Taylor & Francis
Taylor & Francis Group

Journal of Business Analytics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjba20

Using Scouting Reports Text To Predict NCAA → NBA Performance

Philip Maymin

To cite this article: Philip Maymin (2021): Using Scouting Reports Text To Predict NCAA → NBA Performance, Journal of Business Analytics

To link to this article: https://doi.org/10.1080/2573234X.2021.1873077

Published online: 07 Feb 2021.

JOURNAL OF BUSINESS ANALYTICS
https://doi.org/10.1080/2573234X.2021.1873077

Using Scouting Reports Text To Predict NCAA → NBA Performance

Philip Maymin

Department of Analytics, Fairfield University Dolan School of Business, Fairfield, CT USA

ABSTRACT

Draft decisions by National Basketball Association (NBA) teams are notoriously poor. Analytics can help but are often dismissed for being too overfit, complex, risky, and incomplete. To address these concerns, we train separate leave-one-out random forests machine learning models for each collegiate NBA prospect from 2006 through 2019 with a conservative utility function on a novel comprehensive dataset including the raw text of scouting reports, combine measurements, on-court stats, mock draft placements, and more. Despite being unable to draft high school or international players, the resulting model outperforms the actual decisions of all but one NBA team, with an average gain of $100 million. Target shuffling shows that the model does not overfit and feature shuffling shows that handedness and ESPN mock draft rating, but not other mock drafts, are most important. NBA teams may be missing value by not following a disciplined, model-driven, prescriptive analytics approach to decision making.

ARTICLE HISTORY
Received 15 May 2020
Accepted 23 December 2020

KEYWORDS
NBA; basketball; machine learning; text mining

1. Introduction

The National Basketball Association (NBA) is the premier professional men's basketball league in the world, earning more than 8 USD billion in revenue per year (c.f. Wojnarowski & Lowe, 2020). According to the collective bargaining agreement (National Basketball Association and National Basketball Players Association, 2017), approximately half of that money is spent on player salaries. Thirty independent teams compete to sign the best players, subject to a complicated salary cap intended to maintain parity between small-market and large-market teams.

There are two types of salaries in the NBA: rookie scale contracts, where the salaries are effectively dictated by the CBA, and free agent contracts, which are subject to maximum values. Rookies enter the league through a two-round draft system whereby teams that have performed poorly in the preceding season get earlier choices. Rookie contracts are a great deal because the fixed salaries are substantially lower than average league salaries. Teams often trade up for a higher draft pick in order to be able to select the player of their choosing. Indeed, drafting well is one of the most important tasks of an NBA team's front office, and each team employs numerous scouts, invites prospects to workouts, and otherwise spends all year trying to determine which 19-year-olds are likely to become productive professionals.

Most rookies are drafted after playing one year of collegiate basketball (NCAA), though international players are eligible as well. The NCAA reports that each year, approximately 4000 college players are NBA-draft eligible, of which at most sixty, and, because of international players, usually fewer, are drafted (NCAA, 2019). The task facing each of the thirty NBA teams is about ordering their choices for the draft so that when their turn comes, they are ready to select the best available player. While many staff members are involved and assist in data-gathering for the decision, ultimately the choice is made by the general manager of the team, and routinely that decision is not the result of an analytical model.

Indeed, the use of business analytics by sports teams can be described as reluctant at best. On the ticketing side, business analytics is prevalent and pervasive (c.f. Mondello & Kamke, 2014), but on the court or the field, analytics is often used at most as simply another input. A common phrase used by general managers to indicate their approximately equally weighted inputs are "eyes, ears, and numbers," referring to scouts, peers, and analytics.

There may be many reasons why analytics has not been the driver of decision making on sports teams, particularly in decisions regarding player selection: owners may want to participate in the decision making, the market for general managers may be inefficient, or analytics simply may not work.

Specifically, the academic research gap in sports analytics is failing to allay the natural human suspicion of analytics, which has four primary criticisms: overfitting, complexity, riskiness, and incompleteness. This paper conclusively disproves the claim that analytics may not work in sports in player selection by undermining each of those four concerns, thus leaving only behavioural or organisational reasons for underutilisation of business analytics on the basketball side.

CONTACT Philip Maymin philip@maymin.com

Overfitting is the natural concern that a model is using information known now to allegedly predict what it would have done in the past. To address this, we use a separate model for each prospect. No information about player X is in the model used to predict the performance of player X. Furthermore, we utilise the technique of target shuffling (Elder, 2009, 2014) to show that our model does not over-explain.

Complexity is the general "black-box" worry that a model is not understandable and therefore not reliable. To address this, we use a standard random forests machine learning algorithm, and, further, we utilise feature shuffling to investigate variable importances. We also note a simple rule of thumb that human front office decision makers can use.

Riskiness refers to the approach taken in sports analytics generally of seeking to model actual values without regard to business conditions. Specifically, a player who underperforms expectations does far, far more damage to an organisation than the additional value provided by a player who exceeds expectations. To address this, we use a highly skewed utility function to punish underperformance effectively an order of magnitude more than overperformance.

Incompleteness is the critique that valuable information about prospects was not included in the analysis. Specifically, professional scouts often argue that their reports and insights are not incorporated into the analyses. To address this concern, here we use the most comprehensive and complete dataset of prospect information possible, including the raw text of historical scouting reports.

Our results show that a machine learning model trained without hindsight bias would have outperformed human decision making for nearly every single team for the past decade or more. Our approach explicitly addresses each of the four criticisms above with comprehensive data, risk-adjusted performance metrics, explainability enhancements, and out-of-sample testing. The first team to fully adopt business analytics to its most important front office function, namely, choosing which players to draft, will likely reap substantial rewards. There are no longer any justifiable reasons to eschew an overwhelmingly analytical approach to player drafting.

1.1. Literature Review

Descriptive analytics in sports goes back hundreds, if not thousands, of years, and is most famous in baseball, owing to the sabermetrics revolution of Bill James and others (c.f. James, 2010). Some of the innovations include different combinations of box score statistics to determine efficiency, such as runs created or range factor. Others include more complicated formulas to assess an individual player's contribution to a team win or to estimate the number of wins from the number of runs scored and allowed. In basketball, one of the earliest foundational descriptive analytics approaches is Oliver (2004), whose primary contribution was in the four relatively independent factors he identified as being the key components of basketball success: effective field goal percentage, offensive rebounding percentage, free throw rate, and turnover percentage.

Predictive analytics has two phases of history. In the earlier sabermetrics phase, standard linear or logistic regressions were used to forecast future wins or production. Modernly, predictive analytics has turned to machine learning techniques, particularly in applications with access to enormous quantities of data. Miller et al. (2014), for example, analyse the 25 frame-per-second geospatial optical tracking data of made and missed field goal attempts of NBA players to predict scoring efficiency. Jain and Kaur (2017) introduce and use a variant of support vector machines to predict the outcome of basketball games. Lopez and Matthews (2015) use ensembling methods to predict NCAA March Madness outcomes. Harris and Berri (2015) and Berri et al. (2011) conclude what they refer to as the "strange result" that human beings with access to additional qualitative information beyond college performance numbers do worse than their model, but they do not provide an estimate of the difference.

Prescriptive analytics is quite sparse, however. The most natural application in the NBA would be to the draft, the annual system by which the professional teams select players in a predetermined order from a given pool of eligible high school, college, and international players. Even here, most applications have been primarily predictive rather than prescriptive. Kannan et al. (2018), for example, use draft order and college performance history to predict a player's longevity. Harris and Berri (2015) show that the women's professional basketball draft is poorly explained by future productivity, extending a similar result by Berri et al. (2011) on the men's professional basketball draft. Evans (2018) similarly shows that certain factors, such as college turnover percentage, are suboptimally ignored by human drafters, and suggests improvement can therefore be possible, but does not measure the precise amount of possible improvement nor provide a comprehensive prescriptive model. These are not oversights but a necessary result of the non-prescriptive analytics approach which effectively uses the entire sample period to evaluate a hypothesis. By contrast, here we try to create a prescriptive model, without hindsight bias, that can also be backtested.

This relative paucity of prescriptive analytics may partly contribute to, as well as largely result from, the failure by pro teams to adopt a primarily automated approach to decision making. One exception is Maymin (2017) which attempts to automate front-office decision making in draft, free agency, and trades. However, that paper may be prima facie criticised on the basis of all four concerns noted above: the earlier dataset may have been overfit, the model was too complex, the utility function did not reflect business risk factors, and the data lacked other external information, most notably the scouting reports. This paper addresses each of those critiques by extending to a longer and more recent dataset, by using a simpler machine learning model without any hyperparameter tuning as well as noting a simple rule of thumb for human decision makers, by incorporating a highly skewed and more realistic utility function, and by incorporating a host of substantively new additional variables, including the raw text of scouting reports.

