NBA Draft

How do they choose players ?

Giuseppe Cavallaro, Xiaolong Wang, Matteo Mognetti, Piotr Kotylo
Universita' degli studi di Milano-MEF

Introduction

The NBA draft is an annual event dating back to 1950 in which the thirty teams from the National Basketball Association (NBA) can draft players who are eligible and wish to join the league. These are typically college basketball players, but international players are also eligible to be drafted. College players who have finished their four-year college eligibility are automatically eligible for selection, while the underclass men have to declare their eligibility and give up their remaining college eligibility. International players who are at least 23 years old are automatically eligible for selection, while the players younger than 22 have to declare their eligibility. No player may sign with the NBA until he has been eligible for at least one draft. Considering the eligible players, we want to know what are the main features that a player must have in order to be drafted.

Main Objectives

Based on a list of 393 eligible players, from 2012 until 2015, and 20 features for each one. The features are: Height; Weight; G (Games Played); MP(Minutes Played per game); FG(Field Goals made per game); FGA(Field Goals attempted per game); 2P (2-points field goals made per game); 2PA(2-points field goals attempted per game); 3P (3-Points field Goals made per game); 3PA(3-Point field Goals attempted per game); FT (Free throws made per game); FTA(Free throws attempted per game); TRB(Total Rebounds per game); AST(Assists per game); STL(Steals per game); BLK(Blocks per game); TOV(Turnovers per game); PF(Personal fouls per game); PTS(Points per game); Pick (Yes if drafted, otherwise No). Pick is the response variable.

Methods

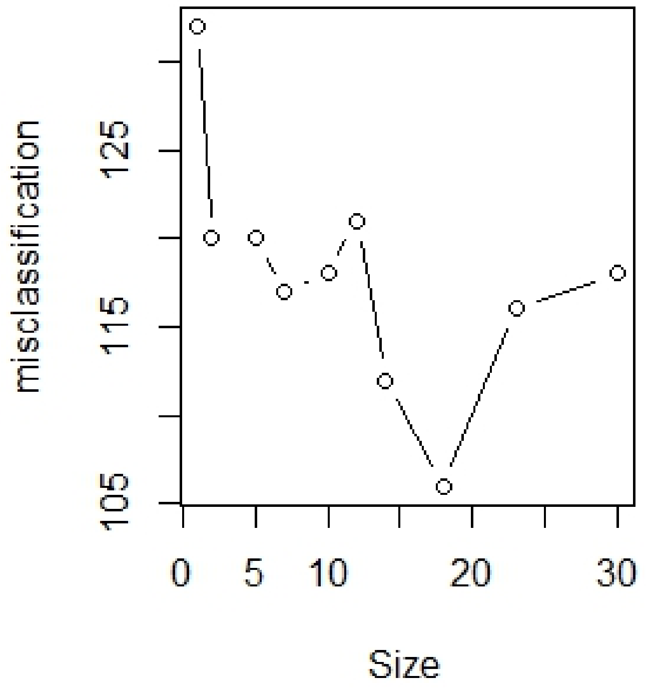
In order to determine which features are preponderant for a player to be drafted, we build a Classification Tree.
Classification Tree is used for categorical response variable, without any assumptions on distributions of parameters, so we try to predict the class probabilities at the leaves. Classification tree partitions a data set into exhaustive (every element of the data set is part of just one node) and non-overlapping nodes. At each node of the tree, the response of interest is summarized by a frequency distribution.
The objective of a split is to increase the homogeneity of the resulting smaller data sets with respect to the target variable. We continually divide the data set by creating node splits of smaller and smaller data sets. We used the Pruning back tree in order to defence against overfitting. We hope that the simpler tree does better when it used to predict new observations, with the test data set. At each step, we remove the split that contributes least to reduce the number of misclassification. In a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. In interpreting the results of a classification tree, we are often interested not only in the class prediction corresponding to a particular terminal node region, but also in the class proportions among the training observations that fall into that region.
Our aim is to assign an observation to a given region to the most commonly occurring class of training observation in that region.
The classification error rate is simply the fraction of the training observations in that region that do not belong to the most common class:

