Game-to-Game Prediction of NBA Players' Points in Relation to Their Season Average

Trpimir Zovak, Ana Šarčević, Mihaela Vranić, Damir Pintar
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
{trpimir.zovak, ana.sarcevic, mihaela.vranic, damir.pintar}@fer.hr

Abstract* – NBA attracts a great deal of attention among sports analysts and sportsbooks regarding the prediction of various outcomes of each game, together with the parameters which affect them. Performance of NBA players is influenced by many unknown and random factors, such as players' psychological condition, social life and injuries. The stated factors hinder game-to-game predictions of players' performance in relation to the expectations set by their past performances. In this paper we leverage the publicly available statistics to create a dataset pertaining to the performance of a single player during a single season. A comparison between points that a player has scored and his current season average was done in order to classify the player's performance as 'over' or 'under'. Using various statistical data concerning previous games of the season, a binary classifier was trained in order to distinguish between those categories for future games. The classifier performed with an accuracy score of 56.7%. Since sportsbooks tend to give 50/50 odds of a player going 'over' or 'under' in relation to their season points per game, these results represent an improvement of 6.7%. Although top features are predominated by offensive statistics (e.g. how many minutes the player plays, how many shots he takes and how strong the offense of his team is), a newly generated feature, which represents tiredness of a player, has shown to be among top 15 informative features.

I. Introduction

Basketball is a sport abundant in statistical data and thus a great candidate for pattern recognition and outcome prediction. NBA (National Basketball Association) is the most watched, media covered and analytically scrutinized basketball league in the world, and for that reason was chosen as the main focus of this paper. The main goal of our research was to collect the necessary data, analyze it and discover patterns which may conform to the existing domain knowledge accumulated throughout the years of watching sports. Additional ideas were derived from previous work done in the field of sports analytics.

In this paper, each game played by a player is labeled with one of the two classes: 'over' or 'under'. Value of the label depends on whether a player scored more or less points in a currently observed game compared to his current season average. The difficulty of this problem lies in its formulation, since the goal is not to predict the performance of a player, which could be done much more successfully, but rather to identify whether he will outperform or underperform in relation to the expectations set by his previously shown performance. Our developed predictive model will showcase a clear advantage over a currently prevalent method of randomly chosen estimation of performance. After providing Related Work in Section II and explaining the possible applications in Section III, collection and preparation of Data was explained in Section IV. The Model, together with the results, is presented in Section V while the ideas for Future Work are discussed in Section VI. Section VII concludes the paper.

II. Related Work

Throughout the years, different authors tried to tackle different problems regarding the NBA. Wheeler gives a good general idea about the features that may be good indicators of player's performance in [1]. Apart from the obvious features, the author discovered two interesting informative features: opponent's shots taken within 10 feet and opponent's shots taken near the rim.

Players' performance can differ depending on how injury-prone they are and how the process of aging affects them [2]. Also, it has been shown that players whose contract is expiring tend to perform above their standards in order to promote themselves for a future contract.

When it comes to predicting who the winning team will be, there are four most indicating features in prediction of the winning team are: defensive rebounds, the points scored by the opposing team, and the number of blocks and assists made by the opposing team. This is a strong indicator that, in order to win a game, it is more important to prevent the opposing team from playing well and scoring points, instead of solely focusing on playing well offensively [3].

Player's performance is most often measured with Player Efficiency Rating (PER). PER sums up positive accomplishments achieved by a player, subtracts the negative accomplishments and returns a per-minute rating of a player's performance. A downside of statistics like PER is a lack of context in which a player performs [4]. For instance, metrics which are based on points differential (e.g. PER, +/-), model the home team's lead dropping from 5 to 0 points in the last minute of the first half in the exact same way that they model the home team's lead dropping from 30 points to 25 points in the last minute of the game. Another problem is that basketball statistical information tends to emphasize offensive performance since there are not nearly as many discrete defensive factors to record in a box score as there are offensive factors. As such, metrics like PER can be biased against defensive specialists. Therefore, the authors propose a win-probability model for evaluating the impact NBA players have on their teams' chances of winning. This allows them to identify players whose high impact is not captured by existing metrics.

