Image Processing Using Convolution Matrices - Multithreaded CPU vs FPGA

Preston Scott
Department of Computer Science
NC State University
Raleigh, NC 27695
pdscott2@ncsu.edu

Abstract

As data mining techniques have overtaken the sports industry, no sport stands more impacted than baseball, which is particularly steeped in statistical history. In this paper, we will examine individual player statistics which summarize the past and test their ability to predict a team's future. Existing popular composite metrics (including Wins Above Replacement) will be evaluated in terms of their predictive power. In addition, we will develop new models to compete with WAR in predicting the future of a team. With an emphasis on methodologies which leverage publicly available data, we will assess just how much of the future can be revealed through an examination of the past.

1 Introduction

Statistical techniques have helped bring scientific rigor to many disciplines in which scientific thinking has been historically absent. A major example of this is the use of statistical techniques in professional sports. Sports decisions which were once thought to require a human expert are now being made through the use of sophisticated algorithms. Major League Baseball (MLB) is a sport which is particularly rich in statistics and in the last decade, teams have began exploiting these statistics to make better decisions.

A challenge in analyzing baseball statistics is consolidating the multitude of available data and deriving meaningful conclusions. Baseball record keepers are notorious for documenting statistics on every aspect of the game - even down to the weather conditions at game time. Obviously, much of this data is meaningless to the task at hand. There have been many attempts to consolidate statistics into composite "power" metrics. One frequently referenced metric is Wins Above Replacement (WAR). WAR endeavors to represent all the attributes of a player (batting, fielding, base running, etc) in a single number that represents the player's worth to the team.

Although WAR was designed to characterize past performance of individual players, WAR can also be used to predict future team performance. A 2009 analysis showed that WAR could predict a team's winning percentage with a correlation factor of 0.83 [1]. In 2016, an artifical intelligence startup called Unanimous A.I. successfully predicted all 8 Major League playoff teams based on their midseason results[2]. The team also correctly predicted that the Chicago Cubs would win their first championship in 108 years.

But if WAR is the industry's best attempt at a consolidated metric, it could also use some improvement. A correlation factor of 0.83 is impressive but not perfect. Moreover, WAR is a vastly complex number to calculate, requiring dozens of inputs - some of which are locked away in proprietary datasets owned by large information technology firms.

The goal of this paper is to take metrics which were designed to summarize individual player history and turn them into predictors of the team future performance. We will start by visiting existing methodologies (WAR in particular) and assess how well they can predict team performance. We will then search more broadly for raw data which can be used to model team performance. Lastly, we will determine how far such a model can be extrapolated in the future (i.e. can a final season outcome be predicted at mid-season?)

2 Methods

Team performance can measured in many different ways. The first question to be answered was whether to use classification or regression to evaluate each team's performance. One obvious choice for quantifying team performance was the number of wins per season. Other choices included Boolean categories (such as whether the team advanced to the playoffs or not) and other categorical classes (such as good, better, best, etc.)

