

# Injury Analysis Based on Machine Learning in NBA Data

Wangwei Wu

Department of Statistics, Sun Yat-sen University, Guangzhou, China

Email: 1025767746@qq.com

**How to cite this paper:** Wu, W.W. (2020) Injury Analysis Based on Machine Learning in NBA Data. *Journal of Data Analysis and Information Processing*, 8, 295-308. <https://doi.org/10.4236/jdaip.2020.84017>

**Received:** July 6, 2020

**Accepted:** November 10, 2020

**Published:** November 13, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

It is a commonplace that the injury plays a vital influence in an NBA match and it may reverse the result of two teams with wide strength disparity. In this article, in order to decrease the uncertainty of the risk in the coming match, we propose a pipeline from gathering data at the player's level including the fundamental statistics and the performance in the match before and data at the team's level including the basic information and the opponent team's status in the match we predict on. Confined to the limited and extremely unbalanced data, our result showed a limited power on injury prediction but it made a not bad result on the injury of the star player in a team. We also analyze the contribution of the factors to our prediction. It demonstrated that player's own performance matters most in their injury. The Principal Component Analysis is also applied to help reduce the dimension of our data and to show the correlation of different features.

## Keywords

Random Forest, Injury, PCA, NBA

---

## 1. Introduction

With the closing of the 2019th NBA final, the Golden State of Warriors was defeated by Toronto raptors by 2:4 which signed a breakdown of a dynasty. During these battles, the injuries coming one after another tear the Warrior apart, leading them not to be able to face the impacts from the Raptors. When it talks to injury of professional athletes, it is a nightmare not only to these players who rely on their health to make a living but also to the people who appreciate these players' performances. It ruined many players who could own a promising future [1] and leads some team to miss their O'Brien Cup by a finger's breadth [2].

In addition, it also may do harm to the assets of the team owners. Therefore, it is in great demand for a classifier to monitor and analyze the players' injury risk in real time in the coming match if they are on the court. With that system, the coach can have a flexible arrangement on the rotation of the athletes.

The former study mainly focused on analyzing the factor related to the injury of players. In Petty DH's work, they interviewed 481 youth pitchers in baseball and found that participants who pitched more than 100 innings in a year were 3.5 times more likely to be injured. In Croisier J L's work, he found that the rate of muscle injury is related to the injury of the athletes. These excellent works could give us some illumination; however, most of them are confined to a specific aspect, which can't give us a complete impression of what leads to injury.

In our work, we build up a pipeline to gather data related to diverse aspects and build up a prediction model and analyze the contribution of different features to filter the most significant ones. We can divide our data into four parts, including the fundamental data of the player, the fundamental data of the team, the relative information of the opponent in the next match, and the player's performance in the matches one week before the match we pay attention to. Then we use the Random Forest, a machine learning method to detect the importance of these factors. And it turns out that the average points in the matches before the next match, the average of minutes and the total number of games the players participated in, age and weight are significant features. Surprisingly, the status of the opponent they will meet counts less, and neither did the number of matches they have attended in the week before. We then use Principal Component Analysis (PCA) method to decrease the dimensions of our model and it displays that many variables are highly relevant. Finally, we do some prediction trials; although it works not very well, it still shows its efficiency in some sense. We hope our model can act as a reference for the coach and manager of the Association to keep their players from injury.

## 2. Background

The relationship between injuries and training status has been widely studied in the sports field. Some researchers interviewed 481 youth pitchers in baseball (aged 9 to 14 years) annually in a 10-year follow-up study. Fisher exact tests were used to investigate risks of injury for pitching more than 100 innings in at least 1 calendar year, starting curveballs before age 13 years, and playing catcher for at least 3 years. And they argue that participants who pitched more than 100 innings in a year were 3.5 times more likely to be injured [3].

In soccer, some researchers find that the rate of muscle injury was significantly increased in subjects with untreated strength imbalances in comparison with players showing no imbalance in preseason by using a standardized concentric and eccentric isokinetic assessment to identify soccer players with strength imbalances [4]. And in NBA, it is said that no correlations were found between injury rate and player demographics, including age, height, weight, and NBA ex-

perience through some descriptive epidemiological study [5].

Apart from those based on traditional statistical methods of prospective study, Alessio Rossi uses some machine learning method to give a effective injury forecasting in soccer with the GPS data. Their classifier can detect 80% of the injuries with about 50% precision, and they give a good trade-off between accuracy and interpretability [6]. Since this paper shows that some overall information in the field is very essential such as distance in meters covered during the training session, it gives us some idea about applying machine learning method but with more instantaneous information such as the latest information in the field.

