Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n2p454

Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms
By Migliorati

Published: 14 October 2020

Electronic Journal of Applied Statistical Analysis
Vol. 13, Issue 02, October 2020, 454-473
DOI: 10.1285/i20705948v13n2p454

Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms

Manlio Migliorati*

University of Brescia, Department of Economics and Management
Via S. Faustino 74/b-25122 Brescia

Published: 14 October 2020

This paper aims to identify the drivers leading to victory for basketball matches in NBA, the american National Basketball Association.
Firstly, a dataset containing box scores and Dean's four factors for regular seasons from 2004-2005 to 2017-2018 has been prepared. Then, box scores and four factors have been used as classification independent variables, to predict the winner of matches involving the Golden State Warriors team. Both CART and Random Forests machine learning techniques have been applied, and quality of fitting analyzed.
Variable importance of fitted models has been studied to identify success drivers showing how, for Golden State Warriors, defense is a key factor to win a game.
At last, these models are shown to be suitable for coaching staff in game preparation, and CART models are shown to be valuable on the basketball court for match interpretation.

keywords: classification, NBA, success drivers, data mining, prediction, machine learning, sport analytics

1 Introduction

This paper aims to show how machine learning techniques can be profitably employed to identify drivers leading to success for basketball matches in NBA, the best basketball

*Corresponding author: m.migliorati017@unibs.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

championship in the world.
Data always played a fundamental role in sport management, constituting the starting point to objectivize analysis in a field where a huge amount of money is invested, but where fortuity plays a great role. So, a data based approach was early adopted in each professional sport (Alamar, 2013; Albert et al., 2017), to face different kind of problems ranging from performance measurement to players selection, from ranking analysis to match preparation.
But it was with the pioneering application described in (Lewis, 2003), centered on Oakland Athletics baseball team, that analytics in sport actually entered the maturity phase, and it became possible to properly speak about data mining in sport. In a fast way, the same approach was widely adopted and adapted, taking into account specific rules and historical data, in all professional sports: hockey, football, soccer...
Basketball milestones of this analytics based approach are pioneering works (Oliver 2004; Kubatko et al. 2007), where concepts as pace and possession were introduced, together with famous "four factors to success": few indexes concentrating an high number of information.
In this paper we are interested in applying data mining to basketball (Bianchi et al., 2017; Zuccolotto and Manisera, 2020), for identifying which are the drivers to victory. To do this, we will focus on the relationship between match outcomes (only win or loss, for basketball) and sets of features offering a snapshot of the match, as already studied in other sports (Carpita et al. 2015; Carpita et al. 2019; Carpita et al. 2020).
Our goal is not to produce outcome predictions for next games to be played; instead, we would like to offer valuable tools to staff coaching, supporting them to decide which moves can be done both before a game and on the basketball court to increase success odds. So, our predictions are made on the base of ex-post information, with the only goal of measuring the goodness of our fitting, and consequently the goodness of the success drivers we will identify.
To do that, we will use two different set of predictors:

1. the so called *box score* analytics, i.e. the classical information (attempted shots, made shots, ...) summarizing a match, see for instance (NBA 2020a; ESPN 2020)

2. the already mentioned *Four Factors*; Oliver identified them in trying to understand how basketball teams win games, and empirically associated them a weight:
   a) Shooting (40%)
   b) Turnovers (25%)
   c) Rebounding (20%)
   d) Free Throws (15%)

   Looking for success drivers, we will look for confirmation about these weights, too.

A large dataset, including 14 NBA regular seasons (from 2004-2005 to 2017-2018), was built, and variables from both box scores and four factors have been used as input for machine learning classification, with the goal of building models predicting the winner

of matches of GSW (Golden State Warriors), champion of season 2017-2018.
Then these models are analyzed in terms of fitting quality and variable importance, to fairly identify success drivers.
Classification models are built using:

- CART (Classification and Regression Trees, Breiman et al. 1984)
- Random Forests (Ho 1995; Breiman 2001)

to couple trees easiness of interpretation, so important for data mining results acceptance and usage, to superior Random Forests fitting qualities.
All analysis in this paper were carried using the R language (Ihaka and Gentleman 1996; R Core Team 2019) ver. 3.6.3 following the tidy approach (Wickham 2014; Wickham et al. 2020).

2 Basketball predictions via machine learning

Considering the large interest in sport betting, it can not be surprising that a great job was made trying to accurately predict games results, see for instance (Bunker and Thabtha 2019; Hubáček et al. 2019). Machine learning techniques have been widely applied, covering all professional sports, from horse races (Davoodi and Khanteymoori, 2010) to hockey (Gu et al., 2016), from American football (Purucker 1996; Kahn 2003; David et al. 2011), to soccer (Tax and Joustra 2015; Min et al. 2008), to give some examples. And that's true for basketball, too.
In Loeffelholz et al. (2009) authors worked on a dataset of 620 NBA games, and used several kinds of ANN (Artificial Neural Networks (Zhang, 2000)) for outcomes prediction; they report a correct winner prediction percentage of 74.33, higher than experts percentage claimed to be 68.67.
In Miljkovic et al. (2010) focus is on predicting the outcomes (and calculating the spread) for 778 NBA games of season 2009-2010, using 141 features as input. Best results are reported for Naive Bayes Classifier (Langley et al., 1992), with an accuracy of 67% .
In Cao (2012) data of 5 NBA seasons were analyzed to produce NBA game outcomes, using ANN, Support Vector Machine (Cortes and Vapnik, 1995), Naïve Bayes and Simple Logistics Classifier (Hosmer and Lemeshow, 2000), with this last approach producing the best prediction accuracy (about 70%).
In a similar way in Beckler et al. (2013) authors used Linear Regression, Support Vector Machines, Logistic Regression and ANN for NBA outcomes prediction, using a dataset including seasons from 1991-1992 to 1996-1997 and reporting an accuracy of 73%.
In Cheng et al. (2016) authors applied the principle of Maximum Entropy (Jaynes, 1957) to predict NBA playoff outcomes for seasons from 2007–08 to 2014–15, using box score information as features, reporting an accuracy of 74.4%.
At last, there are several betting sites suggesting NBA outcomes predictions. As an example, teamranking (2020) proposes predictions about NBA match winners using 4 approaches, built on the base of several sources (historical data, breaking news and trends). For regular season 2017-2018 the maximum accuracy is 74.3%, obtained using

