Bayesian Estimation of an NBA Player's Impact on his Team's Chances of Winning

Aderinto Adesijibomi

5/12/2022

Abstract

Estimations of player performance as key indicators are employed using traditional vendor metrics such as Efficieny. However using this metric provide only estimates of player performance without factoring his performance with respect to other team members. Examples are seen in scenarios where the opponents team members are easy team to face,in such cases the win probability is easily decided before the end of the game and when faced with a tough opponent where probability is decided until the end of the game. This estimates provides a misleading or less informative analysis of player contributions to his team winning games. In this paper a Bayesian regression model is modeled to include players on the court as independent variables, and the response variable is a function of the change in the win probability for every point in time the game. Using this model win probabilities at every point in time is expresses as a combinations of all significant independent variables. Ranking the coefficients of the independent variables. Players who important contributors to the teams winning can be obtained.

Introduction

To understand the concepts of Bayesian learning in sports, one must first understand what analytics in sport is. Sport analytics is a collection and statistical analysis of various types of data, historical, player performance, business operations types of data, that can provide a team or individual with a competitive advantage (Fried and Mumcu 2016). Sport analytics can provide actionable insights by informing players, coaches, managers, and other staff in order to facilitate decision making both during and prior to sporting events. Following the release of the the 2011 film Moneyball, in which Oakland Athletics General Manager Billy Beane relies heavily on the use of analytics to build a competitive team on a low budget, thus the term 'sport analytics' became popular in mainstream sports culture(Wikipedia contributors 2022). Modern sports science is both characterized, challenged by the volume and variety of available data. The rich data gotten from historical data,wearable, makes application of statistical modeling the best methodology for sport analysis (Santos-Fernandez, Wu, and Mengersen 2019). Bayesian can be considered as a statistical technique that relies on Bayes theorem, treat unknown parameters probabilistic and give subjective treatment to probabilities (Bernardo and Smith 2009). One of the most obvious question in basketball analytics and sports generally is determining which players do the most to help their team win games. Traditionally the concepts of analyzing teams in winning have been quantified using a scoring statistics such as points-per-game or true shooting percentage, or a point differential function such as adjusted plus-minus (Rosenbaum 2004). A common metric used in statistical benchmark for comparing the overall value of players is called efficiency (Piper et al. 2022).It is a composite statistic that is derived from basic individual statistics: offensive contributions (points and assists), defensive contributions (steals and blocks). Mathematical formula is derived as

$$\frac{(Points + Rebound + Assists + Steals + Blocks - MissedFieldGoals - MissedFreethrows - Turnovers)}{GamesPlayed} \tag{1}$$

Other metric examples are PER (player efficiency rating), DPR (defensive player rating)

Deshpande and Jensen (2016) asserts while these metrics are informative, and because they do not take into account the context in which players perform, thus they result in false positive probabilities of performance of players in low leverage situations, when the games outcome is essentially decided, for example situations where the win expectancy is unlikely to swing unexpectedly in favor of either home team or away team or in high leverage situations where the outcome of the game hangs on balance. Typical example is seen using game points differential indicators will treat the home team's lead dropping from 5 points to 0 points in the last minute of the first half in the same way model the home team's lead dropping from 30 points to 25 points in the last minute of the second half. The home's team differential is -5 points in both scenarios, but we will see later in this paper, the home team's chance of winning the game dropped from 72 percent to 56 percent in the first scenario while remaining constant at 100 percent in the second scenario. (Deshpande and Jensen 2016) argue that a player's performance in the second scenario has no bearing on the final result and, as a result, should not be compared to that in the first.

Several methods have been used in the past to estimate the concepts of win probability in past years. (Mills and Mills 1970) who evaluated win probability in baseball as win probability added by computing the probability of winning at different times in the game, this is computed by initially estimating the win probability of winning the games at different times in the game and further crediting both pitcher and batter per each plate appearance. Each player contributions are sum over the entire period of the season which can be expressed as the contributions of players to the team winning and losing for the entire season. This is similar to the plus-minus statistics where each player on the court is credited per every basket made and deducted by each opponents basket made, presented as contribution to the team wins if a particular player is on the court (Studeman 2004). This box-score metric is similar to the win probability except that it is computed over the points scored instead of probabilities. A shortcoming of this method is, to capture each player contributions to his team winning one must factor the combination of his team members and the opponents players per each player appearance on the court. In other to fix this weakness in the model, (Rosenbaum 2004) proposed an adjusted plus-minus model as probability of the change in points scored per number of possession in each shift, where the game is divided in to shifts, we the regress this change in points over 100 possessions in every shift to the indicators of every players on the court per shift.

