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 tells 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

Date: December 23, 2015.

Key words and phrases. Sports Prediction, Sabermetrics, Machine Learning, MLB, Baseball, NBA, Basketball, NFL, Football, NHL, Hockey.

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. [5] 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. [44] 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 [56] 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.

[42] 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 [25].

[21] uses a probit regression to predict football games, and [33] uses a neural network to predict football games. [24] uses a Bayesian hierarchical model to predict football games. [6] uses numerous methods to predict NFL games. [54] predicts college football games. [8] uses neural networks to predict both professional and college football games. [27] uses a stochastic process model to rate high school and college football teams.

[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 [52] 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 $b^{(i)}_{2i-1:2i}$ are the visiting and 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, $A^{(y)}_{2i-1,j} = A^{(y)}_{2i-1,k+P} = A^{(y)}_{2i,k+2P} = A^{(y)}_{2i,j+3P} = 1$. Setting up the $A^{(y)}$ matrices in this way allows the model to take home-field advantage into account. Let y_{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

$$D^{(z)}_{l,l} = d^{(z)}_1 + \frac{d^{(z)}_2}{1 + \exp\left(-d^{(z)}_3\left(\frac{l}{L^{(z)}} - d^{(z)}_4\right)\right)},$$

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 $b^{(y)}_{2i-1} - b^{(y)}_{2i}$ 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 [41]. Let

$$X^{(y)} = [D^{(y_{\text{back}})}A^{(y-y_b)}; ...; D^{(1)}A^{(y-1)}; D^{(0)}A^{(y)}],$$

$$Y^{(y)} = [D^{(y_{\text{back}})}b^{(y-y_b)}; ...; D^{(1)}b^{(y-1)}; D^{(0)}b^{(y)}],$$

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

$$K^{(y,i)} = (X^{(y)}_{1:(M^{(y)}+2i-2),:})^t X^{(y)}_{1:(M^{(y)}+2i-2),:},$$

$$w^{(y,i)} = (K^{(y,i)} + I)^{-1}(X^{(y)}_{1:(M^{(y)}+2i-2),:})^t Y^{(y)}_{1:(M^{(y)}+2i-2)}.$$

Typically there is a positive λ 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 $\hat{Y}^{(y)}_{M^{(y)}+2i-1} - \hat{Y}^{(y)}_{M^{(y)}+2i}$ predicts the sign of $b^{(y)}_{2i-1} - b^{(y)}_{2i}$, in other words, which team will win. Remember, the teams playing are contained in rows $(M^{(y)} +$

$2i-1):(M^{(y)}+2i)$ 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

$$K^{(y,1)} = (X^{(y)}_{1:M^{(y)},:})^t X^{(y)}_{1:M^{(y)},:},$$

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

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 $\tilde{u} = (K^{(y,i)} + I)^{-1}\tilde{x}^t$. By the Woodbury Matrix Identity,

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

$$K^{(y,i+1)} = K^{(y,i+1)} + \tilde{x}^t\tilde{x}.$$

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.

	Model	Oracle
MLB Even Years 1986-2015	0.5524	0.5756
MLB Odd Years 1986-2015	0.5480	0.5760
MLB All Years 1986-2015	0.5502	0.5758
NBA Even Years 1986-2015	0.6869	0.6840
NBA Odd Years 1986-2015	0.6773	0.6812
NBA All Years 1986-2015	0.6821	0.6826
NFL Even Years 1986-2015	0.6347	0.7184
NFL Odd Years 1986-2015	0.6237	0.7185
NFL All Years 1986-2015	0.6292	0.7184
NHL Even Years 1986-2015	0.5818	0.6079
NHL Odd Years 1986-2015	0.5819	0.6107
NHL All Years 1986-2015	0.5819	0.6093

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.

