Basketball predictions in the NCAAB and NBA: similarities and differences

Albrecht Zimmermann
albrecht.zimmermann@insa-lyon.fr

February 6, 2016

Abstract

Most work on predicting the outcome of basketball matches so far has focused on NCAAB games. Since NCAAB and professional (NBA) basketball have a number of differences, it is not clear to what degree these results can be transferred. We explore a number of different representations, training settings, and classifiers, and contrast their results on NCAAB and NBA data. We find that adjusted efficiencies work well for the NBA, that the NCAAB regular season is not ideal for training to predict its post-season, the two leagues require classifiers with different bias, and Naïve Bayes predicts the outcome of NBA playoff series well.

1 Introduction

Predicting the outcome of contests in organized sports can be attractive for a number of reasons such as betting on those outcomes, whether in organized sports betting or informally with colleagues and friends, or simply to stimulate conversations about who "should have won".

Due to wide-spread betting during the NCAAB playoffs (or "March Madness"), much work has been undertaken on predicting the outcome of college basketball matches. It is not clear how much of that work can be transferred easily to the NBA. Professional basketball shows several differences to college basketball:

1. Teams are typically closer in skill level, since recruiting and resource bases play a lesser role.
2. Teams play almost all other teams every season.
3. Teams play more games, particularly in the playoffs, where the NCAAB's "one and done" is in sharp contrast to the NBA's best-of-seven series.

To the best of our knowledge, it is not clear how those differences affect the task of learning a predictive model for the sports: the first point implies that prediction becomes harder, whereas the other two indicate that there is more and more reliable data.

Most of the existing work in the field is more or less statistical in nature, with much of it developed in blog posts or web columns. Many problems that can be addressed by statistical methods also offer themselves up as Machine Learning settings, with the expected gain that the burden of specifying the particulars of the model shifts from a statistician to the algorithm. Yet so far there is relatively little such work in the ML literature.

We intend to add to the body of work on sports analytics in the ML community by building on earlier work [13], and evaluating different match representations, learning settings, and classifiers. We compare the results on NCAAB data to those on NBA data, with a particular focus on post-season predictions.

In the next section, we will discuss how to represent teams in terms of their performance statistics, followed by ways of performing prediction in Section 3. In Section 4, we discuss existing work on NCAAB and NBA match prediction. Sections 6 and 7 are given to the evaluation of different prediction settings on

NCAAB and NBA data, respectively, before we compare classifier behavior on those two types of data in more detail in Section 8.

2 Descriptive statistics for teams

In this section, we will discuss the different options for describing basketball teams via the use of game statistics that we evaluate later in the paper. We will begin by a recap of the state-of-the-art, afterwards discussing our own extensions and aggregate statistics over the course of the season.

2.1 State of the art

The most straight-forward way of describing basketball teams in such a way that success in a match can be predicted relate to scoring points – either scoring points offensively or preventing the opponent's scoring defensively. Relatively easy to measure offensive statistics include field goals made (FGM), three-point shots made (3FGM), free throws after fouls (FT), offensive rebounds that provide an additional attempt at scoring (OR), but also turnovers that deprive a team of an opportunity to score (TO). Defensively speaking, there are defensive rebounds that end the opponent's possession and give a team control of the ball (DR), steals that have the same effect and make up part of the opponent's turnovers (STL), and blocks, which prevent the opponent from scoring (BLK). And of course, there are points per game (PPG) and points allowed per game (PAG).

The problem with these statistics is that they are all raw numbers, which limits their expressiveness. If a team collects 30 rebounds in total during a game, we cannot know whether to consider this a good result unless we know how many rebounds were there to be had in the first place. 30 of 40 is obviously a better rebound rate than 30 of 60. Similar statements can be made for field goals and free throws, which is why statistics like offensive rebound rate (ORR), turnover rate (TOR), or field goals attempted (FGA) will paint a better picture. Even in that case, however, such statistics are not normalized: 40 rebounds in a game in which both teams combined to shoot 100 times at the basket is different from 40 rebounds when there were only 80 scoring attempts. For normalization, one can calculate the number of possessions in a given game:

$$Possessions = 0.96*(FGA-OR-TO+(0.475*FTA))$$

and derive teams' points scored/allowed per 100 possessions, deriving offensive and defensive *efficiencies*:

$$\begin{aligned} OE &= \frac{PPG * 100}{Possessions}, \\ DE &= \frac{PAG * 100}{Possessions} \end{aligned} \tag{1}$$

It should be noted that the factor 0.475 is empirically estimated – when first introducing the above formulation for the NBA, Dean Oliver estimated the factor as 0.4 [10].

This is currently the most-used way of describing basketball teams in the NBA. When discussing complete teams or certain line-ups (five (or fewer) player groups), the phrase "points per 100 possessions" makes frequent appearance on sites such as `fivethirtyeight.com`, `www.sbnation.com`, and `hardwoodparoxysm.com`.

While such statistics are normalized w.r.t. the "pace" of a game, they do not take the opponent's quality into account, which can be of particular importance in the college game: a team that puts up impressive offensive statistics against (an) opponent(s) that is (are) weak defensively, should be considered less good than a team that can deliver similar statistics against better-defending opponents. For best expected performance, one should therefore normalize w.r.t. pace, opponent's level, and national average, deriving *adjusted* efficiencies:

$$\begin{aligned} AdjOE &= \frac{OE * avg_{all\ teams}(OE)}{AdjDE_{opponent}}, \\ AdjDE &= \frac{DE * avg_{all\ teams}(DE)}{AdjOE_{opponent}} \end{aligned} \tag{2}$$

The undeniable success of those two statistics, pioneered by Ken Pomeroy [11], in identifying the strongest teams in NCAA basketball have made them the go-to descriptors for NCAA basketball teams.

Dean Oliver has also singled out four statistics as being of particular relevance for a team's success, the so-called "Four Factors" (in order of importance, with their relative weight in parentheses):

Effective field goal percentage (0.4):

$$eFG\% = \frac{FGM + 0.5 \cdot 3FGM}{FGA} \quad (3)$$

Turnover percentage (0.25):

$$TO\% = \frac{TO}{Possessions} \quad (4)$$

Offensive Rebound Percentage (0.2):

$$OR\% = \frac{OR}{(OR + DR_{Opponent})} \quad (5)$$

Free throw rate (0.15):

$$FTR = \frac{FTA}{FGA} \quad (6)$$

2.2 Adjusted Four Factors

In an earlier work [13], we have introduced the idea of adjusting the Four Factors in the same way as efficiencies and evaluated their usefulness for predicting college basketball matches. While multi-layer perceptrons (MLP) achieved better results using the adjusted efficiencies, Naïve Bayes classifiers performed better using the adjusted Four Factors.