III. Motivation and Aplication

A well-known problem, comparable to the one discussed in this paper, is stock investing. The goal of investing is to buy stocks which have more likelihood of price growth. Investors want to know how does the price of a stock compare to the value of a stock so they can decide whether the stock is overvalued or undervalued. If a stock is undervalued, the investors can be more confident that the price of a stock will rise in the future. Real value of a stock is impossible to determine since there is so much information unknown to the investor. Even if all possible information was available, incorporating them correctly would be a challenge. The same goes for predicting outcomes in sports. The market value of a stock rests on an objective foundation (asset value, debt, income, expected growth etc.) the same way that an offer given by a sportsbook rests on a foundation of different sports parameters. These foundations are influenced subjectively by the law of supply and demand. Although this model cannot predict the subjective influence of people betting on a certain event, it can either help the sportsbooks to set a stronger foundation or help a 'player' to exploit those subjective influences.

A variation of this model can be used by players and teams directly. They can see how does each feature impact the player's performance or the performance of the team and work on those features. For example, if they notice that an opposing team defends the rim poorly, they can try and exploit that weakness.

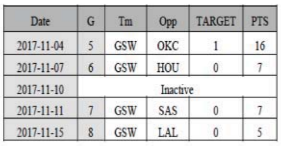
IV. Data

A. Data Collection

In order to predict future events, one often needs to leverage historical data, often in abundant quantity and quality. Basketball Reference has advanced statistics available for all NBA games played in recent history [5]. We used Beautiful Soup library to collect large amounts advanced data for all active teams and players in regular season of 2017-2018 [6]. Each player (and team) was modelled through a separate data frame, where each row represents one game of the season, and each column represents a piece of information about that game such as points scored, minutes played, assists, rebounds etc. An example of a subset of Steph Curry's data frame for season 2017-2018 can be found in Table 1, together with a subset of his team's data frame in Table 2.

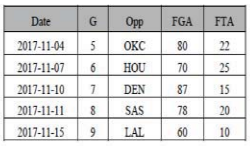
Our goal is to predict the performance of a chosen player Y in a chosen game X. The data will be converted in a way that one instance of the data set contains previously available information about the game X: date of the game, the opposing team, the player's, his team's and the opposing team's past performances.

Table 1. Stephen Curry



Date	G	Tm	Opp	TARGET	PTS
2017-11-04	5	GSW	OKC	1	16
2017-11-07	6	GSW	HOU	0	7
2017-11-10	Inactive				
2017-11-11	7	GSW	SAS	0	7
2017-11-15	8	GSW	LAL	0	5

Table 2. Golden State Warriors



Date	G	Opp	FGA	FTA
2017-11-04	5	OKC	80	22
2017-11-07	6	HOU	70	25
2017-11-10	7	DEN	87	15
2017-11-11	8	SAS	78	20
2017-11-15	9	LAL	60	10

When it comes to real life data, missing values are a common occurrence. For example, some players (for example Dwight Howard) are not good three-point shooters and as such do not even attempt any three-pointers. Looking at one of Howard's games, the following can be seen: 0 three-point shots made (3P), 0 three-point shots attempted (3PA) and a percentage of successful three-point shots (3P%=3P/3PA) with a non-existing value (*NaN*). *NaN* stands for a missing value, since division by zero is not defined. To deal with this, missing values can be replaced by season average, or if season average is also not existent, with 0. After removing all players who have not played at least 17 games with an average of 10 minutes per game, we are left with a total of 432 players.

If a player missed a game for any reason (suspension, coach decision, injury etc.) that game was removed from the data set and information about its deletion was recorded with a new variable. The variable 'missed games' starts with the value of 0 and for every game a player missed, it increases in value by 2, and for every game a player plays it decreases by 1 (with the minimum value of 0). The idea behind this is that if a player missed 1 game, he needs to play 2 games to negate the effect of his absence. This approach is oversimplified and needs to be addressed in future work. One of the problems is that the reason of absence is neglected. A broken nose, a knee injury and a suspension do not have the same effect on player's ability to perform after coming back.

B. Target

For every game X in the season, player's season average was calculated ('PTS_average') for the games played before the game X. Points scored ('PTS') in the game X were then compared to the 'PTS_average'. If 'PTS' was greater than 'PTS_average', the target variable takes the value of 1, if not, it takes the value of 0.