The use of Elder's (2009, 2014)) target shuffling process for assessing NBA draft models has not been present in any of the prior literature.

For team managers, the contributions of this paper allow for a structured, analytical framework for ordering among possible prospects. For team owners, the model provides a benchmark for evaluating front office human decision making. For players and prospects, the insights here can help them better evaluate both their likelihood of being drafted and what skills or other metrics can be improved to bring them the greatest value. For fans, a wider and more consistent adoption of the analytical methods here would improve the quality of competition in the world's premier league.

1.2. Summary of Results

In this paper, we train nearly 1000 leave-one-out individual random forests machine learning models, one for each collegiate NBA draft prospect who was in the ESPN Top 100 prospects list prior to the draft, using their college efficiency and production stats, combine measurement information, high school RSCI and NBA mock draft placements, handedness, ethnicity as estimated from a separate machine learning model based on their name (appeler/ethnicolr on github), and also scouting evaluations and raw scouting text scraped from nbadraft.net for 2006–2019 and processed through term frequency–inverse document frequency (TFIDF) then dimension reduced with latent semantic analysis (LSA).

The forecast variable is an average of three standard win production measures (Win Shares, Wins Produced, and Estimated Wins Added) over each player's three years after the draft, or zero for years they did not play. The utility function for the machine learning model penalises model overestimation far worse than model underestimation. Target shuffling is used to assess model efficacy and feature shuffling is used to assess variable importances.

The correlation between predicted and actual NBA production is 63% and the average error is less than 3 wins per year. Every team except the Denver Nuggets would have benefitted from drafting based on this model rather than the decisions they actually made, even though the model does not have hindsight bias and can only draft NCAA prospects.

The average model pick outperformed the actual pick by 70% and the average team lost out on 100 USD million worth of on-court production. The Minnesota Timberwolves fared the worst of all the teams compared to the model, losing out on 272 USD million worth of on-court production, or about two-thirds of their 2014 Forbes-estimated franchise value.

For 2019, the model valued Cameron Reddish, Brandon Clarke, and Bol Bol much higher than their mocks, suggesting they will be the sleepers of the draft.

2. Data

Table 1 lists the sources and types of data used as input for the projections.

NCAA and NBA data were scraped from realgm.com, combine measurements and Recruiting Services Consensus Index (RSCI, a measure of the prospect's rank at the end of high school) from draftexpress.com and espn.com, and mock drafts from ESPN, draftexpress.com, mynbadraft.com, CNN/SI, and nbadraft.net. NBA Wins Produced numbers through 2012–2013 were provided courtesy of David Berri of wagesofwins.com and the rest were scraped from boxscoregeeks.com.

The British Probability uses the ethnicity predictor for given names of Laohaprapanon and Sood (2019). While almost all the prospect names had "GreaterEuropean,British" as their most likely race, the probability differed across prospects. Figure 1 shows the histogram of the British probability across all prospects.

The scouting numbers for twelve different categories come from nbadraft.net. Each scouting report on a prospect includes a numerical rank on athleticism, size, defence, strength, quickness, leadership, jump shot, NBA readiness, ball handling, potential, passing, and intangibles.

While technically the ranks range from 0 to 10, almost all of the ranks for each category ranged from 5 to 10. The only lower rankings in the dataset were

Table 1. Collegiate Stats, Rankings, Combine Metrics, and Other Inputs.

RSCI	Ranking of the prospect when they were in high school, from draftexpress.com
Mock	Mock drafts: historical projected orderings by nbadraft.net, draftexpress.com, ESPN, CNN/Sports Illustrated, and mynbadraft.com
Awards	The number of awards won by the prospect, from realgm.com
Wingspan, Reach, Max Vertical, Sprint, Agility, Bench, Body Fat, Hand Width, Hand Length, Height, and Weight	Physical measurements from the NBA combine, from draftexpress.com
British Probability	Probability that name is British, using model of github.com/appeler/ethnicolr
Strength of Schedule, Position, Age, Games Played, and Minutes	Strength of schedule, prospect's position (Guard, Forward, Center), age on June 15th of their draft year, games played, and total minutes
Hand	Whether the prospect is right- or left-handed, or unknown
eFG%	Effective field goal percentage
ORB%	Offensive rebound percentage
DRB%	Defensive rebound percentage
AST%	Assist percentage
TOV%	Turnover percentage
STL%	Steal percentage
BLK%	Block percentage
USG%	Usage percentage
Min/PF	Minutes played per personal foul committed
Min/3FGA	Minutes played per three-point field goal attempt
FT/FGA	Free throws attempted per field goal attempt
PER	John Hollinger's Player Efficiency Rating
PPS	Points per shot
ORtg	Dean Oliver's Offensive rating
DRtg	Dean Oliver's Defensive rating
OWS	Offensive win shares
DWS	Defensive win shares
3PT%	Three point field goal accuracy
FT%	Free throw accuracy
Scouting Numbers	1–10 score in: athleticism, size, defence, strength, quickness, leadership, jump shot, NBA readiness, ball handling, potential, passing, and intangibles
Scouting Raw Text	The historical pre-draft scouting report on the prospect,from nbadraft.net

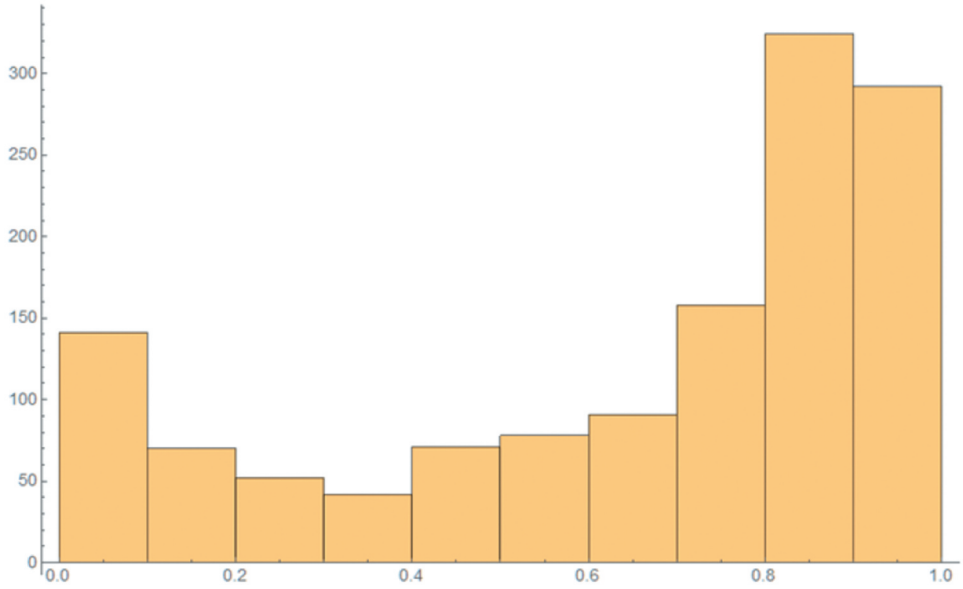


Figure 1. Histogram of the Probability a Prospect's Name is British, per ethnicolr.

four 4's, given to Lou Amundson for his jump shot, to Quincy Douby and Ryan Hollins for strength, and to Steve Novak for athleticism.

Out of the 1,379 prospects in the sample with scouting reports, 289 had missing values for all twelve categories; they were thus excluded from the sample. Of those 289 excluded, only three were projected by ESPN mock drafts to be in the first round: Ryan Anderson, Mitchell Robinson, and Matisse Thybulle. Of the rest, about two-thirds did not appear in the top 100. Of the remaining 1,090 prospects and 12 categories each, there were hardly any missing values at all: only 47 missing values out of 13,080, so less than half of a per cent.

Some prospects had multiple scouting reports across multiple years. In all cases, only the last scouting report prior to the draft was considered. In other words, only information and scouting reports available prior to the time of the drafting decision are used.

The scouting text is structured as follows. Refer to Appendix 1 for a snapshot of an example.

First, a comparison to an NBA player is made. Often this is one player, sometimes two, and occasionally there is no comparison. Of the 1,090 prospects with scouting numbers, 950 had at least one NBA comparison, and 140 did not. However, the presence or absence of a comparison does not appear to be a bias in terms of the scouting evaluations. Table 2 shows the average value for each category for prospects with NBA comparisons ("NBA Comp") and without ("No Comp"), along with a p-value from a T-test. Only Leadership and Jump Shot appear to have a statistically significant difference between the two groups.

In any case, there are very few repeat comparison players: the 1,090 prospects were compared to 790 NBA players. The most frequent comparison was to Ben Gordon, who was the NBA comparison for six prospects (O.J. Mayo, Toney Douglas, Willie Warren, Shelvin Mack, Dion Waiters, and Jamal Murray). Because of this lack of repeatability and therefore relative uselessness in machine learning, the NBA comparison portion of the scouting report text was removed.