To address this problem, (Deshpande and Jensen 2016) proposed a win probability framework and a linear regression model for estimating each player's contribution to his team's overall chance of winning games. He regressed the change in home win probability during each shift of play by diving player on court into 5 home players and 5 away players in other to estimate each team player partial effects on his teams chances of winning. this is expressed as

$$E[y_i \mid h_i, a_i] = \mu_i + \theta_{hi_1} + \cdots + \theta_{hi_{15}} - \theta_{ai_1} - \cdots - \theta_{ai_{15}} \tag{2}$$

where $h_i = h_{i1}, ...h_{i15}$ and $a_i = a_{i1}, ..., a_{i15}$ are indices on corresponding to the home team (h) and away team (a) players, and $\theta_{(1,...488)}$ is the vector of player partial effects.The intercept term μ_i may depend on additional covariates, such as team indicators. Fitting the model in equation 1 is difficult in that we have a large number of players which are not identically independent hence may lead to high degree of collinearity, a method to mitigate this by using regularition of the parameters, depending on the type of regularition choosen, parameters are shrunk as close to zero as possibile. By using Bayesian which involves starting with a prior with mode zero on each player partial effects and using play by play data from season data. with this the model the posterior distribution of the player partial effects is obtained. It is important to note that this purpose of this model does not measure each player unobserved or latent abilities and skill but provide probabilitic measure of each player contribution to their team's winning and these probabilities are context dependent hence it cannot be used for future predictions, since player context can vary from season to season, however it can be used for detailed accounting of each player evaluation metrics as compared to the traditional metrics of EFF, DPR, PER. When combined with traditional metrics, it can help coaches properly distribute playing time and can assist in understanding the amount to which a player's individual performance translates to team wins. A practical way to view this is that after every game, team managers go through lots of playback videos, charts and other metric captured during the game and manually try to correlate them to see which players increased their chance of winning.

Regression Model

As described in the adjusted plus and minus model, each game is divided into shifts of play where we have 10 unique players on the court 5 home team players and 5 away team players. To determine which players were on the court we use play-by-play data obtained from ESPN 2006–2007 season to 2012–2013 season which comprises of 8 seasons . After cleaning the data it was found that 15% of the data were either missing or incomplete for games from the first half of the time window considered. (Deshpande and Jensen 2016) denote the change in the win probability of the home team in the ith shift by y_i, the performance of the 10 players in each ith shift is directly responsible for the change in the win probability, therefore we regress change in the win probability onto parameters of each player on the court during the ith shift to quantify each player response variable as each player contribution to the change in win probability. we model

$$E[y_i \mid h_i, a_i] = \mu_i + \theta_{hi_1} + \cdots + \theta_{hi_{15}} - \theta_{ai_1} - \cdots - \theta_{ai_{15}} + \tau H_i - \tau A_i + \sigma \cdot \epsilon \tag{3}$$

where θ_1 to θ_{488} is a vector of partial effects of every players, τ_1 to τ_{30} is a vector of partial effects of the 30 teams, h_1 to h_{15} and a_1 to a_{15} are indices on tau corresponding to the teams playing in the i_shift and e are the independent normal random variable. μ_i can be viewed as intercept representing home court advantage and sigma is the variability that arises from measuring the uncertainty in y_i

Estimating The Win Probability The Likelihood

To begin fitting the regression model, we must start with estimating the posterior mode for the mean, which will be the probability that the home team wins the game leading with change in points score denoted by L points and time duration T seconds denoted by $P_{T,L}$ and by placing a conjugate $Beta(\alpha_{T,L}, \beta_{T,L})$ prior on $P_{T,L}$ and estimate the posterior mean

$$\hat{P}_{T,L} = \frac{n_{T,L} + \alpha_{T,L}}{N_{T,L} + \alpha_{T,L} + \beta_{T,L}}$$

where $n_{T,L}$ is the number of games which the home team won in this window and $N_{T,L}$ be the number of games in which the home team has led by L points after T seconds modeled as binomial$(N_{T,L}, P_{T,L})$

