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
0169-2070/

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

Please cite this article in press as: Maymin, P.Z., Wage against the machine: A generalized deep-learning market test of dataset value. International Journal of Forecasting (2017), https://doi.org/10.1016/j.ijforecast.2017.09.008.