Are NBA Players' Salaries in Accordance with Their Performance on Court?

Ioanna Papadaki and Michail Tsagris

Abstract Researchers and practitioners ordinarily fit linear models in order to estimate NBA player's salary based on the players' performance on court. On the contrary, we first select the most important determinants or statistics (years of experience in the league, games played, etc.) and utilize them to predict the player salary shares (salaries with regard to the team's payroll) by employing the non-linear Random Forest machine learning algorithm. We are further able to accurately classify whether a player is low or highly paid. Additionally, we avoid the phenomenon of over-fitting observed in most papers by external evaluation of the salary predictions. Based on information collected from three distinct periods, 2017–2019, we identify the important factors that achieve very satisfactory salary predictions and we draw useful conclusions. We conclude that player salary shares exhibit a relatively high (non-linear) accordance with their performance on court.

Keywords NBA · Salaries prediction · Variable selection · Non-linear models

1 Introduction

Professional athletes' field performance and salaries is a topic that has attracted the interest of numerous researchers (Zimmer and Zimmer 2001; Yilmaz and Chatterjee 2003; Olbrecht 2009; Vincent and Eastman 2009; Wiseman and Chatterjee 2010; Garris and Wilkes 2017). The general question of interest is whether players deserve their salaries as depicted by their performance statistics.

Specifically for the NBA (National Basketball Association), Sigler and Sackley (2000) studied the task of salary prediction using data from the 1997–1998 season but with only three predictor variables, rebounds, assists, and points, per game. Ertug and Castellucci (2013) gathered from the 1989–1990 up to the 2004–2005 period and related the players salaries with a set of predictor variables, most of which were

I. Papadaki · M. Tsagris (✉)
Department of Economics, University of Crete, Heraklion, Greece
e-mail: mtsagris@uoc.gr

M. K. Terzioğlu (ed.), *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies*, Contributions to Economics,
https://doi.org/10.1007/978-3-030-85254-2_25

not related to the players' performance on court. More recently, Xiong et al. (2017) performed a similar analysis using more predictor variables measuring the players performance on court for the 2013–2014 season. Sigler and Compton (2018) studied the 2017–2018 season linking the salaries with more predictor variables exposing the players abilities on court.

Based on these (and other) papers, several questions emerge. How can we decide on the predictor variables used to estimate/predict the player salaries? Secondly, are the results obtained from such analyses valid and reliable enough? For instance, can a value of the coefficient of determination as high as 0.6 or 0.7 be a sign of correctness or even suggest that the analysis was successful? We further emphasize that the relationship between the players statistics and their salaries is non-linear and hence linear models are bound to fail in capturing the underlying true association. An additional concern, separate from non-linearity, is model predictability for which internal evaluation has limitations and leads to an over-optimistic performance. These and more matters, discussed later, require delicate treatment which, if not properly addressed, will yield erroneous results.

We deviate from the beaten tracks by first selecting the statistics or determinants with the highest effect on the player salaries. The detection of the important statistics that incorporate the highest amount of information on the players' salaries. We advocate against the use of all available statistics and strongly encourage researchers to apply variable selection algorithms. Not only is it important to determine the appropriate statistics, but it is also crucial for the predictive performance of the models or algorithms applied. We must further decide whether a single set of statistics (*Per game*, *Per 36 min*, etc.) or their combinations contain the highest amount (in linear terms) of information about the salaries and whether feature construction improves our predictions. Do *Advanced* statistics contain more information about the players than the *Per game* or the *Per 36 min* statistics? In either case, do we really require all sets of statistics or a subset of them? Selecting the appropriate statistics, not only removes the noise from the data, but also gives a better insight into the problem.

What can we say about the relationship between the given statistics and the players' salaries? How accurate can our salary predictions be? The second step is to apply sophisticated models to the selected statistics in order to predict the NBA player salaries. Researchers ordinarily apply linear or generalized linear models (e.g., logistic regression), or cluster and discriminant analysis. The downside of these models is their narrow abilities to capture non-linear components when the relationships among the variables are far from linear. Discriminant analysis, for instance, with unequal group covariance matrices, is non-linear but tied to a quadratic function. This gives a higher degree flexibility yet is not flexible enough to capture the associations of interest.

In the next section, we describe the problem of NBA player salary prediction and provide information on the available statistics, how we pre-processed and "cleaned" the data and set the goal of this paper. In Sect. 2, we adumbrate improper approaches that are frequently followed. For instance, employment of linear models or erroneous application of the aforementioned non-linear algorithms to real and simulated data. We describe the tools used for this purpose in the same section; the variable selection we used to select the appropriate statistics and the machine learning algorithms we employed to predict the player salaries. We further show an example of a false analysis and show that all other methods have fallen in the pitfall of over-fitting. In Sect. 3, we delineate a proper approach depicting how to properly evaluate the models by using the cross-validation (CV) procedure and present the results of our analysis. We further explain why our achieved predictive performance is the highest ever achieved. We finally summarize our findings concluding the paper.

1.1 Description of the Data

The starting point of the entire process is to compile all the required information about the players' performance on court, their salaries, the team payrolls as well as other determinants that might prove useful. Our main source of data was basketball-reference.com which is broadly known for providing a great variety of reliable sports statistics. The data acquired were narrowed down to three NBA seasons, 2016–2017, 2017–2018, and 2018–2019. There were available statistics on 486, 540, and 530 players, respectively, for these three seasons.

Throughout the player statistics data accumulated, a multitude of 54 variables[1] provided a plurality of information about each players' performance *Per game*, *Per 36 min*, and *Per 100 possessions*. Those include indexes for field goals, three-point field goals and two-point field goals counted as of total number (FG, 3P, 2P), total number of attempts (FGA, 3PA, 2PA), and percentage of successful attempts (FG%, 3P%, 2P%). The index effective field goal percentage (eFG%), found exclusively on the *Per game* statistics, adjusts for the fact that a three-point field goal is worth one more point than a two-point field goal. In a similar manner, we have free throws (FT), free throw attempts (FTA), and free throw percentage (FT%), offensive rebounds (ORB), defensive rebounds (DRB), and total rebounds (TRB) per game, per 36 min and per 100 possessions. Furthermore, assists (AST), steals (STL), blocks (BLK), turnovers (TOV), personal fouls (PF), and points (PTS) were also included. Among the per 100 team possessions statistics, two more variables were incorporated, offensive rating (ORtg) and defensive rating (DRtg), which are estimates of points produced by players or scored by teams per 100 possessions and of points allowed per 100 possessions, respectively.