2.3 Averaging over the course of the season

To gain a comprehensive picture of a team's performance during the entire season, such statistics would have to be averaged over all games up to the prediction point in time. A simple average would give the same weight to matches that happened at the beginning of the season as to matches that happened the week before the prediction. Since teams' characteristics change over time – players get injured, tactics change – this is clearly undesirable. In addition, we want matches that are close in time to have approximately equal impact on the total average.

In this paper, we therefore consecutively enumerate the *days* of a season – not the *game days* – and use them to weight contributions to the average. As an illustration, imagine a team that played on day 1, 3, 10, and 15 and we want to make a prediction for day 16. For TOR, for instance, we therefore average in the following way (we use superscripts to denote the day on which the statistic has been recorded):

$$TOR_{avg} = \frac{1 \cdot TOR^1 + 3 \cdot TOR^3 + 10 \cdot TOR^{10} + 15 \cdot TOR^{15}}{1 + 3 + 10 + 15 = 28}$$

The impact of the most recent match is more than half in this case, whereas match 1 – two weeks ago – has very little impact. Enumerating game days would give the first match a quarter of the impact of the most recent one, instead.

2.4 Calculating adjusted statistics

Due to the averaging, each team's adjusted statistics are directly influenced by their opponents', and indirectly by those opponents' opponents. As an illustration, consider a schedule like the one shown in Table 1. To calculate $Team_1$'s adjusted offensive efficiency after match 3, we need to know $Team_4$'s adjusted defensive efficiency before the match (i.e. after match 2), which takes the form:

$$\begin{aligned} & AdjDE^2(Team_4) \\ =& \frac{DE^1(Team_4) * avg_{all\ teams}(DE)}{AdjOE^1(Team_3)} \\ +& \frac{DE^2(Team_4) * avg_{all\ teams}(DE)}{AdjOE^2(Team_2)} \end{aligned}$$

We have left out the averaging weights for the sake of readability but as is obvious, we need to add equations estimating the adjusted offensive efficiency for $Team_2$ and $Team_3$, and the further in the season we advance, the more teams are involved.

We do not attempt to solve this set of equations analytically but instead use an iterative algorithm: 1) we use the current values of statistics of teams involved in the right-hand side of equations to calculate new values for teams in the left-hand side. 2) we update all teams' statistics and re-iterate. This is performed until values have stabilized.

Team	Opponents		
$Team_1$	$Team_2$	$Team_3$	$Team_4$
$Team_2$	$Team_1$	$Team_4$	$Team_5$
$Team_3$	$Team_4$	$Team_1$	$Team_6$
$Team_4$	$Team_3$	$Team_2$	$Team_1$
...			

Table 1: Partial example schedule

In practice, we observe that this approach finds two stable solutions at some point, and flips among those two in succeeding iterations. To break this tie, we pick the solution for which the *average* absolute difference between $AdjOE$ and $AdjDE$ is smaller. Let us denote the set of all teams by $\mathcal{T}$, and solutions as set of adjusted offensive and defensive efficiencies:

$$S_i = \{AdjOE_i(T), AdjDE_i(T) \mid \forall T \in \mathcal{T}\}$$

then our tie breaker takes the form:

$$\arg \min_{\{S_1,S_2\}} \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} |AdjOE_i(T) - AdjDE_i(T)|.$$

The same approach is followed when calculating the adjusted Four Factors.

2.5 Employing standard deviations

An important element of teams' performance is consistency. A team can collect an average adjusted offensive efficiency by alternating matches in which it scores much more than expected with ones in which it is virtually shut out. Especially against higher-quality competition, we would expect one of the latter type of matches, and this information can therefore be relevant when trying to predict match outcomes. A straight-forward way of encoding consistency consists of adding the standard deviation of teams statistics over the course of the season.

3 Approaches to prediction

A wide-spread state-of-the-art way of using the derived statistics in predicting match outcomes consists of using the so-called Pythagorean Expectation, e.g.:

$$Win\ Probability = \frac{((Adjusted)\ OE_{avg})^y}{((Adjusted)\ OE_{avg})^y + ((Adjusted)\ DE_{avg})^y}$$

to calculate each team's win probability and predicting that the team with the higher probability wins. More generally, *ranking systems* can by used by ranking the entire pool of teams and predicting for each match-up that the higher ranked team wins.

Many existing approaches are somewhat or even fully hand-crafted. This can be rather high-level, as in *defining* the transition probabilities in LRMC's Markov chain by hand (cf. Section 4), or it can go as far as Ken Pomeroy taking home court advantage into consideration by *multiplying* the home team's stats by 1.014. Also, especially the Pythagorean Expectation seems to be a rather simple model.

Machine Learning promises to address both of these issues: we would expect to be able to *learn* the relative importance of different descriptive measures, in particular if this importance changes for different numerical ranges, and to be able to *learn* their relationships, automatically making the model as difficult (or simple) as needed. We therefore turned to classification learners representing several different paradigms and evaluated their performance.

In a reversal of current practice, explicit prediction of match outcomes could be used to rank teams by predicting the outcome of all hypothetical pairings and ranking teams by number of predicted wins.

Concretely, we evaluate the WEKA [5] implementations of the Multi-Layer Perceptron (MLP), Random Forest (RF), and Naïve Bayes (NB) classifiers. The former two are run with default parameters, whereas for NB we activated kernel estimation.

4 Related Work

The use of the Pythagorean Expectation goes back to Bill James' work on baseball. It was adapted for the use in basketball prediction by numerous analysts, e.g. Daryl Morey, John Hollinger, Ken Pomeroy, and Dean Oliver. The difference between the different approaches comes down to which measures of offensive

and defensive prowess are used and how the exponent is estimated. Dean Oliver also first introduced possession-based analysis formally in his book "Basketball on Paper" [10], although he acknowledges that having seen coaches use such analysis in practice. The same work introduced the "Four Factors".

The adjustment of efficiencies to the opponent's quality is due to Ken Pomeroy who uses them as input in his version of the Pythagorean Expectation to rank NCAAB teams and predict match outcomes. His is far from the only ranking system, however, with other analysts like Jeff Sagarin, Ken Massey or Raymond Cheung running their own web sites and giving their own predictions. Comparisons of the results of different ranking systems can for instance be found at `http://masseyratings.com/cb/compare.htm` or `http://www.raymondcheong.com/rankings/perf13.html`. The worst accuracy for those systems is in the $62\% - 64\%$ range, equivalent to predicting that the home team wins, the best ones achieve up to $74\% - 75\%$. The NCAAB itself uses the so-called Ratings Percentage Index to rank teams, a linear weighted sum of a team's winning percentage, its opponents' winning percentage, and the winning percentage of those opponents' opponents.