Wegmann (2022) derivations is shown below

$$x_1, ..., xn \mid \theta \overset{\text{idd}}{\sim} Bern(\theta)$$

prior

$$prior \sim Beta(\alpha, \beta)$$

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \; for \; 0 \leqslant \theta \leqslant 1$$

posterior

$$p(\theta \mid x_1, ..., xn) \propto p(x_1, ..., xn \mid \theta) p(\theta)$$

$$\theta^s (1-\theta)^f \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}$$

hence the posterior is proportional to $Beta(\alpha + s, \beta + f)$. Take the log sum of the posterior θ equals

$$logP(\theta \mid x) = log \cdot \theta(s + \alpha - 1) \times log \cdot (1 - \theta)$$

$$\frac{\partial logP(\theta \mid x)}{\partial \theta} = \frac{s + \alpha - 1}{\theta} - \frac{f + \beta - 1}{1 - \theta} = 0$$

solving for θ by equating to 0

$$\hat{\theta} = \frac{\alpha + s - 1}{\alpha + \beta + n - 2}$$

at large values of α β it can be written as

$$\hat{\theta} = \frac{\alpha + s}{\alpha + \beta + n}$$

where n is $N_{T,L}$ and s is $n_{T,L}$

$$\hat{P}_{T,L} = \frac{n_{T,L} + \alpha_{T,L}}{N_{T,L} + \alpha_{T,L} + \beta_{T,L}}$$

Bayesion Linear Regression of Player Effects

As stated in the introduction, Bayesian approach is used to for fitting the model in equation 2, and due to the large number of covariates (488) this introduced a high degree of collinearity. Therefore a regularition prior that shrinks component as close to zero is needed to stabilize the model. laplace(L1) prior was chosen because it tend shrink the component towards zero as fast as possible and an existing implementation of gibbs sampler exist in R using the mononvm package. other implementations of regularition are normal or ridge regularition and elastic net regularition, elastic net is better suited for regression problems, the trade off is overhead cost in complexity and the availaility of package suite in R. this is modeled as P is the vector of player players on the court during a specific shift i represented as P^i_j and T be vector of what teams are playing during i shift represented by T^i_j therefore,

$$y_i \mid P^i, T^i \sim \begin{cases} \mathcal{N}(\mu + P^{iT}\theta + T^{iT}\tau\, \sigma^2) & P^i = T^i = 1 \quad for \quad home \quad games \\ 0 & elsewhere \\ \mathcal{N}(\mu + P^{iT}\theta + T^{iT}\tau\, \sigma^2) & P^i = T^i = -1 \quad for \quad away \quad games \end{cases}$$

we place laplace priors on each component of θ and τ conditional on the noise and variability σ^2 the conditional prior densities is expressed as

$$p(\theta, \tau \mid \sigma^2) \propto \Big[(\frac{\lambda}{\sigma})^{488} \times exp\Big(-\frac{\lambda}{2\sigma} \sum_{j=1}^{488} \mid \theta_j \mid \Big) \Big] \times \Big[(\frac{\lambda}{30}) \times exp\Big(-\frac{\lambda}{2\sigma} \sum_{k=1}^{30} \mid \tau_k \mid \Big) \Big]$$

$\lambda > 0$ is the penalty factor that determines the degree of how the model shrinks to zero. $(\mu, \sigma^2, \theta, \tau)$. Initial prior parameter is used as the default starting values for the regression model. We rely on Markov Chain Monte Carlo algorithm by Gibbs in R to simulate the estimate of the posterior distribution. Gibbs sampling is applicable when the joint distribution is unknown or it's difficult to sample directly but the conditional distribution of each variable is known. Gibbs sampling algorithms is one the most popular Markov chain Monte Carlo methods. It produces samples from the posterior distribution by successfully drawing samples from the full conditional distributions of the target distribution. It is special case of MCMC Hasting algorithm in which we don't not have an acceptance or rejection region.