Table 3. History window, size=3, slide=2

Game	Steals	FTA	FGA	FGA-1	FGA-2	FGA-3	TARGET
1	4	12	20	NaN	NaN	NaN	1
2	2	9	15	20	NaN	NaN	0
3	1	9	17	15	20	NaN	1
4	3	4	12	17	15	20	0
5	4	12	14	12	17	15	1
6	2	9	15	14	12	17	0
7	1	9	5	15	14	12	1
8	3	4	8	5	15	14	0

C. *History Window*

With the idea of capturing relevant past information, creation of new ‘history’ features was done. Past information can be captured by looking at every game played before the game X. For each relevant statistic of that game, creation of a new aggregated variable, such as the mean or the median, can be done. Although this does capture the player’s behaviour in the past, it gives the same importance to all the games, neglecting the time factor. Therefore, usage of the history window is suggested. This history window is defined by two parameters: the size of the window and the slide of the window. The size determines how far in the past we want to look, while slide defines by how many steps we want to move the window through the data. Table 3 shows the creation of three new variables: ‘FGA-1’, ‘FGA-2’ and ‘FGA-3’, where ‘FGA-1’ represents the field goal attempts a player took one game before the current one. As it can be seen, *NaN* values are present in the first three examples and as such will not be used for training. With a higher window size, more information is available but also the dimensionality of data increases. With a small window slide, more data will be available, but this causes same information to be present in different instances of the data set, as shown in the green and blue cells in Table 3. More data is preferable which is why the slide of size 1 was chosen. Variables from the games which were played six or more games ago did not seem to carry enough information and therefore a window of size 5 was chosen.

D. *Rest*

Statistics available on Basketball Reference do not offer any information on how well rested the players are. Therefore, a new feature which models this attribute is proposed (1).

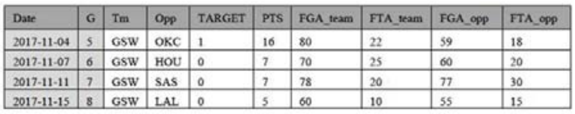
$$rest = 1.5x_1H_1 + 1.3x_2H_2 + 1.2x_3H_3 + 1.1x_4H_4 + x_5H_5 \quad (1)$$

where, x_i takes the value of 1 if a player played a game *i* days before the current game, 0 otherwise, and H_i is a value which helps differentiate "home" from "away" games. If a game, which was played *i* days ago was an away game, H_i takes a value of 1.2, if it was a home game, it takes a value of 1.

The idea behind this formula is that away games take away more energy, both mentally and physically due to travel and absence of friendly audience. Furthermore, the games played yesterday should have a stronger effect on the present in comparison to the games played five days ago.

It needs to be noted that current formula ignores travel distance in away games. For example, Los Angeles Lakers and Los Angeles Clippers play in the same arena which makes the travel distance between them non-existent. On the other hand, Portland and Miami are around 5000km apart. Since we expect that this travel distance may have a significant effect on the predictive power on the modelled variable, in our future research we will try to integrate geographical data in our dataset and further develop the above formula.

Table 4. Data after merging, Stephen Curry



Date	G	Tm	Opp	TARGET	PTS	FGA_team	FTA_team	FGA_opp	FTA_opp
2017-11-04	5	GSW	OKC	1	16	80	22	59	18
2017-11-07	6	GSW	HOU	0	7	70	25	60	20
2017-11-11	7	GSW	SAS	0	7	78	20	77	30
2017-11-15	8	GSW	LAL	0	5	60	10	55	15

E. *Merging Data*

Using the cleaned data, three types of new variables are created for every existing feature: the season average, the average over the last five games and new variables for last five games using the history window. On top of the player data, information about the player’s team and the opposing team was added following the principle showed in Table 4. In the end, every example contains around 1200 variables which include player data, player’s team data and the opposing team data.

V. Model

First games of the season tend to be highly unpredictable and for that reason, the first six games each player has played were removed from the data set.

Before doing any serious modelling it is necessary to split the data into training and test sets. Two approaches were considered. The first approach was to use a chronological split, leaving out few last games of the season of each player for testing and use the rest for training. This raised some concerns since the context is different based on how far the games were in to the season. Perhaps some players’ teams are in the race for the playoffs and may play harder at the end of the season than at the start of the season. Alternatively, maybe some players’ teams are last, with no hope of making the playoffs, and just want to rest, avoid injury and start preparing for the next season. Looking at Fig. 3 it can be seen that time is a factor since volume of games and relations between classes differ from month to month.