In the attempt to obtain a more comprehensive picture of the players' performances, we also consulted the players' *Advanced* statistics, also available on basketball-reference.com. This type of statistics displays variables such as Player Efficiency Rating (PER), a measure of per-minute production standardized so that the league average is 15, True Shooting Percentage (TS%), a measure of shooting efficiency

[1] There were some common variables though such as their age on February 1 of the season (Age), number of years they have played in NBA (EXP), the team (Tm), and the position (Pos) in which they played. Additionally, the total number of games (G) and minutes (MP) they participated and the number of games they were in the starting five (GS) were displayed on almost all types of statistics.

that takes into account two-point field goals, three-point field goals, and free throws, three-point attempt rate (3PAr), which is the percentage of field goals attempts from a three-point range and free throw attempt rate (FTr) which indicates the number of free throw attempts per field goal attempt. In addition, the percentage of available offensive rebounds, defensive rebounds, and total rebounds a player took while he was on the court is estimated using offensive rebound percentage (ORB%), defensive rebound percentage (DRB%), and total rebound percentage (TRB%), respectively.

Advanced statistics further comprise assist percentage (AST%), an estimate of the percentage of teammate field goals a player assisted, steal percentage (STL%), and block percentage (BLK%), estimates of the percentage of opponent possessions that end with a steal by the player and of opponent two-point field goal attempts blocked by the player, together with usage percentage (USG%), an estimate of the percentage of team plays used by a player and turnover percentage (TOV%), an estimate of turnovers committed per 100 plays. Moreover, we have estimates of the number of wins contributed by a player in total (win shares (WS)), due to his offense (offensive win shares (OWS)), due to his defense (defensive win shares (DWS)), and per 48 min (win shares per 48 min (WS/48)) with the last's league average being approximately 10%. We additionally incorporated the offensive box plus/minus (OBPM), defensive box plus/minus (DBPM), and box plus/minus (BPM), which are box score estimates of the offensive, defensive, and total points per 100 possessions a player contributed above a league-average player translated to an average team. Lastly, the value over replacement player (VORP), a box score estimate of the points per 100 team possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team, and prorated to an 82-game season.

Next on our data collection process, we turned to espn.com and hoopshype.com to attain the fundamental information about the players' income and the 30 NBA teams' payrolls for the seasons under investigation.[2] As far as the players' salaries are concerned, the number of available observations was 594, 598, and 503 for the 2016–2017, 2017–2018, and 2018–2019 season, respectively.

It is of major importance to take into account the teams' payroll when predicting the players' salaries, given the variation of this amount among different teams. During 2016–2017, Utah Jazz's payroll was the minimum across NBA with their contracts summing to $80 millions, whereas Cleveland Cavaliers spent the highest amount of money, $130 millions. During the 2017–2018 season, Dallas had the minimum payroll, $85 millions, whereas Charlotte Hornets had the highest payroll of $143 millions with Cleveland Cavaliers having second highest, $137 millions. In our latest season, 2018–2019, Atlanta Hawks had the minimum payroll, whereas Miami Heat had the highest payroll, equal to $79 millions and $153 millions, respectively. Markedly, had Miami Heat's best player signed with Atlanta Hawks he would earn

[2] Hoopshype.com provided us with the choice between the absolute nominal value of each team's payroll or the payroll adjusted for inflation based on the current year (from data provided by the U.S. Department of Labor Bureau of Labor Statistics), to which we chose the first for the sake of correspondence between the base years on affiliated monetary values.

around 65% of his Miami salary, and conversely if Atlanta Hawks' best player was traded to Miami he should get 150% times his current salary.

Sigler and Compton (2018) signified that a player's years on the league is a determinant of equal importance for his salary with his performance, as depicted by his statistics. The maximum amount of salary a player can receive is related to the number of years he has played in the NBA and the amount of the salary cap. During the 2017–2018 season, the maximum salary of a player who had at most 6 years of experience was either $25,500,000 or 25% of the total salary cap, whichever was greater. For a player with 7–9 years of experience, the maximum increased to $30,600,000 or 30% of the salary cap, and for a player with more than 10 years of experience, his maximum contract could reach $35,700,000 or 35% of the salary cap. However, a player can sign a contract for 105% percent of his previous contract, even if the new contract is higher than the league limit. Having said this, we made use of stats.nba.com to include each player's experience (number of years in the NBA league) in our inquiry.

1.2 Cleaning and Pre-processing the Data

The volume of data accumulated needed to be merged into a unified database that would serve the purpose of our analysis. The objective of this process was to associate each player's wage with their statistics, their years of experience, and their team's budget. However, we came across cases of missing information throughout the different sources. For example, there was no available salary or experience for some of the players listed on the statistics database and vice versa. To solve this problem, we solely kept the observations in which we had all three types of information at our disposal. Moreover, some of the players switched teams within the season and, as a result, they were recorded several times on the statistics database, once for every team. In this instance, we preserved the statistics exclusively for the team for which we had information about the salary. On account of better results, it was also deemed necessary to discard all players who participated in less than 10 games during each season, in view of the fact that those observations' contribution to our model's predictability was actually a drawback. Throughout this process of "data cleaning," the remaining observations were 443 players for the 2016–2017 season, 484 players for the 2017–2018 season, and 412 players for the 2018–2019 season. These are considered to be adequate sample sizes.

To make our predictions payroll free, instead of using the nominal wage for each player as the dependent variable on our regression models, we constructed a new variable, the ratio of the player's salary to his team's payroll. This is the players salary share, which will later be used as the dependent variable on our models. The sum of the player salary shares for each team ought to be ≤ 1 and in cases it exceeded 1 it was decided to replace the team's payroll with the aggregation of the team's players' salaries at hand.

1.3 Other Possible Determinants of Salaries

It can be argued that NBA player salaries are not only related to their performance on court but also to publicity and reputation (Ertug and Castellucci 2013). Reputation though is difficult to measure for all players, counting, for example, players' followers in social networks, their contracts with sports companies, promoting activities, etc., and we thus have avoided it.[3] Other contributing factors include player managers that can make hard negotiations with team managers and can achieve higher earnings for their clients. We assert that these factors are projections of the players' image on the court. Highly skilled (and regularly spectacular) players are those who will ordinarily sign contracts with sport companies and will be interviewed more frequently and promoted more, by sports journalists.

A second factor is discrimination, either racial (Kahn and Sherer 1988; Hamilton 1997; Kahn and Shah 2005; Rehnstrom 2009; Wen 2018), nationality-wise (Yang and Lin 2012; Hoffer and Freidel 2014); or exit (Groothuis and Hill 2013). Our personal view is that any alleged discrimination present is fully justifiable by the players' performance. African-American players have better physical skills and are more athletic, which facilitates the quantity of spectacle they offer compared to other players. If those players receive higher salaries simply because they may have better statistics or have a more spectacular type of play, this is by no means evidence of race or country discrimination. Further, foreign players, e.g., Europeans have nourished in a different mentality. American basketball is more athletic than European and usually Europeans require more adjustment time than players drafted from the NCAA. It is perceptible that athletic and physical abilities and the mentality of basketball has caused this alleged discrimination. Investigation of this entails a comparison of the player performance between African-American and white American players and between American and European players, but this is outside the scope of this paper. The same rule applies for the exit discrimination (Groothuis and Hill 2013) who concluded that more athletic players have a higher survival rate in the NBA. We close the discrimination matter by referring to Groothuis and Hill (2013) who used a panel dataset from 1990 to 2008 and failed to find any evidence of either pay or exit discrimination in the NBA.