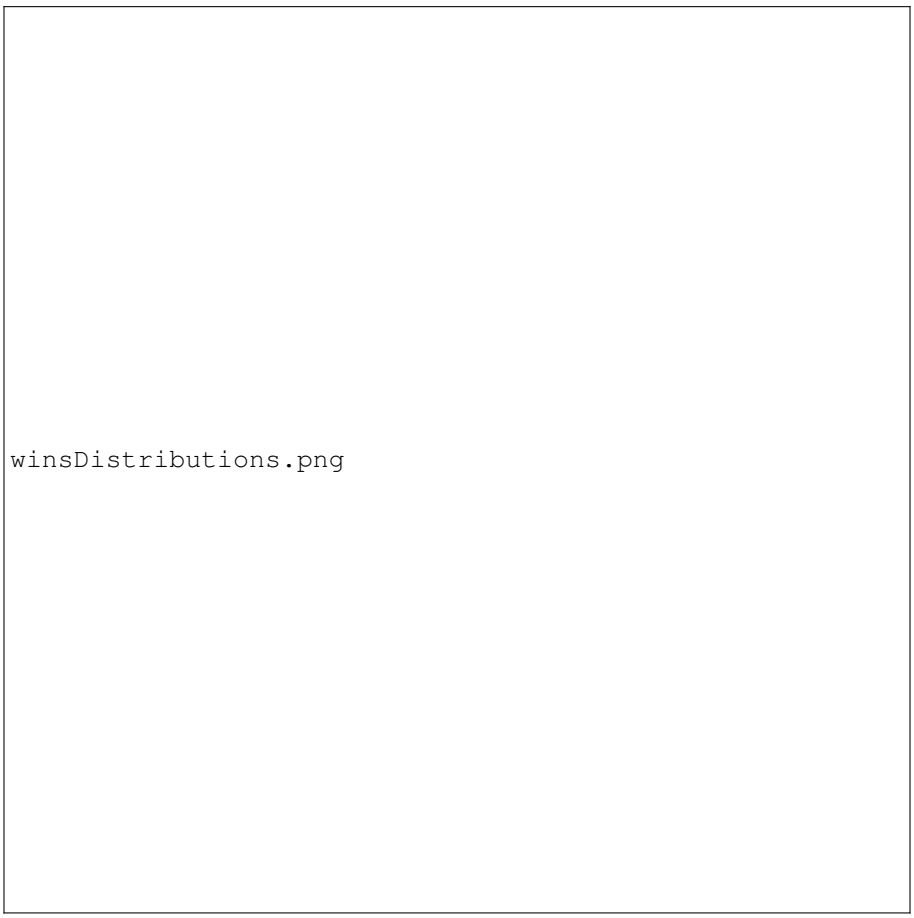

Figure 1: Number of wins per season for all teams from 2001-2014. Left plot shows playoff status true (red) and false (black). Right plot shows histogram of number of wins for all teams for all seasons.

To decide on a proper evaluation methodology, team performance data from years 2001-2014 was analyzed to identify key trends. Figure 1 shows the playoff status versus the number of wins in a season for all teams from 2001-2014. The red and black dots represent teams which did and did not make the playoffs, respectively. As shown, the playoff and non-playoff distributions blend together with no clear delineation between them - a sign that classification based on playoff status would be challenging.

Figure 2 shows a histogram of the number of wins per season for teams from 2001 - 2014. Again, no clear delineation between "good" and "bad" teams was observed. Based on the preliminary analysis, it was decided to analyze the data as a regression problem instead of a classification problem. The number of wins for the season was chosen to be the target Y value to be used for prediction.

3 Plan and Experiment

The main goal of this project was to develop a statistical method of predicting team performance. Along the way, several questions needed to be answered:

- Could any existing methods (RAW in particular) be leveraged as a starting point for the prediction ?
- Could existing methods be improved upon or replaced by a new model developed from the body of publicly available data?
- How far into the future could a prediction be made?

These questions are addressed in the sections below:

3.1 Analyzing Existing Methods

WAR numbers are provided for each player on several popular sports websites [4,5,6]. However, each provider has a slightly different take on the WAR formula and the resultant numbers are different in each case. Therefore, to use WAR as a basis for prediction, a standard WAR formula would need to be selected. Part of this project was to select a standard WAR methodology and then reproduce the requisite WAR numbers to check for accuracy.

Although there are many variations of WAR, they all have common themes. Firstly, metrics for pitchers and positional players must be calculated separately. A pitcher's performance is evaluated with an emphasis on defense and his ability to keep the other team from scoring. The inputs for a pitcher's WAR score include the number of home-runs, walks, and strikeouts he has given up. They are rolled together into a score based on the following equation [4]:

$$PitchingScore = 13 * HR + 3 * (BB + HBP - IBB) - 2 * SO$$

where HR is the number of home runs given up by the pitcher and HR, BB, HBP, IBB, and SO are the number of home runs, walks, hit batsmen, intentional walks, and strikeouts recorded by the pitcher, respectively. This result is standardized according to park and league factors.