The same analysis was done on 2016-2017 season and an identical pattern (Fig. 1) has appeared. This is a strong indicator that in every month of the season except for the last one (April), it is more probable for a player to score less points than his average. This all results in a conclusion that chronological split might not be the optimal choice, which is why we settled for a random split in 80/20 ratio.

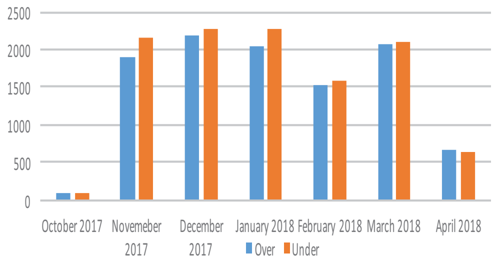


Figure 1. Dependance of time and frequency of over/under

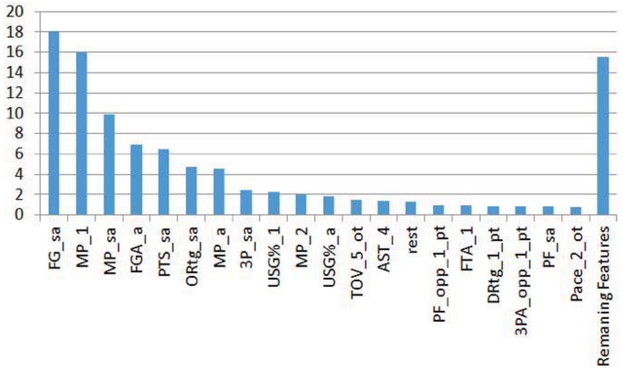


Figure 2. Feature importance

Using Linear SVM with L1 regression, dimensionality of data was reduced to 105 features. The top features are shown in Fig. 2. It is important to distinguish between '_pt' and '_ot' features. The first type ('_pt') represents features that are taken from the players' teams, and '_ot' are features that were taken from the opposing team. 'MP_1' simply represents how many minutes the player played the last game, 'sa' stands for 'season average' and 'a' stands for 'last five games average'.

A description of every statistic can be found in [8]. Two models were tried: Linear SVM and Logistic Regression. In order to prevent overfitting, each model was validated using 5-fold cross validation. All modelling was performed using scikit-learn [7]. After using a grid-search to iterate over the models and parameters, it was found that the Logistic Regression and LinearSVM tend to tperform about equally good, but due to advantage of 0.1%, Logistic Regression was chosen.

Using the random split of data, information about the games in test set could be present in training set due to historical variables. In order to address this concern, the data from season 2016-2017 was collected and used for training, while the data from 2017-2018 was used for testing. This way, we get to keep both datasets representative and independent. The performance of this model did not differ from the performance of the model trained and tested on one-year worth of data. This indicates two great things: patterns in players' behaviour in two consecutive seasons do not differ, and approach of a random data split does not affect the ability of a model to generalize.

Combining datasets from both seasons and then doing a random split, train and test data sets were created, and the process of fine tuning was repeated. The results are presented in Table 5 and Figure 3.

Table 5. Classification report on the test set

	Precision	Recall	F1-score
Under	0.57	0.64	0.60
Over	0.56	0.48	0.51
Average	0.56	0.56	0.56

		Predicted categories	
		Under	Over
True categories	Under	2888	1592
	Over	2200	2000

Figure 3. Confussion matrix

Since no similar models comparable to this model could be found, a comparison was done between the model and the way in which the sportsbooks (sports betting institutions) model the expectations of a player. Sportsbooks tend to offer a certain estimation of a player's points for which they believe that the odds of a player going 'over' or 'under' are 50/50. This estimation is always expressed in a half point number (e.g. 23.5 or 12.5) and a player's season average is used as an anchor. The estimation will differ from a sportsbook to a sportsbook and from a player to a player, but will, in most cases, stay within 1 point range of the player's average. For example, if a player has a season average of 23.4 points per game, it is not uncommon that the sportsbooks will give 50/50 odds of a player scoring more or less points than 24.5 points (1 point higher than his average). One of the reasons for these deviations from the average is player's availability. If a top player is not playing, it is expected from other players to take more shots and score more points than they usually do. Also, sportsbooks tweak their offers based on how people bet on them, and since people tend to bet on a player to score more than his average, rather than less, these offers tend to be higher than players' averages.