As an alternative approach, Kvam *et al.* have proposed a logistic regression/Markov chain model [8]. In this method, each team is represented as a state in a Markov chain and state transitions occur if one team is considered better than its opponent. Logistic regression is used to estimate transition probability parameters from the data. The authors have proposed an updated version using Bayesian estimates [4], and recently published work in which they estimate their method's success in comparison to other ranking schemes [3].

In terms of NBA basketball, Loeffelholz *et al.* [1] have trained a variety of neural networks for predicting NBA match outcomes. The highest accuracy that they report in the abstract is approximately 74%. Oh *et al.* [9] propose building graphical models from play-by-play data that can be used to simulate the flow of a match. They report on exploring counterfactuals: simulating different substitution patterns for particular matches, and observing the resulting outcomes. A similar idea is explored in [2]: using NBA teams' substitution patterns and five-man units' *plus/minus*, they combine Markov chains and ridge regression to predict the outcomes of the 2014/2015 playoff *series*, reporting training set accuracy of 68% (individual matches) and postseason accuracy (complete seasons) of 80%. Finally, Franks *et al.* have proposed exploiting the tracking data nowadays available for NBA matches to derive a new defensive statistic called "counterpoints" [6].

5 Experimental setup

The goal of our experimental evaluation is to gain an understanding about which kind of information is most conducive for achieving high-quality predictions. This involves evaluating

1. different kinds of descriptive statistics,
2. training the classifier on all available data, or only on a subset, and
3. using the preceding season's statistics as stand-in for the first game day's statistics.

We will elaborate those aspects in the following paragraphs. After having found good models, we will explore predictions in more detail in Sections 8 and draw conclusions related to the three differences described in the introduction.

Different statistics We used WEKA's `ChiSquared`, `GainRatio`, `InfoGain`, and `ReliefF` rankers to explore the attributes described in 2, as well as season-level ones (e.g. win percentage), using the "cross-validation" option with ten folds. The results are rather consistent across rankers: the attributes shown in Table 2 are always highly ranked, as are location, and the adjusted efficiencies (adjOEff, adjDEff).

The (adjusted) Four Factors depend to a certain degree on the selected season: while effective field goal percentage (eFG%) is almost always highly ranked, offensive rebound (ORR), turnover (TOR), and free throw rates (FTR) can go from being ranked higher than field goal percentage to unranked and

Statistic	Explanation
Win Percentage	$\frac{\text{Number of Wins}}{\text{Number of Games played}}$
Margin of Victory (MoV)	PPG-PAG in Wins
Possession-adjusted MoV	OEff - DEff in Wins
Possession-adjusted Point Differential	OEff-DEff

Table 2: Base team statistics

vice versa. We explore this phenomenon in more detail by evaluating different attribute combinations in the next sections.

We encode matches with the team statistics of the two opponents, with the opponent with the lexicographically lower name coming first. In addition, matches are described by whether it is a away, home, or neutral court from the perspective of this first opponent, and by the date (name and day of the month). This could be particularly helpful in the case of NCAA basketball, where conference tournaments (in which the competition is of higher quality) occur later in the season.

Classifier evaluation and training data composition We are using two assumptions for our experimental settings. The first is that there is a temporal aspect to basketball prediction. Using matches from later in the season (when trends may have stabilized) to predict matches from earlier could introduce a bias. We therefore do not perform a randomized cross-validation but instead always only train on data of matches that occurred *before* the matches to be predicted.

Second, however, we assume that it is desirable to have as fresh data as possible available for training. To achieve this, we therefore predict matches *day by day*, and add them afterwards to the batch of training data., for the prediction of the following day. As is known from Data Stream classification [7], however, data in a temporal relationship can become "stale". We therefore also evaluate whether limiting training to recent data has an effect.

Filling in first-day statistics On the first day of a season, we have necessarily no statistics for teams available yet are already faced with a prediction task. One way of addressing this consists of using the statistics the team exhibited at the end of the preceding season. Since teams change during the off-season, however – college players graduate/get drafted, NBA players get traded – the preceding season's statistics may be rather misleading. The alternative, i.e. not using the preceding season's statistics, essentially amounts to a coin flip (or more likely, predicting that the home team wins).

6 NCAAB predictions

The data for the experiments on ncaa basketball comes from Ken Pomeroy's page [11], and comprises the seasons 2007/2008-2012/2013. Since we cannot build a model to predict 2007/2008, we predict from 2008/2009 onwards. In earlier work [13], we have found that combining adjusted efficiencies with Four Factors did not give competitive results. Since our manner of calculating the adjustments and averaging have changed, however, we reprise our experiments with adjusted efficiencies and adjusted four factors.

In the first setting, shown in Table 3, we used only a single season's worth of data for training. This means in particular that every time a day's matches are added to the training set, the oldest day's matches are removed.

Representation	Season	MLP	NB	RF
AdjEff	2012/2013	0.7088	0.7018	0.6867
	2011/2012	0.7165	0.7103	0.7031
	2010/2011	0.7070	0.7020	0.6894
	2009/2010	0.7312	0.7207	0.7038
	2008/2009	0.7127	0.7011	0.6841
AdjFF	2012/2013	0.7016	0.6906	0.6842
	2011/2012	0.7098	0.7038	0.6912
	2010/2011	0.6981	0.6899	0.6810
	2009/2010	0.7083	0.7061	0.7016
	2008/2009	0.6960	0.6903	0.6830

Table 3: Predictive accuracies per season (NCAAB), one season worth of data, no filling in first-day statistics

Similarly to our earlier results, we find that the multi-layer perceptron performs better than Naïve

Bayes, with the Random Forest classifier third. We also see that the adjusted efficiencies are superior to adjusting Four Factors, in terms of predictive accuracy. If we use all matches we have available as training data for the day to be predicted (Table 4), we can improve the results slightly.

Representation	Season	MLP	NB	RF
AdjEff	2012/2013	0.7120	0.7024	0.6885
	2011/2012	0.7248	0.7155	0.6946
	2010/2011	0.7115	0.7052	0.6851
	2009/2010	0.7357	0.7318	0.7072
	2008/2009	0.7163	0.7061	0.6907
AdjFF	2012/2013	0.6983	0.6942	0.6795
	2011/2012	0.7191	0.7076	0.6947
	2010/2011	0.7082	0.6933	0.6847
	2009/2010	0.7257	0.7086	0.7038
	2008/2009	0.7050	0.6850	0.6843

Table 4: Predictive accuracies per season (NCAAB), training on all preceding seasons, no first-day statistics

Even though the varying strength of opponents in NCAA basketball should lead to fluctuations in statistics, adding standard deviations to the description does not pay off, as Tables 5 and 6 show.