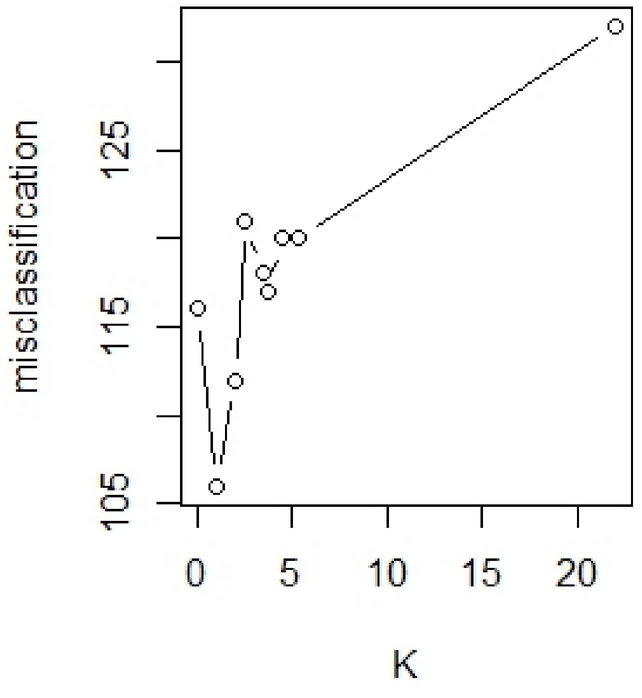
$$E = 1 - max_k(P_{mk}) \tag{1}$$

P_{mk}= represents the proportion of training observations in the m-th region that are from the k-th class.
We use the classification error rate for pruning the tree, because it is preferable for prediction accuracy.
In order to evaluate the performance of a classification tree, we estimate the test error. We split the observations into a training set and a test set, building the tree using the training set and evaluate its performance on the test set. Next we consider whether pruning the tree might lead to improved results.In order to achieve some pruning, we minimize the cost complexity of the tree:
Min:

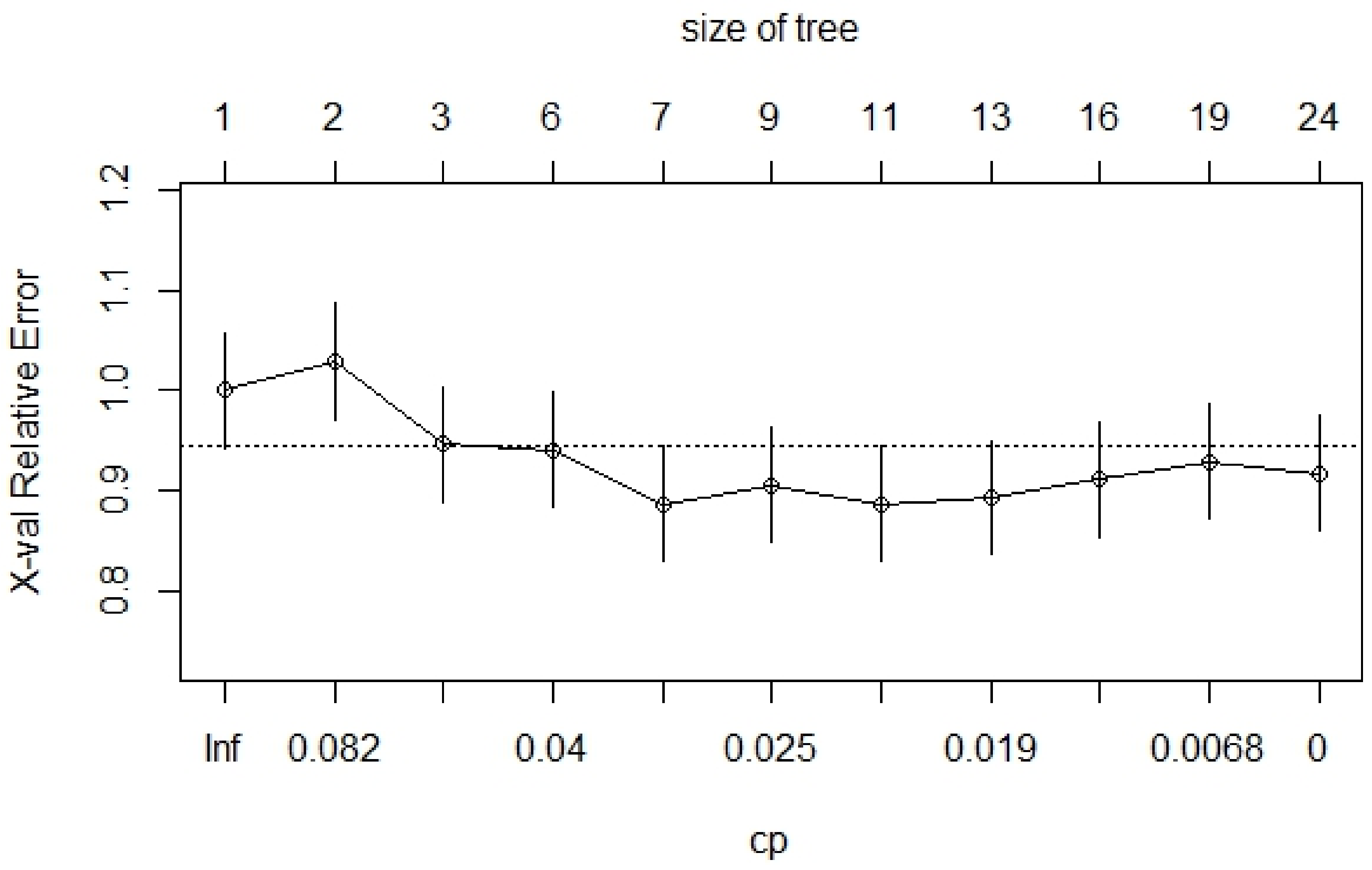
$$D(T) + cp|T| \tag{2}$$

Where: $D(T)$ is the classification error and $|T|$ is the size of the tree , the number of its terminal nodes, and cp is a penalty term.
Performing cross-validation we determine the optimal level of tree complexity, we use the classification error rate to guide the cross-validation and pruning process. The result is that the tree with 8 terminal nodes has a good trade-off between complexity and size. The pruned tree with 8 nodes, applied on the test set, performing correctly the 68% of the observations.