Next, the scouting report has up to four sections: strengths, weaknesses, overall/outlook, and notes. About two-thirds of scouting reports listed strengths and weaknesses, but only about half listed notes and only about a third listed an overall/outlook. Here is where we begin to quantify the text.

First, we simply count the number of characters in each section, divided by the total number of characters in the report. These four computed numbers are appended to the input parameters for each prospect. Figure 2 shows the histograms of these portions for each section type.

Table 2. Presence or Absence of NBA Comparison.

Category	NBA Comp	No Comp	p-value
Athleticism	7.79	7.79	0.995
Size	7.83	7.69	0.166
Defence	7.37	7.27	0.211
Strength	7.32	7.29	0.658
Quickness	7.55	7.52	0.774
Leadership	7.32	7.13	0.008
Jump Shot	7.63	7.38	0.005
NBA Ready	7.45	7.40	0.499
Ball Handling	7.57	7.52	0.509
Potential	7.46	7.36	0.295

Next, we convert each word into its stem using the standard Porter (1980) stemming algorithm. Then, treating each of the four sections separately one-by-one, we apply the term frequency-inverse document frequency (TF-IDF) algorithm (c.f. Jones, 1972; Salton & Buckley, 1988), and then reduce the dimensionality with latent semantic analysis (LSA) so that the otherwise large number of columns does not overwhelm the rest of the input parameters.

Finally, we also append, for each of the four sections, its sentiment score. This sentiment score is calculated with a separate, standard Wolfram Language Documentation (2019) machine learning model of text sentiment. The sentiment score for a sentence indicates the model probability that the underlying sentiment is positive. The overall sentiment for each section is then the average sentiment of the sentences.

The sentences in the raw texts of the scouting reports are not always properly delimited with periods. Sometimes they simply start with a capital letter. Sometimes they are separated by ellipses, either an actual ellipsis or a sequence of three periods. When multiple scouting reports exist, they are concatenated within groups: all the strengths sections are concatenated together, and all the weaknesses sections, and so on. Adjusting for these possibilities, primarily by inserting a period after words other than stopwords that are followed by a capitalised word, there are typically a few dozen sentences per section per prospect.

Thus, the total number of features computed from the raw text of the scouting reports, and not including the twelve separately tabulated skill rankings, is the proportion of total text length for each of the four categories (1 * 4), plus 2 LSA dimensionally reduced TF-IDF vectors for each of the four categories (2 * 4), plus 1 sentiment score for each category (1 * 4), which sums up to 16, and does not overwhelm the 12 skill rankings plus 44 other non-text features for a total of 72 inputs.

3. Method

Figure 3 shows a graphical depiction of the analytical pipeline.

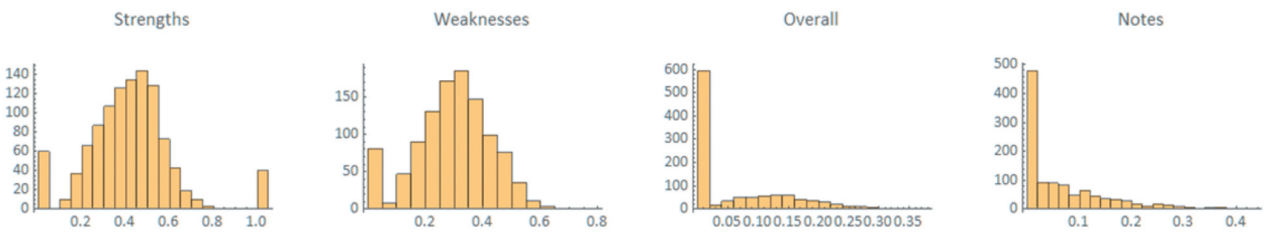


Figure 2. Histograms of the Portion of the Total Report Text Occupied by Each Section Type.

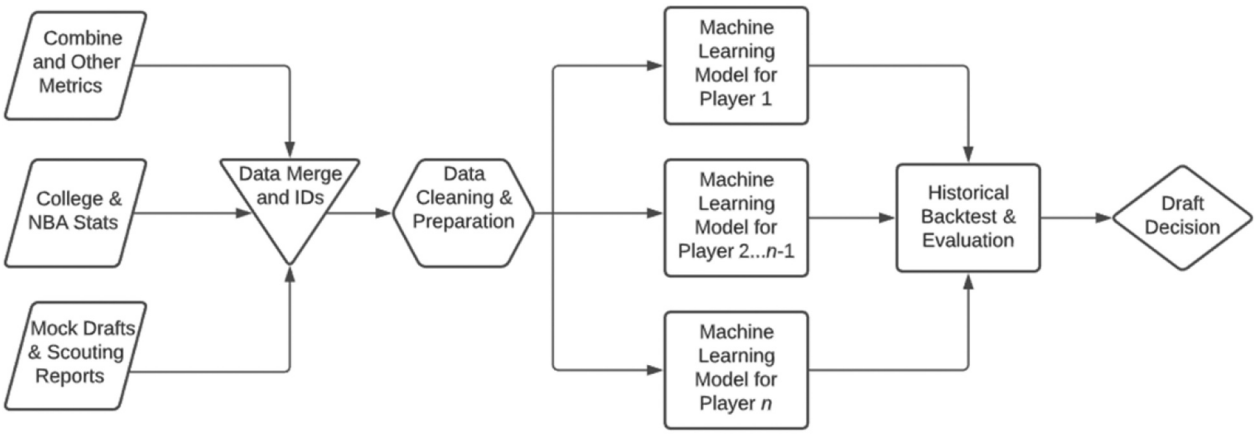


Figure 3. Graphical Depiction of Analytical Pipeline.

For each college prospect in the sample period, a leave-one-out machine learning algorithm is trained on all other prospects, excluding the prospect under consideration from the training set.

The training data for each other prospect includes their collegiate performance, combine measurements, mock draft positions, certain computations based on the raw text of their scouting reports, and other metrics as detailed above in section 2.

That trained model is then applied to the prospect under consideration who had been left out of the training sample. The model's prediction then forms the basis for determining who to draft by comparing to the predictions formed by this same method for other prospects.

3.1. Dependent Variable

The dependent variable used here is Wins Made, a simple average of three standard NBA win-based productivity metrics, over the first three years of the prospect's NBA career, matching the choice of Maymin (2017), and for the same reasons, as described further below: each win-based productivity metric has its own strengths and weaknesses, so averaging them reduces the biases.

3.2. Machine Learning Methods

Because of the large variety and types of data used, a machine learning approach is more applicable than a regression-based one. Multiple machine learning methods were initially tried including standard neural networks, deep neural networks, gradient boosted trees, the self-normalising neural networks of Klambauer et al. (2017), and random forests. These are standard machine learning algorithms for predictive tasks. Because the ultimate size of the dataset is only several hundred rows and several dozen features, none of the neural network approaches fared particularly well, despite significant training time. The overall best models came from random forests, and is the machine learning method used throughout the remainder.

Random forests is an ensemble method where multiple individual decision trees are randomly selected from the list of all features and a vote or average determines the overall prediction. One can think of random forests as an amalgamation of a variety of uncorrelated human analysts, each of whom is looking at different subsets of features in trying to predict the future productivity.

The utility function, or the inverse of the loss function, for random forests is usually symmetrical, when accuracy is the simple goal. In the context of forecasting athletic performance, a surprise to the upside is far less costly than a surprise to the downside. Thus, the utility function we employ is heavily skewed to punish overoptimistic predictions.

Figure 4 shows the dramatic difference. The skewness parameter $\lambda = -0.825$ was chosen so the penalty from a prospect outperforming by five wins per year was approximately one-tenth the penalty of underperforming by five wins per year.

For comparison, the model chose Karl-Anthony Towns as the best overall choice in 2015, predicting he would add 2 wins per year, but after the Timberwolves selected him #1 overall in 2015, he ended up being far better still, adding 15 per year during his first three years. That means the model underestimated his contribution by 13 wins per year.

On the flip side, the biggest underperformer was Anthony Bennett. The model thought he was the fourth best college prospect in the draft, predicting he would average about 1.3 wins per year. The Cavaliers selected him #1 overall in 2013 and he ended up producing negative wins. The model overestimated his contribution by 1.6 wins per year.