For all other players, a combination of defensive and offensive attributes is considered. Batting, base-running, and fielding metrics are all combined to give a composite WAR score. Base-running and fielding have been historically difficult to quantify since there are few statistics which give insight into these activities. Modern technology has come to the rescue, however, and today these metrics are quantified using sophisticated video analysis tools which analyze gameplay and record the positions of all players at all times. Unfortunately most of this data is proprietary and unavailable to the public. With these critical inputs missing, it is difficult to reproduce the WAR score precisely using publicly available data. However, an approximation can be made using a players offensive statistics based on the following equation:

$$BattingScore = \frac{0.69*BB+0.72*HBP+0.89*1B+1.27*2B+1.62*3B+2.10*HR}{AB+BB-IBB+SF+HBP}$$

where BB = base on balls, HBP = hit by pitch, 1B = single, 2B = double, 3B = triple, HR = home run, AB = at bats, IBB = intentional walks and SF = sacrifice flies. The weights on each of the variables above has been assigned based on past likelihoods of potential scenarios influencing a game [4]. This metric is then standardized according to the league average wOBA and then adjusted for park and league factors.

The two equations above were used to calculate an approximation of the WAR score, henceforth known as WAR-lite. The raw data used for the inputs was taken from Sean Lahman's History of Baseball Database [3]. To arrive at a team level WAR-lite score, individual WAR-lite scores for

each pitcher and position player for the team were added together. To eliminate noise, position players with less than 300 plate appearances and pitchers appearing in less than 10 games were excluded. After compiling WAR-lite numbers for each team, regression analysis was used to assess the predictive power of the WAR-lite metric. For comparison, previously calculated WAR scores from an online publication were also analyzed to contrast the WAR-lite formulation with a formal WAR metric calculated using the complete set of inputs.

3.2 Development of New Model

Anticipating some room for improvement, a parallel path for investigation was to leave the WAR (and WAR-lite) metrics aside and develop a new model with the sole purpose of predicting team performance. To do this, we once again used the Lahman database [3] to mine raw statistics for analysis. Since the database was in SQL format, custom queries were created to retrieve the appropriate data for analysis [7]. The feature set was then analyzed using regression analysis to develop a linear combination of raw statistics that would be predictive of team wins. A list of the available features from this database is:

- Offensive Statistics: runs scored, at-bats, hits, doubles, triples, home-runs, walks, strike-outs, stolen bases, caught stealing, hit-by-pitch, sacrifice flies.
- Defensive Statistics: runs allowed, earned runs allowed, earned run average, complete games, shutouts, saves, innings pitched, hits allowed, home runs allowed, walks allowed, strikeouts, errors, double plays, fielding percentage.

During data preprocessing, high correlation was noted between several of the player statistics. This indicated some redundancy in the feature set and a potential for over-emphasizing certain measures. To combat this, principle component analysis was used to reduce the dataset to the most important contributors to variance.

One item missing from the chosen database was a "strength of schedule" metric which rates the quality of the opponents for each team. Despite all efforts, MLB schedules are not distributed fairly. Some teams end up playing strong opponents as many as twenty times per season while easier opponents are not faced at all. It was hypothesized that the use of player performance metrics alone would be naive in the sense that some players may be facing tougher opponents than others. Therefore, some weighting would need to be given to players (and teams) facing stronger opponents throughout the year.

Unfortunately, no easily accessible statistic was found to address the scheduling issue. Therefore, a new methodology was created. A particularly elegant way of expressing the schedules of all teams is within a 30 x 30 matrix, referred to henceforth as S. Each column and row in S represents a different team, for a total of 30 teams. Element $S_{i,j}$ is then the number of times that team i played team j during the season.

$$S_{i,j} = \begin{bmatrix} 0 & 3 & 7 & \cdots & 8 \\ 3 & 0 & 6 & \cdots & 1 \\ 7 & 6 & 0 & .\cdots & 5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 8 & 1 & 5 & \cdots & 0 \end{bmatrix}$$

This matrix representation becomes handy when coupled with another matrix, W, which expresses the winning percentage of each team. If any column j in S is multiplied by the row matrix W, then the result is a single number representing the strength of schedule for team j.

$$[83 \quad 77 \quad 63 \quad \cdots \quad 91] * \begin{bmatrix} 0 \\ 3 \\ 7 \\ \vdots \\ 8 \end{bmatrix} = ScheduleStrength$$

Since this type of matrix schedule formulation did not exist, the matrices for each relevant season needed to be generated. This was done using a Python script which absorbed the scheduling information from a public website, summed the opponent matchups for each team, and generated the final matrix [7].

