Stats 415 Final Project:
Modeling NBA Win Probability Using Historical Data

Christopher Lanctot, Jimmy Miron, Dimitry Slavin

Abstract— **This paper addresses modeling and predicting NBA teams' win probability based on past game statistics. A scraper written in Python was used to crawl ESPN.com to collect team and player statistics into a dataset. This dataset was then cleaned and transformed to generate metrics whose moving averages serve as predictors in various logistic regression models.**

First, single variable logistic regression analysis is applied in order to identify each metric's effect on win probability. The coefficients' magnitude from single variable logistic regression are then compared in order to determine the metrics' relative importance. Then, multivariate logistic regression analysis is employed to create a mathematical model that is capable of predicting a team's win probability based on the moving average of five metrics— points differential, defensive efficiency, offensive efficiency, adjusted field goal percentage, and assist ratio. Finally, principle component analysis is used in combination with multivariate logistic regression. Surprisingly, PCA did not improve the accuracy of prediction from the initially achieved 67%.

I. INTRODUCTION

Since the 1980s, the National Basketball Association (NBA) has cemented itself as a major player in the world of professional sports. With the growth of cable television, games are watched by millions of fans both domestically and internationally.

Furthermore, with 30 teams each playing 82 games per season, professional basketball has become one of the most common avenues for sports' gamblers to place bets. Professionals with years of experience and a strong understanding of the sport generate gambling odds for each game, which can then be bet on through online gambling houses or casinos. Many die-hard basketball fans with good instincts have profited from betting against these odds. But how much does instinct *really* have to do with it?

In this report, we will show that even with relatively little experience watching and analyzing basketball, statistical learning methods can be used to systematically predict the outcome of games.

II. Data Analysis Methods

A. Research Questions

Our overarching goal for the project was to accurately predict the outcome of NBA games, but in order to guide our analysis we asked four main research questions:

1. What are the most influential team metrics/statistics for predicting the probability of a team winning?
2. Does the importance of these metrics relative to each other change from season to season?
3. What combination of predictors for a logistic regression model yields the best prediction results?
4. If we were to use moving averages of metrics as predictors in a logistic regression model, what is the ideal look-back time?

B. Data Acquisition

Since there was no readily available and free dataset of NBA game and player statistics, we created our own. With the help of a scraper written in Python, we collected team and player statistics for every NBA regular season game from 2004 to 2015 from ESPN.com.

The scraper gathered data using the following three basic steps:

1. Collect team names and associated team home page URLs
2. For each team, collect a list of game IDs corresponding to the games that the team played between 2004 and 2015
3. For each game ID, collect the box-score player statistics and save all the statistics into one CSV file

The final CSV file contained 317,868 rows, each row corresponding to a particular player playing in a particular game. In total, the scraper collected data for 13,106 games.

C. Data Cleaning

Due to various typos and inaccuracies in the way the game statistics were presented on ESPN.com, some of the collected data was unusable. We implemented a number of strategies to minimize the amount of unusable data.

An example of a reoccurring inaccuracy and the strategy we used to resolve it is described below:

For some of the games for which we collected data, the scraper returned 'Los Angeles' as the team name, not distinguishing between whether the team was the Los Angeles Clippers or the Los Angeles Lakers. Not wanting to throw out the game data for which this was the case, we developed an algorithm that could extrapolate the names of the teams playing in a given game based on the list of players playing in that game.

Nevertheless, some data was still unusable due to non-systematic inaccuracies. After data cleaning and interpolation our dataset contained 12,995 rows, meaning that only 111 games were necessarily withheld. Aggregate data loss, therefore, was only 0.85%.

D. Dataset

As previously stated, the dataset that we worked with contained statistics for 12,995 games. We looked at a total of 11 seasons (2004-05 to 2014-15), which results in roughly 1,200 games per season[1]. After aggregating game statistics for each player to game statistics for each team, we had 18 predictors for each team in each game. Table I summarizes the attributes of the dataset.

The statistics described in Table I were used to derive the large majority of advanced NBA statistics. These advanced metrics are commonly accepted to serve as better indicators for team performance, so we calculated them from our existing predictors and appended them to our dataset. These advanced statistics are summarized in Table II.

