# Project Milestone: Predicting NBA Player Performance

Adam Tong, John Conley, and Matt Girouard

October 28, 2014

## 1 Introduction

The objective of our project is to use historical NBA player data by season to predict a player's statistical performance for a particular season. We hypothesize that a player's statistics can be predicted by his performance in previous years and by the historical career trajectories of other, similar players.

At this milestone, we are comfortably on the way to achieving our goals. Here we present our data, models, results, and completion goals.

## 2 Data

We sourced all seasons from all players in basketball history, from 1946-47 to the present. This summed up to:

- 4217 Players
- 25413 Total Player Seasons

For each player season, we collected all of the "raw" statistics tracked for the player. During the late 70's, the NBA added a few measurements (turnovers) and changed a few rules (the three-point shot), and as such, some statistics are missing for players before 1979.

- Player Age
- Year
- Player Number of Years of Experience
- Team
- League (NBA/ABA)
- Games Played

- Minutes Played
- Field Goals
- Field Goal Attempts
- Field Goal Percentage
- Three Point Field Goals
- Three Point Field Goal Attempts
- Three Point Field Goal Percentage
- Free Throws
- Free Throw Attempts
- Free Throw Percentage
- Offensive Rebounds
- Defensive Rebounds
- Total Rebounds
- Assists
- Steals
- Blocks
- Turnovers
- Fouls
- Points

## 2.1 Data Normalization and Massaging

Thus far, our regressions have been based on per 36 minute data. 36 minutes is traditionally viewed as the average playing time per game of an average starter, and thus normalizing by minutes is a good way to map playing-time-agnostic averages. Playing time is notoriously hard to predict in the NBA.

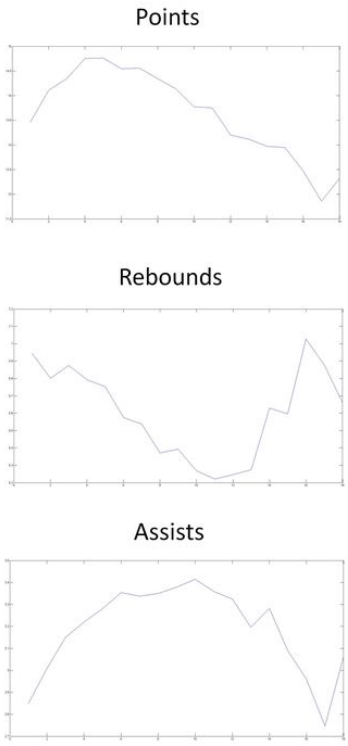
One problem with per-minute data, however, is that it fails to account for pace. The pace of a basketball game is the number of possessions of the game; intuitively, a higher pace means more field goal attempts and inflated statistics compared to a slower paced game. For instance, one historical season could have an average of 100 shots per game per team, and

another could have an average of 80 (this data is actually factual: the late 60s and early 70s had extremely fast-paced basketball, and the early 90s had comparatively slower-paced games). Normalizing for pace is now very standard among advanced basketball measurements because of how much the style of the game changes over teams and time.

We want to use both of these measures in our models. Of course, using one type of normalized data precludes using another type of normalized data - comparing minutes to possessions is like apples to oranges - so we would have to re-train the model on th different types of normalized data.

## 2.2 Trends

One of our main hypotheses is that a player's upcoming performance can be predicted by historical career trends of other players. Here, we show a simple visual representation of the three most recorded stats in basketball: points, rebounds, and assists. We took the average measurements of all rookie seasons in recorded history, then all sophomore seasons, and so forth.




These graphs are measured over the first 18 years of a player's career. The longest NBA career has lasted 21 seasons, but the data after 18 seasons is highly suitable to error because of the small number of players who have had careers over 18 years. The X-axis tracks the year

of the career, the Y-axis tracks, from top to bottom, the per-36 points, rebounds, and assists.