Wegmann (2022) explained Gibbs sampling using MCMC algorithm using Bayesian variable selection, lets assumes a linear regression with β coefficients representing the partial effects of 488 players expressed as

$$y = \beta_0 + \beta_1 \cdot x_1 +, .., \beta_p \cdot x_p + \epsilon$$

. Identifying what player affects the response variable y will be a linear combinations of which β that are non zeros. Next we introduce a Bernoulli distribution to combine all successes where β values are all 1s, this values will be used as our prior or proposal distribution. Posterior probability can be evaluated as

$$p(\mathcal{I} \mid y, X) \propto p(y \mid X, \mathcal{I})(p\mathcal{I})$$

where $\mathcal{I}$ represents the matrix combinations of non zero coefficients of the linear model. The marginal likelihood $p(y \mid X, \mathcal{I})$ is the integral of

$$\int p(y \mid X, \mathcal{I}, \beta) p(\beta \mid X, \mathcal{I}) \partial \beta$$

. The joint prior parameters of β and σ^2 can be represented as

$$\sigma^2 \sim \mathcal{I}nv - \chi^2(v_0, \sigma_0^2)$$

$$\beta_{\mathcal{I}} \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2 \Omega_{\mathcal{I},0}^{-1})$$

. Hence marginal likelihood of y conditional on X and $\mathcal{I}$ is thus expressed as

$$p(\mathcal{I} \mid y, X) \propto \mid X_{\mathcal{I}}' X_{\mathcal{I}} + \Omega_{\mathcal{I},0}^{-1} \mid^{-\frac{1}{2}} \mid \Omega_{\mathcal{I},0}^{-1} \mid^{\frac{1}{2}} (v_0 \sigma_0^2 + \mathcal{R}SS_{\mathcal{I}})^{\frac{-v_0+n-1}{2}}$$

and $\mathcal{R}SS_{\mathcal{I}}$ is the residual sum of squares. Due to the multivariate matrix $\mathcal{I}$, it will computational in efficient to draw posterior samples, hence we implement Gibbs sampling which is special case of MCMC HA algorithm. in gibbs sampling we simulate draws from $p(\mathcal{I} \mid y, X)$

$$Draw I_1 \mid \mathcal{I}_{-1}, y, X$$

$$Draw I_2 \mid \mathcal{I}_{-2}, y, X$$

$$\cdots$$

$$Draw I_p \mid \mathcal{I}_{-p}, y, X$$

Assumptions

Change In Win Probability $\mathcal{Y}_i$ idd

To fit the regression model we assume $y_i s$ are independently identical, but this is not the case as all 10 players on the court are changed at every shift of substitutions, it's therefore expected to be some autocorrelation between y_i and y_{i+1} plotted in time series shown in plot a negative autocorrelation of -10 is seen constant across all lag, with no change observed at larger lags. The lack of strong correlation between y_i and y_{i+1} makes assumption $\overset{\text{idd}}{\sim}$ feasible while the assumption is not practically possible

$\mathcal{Y}_i$ is Guassian

The multivariate random variables conditional on (P_i, T_i), y_i is Gaussian with constant variance, hence the linear combinations of the $(\mu, \sigma^2, \theta, \tau)$ must have a normal distribution.

Thirdly as y_i depends on the length of each shifts, it is acceptable to conclude that change in y_i is function of time duration and lead. As shown in figure 1, it can be seen clearly change in win probability varies at different time shifts, however the overall correlation is very small between y_i and y_{i+1}, it can be said that a player performance in shift 15s in which his player leads by 20% is the same as the impact his team has in winning if the lead time extends for 30s.

Posterior Analysis

Gibbs algorithm in r implemented using the monomvn package in R studio by sampling from the full posterior distribution of $(\mu, \sigma^2, \theta, \tau)$. The marginal posterior distribution of each variable is plotted and represented as the marginal probabilities of each player's partial effect as shown the figure below.