A third possible determinant factor is TV contracts,[4] which were deemed as not important by Kelly (2017). Kelly (2017) applied a linear regression model where a subset of the TV contracts relevant variables were statistically highly significant, yet the goodness of fit of the model was very poor.[5]

[3] For example, to measure a player's level of spectacle we would have to collect the number of dunks, the number of alley-hoops, the number of fake movements, the number of ankle-breaking phases, or any other spectacular movements.

[4] TV contracts contribute to the NBA revenues which determines the salary cap.

[5] This is another case that exemplifies why non-linear models are necessary to yield more accurate predictions.

2 Salary Prediction

We will now illustrate some incorrect approaches that are ordinarily followed by researchers. Subsequently, we describe the pipeline for selecting the most important performance determinants that affect the player salaries and how to make the most of the predictive capabilities of those determinants.

2.1 Criticism of Some Current Approaches

Sigler and Sackley (2000) related some player statistics (points, rebounds, and assists per game) with the player salaries using a linear regression model and computed an unsurprisingly low coefficient of determination (R^2). Ertug and Castellucci (2013) used a linear regression model to estimate the team revenues attributed to ticket sales computing high R^2 values for two models, 0.75 and 0.77. When it came to estimating player salaries, their linear models had a low fit though ($R^2 = 0.30$ and $R^2 = 0.31$).[6]

The fact that researchers do not select the important determinants prior to fitting the model is an example of what not to do. Ertug and Castellucci (2013), for example, in their seemingly optimal model for the team's ticket-based revenue used 13 variables, out of which only four variables were statistically significant and two of them were highly statistically significant. Retaining the other nine variables in the model does not add but removes value from it in a threefold manner. (a) It is known that the addition of variables in the model leads to higher R^2 values. Thus, the reported value of 0.77 for the R^2 is an over-estimate of the true R^2 of their model. (b) This practice makes the model unnecessarily more complex and (c), in fact, deteriorates the predictive performance of the model. This is associated with the curse of dimensionality (Hastie et al. 2009) and is the main reason why variable selection is necessary, to remove the irrelevant variables that add noise and no information. As an analogue of this task we refer to national teams who select their best among the all-star players when participate in international championships (continental and universal).

Attempting to model the non-linear relationships of the statistics with the player salaries via adoption of a linear model is a policy that should be avoided. A preferable strategy would be to add of square terms in some variables, as this may improve the performance of the model, but perhaps not significantly. Linear models will encapsulate, to some degree, the trend in the variables, but they definitely cannot be used for safe prediction. The following example from basketball suits to convey our message. NBA teams select the most talented rookie players, but solely talent is not enough. It is training that will take those players to the next level and the better the "material" in a team hands the higher its chances to win the championship.

Assessing the goodness of fit of a model via the R^2 is a criticism raising strategy. The performance of a model that has been constructed on some data must be tested

[6] In the game of hockey, Vincent and Eastman (2009) applied a quantile regression instead, a robust to outliers regression model, but still linear.

on different data that the model has never "seen." During training, players test their abilities against one another, but soon they understand each other's play and perhaps some players perform very well, but only during practice. Players are not getting paid to play well during practice with their teammates, but to play well against new players whose team systems or play they have not seen. Players are always evaluated externally and not internally.

Not all researchers though fall into the aforementioned pitfalls.[7] Wiseman and Chatterjee (2010) performed a variable selection procedure in order to predict the American League Baseball player salaries and also included a quadratic effect in the years of major league service. However, reporting an internal (and hence over-optimistic) R^2 as high as 70.1% was an incorrect decision made by those researchers.

2.2 Variable Selection and Prediction Algorithms

To further assist the comprehension of the analysis, we will narrate the LASSO (Least Angle Selection and Shrinkage Operator) variable selection and the Random Forest (RF) algorithm.[8]

2.2.1 Least Absolute Shrinkage and Selection Operator

The Pearson correlations between NBA player salary shares and performance measures were deemed statistically significant, but not all of them remain significant when all predictor variables enter a regression model. The LASSO algorithm (Tibshirani 1996) will facilitate the selection of the most important performance measures.

LASSO is a regression model that simultaneously performs variable selection and regularization of the relevant coefficients. It improves the predictive performance by shrinking the regression coefficients so as to reduce over-fitting, and performs variable selection by setting some of them to zero hence discarding variables that are responsible for large variance, therefore making the model more interpretable. LASSO minimizes the following penalized sum of squares:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|, \tag{1}$$

[7] Criticizing all available papers is outside the scope of this paper and hence we do not pursuit this further.

[8] We performed the analysis using the open-source software *R* (R Core Team 2020). LASSO is implemented in the R package *glmnet* (Friedman et al. 2010), while RF is implemented in *ranger* (Wright and Ziegler 2017).

where y_i is the i-th response value, x_{ij} denotes the i-th value of the j-th predictor variable, n denotes the sample size, and p is the number of predictor variables. Fine-tuning of the penalty parameter λ is essential since it determines the amount of regularization, the strength of shrinkage, and, ultimately, the number of variables selected for inclusion in the final model. Such is achieved through a cross-validation procedure, where the value λ yielding the lowest estimated prediction error is preferred.

2.2.2 Random Forests

The RF algorithm is a fast and flexible data mining approach well suited for high-dimensional data. The algorithm is built upon creating many classification or regression[9] trees. According to Breiman (2001), RF randomly draws a subset of variables and a bootstrap sample[10] and uses only this subset of features to grow a single tree. The process of randomly selecting variables and bootstrap samples is repeated multiple times and the results are aggregated. By creating many random trees (500 or 1000, for instance), one ends up with a random forest.

As stated in the Introduction, the relationship between the player salaries (shares) and their performance on court is not expected to be linear, hence the RF algorithm will allow us to capture the non-linear components of this relationship.

2.3 A Note on the Response Variable

We stated earlier that we converted the player salaries into (payroll free) percentages that are on the same basis for everyone. The implications of this transformation to LASSO, which employs a linear regression model, hence a normal distribution, are obvious. Unlike the normal distribution whose support is unbounded, the percentages lie within a restricted range of values. Additionally, predictions with LASSO are not constrained to lie within that plausible range. Not correct specification of a distribution or of a regression that takes into account the space where the response variable is defined is another frequent mistake of researchers and practitioners.

We refrained from using the salary shares in LASSO and transformed the response prior to employing LASSO using the logit transformation $y^* = \log \frac{y}{1-y}$, where y denotes the player salary shares. The logit transformation is well defined when $y \neq 0$ and $y \neq 1$, which holds true in our case. This transformation is not obligatory when RF are used since the predicted values are in fact weighted averages of the observed values (Lin and Jeon 2006).

