.

[2] Czasonis, Megan, Mark Kritzman, and David Turkington. 2022. *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, no. 1: 27–46.

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

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

[6] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.. 2017. "Attention Is All You Need." *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.*

Appendix: Convergence of Relevance to Other Prediction Methods

Convergence to Linear Regression Analysis

The prediction equation corresponding to full sample linear regression equals:

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

Expanding the expression for relevance gives:

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

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

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

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

Where:

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

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:

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

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

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

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 relevance-based prediction, 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:

$$q_t k_i' = x_t W^Q W^K x_i' \tag{A13}$$

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 relevance-based prediction described in this article.