Representation	Season	MLP	NB	RF
AdjEff	2012/2013	0.7120	0.6906	0.6902
	2011/2012	0.7089	0.6998	0.6946
	2010/2011	0.7030	0.6998	0.6853
	2009/2010	0.7174	0.7129	0.7024
	2008/2009	0.7075	0.6947	0.6903
AdjFF	2012/2013	0.6848	0.6887	0.6844
	2011/2012	0.6936	0.6994	0.6970
	2010/2011	0.6968	0.6920	0.6730
	2009/2010	0.7138	0.7094	0.6994
	2008/2009	0.6887	0.6936	0.6828

Table 5: Predictive accuracies per season (NCAAB), one season worth of data, no filling in first-day statistics, standard deviations

Similarly, we do not find any benefit to using the preceding season's statistics to fill in team statistics for their first match, an indicator that teams change too much in the off-season, and omit the results here.

Representation	Season	MLP	NB	RF
AdjEff	2012/2013	0.7090	0.6910	0.6960
	2011/2012	0.7174	0.7066	0.6931
	2010/2011	0.7121	0.6998	0.6896
	2009/2010	0.7346	0.7244	0.7051
	2008/2009	0.7128	0.6874	0.6909
AdjFF	2012/2013	0.7061	0.6919	0.6827
	2011/2012	0.7179	0.7025	0.6959
	2010/2011	0.7065	0.6963	0.6753
	2009/2010	0.7233	0.7151	0.6944
	2008/2009	0.7006	0.6971	0.6883

Table 6: Predictive accuracies per season (NCAAB), all preceding seasons for training, no filling in first-day statistics, standard deviations

7 Predictive Results on NBA data

For our experiments in NBA match data, we crawled the information for seasons 2007/2008-2014/2015 from `www.basketball-reference.com`. Since we do not have data to build a model to predict 2007/2008, all tests were run only from 2008/2009 onwards.

Our first experimental comparison can be found in Table 7: in that setting, the classifier is trained on only the season preceding the one to be predicted. Furthermore, every time a game day is added to the training data, the oldest one is removed. In this manner, the model should always be based on the most recent information, without trends from older seasons acting in a distorting manner. We have evaluated three representations (in addition to teams' and matches' base statistics) – Adjusted Efficiencies (AdjEff), Adjusted Efficiencies and Four Factors (AdjEff-FF), and Adjusted Four Factors (AdjFF). The statistics of Teams on the first day of the season are not filled in.

All in all, this experiment indicates that Adjusted Efficiencies are the most effective encoding. Especially Naïve Bayes performs well on this representation (and outperforms the MLP – contrary to the results on the NCAA data). Interestingly, Random Forests recover in comparison to the MLP once representations involve the Four Factors – adjusted or not – those seem to be more useful to a tree model than

Representation	Season	MLP	NB	RF
AdjEff	2014/2015	0.6209	0.6552	0.6056
	2013/2014	0.6133	0.6300	0.5997
	2012/2013	0.5906	0.6446	0.5799
	2011/2012	0.6210	0.6294	0.6201
	2010/2011	0.6148	0.6606	0.6003
	2009/2010	0.6189	0.6524	0.6204
	2008/2009	0.6433	0.6806	0.6312
AdjEff-FF	2014/2015	0.6079	0.6430	0.6018
	2013/2014	0.5709	0.6187	0.5929
	2012/2013	0.5807	0.6476	0.6020
	2011/2012	0.5819	0.6313	0.6089
	2010/2011	0.6133	0.6522	0.6301
	2009/2010	0.5747	0.6418	0.6021
	2008/2009	0.6137	0.6586	0.6160
AdjFF	2014/2015	0.5927	0.6369	0.6133
	2013/2014	0.5982	0.6255	0.6133
	2012/2013	0.5776	0.6416	0.6012
	2011/2012	0.5736	0.6238	0.5847
	2010/2011	0.6240	0.6491	0.6270
	2009/2010	0.6052	0.6402	0.6235
	2008/2009	0.6274	0.6532	0.6297

Table 7: Predictive accuracies per season (NBA), one season worth of training data, no filling in first-day statistics

Representation	Season	MLP	NB	RF
AdjEff	2014/2015	0.6438	0.6598	0.5988
	2013/2014	0.6240	0.6384	0.6035
	2012/2013	0.6187	0.6499	0.6005
	2011/2012	0.6331	0.6378	0.6006
	2010/2011	0.6506	0.6636	0.6293
	2009/2010	0.6494	0.6456	0.6151
	2008/2009	0.6608	0.6677	0.6395
AdjEff-FF	2014/2015	0.6346	0.6552	0.6323
	2013/2014	0.6315	0.6338	0.6065
	2012/2013	0.6134	0.6454	0.6058
	2011/2012	0.5857	0.6248	0.6071
	2010/2011	0.6392	0.6545	0.6186
	2009/2010	0.6159	0.6402	0.6349
	2008/2009	0.6251	0.6586	0.6129
AdjFF	2014/2015	0.6262	0.6590	0.6140
	2013/2014	0.6414	0.6338	0.5914
	2012/2013	0.6050	0.6431	0.6134
	2011/2012	0.5903	0.6313	0.6071
	2010/2011	0.6323	0.6560	0.6369
	2009/2010	0.6319	0.6433	0.6067
	2008/2009	0.6274	0.6563	0.6251

Table 8: Predictive accuracies per season (NBA), training on all preceding seasons, no filling in

the efficiencies. It is also remarkable that the order of seasons according to predictive accuracy changes – not only between classifiers but also between representations. In contrast to this, the order remained relatively stable for NCAA data. Given that the training data in this setting were limited, using more data might make a difference.

To explore this direction, and because it is not clear how to decide how many prior seasons to use for training, we keep all seasons' data for training in the setting that we show in Table 8.

In fact, MLP and RF profit from using more data, NB less so, while continuing to perform best on average. It is somewhat surprising to see a relative large negative change in its predictive accuracy for 2008/2009, AdjEff, since this is the season that was predicted with the model having the least amount of data. The impact of additional data is more pronounced for the representations that involve the Four Factors.

Experiments with a single season of training data and standard deviations led to a deterioration, so we refrain from reporting them here. Instead, Table 9 shows what happens when we augment the preceding setting with standard deviations, indicating the consistency of teams' performance indicators.