[9] Depending on the nature of the response variable.

[10] Sample with replacement of the same size.

2.4 Distribution of the NBA Player Salary Shares

Figure 1 shows the kernel density estimates of the distributions of the salary shares for each season. The differences are rather small and indeed there is no evidence to support that the distributions vary statistically significantly across the three seasons (p-value = 0.9263).

2.5 Internal Evaluation in Our Datasets

We now illustrate the internal evaluation of RF in our datasets and manifest the over-rated performance they seem to be possessing. We standardized our predictor variables prior to the analysis in order to transform all variables into the same scale.[11] We implemented the tenfold CV procedure (described in the next section) to tune the penalty parameter (λ) of LASSO. We then performed LASSO penalization using the chosen value of λ to select the most important factors which were plugged into the RF algorithm.

We evaluate the predictive performance of RF by contrasting each set of predictions (one set for each hyperparameter) against the true salary shares using the Pearson correlation coefficient (PCC) and the percentage of variance explained (PVE).[12]

$$PCC = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \left(\tilde{y}_i - \bar{\tilde{y}}\right)}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} \left(\tilde{y}_i - \bar{\tilde{y}}\right)^2}} \tag{2}$$

$$PVE = 1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{3}$$

where $\tilde{y}_i$ refers to the predicted value of the i-th observation and $\bar{y}$ denotes the mean value. The PVE values for each set of statistics appear in Table 1. We also report the PVE values of LASSO as a comparison of the performance of a linear and of a non-linear algorithm.

Unlike the mean squared error (MSE) or mean absolute error (MAE) the aforementioned metrics have a benchmark value to compare against. The raw values of MAE, or MSE, do not reflect the performance of the model relative to model-free average predictions. On the contrary, for both PCC (2) and PVE (3) their maximum value is 1 indicating excellent predictive performance, whereas the minimal value equal of 0 refers to completely random predictions. Higher values of PCC (2) indicate a higher number of correct model-based predicted orderings, whereas higher values

[11] We explained why this pre-processing step is incorrect in Sect. 3.3.

[12] PVE is equivalent to R^2 for the linear models.

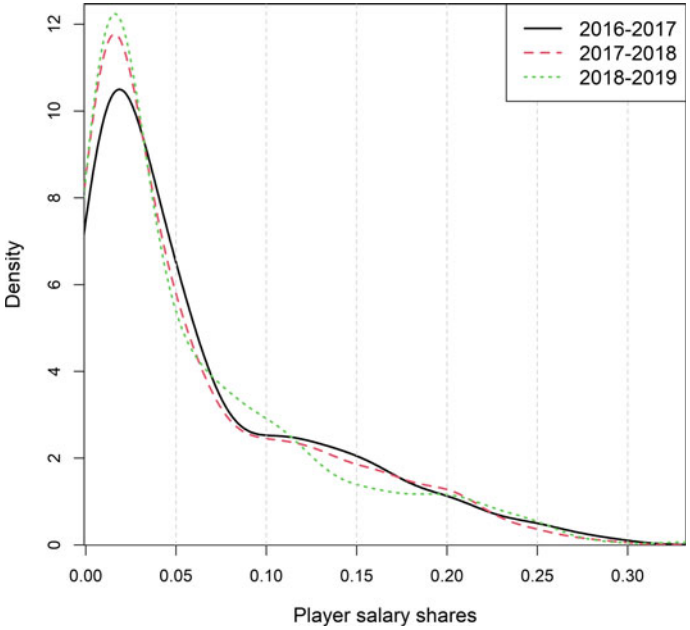


Fig. 1 Kernel density estimates of the NBA player salary shares across the three seasons

Table 1 PVE values for each set of statistics and each algorithm across the three seasons

Season	2016–2017		2017–2018		2018–2019	
Statistics	RF	LASSO	RF	LASSO	RF	LASSO
Per game	0.910	0.420	0.869	0.246	0.884	0.220
Per 36 min	0.901	0.284	0.866	0.223	0.826	0.210
Per 100 possessions	0.900	0.305	0.866	0.223	0.826	0.164
Advanced statistics	0.848	0.163	0.842	0.139	0.862	0.181

of PVE (3) indicate that on average the errors of the model-based predictions are much less than the errors of random, model-free predictions.

We used the PVE[13] (3) for model assessment and perhaps the only safe conclusion we can draw from Table 1 is that non-linear models have superseded the linear model of LASSO. RF always produced PVE values above 0.8 (or 80%) indicating an excellent fit. If we compared these PVE values against the R^2 values reported in previous papers we would be delighted, not only because we outperformed their fit, but also because our PVE values are remarkably high. We will repeat ourselves

[13] The predicted values of LASSO were first back-transformed to percentages using the inverse of the logit transformation, $y = \frac{1}{1+e^{-y*}}$, and y was used to compute the PVE.

that the cost of this high PVE is interpretability. RF does not produce a coefficient for each predictor variable that could reflect the variable's (marginal) effect on the salary shares.

2.6 Model Complexity

The last, but equally important, point to take into account is model complexity. Fitting a highly complex non-linear model does not necessarily yield better prediction. To demonstrate this, we expose below a short script written in R[14] evaluating, internally, a model's performance. We randomly generated a set of 20 predictor variables and a random response variable. We then applied a non-linear model (projection pursuit regression[15]), where each time we increased the complexity of the model and computed the PCC (2) between the observed and fitted values.

```
set.seed(12345)
## generate random predictors
x <- matrix( rnorm(400 * 20), ncol = 20 )
## generate random response
y <- rnorm(400)
pcc <- numeric(10)
for (i in 1:10) {
  ## perform projection pursuit regression
  mod <- ppr(y ˜ x, nterms = i)
  ## compute PCC
  pcc[i] <- cor( fitted(mod), y)
}
round(pcc, 3)
0.443 0.575 0.644 0.666 0.748 0.736 0.844 0.933 0.860 0.950
```

Evidently, the PCC between the observed and fitted values increases with model complexity. Further, surprisingly enough, we managed to obtain a high level of correlation when in fact there is no relationship between the response and the predictor variables. This again points out that an internal evaluation draws no safe conclusions as over-fitting occurs. A second source of complexity comes from the fact that we used all available 20 predictor variables and not a subset of them. This is an extra reason why we should have performed variable selection prior to estimating the predictive performance. Penalizing for complexity, e.g., via Bayesian Information

[14] The example is reproducible and will always yield the same results.

[15] We tried this model in our analysis but the results were not that accurate and hence we omitted them.

Criterion could have avoided this phenomenon, but even then, internal evaluation would over-estimate the true performance of the model.