### 3. Method

Random Forest classifier is a kind of ensemble machine learning method (other two famous algorithms are boosting and bagging [7]) widely applied in classification work. The random forest classifier consists of a combination of tree classifiers (another machine learning method called C4.5, where every variable can be regarded as a leaf in a tree [8]) where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector. The random forest classifier used for this study consists of using randomly selected features or a combination of features at each node to grow a tree. Any examples (pixels) are classified by taking the most popular voted class from all the tree predictors in the forest. Design of a decision tree required the choice of an attribute selection measure and a pruning method. There are many approaches to the selection of attributes used for decision tree induction and most approaches assign a quality measure directly to the attribute. The most frequently used attribute selection measures in decision tree induction are the Information Gain Ratio criterion and the Gini Index. The random forest classifier uses the Gini Index as an attribute selection measure, which measures the impurity of an attribute with respect to the classes. For a given training set  $T$ , selecting one case (pixel) at random and saying that it belongs to some class  $C_p$  the Gini index can be written as:

$$\sum_{i \neq j} f(C_i, T) / |T| \left( f(C_j, T) / |T| \right) \quad (3.1)$$

where  $f(C_i, T) / |T|$  is the probability that the selected case belongs to class  $C_i$ .

Each time a tree is grown to the maximum depth on new training data using a combination of features. These fully grown trees are not pruned. As the number of trees increases, the generalization error always converges even without pruning the tree and overfitting is not a problem because of the Strong Law of Large Numbers. The number of features used at each node to generate a tree and the number of trees to be grown are two user defined parameters required to generate a random forest classifier. At each node, only selected features are searched for the best split. Thus, the random forest classifier consists of  $N$  trees, where  $N$  is the number of trees to be grown, which can be any value defined by the user.

To classify a new dataset, each case of the datasets is passed down to each of the  $N$  trees. The forest chooses a class having the most out of  $N$  votes, for that case.

We choose random forest as our main method for this analysis not only because it is a relatively stable machine learning method and has kind of resistance to unbalanced data but also the model from it would not lack in interpretation for this algorithm comes from CART, another method known for its Interpretability. We can use this interpretability to do some factor analysis for our features.

## 4. Analysis and Result

### 4.1. Original Data

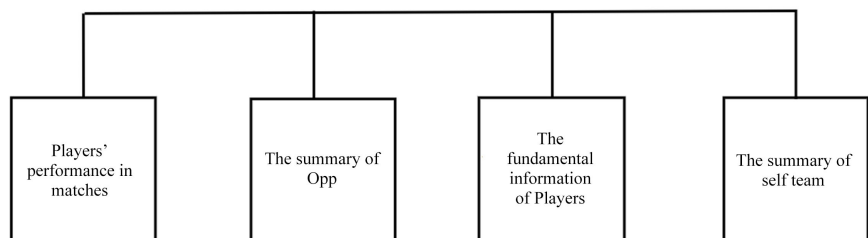
To gather the data needed for this analysis, we collect injury data from pro sports transactions website [9] where includes all the injury recordings from 2011 to 2019 in NBA. Moreover, we removed all the injuries not happening in the court, because anyone familiar with the NBA can tell that many injuries not at the field would be just an excuse for rotation, but in fact the player is in good state. We get other information we need from BASKETBALL REFERENCE website [10] containing the matches in 2015-2016 season and the information of players who played at least 20 minutes averagely. All the information can be classified into four parts shown in **Figure 1**.

Some basic introduction of the data is as follows:

- 1) We only adopt the data of Season 2016 for analysis.
- 2) The threshold for players we pay attention to is that they must have an average playing time of at least 20 minutes.
- 3) In that condition, there are 277 players including the number of some players traded to other teams.
- 4) Only 27 injuries of 824 in total are happened on the field and among these 277 players.
- 5) We have 13,975 recorded data bars in all.

We can tell that it is completely unbalanced data. The injury cases are only 27 but we have 13,975 cases in total. We could not rely on it to give us an accurate prediction even we neglect the latent variables out of court playing a significant role in players' injury. However, we can still do some factor analysis through it.

A sample data is shown below: the meaning of the feature's name will be introduced in the following section.



**Figure 1.** Four parts of the features applied in random forest.

1) The player's performance in match.