decision tree on data of march games.
The perspective of our job is different, and not so far from Thabtah et al. (2019). Working on a dataset of 430 observations focused on NBA final games from 1980 to 2017, authors applied 3 machine learning algorithms (Naïve Bayes, ANN and Logistic Model Trees (Landwehr et al., 2005)) for predicting outcomes using box score information as predictors. They report an accuracy of more then 80% in best cases, and concluded how defensive rebounds are the more influential feature.
We are not interested in predicting in advance the game winner, but in using fitting quality of prediction models as a guarantee for the quality of our analysis. So we will use a large (thousands instead of hundreds) dataset, and our prediction models will be based on ex post information, to have high accuracy predictions. These models will be analyzed in terms of variable importance for identifying success drivers, and the high fitting quality of prediction models will be the guarantee that our analysis about variable importance is fair, too.

3 Basic basketball analytics

Among several others analytics (see NBA (2020b)), basketball match analysis can be approached using:

1. box score analytics, a set of indicators summarizing the trends of a match. For our analysis purposes, box store contains 13 information to be used as predictors:
   - 1.1 PTS: points made
   - 1.2 P2A: 2 points field goals attempted
   - 1.3 P2M: 2 points field goals made
   - 1.4 P3A: 3 points field goals attempted
   - 1.5 P3M: 3 points field goals made
   - 1.6 FTA: free throws attempted
   - 1.7 FTM: free throws made
   - 1.8 OREB: offensive rebounds
   - 1.9 DREB: defensive rebounds
   - 1.10 AST: assists (passage of the ball leading to a field goal score)
   - 1.11 TOV: turnovers (loss of ball possession)
   - 1.12 STL: steals (stealing ball to opponent)
   - 1.13 BLK: blocks (deflecting a field goal attempt)

   and these information are available for both teams involved in the match.

2. four factors[1], a set of derived statistics that, following (Oliver, 2004), are key to success. Starting from the concept of *possession*, i.e. the number of times a team gains control of the ball during a match, defined as

$$Poss = (P2A + P3A) + 0.44 * FTA - OREB + TOV \tag{1}$$

it is possible to define the four factors:

a) Shooting, measured by effective Field Goals percentage:

$$eFG\% = (P2M + 1.5 * P3M)/(P2A + P3A) \tag{2}$$

b) Turnovers ratio, the number of turnover (i.e. loss of ball) per possession

$$TO\% = TOV/POSS \tag{3}$$

c) Rebounding, defined by rebounding percentage:

$$Reb\% = OREB/(OREB + DREB) \tag{4}$$

d) Free throws rate

$$FT\_rate = FTM/(P2A + P3A) \tag{5}$$

In effect, for a match we must calculate four factors both for a team and for its opponent, arriving to 8 statistics that will be addressed as follows:

1. shooting
   - team effective field goals percentage

$$eFG.team = (P2M.team + 1.5 * P3M.team)/(P2A.team + P3A.team) \tag{6}$$

   - opponent effective field goals percentage

$$eFG.opp = (P2M.opp + 1.5 * P3M.opp)/(P2A.opp + P3A.opp) \tag{7}$$

2. turnovers ratio
   - team turnovers ratio

$$TO.team = TOV.team/POSS.team \tag{8}$$

   - opponent turnovers ratio

$$TO.opp = TOV.opp/POSS.opp \tag{9}$$

[1] for this paper four factors were calculated via the R package BasketballAnalyzeR, as described in (Zuccolotto and Manisera, 2020)

3. rebounding percentage

   - team offensive rebounding percentage

$$Reb.off.team = OREB.team/(OREB.team + DREB.opp) \tag{10}$$

   - team defensive rebounding percentage

$$Reb.def.team = DREB.team/(OREB.opp + DREB.team) \tag{11}$$

4. free throws rate
   - team free throws rate

$$FT.rate.team = FTM.team/(P2A.team + P3A.team) \tag{12}$$

   - opponent free throws rate

$$FT.rate.opp = FTM.opp/(P2A.opp + P3A.opp) \tag{13}$$

4 The dataset

Current rules for regular season were adopted in season 2004-2005[2]: NBA championship was subdivided in 2 conferences (est and west); each conference is compound by 3 divisions, and each division includes 5 teams, so overall there are 30 teams in NBA.
Each season starts with a regular season involving all teams, and each team plays 82 games. Regular season is followed by playoff, where only the best 16 teams fight to gain the final.
In this scenario our dataset includes 14 seasons, for a total of about 17.000 matches of regular seasons.
What we are going to do is to focus on 1 team, Golden State Warriors (the winner of season 2017-2018):

- for the 14 regulars seasons from 2004-2005 to 2017-2018 we have 1130 games involving GSW.
- 90% of our observations, randomly chosen, will be used for training our classification models
- the remaining 10% will be used for testing

5 CART and Random Forests

Just few words to remember what CART (Classification And Regression Tree, Breiman et al. 1984) and Random Forests (Ho 1995; Breiman 2001) are.

