

Forecasting Basketball Games Using Bayes Networks

April 27, 2011

1 Problem Definition

In recent years, sports analytics has evolved into an important field. Owners spend millions upon millions of dollars in their teams and want to avoid spending money on players that will not help the team's performance. Analysts use the majority of their time trying to pick which teams will win the championship every year and fans want to win their own fantasy seasons for bragging rights. With "Sabermetrics," the baseball statistical community has essentially worked their game down to a science. Ask an analyst and they can inform you who will have a big year, predict who will win the Cy Young award, and who will win the batting title. The same cannot be said for basketball, which is a sport that relies more heavily on interaction between teammates than baseball. This causes a problem in determining how individual player statistics factor into predicting wins and losses, which is what we intend to solve.

2 Related Work

Efforts into predicting basketball games thus far have not been huge successes. Using different machine learning approaches [1] were able to achieve a good accuracy, reaching their best results using the naïve bayes approach, with an accuracy of 67%. On the other hand, their approach was not flexible for projecting how trades and free agent signings would effect teams as they used only team statistics.

[2] tried a different approach, using Weighted Likelihood to predict team winning percentages for the entire season. They used weighted probability distributions on team statistics, obtaining a very good accuracy on some tests, but also obtaining completely wrong results in other instances. This high variability, and the fact that it is not possible to predict the result of a single game during the championship, makes this approach useless for our purposes.

[3] used a Markov Logic Network and Bayesian Logic to predict the outcome of games in two different ways. The MLN fared better with an average accuracy of 76%

while Bayesian Logic only attained 64%. However, their inference scheme seemed to ignore many of the variables involved in a game that we would like to include.

3 Approach

3.1 Gathering Statistics

All of our statistics were collected from basketball-reference.com. We approached the problem with individual player statistics in mind to be able to account for player movement and injuries during the season. We gathered both basic and advanced metrics for all individuals playing in the 2008-2009 NBA season. The basic statistics were simple measures that can be viewed as they occur during a game (such as the number of field goals made or the number of steals). The more advanced metrics were formulas that were often created by experts, including the Player Efficiency Rating (PER) as calculated by John Hollinger, which is a well known and fairly effective attempt to measure a player's per minute effectiveness.

As far as team statistics are concerned, we did not use as many as other prediction schemes have. This is because we wanted to be able to account for player movement and injuries as the season progressed. The only team statistics we used were the average number of points a team scored and the average number of points scored against it. There was a simple correlation in that season's playoff results where the team that scored more points and gave up fewer often won the series (as one would expect).

3.2 Bayesian Network

We tested two separate approaches to predicting the outcome of games. The first was to use only individual player statistics and the second was to use a combination of individual and simple team metrics.

3.2.1 Individual Matchups

To compare players, we created a Bayes Network which matched up all of the players on the two teams and attempted to predict which player would win the individual matchup (pitting the player from the first team's offense

and pitting it against the player from the second team’s defense and vice versa). Initially, the offensive statistics used were a combination of the number of field goals and three pointers attempted on offense and the number of defensive rebounds, steals, and blocks on defense. This way we measured the offensive player’s usage against the defensive player’s involvement.

In our second approach, we added in unconventional statistics to see what effect they could have on the prediction algorithm. We compared player heights, weights, and ages. The heights and weights were straightforward comparisons, where the taller and the lighter players were granted an edge. The age of a player was slightly different as, while looking at the statistics, we found that many NBA players seem to peak statistically at the age of 26.

Finally, we used the PER of each player, matching up each of the most efficient players on the two teams, then the second most efficient, through the number of players that the smaller roster had. While this is not an entirely accurate comparison, it did match one team’s best player to another, which can be what the end of an NBA game comes down to. This is the individual matchup system that performed the best, so we used it as the basis for the individual and team combined approach. The equation used to calculate the probability of the home team’s player winning this matchup is shown in (1) below.

$$0.5 + \frac{\min(\max(PER_1 - PER_2, -20), 20)}{40} \quad (1)$$

3.2.2 Individual and Team

For the hybrid approach considering both individual player and team past performance, we added very little in terms of extra calculations. This was intentional, as a simple team metric derived from the past performance of a team is only useful if the players on the team have not changed at all. Thus, we decided to base the team

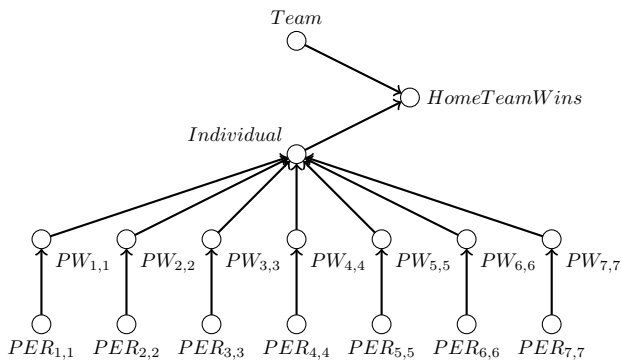


Figure 1: The Team+Individual Bayes Network.