Feature name	Value	Feature Name	Value	Feature Name	Value
<b>name</b>	Steven Adams	<b>3P</b>	0	<b>PF</b>	4
<b>G</b>	1	<b>3PA</b>	0	<b>PTS</b>	6
<b>Date</b>	20160102	<b>3P%</b>		<b>GmSc</b>	4.7
<b>Age</b>	22 - 100	<b>FT</b>	0	<b>+/-</b>	4
<b>Tm</b>	OKC	<b>FTA</b>	0		
<b>xa0</b>		<b>FT%</b>			
<b>Opp</b>	SAS	<b>ORB</b>	1		
<b>xa1</b>	W (+6)	<b>DRB</b>	6		
<b>GS</b>	1	<b>TRB</b>	7		
<b>MP</b>	29:34:00	<b>AST</b>	1		
<b>FG</b>	3	<b>STL</b>	0		
<b>FGA</b>	4	<b>BLK</b>	1		
<b>FG%</b>	0.75	<b>TOV</b>	2		

2) The summary of the team (self and the opponent).

Feature Name	Value	Feature Name	Value	Feature Name	Value	Feature Name	Value
<b>Rk</b>	1	<b>FT%</b>	0.788	<b>MOV</b>	11.63	<b>TOV%</b>	13.5
<b>Team</b>	Golden State Warriors*	<b>ORB</b>	770	<b>SOS</b>	-0.28	<b>DRB%</b>	74.9
<b>G</b>	82	<b>DRB</b>	2873	<b>SRS</b>	11.35	<b>FT/FGA</b>	0.198
<b>MP</b>	19,780	<b>TRB</b>	3643	<b>ORtg</b>	115.6	<b>Arena</b>	Oracle Arena
<b>FG</b>	3532	<b>AST</b>	2491	<b>DRtg</b>	104	<b>Attend.</b>	803,436
<b>FGA</b>	7140	<b>STL</b>	785	<b>NRtg</b>	11.6	<b>Attend./G</b>	19,596
<b>FG%</b>	0.495	<b>BLK</b>	555	<b>Pace</b>	99.8		
<b>3P</b>	982	<b>TOV</b>	1211	<b>FTr</b>	0.259		
<b>3PA</b>	2562	<b>PF</b>	1585	<b>3PAr</b>	0.359		
<b>3P%</b>	0.383	<b>PTS</b>	9503	<b>TS%</b>	0.597		
<b>2P</b>	2550	<b>Age</b>	28.2	<b>eFG%</b>	0.563		
<b>2PA</b>	4578	<b>W</b>	67	<b>TOV%</b>	13.2		
<b>2P%</b>	0.557	<b>L</b>	15	<b>ORB%</b>	22.8		
<b>FT</b>	1457	<b>PW</b>	67	<b>FT/FGA</b>	0.204		
<b>FTA</b>	1850	<b>PL</b>	15	<b>eFG%</b>	0.486		

## 3) The fundamental status of the player.

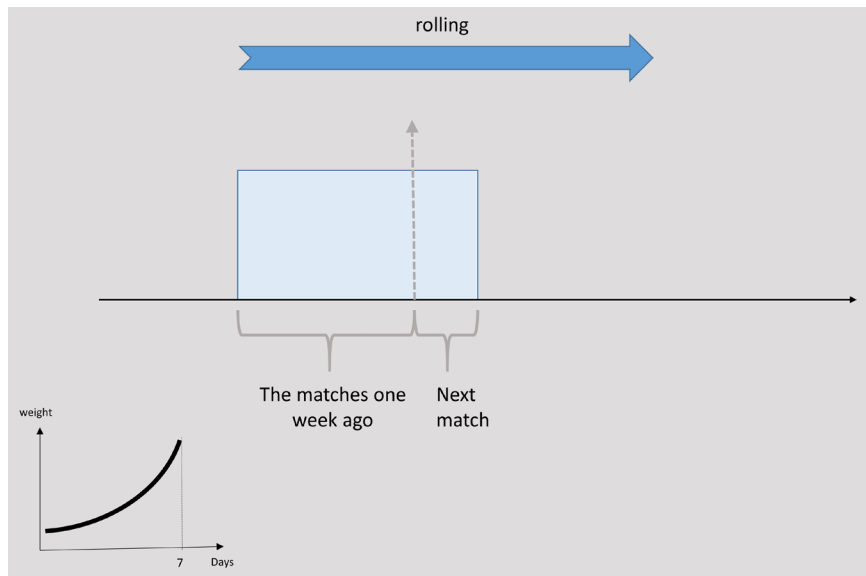
Feature Name	Value	Feature Name	Value
<b>Rk</b>	3	<b>OWS</b>	3.3
<b>Player</b>	StevenAdams	<b>DWS</b>	3.1
<b>Pos</b>	C	<b>WS</b>	6.5
<b>Age</b>	23	<b>WS/48</b>	0.13
<b>Tm</b>	OKC	<b>OBPM</b>	-0.7
<b>G</b>	80	<b>DBPM</b>	1.2
<b>MP</b>	2389	<b>BPM</b>	0.6
<b>PER</b>	16.5	<b>VORP</b>	1.5
<b>TS%</b>	0.589	<b>shoots</b>	Right
<b>3PAr</b>	0.002	<b>height</b>	213 cm
<b>FTr</b>	0.392	<b>weight</b>	120 kg
<b>ORB%</b>	13	<b>BLK%</b>	2.6
<b>DRB%</b>	15.4	<b>TOV%</b>	16
<b>TRB%</b>	14.2	<b>USG%</b>	16.2
<b>AST%</b>	5.4		
<b>STL%</b>	1.8		

