International Journal of Forecasting 35 (2019) 776–782

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Wage against the machine: A generalized deep-learning market test of dataset value

Philip Z. Maymin *

Vantage Sports, United States
University of Bridgeport, Mandeville Hall 217b, 126 Park Avenue, Bridgeport, CT 06604, United States

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Sports forecasting
Gambling
Wagering
Data
Analytics

ABSTRACT

How can you tell whether a particular sports dataset really adds value, particularly with regard to betting effectiveness? The method introduced in this paper provides a way for any analyst in almost any sport to attempt to determine the additional value of almost any dataset. It relies on the use of deep learning, comprehensive historical box score statistics, and the existence of betting markets. When the method is applied as an illustration to a novel dataset for the NBA, it is shown to provide more information than regular box score statistics alone, and appears to generate above-breakeven wagering profits.

1. Introduction

How can you tell whether a particular sports dataset really adds value?

This is a new concern. Until recently, there were so few datasets that anything different almost always added value. In the past few years, though, so many new datasets have emerged across all major sports—including data derived from optical tracking, body sensors, computer vision, and GPS and RFID location systems (see Barlow, 2015)—that it is no longer clear whether the new datasets make any marginal contribution at all relative to what we already had before. However, we do not have good analytics for deciding which datasets add enough value to warrant further investment and which do not. Our industry's earlier thirst for data has been quenched and we are now at risk of drowning.

There are several difficulties in deciding whether an additional piece of data adds value to an existing corpus of knowledge, because the important issue for practitioners is not the data itself but the insights available from it. One difficulty is *consistency*: if you ask one genius to extract all possible insights from dataset X, and another genius to extract all possible insights from datasets $X + Y$, the first genius may be smarter or luckier or both, and get more insights from less data, in which case we would erroneously conclude that dataset Y is not necessary; or the second genius might get more insights, but have obtained those insights from X as well. Another difficulty is *congruity*: one dataset might be raw video footage while another is textual scouting reports; the processes by which insights are extracted are likely to differ substantially between the two, thus adding another layer of potential noise. The third difficulty is *comparability*: if the two geniuses come up with different insights, how can we decide which are more important, or whether they complement each other?

These issues apply to all questions of dataset evaluation. In many sports, though, we are blessed with one recent machine learning innovation and two natural phenomena that we can harness to answer all three difficulties.

To address consistency, we will use a deep-learning algorithm to extract insights automatically from both the original and augmented datasets. This ensures that an equal amount of machine intelligence is applied to both. Deep learning is a term for artificial hierarchical neural networks that have proven recently to be remarkably robust

* Correspondence to: University of Bridgeport, Mandeville Hall 217b, 126 Park Avenue, Bridgeport, CT 06604, United States.
E-mail address: philip@maymin.com.

https://doi.org/10.1016/j.ijforecast.2017.09.008

and effective algorithms in various domains; see Schmidhuber (2015) for an overview and survey of their numerous victories in pattern recognition and machine learning. Roughly speaking, deep learning differs from other machine learning techniques in that it seems to be the best at mimicking the human mind for learning complex hierarchical patterns from past examples, and it has set many modern records, such as beating humans in the game of Go, image recognition, automatic captioning, and more.

To address congruity, we will use quantitative summary statistics drawn from the datasets, so that we are essentially comparing one enhanced box score with another. This puts the datasets on an equal footing. One of the advantages of deep learning is the ability to use large numbers of factors, meaning that we do not need to restrict the number of columns from either source, but can instead use essentially all available information from both.

To address comparability, we rely on a convenient and beautiful natural phenomenon in sports: the existence of robust and healthy betting markets. This is the primary distinguishing characteristic of sports datasets that allows us to use the approach presented here; for example, there is no known predictive market for the evaluation of medical datasets. Even in sports, if the new data cannot help you make more money than the old data could, it is possible that they might still be useful in an explanatory or other role; but if the new data *can* improve predictability in sports markets, then we know *for sure* that they have significantly and substantially more value than the old.

1.1. Novelty of research

The issue of evaluating datasets in a sea of available choices is a novel one, as is the solution presented here. Of course, research into the evaluation of which of several machine learning models is best has been done; Fawcett (2006) provides a recent introduction to a standard approach. Research into deep learning is also growing rapidly; see Schmidhuber (2015) for a recent overview, as noted above.