The results can at best considered inconclusive but generally speaking, including standard deviations degrades predictive accuracy. At first, this is surprising, but given that team statistics are already only approximations (with adjustment and averaging potentially introducing errors), calculating standard deviations over a season (or several of them) might very well increase the effect of any errors.

Using the end-of-season statistics of the preceding season as fill-in for the statistics of the first match of teams has the potential to add some percentage points but can also misrepresent the actual strength of a team in the new season. Unfortunately, we have found that the latter effect seems to hold. We report only one setting (all seasons as training data, Table

Representation	Season	MLP	NB	RF
AdjEff	2014/2015	0.6499	0.6522	0.6178
	2013/2014	0.6323	0.6414	0.5883
	2012/2013	0.6286	0.6431	0.6073
	2011/2012	0.5978	0.6192	0.6257
	2010/2011	0.6278	0.6506	0.6270
	2009/2010	0.6463	0.6357	0.6220
	2008/2009	0.6327	0.6722	0.6335
AdjEff-FF	2014/2015	0.6278	0.6491	0.6110
	2013/2014	0.6149	0.6323	0.6096
	2012/2013	0.5997	0.6377	0.6081
	2011/2012	0.6155	0.6229	0.6071
	2010/2011	0.6331	0.6506	0.6171
	2009/2010	0.6044	0.6357	0.6067
	2008/2009	0.6160	0.6525	0.6084
AdjFF	2014/2015	0.6438	0.6560	0.6293
	2013/2014	0.5906	0.6384	0.5936
	2012/2013	0.6058	0.6431	0.5982
	2011/2012	0.5773	0.6145	0.5875
	2010/2011	0.6598	0.6506	0.6217
	2009/2010	0.6258	0.6319	0.6212
	2008/2009	0.6144	0.6593	0.6167

Table 9: Predictive accuracies per season, training on all preceding seasons (NBA), no filling in, standard deviations

Representation	Season	MLP	NB	RF
AdjEff	2014/2015	0.6369	0.6659	0.6156
	2013/2014	0.6331	0.6331	0.5914
	2012/2013	0.6195	0.6499	0.5959
	2011/2012	0.6304	0.6341	0.6080
	2010/2011	0.6438	0.6697	0.6491
	2009/2010	0.6509	0.6425	0.6151
	2008/2009	0.6586	0.6700	0.6289
AdjEff-FF	2014/2015	0.6178	0.6613	0.6301
	2013/2014	0.6080	0.6376	0.6073
	2012/2013	0.6081	0.6476	0.5974
	2011/2012	0.6024	0.6238	0.6145
	2010/2011	0.6201	0.6560	0.6423
	2009/2010	0.6136	0.6479	0.6303
	2008/2009	0.6274	0.6593	0.6304
AdjFF	2014/2015	0.6293	0.6682	0.6171
	2013/2014	0.6209	0.6361	0.6080
	2012/2013	0.5951	0.6499	0.6172
	2011/2012	0.6080	0.6369	0.6127
	2010/2011	0.6461	0.6583	0.6293
	2009/2010	0.6410	0.6418	0.6341
	2008/2009	0.6456	0.6624	0.6411

Table 10: Predictive accuracies per season, training on all preceding seasons (NBA), filled in first-day statistics

10), and while there are occasional improvements, the effect is not consistent.

In light of our results, the accuracy reported in [1] (74.33%) is surprising. Unfortunately, we have not been able to access a full version of that paper but the abstract (where the accuracy is mentioned) notes that the authors performed extensive tuning of the neural networks used.

8 Comparing prediction curves

The numbers we have shown in tabulated form in the preceding sections summarize the accuracy of classifiers over the entire season. As we explained in Section 5, however, we classify each day's matches separately, before we relearn the classifier with that day's data added. We can therefore explore the development of predictive accuracy over the course of a season to gain better insight into classifier behavior.

We will use this information to explore the three aspects that we highlighted in the introduction. We begin by looking in more detail at the phenomen that a) professional teams play more games during a season than university teams, and b) the skill differential among professional teams is smaller than between university teams (and professional teams play all possible opponents at least once).

Concretely, professional teams play 82 games before the post-season, university teams only about 30. Since team statistics are averaged over all matches played before the test match, we would therefore assume that descriptors for professional teams are more representative than for university teams and predictions should therefore stabilize better. If, on the other hand, the skill differential has a larger impact, NCAAB predictions should be quicker to stabilize.

Figures 1 and 2 (3 and 4) show the cumulative accuracy over the course of a full season achieved by Naïve Bayes (MLP) for the NBA and NCAAB re-

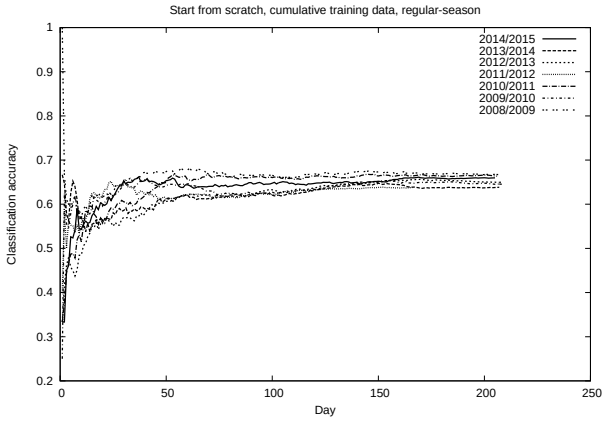

Figure 1: Accuracy on NBA data as a function of game days, NB

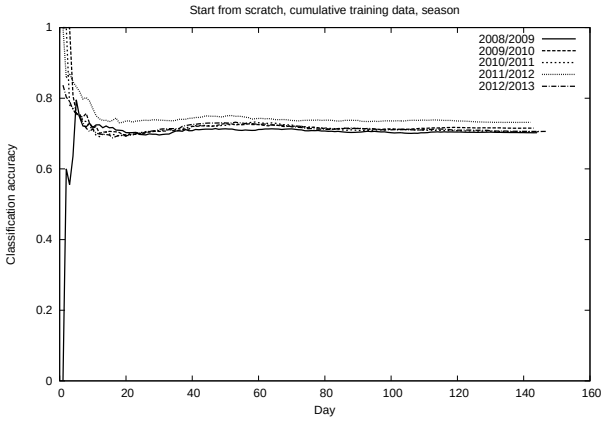

Figure 2: Accuracy on NCAAB data as a function of game days, NB

spectively.[1] The x-axis is labeled with the number of games days – since not all teams play on the same day, there are siginificantly more game days per season than per team. In comparing the NB curves, we see that they stabilize much more quickly for NCAAB than for the NBA. For the NBA, the classifier need about half a season (41 games), whereas for NCAAB the classifier arrives at this point at about a third of the season (10 games).