## 4.2. Data Processing and Data Introduction

The raw data is not suitable for putting into the model. We should transform them into some characters useful for our research beforehand. The preprocessing methods in this section mainly referred to sliding window with moving weighted average and missing value imputation.

### 4.2.1. Players' Performance in Matches

Since we pay attention to factors that happen just before the match, so the matches the player played before the match focused on is in our consideration. To be convenient, we could assume that all the matches in one week before are important. And we adopt a method in Time series called moving weighted average [11]. In this method, we assume that the performances in each match obey an exponential distribution—the closer to the focused match, the larger its weight is. And we could use this adjustment sliding through all the matches that players played in this season to produce our processed data as shown in **Figure 2**. Besides that, we abstract some other information useful for our research from the rolling data flow, for example, we can obtain whether there are some back to back matches during seven days' matches, and how many days players are out of court, and how many home and away home matches the players attend respectively. The detailed introduction is shown in **Table 1**.



**Figure 2.** The workflow of the sliding window.

**Table 1.** The characters of performance in matches.

name	description	method of disposal
<b>Points_AVG</b>	Average of points the player gets during these matches	
<b>MP_AVG</b>	Average time of playing during these matches	
<b>FG_AVG</b>	Average field goals of the player during these matches	
<b>FGA_AVG</b>	Average field goal attempts during these matches	
<b>TRB_AVG</b>	Average rebounds of the player during these matches	use weighted average
<b>AST_AVG</b>	Average assists of the player during these matches	
<b>STL_AVG</b>	Average of Steal during theses matches	
<b>BLK_AVG</b>	Average of blocks during these matches	
<b>P3_AVG</b>	Average of 3-Point field goals during these matches	
<b>PA3_AVG</b>	Average of 3-point field goal attempts during these matches	
<b>+/-_AVG</b>	Average of +/- of the player during these matches	use average directly
<b>M_days</b>	Total number of matches during these days	
<b>Ms_days</b>	Total number of matches starting during these days	just sum up
<b>Home_days</b>	Total number of matches held at home	
<b>Host_days</b>	Total number of matches held away from home	
<b>B2B</b>	Whether the routine includes back to back games	

#### 4.2.2. The Fundamental of Players

In this section, we pay attention to players’ fundamental state. For example, we draw attention to their weight, height, and their shooting habits—do right or left hands influence the risk of injury? And we also have a summary of players’ performance in their career. Therefore, we can have a general impression about what style of players get hurt more easily.

There are also some advanced indexes which combined some fundamental data into an index giving a summary of some specific behavior. For example, *USG%* is an index which measuring the percentage of team plays used by a player while he was on the floor. The detailed introduction of these features is demonstrated in [Table 2](#).