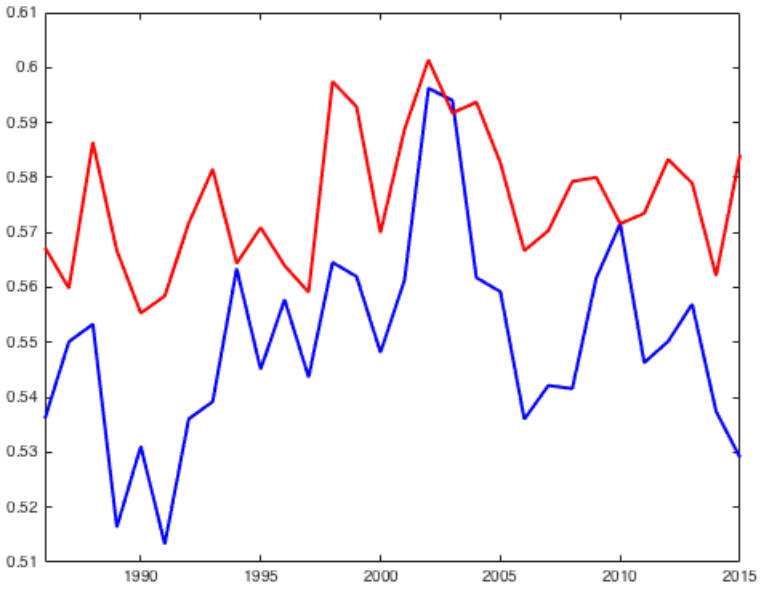

FIGURE 1. MLB: Our model (blue) vs. the oracle (red).

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

[1] http://www.baseball-reference.com
[2] http://www.basketball-reference.com
[3] http://www.pro-football-reference.com
[4] http://www.hockey-reference.com
[5] David Abuaf, Tony Chen, Alexander Trifunac, "Can One Predict Next Year's Winning Percentage using OLS Regression on Baseball Statistics?," available online.
[6] Eduardo C. Balreira, Brian K. Miceli, Thomas Tegtmeyer, "An Oracle Method to Predict NFL Games," *Journal of Quantitative Analysis in Sports*, 10, 183-196. 2014.
[7] Matthew Beckler, Hongfei Wang, Michael Papamichael, "NBA Oracle," available online.
[8] Andrew D. Blaikie, Gabriel J. Abud, John A. David, R. Drew Pasteur, "NFL & NCAA Football Prediction using Artificial Neural Networks," *Proceedings of the 2011 Midstates Conference on Undergraduate Research in Computer Science and Mathematics*.
[9] Jack David Blundell, "Numerical Algorithms for Predicting Sports Results," available online.
[10] Mark Brown, Joel Sokol, "An Improved LRMC Method for NCAA Basketball Prediction," *J. Quant. Anal. Sports*, Volume 6, Issue 3, 2010, Article 4.
[11] Bruce Bukiet, Elliotte Rusty Harold and Jose Luis Palacios, "A Markov Chain Approach to Baseball," *Operations Research*, Vol. 45, No. 1 (Jan. - Feb., 1997), pp. 14-23.
[12] Samuel E Buttrey, Alan R Washburn, Wilson L Price, "Estimating NHL Scoring Rates," Volume 7, Issue 3 (Jul 2011), *Journal of Quantitative Analysis in Sports*.
[13] Chenjie Cao, "Sports Data Mining Technology Used in Basketball Outcome Prediction," Master's Dissertation, Dublin Institute of Technology.
[14] Bradley P. Carlin, "Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information," *The American Statistician*, 50:1, 39-43, 1996.

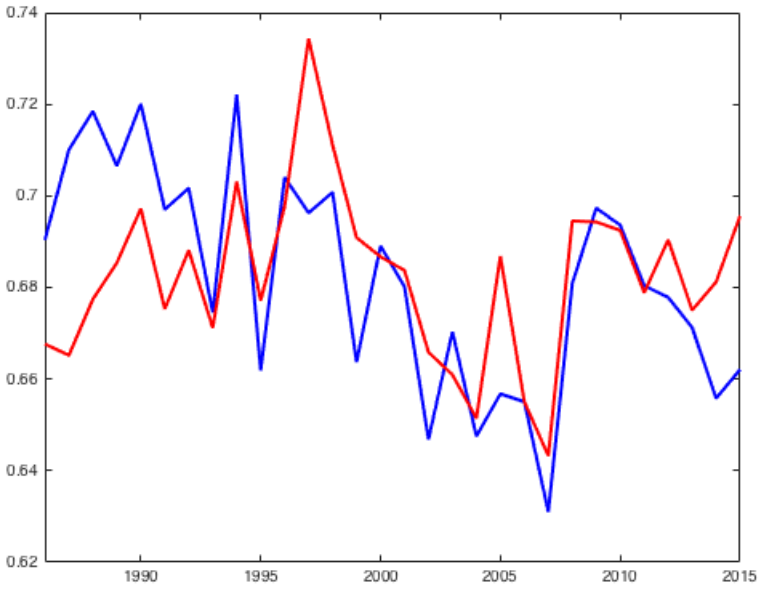

FIGURE 2. NBA: Our model (blue) vs. the oracle (red).