3 A More Valid Approach

The previous example indicates how we can get trapped in over-fitting. The reported PVE values refer to the internal evaluation of the models because these are internal PVE values. We will elucidate the correct way to estimate a model's predictive performance (external evaluation) using the k-fold CV protocol. To obtain an unbiased estimate of the predictive performance we need large sample sizes, a condition we meet because we have information on hundreds of players at each season. Finally, we will demonstrate that the observed performance metrics in Table 1 are actually very high and far from reality.

3.1 The k-Fold CV

The k-fold CV protocol splits the data into k mutually exclusive groups, termed folds. The ordinary value of k, which we also used, is $k = 10$, yielding the 10-fold CV. We select one fold and leave it aside to play the role of the test set. The remaining nine folds are combined into what is called the training set. We standardize the predictor variables of the training set only. We then perform variable selection using LASSO and feed the RF algorithm with the selected variables. We use the same selected predictor variables from the test set and we scale them using the means and standard deviations of the same predictor variables from the training set. We use these scaled predictor variables of the test set to predict the values of the response variable (player salary shares) of the test set. For RF we used a range of splits of variables,[16] thus, we end up with multiple predictions, one set of predictions for each hyperparameter, whose predictive performance we compute.

We subsequently select another fold to play the role of the test set and insert the previous fold (previous test set) into the training set and repeat the pre-described pipeline. The process is repeated until all folds have played the role of the test set. In the end, we compute the average predictive performance of RF with each hyperparameter and choose the hyperparameter that yields the highest predictive performance.

[16] These are termed hyperparameters and need to be tuned.

3.2 The Essence of CV

The importance and necessity of any CV protocol can be further appreciated through an investment example. Assume an NBA team manager or team owner who wishes to invest their money on some market, stock exchange, mutual or pension funds, real estate, etc. There are two available investment companies residing in the building right next to his/her. Company A has a long record of remarkably high PVE values in their models. The company gathers the prices spanning from several days ago up to today and fits variable selection and machine learning algorithms and computes the PVE values of the models/algorithms using the same data. It shows no record of predicting future prices though. Company B, on the other hand, applies a different strategy. It again uses historical data, but keeps the old ones for model building and training and treats the most recent ones as the future that must be predicted. The PVE values (of the future predictions) of company B are, perhaps significantly, lower than the PVE values (of the past and present predictions) of company A but are safer predictions of the future. Company A implements the wrong approach described in Sect. 2.1, whereas company B implements the correct strategy described in Sect. 3.1.

3.3 The Importance of Processing the Data in the Training Set

CV can be seen as a simulation of realistic scenarios. Let us denote the training set by *present* and the test set by *future*. We observe the present and attempt to predict the future. We process (standardize) the data in the training set (present) and use those means and standard deviations to scale the test data (future). Had we standardized the data from the beginning would deviate from the realistic scenario as we would have allowed information from the future to flow into the present. Thus, attempting to transform the data into the same scale prior to performing any CV protocol is erroneous and should be avoided.

But why is standardization so important? Numerous variables listed on the *Per game*, *Per 36 min*, and *Per 100 possessions* refer to percentages therefore deviate between 0 and 1, games played (G) can reach values as high as 82 and players' ages vary between 19 and 42. Furthermore, three-point field goals per 100 team possessions (3P), for example, span between 0 and 7.2 and total rebounds per 100 team possessions (TRB) between 3.0 and 23.8. Likewise, the majority of the *Advanced* statistics are estimates of percentages, while Win Shares (WS) are measured on a scale of -1.7 to 15.4 and Box Plus/Minus (BMP) of -5.7 to 11.1, just to name a few. Standardization is a necessary processing strategy in order to prevent our results from being strongly affected by the scale of measurement of the variables.[17]

[17] This is the reason why VS algorithms, such as LASSO, require standardized data.

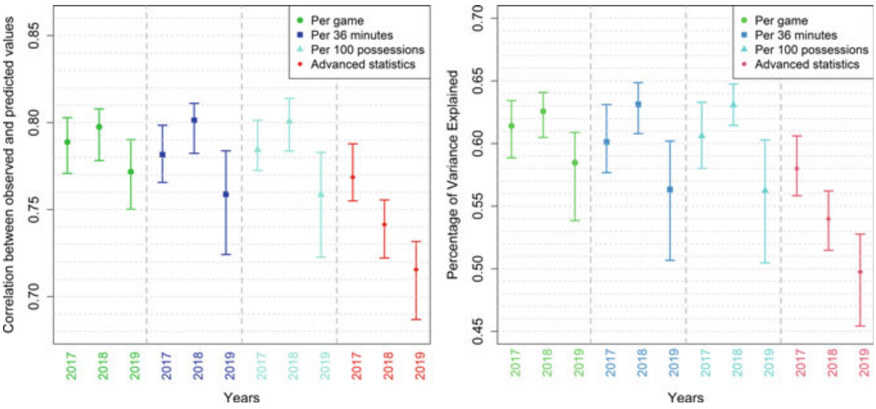
3.4 Results of the 10-Fold CV Protocol

Unarguably partitioning the data into 10-folds contains an inherent variability as different partitions will give different results. To robustify our inference against this uncertainty, we repeated the 10-fold CV procedure (variable selection and predictive performance estimation) 50 times and report the aggregated predictive performance of the RF algorithm.

Figure 2 contains the average PCC (2) and PVE (3) for every season using either set of statistics (*Per game*, *Per 36 min*, *Per 100 possessions*, and *Advanced* statistics), along with the corresponding 95% confidence intervals. Overall, use of the *Advanced* statistics resulted in the worst performance among all datasets while the *Per 36 min* and *Per 100 possessions* portrayed a very similar picture, perhaps due to the fact that LASSO was selecting the same statistics. The *Per game* statistics evidently gave the optimal predictions overall with an exception for the season 2017–2018, whose predictions ranked second best with the difference being tiny. The *Per game* statistics also dominated in terms of variance of the predictive performance. The length of the 95% confidence intervals for the true predictive performances are always the shortest, indicating higher stability.

There is a common pattern among the first three sets of statistics. We observe an increase in the predictability as we move from the 2016–2017 to the 2017–2018 season which then decays as we move to the 2018–2019 season. Further, in the last season, we observe the highest variability in the predictive performance and the confidence intervals are the widest observed across the three seasons.

