

# Predictive Modeling and Statistical Analysis in Sports

Michael Cohen and Matthew Sloane

May 4, 2015

## **Abstract**

In order to fully examine the breadth of possibility in sports data analysis, we look at two areas of differing scope.

The first is scalable and generalizable, exploring the expected performance of athletes in track and field events by means of a data-centric model. It takes Athletic.net data using an in-house web scraper and fits past results against a corpus in order to predict future results in the same race distance. Furthermore, it builds a recommendation system based on composite athlete profiles and displays similar athletes.

The second is more specific, examining the effect of momentum in professional NBA basketball with respect to both players and teams. Momentum is defined via a set of enumerable conditions and its model is created by means of a variation of the same in-house web scraper.

Together, these experimental analyses are represented in the same user interface. As such, they demonstrate the usefulness and applicability of predictive modeling across many subfields within sports-data science.

## **Acknowledgements**

We would like to dedicate this paper to Professor Melanie Wu for her guidance and support, as well as Kenton Freemuth and the Pomona-Pitzer Men's and Women's Track and Field teams.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Our Project . . . . .	5
1.2	Background . . . . .	6
1.2.1	Player Analytics . . . . .	7
1.2.1.1	Fitting Performances . . . . .	9
1.2.1.2	Quantification . . . . .	12
1.2.2	Data Sources . . . . .	13
1.2.2.1	Predictions . . . . .	15
1.2.2.2	Evaluation of Qualitative Concepts . . . . .	17
1.3	Contributions . . . . .	21
<b>2</b>	<b>System Overview</b>	<b>23</b>
2.1	Scraping . . . . .	23
2.2	Parsing . . . . .	24
2.3	Cleaning . . . . .	25
2.4	Normalization . . . . .	25
2.5	Archiving . . . . .	27
2.6	Data Recovery . . . . .	27
<b>3</b>	<b>Track and Field</b>	<b>29</b>
3.1	Design and Interface . . . . .	30
3.2	Implementation . . . . .	33
3.2.1	Prediction . . . . .	34

3.2.2	Recommendation . . . . .	39
3.3	Results . . . . .	44
3.3.1	Prediction . . . . .	44
3.3.2	User Study of Recommendation . . . . .	48
3.4	Discussion . . . . .	53
<b>4</b>	<b>Basketball</b>	<b>55</b>
4.1	Design and Interface . . . . .	56
4.2	Implementation . . . . .	57
4.3	Results . . . . .	60
4.4	Discussion . . . . .	64
<b>5</b>	<b>Conclusion</b>	<b>66</b>
	<b>Bibliography</b>	<b>68</b>

# 1 Introduction

## 1.1 Our Project

Our project investigates the realms of predictive modeling, recommendation systems, and statistical analysis within sports.

The first two areas are covered using track and field data from Athletic.Net, since track and field lends itself to structured data and tangible predictions by its very nature. Furthermore, because there are no set similarity measures between track and field athletes, it becomes an intriguing challenge to produce clusters of similar athletes. In both cases, k-means clustering and significant preprocessing is used as the underlying mechanism. The predictive model is evaluated through actual fourth-year personal records, while the recommendation system is evaluated through questionnaires completed by actual varsity track and field high school and collegiate athletes.

The third area is explored through basketball data from BasketballReference.com. It analyses the effect of momentum in professional basketball games at both the player and team levels. The performances of players and teams in games is contextualized against their overall field-goal levels, among other statistical measures. Regressions are performed that examine whether a player performs better or worse after making consecutive shots, and after missing multiple shots in a row.

Together, these three components are wrapped up in a single user in-

terface that prescribes to modern design principles. It is flat, sleek, and minimalistic, and illustrates the feasibility of making sports data analytics tools accessible to parties worldwide.

## 1.2 Background

The field of sports data analytics has a wide range of use cases, from evaluating player performances to predicting the likelihood of injury. Recent celebrated work includes Michael Lewis's book *Moneyball*, which addresses the question of how the Oakland Athletics baseball team achieved such great success with one of the league's smallest bankrolls. One of the key conclusions was that players are often credited for things outside of their control. For example, a pitcher's win-loss ratio is largely an indicator of how well his surrounding teammates perform, but this fact is often overlooked [Dav06].

There is also a significant amount of available data by virtue of the statistics collected and tracked by third parties. For example, the National Basketball Association (NBA) records basic information such as player points per game, number of assists, and number of minutes per game, but also more advanced statistics like evening home game wins, net points scored while on the court, and player efficiency ratings. This is particularly true in the United States, but is also becoming a factor overseas in places such as the English Premier League (EPL), where player pass accuracies are recorded alongside visual depictions of shots taken, and by whom.

Another breakthrough in advanced analytics has recently occurred due

to the installation of cameras in arenas, currently used in both the NBA and EPL. Players' x and y coordinates and the ball's x, y, and z coordinates are tracked by cameras from a number of angles, and their locations are tracked at very high speeds (twenty-five frames per second in the NBA and ten frames per second in the EPL). This allows for the game to be analyzed in ways not previously possible with only box score and play-by-play data [Dav14].