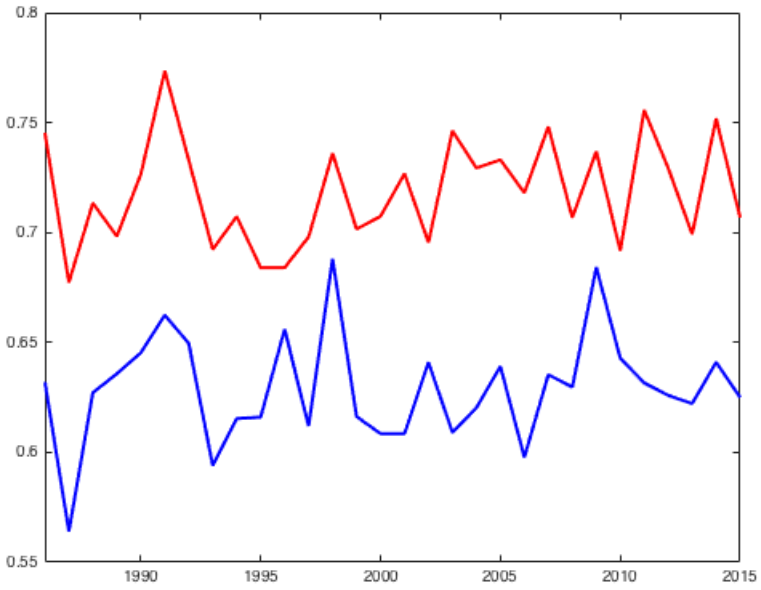

FIGURE 3. NFL: Our model (blue) vs. the oracle (red).

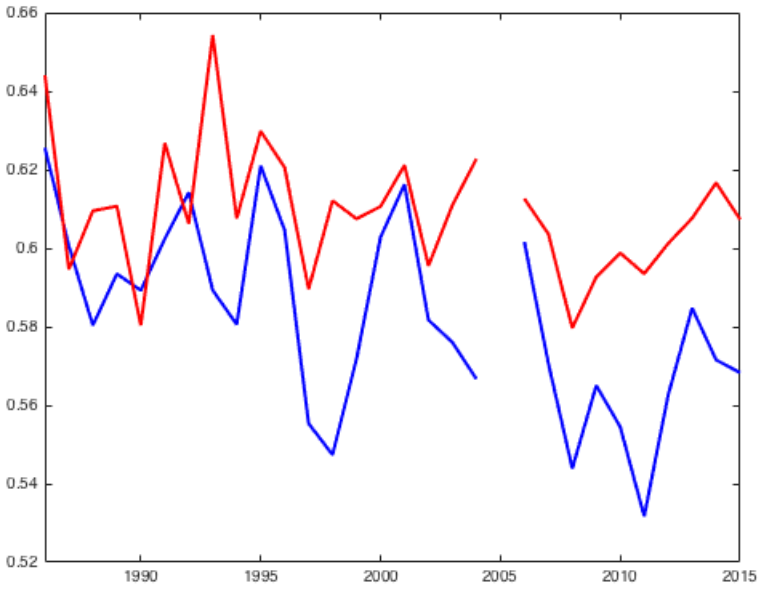

FIGURE 4. NHL: Our model (blue) vs. the oracle (red).