Here, though, we fix the machine learning algorithm to be deep learning, and instead vary the datasets. Furthermore, we take the practitioner's viewpoint by using an established dataset as the base and augmenting it with new data to test whether the marginal contribution is significant or not. Finally, we compare the result with the betting markets to see whether or not the new data does a better job of predicting outcomes. Deep learning was chosen because of its broad success in many areas, as noted above.

1.2. Academic rigor/validity of the model

We ensure the model's validity by using a standard deep learning algorithm applied to historical data that has not been exposed to betting markets to evaluate the performance in future wagering. Further, we roll the model forward on a daily basis, avoiding lookahead bias and maintaining a strict out-of-sample test. Finally, the same model is applied to previously unseen results, namely the 2015–2016 National Basketball Association (NBA) season, and the results continue to be substantially and significantly above break-even, without any modification to the model. Thus, the model passes the ultimate test of model validity.

1.3. Reproducibility

Everything shown in this paper is reproducible. The data on betting markets are easily available through a range of sources; the NBA's boxscore and similar data are available through their website; the deep learning algorithm uses the free open-source h2o library; and the augmented data are routinely made available both to researchers and to writers (see Csapo & Raab, 2014). Finally, because the data are objective and well-defined, they could, in principle, be re-collected from video footage by anyone.

1.4. Application and interest/impact

The particular application in this paper is to the NBA. Extensions to other professional basketball leagues around the world, or to college basketball, would be straightforward. Extensions to other sports would take longer since one must first develop the augmented dataset, but, in principle, there is no obstacle.

Further, in addition to evaluating the dataset considered here, the approach is viable for *any* such question on *any* dataset. The only requirements are that the old and new datasets be in the same form (i.e., quantitative columns of information), and that there exist market forecasting results that the data could help predict. Note however that, even with this approach, it would still be possible for a particularly subtle pattern or value of the dataset to remain undetected.

Thus, the approach presented here has an impact for virtually all modern and popular sports.

2. Data

Datasets need to be combined with intelligence in order for actionable value to be derived. The novel method proposed here involves the standardization of intelligence across datasets by using deep learning, a machine learning algorithm that mimics human intelligence by using high-level hierarchical abstractions and structures. Deep learning is used to try to beat historical sports wagering lines. If the original dataset does not beat the market lines but the augmented dataset does, then the additional data conclusively add value.

The specific dataset used here is from Vantage Sports, where highly trained human analysts tabulate dozens of unique metrics for every NBA game, including whether a hand was up on defense for each field goal attempt, whether a screen was used or rejected, solid or not solid, split or not split, and more. See Table 1 for a comparison of this dataset with the boxscore and optical datasets.

The original dataset is all publicly available NBA data, including boxscore and optical data. The augmented dataset adds the Vantage data as well. The Vegas lines used are the closing lines, which are the hardest to beat. Note that, although injuries are not included in any of the data sets, they are certainly important, and a clean injury dataset would probably improve the results further.

In terms of typical file sizes, rows, and numbers of data points, all on a per-game basis, boxscore and play-by-play

Table 1
Data vs. information vs. knowledge for various basketball datasets.

Data source	File size	Rows	Data points	Information	Knowledge
Boxscores and play-by-play	25k	∼100	∼700	Medium	Medium
Optical data	40,000k	∼100,000	∼2,000,000	Low	Medium
Vantage data	500k	∼3000	∼16,000	High	High

data have the lowest values and optical data the highest, while the Vantage dataset is in the middle.

As the boxscore and play-by-play data include some basic information about every possession, the file is typically about 100 rows, with about 700 data points.

The optical dataset includes two-dimensional court coordinates for all ten players and three-dimensional coordinates for the ball, both at 25 frames per second. However, since not all players are tracked at all times (for example, during free throws), the overall number of coordinate information rows is usually less than the theoretical maximum of (a) 2 coordinates $\times$ 10 players tracked $\times$ 25 frames per second $\times$ 60 s $\times$ 48 min plus (b) 3 coordinates for the ball $\times$ 25 frames per second $\times$ 60 s $\times$ 48 min, which is about one and a half million.