[2]and this uniformity of rules is the reason why our dataset starts from this season; instead, playoff access rules change more often, last time in 2016

5.1 CART

CART is a kind of binary decision tree, built binary splitting (in a recursive way) a population in several regions. Splits are decided on the base of predictor's values minimizing a cost function, and are recursively applied until a stop condition is fired; at the end of the process, each region will have associated a constant response.
In a tree:

- nodes represent predictors, and one predictor value is chosen to split the node in 2 branches.
- for a new observation, a branch is chosen on the base of its predictor value
- this action is repeated until the new observation reaches a tree leaf (i.e. a node without children)
- the constant value associated to that leaf is the prediction for the new observation

CART can be used for both classification (when the output variable is qualitative, our case in predicting game winner) and regression (when the output variable is quantitative): classification of new observations uses the mode of the population of a leaf, regression the mean.
CART plays an important role in machine learning, not only because of their not so bad predictive power, but mainly because of the ease in understanding its results.
In general with CART trees attention must be payed to avoid *overfitting*, i.e. building a model too much tailored on training data, and consequently not particularly suitable in managing new observations.

5.2 Random Forests

Random Forests is an ensemble learning technique, and can be used both for classification and regression. It is based on the idea of building a huge number of different decision trees, where tree splits are evaluated on the base of a subset of randomly chosen predictors.
In this way it is possible to generate trees where:

- also weak predictors can play a role
- more important predictors can be excluded from the model

mitigating in this way the classical tree overfitting problems.
Random Forests have a good predictive power, but their results are not so simple to be interpreted as for trees.

In the following we will use both CART and Random Forests for classification, looking for an answer to the question: "will Golden State Warriors win this game?".

6 CART box score based

Our first model is a CART tree built using box score variables as predictors.
A first application using all the 13 box score variables for both teams produces only an obvious conclusion: points are by far the more important variable, and Golden State Warriors will win a game when its score is greater than the opponent score.
So, in order to find something of more interesting, we exclude from predictors both PTS(points) and AST(number of assist) variables.
The tree (a bit pruned to limit overfitting) obtained from training data seems more interesting, and is depicted in figure 1:

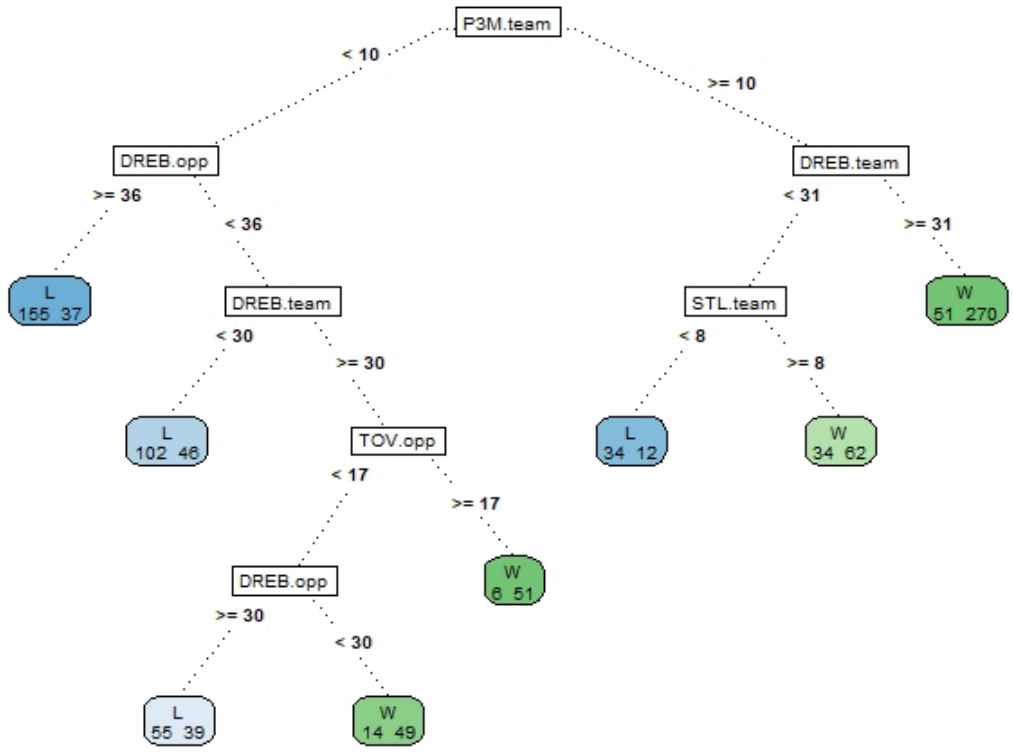

Figure 1: training CART based on box score predictors without PTS and AST

In this tree:

- nodes represent predictors: so, the root of the tree, the P3M.team node, represents the *made 3-points shots* input
- for classifying a new observation the tree is crossed on the base of the value of the predictor: starting from the root of the tree, we will descend to the left if the made

3-points shots of new observation are less than 10, otherwise we will descend to the right

- leaves contains the classification result (Winner (green) or Loser (blue)), together with frequency of the 2 possible classification results. Color is darker depending on how a frequency is greater than the other one.