We can see a clear trend over a player's career: points peak early in a career, then trends downward, rebounds have a distinct U-shape, and assists plateau around the middle of a player's career. This gives us a good indication that our hypothesis is correct, and if we can

### 2.3 Sources

All of our data was scraped from Basketball Reference (www.basketball-reference.com). We used the Scrapy Python framework to write the web scraper, and MATLAB and Excel to format the data.

## 3 Models

Our models so far have all been trained on per-36 minute data. Further, our models all use the player's number of years in the league as the baseline for comparison (instead of the player's age). We are using the MARS regression algorithm, which is essentially a piecewise linear regression, and comparing it to the standard least-squares regression.

### 3.1 Multivariate Adaptive Regression Splines (MARS)

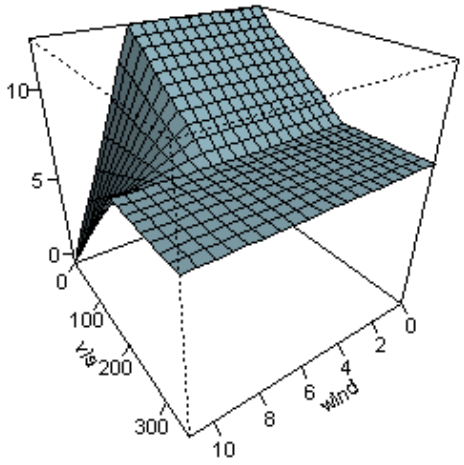



Figure 1: A visual representation of the MARS algorithm; we can clearly see the distinctive "hinges."

The MARS algorithm relies on hinge functions, which are functions of the form $h(x) = max(x_j - t, 0)$ and $h(x) = max(t - x_j, 0)$. $x$ is a vector of length $n$, and $t$ is one of the

observed values in the training set. We will refer to $t$ as a knot.

The MARS algorithm consists of two major phases: the forward pass and the backward pass. In the forward pass we initialize the predictor function to $\hat{f}(x) = \beta_0$, where $\beta_0$ is the mean of Ytrain. We then begin iterating. On each iteration we loop through the set of knots and see which knot reduces error the most. We then add the pair of hinge functions corresponding to this knot to the $\hat{f}(x)$. We stop looping when we have reached the maximum allowable number of terms (set by the user) or when additional terms no longer reduce error. The backward pass goes through each term in $\hat{f}(x)$ and deletes the least effective term until GCV is optimal.

# 4 Results

For each player, we want to use that player's entire career up to the season to be predicted. For example, we use season 1 to predict season 2, seasons 1 and 2 to predict season 3, and so on. Right now, we combine multiple seasons by simply averaging the per-36-minute values across all statistics. We also input into the model the most recent season before the one we want to predict; for example, if we combined seasons 1 and 2 to predict season 3, then we would imput "2" as a parameter as well.

## 4.1 MARS

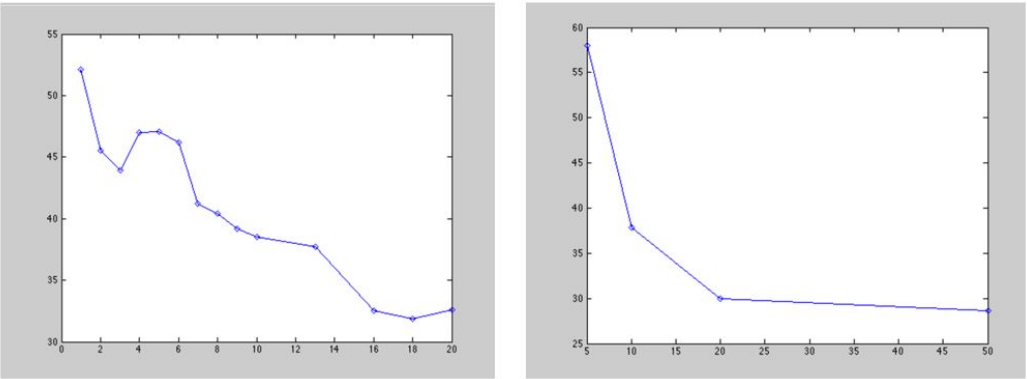



Figure 2: MARS forward pass on points per 36. The model predicts the points output of an upcoming season with the player's previous years' stats as inputs. X-axis: maximum number of terms allowed; Y-axis: mean-squared error. The left image has a smaller X-axis range and the right image a larger; they are otherwise the same model.