Notice that each advanced statistic can be derived from the statistics shown in Table I via the formula in the Calculation column.

Name of Dataset Attribute	Type of Variable	Description of Professor Attribute
game_id	Identifying	Identification Number *(unique to each game)*
season	Identifying	Season to which game corresponds to (04-05 to 14-15)
team_id	Identifying	Indentication Number (unique to each team)
home_01	Indicator	Whether or not a team was playing at home (1 if yes, 0 if no)
win_01	Indicator	Whether or not a team won the game (1 if yes, 0 if no)
PTSF	Numerical	Points For
PTSA	Numerical	Points Against
AST	Numerical	Assists
BLK	Numerical	Blocks
DREB	Numerical	Defensive Rebounds
OREB	Numerical	Offensive Rebounds
REB	Numerical	Total Rebounds (=OREB + DREB)
MIN	Numerical	Minutes Played (=240 unless game went into overtime)
PF	Numerical	Personal Fouls
STL	Numerical	Steals
TO	Numerical	Turnovers
3PM	Numerical	Three Pointers Made
3PA	Numerical	Three Pointers Attempted
FGM	Numerical	Field Goals Made
FGA	Numerical	Field Goals Attempted
FTM	Numerical	Free Throws Made
FTA	Numerical	Free Throws Attempted

Table I: Summary of Dataset Attributes

Variable	Description	Calculation
PTS_diff	Score Differential (Negative Number Indicates Loss)	PTSF - PTSA
PACE	Estimated Number of Possessions (Same for Each Team)	0.96*[FGA + TO + 0.44*FTA - OREB]
AST_ratio	Number of assists a team has per 100 possessions	(AST / PACE) * 100
DEF_eff	Number of points a team allows per 100 possessions	(PTSA / PACE) * 100
OFF_eff	Number of points a team scores per 100 possessions	(PTSF / PACE) * 100
FGP_adj	Adjusted Field Goal Percentage	[(PTS - FTM) / FGA] / 2
TO_rate	Percentage of team's possessions that end in a turnover	TO / PACE
OREB_p	Offensive Rebound Rate	OREB / (DREB + OREB)
DREB_p	Defensive Rebound Rate	DREB / (DREB + OREB)

Table II: Summary of Advanced Statistics

[1] Note that this estimate is not completely accurate because the 2011-12 NBA season endured a lockout period, cutting the number of games from 82 per season to 66.

E. Descriptive Statistics and Exploratory Data Analysis

Since the advanced statistics in Table II are commonly accepted to be better at describing team performance in a game, the rest of the analysis focuses on them. Before using the advanced statistics as predictors in our logistic regression model though, we got a feel for the data by looking at some descriptive statistics.

In order to better understand the predictors, Table III below displays the descriptive statistics of the advanced metrics that we calculated:

Stat	count	mean	std	median
PTS_diff	25990	0.00	13.36	0.00
PACE	25990	92.13	5.99	91.67
AST_ratio	25990	16.75	3.32	16.74
DEF_eff	25990	107.43	11.49	107.49
OFF_eff	25990	107.43	11.49	107.49
FGP_adj	25990	0.50	0.06	0.49
TO_rate	25990	0.15	0.04	0.15
DREB_p	25990	0.74	0.08	0.74
OREB_p	25990	0.26	0.08	0.26
FT_rate	25990	0.23	0.09	0.22

Table III: Descriptive Statistics of Advanced Metrics

Note that the count in Table III is double the amount of games in our dataset because each game statistic has a value for both the home and away team.

III. Statistical Learning Methods

A. Single Variable Logistic Regression

In order to determine the individual effect of the metrics on the outcome of a game we first ran separate logistic regressions on each metric against the win/loss indicator of the home team. The general logistic function used in the regressions is shown below:

$$F(x_i) = \frac{1}{1 + e^{-\left(\beta_0 + \beta_j^{h,n} x_j^{h,n} + \beta_j^{a,n} x_j^{a,n}\right)}}$$

$$= P\left(home\ team\ wins \middle| n, x_j^{h,n}, x_j^{a,n}\right)$$

where $x_j^{h,n}$ represents the ***n-game moving average*** of the j^{th} statistic for the home team, $x_j^{a,n}$ represents the same moving average statistic for the away team, and n represents the ***look-back***, which is defined as the number of past values[2] included in the moving average[3]. Also note that a positive value for β_j implies that an increase in x_j increases the probability of the home team winning.