The Vantage dataset has fewer data points than optical but more information, because it reports not merely location data that needs to be processed into basketball intelligence, but the actionable information itself: was there a screen, was it used, was there a cut, was there a closeout opportunity, did the closeout happen, did the defender keep his player in front of him, etc. Both the process and the output of Vantage data are discussed in more detail below.

The differences between data, information, and knowledge have long been recognized in the field of knowledge management and information science; see Zins (2007) for a conceptual review and multiple definitions of what they call "these three key concepts". The three concepts are often visualized as a pyramid, with data on the bottom, information above it, and knowledge above that, implying that information adds meaning to the arbitrariness of data, and knowledge provides context. A fourth triangle, wisdom, is sometimes added above knowledge to indicate proper decision-making given the knowledge below.

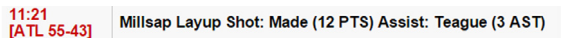
The data points are the raw numbers coming from the three sources, and for sheer volume, optical has the most and play-by-play the fewest, while Vantage is in the middle. However, the optical dataset contains the least information, because the information embedded in where a particular player was standing at a given time is very small, especially when compared with the virtually identical numbers coming both before and after. In contrast, the boxscore and play-by-play datasets contain more information: knowing that a particular player scored or assisted in a basket is a small amount of data but a larger amount of information. The Vantage data contains still more information, as it includes not only who took the shot and who assisted, but also who defended, where the shot was taken, what kind of shot it was (e.g., turnaround, layup, fadeaway, hook shot, floater, etc.), and the nature of the defense (was a hand up, was the shot pressured, etc.).

Knowledge can be viewed as the insights extracted from the meaningful information. There is indeed some knowledge to be extracted from the boxscore and play-by-play statistics, and much valuable work over the past several decades has focused on just that, using a range of metrics from boxscore metrics such as wins produced (Berri, Schmidt, & Brook, 2007) to various forms of plus-minus metrics (Engelmann, 2015). Some knowledge can be extracted from the optical data as well; see for example McQueen et al.'s (2014) attempt to use machine learning to recognize on-ball screens from the optical data.

Note that any knowledge that might possibly be extracted from the optical data is only a subset of the information that is available from the Vantage dataset directly. The information available from Vantage is greatest of all, because it includes actionable basketball facts with embedded meaning. Furthermore, the additional knowledge that can be extracted from that large amount of information is itself large, since adding context can help in ranking players and teams based on the important metrics, evaluating performances, aiding development, and searching for underrated players, among other things.

A two-second clip exemplifies the differences among these three data sources. The clip, annotated with Vantage Sports data and described in interviews, is available from Abbott (2015).

At the start of the second half of the March 6, 2015, game against the visiting Cleveland Cavaliers, Jeff Teague of the Atlanta Hawks started a play that eventually led to a layup by Paul Millsap. The complete description of those two seconds in the official play-by-play reads as follows:



This description contains four pieces of data: the scorer, the passer, the shot outcome, and the shot type. These are also pieces of information. In contrast, the corresponding optical data has about 50 pairs of location data points for each of the ten players plus 50 triplets of location data points for the ball, a total of about 650 data points. However, none of that data is information. The corresponding Vantage data points are shown in Table 2, consisting of 53 data points, all of which are also pieces of information.

2.1. NBA data

The NBA dataset is sourced from the nba.com website, as along with some commonly calculated additional information such as scheduling (back-to-back indicators, rest days, etc.). Overall, about 50 metrics per game comprise

Table 2
Vantage Sports data for the Jeff Teague assist to Paul Millsap play.