Moving on to the MLP, we see the clear difference between a classifier with a strong bias, and one exhibiting strong variance. The neural network shows much bigger oscillations than the Naïve Bayes, occasionally outperforming NB at the same relative (early) point of the NBA season. But as in the case of the NB, predictions for the NCAAB stabilize more quickly. Additionally, the swings for the MLP are much smaller. This implies easier decision boundaries, i.e. the larger relative strength differential of which we wrote in the introduction.

8.1 Post-season predictions

Compared to the post-season, regular season matches are relatively easier to predict. In the post-season, teams will be mostly more similar in strength, in the NCAAB case probably much more similar. If this is the case, we would expect that the models learned on full seasons (including past post-seasons and the current post-season up to the prediction date) would perform clearly worse in the post-season for the NCAAB, only somewhat worse in the NBA.

There are also differences in the number of teams involved in the post-season and the number of games played:

- In the NCAAB, 64 teams reach the post-season (in recent years 68, eight of which play in a "play-in" weekend to complete the field). Each pairing of teams only play each other once, for a total of 61 games.
- In the NBA, 16 teams reach the post-season. Each pairing plays best-of-seven series, meaning that a team needs to win four times to move on to the next round. The whole post-season can therefore have between 60 and 105 games.

In both leagues, teams are ranked within a particular group (4 regions in the NCAAB, 2 conferences in the NBA) according to their performance during the regular season. Higher-ranked teams get rewarded by playing lower-ranked ones. Given the larger field in the NCAAB, a first-ranked, i.e. best, team in a region plays the 16th-best. In the NBA, the pairing is best

[1] The end point of each curve corresponds to the results reported in Tables 8 and 4.

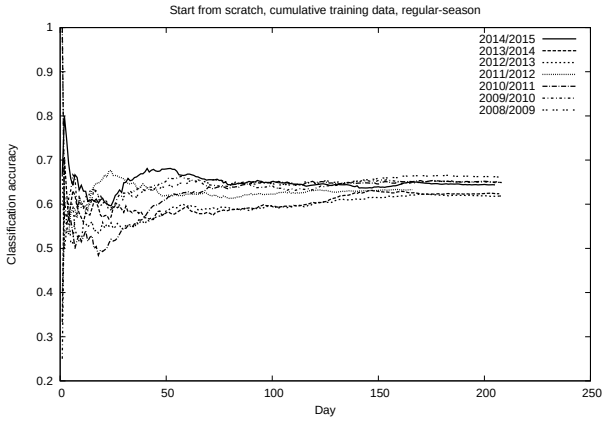

Figure 3: Accuracy on NBA data as a function of game days, MLP

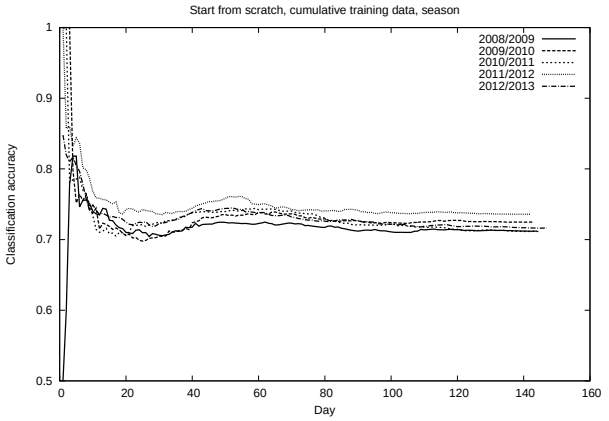

Figure 4: Accuracy on NCAAB data as a function of game days, MLP

and 8th-best. As a result of this, early rounds should still be easier to predict in the NCAAB than in the NBA. However, the one-shot nature of the NCAAB means that a higher-ranked team that slips up cannot recover – such upsets make up part of the attraction of the post-season. In the NBA, the format will typically favor the better team.

Post-season predictions are therefore affected both by the effects considered in the preceding section, and the different post-season formats.

The NCAAB post-season curves, shown in Figures 5 and 6, show that with the exception of 2008/2009 (when notably all four first-ranked teams reached the semi-final round), the cumulative accuracy of post-season predictions are below the regular season accuracy. This is the case by a rather wide margin for NB, but still noticable for the MLP.

Of particular interest is the curve for the MLP for the 2010/2011 post-season. By day three, the 32 matches of the first round (which we expected to be easiest to predict) have been decided, i.e. more than half the total of the post-season, and the MLP boasts an accuracy of more than 80%. In the succeeding rounds, the cumulative accuracy deteriorates consistently, reaching the lowest point on day 11 before recovering slightly with the correct prediction of the final match. After the first round, the MLP was effectively not better than chance. This aligns with what happened in the first round of that season's post-season when the higher-ranked teams mostly won their first-round matches.

Hidden in this is a remarkable fact: between the NB and MLP classifiers, there is only a single incorrectly predicted champion! Unfortunately, such post-season results would be useless for the task of winning the office pool, one of the main motivations in the non-academic world for predicting the NCAAB post-season.

NBA post-season curves (Figures 7 and 8) show less deviation from the regular season cumulative accuracy than in the case of the NCAAB. Accuracies are roughly the same for 2010/11, 2014/15 (NB), 2009/10 (MLP), 2012/13 (MLP), better for 2011/12, 2009/10 (MLP), and when they are worse (2008/09, 2013/14, 2014/15 (MLP), 2009/2010 (NB), 2012/2013 (NB)), the drop-off is not as steep as in the NCAAB case.

They do, however, show much more variation than regular season curves, especially in the beginning. At one end of this is the 2012/13 season, a curve that starts out with 100% accuracy for the first two game days for either classifier. The first round that year saw two of the eight pairings won 4-0 by the higher-seeded team, yet also featured an "upset" (sixth over third), bringing the cumulative accuray down. None of the higher-seeded teams lost its first (home) game