To incorporate the newly created strength of schedule metric into the prediction, the full feature set was first modeled without a scheduling variable. Then the results of that model (predicted wins) were used as the W matrix for the strength of scheduling calculations. The strength of schedule numbers for each team were then added as a new column in the feature set and the entire data set was re-modeled.

3.3 Extension of Model

Although the prediction of team winning percentage from individual player statistics is academically interesting, the practical value of such a prediction is limited. If one must wait for an entire season of statistics to come in before making a prediction, then the records of each team will already be known. Therefore, to extend the usefulness of the model, final season results were predicted using the player statistics from mid-season.

Since most datasets report only full season results, it was difficult to find mid-season data for all the years represented in the training data set. Therefore, training a new model based on mid-year data was not possible. Instead, the same full season model used by just extrapolating the mid-season test set data to represent an entire year. Several methods of extrapolation were attempted, but the best results were obtained using simple linear extrapolation.

4 Results

4.1 Prediction Using Existing Methods

Due to the difficulties in accessing proprietary information, an exact reproduction of the WAR metric was not possible. However, WAR-lite approximations calculated using publicly available components were analyzed, as shown in Figure 2 (left). As shown, there is very poor correlation of the WAR metric to number of team wins. It was speculated that part of the poor correlation was due to the elimination of proprietary inputs of the WAR formula. To address this, previously calculated values were used from an online publication [4]. These values were also plotted against winning percentage (Figure 2, right). The predictive value of WAR was greatly improved using the fully formulated WAR values, but predictions were still highly erratic in some cases.

4.2 Prediction Using New Model

With the predictive power of WAR falling short of expectations, a new model was developed to provide better insight into team performance. The model was based on a feature set of 26 player statistics mined from the Lahman baseball database [3]. After some feature size reduction using principle component analysis, linear regression was used to fit raw player statistics with team winning performance. Despite all intuition to the contrary, the addition of a strength of schedule component was not helpful to the analysis. This component seemed to show no correlation to the actual team performance and only served to worsen the RMSE of the final prediction. Therefore, this component was not included in the final analysis.

The fit of the model to the training data is shown in Figure 3. Also shown is the results of the test data model prediction.

Turning these predications into practical implications, the actual standings for the 2014 MLB season are shown in Table 1 (at the end of the paper), along with the predictions from the model. Although it did not make sense to pursue this problem as a classification problem, it was still interesting to see the classifications that could be inferred after the wins prediction was complete. Based on the predicted wins, 5 out of 6 division winners were correctly predicted, along with 9 out of 10 playoff teams.