So, a new observation is classified finding the more suitable leaf (and using associated constant as prediction) on the base of its predictor values, crossing the tree starting from the root.
The meaning of the pink path in figure 2:

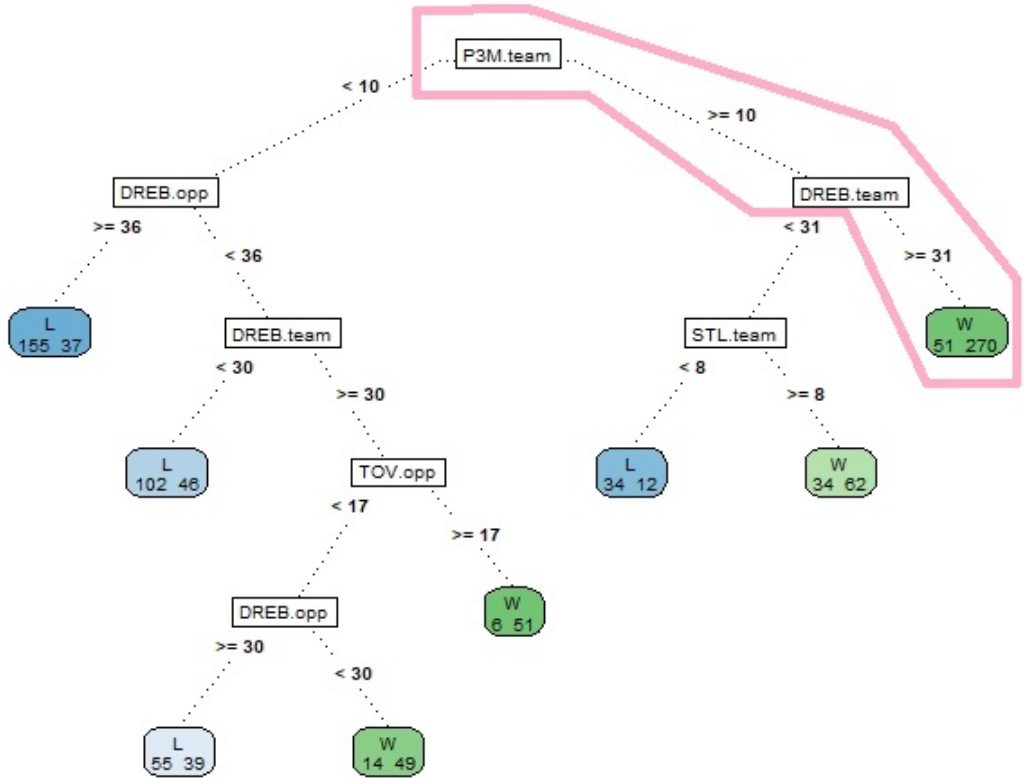

Figure 2: a path in CART built using box score predictors

is that Golden State Warriors will be predicted to win its match when they make at least 10 3-point shots ($P3M.team \geq 10$) and 31 defensive rebounds ($DREB.team \geq 31$): a good example of the easiness of tree understanding we talked about.
Moreover, that tree offers a clear example of the reason why this kind of tools can be really valuable to staff coaching.

Looking to the left subtree, we can see how the situation seems against GSW when the number of 3-point shots is lower than 10 ($P3M.team < 10$), because many leaves predict with high probability a GSW defeat. But, in effect, the tree shows how there are 2 more little possibilities that can be tried if the opponent is not too strong in defensive rebounds ($DREB.opp < 36$). GSW should pay great attention to defense, trying to get an high number of defensive rebounds ($DREB.team \geq 30$), together with an aggressive attitude inducing opponent or to lose many balls ($TOV.opp \geq 17$)), either to take few defensive rebounds ($DREB.opp < 30$).
So, the GSW coach should apply game plans bringing his team to green (i.e. victory) leaves.
Only 5 predictors (3-point shots, team defensive rebounds, opponent defensive rebounds, team stolen balls, opponent lost balls), among the 22 available, were used to build up the tree: this is a good first step in the direction of identifying the success drivers we are looking for, but first we have to verify how much the model is suitable.
To assess fitting quality, the CART model is tested predicting the final results for the remaining 10% of our observations, and comparing these predictions to actual results.
In other words, we built a confusion table for predictions and actual results, producing following table:

Table 1: confusion matrix for CART box score without PTS and AST

		Prediction		
GSW winner?		no	yes	Total
Actual	no	34	10	44
	yes	22	47	69
	Total	56	57	113

and accuracy (defined as the ratio between correctly classified observations and overall number of observations) for this model is 0.7168, high enough for safely analyzing variable importance to detect success drivers. Variable importance for CART models is calculated[3] using not only effective splits (i.e. predictors effectively employed for splitting), but also surrogate splits (i.e. predictors to be used in case of missing values) and competing splits (i.e. predictors evaluated but not chosen for splits), to have a view as complete as possible.
In this way we can trace not only variables appearing in tree plot, but also other variables playing a meaningful role in tree building. For CART model above, variable importance is depicted in following figure:

[3]see R caret package documentation for details (Khun, 2020)