Jeff Teague: On-ball screen received
Location: Top key (3pt)
Outcome of received on-ball screen: Assist
Split?: Yes
Use of screen: Use screen
Kyrie irving: Screen for offensive player
On-ball screen defender response: Switch
On-ball screen keep-in-front: Not applicable
Location: Top key (3pt)
Screen outcome: Assist
Timofey Mozgov: On-ball screen by offensive player
Defender response to on-ball screen: Hedge or hard show
Keep-in-Front: No Keep-in-Front
Location: Top key (3pt)
Outcome of a no Keep-in-Front: Straight to basket
Screen outcome: Assist
Secondary on-ball screen defender response: Double-Team
Split?: Yes
Jeff Teague: Drive
Drive left: Drive left
Starting location: Right wing (3pt)
Ending location: High post
Al Horford: On-ball screen set
Location: Top key (3pt)
Outcome of set on-ball screen: Assist
Outcome for screen setter: Roll
Receives ball?: No
Screen effort: Reroute defender
Kevin love: Help attempt
Help/Double-Team outcome: Assist
Keep-in-Front during help attempt: Keep-in-Front
Location: Center low post
Kevin love: FG attempt against
Closeout situation: No
Location: Right low post
Made shot: Yes
Shot clock: Shot clock at 5
Shot defense: Pressured
Shooter: Paul Millsap
Paul Millsap: Pre-acquisition action
Move: Post up
Paul Millsap: Shot attempt
Backboard: Yes
Defender (1): Kevin love
Defender (2): Timofey Mozgov
Dribbles: 0 Dribbles
Location: Center low post
Post-Acquisition location: Center low post
Pre-Acquisition location: Left hash
Made shot: Yes
Release: Reverse layup
Shot clock: Shot clock at 5
Shot defense: Pressured

the "standard" dataset, prior to augmenting. Due to space concerns, Table 3 lists the standard abbreviations used in the dataset. The exact definitions are easy to work out, but can be found on their website in case of uncertainty. The data collected are for the 2014–2015 NBA season.

Note that the standard metrics include certain metrics that are derived from the optical data and are made available on a per-game basis. These include items such as the total distance run, total touches, secondary assists, passes, and contested, uncontested and defended at the rim field goal attempts and makes. However, it should be noted that these definitions of contested and defended are based purely on proximity, and do not distinguish between defenders that are contesting a shot actively with a hand up and those who just happen to be nearby.

2.2. Vantage Sports data

Vantage Sports captures data from broadcast footage using a large team of fully trained full-time employees. The analysts tag every tracked event for every player in the game, after which the tags are cross-referenced and cross-validated to ensure their validity.

The tags that are tracked by human eyes are intended to represent the critical pieces of actionable basketball intelligence that a coach, player, or general manager would want to know about a game.

One example of this is contested shots. It is common knowledge that having a hand up is the key to a good shot defense (see Csapo & Raab, 2014), but no previous data source has made that information available: it is not in any boxscore, play-by-play, or optical database. However, Vantage Sports has this data for every shot attempt, by every player, on every team, in every game.

As another example, Vantage tracks whether a pass was made to an open shot, regardless of whether or not the shooter made it, because the passer should be rewarded for making the correct pass regardless of the bounce of the ball. Vantage also tracks active pressure (meaning actively moving hands, not just proximity) on the perimeter, on sidelines, and on inbounds passes; rebounding efforts and opportunities; screen offense and defense—did they hedge? did they do a hard show? did the ballhandler split the screen? and other subtags; close-out opportunities; cuts; etc. Table 4 lists a representative sample of the metrics in the augmented dataset. The metrics are spelled out here because they are unique. Further information is available on the Vantage Sports enterprise website. The data in this augmented dataset are also for the 2014–2015 season.

2.3. Betting markets

Historical betting market data for the 2014–2015 NBA season can be sourced from a range of websites, e.g. vegasinsiders.com. It is important to note that only closing lines are used in historical testing. These are widely considered to be the hardest lines to beat, as they represent the market's best and final forecast; see for instance the paper by Dare, Dennis, and Paul (2015), which shows that opening NBA lines contain substantial biases when a high-quality player is absent, but that all of the biases are eventually removed so that the closing line is a fair 50–50 bet.

There are two kinds of standard bets: spread bets and over/unders. Spread bets predict that one of the teams will win by a certain minimum margin of points. Over/unders predict that the total points scored by the two teams combined will exceed some threshold. More complex bets are

Table 3
Metrics in the standard dataset.