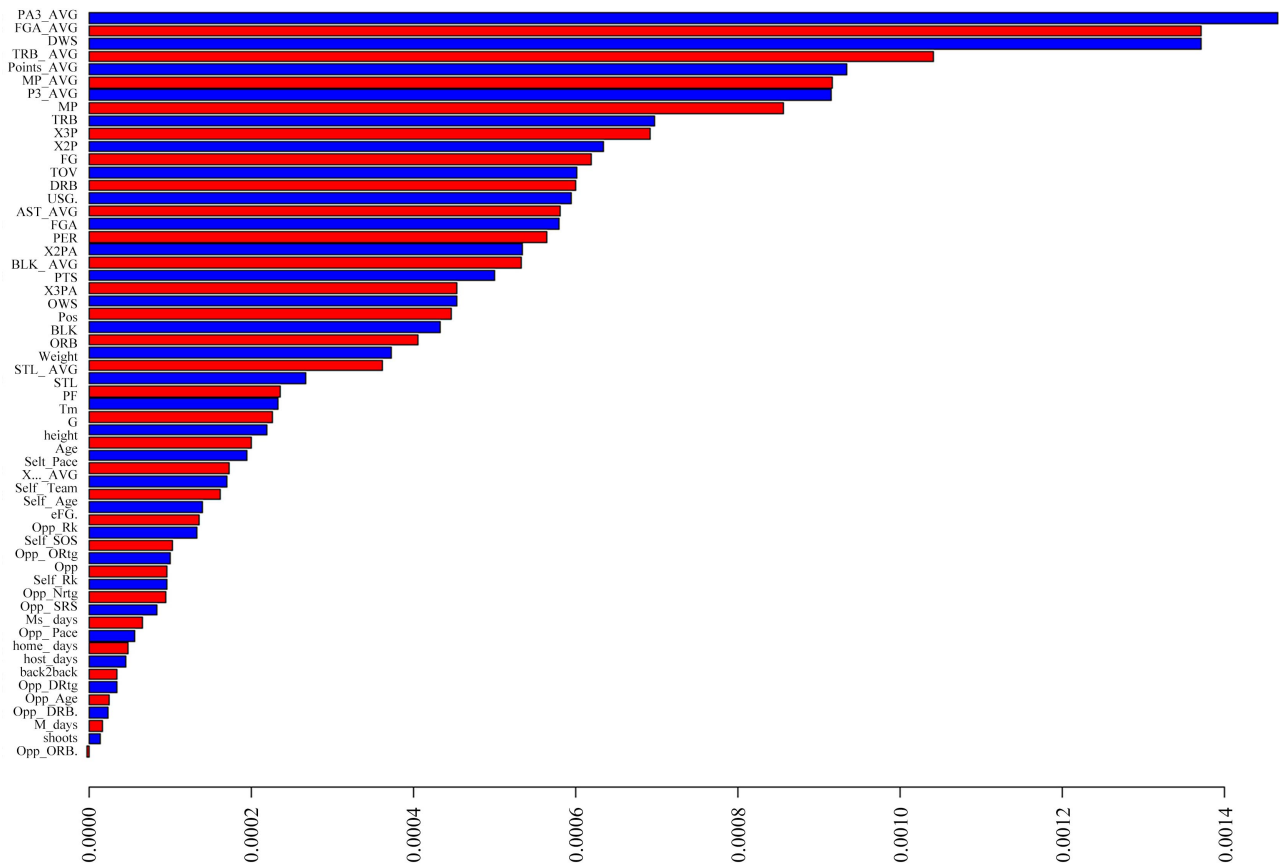
### 4.2.3. The Self and Opponent Team's State

Except for player's state, we should also consider in the team-level, which include not only the team the player stays in but also the team he will face in the next match, could also contribute to the player's injury in some cases. For example, if the style of player is in fast pace, then some players may get hurt during rapid shifts. And if the opponent is skilled in defensive strength, then the player may get hurt easily too. Therefore, for self and opponent team, we should consider different index. And many of them are assessed by ESPN [12]. And all the variables are displayed in Appendix ([Table A1](#)).

## 4.3. The Result

### 4.3.1. The Variable Importance

We used the permutation test in random forest to evaluate the importance of all the features in our model. And the result is shown in [Figure 3](#).



**Figure 3.** The features importance of the random forest model based on the permutation method.



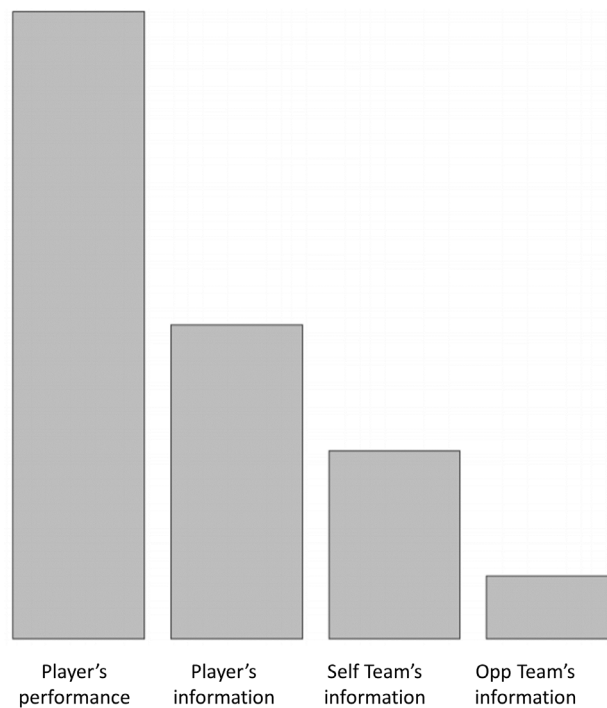
**Table 2.** The characters of the fundamental of players.