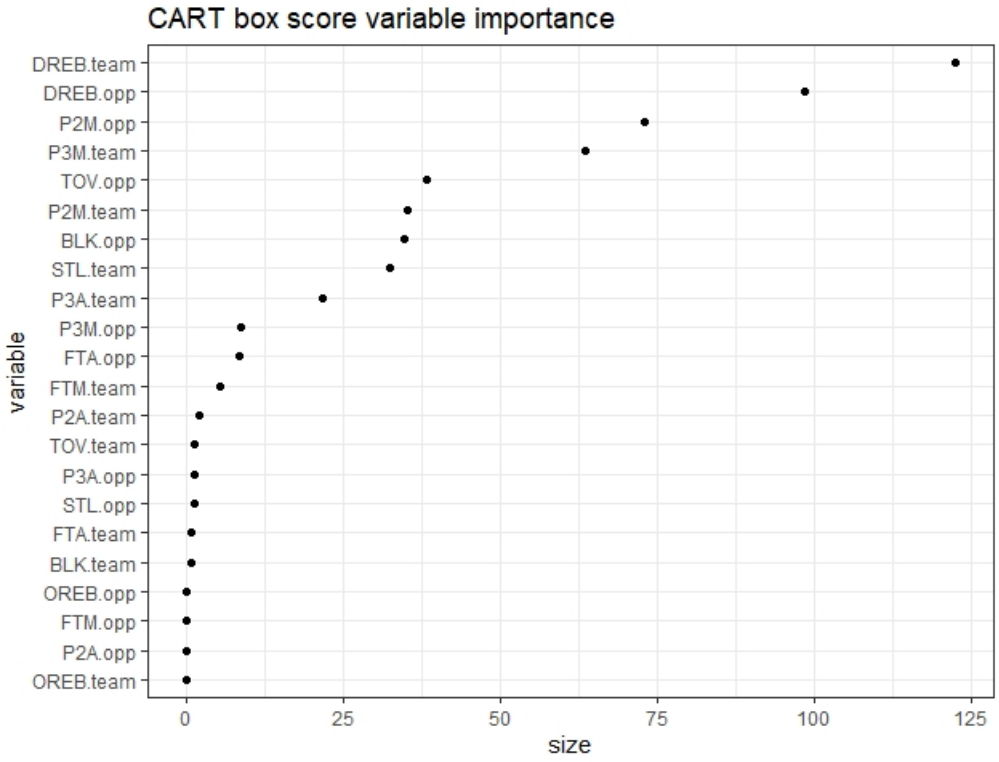

Figure 3: variable importance for CART box score without PTS and AST

Looking to the plot, it is clear that defense is a key factor for GSW successes, and (but this is more obvious) to win it is necessary to made 2 and 3 point shots.
Moreover, it is possible to note how offensive rebounds, a variable which is normally considered as really important (and in effect it important in some game situations, as shown in (Zuccolotto et al., 2017)), are neglected in our model, also considering a definition of importance as large as possible as we did.

7 CART four factors based

Second model is still built using CART, but with four factors as predictors. The (pruned) training tree for our dataset is depicted in figure 4:

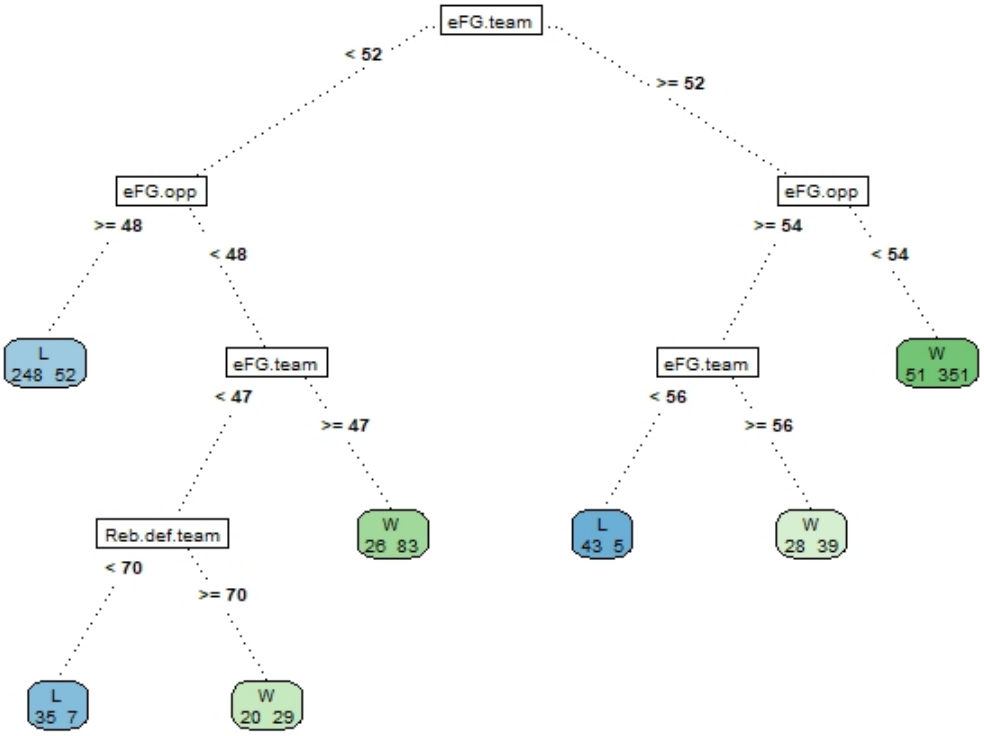

Figure 4: CART built using full four factors

In this case the model is built using just three predictors (eFG.team, i.e. GSW shooting factor, eFG.opp, i.e. opponent shooting factor, and Reb.def.team, i.e. GSW defensive rebound factor) among the 8 variables available.
This tree does not contain amazing information; mainly, it shows how shooting factor plays a primary role as success driver (a first confirmation of Oliver's weights definition). Just an interesting note: defensive rebounds seem to have high importance (node Reb.Def.team on the left), confirming what we observed analyzing box score tree.
In any case, the presence of shooting factors maybe hides other useful information, so we build a model without them, obtaining following tree:

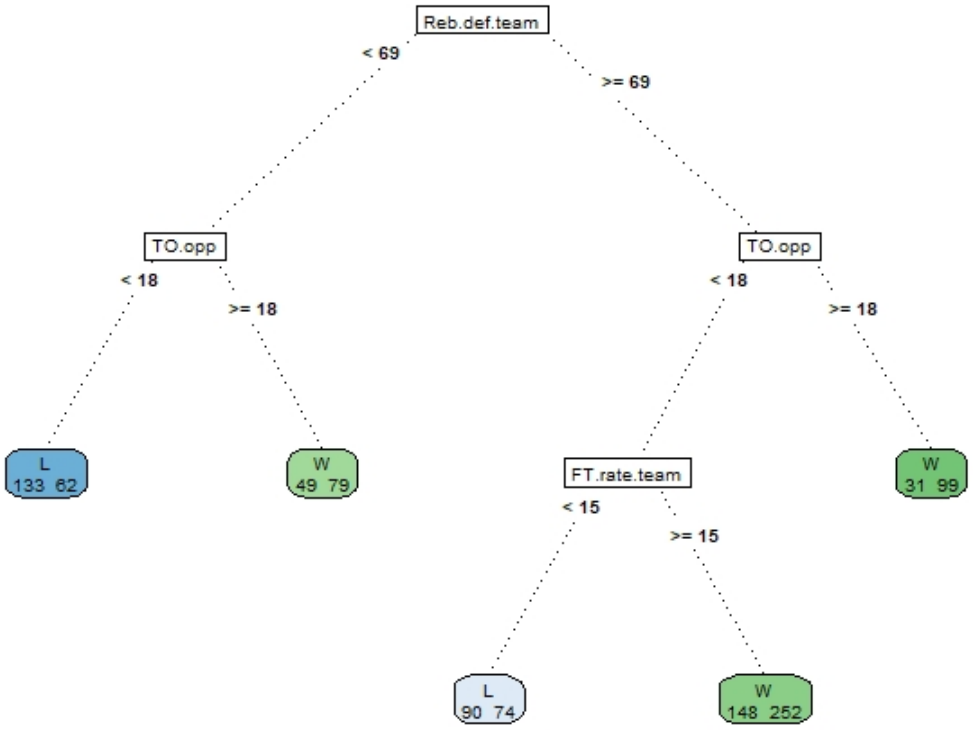

Figure 5: CART built using four factors without shooting

and this tree confirms conclusions about importance of GSW defense made using box scores model based: they must fight for rebounds in their own area, and play in an aggressive way to induce opponents to lose the ball.
This tree offers another path to be interpreted: the FT.rate.team (i.e. the number of free throws respect to sum of 2 and 3 point shots) becomes important in situations where GSW is strong on defensive rebounds, but not strong enough in frequently inducing opponent to lose the ball. In these situations, it is appropriate to play to gain free throws with respect to other possibilities way to conclude the action; how do it depends on game peculiarities (for instance coaching staff can decide to play in attack, and/or inducing opponents to make fouls), but the target the coaching staff should get is clear. Again, tree can be a valuable tool to coaching staff, both to prepare the match and to react in the right way to situations happening on the court.
What's about the quality of the fitting? Confusion table is the following:

Table 2: confusion matrix for CART four factors without shooting

		Prediction		
GSW winner?		no	yes	Total
Actual	no	18	26	44
	yes	11	58	69
	Total	29	84	113

with an accuracy of 0.6726: not so bad, taking into account we are not considering shooting, the most important factor, and high enough to let this analysis be used as decision support tool by coaching staff.

8 Random Forests box score based

In this section we will analyze models built using Random Forests, to have another source of information with high quality fitting.
By construction Random Forests enable emerging of less weight variables, but the presence of variables PTS and AST is too cumbersome; so, as for CART, we prefer to make box score based predictions without them. Results can be found in following table:

Table 3: confusion matrix for Random Forests box score without PTS and AST

		Prediction		
GSW winner?		no	yes	Total
Actual	no	40	4	44
	yes	6	63	69
	Total	46	67	113

Accuracy is about 0.9: the model is really suitable, also if we don't use points and assists. Variable importance, in terms of mean decrease both in accuracy node impurity, is depicted in figure below:

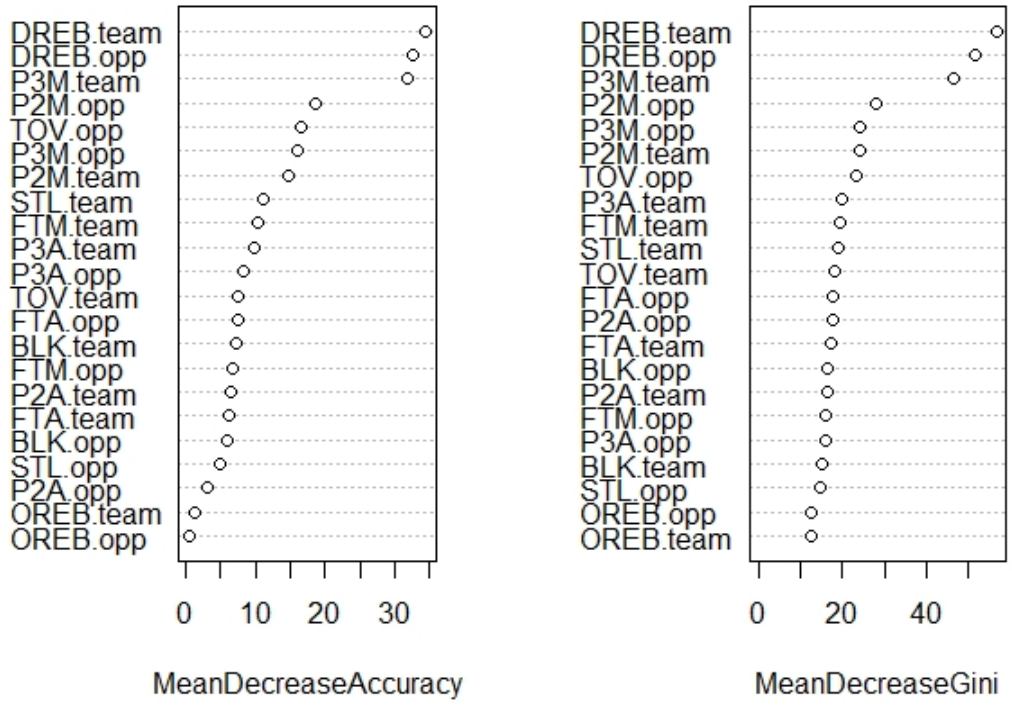

Figure 6: Random Forests box score, mean decrease

Again, we find a confirmation that defense is really important for GSW to win: among more important variables we have defensive rebounds and opponents turnover. As it is easy to guess, made shots are important, too. At last, it is confirmed the low importance of offensive rebounds.
So, also this analysis is aligned with conclusions we made on the base of CART model results.

9 Random Forests four factors based

Our last model: Random Forests using four factors as features. Prediction results are summarized in following table:

Table 4: confusion matrix for Random Forests four factors

		Prediction		
GSW winner?		no	yes	Total
Actual	no	40	4	44
	yes	3	66	69
	Total	43	70	113

This model has the highest accuracy (0.94). In this case we used all 8 predictors, because the model without shooting, the most important among four factors, has an accuracy equals to 0.6018, not so high, without offering new infos about success drivers.

For this model variable importance is the following:

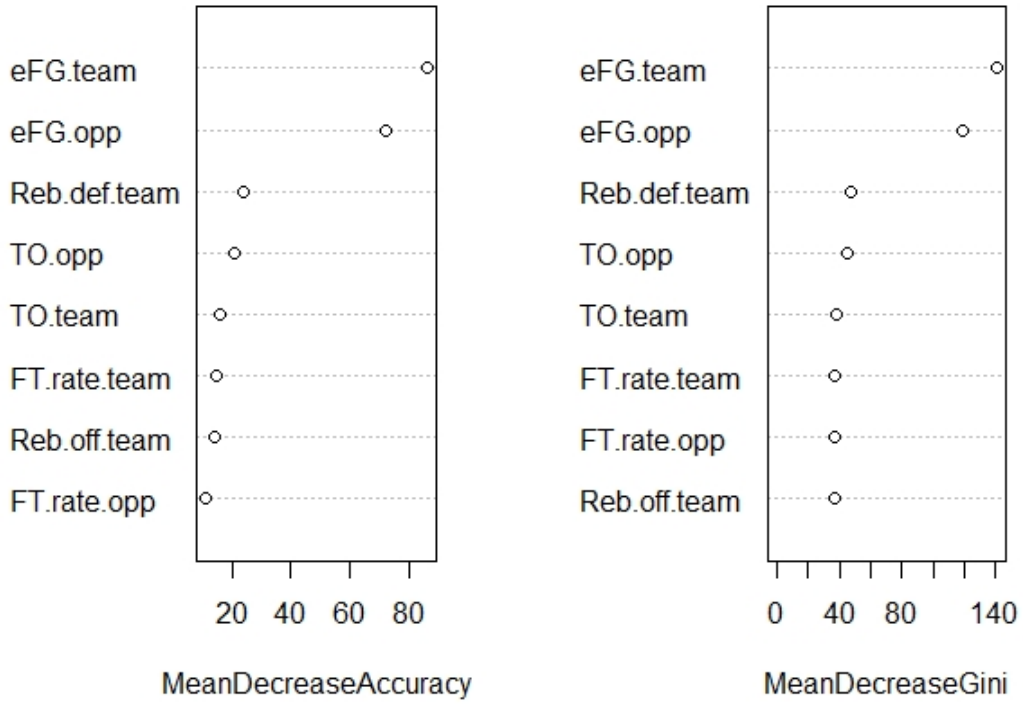

Figure 7: RF four factors, variable importance

As we can imagine shooting factors, both offensive and defensive, play a primary role; moreover, we have another confirmation about importance of defense and low importance of offensive rebounds, as we argued interpreting other models.

10 Conclusions

In this paper CART and Random Forests machine learning techniques have been applied to a large dataset of NBA games, with the purpose of detecting success factors. In particular we focused on Golden State Warriors regular seasons from 2004-2005 to 2017-2018, sharing the same rules framework, looking for their victories' drivers.
Our classification models were built using as predictors both box score and four factors, on the base of ex-post data, showing high quality fittings:

Table 5: comparing prediction measures

	CART Box Score no PTS and AST	CART Four Factors no shooting	Random Forests Box Score no PTS and AST	Random Forests Four Factors
sensitivity	0.6812	0.8406	0.9130	0.9565
specificity	0.7727	0.4091	0.9091	0.9091
accuracy	0.7168	0.6726	0.9115	0.9381
recall	0.6812	0.8406	0.9130	0.9565
precision	0.8246	0.6905	0.9403	0.9429
F_measure	0.7461	0.7582	0.9264	0.9497

and this high quality fittings are a guarantee we can fairly analyze models in terms of variable importance, to detect the success drivers we are looking for.
In the case of Golden State Warriors we detect how, after shooting factors, more important success factors are related to defense (defensive rebounds and opponent turnovers). Instead, it seems that offensive rebounds are not so important, confirming the famous saying 'offense sell tickets, defense wins championships".
With respect to Oliver's four factors weighting, it is confirmed how shooting factor is the most important success driver but, in our dataset, defensive rebounds seem to be more important than turnover. Instead, it is confirmed also the lower importance of free throws (but we verified how they become important in particular situations).
Models we built can be useful to coaching staff in preparing games; moreover CART trees, thanks to their understandability, can be useful also on the court, to interpret the game and try to drive the team to advantageous situations.