Tables 2 and 3 present the optimal (average) predictive performances of the RF using each set of statistics across the three seasons when LASSO variable selection has been applied prior to RF and when all statistics were fed into the RF. The PCC values are remarkably high, lying in the range of 0.7–0.8. The PVE values are lower,



(a) PCC (2)

(b) PVE (3)

Fig. 2 Predictive performance metrics using each set of statistics across the three seasons

Table 2 PCC values for each set of statistics across the three seasons

	With LASSO			Without LASSO		
Statistics	2016–2017	2017–2018	2018–2019	2016–2017	2017–2018	2018–2019
Per game	0.789	0.798	0.772	0.783	0.794	0.758
Per 36 min	0.781	0.801	0.759	0.772	0.781	0.750
Per 100 possessions	0.784	0.801	0.759	0.769	0.778	0.747
Advanced	0.769	0.741	0.716	0.742	0.752	0.718

Table 3 PVE values for each set of statistics across the three seasons

	With LASSO			Without LASSO		
Statistics	2016–2017	2017–2018	2018–2019	2016–2017	2017–2018	2018–2019
Per game	0.614	0.626	0.585	0.600	0.616	0.561
Per 36 min	0.601	0.631	0.563	0.577	0.589	0.540
Per 100 possessions	0.606	0.631	0.562	0.572	0.584	0.534
Advanced	0.580	0.540	0.498	0.534	0.548	0.499

as expected, yet these figures are high in comparison to prior research and most importantly, they were produced by external and not internal evaluation. Further, these two tables clearly visualize the essence of variable selection prior to using RF. The predictive performance changes with and without variable selection changes slightly, but the advantage of LASSO is that it identifies the most important statistics, presented in Table 4.

Table 4 contains the statistics that were most frequently selected by LASSO throughout the 50 repetitions of the 10-fold CV. Overall, experience and minutes played of each player were the two statistics that were always selected regardless of the set of statistics and the year of play. The third statistic was either the games played or the games started, followed by the points, the defensive rebounds, and the field goals attempted. In the Advanced statistics, the USG (an estimate of the percentage of team plays used by a player) and OBPM (box score estimate of the offensive, defensive, and total points *Per 100 possessions* a player contributed above a league-average player translated to an average team). Excluding the *Advanced* statistics, we can see that there seems to be a stability in the selected statistics across the three seasons. In the *Per game*, the last season only substitutes the games played, the filed goal attempts, and the defensive rebounds with the points scored. A common feature with the *Per 36 min* is that defensive rebounds do not seem to play a significant role in the last season. When it comes to the *Per 100 possessions*, defensive rebounds never seem to contribute to the salary of the players.

Table 4 Most important statistics per set of statistics across the three seasons

Statistics	2016–2017	2017–2018	2018–2019
Per game	EXP, MP, G, FGA, DRB	EXP, MP, G, FGA, DRB	EXP, MP, PTS
Per 36 min	EXP, MP, GS, DRB, PTS	EXP, MP, GS, DRB, PTS	EXP, MP, GS, PTS
Per 100 possessions	EXP, MP, GS, PTS	EXP, MP, GS, PTS	EXP, MP, GS, PTS
Advanced	EXP, MP, USG, OBPM	EXP, MP, USG, OBPM	EXP, MP, USG, OBPM

The use of *Advanced* statistics did not yield better results than the use of the *Per game* statistics, in fact, the former dataset produced the worst results. We highlight that the PER index is included in the *Advanced* statistics.

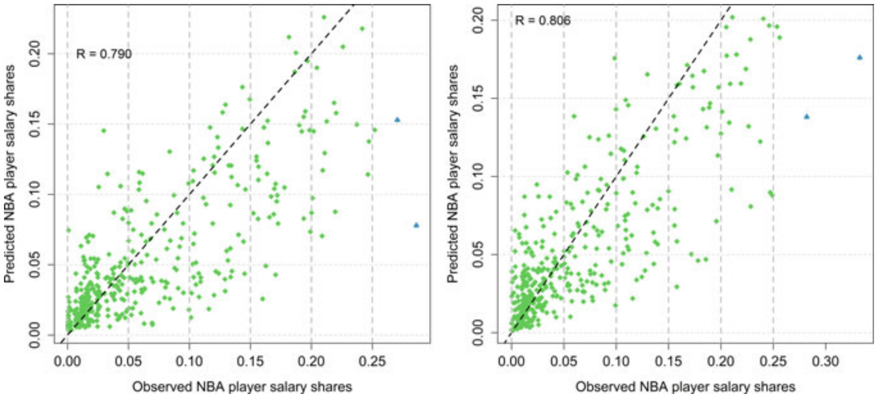
3.5 Testing the Predictability of the RF Algorithm

In order to show the validity of the PCC and PVE values reported in Tables 2 and 3, respectively, we used the identified statistics from the 2016–2017 season, fitted an RF, and predicted the salary shares of the 2017–2017 season. We repeated the same task using the statistics from the 2017–2018 season to predict the salary shares of 2018–2019. This way the next season's data played the role of the validation set, a new dataset that the algorithms never "saw" during the CV protocol. The PVE values were 0.624 and 0.650, respectively, while the PCC values were equal to 0.790 and 0.806, respectively. The observed and the predicted salary shares are displayed in Fig. 3.

3.6 NBA Player Salary Share Classes

Let us now provide some in-depth statistics regarding the player salaries. Table 5 shows the distribution of the player salaries across the three seasons.

During the 2016–2017 season, the six highest paid players (in terms of team's payroll share) were James Harden (Houston Rockets point guard, 29.18%), Al Horford (Boston Celtics center, 28.40%), Russell Westbrook (Oklahoma City Thunder point guard, 26.75%), Kevin Durant (Golden State Warriors power forward, 26.13%), Brook Lopez (Brooklyn Nets center, 25.69%), and Dwyane Wade (Chicago Bulls shooting guard, 25.08%). Among them, James Harden was second in the points per game (29.1), Russell Westbrook and Kevin Durant's statistics justify their salaries. Surprisingly enough, Al Horford's statistics do not match his salary, as he was scor-



(a) 2017-2018

(b) 2018-2019

Fig. 3 Observed versus predicted player salary shares for 2017–2018 and 2018–2019

Table 5 Distribution of player salaries across the three seasons

	Player salaries in percent of the team's payroll					
Season	[0, 5%)	[5, 10%)	[10, 15%)	[15, 20%)	[20, 25%)	[25, 30%]
2016–2017	239	73	48	32	17	6
2017–2018	257	66	49	33	17	3
2018–2019	231	75	39	25	18	4

ing 14 points per game despite playing 32 min. The same is true for brook Lopez and Dwyane Wade whose statistics are rather low.

During the 2017–2018 only three players received more than 25% of the team's payroll, Paul Millsap (Denver Nuggets power forward, 28.61%), Harrison Barnes (Dallas Mavericks power forward, 27.05%), and Stephen Curry (Golden State Warriors point guard, 25.20%). Stephen Curry was scoring an average of 26.4 points per game, whereas Paul Millsap and Harrison Barnes were as low as 14.6 and 18.9 points per game, despite playing 30 and 34 min per game, respectively.

The four highest paid players (in terms of salary shares) for the last season, 2018–2019, were Lebron James (Los Angeles Lakers small forward, 33.25%), Chris Paul (Houston Rockets point guard, 28.19%), Stephen Curry Golden State Warriors point guard, 25.60%), and Blake Griffin (Detroit Pistons, power forward, 25.36%). Chris Paul was the only one among those 4 to score less than 20 points per game (15.6) even though he was playing 32 min per game. He was giving 8.2 assists per game and stealing the ball 2 times per game, yet these statistics do not match that large salary share.