With the asymmetric utility function chosen, both of those choices would be penalised by approximately

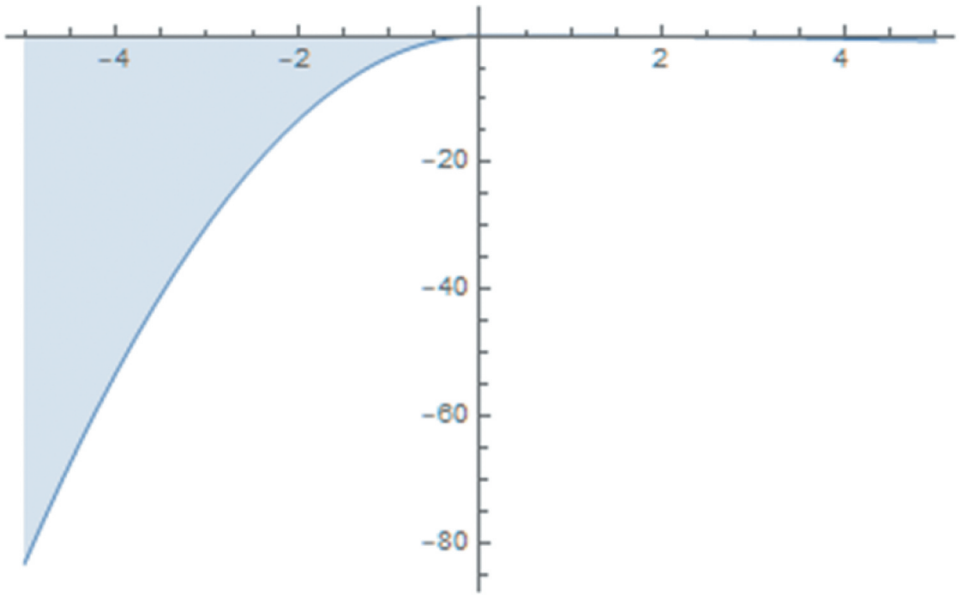


Figure 4. **Asymmetric Utility**: a = **actual value**, p = **predicted value**, λ = **skewness parameter** $\mathcal{U}(a,p,\lambda) = -(p-a)^2(\text{Sign}(p-a)+\lambda)^2$.

the same amount, despite the vastly different magnitude. (In fact, the Bennett overestimation would be penalised 64% more.)

3.3. Win Values

Several metrics attempt to allocate a team's wins among its roster. Here we use the average of three prominent such metrics. Wins Produced (WP) (Berri, 2011), Win Shares (WS) (Kubatko, 2019), and Estimated Wins Added (EWA) (Hollinger, 2019). Each one uses box-score-level data to estimate production, and each one arguably has certain tilts. WP, for example, tends to weigh rebounding ability more heavily, while EWA tends to reward inefficient shooting in some circumstances. Taking the average and calling it Wins Made (WM) aims to reduce their individual biases. To be sure, the entire analysis could easily be redone with any of these individual metrics, or any other metric.

Wins are converted into dollars using the following standard computation. Over the sample period in question, league revenue totalled about 4 USD billion, and players received half of that, or about 2 USD billion. There are 82 games for each of 30 teams, so the number of wins available is 82 * 30/2 = 1,230. Therefore, on average, a win is worth about 2 USD billion divided by 1,230, or about 1.65 USD million. These numbers will increase in the future, relative to the time period examined here.

3.4. Leave-One-Out Machine Learning

Leave-one-out-cross-validation (LOOCV) is one type of cross-validation technique used to train better machine learning models, and it involves cross-validating a model on every individual input. Here, however, the method is used to maximise training ability on a per-prospect basis without using correlated information. Obviously, if the prospect is in the training set, the model will overfit because it will learn what his eventual output was. Thus, instead of training a single overall model, here we train nearly 1000 individual models.

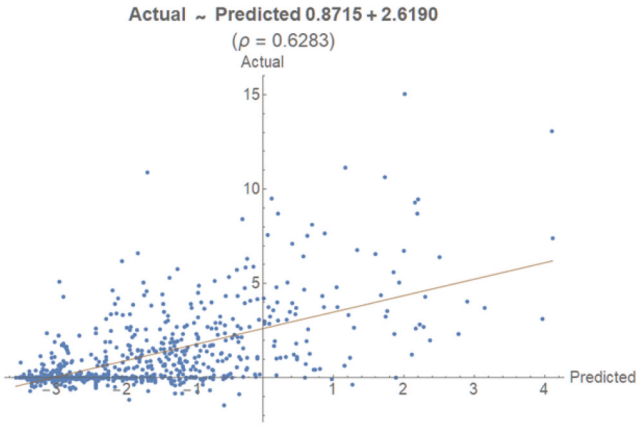
Such a technique would not be appropriate in all circumstances. In modelling financial time series, for example, leaving out one day of prices but using both past and future prices introduces inevitable hindsight bias, because prices are highly auto-correlated. Such a technique would also not be appropriate across multiple decades of prospects because the skills, stats, and rules of the 1970s are different than they are today.

In our case, however, it is entirely appropriate, because (1) the game, while it did evolve over the time period in question, did not change drastically, and (2) the performance of a player drafted years later than a particular prospect adds no predictive value. To put it another way, this approach effectively assumes that a player plucked from the year he was drafted and placed on a similar team in a similar role but in a different year would likely perform approximately the same.

4. Results

The correlation between predicted average Wins Made (WM) over a prospect's first three NBA seasons, and the actual WM, is 63%, varying with the prospect's position, as shown in Figure 5.

Following Shmueli and Koppius (2011), correlation and similar metrics may not be as relevant in predictive studies; thus, Figure 6 shows the highly right-skewed



Position	Prospects	Correlation
GF	34	0.4632
SG	105	0.4989
C	84	0.5995
G	143	0.6144
SF	89	0.6221
PF	122	0.6255
F	134	0.6673
PG	72	0.7111
FC	34	0.7201

Figure 5. Correlation between Predicted Wins Made (WM) and Actual Wins Made (WM).

distribution of the actual minus predicted values. This skewness is as expected from our utility function that penalises overestimation more severely than underestimation. The mean error is 2.86 wins per year; the mean absolute error is 2.89 wins per year.

The model reveals a simple rule of thumb for draft decision makers. Figure 7 shows the predicted wins made as a function of the quantiles across each feature. For example, the first dot, for the median (quantile = 0.5), is computed as the predicted wins for a hypothetical player whose input characteristics are all equal to the median of that characteristic across the entire sample. (For the two non-numeric input features, we assume the player is a right-handed forward.) Because of the skewed utility function, a positive prediction is not made until the hypothetical player is above the 95th percentile in all of the inputs. However, the simple rule of thumb is consistent across all quantiles: merely computing each player's average quantile across their features goes a long way to evaluating the player with the formal machine learning model.

4.1. Average Draft Pick Performance

One way to evaluate the model is to consider how its picks did against actual human picks. Figure 8 shows every draft pick plotted with the average productivity of the actual human choices along the x-axis and the average productivity of the model choices along the y-axis. The model substantially outperformed the actual human decisions in every draft pick, on average across the time period.

4.2. Team Performance

Table 3 shows how each individual team would have fared had they followed the machine learning model forecasts instead of their own human judgement.

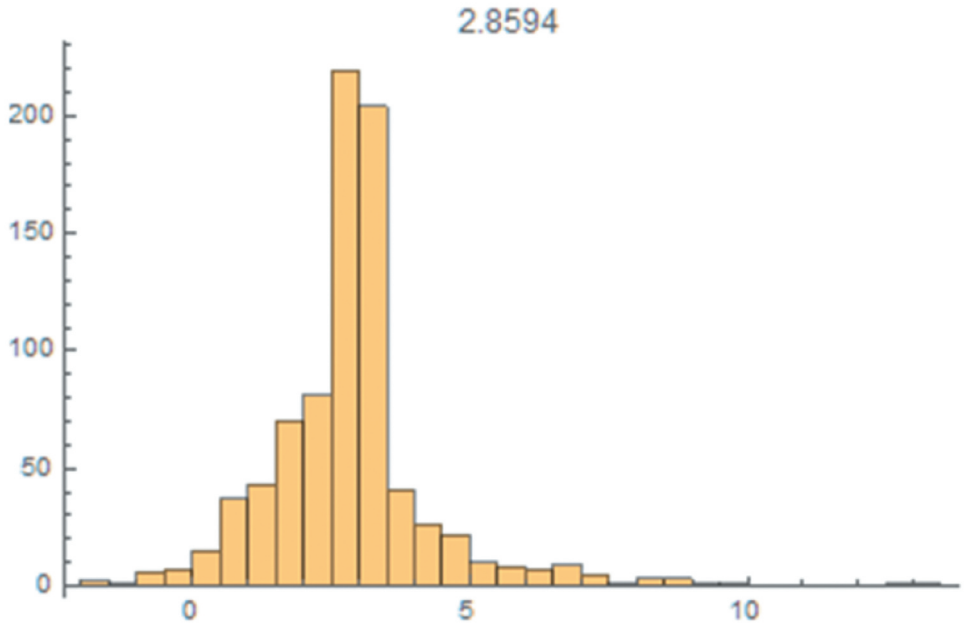


Figure 6. Histogram of Net Production (Actual Minus Predicted Wins Made); Mean Displayed.

