STATISTICS-FREE SPORTS PREDICTION

ALEXANDER DUBBS

Abstract. We use a simple machine learning model, logistically-weighted regularized linear least squares regression, in order to predict baseball, basketball, football, and hockey games. We do so using only the thirty-year record of which visiting teams played which home teams, on what date, and what the final score was. No real "statistics" are used. The method works best in basketball, likely because it is high-scoring and has long seasons. It works better in football and hockey than in baseball, but in baseball the predictions are closer to a theoretical optimum. The football predictions, while good, can in principle be made much better, and the hockey predictions can be made somewhat better. These findings tell us that in basketball, most statistics are subsumed by the scores of the games, whereas in baseball, football, and hockey, further study of game and player statistics is necessary to predict games as well as can be done.

1. Introduction

There is a long tradition in statistics in predicting many aspects of athletic events, in particular which teams will win, which players are the best, and the propensity of players to become injured. The tradition began in baseball, and was glorified in *Moneyball* [35], but it has now extended to almost all other major sports. With the growing popularity of sports gambling and “fantasy” sites, there is more demand than ever for statistical information about which players will succeed and which teams will win.

This paper uses a simple, weighted, and penalized regression model (see [41]) to predict the outcome of MLB, NBA, NFL, and NHL games, using data going back more than thirty years scraped from the websites [1], [2], [3], and [4]. It bears some resemblance to the model in [22], except it measures the ability of teams over games instead of players over possessions, and it takes into account home-field advantage. We intentionally limit our data use to the date, home and visiting teams, and score of each game, and we compare our predictions to a theoretically near-optimal indicator. Doing so tells us what statistical information is contained just in the scores, and whether what are commonly referred to as “statistics” have real predictive power. In basketball, the statistics are largely made unnecessary by the record of game scores, whereas in football this is clearly not the case. Baseball and hockey lie somewhere in the middle. This is likely because basketball has long seasons and high-scoring games, whereas baseball and hockey have long seasons but low-scoring games. Football has short seasons and is effectively “low-scoring,” because what matters is the number of scores that take place, not the scores’ point.
values. The models are trained on the even-numbered years and tested on the odd-numbered years.

The theoretically near-optimal indicator works as follows: Since our data is historical, we can predict every game by looking at the eventual end-of-season rankings and always bet that the eventually higher-ranked team will win. This estimator does not adjust for schedule difficulty, but is nonetheless very hard to beat. We demonstrate the performance of our penalized regression compared to this estimator. Furthermore, we show how it can be computed quickly using the Woodbury Matrix Identity.

Earnshaw Cook published the first major work on sabermetrics (baseball statistics) in 1964, [20]. [55] uses a Bayesian hierarchical model to predict Major League Baseball games, and [36] does so using ensemble learning. [7] and [38] predict baseball games using a number of statistics. [30] uses a k-nearest-neighbor algorithm to predict Korean baseball games. [11] models individual baseball games as Markov processes and studies many aspects of the game, including batting order, but does not post predictions. [12] studies winning and losing streaks in baseball. [31] and [32] use Bayesian hierarchical models to study hitting performance in baseball. [29] believes the most important trait in a baseball player is his propensity to get on base.

The first model for predicting the outcome of professional basketball games appeared in [28]. [17] and [40] use player position data to predict how likely an NBA team is to score on a given possession. [40] uses many other statistics as well, and achieves better results than we do in basketball prediction, but over a shorter time period using many more statistics. [49] does about as well as we do at NBA prediction but also over a very short time span and using many statistics. [7] and [50] use a simple models to predict NBA games, [39] uses a Kalman Filter, and [51] uses a Naive Bayes predictor. [13] surveys many NBA prediction methods. [18] predicts the betting line in NBA games. [19] predicts the likelihood of making a three-pointer using a logistic regression. [15] and [23] study the “hot hand” effect, in which they do not believe. [47] studies how to win the playoffs.

[44] uses several machine learning methods, decision trees, rule learners, neural networks, naive Bayes, and random forests, and many statistics to predict NCAA basketball games. [10], [14], [16], [45], and [46] use different methods to predict the NCAA men’s basketball. In fact, the Journal of Quantitative Analysis in Sports ran an entire issue on NCAA prediction in 2015 [24].


[48] explains the extent to which casino betting lines predict NFL games, which raises interesting questions about the power of democracy in prediction. [50] predicts the betting lines. [43] predicts NFL games using Twitter, another democratic approach. [37] studies the NFL draft.