The aforementioned players were evidently receiving a remarkably high share of their team's payroll, more than a quarter. Lebron James received an excessively high share, more than a third of Los Angeles Lakers' payroll, during the 2018–2019 season. We have no evidence to conclude that the highest paid players belong to the champion team. Kevin Durant won the NBA championship with the Golden State Warriors in 2017 and Stephen Curry was a member of the same team that won the championship in 2018. Toronto Raptors won the championship, but their best player[18] is not in the aforementioned list. These players are not the best among NBA and this small piece of information markedly shows that salaries are not always affected by statistics, hence partially explaining why salary prediction is hard to do with only performance statistics.

The blue triangles in Fig. 3a correspond to Harrison Barnes (up) and Paul Millsap (down) indicating that RF predicted correctly that their salary shares should be lower than what they actually received. The blue triangles in Fig. 3b correspond to Lebron James (up) and Chris Paul (down). According to RF, Lebron James was rather overpaid during that year, whereas Chris Paul was evidently over-paid as depicted by his statistics.

3.6.1 Salary Share Class Prediction

Table 5 transparently presents that most NBA players receive a small percentage (at most 5%) of the team's payroll. This led us to the second part of our analysis that of discriminating between the low and the higher paid players. To this end, we employed the LASSO and RF algorithms again. In this scenario, LASSO selects the most appropriate statistics by minimizing a more appropriate penalized function

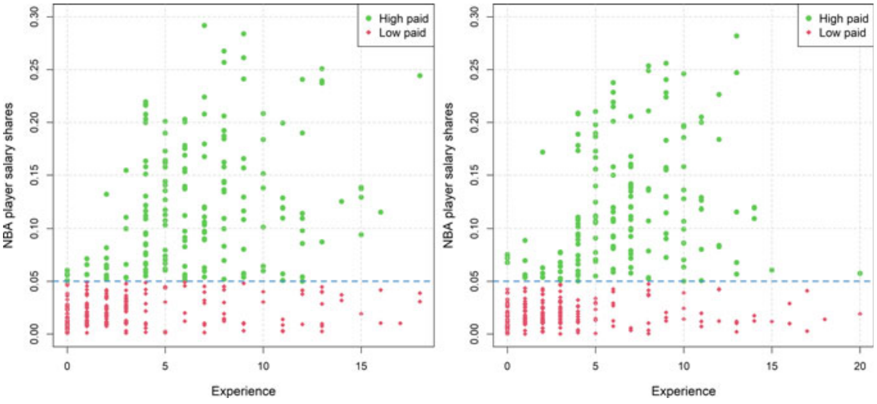
$$\sum_{i=1}^{n}\left[C_i\sum_{j=1}^{p}\beta_j x_{ij}-\log\left(1+e^{\sum_{j=1}^{p}\beta_j x_{ij}}\right)\right]+\lambda\sum_{j=1}^{p}|\beta_j|, \tag{4}$$

where C_i takes two values, 0 and 1 corresponding to players receiving lower or more than 5% of their team's payroll, respectively.

Having mentioned earlier that the number of years in the NBA affects the player salaries we visualize their relationship in Fig. 4.[19] Their relationship is clearly non-linear, the Pearson correlations are rather low (0.45 and 0.42, respectively) and there is no apparent threshold to separate the low from the highly paid players. We cannot visually distinguish, in a straightforward manner, the low from the highly paid players. Further, broadly speaking, there is tendency for the salaries to increase, as expected, with thee years of service in the league but percentage-wise this is not true for all players.

[18] Kawhi Leonard was receiving 16.78% of Toronto Raptors' payroll.

[19] We present this information for the 2016–2017 and 2018–2019 seasons only, due to space limitations. The scatter plot for the 2017–2018 season, and the scatter plots for the number of games played and the number of games the players were in the starting five were similar and hence omitted.



(a) 2016-2017 season

(b) 2018-2019 season

Fig. 4 Player salary shares against the number of years in the NBA

3.6.2 Assessment of the Classification Task

We utilized the Area Under the Curve (AUC) to evaluate the classification performance. In this context, AUC shows the probability of correctly classifying a player to the class or group (low or highly paid players) he belongs to. AUC lies within 0 and 1, where 0.5 denotes random assignment. In contrast to the proportion of correctly classified players, AUC is not affected by the distribution of the two groups.

We implemented the same (50 times) repeated 10-fold CV protocol and present the results in Fig. 5. Once again, the *Per game* statistics resulted in the optimal predictive performance, for which the average AUC was always greater than 0.80, whereas the *advanced* statistics yielded the lowest predictive performance. To appreciate the significance of this high value, we can give the following interpretation. The years of experience of a player in the league and the average number of minutes he played for a given season allow to classify him to the low- or high-paid group with a probability equal to 0.8.

In terms or the selected statistics, LASSO was consistently selecting the same statistics as can be seen in Table 6. The number of years the players in the NBA, the average minutes they played in each game, and the number of games they were in the starting five were the most important statistics throughout the datasets and the three seasons.

As a second, validation, step we used the selected statistics from the 2016–2017 season, namely, the number of years in the league and the minutes played, fed them into an RF using the statistics of 2017–2018, and predicted the salary share class of that season. We repeated this task to predict the salary share classes of the 2018–2019 season. The reasoning behind is to test the algorithm's ability to predict the next season's salary share classes. The AUC values for the 2017–2018 and the 2018–2019 predictions were 0.811 and 0.841, respectively, corroborating the results of the CV process.

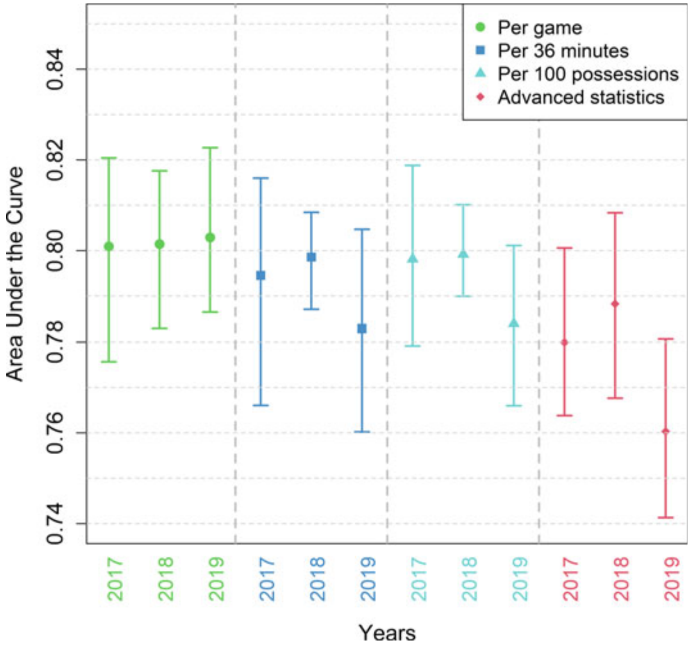


Fig. 5 AUC using each set of statistics across the three seasons

Table 6 Most important statistics per set of statistics across the three seasons

Statistics	2016–2017	2017–2018	2018–2019
Per game	EXP, MP	EXP, MP, GS	EXP, MP
Per 36 min	EXP, MP, GS	EXP, GS	EXP, MP, GS
Per 100 possessions	EXP, MP, GS	EXP, GS	EXP, MP, GS
Advanced	EXP, MP, OBPM	EXP, MP, WS	EXP, MP, WS