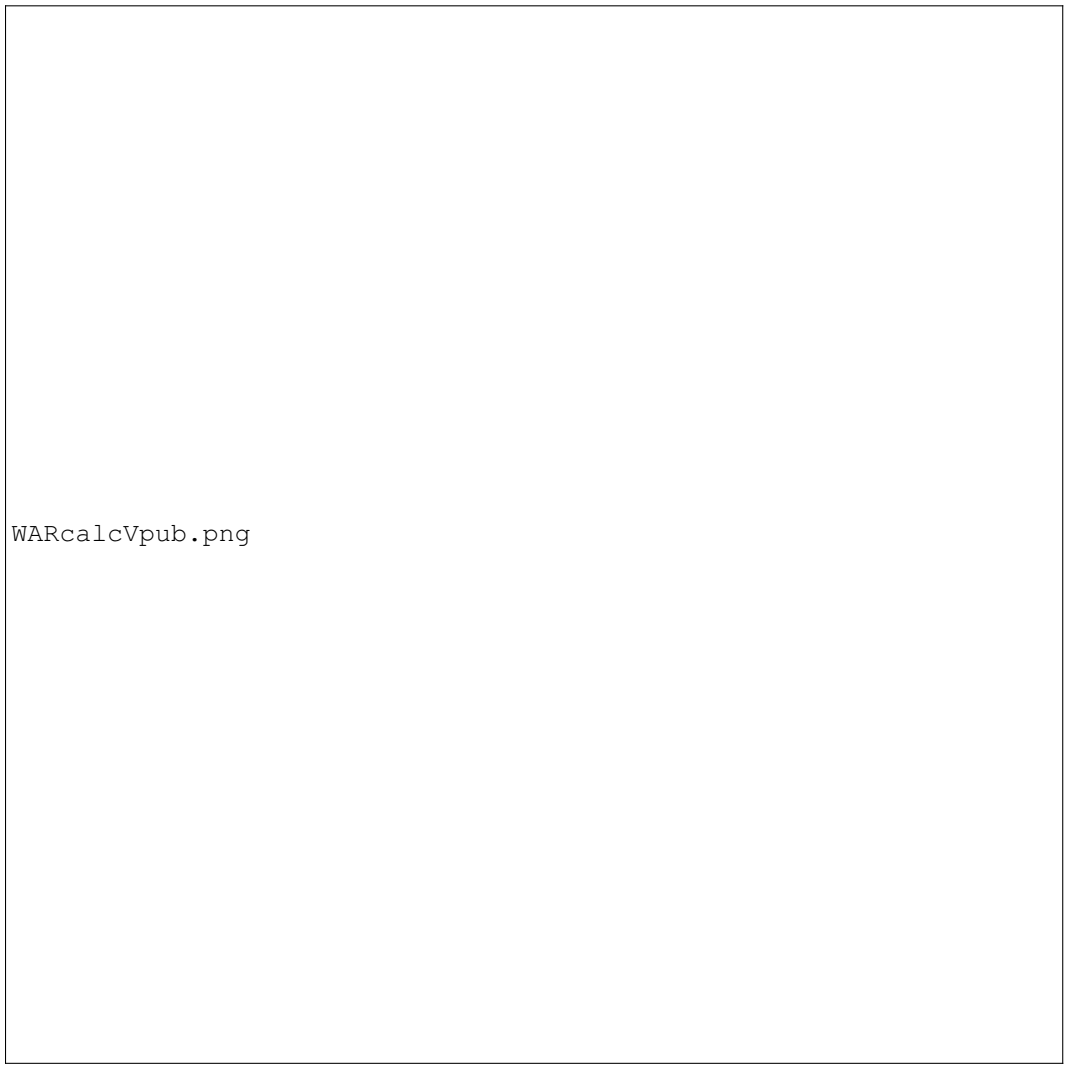

Figure 2: Calculated WAR-lite (left) and published WAR (right) values vs number of team wins.

4.3 Prediction Using the New Model Using Mid-Season Data

As discussed, a prediction which must wait for the results to come in before it can be calculated holds little practical value. As such, a prediction of the final season standings was made using only the data that was available at mid-season for that year (2014). The results of this prediction are shown in Table 2 (at the end of the paper).

5 Conclusions

In the first part of this project, we attempted to use an existing methodology (WAR) to make predictions about team performance. Although it was possible to make a reasonable prediction using published WAR scores, it was impossible to reproduce the WAR calculations because of the proprietary nature of some inputs. The "best attempt" WAR-lite metric proved to have very little predictive power. Therefore, at least as far as open methodologies are concerned, WAR proved to be inadequate.

In the second part of the project, a new model was developed using a linear combination of 26 publicly available individual player baseball statistics. The new model resulted in an RMSE of 2.5 on test set data and also correctly predicted 9 of the 10 playoff teams for the 2014 MLB season. One