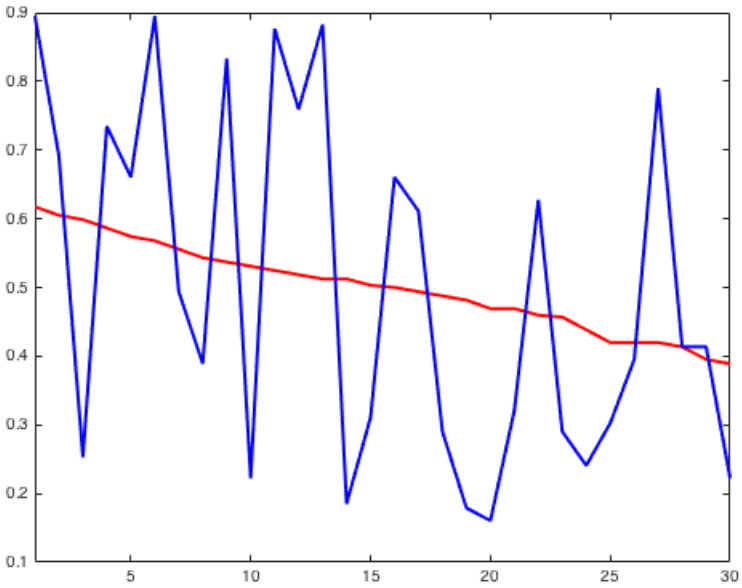

FIGURE 5. MLB: Percentage wins (red) vs. predicted percentage wins (blue).

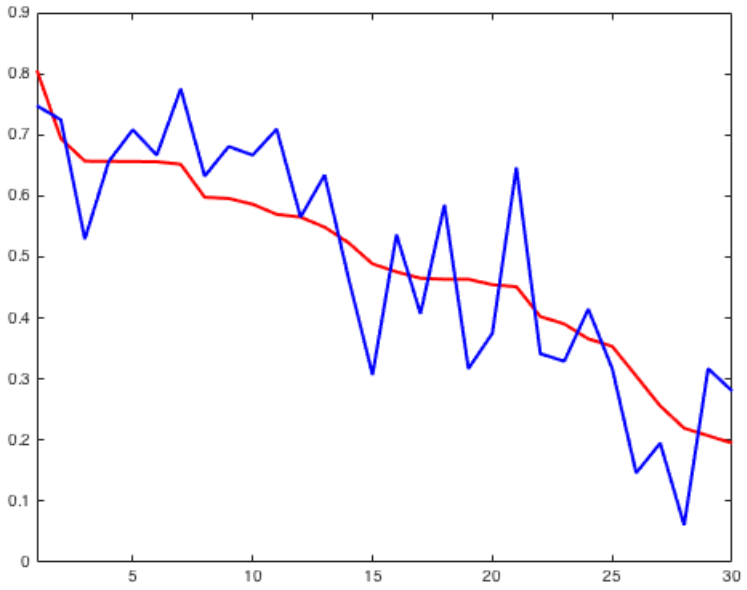

FIGURE 6. NBA: Percentage wins (red) vs. predicted percentage wins (blue).

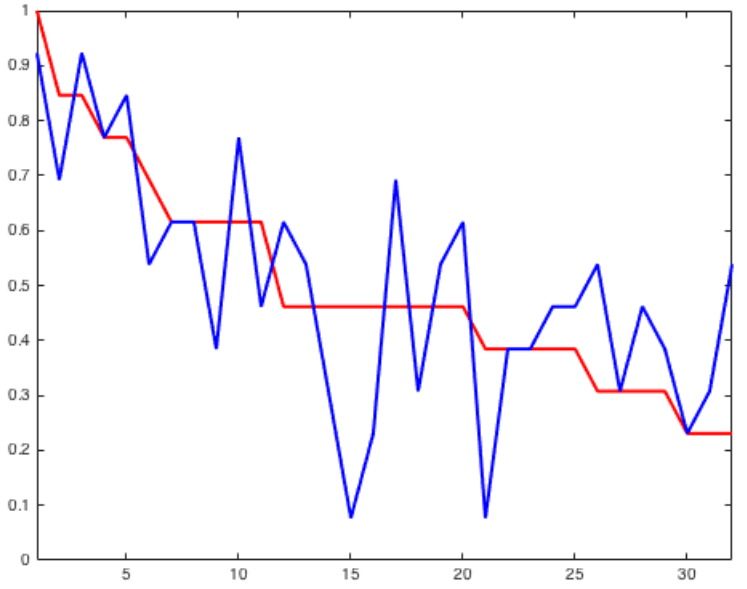

FIGURE 7. NFL: Percentage wins (red) vs. predicted percentage wins (blue).