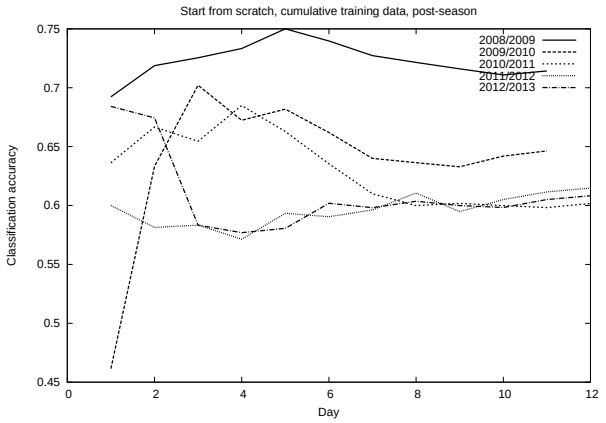

Figure 5: Accuracy for NCAAB post-season matches as a function of game days, NB

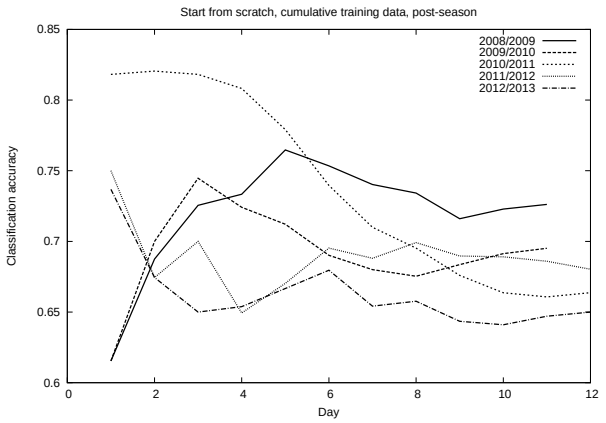

Figure 6: Accuracy for NCAAB post-season matches as a function of game days, MLP

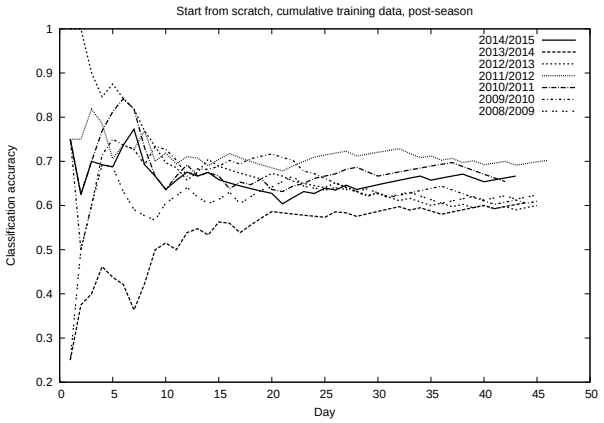

Figure 7: Accuracy for NBA post-season matches as a function of game days, NB

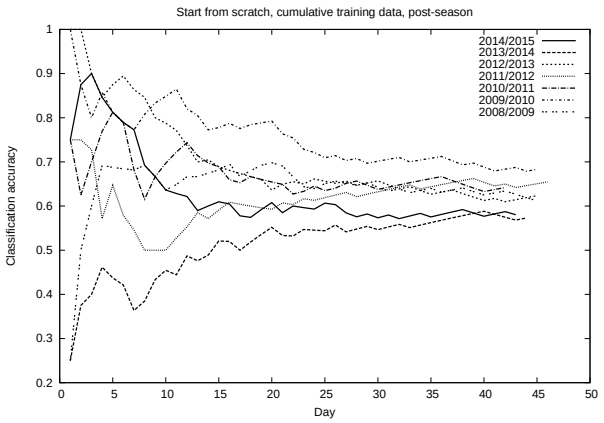

Figure 8: Accuracy for NBA post-season matches as a function of game days, MLP

Season	ANN	NB	# finals games
2014/2015	5/15	11/15	6 (0.66)
2013/2014	5/15	11/15	5 (0.8)
2012/2013	3/15	9/15	7 (0.57)
2011/2012	3/15	7/15	5 (0.8)
2010/2011	2/15	8/15	6 (0.66)
2009/2010	3/15	7/15	7 (0.57)
2008/2009	0/15	11/15	5 (0.8)

Table 11: NBA postseason series-level accuracies, and finals characteristics

in the first round. On the other end is the 2013/14 season, which saw three lower-seeded teams win their first-round series, as well as four higher-seeded teams needing seven games to win. In addition, in three of those series, the lower-seeded team won the first (away) game, another ingredient for the low early accuracy in the predictions for that season.

We have written above that the best-of-seven format of the NBA playoffs allows the stronger team to recover from off-games. If that were the case, predictors should do well at a series level, i.e. predict the four wins of the eventual winner.

Table 11 shows for each season, how many play-off series each classifier predicted correctly. Notably, the ANN never predicts more than 33% correctly yet its cumulative post-season accuracies were much higher than that. The reason is that it actually manages to correctly predict wins by the eventual loser of a series, something that the NB has problems with. The latter, on the other hand, predicts 11/15 (73.33%) for three seasons.

Finally, there are the Finals series, offering a particularly attractive test case: they consist of at most seven matches, with the number of actually played games an indicator of the relative strength of the two teams. Table 11 lists for each season the number of finals games giving us an upper bound on the accuracy when always predicting the series winner as match winner.

Figure 9 shows that NB shows this behavior (correctly predicting the eventual champion's wins, incorrectly predicting the eventual series loser's) for

- 2014/2015, when no one expected the Cleveland Cavaliers to win games 2 and 3, and the Golden State Warriors won the championship,
- 2013/2014, when the San Antonio Spurs won four matches by at least 15 points each, and lost one match by 2,
- 2012/2013, when it correctly liked the Miami Heat better than the San Antonio Spurs, although the latter's finals loss was considered somewhat of an upset at the time,
- 2008/2009, when the Orlando Magic were overmatched in the finals

but off

- (badly off) for the 2010/2011 series when Las Vegas (along with many others) clearly expected Miami to win, and
- for the competitive 2009/2010 series.

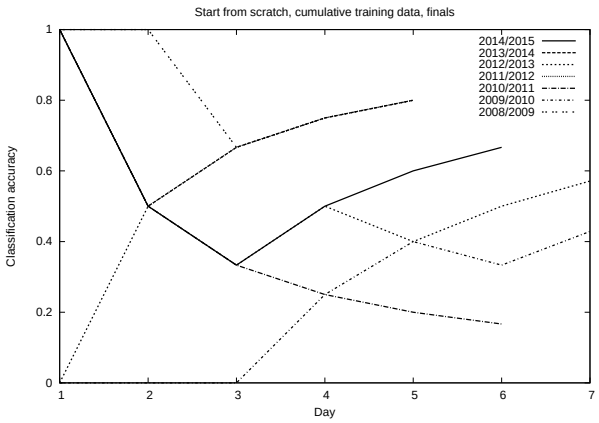

Figure 9: Accuracy for NBA finals matches as a function of game days, NB

2011/2012 is a particular case: while NB predicts four games correctly, one is for Oklahoma City, and only three for the eventual champion (and clearly superior team) Miami.