There is also some past work done on NHL hockey, including one paper on game prediction [72] using neural networks, and one paper on scoring rates [12]. There are other papers using various factors to predict hockey games [53], [34].
[9] and [26] survey a group of machine learning methods used in sports prediction in general. There is also a wealth of research on the statistics of soccer games.

2. The Model

For \( y \) denoting the year, let \( b(y) \) be a vector such that \( b_{2i-1}^{(y)} \) are the visiting home scores, respectively, in the \( i \)-th game of the year \( y \) season. Let \( L^{(y)} \) be twice the number of games in year \( y \) and let \( P \) be the total number of teams that have played in either the MLB, NBA, NFL, or NHL since the 1986 season. Let \( A^{(y)} \) be a \( L^{(y)} \times P \) matrix that is all zeros except that if teams \( j \) and \( k \) are the visitors and home teams in game \( i \) of the year \( y \) season, \( A_{21-1,j}^{(y)} = A_{21-1,k+p}^{(y)} = A_{21,k+2p}^{(y)} = A_{21,j+3p}^{(y)} = 1 \). Setting up the \( A^{(y)} \) matrices in this way allows the model to take home-field advantage into account. Let \( y_{\text{back}} \) be the number of seasons considered other than the current season used to predict the current season, it is sport-dependent. In baseball, football, and hockey \( y_{\text{back}} = 4 \), and in basketball \( y_{\text{back}} = 2 \).

Let \( D^{(z)} \) be a diagonal matrix such that

\[
D^{(z)}_{i,i} = \frac{d^{(z)}_2}{1 + \exp \left( -d^{(z)}_1 \left( \frac{y_{i,k} - y_{i,j} - d^{(z)}_4}{L^{(y)}} \right) \right)},
\]

where the \( d^{(z)}_{1:4} \) are tuning parameters picked by maximizing the predictivity of the upcoming model on even-numbered years. We pick the logistic curve because if its versatility; it can model a line, a concave-up curve, a concave-down curve, and a step function.

For matrices \( U^{(1)}, \ldots, U^{(n)} \), let their vertical concatenation be \( [U^{(1)}; \ldots; U^{(n-1)}; U^{(n)}] \), with \( U^{(1)} \) on top. We will now explain how to predict whether \( b'_{2i-1}^{(y)} - b'_{2i}^{(y)} \) is positive or negative using only historical data (if it is zero we say that we predicted it correctly one half of one time). We use a weighted regularized linear least squares regression, information about them can be found in [11]. Let

\[
X^{(y)} = \left[ D^{(y_{\text{back}})} A^{(y-y_{\text{back}})}; \ldots; D^{(1)} A^{(y-y_{\text{back}})}; D^{(0)} A^{(y)} \right],
\]

\[
Y^{(y)} = \left[ D^{(y_{\text{back}})} b^{(y-y_{\text{back}})}; \ldots; D^{(1)} b^{(y-y_{\text{back}})}; D^{(0)} b^{(y)} \right],
\]

and let \( M^{(y)} = L^{(y-y_{\text{back}})} + \ldots + L^{(y-1)} + L^{(y)} \). Let

\[
K^{(y,i)} = (X^{(y)} (1:(M^{(y)}+2i-2));:) \cdot X^{(y)} (1:(M^{(y)}+2i-2));:) ^{-1}
\]

\[
w^{(y,i)} = (K^{(y,i)} + I)^{-1} (X^{(y)} (1:(M^{(y)}+2i-2));:) \cdot Y^{(y)} (1:(M^{(y)}+2i-2));:)
\]

Typically there is a positive \( \lambda \) parameter in front of the \( I \), we omit it for it is absorbed by the \( D^{(z)} \). Typically also \( X^{(y)} \) and \( Y^{(y)} \) would be centered. Empirically this appears unnecessary for our problem. It is not necessary to invert the whole matrix, Gaussian elimination may be used (backslash in MATLAB). To do prediction, set

\[
\hat{Y}^{(y)} (M^{(y)}+2i-1:(M^{(y)}+2i)) = X^{(y)} (M^{(y)}+2i-1:(M^{(y)}+2i));: w^{(y,i)}
\]

The sign of \( \hat{Y}^{(y)} (M^{(y)}+2i-1 - Y^{(y)} (M^{(y)}+2i)) \) predicts the sign of \( b'_{2i-1}^{(y)} - b'_{2i}^{(y)} \), in other words, which team will win. Remember, the teams playing are contained in rows \( (M^{(y)} + \ldots + M^{(y+2i-1)})) \)
\[ 2i - 1 : (M(y) + 2i) \] of \( X(y) \). The entries in the \( D(z) \) are picked so that the sum of the model’s correct predictions is as high as possible on even years.