Together with the fact that sportsbooks do not give estimations for every player in the league, but only for a certain subset of top players, these facts make the comparison between their models and the model presented in this paper, unfair. Performance of an F1 score of 56.3% and accuracy of 56.7%, presents an improvement of 6.7% in relation to the 50/50 model. Although this comparison is not fair, it is the only one that can be made at this moment.

VI. Future Work

Defensive strength of the opponent (team or a player) could help the performance of the model significantly. Some defensive statistics are present in the model, but there is a lot more information that could be helpful: Who is guarding the player? How well does the opposing team defend the position of that player? If a player scores most points by shooting three-pointers, how well does the opposing team defend the three-point shot? If a player scores a lot of points by shooting free throws, how often does the opposing team foul etc.? Getting all this information could be challenging, but if successful, very helpful.

At the moment of writing this paper, Dallas Mavericks have the best home record and the worst away record in the league. This indicates that teams, therefore players, perhaps Ado not perform the same when playing away and when playing home. An entire new set of features could be created distinguishing statistics of home games and statistics of away games. This would yield a huge amount of new features.

Some teams like Golden State Warriors find themselves leading by 30 points differential (in basketball terms, a blowout) coming into the fourth quarter. Since they are leading by 30, there is no need for starting players to play and risk an injury, therefore the bench players are given a chance to play. This results in some of the players playing fewer minutes than they are used to, taking less shots and scoring less points. Engineering a new feature which indicates the likelihood of a team to blowout an opposing team, or to be blown out by an opposing team could improve the performance of the model.

The same way how data was combined using two seasons, collection of more consecutive seasons could be used for training the model.

On top of everything stated above, there are a lot of possibilities for new features, not only using statistics but using information from social media and media in general. All of that could be examined in the future given the necessary time.

VII. Conclusion

The goal of our research was to identify when the player will outperform and when will he underperform in relation to the expectations of his performance set by past performances. In our case, those expectations were set to be the player's season average.

Using available data, with the help of history window and new 'rest' feature, we managed to achieve an accuracy score of 56.7%. This represents a 6.7% improvement to the 50/50 model. A lot of room is left for improvement and we believe that the inclusion of separate *home* and *away* features, together with more advanced defensive statistics, could have a significant impact on the results.

Two features were engineered by the author: 'missed games' and 'rest'. The first of the two had no impact on the performance of the model. This seems intuitively wrong since it must be significant if a player has missed a certain number of games together with the cause of that. Although 'rest' seems to be informative, coefficients which were used to calculate 'rest' came from intuition and tuning of those coefficients needs to be addressed. The 'rest' feature's formula can be further developed by collection and integration of additional data, most notably the distance that team needs to travel in order to get to the game.

In order to keep as much data as possible, history window slide of 1 was used. Since the importance of history features such as 'MP_5' was near 0, a history window of size 5 was used. The tuning of these hyperparameters will also be addressed in the future research.

Possible applications of this work is seen in sportbooks domain and direct application by teams and players in order to improve their game or exploit the weakneses of the opposing team.

Acknowledgment

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

References

[1] Kevin Wheeler, Predicting NBA Player Performance, http://cs229.stanford.edu/proj2012/Wheeler-PredictingNBAPlayerPerformance.pdf [4.12.2018.]

[2] Douglas Hwang, Forecasting NBA Player Performance using a Weibull-Gamma Statistical Timing Model, 2012, http://www.sloansportsconference.com/wp-content/uploads/2012/02/46-Forecasting-NBA-Player-Performance_DouglasHwang.pdf [30.11.2018.]

[3] Matthew Beckler, Hongfei Wang, Michael Papamichael, NBA Oracle,https://www.mbeckler.org/coursework/20082009/10701_report.pdf [2.12.2018.]

[4] [Sameer K. Deshpande, Shane T. Jensen, Estimating an NBA player's impact on his team's chances of winning, 2016, http://wwwstat.wharton.upenn.edu/~stjensen/papers/shanejensen.basketball2016.pdf [1.12.2018.]

[5] https://www.basketball-reference.com/ [19.1.2018.]

[6] https://www.crummy.com/software/BeautifulSoup/bs4/doc/ [4.12.2018.]

[7] F. Pedregosa etl al. 2011. Scikit-learn: Machine learning in python. Hournal of machine learning research

[8] https://www.basketball-reference.com/about/glossary.html [20.1.2019.]