Analytical strategies are shared between different teams largely due to the high turnover rate among coaches and managers. This allows for rapid progression and implementation of various statistical analyses, which is partially responsible for the boom in the sports analytics industry over the past decade. One example that illustrates this lies in the MIT Sloan Sports Analytics Conference, an annual event that discusses recent developments in sports data analytics. In 2007, there were 175 attendees. However, in 2013, there were over 2200 attendees; this is an increase of over 1200 percent [Wil14].

Historically, predictive sports modeling has been accomplished through mathematical, theoretical models, based on human intuition and other primitive means. However, with the recent technological advances in modern analytics, opportunities have arisen for a transition into data-driven modeling.

### **1.2.1 Player Analytics**

Performance modeling and analysis have a vast number of capabilities, namely for player evaluation. With this knowledge, a player can be examined

within the context of his or her team and league. Managers and coaches can subsequently use this information to make educated decisions.

One such use case is regarding the optimization of salary for a given player. That is, given a player's yearly performance, has he or she met expectations? Is he or she a financially viable investment? In order to answer these questions, it is necessary to develop metrics through which to compare and measure players.

Another feature is player evaluation for the draft. In leagues such as the the National Football League (NFL), the National Hockey League (NHL), and the NBA, the draft is a yearly event where players hoping to compete professionally are sequentially chosen by teams from around the league. Their history and development are scrutinized by analysts and coaches in the months leading up to the draft. Websites such as Draft Express exist solely to provide information about incoming talent by means of scouting reports, statistics, blogs, and relevant announcements [Exp14].

A third reason behind player analysis is due to the usefulness of taking opponent game tactics and finding ways to undermine and overcome them. Optimal starting player lineups can be determined on a game-by-game basis based on the strategies employed by particular teams. Past results can also be analyzed in the context of utilized tactics in order to adapt and learn from previous mistakes.

The final and most extensive area of study is descriptive player analytics, taking players from around the league and describing them through various



features and extensions. For example, the NFL's quarterback passer rating was developed by means of several existing statistics, such as yards gained, number of touchdowns, and number of interceptions. Another metric, the aptly-named assist-to-turnover ratio, is widely used to help illustrate a point guard's performance in a given basketball game [Dav14].

#### **1.2.1.1 Fitting Performances**

Current player performances help orient a team and adjust gametime strategies and team management composition in the here and now. However, another worthwhile realm is the field of evaluating future player performances. Here, it becomes possible to answer the following questions with high probability: will this player be producing in five years time? What is this player's ceiling in terms of overall potential? How will this player perform in an alternate environment? How will some environment affect the player or teams performance? How can we determine a player's performance trajectory?

This is the science of fitting athletic performances and estimating future results. One of the most valuable aspects of predictive analytics lies in its cost-saving ability. Players that are likely to sustain career-ending injuries can be avoided, and low-risk players can be sought after, resulting in benefits in the long run.

Significant time and effort has been dedicated to exploring the hypothetical limitations of world-class Olympic athletes. In the 1960s, researcher Michael Deakin presented a mathematical model in order to estimate an

upper-bounds of athletic performance in the mile [Hea06] [Hop05]. One decade later, Gerry Purdy formulated the Purdy Points model, which analyzes athletic performances and attempts to contextualize them amongst themselves. In other words, it looks at races of different distances and assigns relative values [Dea67].

More recently, work has been done not to ascertain Olympian limitations, but rather the limitations of individual athletes. This often involves taking performances of similar athletes and formulating a model in order to make sound predictions [Whi07]. Given the statistical nature of track and field, where results are cleanly measured in meters, minutes, and seconds, how can we measure the progression of athletes?

Over the years, there have been many attempts to evaluate and make predictions regarding track and field athletic performances. There are several mathematical models that extrapolate on past marks and results in order to predict new ones. The Riegel model is one such formula [Whi07]. It was first introduced in an edition of the *Runner's World* magazine in 1977. It was stated as follows:

$$\text{new time} = \text{original time} * \left(\frac{\text{new\_distance}}{\text{original\_distance}}\right)^{1.06}$$

One of the reasons for the popularity of this formula lies in its clarity. It says that doubling the distance of a race will result in a speed decrease of approximately four percent. However, it does make several simplifying assumptions, which are detrimental to the model. One assumption is that the athlete is not biased towards either speed or endurance. This means that

the model might not fit a sprinter or a marathoner as well as it would for a five kilometer or ten kilometer runner. Another reason is that the model is best for races of a duration between three and a half minutes to four hours. This is a similar assumption, but better quantifies the model’s limitations.