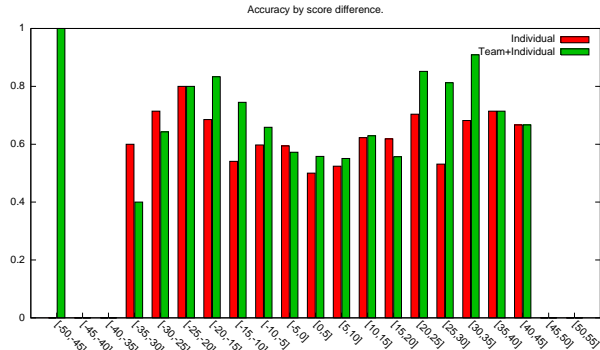


Figure 2: Accuracy by score difference for the two Bayes Networks.

comparison entirely on the average number of points the team scored and the average number of points the team allowed. While it can be argued that this metric will also change with the personnel on the floor, we opted to think of it as reflecting the style of the coach. An offense that is based on fast breaks is more likely to have a high number of points per game than an offense in a halfcourt offense. The comparison was calculated as the ranking of the team in each category less the ranking of their opponent in the same category. This was then normalized to a figure between -0.5 and 0.5 and added to 0.5 , in a similar way as in (1), which provided a method of determining a probability that the home team would win. Since we still wanted to focus on the individual player matchups, we weighted the conclusion that those came to as 70% and the decision from the team comparison as only 30%. Admittedly, these figures are arbitrary for the moment, but given more time we can find the appropriate balance.

3.3 Elimination Order

In determining whether the home team would win the game, we had to decide what order to eliminate the factors in. Thankfully, our network all flows in the same direction, so it was simple to eliminate all of the PER factors, then all of the individual comparison factors, and then the team comparison factor to draw our conclusion.

4 Evaluation

We abandoned our first implementation using offensive/defensive abilities, as well as metric such as height, weight and age, since it came out they did not perform well. In Table 1 you can see the accuracy results on the two Bayesian Networks, the Individual and the Team+Individual, as well as the best result obtained with our first approach. Table 1 shows also some other infor-

Team&Games	Season 2009/2010
Statistics	Season 2008/2009
#Teams	30
#Players	445
#Games	1230
Accuracy [First attempt]	50.16%
Accuracy [Individual]	58.70%
Accuracy [Team+Individual]	63.98%

Table 1: Information about the dataset used, as well as the final accuracy of the tested Bayesian Networks.

mation about our dataset. We used statistics about 445 players to predict the outcome of 1230 games, reaching a top accuracy of almost 64%. In Figure 2 is shown the accuracy percentage of the two Bayesian Networks, grouped by score difference. A negative difference means that the home team lost the game. The bars indicate the percentage of games ended with the given score difference that have been correctly predicted by the two algorithms.

5 Discussion

While we did not achieve our original stated goal of beating the 76% of [3], we did tie the previous accuracy of their Bayesian Logic approach with far fewer variables. Simply using PRE statistics about players we were able to reach an accuracy of 58%, way higher than our first approach using many different metrics. Moreover, adding very few information about team statistics, we were able to reach an accuracy of 64%. We believe with more time, and more knowledge about metrics and sport statistics, we may be able to successfully apply Bayesian Networks to basketball games forecasting. The simplicity of the metrics and conditional probabilities used in our approach and the obtained results are something really surprising.

It is interesting to notice the chart in Figure 2: while it is expected to have a drop in accuracy when the score difference is close to zero, which denotes a tight game, we can not say the same for the lower accuracy at the tails; we explain this with the fact that there are less games ending with a huge score difference, and that in many case they are games where something unpredicted happened in one of the two teams, usually an injured key player, facts that are not taken into account by the Bayesian Network.

One major improvement that can be made to our project would be to have it account for rookies that are entering the league in the season being predicted. There are 60 rookies drafted before each season, in addition to many called up from the developmental league and and this makes for far too many wild cards in the matchups.

This could be handled with another feature we would like to add, which is the ability to incorporate live data from the season being predicted into the system. By the end of the predictions, the data we are using for the matchups is one year old and fresh new data has been developed. If we could integrate the statistics with games that have already taken place this season, we would be able to draw more accurate conclusions.

References

- [1] DRAGAN MILJKOVIĆ, LJUBIŠA GAJIĆ, A. K. Z. K. The use of data mining for basketball matches outcomes prediction. In *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on* (2010), IEEE, pp. 309–312.
- [2] HU, F., AND ZIDEK, J. Forecasting NBA basketball playoff outcomes using the weighted likelihood. *Lecture Notes-Monograph Series 45* (2004), 385–395.
- [3] ORENDORFF, D., AND JOHNSON, T. First-Order Probabilistic Models for Predicting the Winners of Professional Basketball Games.