Wage Against the Machine: A Generalized
Deep-Learning Market Test of Dataset Value

Philip Z. Maymin, Chief Analytics Officer, Vantage Sports
Associate Professor of Analytics and Finance, University of Bridgeport

1. Introduction

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

This is a new concern. Until recently, there were so few datasets, anything different almost always added value. But in the past few years, 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 (cf. [1])—that it is no longer clear if the new datasets have any marginal contribution at all relative to what we already had before. Yet 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.

Deciding if an additional piece of data adds value to an already existing corpus of knowledge has several difficulties, because what matters to 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, if the first genius is smarter or luckier or both, he may get more insights from less data, and we would erroneously conclude that dataset Y is not necessary; or the second genius might get more insights, but he would have gotten those insights from X as well. Another difficulty is congruity: one dataset might be raw video footage while another is textual scouting reports; the process by which insights are extracted are likely to differ substantially between the two, adding another layer of potential noise. And the third difficulty is comparability: if the two geniuses come up with different insights, how can we decide which ones are more important, or if they complement each other?

Those issues apply to all questions of dataset evaluation. But in many sports, 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 automatically extract insight from both the original and the augmented dataset. This ensures that an equal amount of machine intelligence is applied to both. Deep learning is a term for artificial hierarchical neural networks that recently has proven to be a remarkably robust and effective algorithm in various domains; c.f. [2] for an overview and survey of its numerous victories in pattern recognition and machine learning.

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 equal footing. One of the advantages of deep learning is the ability to use large numbers of factors, so 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 allow us to use the approach presented here. There is no known predictive market for evaluating medical datasets, for example. And even in sports, if the new data cannot help you make more money than the old data could, it is possible the data could still be useful in an explanatory or other role; but if the new data can improve predictability in sports markets, then we will know for sure that the new data has significantly and substantially more value than the old.

1.1. Novelty of research
The question of evaluating datasets in a sea of available choices is a novel one. The solution presented here is too. Research into evaluating which of several machine learning models is best has been done, of course; c.f. [3] for a recent introduction to a standard approach. And research into deep learning is hot and growing; c.f. [2] for a recent overview, as noted above.

But here we are fixing the machine learning algorithm as deep learning, and instead varying the datasets. Further, we take the practitioner’s viewpoint in using an established dataset as the base, and augmenting it with new data to test whether the marginal contribution is significant or not. And finally, we compare to the betting markets to see if the new data does a better job of predicting outcomes.

1.2. Academic rigor / validity of 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 on a daily basis, ensuring no lookahead bias and maintaining a strict out-of-sample test. Finally, the same model was applied to previously unseen results, namely the 2015-2016 National Basketball Association (NBA) season, and the results continued to be substantially and significantly above break-even, without any modification to the model. Thus the model passed the ultimate test of model validity.

1.3. Reproducibility
Everything shown in this paper is reproducible. The source of data for betting markets is easily available through multiple 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 is routinely made available both to researchers and to writers (c.f. [4]), and because the data is objective and well-defined, it can in principle be re-collected by anyone from video footage.

1.4. Application and Interest / Impact
The particular application in this paper is in the NBA. Extensions to other professional basketball leagues around the world would be straightforward, as well as to college basketball. Extensions to other sports would take longer to first develop the augmented dataset, but in principle, there is no obstacle.

Further, beyond evaluating the given dataset shown here, the approach is viable for any such question for any dataset. The only requirements are that both the new and the old datasets are in the same form (i.e., quantitative columns of information) and that there exists a market forecasting results that the data could help predict.

Thus, the approach presented here impacts virtually all modern and popular sports.
2. Data

Datasets need to be combined with intelligence to derive actionable value. The novel method proposed here is to standardize the intelligence across datasets by using deep learning, a machine learning algorithm that mimics human intelligence by using high-level hierarchical abstractions and structures. The 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 adds value.

The specific dataset used here is from Vantage Sports, where dozens of unique metrics are tabulated by highly trained human analysts 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 with boxscore and optical datasets.

The original dataset is all publicly available NBA data, including box score 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.

In terms of typical file size, number of rows, and number of data points, all on a per-game basis, boxscore and play-by-play data have the lowest, optical data the highest, and Vantage data is in the middle.

Boxscore and play-by-play data include some basic information about every possession, so the file is typically about 100 rows and has about 700 data points.

Optical data includes two-dimensional court coordinates for all ten players and three-dimensional coordinates for the ball, both at 25 frames per second. But not all players are tracked, for example, during free throws, so the overall number of rows usually is less than the calculated maximum of $2 \cdot 10 \cdot 25 \cdot 60 \cdot 48 + 3 \cdot 25 \cdot 60 \cdot 48$, which is about one and a half million, but could be as large as two million if the referee location data were also made available.