Scheduling: Number of games in the last five days, back-to-back, rest days, list of referees
Regular: PTS_OFF_TOV,PTS,PTS_PAINT,PTS_2ND_CHANCE,PTS_FB,LARGEST_LEAD, FTA_RATE,TM_TOV_PCT,OREB_PCT,BLK,PF,FTA,FTM,STL,TO,PLUS_MINUS
Portions: PCT_PTS_2PT,PCT_PTS_2PT_MR,PCT_PTS_3PT,PCT_PTS_FB,PCT_PTS_FT, PCT_PTS_OFF_TOV,PCT_PTS_PAINT,PCT_AST_2PM,PCT_UAST_2PM, PCT_AST_3PM,PCT_UAST_3PM,PCT_AST_FGM,PCT_UAST_FGM
Ratios: OFF_RATING,PIE,CFG_PCT,UFG_PCT,FG_PCT,DFG_PCT,EFG_PCT,FG3_PCT, FT_PCT,PCT_FGA_2PT,PCT_FGA_3PT,TS_PCT,AST_PCT,AST_RATIO
Optical: DIST,ORBC,DRBC,TCHS,SAST,FTAST,PASS,AST,CFGM,CFGA,UFGM,UFGA, DFGM,DFGA

Table 4
Metrics in the augmented dataset.

Scoring: Contest+ FG%, Open+ Freq., Open+ FG%, Points per Chance, …
Shot defense: Block-to-Possession, Contest+, Points against per shot, Shots per chance, …
Movement and involvement: Offensive activity rate, Cut efficiency, Touches per chance, …
Turnovers, Disruptions, Fouling: In-Air TO%, Unforced TO%, Effective bump%, Front post D%, Pressure rate, …
Passing: Indirect pass rate, Assist+ screen%, Deflected-pass rate, True facilitation, …
Rebounding: O/DBlockouts per 100 Opps, O/DReb pursuit rate, …
Screening offense and defense: Screens received/Set per chance, Split%, Solid screen%, KIF%, Hedge%, …

also available, but for simplicity, the model is only ever trained on these two, the most standard bets.

Note also that the standard betting cost is assumed throughout this paper: losses cost 10% more than wins pay. For example, a \$110 bet for the over/under to exceed 200 points will result in a \$100 gain paid to you if the total is 201 or greater, a \$110 loss paid by you if the total is 199 or lower, and the return of your original bet if the total is exactly 200. (This tie situation is called a "push", and you are placed in the same economic position as if your bet had never been placed at all.) Thus, the breakeven probability is $11/(11+10) = 52.381\%$.

(Finally, note that many sportsbooks have recently begun offering even better odds of \$105 on losses to \$100 on wins, leading to a breakeven probability of only $10.5/(10.5+10) = 51.2195\%$.)

3. Methods

Conceptually, the comparison for each dataset is done as follows.

We begin with the 50th day of the season in order to have a base of data from which to start, and create the following training table each day. The columns indicate the date of the game, the two teams, the market betting lines, and a 20-game moving average of the (standard or augmented) metrics for each team. We run two deep learning algorithms for each dataset: one for predicting the actual resulting spreads, and one for predicting the actual resulting over/unders.

We find the best fitting model using a deep learning algorithm with the standard parameters of h2o (Candel, 2015). Taking that best fitting model, we then apply it to that day's games to establish a prediction. The following day, we extend the training set by one day, re-train, and re-predict. In both cases, predictions that exceed equiprobability significantly with statistical confidence at the 5% level are assumed to be placed bets, while those that do not are assumed to be skipped bets. This allows algorithms to decide not only whether to bet over or under, but also whether to bet at all. Again, the same procedure is followed for both the standard dataset and the augmented dataset. We then conglomerate the two strategies within each dataset (spreads and over/unders) in order to compare the two time series. In other words, for either dataset, given a prediction on a bet, we test to see whether the probability of winning differs statistically significantly from 50%. If it does, it is a bet in that direction; otherwise, no bet.

4. Results

In using this approach to evaluate the marginal value of an additional dataset (namely parallel deep learning on the two datasets, each attempting to outpredict the market), there are four possible results.

First, it is possible that neither the original nor augmented datasets can beat the breakeven probability of 52.381%. In this case, it would be difficult to decide whether the augmented data are marginally valuable, since even a random coin toss can achieve a 50% probability. This is generally the most likely scenario, as it is actually quite difficult to beat the markets. Thus, we would not be able to simply conclude that the data are more valuable.