name	description	Notes
<b>Pos</b>	The position the player plays in	There are 5 positions in total.
<b>MP</b>	The average minutes the player plays during the career	
<b>PTS</b>	The average points the player gets during the career	
<b>STL</b>	The average steals the player gets during the career	
<b>BLK</b>	The average blocks the player gets during the career	
<b>TOV</b>	The average turnovers the player gets during the career	
<b>eFG%</b>	Effective Field goal percentage	
<b>Tm</b>	The team the player stays in	
<b>USG%</b>	Usage percentage	An estimate of the percentage of team plays used by a player while he was on the floor.
<b>OWS</b>	Offensive win shares	An estimate of the number of wins contributed by a player due to his offense.
<b>DWS</b>	Defensive win shares	An estimate of the number of wins contributed by a player due to his defense.
<b>PER</b>	Player efficiency rating	A measure of per-minute production standardized such that the league average is 15.
<b>weight</b>	Their weight in that season	
<b>height</b>	Their height in that season	
<b>shoots</b>	Their shooting behavior include right and left	

It is not surprising to see that the player's performance in recent days is very important in their risk of injury. In fact, the latest indexes are more important than the other three parts of general descriptions. We also summarize the average importance of every part in **Figure 4**. We can tell that the overall information of one player is influential too, but which team they will face in next match is least significant.

From this chart, we can tell that:

- 1) When a player tries more shoots especially in 3-point shoot, they can get hurt in a larger possibility.
- 2) The more time the players play in the field, the more chance they may get hurt.
- 3) Interesting enough, the information of player themselves, DWS, an advanced index indicating the number of wins contributed by a player due to his defense, leads to the chance of injury. We can speculate that a fierce defense may cause damage to the defense player himself.
- 4) Weight and height are not as important as I thought.
- 5) The frequency of routine is not important equally. Even back to back games may not lead to high risk of injury.



**Figure 4.** The importance of four parts.

#### 4.3.2. Factor Analysis

We can tell that we have too many variables in our model and in fact, many of them are in high correlation. It is beyond doubt that when a player plays more time, they have more chance to give goals attempts and get more points. So, when we use PCA or some other techniques to shrink the scale of our variables, we can make our model displaying more briefly.

We can tell that these variables can be divided into 10 more significant dimensions as shown in **Figure 5**. And some variables that contribute to the first dimension are just the ones important in influencing the injury of players, which is shown in **Figure 6**. Points\_Avg, Mp\_AVG, FGA\_AVG ... they are all some import variables in our Random Forest model and they are just in high correlation as we predict. However, opp\_Rk and opp\_SRS are two variables essential in third dimension, but they contribute least in the random forest model.

#### 4.3.3. Prediction

We collected the relative NBA data in 2017 as our testing set which contained 10,199 pieces of items but only 15 of them were injured on the court, which means it showed a more serious unbalance than the training set. It is not beyond our expectations that our model did not perform well on the test dataset, since the data is so unbalanced after some data clean. However, even in this situation, we still have some correct predictions on injury events and most of them are the star players in the team which should deserve more attention than others, as you can see in Appendix (**Table A2**), which means this method of analyzing injury occurred in the field does work.

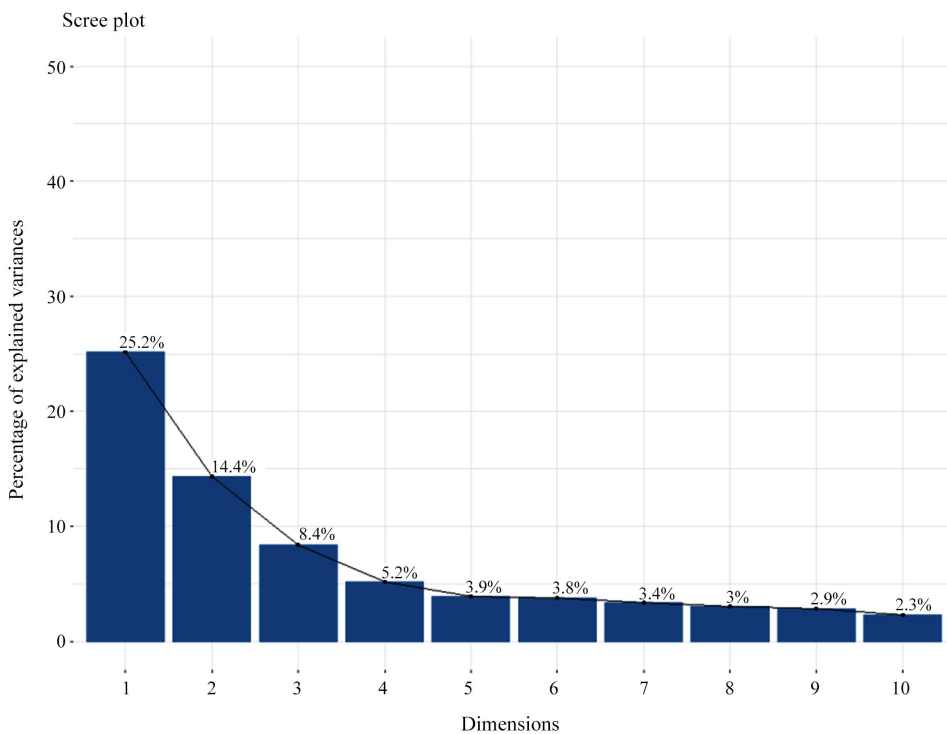


Figure 5. 10 dimensions of PCA.