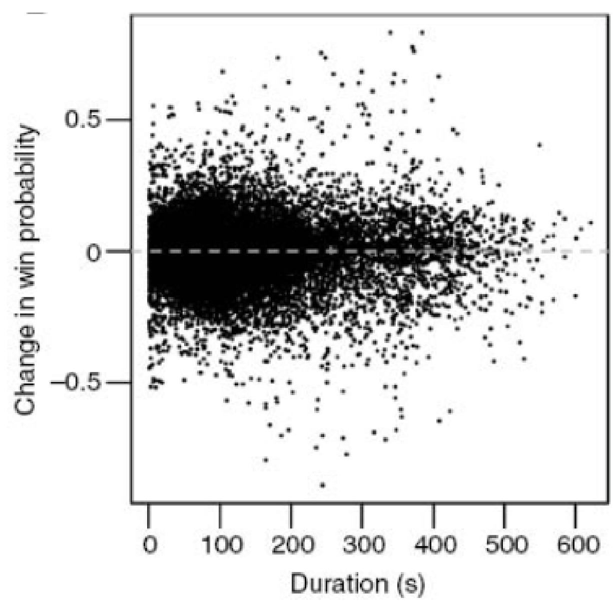

Figure 1: Change in win probability with different lead times (Deshpande and Jensen 2016)

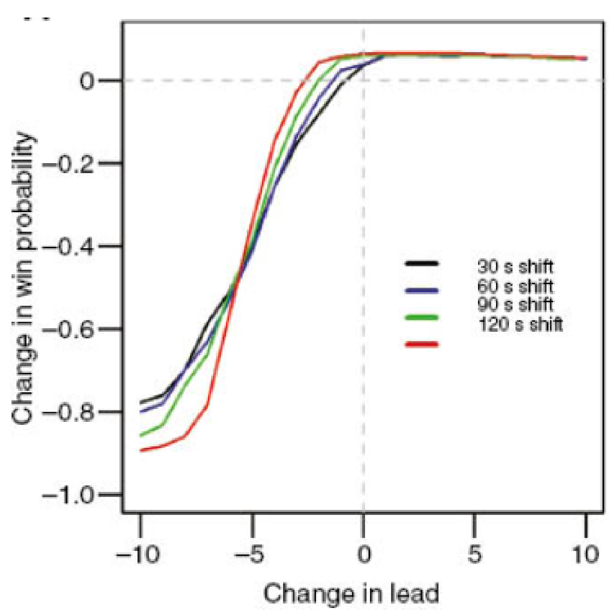

Figure 2: Change in win probability with different time durations (Deshpande and Jensen 2016)

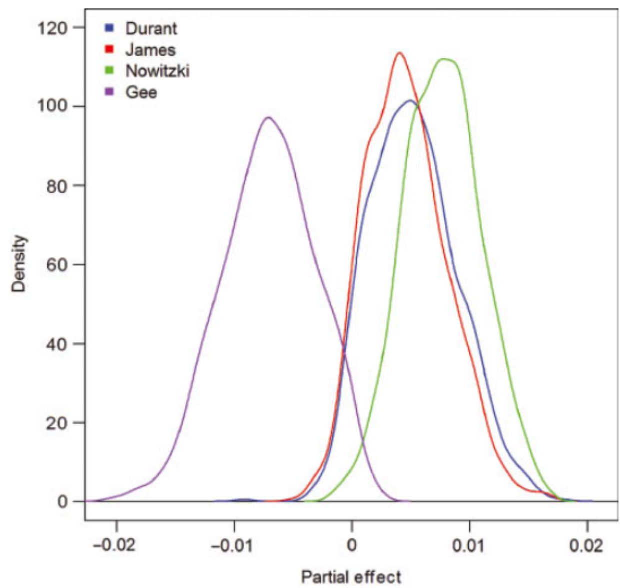

Figure 3: Approximate posterior densities of several players' partial effects(Deshpande and Jensen 2016)

From the sample plot of the posterior probabilities of player partial effects, it is observed that the the variability of y_i is within the range of [-0.2, 0.2], this implies that the marginal posterior estimates of one partial effect cannot determine overall win probability, thus the bias the model is reduced a close to zero as possible. A possible cause of this is the regularization introduced which shrinks the components of covariates as close to zero as possible. However due to the small values of the posterior estimates, enough dissemblance is seen which is used to plot the posterior probabilities. Looking the at plot we can see evaluate each player contribution to his team winning in games. figure below show the various plot of player effect at various times over the course of the season.