Second, it is possible that both the original and augmented datasets can beat the breakeven probability. This is the least likely scenario, as selection bias suggests that you are unlikely to be searching for augmented data if you are already able to beat the breakeven probability. Nevertheless, it would still be possible to run a statistical test in this case in order to determine whether the marginal contribution makes it worthwhile to change the predictive algorithm and adopt the augmented dataset.

Third, it is possible that the original dataset exceeds the breakeven probability but the augmented dataset does not. This is quite unlikely in a general deep learning setting, as the algorithm does not typically perform worse with more inputs. If this is the result, it is most likely to be due to a random perturbation (or an error in the machine learning algorithm) rather than a significant decline.

Fourth, it is possible that the original dataset falls short of the breakeven probability but the augmented dataset exceeds it. This is the most exciting scenario, as we can then determine conclusively that the augmented dataset does provide more value.

The results from the data above are as follows. There were slightly under 1000 possible wagers for each type of bet for each of the datasets over the sample period. Each wager involved risking \$110 to win \$100. Among spread bets, 257 won, 223 lost, and 372 were not bet, for a winning percentage of 53.54%. Among over/unders, 250 won, 210 lost, and 392 were not bet, for a winning percentage of 54.35%.

Starting with an assumed initial bankroll of \$5000, the daily rolling deep learning algorithm using the standard dataset is correct on 49% of wagers and ends the season with \$1700. This means that the standard dataset combined with deep learning is unable to do better than a coin toss. (Indeed, the 49% is not statistically significantly different from 50%.) This should not be surprising, as the betting markets are quite efficient, and we should expect them to incorporate all standard publicly available information.

Using Vantage data, the algorithm is correct on 54% of wagers and ends the season with \$6500. The difference is highly statistically significant (p-value < 0.01), and exceeds the breakeven probability. See Fig. 1 for a time series graph of the two strategies combined.

The same comparison can be made across only the spread trades or only the over/unders, and a similar picture emerges. The augmented dataset improves each of those kinds of bets, relative to the standard dataset.

By construction, all of these results are out-of-sample.

Furthermore, the model was also tested in "paper" trading for the start of the 2015–2016 season, making it entirely out of sample. Without any modification to the deep learning model for the augmented dataset, but simply continuing as if the new season were an extension of the old, the model continued to outperform the breakeven probabilities, achieving a reported winning probability of 56.65% for bets on spreads and over/unders for the period until December 13, 2015.

5. Conclusion

The method introduced in this paper provides a way for any analyst in almost any sport to determine the additional value of almost any dataset.

The method is uniquely suited for sports analytics because it requires both justifiable datasets and an associated, liquid wagering market that is likely to have pockets of inefficiency. It would not apply to random data—deep learning can't predict the next coin toss—and it probably would not apply to financial events where the markets are likely to be very efficient. In the world of sports analytics, though, the method outlined here can be used to have a substantive and permanent impact on any sport that has a healthy wagering market associated with it.

When we apply the method to NBA betting markets with both a standard, publicly-available dataset and an augmented one that incorporates data from Vantage Sports, we find that a rolling deep learning model with the augmented data substantially and significantly outperforms a similar machine learning model with the standard data over the 2014–2015 season. Furthermore, the performance when using the augmented data is above the betting breakeven probability.

Finally, the same model without modifications continues to outperform others in subsequent markets and games, yielding a winning probability that is in excess of 56% for bets on games between the start of the season on October 27, 2015, until December 13, 2015.

Such a result is unprecedented and remarkable for two reasons: first, an efficient market is notoriously difficult to beat; and second, the result provides a conclusive demonstration of the power of the approach presented in this paper. It also provides a new type of framework for investigating the profitability of a betting scheme.

Extensions to other professional or collegiate leagues would be straightforward. In addition, extensions to other professional or collegiate sports could also be pursued. Finally, one caveat to remember, as discussed above, is that it is possible that there could be value in the augmented dataset but that the approach here might fail to find it; since deep learning does seem to learn in the same way as a human, it would be possible, as with a human, for it to fail to find an existing pattern.