Vantage data has fewer data points than optical but more information, because it does not simply report location data that needs to be processed into basketball intelligence; instead it reports the actionable information directly: 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. A more detailed discussion of both the process and the output of Vantage data will be given below.

| 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      | ~3,000 | ~16,000     | High        | High      |
The difference between data, information, and knowledge has long been recognized in the field of knowledge management and information science, c.f. [5] for a conceptual review and multiple definitions of what they call "these three key concepts." These three concepts are often visualized as a pyramid with data on the bottom, information above it, and knowledge above that, the implication being 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 by sheer volume, optical has the most, play-by-play the least, and Vantage in the middle. However, in terms of information, the optical data has the least, because the information embedded in where a particular player was standing is very small, especially when compared with the virtually identical numbers coming both before and after. By contrast, the information in the boxscore and play-by-play data is higher: knowing that a particular player scored or assisted in a basket is a small amount of data but a larger amount of information. The information from the Vantage data source is higher still 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 may be viewed as the insights extracted from the meaningful information. There is indeed some amount of knowledge to be extracted from box score and play-by-play statistics, and over the past several decades, much valuable work has done just that, from boxscore-metrics such as wins produced [6] to various forms of plus-minus metrics [7]. And there is knowledge to be extracted from the optical data as well, such as attempts [8] to use machine learning to recognize on-ball screens from the optical data.

Note that the knowledge that might possibly be extracted from the optical data is a subset of the information directly available from the Vantage data. The information available from Vantage is highest of all because it includes actionable basketball facts with embedded meaning. Further, the additional knowledge available to be extracted from that high amount of information is itself high, because, among other things, adding context can help rank players and teams based on the important metrics, to evaluate performance, aid development, and search for underrated players.

A two-second clip exemplifies the differences among these three data sources. The clip, annotated with Vantage Sports data and described with interviews, is available at [9].

To start 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. In fact, the complete description of those two seconds in the official play-by-play reads as follows:

11:21 [ATL 55-43] Millsap Layup Shot: Made (12 PTS) Assist: Teague (3 AST)

This description has four pieces of data: the scorer, the passer, the shot outcome, and the shot type. Those are also pieces of information. By contrast, the corresponding optical data has about 50 pairs of location data for each of the ten players plus 50 triplets of location data for the ball, or about 650 data points. But none of that data is information. The corresponding Vantage data is shown on the next page in Table 2. It contains 53 points of data, which are also all points of information.
| Event Description | Details |
|-------------------|---------|
| **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 |
2.1. NBA data
The NBA dataset is sourced from the nba.com website, as well as some commonly calculated additional information such as scheduling (back-to-back indicators, rest days, etc.). Overall, about 50 metrics per game comprise the “standard” dataset, prior to augmenting. For space concerns, Table 3 lists the standard abbreviations used in the dataset. The exact definitions can be easily understood or found on their website. The data collected is for the 2014-2015 NBA season.

Table 3: Metrics in the standard dataset

| Scheduling: | Number of games in last 5 days, back-to-back, rest days, list of referees |
| Regular:    | PTS_OFF_TOV,PTS,PTS_PAINT,PTS_2ND_CHANGE,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,PCTPTS_OFFSET_TOV,PCTPTS OFFSET_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,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,CFG,CFG,UFGM,UFGA,DFGM,DFGA |

Notice that the standard metrics include certain metrics derived from the optical data and made available on a per-game basis. These include items such as total distance run, total touches, secondary assists, passes, and contested, uncontested, and defended at the rim field goal attempts and makes. Note though that these definitions of contested and defended are purely proximity based and do not distinguish between defenders actively contesting a shot 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. Those tags are then cross-referenced and cross-validated to ensure validity.

The tags that are tracked by human eyeballs 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 is contested shots. It is common knowledge that having a hand up is the key to good shot defense, c.f. [4]. But no data source makes that information available. It is not in any box score, play-by-play, or optical database. Vantage Sports has this data, for every shot attempt, by every player, on every team, in every game.

As another example, Vantage tracks if a pass was made to an open shot, 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 3 lists a representative sample of the metrics in the augmented dataset. Here the metrics are spelled out because they are unique. Further information is available on the Vantage Sports enterprise website. The data for this augmented dataset is also for the 2014-2015 season.

Table 3: 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/DRP Embarrass Rate,... |
| Screening Offense and Defense: | Screens Received/Set per Chance, Split%, Solid Screen%, KIF%, Hedge%, ... |

2.3. Betting markets

Data for historical betting markets for the 2014-2015 NBA season can be sourced from various 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; c.f. [10] which shows that opening NBA lines have substantial biases when a high-quality player is absent, but that all the biases are eventually removed and 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 both teams combined will exceed some threshold. More complex bets are also available, but for simplicity, the model in all cases is trained only on these two, the most standard bets.

Note also that assumed throughout this paper is the standard betting cost: losses pay 10 percent more than wins. 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 $11/(11 + 10) = 52.381\%$. 
3. Methods

Here is how the comparison is done, conceptually, for each dataset.

We begin with the 50th day of the season in order to have a base of data from which to start.

Each day, we create the following training table. Columns indicate the date of the game, the two teams, the market betting lines, and a 20-day moving average of the (standard or augmented) metrics for each team. We run two deep learning algorithms for each dataset: one to predict the actual resulting spreads, and one to predict the actual resulting over/unders.