There are several things to emphasize here. First, notice that we use a moving average of the statistics as predictors. This was done because it is an effective way to reduce variance in short-term fluctuations and capture the true performance of a team. Consequently, we generated multiple versions of our predictors based on (i) different values of n, where

$$n \in \{7,14,21,28,35\}$$

and (ii) two different methods of calculating the moving average: simple moving average (SMA) vs. exponential moving average (EMA).

Second, before running the regressions we normalized the predictors (i.e. subtracted the mean and divided by the standard deviation). The normalization of predictors is usually unnecessary in regression, however, since the goal is to determine the importance of the predictors relative to each other, normalization is key. Normalization allows us to compare the magnitudes of the β vectors generated for each statistic. Table IV below display the comparison of beta coefficients for a 21-game simple and exponential moving average, respectively:

Stat	beta_0 SMA (n = 21)	beta_h SMA (n = 21)	beta_a SMA (n = 21)	beta magnitude
PTS_diff	0.498	0.666	-0.621	1.038
DEF_eff	0.458	-0.498	0.456	0.816
OFF_eff	0.445	0.463	-0.431	0.774
FGP_adj	0.437	0.403	-0.379	0.705
AST_ratio	0.422	0.273	-0.211	0.545
TO_rate	0.412	-0.185	0.207	0.497
OREB_p	0.414	-0.186	0.200	0.496
FT_rate	0.409	0.162	-0.111	0.454
PACE	0.406	-0.082	0.069	0.420
Stat	**beta_0 EMA (n = 21)**	**beta_h EMA (n = 21)**	**beta_a EMA (n = 21)**	**beta magnitude**
PTS_diff	0.499	0.680	-0.629	1.052
DEF_eff	0.457	-0.492	0.446	0.806
OFF_eff	0.445	0.482	-0.452	0.797
FGP_adj	0.437	0.421	-0.404	0.729

[2] To be clear, the moving average does not include the statistics from the game for which the probability is being estimated. This is a slight variation on a traditional moving average.

[3] Note that the first n games of every season are used to initialize the moving averages

AST_ratio	0.422	0.277	-0.233	0.556
TO_rate	0.412	-0.188	0.209	0.499
OREB_p	0.414	-0.187	0.205	0.498
FT_rate	0.410	0.174	-0.118	0.461
PACE	0.407	-0.088	0.086	0.425

Table IV: 21-Game SMA and EMA Coefficient Comparison

Both of these tables are sorted by the coefficient vector's magnitude in descending order. It can be inferred that, due to it having the largest magnitude, score differential is the strongest predictor. This makes sense; if a team consistently beats their opponents by a large margin of victory, it's intuitive to expect that the team will win its next game. The next three predictors— offensive efficiency, defensive efficiency, adjusted field goal percentage— all share similar magnitudes, which are also comparatively large. After adjusted field goal percentage there is a significant drop in magnitude.

Before we compare further, it's important to note the purpose of regressing on the PACE variable. As noted in Table II, PACE, the number of possessions, is the same for the home and away teams. This is because both teams will have the same amount of possessions per game[4]. Since we are calculating a moving average of the statistics over past games, the PACE values for the home and away team will not be entirely the same, however, it is constant between teams and has relatively little influence over the game's outcome. Consequently, the PACE regression is viewed as a control; its' coefficient magnitude can be used as a benchmark for other predictors to gauge how much noise the logistic regression model is capturing.

In a multivariate regression, we would ideally like to only use predictors that have a major influence on the outcome of a game. If a (normalized) predictor has coefficient magnitude values that are just above or even smaller than that of the PACE predictor, we would not want to use it in our multivariate regression.