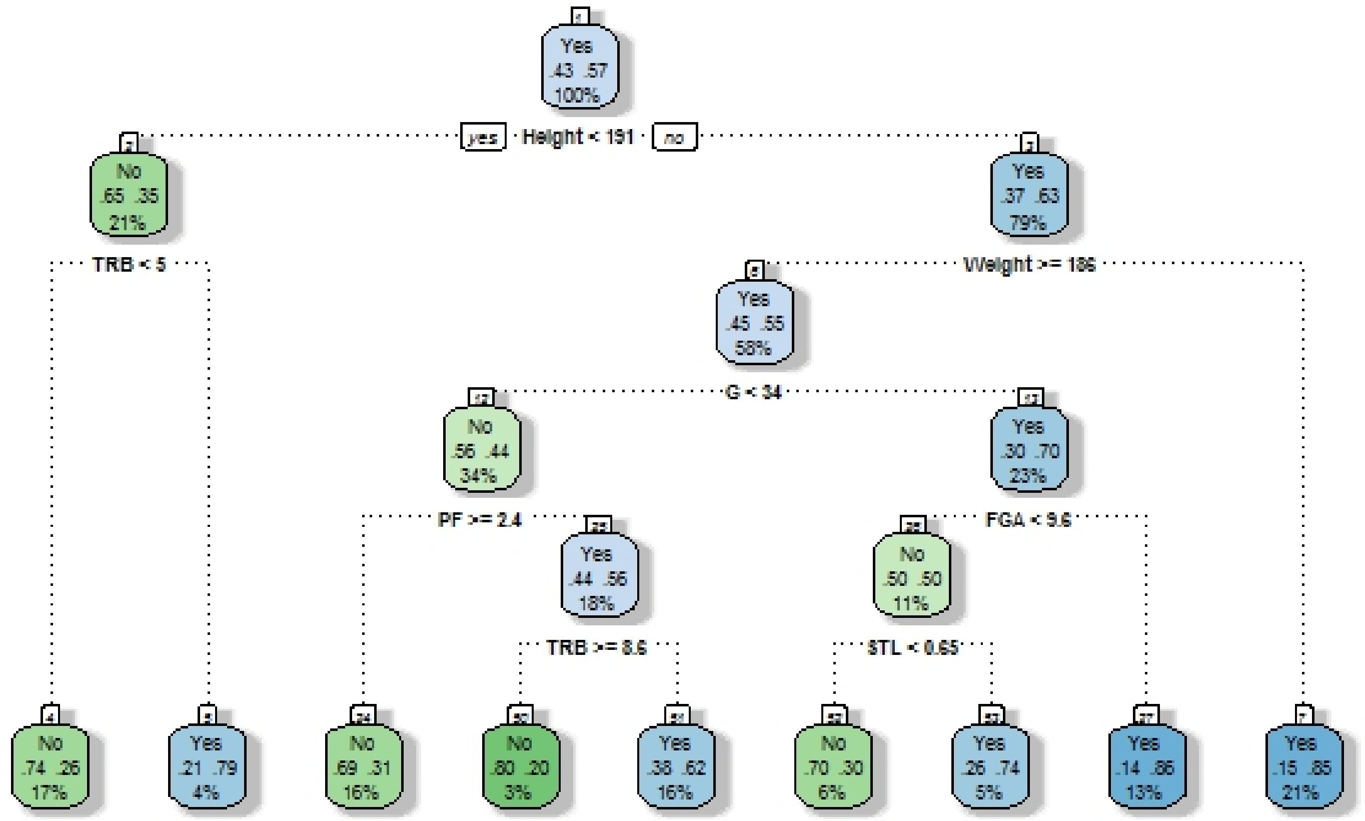


K is the number of folds of the cross-validation.



The cp represents the cost-complexity parameter of our model.
The x-val represents the number of cross-validations.

Results



Conclusion

In our analysis all players are eligible, with young age, and all they play for a team. The higher probability to be drafted is associated with the players well-build, so the possibility of improve their technical characteristics, being them good choices. The numbers of games played is a main features, because it shows the attitude of the player and his position in the team, good players play more. The personal foals variable, together with the total rebounds, can be considered as a relevant features for defence players, so they are characteristics that a good defence player should have. A player with a high field goals attempted and a good height has good probability of being drafted. Meanwhile, in order to have some chance to be drafted, if he hasn't good attitude to throw, it should have good attitude at stealing the balls. The players which aren't so tall, but have a very good attitude in taking the rebounds could be drafted, but the probability is really poor. As we can aspect the height is an important feature for a player, but isn't enough, together with it other attitude should have a player in order to be drafted. A critical point of our model is that the main features reported don't explain player drafted with very good game vision. This is a preponderant feature for a play-maker, so the quality of a player is a feature that isn't measurable and that should be determined scouting player by player.

References

An Introduction to Statistical Learning, Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013, Springer.
The Elements of Statistical Learning, Authors: Trevor Hastie, Robert Tibshirani, Jerome Friedman,2008, Springer.
$http://espn.go.com/mens-college-basketball/statistics/player/_/stat/scoring-per-game/sort/avgPoints/year/2015$
$http://www.basketball-reference.com/draft/NBA_2014.html$
$http://www.ncaa.com/rankings/basketball-men/d1/associated-press$