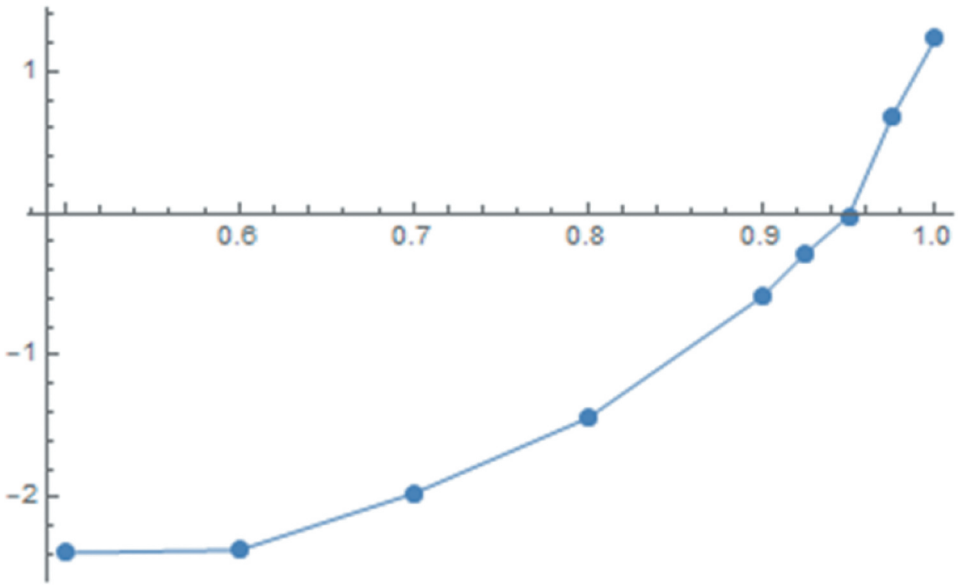


Figure 7. Predicted Wins Made Versus Feature Quantiles.

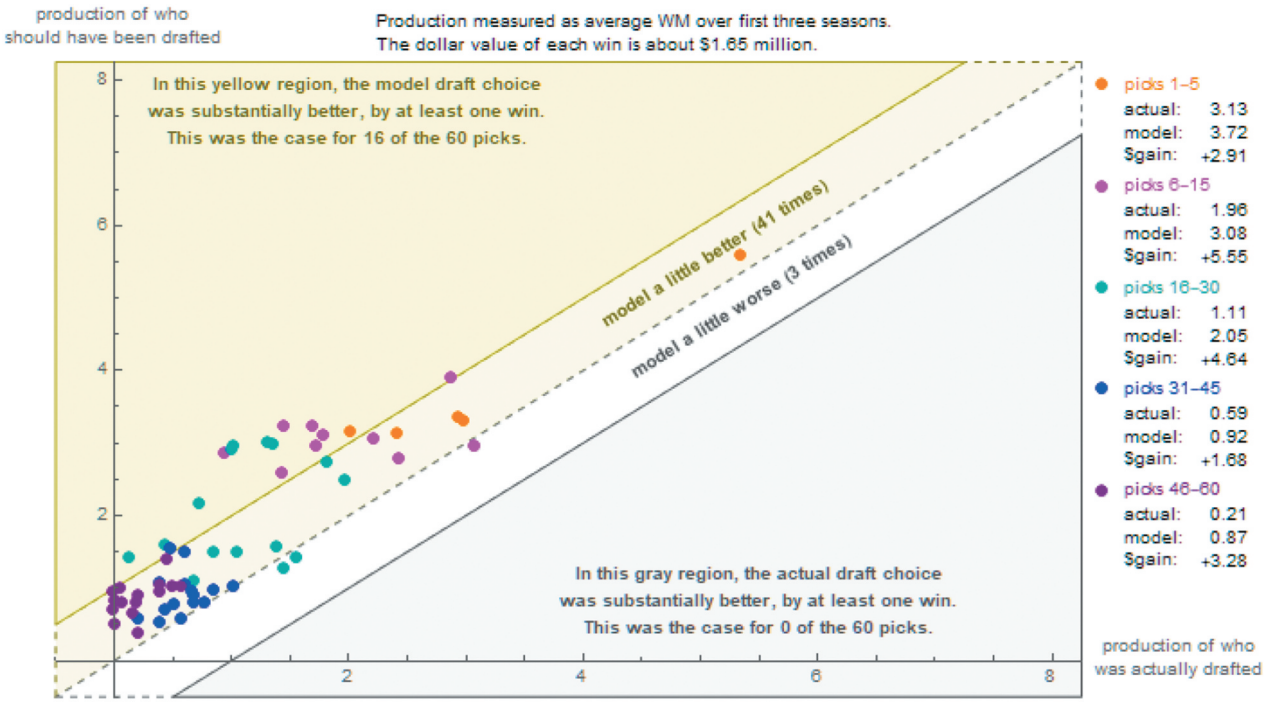


Figure 8. Scatterplot of Net Production (Actual Minus Predicted Wins Made) for Model versus Actual Draft Choices; Mean Displayed.

These numbers are "partial-equilibrium" results in the sense that, of course, only one team can follow the advice of the model.

Minnesota lost out on nearly two-thirds of its entire franchise value due to its drafting choices, despite the fact that they drafted one of the best performing players of the time period, Karl Anthony-Towns, in agreement with the model pick for first overall pick that year.

The average team lost out on 100 USD million worth of value, and the average model pick outperformed the human pick by over 70%.

The Denver Nuggets were the only team to have a better historical drafting record than the model.

4.3. Target Shuffling

In evaluating a machine learning model, the biggest fear by human decision makers in a sports context is that of overfitting. How can we know that the black box we found is indeed a robust explanation that will continue into the future rather than merely reflect an artefact of past data?

Table 3. Historical Team Performance Compared with the Model.

	Team	Total Picks	Avg. Production	Avg. Production	Avg. Production	from Model (total lost profits)	Team Value Forbes (2014)	Team Value
1	MIN	34	1.34	2.96	1.62	$27,24,68,485	$43,00,00,000	63.36%
2	GOS	25	0.95	2.72	1.77	$21,88,92,031	$75,00,00,000	29.19%
3	MEM	31	0.92	1.98	1.07	$16,40,08,911	$45,30,00,000	36.21%
4	MIL	26	1.29	2.47	1.18	$15,23,67,361	$40,50,00,000	37.62%
5	PHX	29	0.43	1.47	1.04	$14,99,40,231	$56,50,00,000	26.54%
6	PHL	40	1.09	1.84	0.75	$14,88,97,564	$46,90,00,000	31.75%
7	CLE	32	0.84	1.70	0.86	$13,69,59,020	$51,50,00,000	26.59%
8	TOR	18	1.52	2.97	1.45	$12,93,54,239	$52,00,00,000	24.88%
9	OKC	34	1.31	2.06	0.75	$12,66,96,136	$59,00,00,000	21.47%
10	SAC	34	0.97	1.70	0.73	$12,27,79,990	$55,00,00,000	22.32%
11	NOP	24	1.29	2.33	1.03	$12,25,08,200	$42,00,00,000	29.17%
12	ATL	27	0.98	1.82	0.84	$11,20,05,530	$42,50,00,000	26.35%
13	CHI	25	1.57	2.38	0.81	$10,07,28,829	$1,00,00,00,000	10.07%
14	NYK	28	0.98	1.69	0.71	$9,78,58,657	$1,40,00,00,000	6.99%
15	ORL	23	0.92	1.72	0.80	$9,09,50,588	$56,00,00,000	16.24%
16	BRK	31	1.18	1.77	0.59	$9,07,43,546	$78,00,00,000	11.63%
17	LAC	25	0.86	1.49	0.64	$7,88,90,878	$57,50,00,000	13.72%
18	DAL	21	0.52	1.28	0.76	$7,87,12,023	$76,50,00,000	10.29%
19	WAS	21	0.98	1.71	0.73	$7,61,50,004	$48,50,00,000	15.70%
20	SAN	31	0.68	1.15	0.47	$7,25,44,739	$66,00,00,000	10.99%
21	BOS	38	0.88	1.24	0.37	$6,90,09,907	$87,50,00,000	7.89%
22	UTH	32	1.21	1.63	0.42	$6,60,40,656	$52,50,00,000	12.58%
23	IND	23	1.09	1.66	0.57	$6,46,01,483	$47,50,00,000	13.60%
24	LAL	29	0.71	1.16	0.45	$6,42,07,345	$1,35,00,00,000	4.76%
25	MIA	17	0.82	1.51	0.68	$5,76,18,211	$77,00,00,000	7.48%
26	HOU	28	1.29	1.67	0.38	$5,30,81,036	$77,50,00,000	6.85%
27	CHA	24	1.10	1.52	0.42	$4,94,56,892	$41,00,00,000	12.06%
28	POR	32	1.54	1.82	0.28	$4,49,04,522	$58,70,00,000	7.65%
29	DET	33	1.16	1.39	0.22	$3,64,87,015	$45,00,00,000	8.11%
30	DEN	25	1.41	0.98	-0.42	($5,23,87,972)	$49,50,00,000	-10.58%

One excellent but rarely used technique is Elder's target shuffling (Elder, 2009, 2014). The technique applies to any black-box-generating-process. In our case, the process is the machine learning model on the data described above.