Accordingly, we decided to only use the top five predictors for the multivariate regression discussed in the next section as the 6th through 9th advanced predictors had roughly the same magnitude of coefficients as the PACE predictor.

In order to further justify our choice of predictors we performed an additional test: we analyzed the single variable regression coefficients separately for each season to ensure that the predictors used in the model are temporally homogenous. That is, to ensure no major trends affect the statistical significance of our predictors over time. To do this, we performed the logistic regression for each predictor for each season (i.e. 110 more times). These results are shown in Figure 1, which displays the coefficient magnitude for each predictor by season.

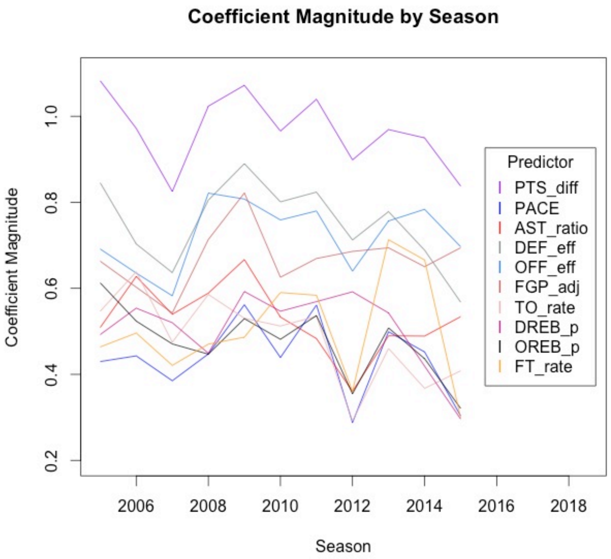

Figure 1: Coefficient Magnitude by Season (EMA, n = 21)

Notice that the top four predictors consistently have high coefficient magnitudes throughout the 11 seasons. Assist ratio does decrease from the 2009 season to the 2012 season but overall follows no major transitional trends. Also, given its influence in early seasons (04-05 to 08-09) and its recent increase it is ranked as a top predictor from our chosen scheme relative to the bottom four predictors. Hence, we found it acceptable to include this predictor in our multivariate regression.

Another important thing to mention is the general consistency of the magnitude rankings through time. Score differential remains the most influential metric for all eleven seasons, and the next three metrics, although slightly less consistent, still generally uphold the same ranking through time. This ranking consistency serves as a justification for training our multivariate regression model on the full eleven years worth of data, rather than training on each season separately. In conclusion, the team statistics that determined a game's outcome appear to be roughly the same throughout the entire eleven-year period being analyzed.

IV. Multivariate Logistic Regression

Feeling more confident about our choice in predictors, we trained ten multivariate logistic regression models, each corresponding to a unique

[4] ignoring differences of +/- 1

combination of n (look-back value) and moving average method (EMA or SMA). We then performed 10-fold cross validation on these models in order to identify which model attained the best accuracy. Figure 2 below shows the misclassification rate of all ten models:

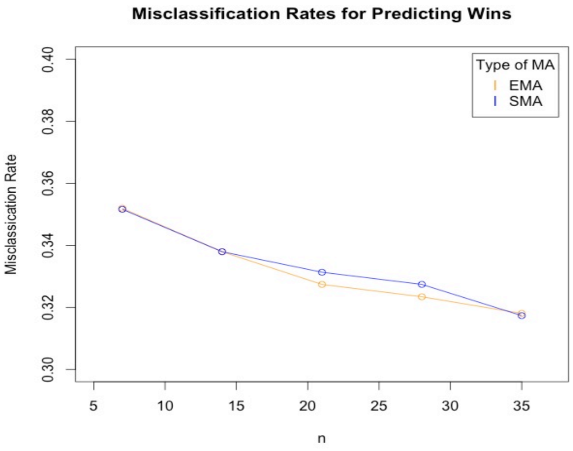

Figure 2: Misclassification of Game Outcomes- Comparing Ten Different Models

The first thing to notice here is how similar the misclassification rates are. Although an increase in the look-back improves prediction accuracy, the effect is rather weak. Increasing the look-back from 5 games to 35 games yields an error improvement of less than 5 percent.