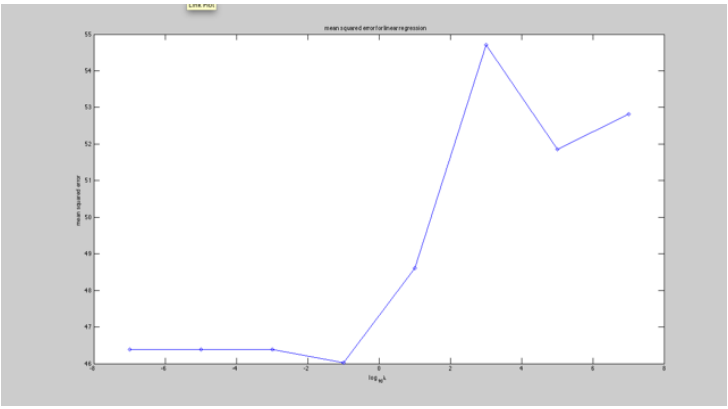


Figure 3: A linear regression with the same properties as the above MARS regression. The X-axis is the log of lambda and the Y-axis is mean-squared error of the predicted points value.

We implemented the forward pass of the MARS algorithm. As we can see in the plots above, the algorithm performs much better than a simple linear regression, and in fact has a MSE of less than 30, which can be considered good performance. We have yet to implement the backward pass of MARS, which would allow us to greatly increase the complexity of the model without producing overfitting, hopefully leading to higher predictive power.

# 5 Project Completion

Many of our goals for completion have been mentioned above, and we want to summarize them here. All of these changes are to the data, not the model, so the implementation cost is expected to be relatively low.

We want to re-run models on data normalized in different ways. So far we have exclusively used per-36 minute data, and we would like to run the models on possession-normalized data and raw data (season totals) as well.

We want to try standardizing by player age instead of player NBA experience. Currently our models use each player's years of experience in the league as the baseline: all players' first years in the league are considered the same, regardless of whether the player was 22 or 32. It may be the case that a player's age is a better predictor of performance trends than a player's NBA experience, since there are many high-quality international leagues and since a player's deterioration has much to do with his age. Re-running the same models with a different baseline is simple, but requires more data manipulation.

We want to output a universal, era-and-position-agnostic player rating that allows us to easily output one error universal error percentage. Many advanced statistics have been created over the past few years; two of the most popular and important ones are player

efficiency rating (PER) and win shares (WS). Our objective is to use the models to predict a player's statistics, then use the PER and WS formulas to output an error percentage. Both of these statistics require league-wide average statistics to calculate, and thus we will have to scrape this new data as well. However, after scraping this data, implementing this formula calculation should not be extremely difficult or time consuming.

Currently, we are using the entire training set to calculate the model parameters. Slicing the training set in different ways may lead to better results. A couple ways we can do this are by:

- Era - certain events in league history have had a great impact on the game and thus our predictive models. The most notable point we want to consider is the introduction of the three-point shot in 1979, which changed the game immeasurably. Training on data only after 1979 may lead to better results.
- League - Prior to 1976, there were two professional basketball leagues in the US: the ABA and the NBA. Each league had its own distinct rules and flavor: for instance, the ABA invented three-point shot and implemented it before the NBA implementation in 1979. We may achieve better results by splitting the training set by league.
- Player Position - the models may have higher predictive power if we split the data by position, for instance if we predict point guard statistics with only historical point guard data. This is intuitive because all five positions on a team vary widely in function and therefore statistical output.
- Player Similarity - if time allows, we want to explore an unsupervised classification of players into categories based on similar performance trends. For instance, some players may have more of a "peaked" performance over their first seven years and would lead to a better trained model for a similar player and other players may have a more "cyclical" career.