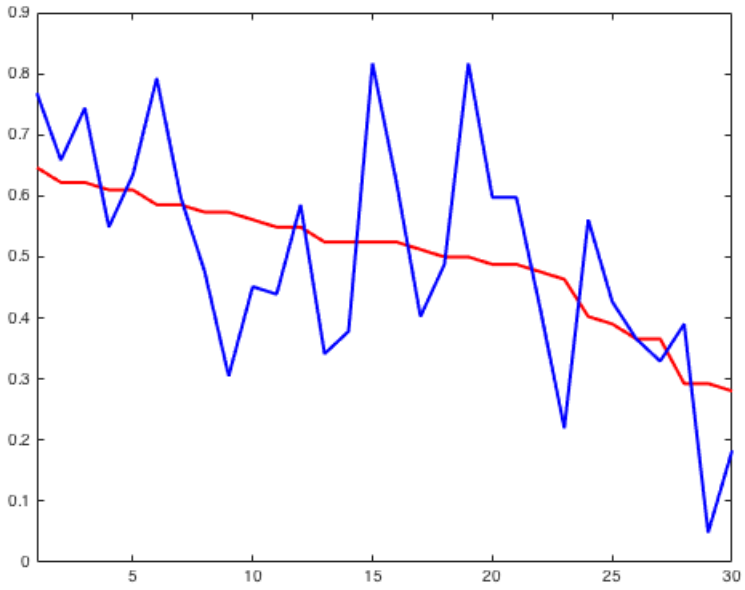

FIGURE 8. NHL: Percentage wins (red) vs. predicted percentage wins (blue).

[15] Eugene M. Caruso, Nicholas Epley, "Hot Hands and Cool Machines: Perceived Intentionality in the Prediction of Streaks," available online.
[16] Steven B. Caudill, "Predicting discrete outcomes with the maximum score estimator: the case of the NCAA men's basketball tournament," *International Journal of Forecasting* 19 (2003) 313-317.
[17] Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry, "A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes," arXiv, 2015.
[18] Bryan Cheng, Kevin Dade, Michael Lipman, Cody Mills, "Predicting the Betting Line in NBA Games," available online.
[19] Torin K. Clark, Aaron W. Johnson, Alexander J. Stimpson, "Going for Three: Predicting the Likelihood of Field Goal Success with Logistic Regression," March 1-2, 2013, 7-th annual MIT Sloan Sports Analytics Conference.
[20] Earnshaw Cook, *Percentage Baseball*, Waverly Press (1964).
[21] David N. DeJong, "Using Past Performance to Predict NFL Outcomes: A Chartist Approach," 1997, available online.
[22] Paul Fearnhead and Benjamin Matthew Taylor. "On estimating the ability of nba players." *Journal of Quantitative Analysis in Sports*, 47(3):1298, July 2011.
[23] Robert Vallone and Amos Tversky, "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology*, 17, 295-314 (1985).
[24] Mark E. Glickman and Hal S. Stern, "A State-Space Model for National Football League Scores," 1998, *Journal of the American Statistical Association*, March 1998, Vol. 93, No. 441.
[25] Mark E. Glickman and Jeff Sonas, "Introduction to the NCAA men's basketball prediction methods issue," *J. Quant. Anal. Sports* 2015; 11(1): 1-3.
[26] Maral Haghighat, Hamid Rastegari, and Nasim Nourafza, "A Review of Data Mining Techniques for Result Prediction in Sports," *ACSIJ*, Vol. 2, Issue 5, No. 6, November 2013.