The basic idea of Elder's target shuffling is to treat the entire machine learning process as a function: randomly rearrange the output variable among the instances, run the black-box-generating-function, and measure the strength of the resulting prediction. Repeating that multiple times yields a distribution of predictive strength, relative to which the original strength can be compared. Target shuffling thus measures the function's sensitivity to the outcome variable.

One way to think about Elder's technique is that it addresses the charge of over-explanation or over-fitting: is the black-box-generating-process so general that it could explain or fit *any* output data? If so, then it effectively explains too much, and is not useful.

Figure 9 shows the histogram of the correlation of the predicted with the actual wins made across 100 target shuffling simulations, overlaid with the actual correlation observed in the original unshuffled data. The effective p-value here is zero: the maximum

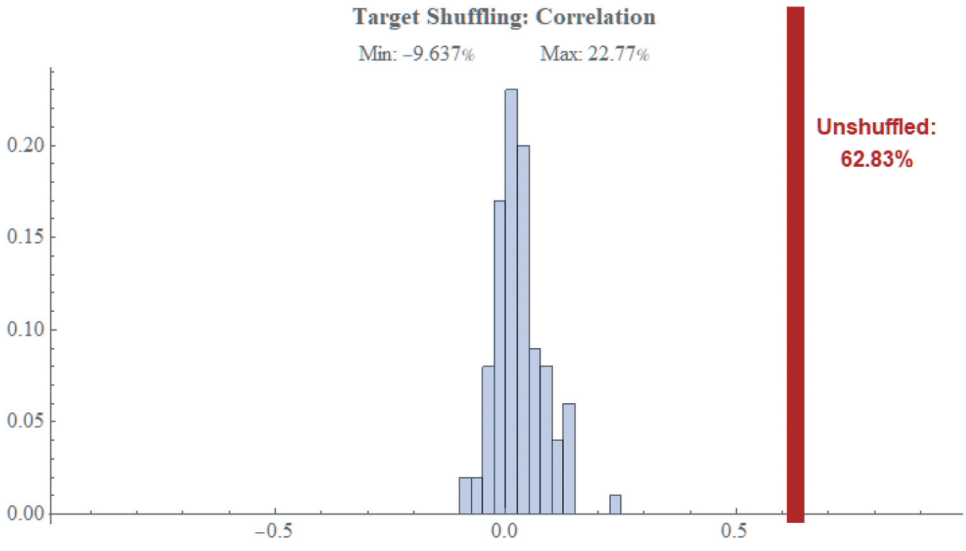


Figure 9. Target Shuffling Histogram of Correlation of Actual and Predicted Wins Made.

correlation observed among target shuffled datasets was less than 23% while the original correlation was nearly 63%.

4.4. Feature Shuffling

While target shuffling addresses the concern of over-explanation, feature shuffling addresses the concern of over-specification. Similarly to target shuffling, we can shuffle each the column of each feature's data across all instances to get a measure of how important that feature is to the model.

Unlike target shuffling, however, no new models need to be generated. When the output variable is shuffled, then a new model needs to be generated for each simulation. When an input variable is shuffled, however, the original model merely needs to be re-run on the new input dataset.

Figure 10 shows the feature shuffling bar chart of variable importances, calculated as the difference in correlation between predicted and actual wins made. Every feature has some importance, but the two single most important features are the handedness of the player and his ranking on the ESPN mock draft. Interestingly, the other mock drafts add marginally little value relative to the ESPN one.

Exploring the handedness features a little further, there are 277 right-handed players, 26 left-handed players, one ambidextrous player (Tristan Thompson), and 521 missing values. What happens if those four values are replaced with each of the others in turn?

Table 4 lists the resulting change in correlations for the various subsets of handedness listed. (Note that because there is only one ambidextrous player in the dataset, correlations are uncomputable, but the forecast wins made increased if the player was deemed left-handed, and increased even further if the player was deemed right-handed.)

Because each row examines only the subset of data matching that handedness, the diagonal elements can

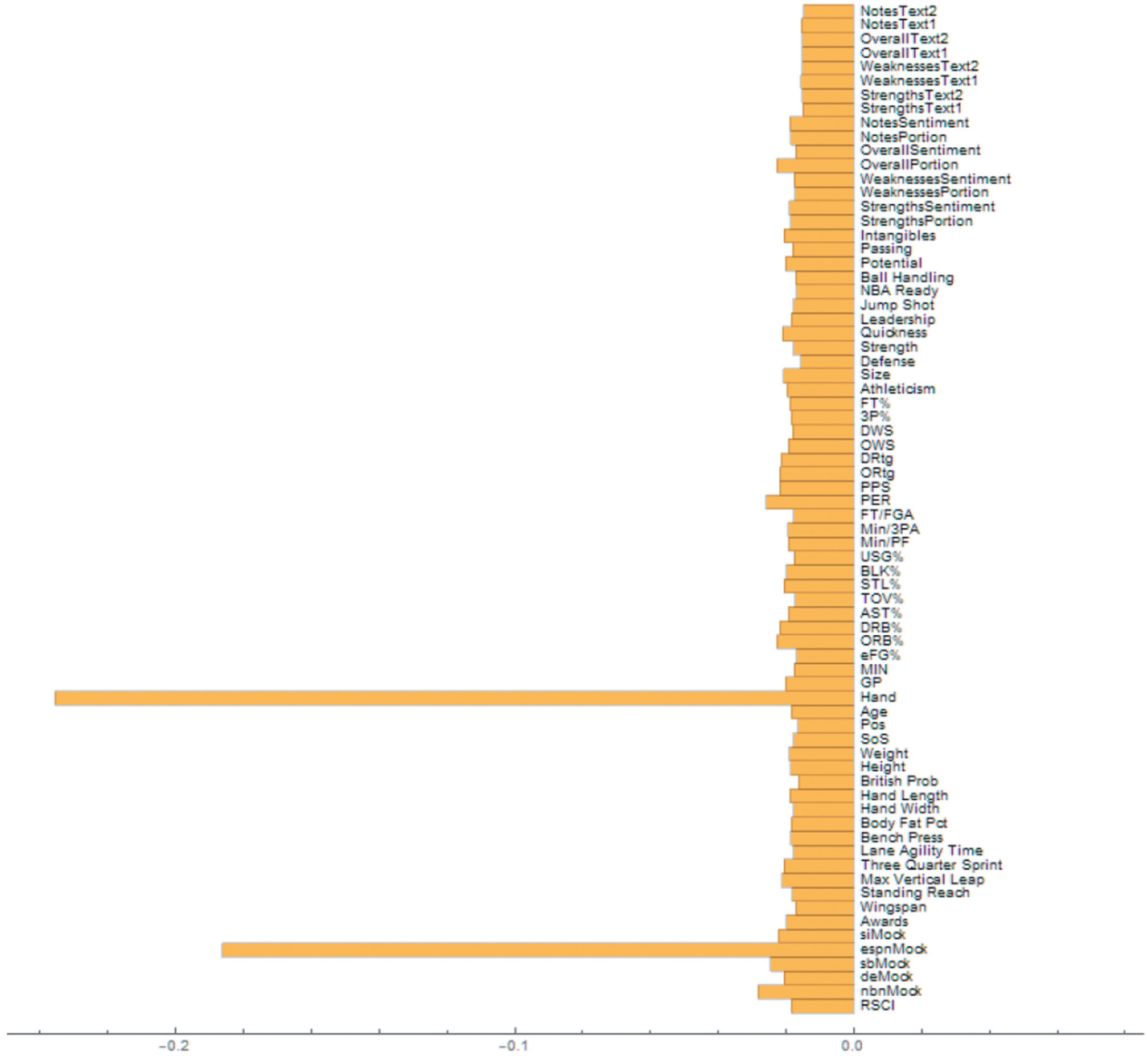


Figure 10. Feature Shuffling Bar Chart of Variable Importances.

Table 4. Change in Predicted vs. Actual Correlation for Changes in Handedness.

	To: Left	To: Right	To: Missing	To: Both
From: Left (26)	−17%	−23%	−30%	−32%
From: Right (277)	0%	0%	−10%	−9%
From: Missing (521)	−35%	−38%	−22%	−23%

also differ from the overall correlation. The left-handed subset has a 17% lower correlation while the missing-handed subset has a 22% lower correlation.