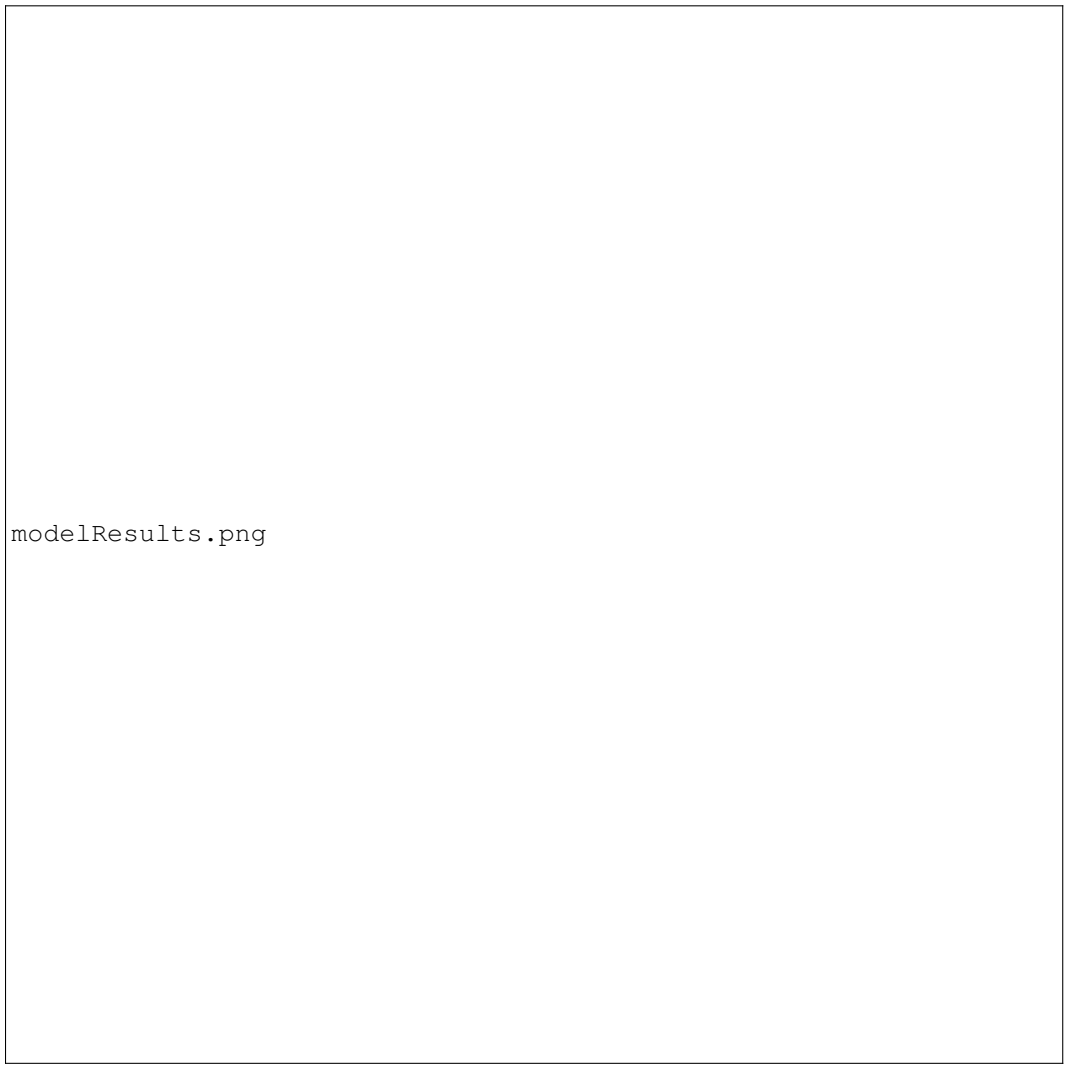

Figure 3: Linear regression model using training data.

surprising result was the lack of correlation between strength of schedule and number of team wins. This was even more disappointing given the large amount of time dedicated to addressing the issue. More extensive analysis may reveal a better way to incorporate strength of schedule to improve the predictions. Moreover, the matrix multiplication approach may be used to address other problems more elegantly.

The third part of the project endeavored to extend the usefulness of the new model by making year end predictions based on mid-season data. The prediction results were disappointing but could be improved immensely by addressing several potential issues. The majority of the error was probably due to the extrapolation required to turn mid-season data into an input for the full season model. If a better mid-season data source could be identified, a better approach would be to train a new model on mid-season data alone versus using extrapolation. If extrapolation must be used, a better approach would be to analyze each feature individually to see how it drifts from mid-season to end-of-season. Such an analysis would likely reveal that linear extrapolation is too crude to be used as a transformation method.

References

[1] Cameron, D. (2009) WAR: It Works, http://www.fangraphs.com/blogs/war-it-works/ .

[2] Unanimous A.I. Corporate Press Release (2016) A.I. Makes Yet Another Remarkable Prediction, http://www.marketwired.com/press-release/ai-makes-yet-another-remarkable-prediction-2172570.htm .