The second thing to notice is the similarity between using a simple moving average versus an exponential moving average. Weighting more recent team performance more heavily generally does not affect the model's ability to predict. This conclusion somewhat disvalues the common betting strategy of using a team's recent 'hot streak' to justify an increase in its odds of winning its next game.

As far as choosing a model to employ, we decided that a 21-game look back with an exponential moving average achieves a satisfactory balance between model applicability and accuracy. Table V below displays the 11 coefficient values of the model for the normalized predictors:

Coefficient	Home	Away
Intercept	0.479	NA
PTS_diff	1.019	-0.132
AST_ratio	0.029	-0.002
DEF_eff	0.276	0.323
OFF_eff	-0.283	-0.284
FGP_adj	-0.028	-0.004

Table V: Multivariate Regression Model Coefficient Values (21-Game EMA)

V. Principle Component Analysis

Now that the data is well understood, principle component analysis was performed on the full set of original predictors. Following our decision from the cross validation performed in the previous section, we used the 21-game exponential moving average of each statistic as input to the PCA.

After identifying the set of principle components, we generated a set of multivariate logistic regression models. The set of models was made by iteratively increasing the number of principle components that were regressed upon.

34 principle components were identified, corresponding to 34 unique logistic regression models. Figure 3 is a scree bar plot of the top 25 principle components, superimposed onto a plot of the misclassification error of the 25 corresponding logistic regression models. Superposition was chosen to demonstrate the unison in behavior between the amount of variance captured and the error rate.

As we expect, the misclassification error decreases as the number of principle components used in the model increases. It is quite surprising, however, that using the principle components as predictors in the logistic regression yields little to no improvement in predictive accuracy. The misclassification error of the 21-game EMA multivariate logistic regression model from the previous section— a model that uses only ten predictors— achieves practically the same error rate as all 24 of the models depicted in Figure 3. Model interpretability was sacrificed for little to no gain in predictive accuracy.

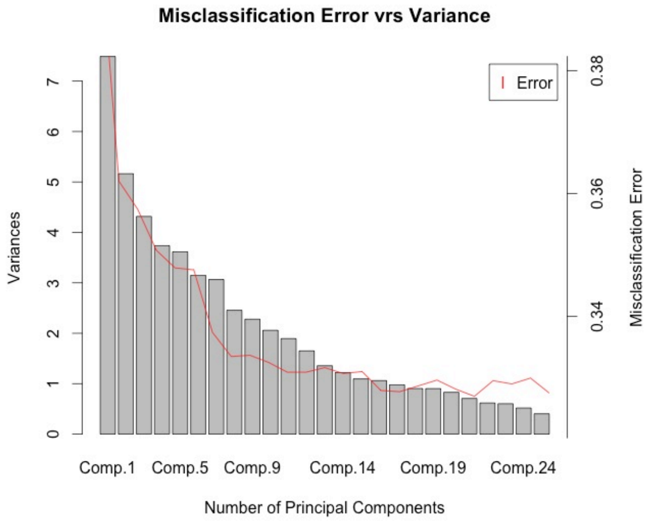

Figure 3: Scree Plot and Logistic Regression Misclassification Error

This result further justifies using the more simple logistic regression model from the previous section. It also provides support for the hypothesis that there seems to be a maximum level of predictive accuracy that one can achieve given the current data. A rough estimate for this supposed maximum level is 70%.

VI. Summary

We have shown that it is possible to effectively predict the outcome of NBA games using logistic regression on past game statistics.

At the beginning of the analysis we presented a method for identifying the most influential statistics for predicting the outcome of a game, concluding that teams' past score differential is consistently the most important metric.

Then, a multivariate logistic regression model was generated based on the top five most influential statistics, achieving a relatively low misclassification error rate of approximately 33%, yet still maintaining both simplicity and model interpretability.

Finally, we compared this simple model to a much more complicated logistic regression model whose predictors were determined using Principal Component Analysis. To our surprise, the more complex model was not able to achieve a higher predictive accuracy, suggesting that there may be an inherent limit on statistical learning methods' ability to predict the outcome of NBA games given the current data set.