[27] David Harville, "The Use of Linear-Model Methodology to Rate High School or College Football Teams," *Journal of the American Statistical Association*, Vol. 72. No. 358 (Jun., 1977), pp. 278-289.
[28] Evan Heit, Paul C. Price, Gordon H. Bower, "A Model for Predicting the Outcomes of Basketball Games," *Applied Cognitive Psychology*, Vol. 8, 621-639 (1994).
[29] Adam Houser, "Which Baseball Statistic Is the Most Important When Determining Team Success?," The Park Place Economist, Volume XIII.
[30] Wu-In Jang, Aziz Nasridinov, Young-Ho Park, "Analyzing and Predicting Patterns in Baseball Data using Machine Learning Techniques," *Advanced Science and Technology Letters*, Vol. 62 (Sensor 2014), pp. 37-40.
[31] Shane T. Jensen, Blakeley B. McShane, and Abraham J. Wyner, "Hierarchical Bayesian Modeling of Hitting Performance in Baseball," *Bayesian Analysis* (2009).
[32] Wenhua Jiang and Cun-Hui Zhang, "Empirical Bayes in-season prediction of baseball batting averages," Vol. 6 (2010) 263-273.
[33] Joshua Kahn, "Neural Network Prediction of NFL Football Games," 2003, available online.
[34] Benjamin Leard, Joanne M. Doyle, "The Effect of Home Advantage, Momentum, and Fighting on Winning in the National Hockey League," *Journal of Sports Economics*, October, 2011, Vol. 12 No. 5 538-560.
[35] Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*, W. W. Norton & Company, New York, NY, 2004.
[36] Arlo Lyle, "Baseball Prediction Using Ensemble Learning," available online.
[37] Kimberly J. McGee and Lee N. Burkett, "The National Football League Combine: A Reliable Predictor of Draft Status?," *Journal of Strength and Conditioning Research*, 2003, 17(1), 6-11.
[38] Dennis Moy, "Regression Planes to Improve the Pythagorean Percentage: A regression model using common baseball statistics to project offensive and defensive efficiency," undergraduate thesis, U.C. Berkeley.
[39] Jirka Poropudas, "Kalman Filter Algorithm for Rating and Prediction in Basketball," Master's Thesis, University of Helsinki, Faculty of Social Sciences, 2011.
[40] , "Modelling the NBA to Make Better Predictions," Master's Thesis, MIT, 2013.
[41] Ryan Rifkin, Gene Yeo and Tomaso Poggio, book chapter, "Regularized Least-Squares Classification," http://cbcl.mit.edu/publications/ps/rlsc.pdf
[42] Zifan Shi, Sruthi Moorthy, Albrecht Zimmerman, "Predicting NCAAB match outcomes using ML techniques - some results and lessons learned," arXiv, 2013.
[43] Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A. Smith, "Predicting the NFL Using Twitter," arXiv, 2013.
[44] C. Sire and S. Redner, "Understanding baseball team standings and streaks," Eur. Phys. J. B 67, 473-481 (2009).
[45] H.O. Stekler and Andrew Klein, "Predicting the Outcomes of NCAA Basketball Championship Game," Research Program on Forecasting, Working Paper No. 2011-003.
[46] Sara Stoudt, Loren Santana, Ben Baumer, "In Pursuit of Perfection: An Ensemble Method for Predicting March Madness Match-Up Probabilities," available online.
[47] Michael R. Summers, "How to Win in the NBA Playoffs: A Statistical Analysis," *American Journal of Management*, vol. 13(3) 2013.
[48] Greg Szalkowski and Michael L. Nelson, "The Performance of Betting Lines for Predicting the Outcome of NFL Games," arXiv, 2012.
[49] Renato Amorim Torres, "Prediction of NBA games based on Machine Learning Methods," 2013, available on the internet.
[50] Jim Warner, "Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line," available on the internet.
[51] Na Wei, "Predicting the outcome of NBA playoffs using the Naive Bayes Algorithms," available on the internet.
[52] Joshua Weissbock, "Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data," Master's Thesis, University of Ottawa, 2014.
[53] Joshua Weissbock, Diana Inkpen, "Combining Textual Pre-game Reports and Statistical Data for Predicting Success in the National Hockey League," *Advances in Artificial Intelligence*, Volume 8436 of the series *Lecture Notes in Computer Science* pp. 251-262, 2014.

[54] Brady T. West, Madhur Lamsal, "A New Application of Linear Modeling in the Prediction of College Football Bowl Outcomes and the Development of Team Ratings," *J. Quant. Anal. Sports*, Volume 4, Issue 3, 2008, Article 3.

[55] Tae Young Yang and Tim Swartz, "A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball," *Journal of Data Science*, 2(2004), 61-73.

[56] Yuanhao (Stanley) Yang, "Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Baseball Statistics," undergraduate thesis, U.C. Berkeley.

Depts. of Mathematics and CSE, University of Michigan
E-mail address: alex.dubbs@gmail.com