[3] The History of Baseball (2016), http://www.seanlahman.com/baseball-archive/statistics/ .

[4] What is WAR? (2016), http://www.fangraphs.com/library/misc/war/

[5] Baseball-Reference.com WAR Explained (2013),

[6] Baumer, B.S. & Jensen, S.T. & Matthews, G. J. (2015) openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball. (arXiv:1312.7158v3), https://arxiv.org/pdf/1312.7158v3.pdf .

[7] Code for MLB Data Analysis Project https://github.com/pdscott/MLBDataProject

Table 1: 2014 MLB Standings with Actual and Predicted Win Performance

2014 ACTUAL TEAM WINS VS MODEL

Division	Team	Actual Wins	Predicted Wins	Notes
AL East				
	Baltimore Orioles	96	95.87	✓⋆
	New York Yankees	84	84.11	
	Toronto Blue Jays	83	88.42	
	Tampa Bay Rays	77	83.53	
	Boston Red Sox	71	71.99	
AL Central				
	Detroit Tigers	90	87.46	✓⋆
	Kansas City Royals	89	85.47	
	Cleveland Indians	73	67.14	
	Chicago White Sox	73	67.14	
	Minnesota Twins	70	72.28	
AL West				
	Los Angeles Angels	98	95.92	✓⋆
	Oakland Athletics	88	91.17	⋆
	Seattle Mariners	87	87.19	
	Houston Astros	70	67.98	
	Texas Rangers	67	63.81	
NL East				
	Washington Nationals	96	97.77	✓⋆
	New York Mets	79	78.80	
	Atlanta Braves	79	80.38	
	Miami Marlins	77	76.64	
	Philadelphia Phillies	73	73.62	
NL Central				
	St. Louis Cardinals	90	88.34	⋆
	Pittsburgh Pirates	88	88.74	⋆
	Milwaukee Brewers	82	81.64	
	Cincinnati Reds	76	77.76	
	Chicago Cubs	73	71.98	
NL West				
	Los Angeles Dodgers	94	94.15	✓⋆
	San Francisco Giants	88	87.00	⋆
	San Diego Padres	77	73.90	
	Colorado Rockies	66	67.36	
	Arizona Diamondbacks	64	63.47	

✓ : Division winner correctly predicted (5 out of 6 correct)
⋆ : Playoff team correctly predicted (9 out of 10 correct)

Table 2: 2014 MLB Standings with Actual and Predicted Win Performance

2014 ACTUAL TEAM WINS VS MODEL

Division	Team	Actual Wins	Predicted Wins	Notes
AL East				
	Baltimore Orioles	96	87.67	
	New York Yankees	84	80.25	
	Toronto Blue Jays	83	88.34	
	Tampa Bay Rays	77	70.78	
	Boston Red Sox	71	70.12	
AL Central				
	Detroit Tigers	90	102.78	✓⋆
	Kansas City Royals	89	82.18	
	Cleveland Indians	85	83.55	
	Chicago White Sox	73	71.29	
	Minnesota Twins	70	80.34	
AL West				
	Los Angeles Angels	98	99.04	✓⋆
	Oakland Athletics	88	96.26	⋆
	Seattle Mariners	87	89.02	
	Houston Astros	70	65.28	
	Texas Rangers	67	69.91	
NL East				
	Washington Nationals	96	91.21	✓⋆
	New York Mets	79	75.63	
	Atlanta Braves	79	86.53	
	Miami Marlins	77	80.05	
	Philadelphia Phillies	73	65.56	
NL Central				
	St. Louis Cardinals	90	86.29	⋆
	Pittsburgh Pirates	88	79.01	
	Milwaukee Brewers	82	83.36	
	Cincinnati Reds	76	88.32	
	Chicago Cubs	73	69.64	
NL West				
	Los Angeles Dodgers	94	87.54	✓⋆
	San Francisco Giants	88	84.31	
	San Diego Padres	77	70.37	
	Colorado Rockies	66	76.47	
	Arizona Diamondbacks	64	67.95	

✓ : Division winner correctly predicted (4 out of 6 correct)
⋆ : Playoff team correctly predicted (6 out of 10 correct)