Next steps will concern the application of the described approach to other teams and basketball championships (e.g. italian Lega Basket), to verify if and how success factors are team and championship specific, and to investigate models and drivers for score differences.

Acknowledgment

I'd like to thank Prof. Marica Manisera, Prof. Paola Zuccolotto and Prof. Maurizio Carpita, from the department of Economics and Management of University of Brescia, for their invaluable suggestions and for the possibility they offered me to work in such a really interesting domain as analytics applied to sport is.

References

Alamar, B. C. (2013). *Sports analytics: A guide for coaches, managers, and other decision makers*. Columbia University Press.

Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H. (2017). *Handbook of Statistical Methods and Analyses in Sports*. CRC Press.

Beckler, M., Wang, H., and Papamichael, M. (2013). Nba oracle. https://www.mbeckler.org/coursework/2008-2009/10701_report.pdf.

Bianchi, F., Facchinetti, T., and Zuccolotto, P. (2017). Role revolution: towards a new meaning of positions in basketball. *Electronic Journal of Applied Statistical Analysis*, 10(3):712–734.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Bunker, R. P. and Thabtha, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction. masters dissertation; technological university dublin. http://arrow.dit.ie/cgi/viewcontent.cgi?article=1040&context=scschcomdis.

Carpita, M., Ciavolino, E., and Pasca, P. (2019). Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, 19(1):74–101.

Carpita, M., Ciavolino, E., and Pasca, P. (2020). Players' role-based performance composite indicators of soccer teams: A statistical perspective. *Social Indicators Research*. https://doi.org/10.1007/s11205-020-02323-w.

Carpita, M., Sandri, M., Simonetto, A., and Zuccolotto, P. (2015). Discovering the drivers of football match outcomes with data mining. *Quality Technology & Quantitative Management*, 12(4):537–553.

Cheng, G., Zhang, Z., Kyebambe, M., and Kimbugwe, N. (2016). Predicting the outcome of nba playoffs based on the maximum entropy principle. *Entropy*, 18(12).

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *machine learning*, 20:273–297.

David, J., Pasteur, R. D., and Ahmad, M. . J. M. (2011). Nfl prediction using committees of artificial neural networks. *Journal of Quantitative Analysis in Sports*, 7(2):1–15.

Davoodi, E. and Khanteymoori, A. (2010). Horse racing prediction using artificial neural networks. *recent advances in neural networks, fuzzy systems & evolutionary computing*, pages 155–160.

ESPN (2020). Espn nba statistics. https://www.espn.com/nba/stats.

Gu, W., Saaty, T., and Whitaker, R. (2016). Expert system for ice hockey game prediction: Data mining with human judgment. *International Journal of Information Technology & Decision Making*, 15(04):763–789.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document, Analysis and Recognition*, pages 278–282. IEEE.

Hosmer, D. and Lemeshow, S. (2000). *Applied logistic regression*. John Wiley & Sons, Inc.

Hubáček, O., Sourek, G., and Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysys and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Jaynes, E. (1957). Information theory and statistical mechanics. *the physical review*, 106(4):620–630.

Kahn, J. (2003). Neural network prediction of nfl football games. http://homepages.cae.wisc.edu/ ece539/project/f03/kahn.pdf.

Khun, M. (2020). *package "caret"*. https://cran.r-project.org/web/packages/caret/caret.pdf.

Kubatko, J., Oliver, D., Pelton, K., and Rosenbaum, D. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3(3):1–22.

Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine Learning*, 59:161–205.

Langley, P., Iba, W., and Thompson, K. (1992). An analysis of bayesian classifiers. In *Proceedings of theTenth National Conference on Artificial Intelligence*, pages 399–406. AAAI.

Lewis, M. (2003). *Moneyball : the art of winning an unfair game*. W.W. Norton.

Loeffelholz, B., Bednar, E., and Bauer, K. (2009). Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1):1–17.

Miljkovic, D., Gajic, L., Kovacevic, A., and Konjovic, Z. (2010). The use of data mining for basketball matches outcomes prediction. In *IEEE 8th international symposium on intelligent systems and informatics*, pages 309–312. IEEE.

Min, B., Kim, J., Choe, C., Eom, H., and McKay, R. I. (2008). A compound framework for sports results prediction: a football case study. *Knowledge-Based Systems*, 12:551–562.

NBA (2020a). Nba statistics. https://stats.nba.com/teams/traditional.

NBA (2020b). Nba statistics glossary. https://stats.nba.com/help/glossary/.

Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books inc.

Purucker, M. (1996). Neural network quarterbacking. *IEEE Potentials*, 15:9–15.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tax, N. and Joustra, Y. (2015). Predicting the dutch football competition using public data: A machine learning approach. *IEEE Transactions on Knowledge and Data Engineering*, 10(10):1–13.

teamranking (2020). teamranking predictions. https://www.teamrankings.com/nba/betting-models/detailed-splits/.

Thabtah, F., Zhang, L., and Abdelhamid, N. (2019). Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*, 6(1):103–116.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.

Wickham, H., François, R., Henry, L., and Müller, K. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5.

Zhang, P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, 30(4):451–462.

Zuccolotto, P. and Manisera, M. (2020). *Basketball Data Science – with Applications in R*. Chapman and Hall/CRC.

Zuccolotto, P., Manisera, M., and Sandri, M. (2017). Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International journal of sports science & coaching*, 13(4):569–589.