Comparing Players

Directly comparing the partial effects of players is not possible because of the multivariate characteristics of the contexts in which they played. To group the player effects with similar player effects, Mahalanobis distance between leverage profiles of each player was computed, leverage profiles such as total number of shifts in which players played, player's team average win probability at the start of the shift,average duration of shift,and length of shift. Doing this we compared players playing with the same leverage profiles. Table 1 shows the player comparison table with similar leverage profiles shown in parenthesis. Table 6 shows a boxplot comparison of posterior distribution of player partial effects. As shown in the posterior analysis that player effects can be viewed as his contributions to his teams winning in games. Several key performance indicator can also be deduced from posterior samples. Impact ranking, where we rank all players based on the mean of their posterior probabilities.

Discussion

This study identifies the importance of adding features of context in which every player plays in estimating player impact on his teams chances of winning games. Using Bayesian statistical modelling combined with linear regression, probabilities of each player in any game which influenced a change in overall probability of team winning in games can be derived. Comparing traditional box score metrics which very often can be

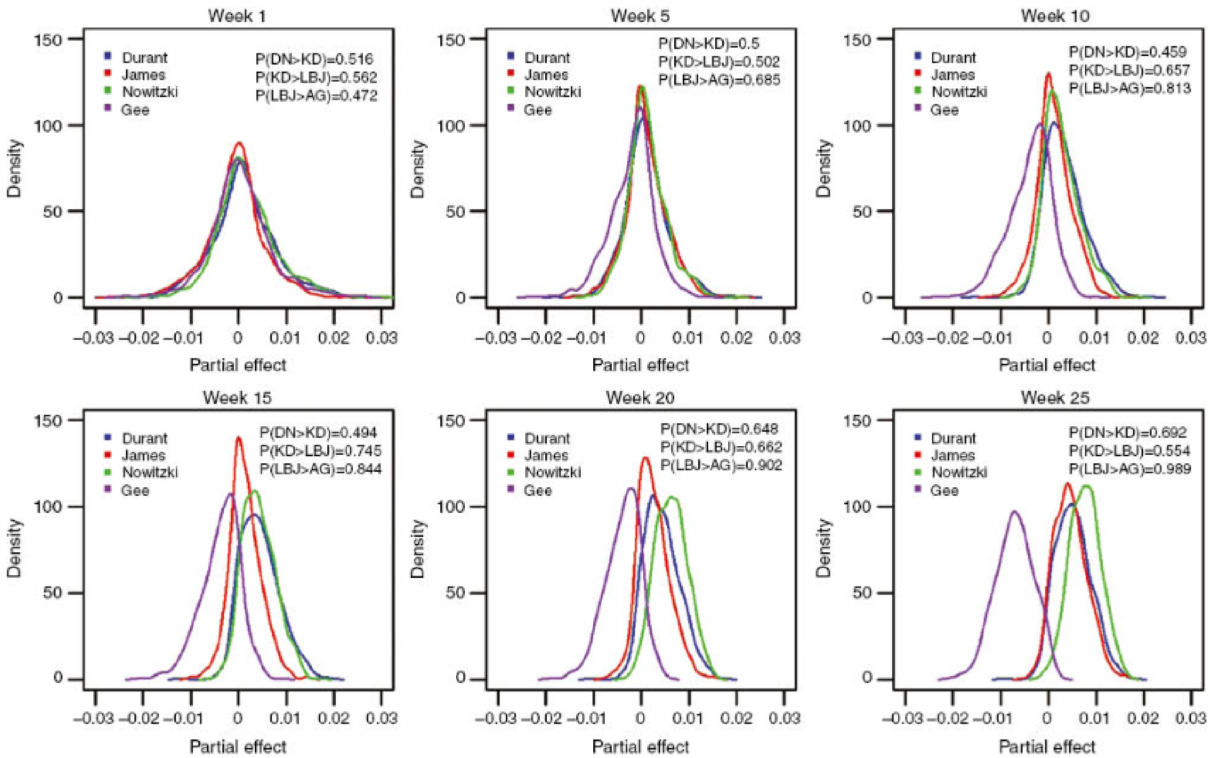

Figure 4: Approximate posterior densities of Kevin Durant's, LeBron James', and Dirk Nowitzki's partial effect as the season progresses(Deshpande and Jensen 2016)