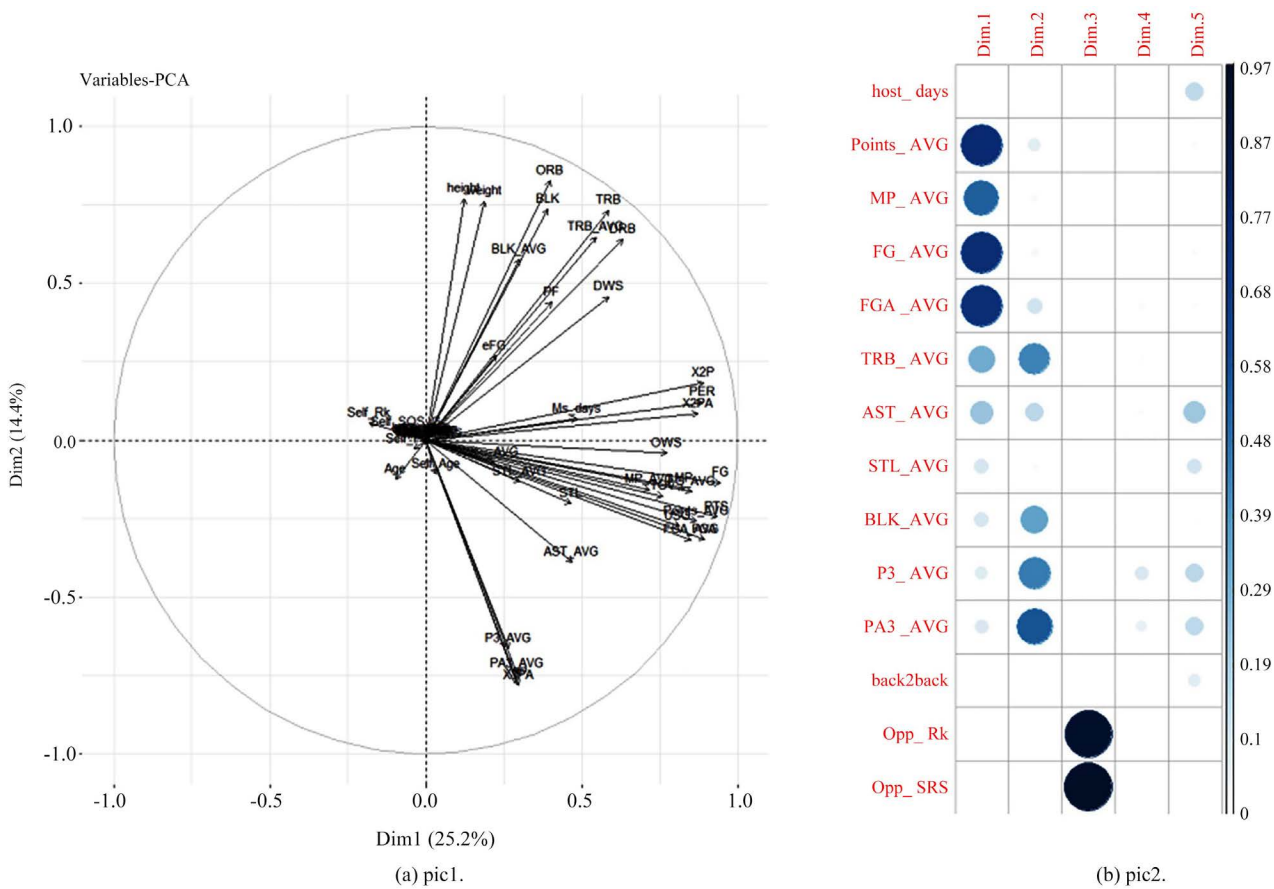


Figure 6. The pie chart of the former two PCA dimensions and the contributions of different features to the dimensions.