Left-handedness is somewhat more informative than right-handedness, when compared to missing data: when replaced with missing data, left-handedness correlation drops a further 13% below the 17% drop in that subset originally while right-handedness drops only 10%. Replacing missing data with left-handedness or right-handedness drops the correlation a further 13–15% below the 22% drop it already has from being a subset. Ambidextrousness appears to be the same as missing data.

We might also ask how the ESPN mock draft importance changed over time. Figure 11 shows the correlation between predicted and actual wins made for each year's draft, as well as the component due to the yearly ESPN feature shuffled evaluations. Shuffling the ESPN mock draft ratings each year consistently reduces the correlation, by as much as half in 2014. In other words, the ESPN mock draft is a consistently important variable in the draft model.

4.5. Comparison to Prior Models

The closest comparison model is Maymin (2017), which found similar value to automated projections as the present paper. However, that dataset covered data from 2003 to 2015 and did not include the raw text of scouting reports. One might imagine that, as with all progress, earlier years such as 2003–2006 offered greater rewards to automation than later years such as 2015–2019.

For example, Figure 12 shows the excess difference in average production by the top ten players projected by the model net of the average production of the top ten players actually drafted. (Note again that some of the actually drafted players were not available as choices to the model, such as international players.) Earlier years do indeed predominate; however, in this new model, the later years are promising as well, and of course the three-year results of 2018 and 2019 are yet to be decided: production almost always increases after the rookie year.

5. Conclusions

Training separate leave-one-out machine learning models for each NBA prospect using the raw text of scouting reports along with all the additional data and stats results in a model that outperforms the actual historical decision making of every single NBA team except one (Denver Nuggets), and by a substantial amount. NBA teams appear to be missing value in the draft by not following a disciplined, model-driven approach to decision making.

There are two possible approaches to the application of these results.

One, the more academic one, would be to continue digging into the model results and doing further dimension reduction to determine which subset of features contributes the most to these results. This approach would actually violate the main message of the results: human beings are not as good at judging future athletic production as machines.

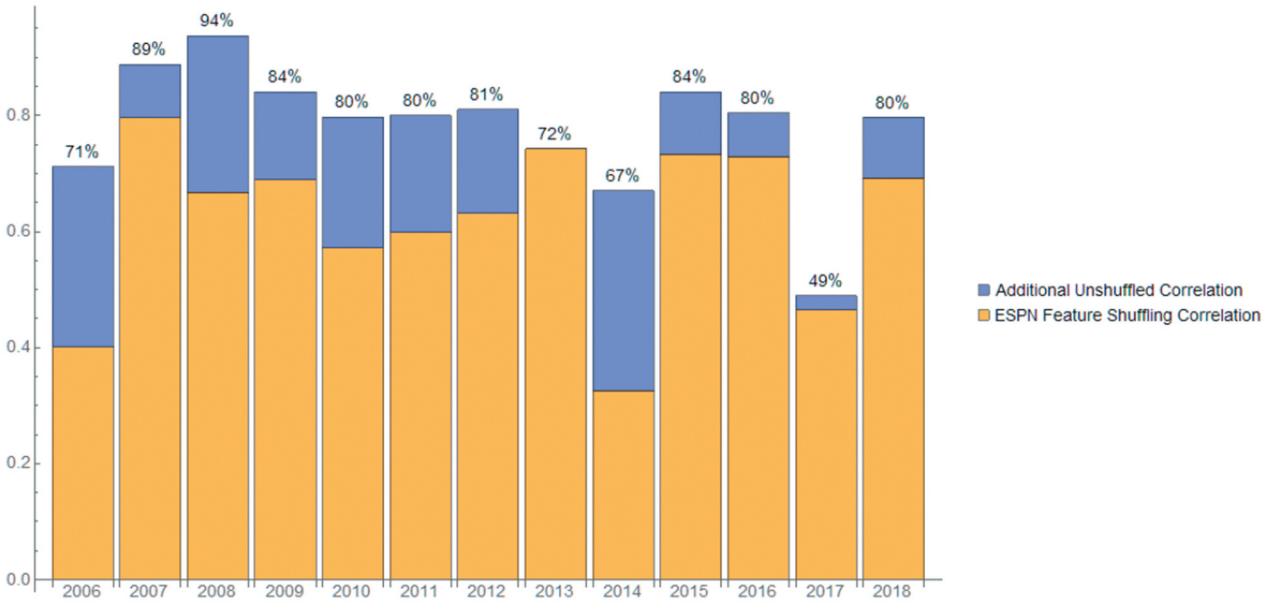


Figure 11. Yearly Predicted vs. Actual Correlation for ESPN Mock Draft Feature Shuffling.

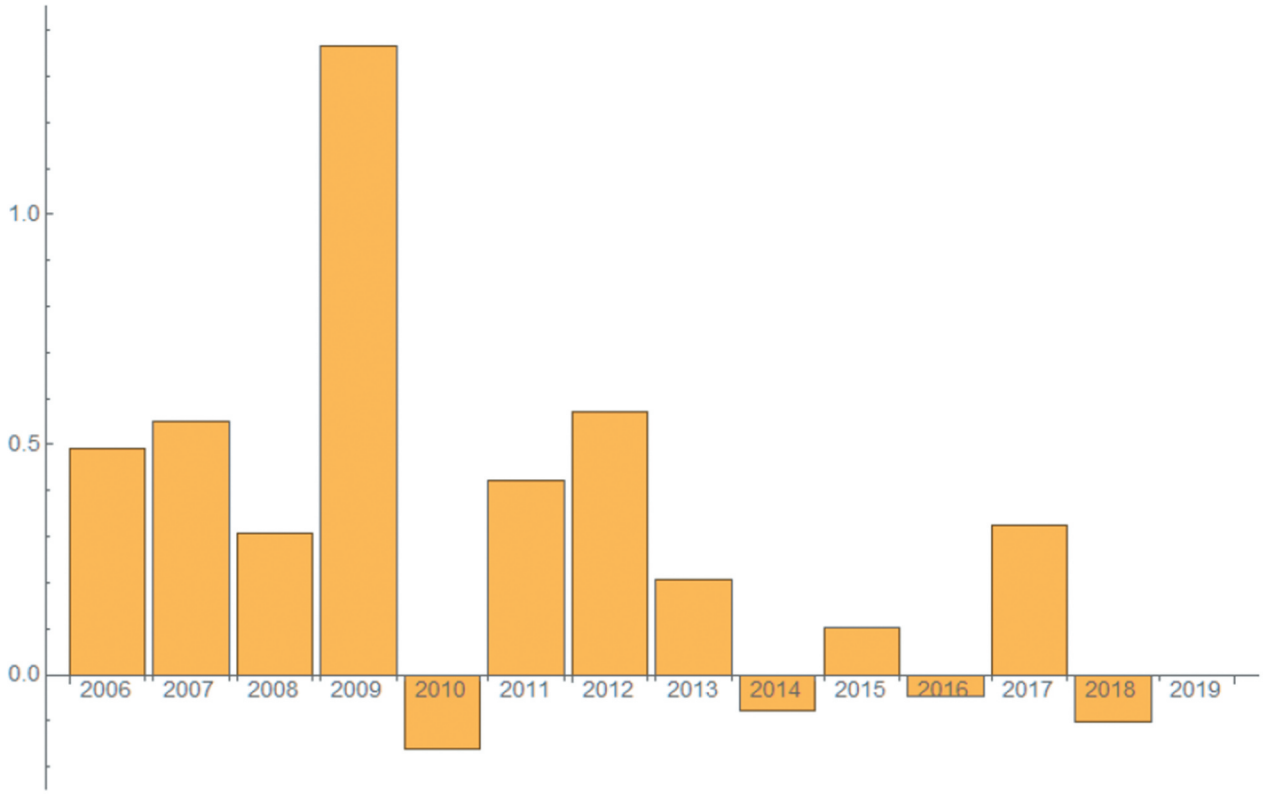


Figure 12. Excess Wins Made for Top Ten Model Choices vs. Top Ten Drafted Players.

The subtext of the academic approach is to create a simpler model for a human to evaluate and integrate into their decision making; but we know human decision making is flawed, particularly in such a psychologically entrapping field as sports, where one meets with prospects, friends, parents, extended family, and more, and behavioural and perhaps irrational patterns of thinking are the norm more than the exception.

The other possible approach to the application of these results is the practical one. It is a "kitchen sink" approach of continuing to add variables to the analysis and continuing to improve the machine prediction. Thus, rather than looking to trim the feature space, this approach would add to it instead.

Prospects undergo a battery of psychological and other testing that remains confidential but is in the possession of the teams. In addition to formal tests, informal assessments of character form another possible data source to add to the mix. How did the prospect perform during his visit to the team's training facility? What are their sleeping patterns like? What measures of intelligence or academic achievement can we include?