We find the best fit model with a deep learning algorithm using the standard parameters of h2o [11]. Taking that best fit model, we 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. Predictions, in both cases, that significantly exceed equiprobability with statistical confidence are assumed to be placed bets; those that do not are assumed to be skipped bets. This allows for both algorithms to decide not only whether to bet over or under, but also allows them the ability to choose not to bet at all. Again, this is the same procedure for both the standard dataset and the augmented dataset.

Then we conglomerate the two strategies within each dataset (spreads and over/unders) to compare one time series against the other.

4. Results

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

First, it is possible that neither the original nor the augmented dataset can beat the breakeven probability of 52.381%. In this case, it would be difficult to decide if the augmented data is marginally valuable, since even a random coin toss can achieve a 50% probability. This is the most likely scenario in general. It is after all quite difficult to beat the markets. In this case, we would not be able to simply conclude that the data is more valuable.

Second, it is possible that both the original and the augmented dataset can beat the breakeven probability. This is the least likely scenario. Essentially by selection bias, if you are already able to beat the breakeven probability, you are unlikely to be searching for augmented data. Nevertheless, in this case, it would still be possible to run a statistical test to evaluate if the marginal contribution is worth changing the predictive algorithm and adopting the augmented dataset.

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

Fourth, it is possible that the original dataset fell short of the breakeven probability, while the augmented dataset exceeded it. This is the most exciting scenario as we would be able to then conclusively determine that the augmented dataset did provide more value.
Here are the results from the data above.

Starting with an assumed initial bankroll of $5,000, the daily rolling deep learning algorithm using the standard dataset is correct on 49% of wagers and ends the season with $1,700. 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 indeed quite efficient, and we should expect that they incorporate all standard publicly available information.

Using Vantage data, the algorithm is correct on 54% of wagers and ends the season with $6,500. The difference is highly statistically significant (p-value < 0.01). And it exceeds the breakeven probability. See Figure 1 for a time series graph of both strategies.

Figure 1: Time series of deep learning strategies using the **standard** and **augmented** datasets

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

All of these results are by construction out-of-sample.

Furthermore, the model was also tested in paper trading for the start of the 2015-2016 season. Without any modification to the deep learning model for the augmented dataset, simply by continuing as if the new season were an extension of the old, the model continued outperforming the breakeven probabilities and had a reported winning probability of 56.65% for bets on spreads and over/unders through 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 justifiable datasets as well as an associated and 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 would probably not apply to financial events where the markets are likely very efficient. In the world of sports analytics, however, the method outlined here can be used to substantively and permanently impact any sport that has a healthy wagering market associated with it.

Applying the method to NBA betting markets with a standard dataset available publicly as well as an augmented one incorporating 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 with the augmented data is above the betting breakeven probability.

Finally, the same model without modification continues to outperform in subsequent markets and games, yielding a winning probability in excess of 56 percent for bets on games from the start of the season on October 27, 2015 through December 13, 2015.

Such a result is unprecedented and remarkable for two reasons: first, it is notoriously difficult to beat an efficient market; and second, it demonstrates conclusively the power of the approach presented in this paper.

Extensions to other professional or collegiate leagues would be a straightforward application. In addition, extensions to other professional or collegiate sports could also be pursued.

References

[1] Barlow, Janine M. (2015). Data Analytics in Sports. O'Reilly Media: USA. First edition August 2015.
[2] Schmidhuber, Jürgen. (January 2015). “Deep learning in neural networks: An overview,” Neural Networks 61, 85-117.
[3] Fawcett, Tom (2006). “An introduction to ROC analysis,” Pattern Recognition Letters 27:8, 861-874.
[4] Csapo, Peter; Raab, Markus (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. doi:10.1371/journal.pone.0114184
[5] Zins, Chaim (2007). “Conceptual approaches for defining data, information, and knowledge,” J. Am. Soc. Inf. Sci. 58:4, 479-493.
[6] Berri, David; Schmidt, Martin; Brook, Stacey (2007). The Wages of Wins: Taking Measure of the Many Myths in Modern Sport, Updated Edition. Stanford Business Books.
[7] Engelmann, Jeremias (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.
[8] McQueen, Armand; Wiens, Jenna; Guttag, John (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 December 13, 2015.
[9] Abbott, Henry (May 5, 2015). “Jeff Teague on art of the playmaker,” ESPN TrueHoop, available online at http://espn.go.com/blog/truehoop/post/_/id/73055/jeff-teague-on-art-and-science-of-the-playmaker, retrieved on December 13, 2015.
[10] Dare, William H.; Dennis, Steven A.; Paul, Rodney J. (2015). “Player absence and betting lines in the NBA,” Finance Research Letters 13, 130-136.
[11] Candel, Arno (February 26, 2015). “The Definitive Performance Tuning Guide for H2O Deep Learning,” retrieved from http://h2o.ai/blog/2015/02/deep-learning-performance/ on December 14, 2015.