Acknowledgments

I am grateful for the contributions of each member of the Vantage Sports team, especially Brett McDonald, Chase Exon, Cameron Tangney, Mark Jansen, Scott Snider, and Brandon Sedgwick, and the many dozens of analysts, and a special additional thank you to Chase Exon for the catchy title.

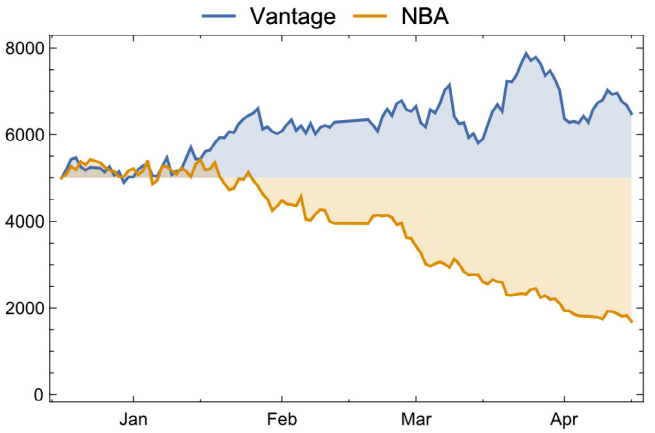


Fig. 1. Cumulative balance of deep learning strategies using the standard (yellow) and augmented (blue) datasets for 2014–2015. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

Abbott, H. (2015). Jeff Teague on art of the playmaker. ESPN TrueHoop, May 5, available online at http://espn.go.com/blog/truehoop/post/_/id/73055/jeff-teague-on-art-and-science-of-the-playmaker, (retrieved on 13-12-15).

Barlow, J. M. (2015). *Data analytics in sports*. USA: O'Reilly Media.

Berri, D., Schmidt, M., & Brook, S. (2007). The wages of wins: taking measure of the many myths in modern sport (updated edition). Stanford Business Books.

Candel, A. (2015). The definitive performance tuning guide for H2O deep learning. February 26, retrieved from http://h2o.ai/blog/2015/02/deep-learning-performance/ (on 14-12-15).

Csapo, P., & Raab, M. (2014). Hand down, man down: analysis of defensive adjustments in response to the hot hand in basketball using novel defense metrics. *PLoS One*, *9*(12), e114184.

Dare, W. H., Dennis, S. A., & Paul, R. J. (2015). Player absence and betting lines in the NBA. *Finance Research Letters*, *13*, 130–136.

Engelmann, J. (2015). Estimating a player's influence on his teammates' BoxScore statistics using a modified adjusted +/- framework. Presented at the 2015 New England Symposium on Statistics in Sports, video available at https://youtu.be/OuC0YZTADcE.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

McQueen, A., Wiens, J., & Guttag, J. 2014. Automatically recognizing on-ball screens. presented at the 8th annual MIT Sloan Sports Analytics Conference, paper available at http://www-personal.umich.edu/~wiensj/papers/SSAC2014.pdf (retrieved on 13-12-15).

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, *61*, 85–117.

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the Association for Information Science and Technology*, *58*(4), 479–493.

Philip Z. Maymin is Associate Professor of Analytics and Finance at the University of Bridgeport Trefz School of Business. He is also the founding managing editor of Algorithmic Finance and the co-founder and co-editor-in-chief of the Journal of Sports Analytics. He has also been an analytics consultant with several NBA teams and is the Chief Analytics Officer for Vantage Sports. He holds a Ph.D. in Finance from the University of Chicago, a Master's in Applied Mathematics from Harvard University, and a Bachelor's in Computer Science from Harvard University. He also holds a J.D. and is an attorney-at-law admitted to practice in California. He has been a portfolio manager at Long-Term Capital Management, Ellington Management Group, and his own hedge fund, Maymin Capital Management. He has also been a policy scholar for a free market think tank, a Justice of the Peace, a Congressional candidate, an Assistant Professor of Finance and Risk Engineering at the NYU School of Engineering, and an award-winning journalist. He was a finalist for the 2010 Bastiat Prize for Online Journalism. He was awarded a Wolfram Innovator Award in 2015. He won the Wolfram Live Coding Challenge in 2016 and the Wolfram One-Liner Competition in 2015 and 2016.