## 5. Conclusions & Discussion

In our work, we use some machine learning methods, or to be more specific, the Random Forest method to build up a model to analyze the latent factors that have some correlation with the injury of players in NBA. We find it is in high correlation with the players' performance on the most recent matches the player has played. Playing more actively and taking more trials on shooting, they are more easily involved in risks of injury in the next match. It is something that coaches can pay attention to. Besides that, we also exclude some factors which seem to be relative to injuries. The weight and height is part of them. And more counterintuitive, it seems that the intensive routine won't result in injury too. Even when player gets some back to back games, their chance for injury remains constant.

We also do some job to decrease the scale of variables in our model. We find that some of the most important variables in our model are highly relevant to each other, which means we don't need to consider all of them simultaneously but combine them into some new factors, which will make the model look concise and still won't lack interpretation.

As we have said, the dataset is so unbalanced and the injury events are so rare compared to the not injured data items. So, others can try to use more data of past seasons to train a more sensible and accurate model to predict the injury event in the future. And from another perspective, maybe there exist other factors we neglect which counts more in the injury of players. Since getting hurt is such a rare event, we can hardly get enough factors as we hope to get an ideal result; however, the more we know, the better we can do in prediction. And such a method can also be used in other fields, especially in some one-to-one sports, such as tennis and badminton. In this field, the variables we need to consider are less so we can speculate that this method can be applied with a more ideal result.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

- [1] The 5 Best NBA Careers Ruined by Injuries. <https://bleacherreport.com/articles/2795270-the-5-best-nba-careers-ruined-by-injuries#slide1>
- [2] Vulnerable Warriors Are Losing the Injury Lottery in NBA Finals. <https://www.thestar.com/sports/raptors/opinion/2019/06/04/vulnerable-warriors-are-losing-the-injury-lottery-in-nba-finals.html>
- [3] Petty, D.H., Andrews, J.R., Fleisig, G.S., *et al.* (2004) Ulnar Collateral Ligament Reconstruction in High School Baseball Players: Clinical Results and Injury Risk Factors. *The American Journal of Sports Medicine*, **32**, 1158-1164. <https://doi.org/10.1177/0363546503262166>
- [4] Croisier, J.L., Ganteaume, S., Binet, J., *et al.* (2008) Strength Imbalances and Preven-

- 
- tion of Hamstring Injury in Professional Soccer Players: A Prospective Study. *The American Journal of Sports Medicine*, **36**, 1469-1475.  
<https://doi.org/10.1177/0363546508316764>
- [5] Drakos, M.C., Domb, B., Starkey, C., *et al.* (2010) Injury in the National Basketball Association: A 17-Year Overview. *Sports Health*, **2**, 284-290.  
<https://doi.org/10.1177/1941738109357303>
- [6] Rossi, A., Pappalardo, L., Cintia, P., *et al.* (2018) Effective Injury Forecasting in Soccer with GPS Training Data and Machine Learning. *PloS ONE*, **13**, e0201264.  
<https://doi.org/10.1371/journal.pone.0201264>
- [7] Quinlan, J.R. (1996) Bagging, Boosting, and C4. 5. *AAAI/IAAI*, **1**, 725-730.
- [8] Quinlan, J.R. (2014) C4. 5: Programs for Machine Learning. Elsevier, Amsterdam.
- [9] <https://www.prosportstransactions.com/basketball/>
- [10] <https://www.basketball-reference.com/>
- [11] Shumway, R.H. and Stoffer, D.S. (2017) Time Series Analysis and Its Applications: With R Examples. Springer, New York.
- [12] <http://espn.go.com/>

## Appendix

**Table A1.** The characters of team's state.

name	description	NOTES	type
SOS	Strength of Schedule	A rating of strength of schedule. The rating is denominated in points above/below average, where zero is average.	
Team_Pace	Measure the pace of the team.	An estimate of possessions per 48 minutes.	self team
Age	Average age of the players in the team		
Rk	The ranking of the team in that season		
SRS	Simple Rating System	A team rating that takes into account average point differential and strength of schedule. The rating is denominated in points above/below average, where zero is average.	
ORtg	Offensive Rating	An estimate of points produced (players) or scored (teams) per 100 possessions.	
DRtg	Defensive Rating	An estimate of points allowed per 100 possessions.	
NRtg	Net Rating	An estimate of point differential per 100 possessions.	opp team
Oppo_Pace	Measure the pace of the opp team	An estimate of possessions per 48 minutes.	
ORB%	Offensive Rebound Percentage	An estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor.	
DRB%	Defensive Rebound Percentage	An estimate of the percentage of available defensive rebounds a player grabbed while he was on the floor.	
Opp_Rk	The ranking of the opp team in that season		

**Table A2.** Performance in prediction.

	injury	normal
injury_pre	4	58
normal_pre	11	10,126