The MLP classifier, by contrast, does not predict a single finals series correctly but predicts enough wins of the series losers to achieve acceptable accuracies for several years.

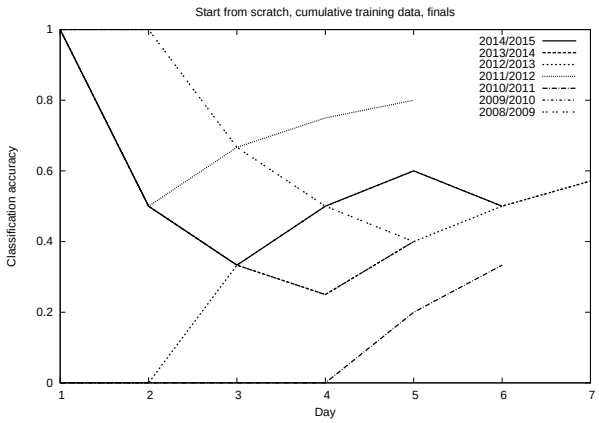

Figure 10: Accuracy for NBA finals matches as a function of game days, MLP

8.2 Discussion and perspectives

The experimental evaluation shows that the three phenomena we discussed in the introduction:

1. fewer matches in university basketball
2. wider skill spectrum during the regular season in the NCAAB, and
3. strongly differing post-season formats,

have noticable effects on predicting the respective leagues. The skill difference means that accuracies in the NCAAB stabilize relatively quickly but combined with the smaller number of games available for estimating statistics, this handicaps the models when it comes to predicting the post-season. This problem is exacerbated by the fact that lower-ranked teams need only one win to upset a higher-ranked one.

A way of addressing this issue could be to train classifiers on a subset of available matches. Limiting training data exclusively to post-season matches seems too restrictive to us: for the 2008/2009 season, for instance, we would have only 61 games available. Alternatively, one could try and augment these data with all regular season matches between the eventual post-season participants. Since NCAAB teams play only around 30 opponents, however, there are not too many of those available. We have recently shown that one can use clustering [12] to identify clusters of approximately equally strong teams. This could be used to identify good training instances in the data, beyond matches of post-season participants.

The models trained on the NBA regular season transfer more easily to the post-season, since there are more games for estimating statistics, closer skill level, and encounters between all teams. As a result, post-season accuracies are more in line with overall ones, even though the post-seasons stays harder to predict than for college teams. At the level of series winner predictions, however, the best-of-seven format helps the NB classifier to do rather well.

9 Conclusions and future work

In this paper, we have compared predictions on NCAAB and NBA match data in terms of different representations, predictive settings, and used classifiers. We have found that the adjusted efficiencies used in predicting college matches can be transferred to the case of NBA basketball, and turn out to be the most useful representation for that data. Other statistics and augmentation, for instance with standard deviations, proved less effective.

The particularities of the NCAAB – fewer matches, wider skill distribution, different playoff format – means that its regular season is relatively easy to predict but is not an optimal training set for predicting the post-season. In the NBA, the model can be more easily transferred but the league remains somewhat harder to predict, at least on the level of individual games.

NCAAB data is more effectively modeled by a neural network, a learner with a weak bias that can fit any decision boundary, whereas on the NBA data, a learner with strong bias – Naïve Bayes – is more effective. This is particularly pronounced when predicting the winner of playoff series in the NBA, where NB easily outperforms MLP.

A general issue is that our experimental results do not show a single superior option for prediction – neither in the case of the NCAAB nor in the case of the NBA. In fact, changes to the representation can change which seasons are predicted well, and

which classifiers perform best. One possible future option consists of using these different models (derived from different representations and classifiers), as ensembles. This will require solving the question how to weight each model's input and how to avoid correlated errors. Preliminary results of research we are currently undertaking indicate a second direction: some models seem to perform better in the early season, whereas others need time to ramp up. Identifying the best point to switch would allow to combine models. A third option that we intend to explore is to generate counterfactual data, inspired by Oh *et al.*, and using these data to train classifiers on larger sets of representative matches.

Finally, this study used cumulative accuracy as a performance measure to compare different options. In practical terms, the most immediate practical use of a predictive model would lie in helping to place sports bets. This requires that models be either a) better than betting odds set by sports books, or b) be close enough in accuracy that the times when the model is right and the odds wrong outweigh the times the opposite is true. Since this depends on the actual odds given, correctly predicting big upsets pays out more, evaluating predictive models in this manner requires additional data collection of closing odds for the matches under consideration.

References

[1] Loeffelholz Bernard, Bednar Earl, and Bauer Kenneth W. Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1):1–17, January 2009.

[2] Harish S Bhat, Li-Hsuan Huang, and Sebastian Rodriguez. Learning stochastic models for basketball substitutions from play-by-play data. In *MLSA15, Workshop at ECML/PKDD*, 2015.

[3] Mark Brown, Paul Kvam, George Nemhauser, and Joel Sokol. Insights from the LRMC method for NCAA tournament predictions. In *MIT Sloan Sports Conference*, March 2012.

[4] Mark Brown and Joel Sokol. An improved LRMC method for NCAA basketball predictions. *Journal of Quantitative Analysis in Sports*, 6(3), 2010.

[5] Eibe Frank and Ian H. Witten. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 1999.

[6] Alexander Franks, Andrew Miller, Luke Bornn, and Kirk Goldsberry. Counterpoints: Advanced defensive metrics for NBA basketball. MIT Sloan Sports Analytics Conference, 2015.

[7] Mohamed Medhat Gaber, Arkady B. Zaslavsky, and Shonali Krishnaswamy. A survey of classification methods in data streams. In Charu C. Aggarwal, editor, *Data Streams - Models and Algorithms*, volume 31 of *Advances in Database Systems*, pages 39–59. Springer, 2007.

[8] S. P. Kvam and J. S. Sokol. A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics*, 53:788–803, 2006.

[9] Min-hwan Oh, Suraj Keshri, and Garud Iyengar. Graphical model for basketball match simulation. In *MIT Sloan Sports Analytics Conference*, 2015.

[10] Dean Oliver. *Basketball on Paper.* Brassey's, Inc., 2002.

[11] K. Pomeroy. Advanced analysis of college basketball. `http://kenpom.com`.

[12] Albrecht Zimmermann. Exploring chance in ncaa basketball (originally in "MLSA15", workshop at ECML/PKDD 2015). *arXiv preprint arXiv:1508.04688*, 2015.

[13] Albrecht Zimmermann, Sruthi Moorthy, and Zifan Shi. Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned (originally in "MLSA13", workshop at ECML/PKDD 2013). *arXiv preprint arXiv:1310.3607*, 2013.