This process can be accelerated. First compute

\[
K(y, i) = (X(y)_{1:M(y), :})^t X(y)_{1:M(y), :},
\]

\[
w(y, i) = (K(y, i) + I)^{-1}(X(y)_{1:M(y), :})^t Y(y)_{1:M(y), :}.
\]

Hypothesize that we know \( K(y, i) \) and \( w(y, i) \), we will find them for \( i + 1 \). Let \( \tilde{x} = X(y)_{(M(y) + 2i - 3); (M(y) + 2i - 2), :} \), \( \tilde{y} = Y(y)_{(M(y) + 2i - 3); (M(y) + 2i - 2), :} \) and \( \tilde{u} = (K(y, i) + I)^{-1}\tilde{x} \).

By the Woodbury Matrix Identity,

\[
w(y, i+1) = (I - \tilde{u}(I_{2 \times 2} + \tilde{x}\tilde{u})^{-1}\tilde{x}\tilde{u})^{-1}(\tilde{w} + \tilde{u}\tilde{y}),
\]

\[
K(y, i+1) = K(y, i+1) + \tilde{x}\tilde{x}.
\]

3. Results

The following table shows the results of the model on all four sports, on the even years on which it was trained, on the odd years, and on all years. The form of the results is the probability of correctly predicting the winner of a game. The “Model” column denotes the performance of our model, whereas the “Oracle” column denotes the performance of the theoretically hard-to-beat model described in the introduction which uses information from the future to predict the past.

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLB Even Years</td>
<td>0.5524</td>
<td>0.5756</td>
</tr>
<tr>
<td>MLB Odd Years</td>
<td>0.5480</td>
<td>0.5760</td>
</tr>
<tr>
<td>MLB All Years</td>
<td>0.5502</td>
<td>0.5758</td>
</tr>
<tr>
<td>NBA Even Years</td>
<td>0.6869</td>
<td>0.6840</td>
</tr>
<tr>
<td>NBA Odd Years</td>
<td>0.6773</td>
<td>0.6812</td>
</tr>
<tr>
<td>NBA All Years</td>
<td>0.6821</td>
<td>0.6826</td>
</tr>
<tr>
<td>NFL Even Years</td>
<td>0.6347</td>
<td>0.7184</td>
</tr>
<tr>
<td>NFL Odd Years</td>
<td>0.6237</td>
<td>0.7185</td>
</tr>
<tr>
<td>NFL All Years</td>
<td>0.6292</td>
<td>0.7184</td>
</tr>
<tr>
<td>NHL Even Years</td>
<td>0.5818</td>
<td>0.6079</td>
</tr>
<tr>
<td>NHL Odd Years</td>
<td>0.5819</td>
<td>0.6107</td>
</tr>
<tr>
<td>NHL All Years</td>
<td>0.5819</td>
<td>0.6093</td>
</tr>
</tbody>
</table>

The first four figures show the performance of our model (in blue) vs. the oracle (in red) in every year from 1986-2015. The performance is measured by the ratio of games predicted correctly. The figures are in the order MLB, NBA, NFL, NHL. They show that the model performs well in basketball, which has long seasons and high-scoring games. It performs passably in baseball and hockey and poorly in football. These results indicate that most basketball statistics are subsumed by the game scores. This is somewhat the case in baseball and hockey and not the case in football. The hockey graph “jumps” during the strike in the 2005 season.

The second four figures show the percentage of times that each team won in the 2015 season in red and the percentage of times they were predicted to win in blue.
The x-axis is the end-of-season ranking of the team where 1 (leftmost) is the best. The figures are in the order MLB, NBA, NFL, NHL.

References

Figure 2. NBA: Our model (blue) vs. the oracle (red).

Figure 3. NFL: Our model (blue) vs. the oracle (red).
Figure 4. NHL: Our model (blue) vs. the oracle (red).

Figure 5. MLB: Percentage wins (red) vs. predicted percentage wins (blue).
Figure 6. NBA: Percentage wins (red) vs. predicted percentage wins (blue).

Figure 7. NFL: Percentage wins (red) vs. predicted percentage wins (blue).
Figure 8. NHL: Percentage wins (red) vs. predicted percentage wins (blue).


[48] Na Wei, “Predicting the outcome of NBA playoffs using the Naive Bayes Algorithms,” available on the internet.


DEPTS. OF MATHEMATICS AND CSE, UNIVERSITY OF MICHIGAN

E-mail address: alex.dubbs@gmail.com