3.7 Further Analysis

We further performed other variable selection algorithms (γ-OMP, Tsagris et al. 2020) and non-linear prediction algorithms such as projection pursuit (Friedman and Stuetzle 1981) and $k - NN$ (Altman 1992) but their results were sub-optimal and hence omitted. Another strategy was to construct more variables for each dataset, such as square and cubic transformation of each variable, along with all pairwise products of the variables. A second strategy was to combine all variables and the third strategy was to use all variables for each dataset and ignore the variable selection phase. None of these strategies improved the predictive performance of the RF.

4 Conclusions

The relationship between NBA player statistics and their salaries (expressed as percentage of the team's payroll) is evidently non-linear and we showed the necessity to apply non-linear models and algorithms. Using real and simulated data we showed the erroneous decisions that can be made when applying linear models. We demonstrated that non-linear models will yield over-optimistic results when they are internally validated. We then described the correct approach to investigate the relationship between a response and many predictor variables and how to correctly estimate a model's predictive performance.

Using the LASSO variable selection we managed to detect the important factors (statistics) that are mostly associated with the NBA player salaries and utilizing the RF non-linear algorithm we predicted the player salaries satisfactorily enough. The level of achieved accuracy is, to the best of our knowledge, the highest ever observed. The validity of the variable selection process and non-linear prediction was evaluated using a repeated cross-validation protocol yielding reliable results.

Predicting NBA player salaries using information on the players' performance on court yields predictions whose accuracy is satisfactory but not as high as one would expect. We argue that key factors mentioned in the manuscript, such as popularity, quantity of spectacle offered, etc. could improve the accuracy of the salary predictions significantly. Another future idea is to switch direction. Instead of investigating the present, whether the players are getting paid according to what they perform on court, one should investigate their future salaries. The level of the contract of a free agent depends not only on his record but also on many factors, such as his age, his playing position, and the available teams among others. For instance, a power forward/center with high performance will sign with a team that is looking for a power forward/center. Additionally, among those teams interested in that player, one must see their salary cap and the players already in that team in order obtain a better picture. Further, we did not include more personal information, such as whether a player is an All-Star, if he is a member of the all-NBA team or the NBA All-Defensive Team, etc. Examination of all those factors could yield more accurate salary predictions than those presented in this paper. Adoption of more complex machine learning algorithms, such as SVM (Drucker et al. 1997) or gradient boosting (Friedman 2001), is another possibility worth exploring.

We close this paper by posing a question. Is it possible that more than one combination of statistics facilitate the prediction of the NBA player salaries? Evidently, the minutes played, the field goals attempted, and the points scored are correlated. By observing the selected statistics in Table 4 we saw that the points scored, substituted the games played, and the field goals attempted, only for the last season. This could be evidence that the variable selection task returns one solution among the many.

Acknowledgements We are grateful to Stefanos Fafalios, Nikolaos Pandis, Eleftherios Pavlos, and Zacharias Papadovasilakis for reading earlier versions of this manuscript and providing constructive feedback.

References

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 46:175–185

Breiman L (2001) Random forests. Mach Learn 45:5–32

Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: Advances in neural information processing systems, pp 155–161

Ertug G, Castellucci F (2013) Getting what you need: how reputation and status affect team performance, hiring, and salaries in the NBA. Acad Manag J 56:407–431

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Friedman JH, Stuetzle W (1981) Projection pursuit regression. J Am Stat Assoc 76:817–823

Garris M, Wilkes B (2017) Soccernomics: salaries for World Cup Soccer athletes. Int J Acad Bus World 11:103–110

Groothuis PA, Hill JR (2013) Pay discrimination, exit discrimination or both? Another look at an old issue using NBA data. J Sports Econ 14:171–185

Hamilton BH (1997) Racial discrimination and professional basketball salaries in the 1990s. Appl Econ 29:287–296

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media

Hoffer AJ, Freidel R (2014) Does salary discrimination persist for foreign athletes in the NBA? Appl Econ Lett 21:1–5

Kahn LM, Shah M (2005) Race, compensation and contract length in the NBA: 2001–2002. Ind Relat: J Econ Soc 44:444–462

Kahn LM, Sherer PD (1988) Racial differences in professional basketball players' compensation. J Law Econ 6:40–61

Kelly T (2017) Effects of TV contracts on NBA salaries. Technical report, Department of Economics, Colgate University, USA

Lin Y, Jeon Y (2006) Random forests and adaptive nearest neighbors. J Am Stat Assoc 101:578–590

Olbrecht A (2009) Do academically deficient scholarship athletes earn higher wages subsequent to graduation? Econ Educ Rev 28:611–619

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rehnstrom K (2009) Racial salary discrimination in the NBA: 2008–2009. Major Themes Econ 11:1–16

Sigler K, Compton W (2018) NBA players' pay and performance: what counts? Sport J

Sigler KJ, Sackley WH (2000) NBA players: are they paid for performance? Manag Finance 26:46–51

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B (Methodol) 58:267–288

Tsagris M, Papadovasilakis Z, Lakiotaki K, Tsamardinos I (2020) The γ-OMP algorithm for feature selection with application to gene expression data. IEEE/ACM Trans Comput Biol Bioinform (accepted for publication)

Vincent C, Eastman B (2009) Determinants of pay in the NHL: a quantile regression approach. J Sports Econ 10:256–277

Wen R (2018) Does racial discrimination exist within the NBA? An analysis based on salary-per-contribution. Soc Sci Q 99:933–944

Wiseman F, Chatterjee S (2010) Negotiating salaries through quantile regression. J Quant Anal Sports 6

Wright MN, Ziegler A (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1–17

Xiong R, Greene M, Tanielian V, Ulibarri J (2017) Research on the relationship between salary and performance of professional basketball team (NBA). In: Proceedings of the 8th international conference on E-business, management and economics, pp 55–61

Yang CH, Lin HY (2012) Is there salary discrimination by nationality in the NBA? Foreign talent or foreign market. J Sports Econ 13:53–75

Yilmaz M, Chatterjee S (2003) Salaries, performance, and owners' goals in major league baseball: a view through data. J Manag Issues 15:243–255

Zimmer MH, Zimmer M (2001) Athletes as entertainers: a comparative study of earnings profiles. J Sport Soc Issues 25:202–215