Player	**Similar players**	
LeBron James	DeAndre Jordan (0.025)	Kevin Durant (0.055)
	Blake Griffin (0.082)	Stephen Curry (0.204)
Chris Paul	Shawn Marion (0.081)	Courtney Lee (0.103)
	Terrence Ross (0.126)	Chris Bosh (0.141)
Kyrie Irving	DeMarcus Cousins (0.080)	Tristan Thompson (0.087)
	Brandon Bass (0.099)	Randy Foye (0.109)
Zach Randolph	Jimmy Butler (0.020)	David West (0.045)
	Mike Conley (0.063)	George Hill (0.073)

Figure 5: Players with similar leverage profiles(Deshpande and Jensen 2016)

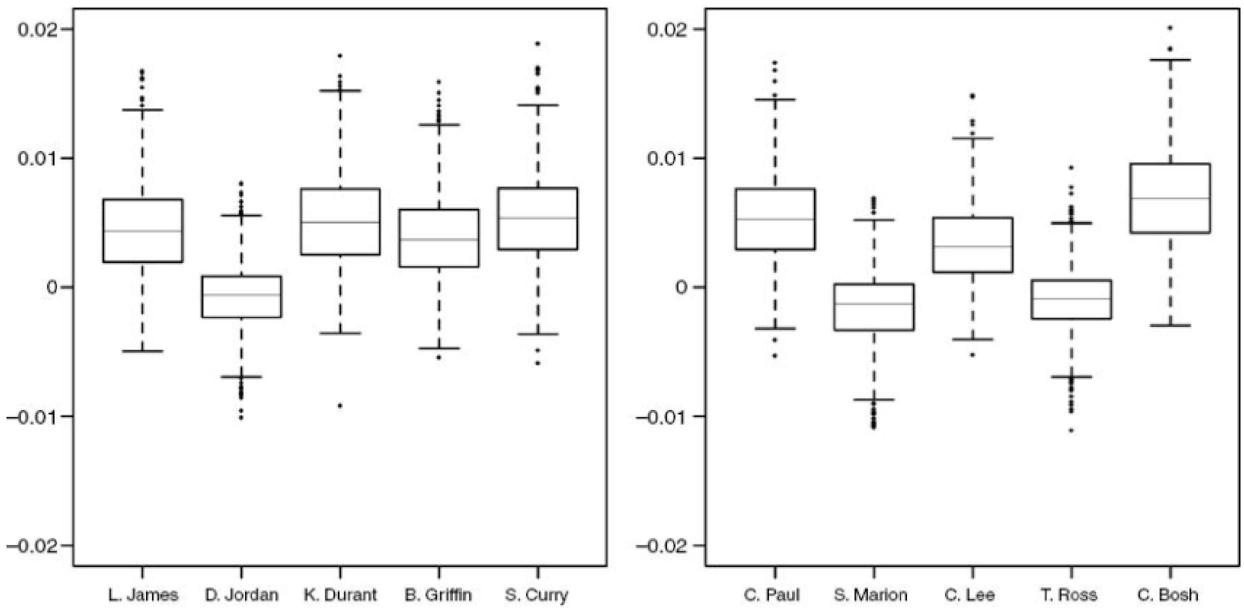

Figure 6: Comparison box plots of partial effects of players with similar leverage profiles.(Deshpande and Jensen 2016)