The methods used here are limited by evaluating only collegiate prospects, by supposing that the game itself, while constantly evolving, remained sufficiently similar over the sample period to justify comparisons across years, and by presuming that professional production can be measured with any of the metrics used. Additional methods and analysis would be required for international or other non-collegiate players. Careful adjustment would be required if substantial rule changes drastically alter how prior athletes would have performed in the new environment. And alternative measures of productivity, if developed, would need to be used as the predictive variable in question in the framework.

Potential future research could investigate optimal combinations of human and computer-based decision making. Thus, the tools here could be used not only as an automated decision maker, but also as a decision support tool. Care should be maintained though to prevent human decision makers from merely overriding the model on a whim; rather, the human contribution should also be tracked and evaluated to inspect it both for possible biases and hidden talent and value.

In applications to areas outside of sports, it is useful to remember how difficult it is for team owners and general managers to rely strongly on an analytical approach to decision making. It results in a feeling of a loss of control, and that feeling causes the humans to generate criticisms of any analytical approach at all. These standard criticisms are: overfitting, complexity, riskiness, and incompleteness. We have shown here that these criticisms can be overcome in draft decision making, yet, most likely, most teams will still continue to refuse to adopt an analytical approach. It is useful to bear their example in

mind as we evaluate analytical approaches in areas outside of sports: are the reasons for eschewing analytics themselves sound, or are they merely excuses like with sports teams?

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Philip Maymin http://orcid.org/0000-0002-3926-8720

References

Berri, D. J. (2011). "*Wins Produced Comes Back Better and Stronger!*" WagesOfWins.com. Retrieved November 29, 2019 from http://wagesofwins.com/2011/12/11/wins-produced-comes-back-better-and-stronger.

Berri, D. J., Brook, S. L., & Fenn, A. J. (2011). From college to the pros: Predicting the NBA amateur player draft. *Journal of Productivity Analysis*, *35*(1), 25–35. https://doi.org/10.1007/s11123-010-0187-x

Elder, J. (2009). "Target Shuffling to Estimate Baseline Performance.". In R. Nisbet, J. Elder, & G. Miner (Eds.), *Handbook of Statistical Analysis and Data Mining Applications* (1st ed., pp. 297–300). Academic Press.

Elder, J. (2014). *"Evaluate the Validity of Your Discovery with Target Shuffling."*. Elder Research: White Paper.

Evans, B. A. (2018). From college to the NBA: What determines a player's success and what characteristics are NBA franchises overlooking? *Applied Economics Letters*, *25*(5), 300–304. https://doi.org/10.1080/13504851.2017.1319551

Harris, J., & Berri, D. J. (2015). Predicting the WNBA draft: What matters most from college performance? *International Journal of Sport Finance*, *10*(4), 299–309. https://www.researchgate.net/publication/283865124_Predicting_the_WNBA_Draft_What_Matters_Most_from_College_Performance

Hollinger, J. (2019). *"Hollinger NBA Stats."* ESPN. Retrieved on November 29, 2019 from http://insider.espn.com/nba/hollinger/statistics.

Jain, S., & Kaur, H. (2017). "Machine learning approaches to predict basketball game outcome." *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall)*, 1–7, Dehradun, India.

James, B. (2010). *"The New Bill James Historical Baseball abstract."*. Free Press.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28* (1), 11–21. https://doi.org/10.1108/eb026526

Kannan, A., Kolovich, B., Lawrence, B., & Rafiqi, S. (2018). Predicting National Basketball Association Success: A Machine Learning Approach. *SMU Data Science Review*, *1* (3). Article 7. https://scholar.smu.edu/datasciencereview/vol1/iss3/7/

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-Normalizing Neural Networks. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 972–981. https://dl.acm.org/doi/10.5555/3294771.3294864

Kubatko, J. (2019). "*NBA Win Shares*." *Basketball Reference*. Retrieved on November 29, 2019 from https://www.basketball-reference.com/about/ws.html.

Laohaprapanon, S., & Sood, G. (2019). *"Ethnicolr: Predict Race and Ethnicity From Name."* Retrieved on June 8, 2019 and November 28, 2019 from https://github.com/appeler/ethnicolr.

Lopez, M. J., & Matthews, G. J. (2015). Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, *11*(1), 5–12. https://doi.org/10.1515/jqas-2014-0058

Maymin, P. Z. (2017). The Automated General Manager: Can an Algorithmic System for Drafts, Trades, and Free Agency Outperform Human Front Offices? *Journal of Global Sport Management*, *2*(4), 234–249. https://doi.org/10.1080/24704067.2017.1389248

Miller, A., Bornn, L., Adams, R., & Goldsberry, K. (2014). "Factorized point process intensities: A spatial analysis of professional basketball." *International Conference on Machine Learning*, 235–243. http://proceedings.mlr.press/v32/miller14.html

Mondello, M., & Kamke, C. (2014). The Introduction and Application of Sports Analytics in Professional Sport Organizations. *Journal of Applied Sport Management*, *6*(2), 2. https://js.sagamorepub.com/jasm/article/view/4035

National Basketball Association and National Basketball Players Association (2017). *"Collective Bargaining Agreement."* Retrieved on December 7, 2020 from https://nbpa.com/cba.

Oliver, D. (2004). *"Basketball on Paper."*. Potomac Books.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Research, N. C. A. A. (2019). *"Men's Basketball: Probability of competing beyond high school."* Retrieved on December 7, 2020 from http://www.ncaa.org/about/resources/research/mens-basketball-probability-competing-beyond-high-school

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Shmueli, S., & Koppius, K. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35* (3), 553–572. https://doi.org/10.2307/23042796

Wojnarowski, A., & Lowe, Z. (2020). *"NBA revenue for 2019-20 season dropped 10% to $8.3 billion, sources say"*. ESPN: Wolfram Research. https://espn.com/nba/story/_/id/30211678 on December 7, 2020.

Wolfram Language Documentation (2019). *"Sentiment: Built-in Classifier."* Retrieved on November 29, 2019 from https://reference.wolfram.com/language/ref/classifier/Sentiment.html.

Appendix

NBA Comparison: Gordon Hayward

Strengths: Middleton is a late blooming prospect with solid size for a SF at 6'7 210, and he's still growing into his body. . .Silky smooth operator on offense, and can score effectively at all 3 levels. . .Has a sweet jump shot. . .Soft touch, and is consistent out to about 22 feet and shows raw NBA range. . .Able to make shots off screens, spotting up or off the dribble. . .Functional in the triple threat. . .Likes to operate in the mid-range area, most comfortable from 15-18 feet, and utilizes a nice pull-up game and an interesting blend of runners in his arsenal. . .Middleton can stick the 3 ball with his feet set, and he has the release speed and form to get his shot off against contests effectively. . .Plays the game at his own pace, and has a very unique and unorthodox style of play. . .Very unselfish and won't go outside of the offense to get his own shots or production very often, which will benefit his eventual transition to the pros. . .Doesn't need to dominate the ball at all to make his presence felt on the court. . .Nice passing skills. . .Will contribute some on the glass. . .Appears to be very coachable, and he still has upside remaining as he get stronger and fills out his frame. . .

Weaknesses: Has had a lot of work to do to come close to regaining the momentum he had before the season as a prospect. . .Had an ankle injury that nagged him all of his Jr. season was never 100% and comfortable under the new Texas A&M coaching staff. . . Didn't really show any signs of improvement in any aspects of his game. . .Middleton right now needs to work on getting stronger to maximize his athleticism and ability to withstand tough defense. . .There are times when he can get pushed around on both ends by physical play. . .Middleton also would be well-served to continue to work on his ball-handling some more, were he's currently mediocre. . .He does sporadically make moves with the ball that show his potential, but he's not consistent with it yet and he's not a guy who operates all that effectively in pure isolation sets. . .Doesn't handle double-teams or extra defensive attention all that well. . .Can struggle to finish around the rim at times due to a lack of explosiveness and strength. . .Could stand to get more aggressive and play with more urgency. . .Middleton is a middle-of-the-road defensive player, and he can struggle with on-ball defense at times. . .A bit upright in his stance and middling lateral quickness. . .

Overall: Middleton's late blooming status and promising Soph. season making him a prospect worth keeping an eye on. . .His lackluster Jr. season, despite injury, has dropped his stock from a possible lottery selection to a 2nd rounder. . . He's a good shooter with a smooth, unselfish floor game and the tools that project well to the next level as a complementary piece. . .He isn't the best defender, but as he grows into his body he can be an even better player and prospect, as it looks like he hasn't peaked physically or athletically yet. . .Still has upside as a prospect and could be a big value pick for a team

Notes: Measured 6'7.5 (in shoes) 211 lbs, with a 6'10 wingspan at the 2011 Kevin Durant Skills Academy.