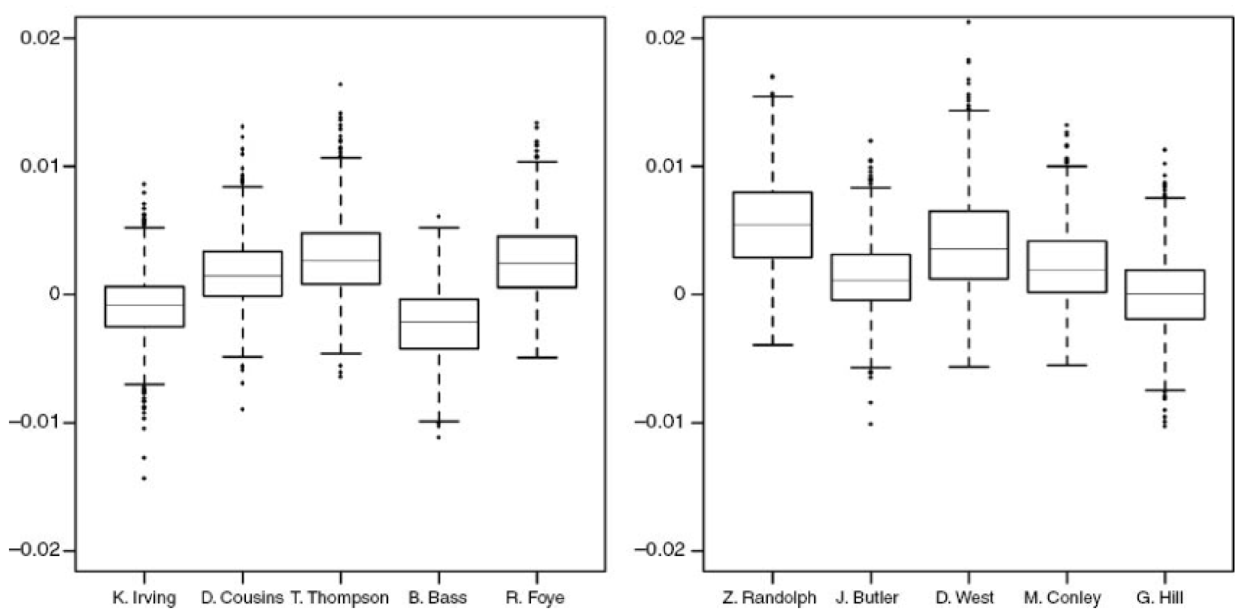

Figure 7: Comparison box plots of partial effects of players with similar leverage profiles.(Deshpande and Jensen 2016)

vendor agnostic usually provides performance estimations of several types of key indicators with Bayesian models which factors in player contexts and hence provide a more correlated view of player effect and his team effects in every game. Both boxscore metric and Bayesian models can be used congruently where boxscore metric provide individual player metrics and Bayesian models provide a holistic view of each player performance conditional on his team members

Reference

Bernardo, José M, and Adrian FM Smith. 2009. *Bayesian Theory*. Vol. 405. John Wiley & Sons.

Deshpande, Sameer K, and Shane T Jensen. 2016. "Estimating an NBA Player's Impact on His Team's Chances of Winning." *Journal of Quantitative Analysis in Sports* 12 (2): 51–72.

Fried, Gil, and Ceyda Mumcu. 2016. *Sport Analytics: A Data-Driven Approach to Sport Business and Management*. Taylor & Francis.

Mills, Eldon G, and Harlan D Mills. 1970. "Player Win Averages: A Complete Guide to Winning Baseball Players."

Piper, Jasmine, Bill Siddiq, Hiroki Matsuo, Grant Morgan, Ali Boolani, and Andreas Stamatis. 2022. "Predicting in-Season Performance from Pre-Season Power and Gait Metrics in Collegiate Basketball: Preliminary Evidence." In *International Journal of Exercise Science: Conference Proceedings*, 15:2. 1.

Rosenbaum, Dan T. 2004. "Measuring How NBA Players Help Their Teams Win." *82Games. Com (Http://Www. 82games. Com/Comm30. Htm)*, 4–30.

Santos-Fernandez, Edgar, Paul Wu, and Kerrie L Mengersen. 2019. "Bayesian Statistics Meets Sports: A Comprehensive Review." *Journal of Quantitative Analysis in Sports* 15 (4): 289–312.

Studeman, D. 2004. "The One about Win Probability." *Hardball Times*.

Wegmann, Berti. 2022. *Bayesian Learning: Bayesian Model Comparison. Variable Selection.* Department of Computer; Information Science Linkoping University.

Wikipedia contributors. 2022. "Sports Analytics — Wikipedia, the Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Sports_analytics&oldid=1075510338.