Alternatively, the VO<sub>2</sub> Max Model, developed by researchers Jack Daniels and Jimmy Gilberts in 1979, uses oxygen consumption, time, and velocity in order to predict results for distances [DG79]. Their model is as follows:

$$\text{VO}_2 \text{ Max} = \frac{-4.60+0.182*velocity+0.000104*velocity^2}{0.8+0.189e^{-0.0128*time}+0.299*e^{-0.193*time}}$$

Here, the greatest assumption is that the race distance is not primarily anaerobic, since oxygen capacity and consumption relates to aerobic exercise. Therefore, the model is most applicable to races over eight-hundred meters. Since velocity is a measure of distance over time, an athlete’s VO<sub>2</sub> Max and a new distance can be used in order to predict a duration. Other models such as the McMillan running calculator are commonly used, but their measures are not publicly released and therefore are less useful for our own purposes [Whi07].

Still, the majority of these models share a common detail: they were all developed several decades ago, and they handle extrapolating to different race distances. As illustrated in Section 1.2, the recent abundance of data has allowed for a new, practical approach: data-driven modeling - this will enable us to investigate predictions on already-competed race distances using progressions of existing athletes.

### 1.2.1.2 Quantification

Quantification is the act of counting and measuring that maps human-sense observations and experiences into members of some set of numbers [Wiki09]. Tests in school, for example, are an attempt to quantify a student's performance in a given subject. Similarly, sports events quantify how productive an individual or group of individuals is at a given activity.

Proper quantification of a given principle or concept is essential to analysis of any field. Answering compelling questions almost always involves a non-trivial quantification process. A common problem is deciding how to project human output. In the case of sports, this could manifest itself as a team deciding whom to draft or for whom to trade. Regardless of the specifics, however, there are two things that need to be defined in order to properly evaluate these kinds of questions:

1. a set of features or criteria that are either already quantified or can be further broken down into quantified data
2. an algorithm that takes said quantified criteria as input and returns a value in the pre-specified domain

Momentum has been considered a significant factor in the outcome of sports contests by players, coaches, and fans since long before the recent big data revolution. Without underlying data to back it up, momentum was something that was felt by viewers and athletes, but not necessarily well-defined. After a sequence of good plays by team A and a sequence of bad plays by team B, there is thought to be momentum in favor of team

A. Or, after an individual player makes several consecutive good plays, he or she is thought to be on a roll. Whether looking at it from a team's or individual's perspective, acknowledging that there is momentum implies that the expected success rate in the upcoming plays is higher than under normal circumstances.

The notion of momentum has been explored heavily in sports. While there is a general set of features that are universally used to determine momentum, turning these into specific criteria that everyone agrees upon remains challenging. A study with certain criteria may give evidence towards one belief, while another study with different criteria can yield potentially contradicting results. Another issue is the existence of dependencies within the feature set. The feature set should take a certain form if momentum exists and take another form if momentum doesn't exist, without being influenced by external factors. We will discuss previous research in sports momentum further in section 1.2.2.

### **1.2.2 Data Sources**

The digital universe contains over 4.4 trillion gigabytes of data and is increasing at an exponential rate. In two years, this number is projected to double [Sol14]. Therefore, data-driven modeling has become more and more feasible as data becomes more readily available.

Websites such as Basketball Reference and Pro Football Reference are dedicated to maintaining data stores of everything from player shots to

points to even play-by-play notes. Hence, they possess massive databases of statistics that can be easily utilized and repurposed by the average user [Dav06]. In fact, Basketball Reference does not prevent web scraping and there are numerous web scrapers on Github and other source code repositories that can accomplish most any data collection task from the website.

Still, Basketball Reference produces more than just accessible, structured data. For example, it has an in-house calculation system that determines similarity scores between NBA players. It does not attempt to deduce similarities in playing style — rather, it looks at the similarity of shape and quality between two players' careers. In other words, if two players started off with three productive years, followed by two years in a slump, they are more likely to be categorized together. This classification is done in a position-by-position basis [Ref14]. These similarity scores are made available by the website. Yet, the majority of the website is devoted to simply compiling structured data on games, players, and teams. Data analysts are able to utilize this data through methods such as web scraping and the analysis thus becomes a more collaborative process. It is then possible to efficiently perform more complex analyses and determine predictions and patterns.

This is the crux of data-centric modeling. There are a number of online databases with easily obtainable content. As mentioned above, Basketball Reference has a number of statistics and metrics that allow data analysts to extrapolate and make groundbreaking connections. Other data sources, such as Athletic.Net, have results for over 4.9 million high school athletes, and are chronologically ordered in a manner conducive for predictive mod-

eling [Ref14]. With the relative ease of data collection and the abundance of available data points, it is feasible to formulate a model that can make accurate predictions.

### 1.2.2.1 Predictions

One of our fields of study in predictive sports analytics is scalable and generalizable, exploring the expected performance of athletes in track and field given previous results. The second field of study examines the effect of momentum in professional basketball at both the player and team level. From there, additional predictions can be made: for example, given a player's momentum at a specific point in the game, what is the likelihood that he or she will make the subsequent shot? Together, these experimental models demonstrate the breadth and utility of predictive analysis within sports data science.

Athletes are constantly compared to other athletes when referenced by commentators across all sports. Therefore, it seems logical to use the statistics of many athletes in order to predict those of another. For example, if Athlete X appears to be following in the footsteps of Athlete Y, then maybe it would make sense to use Athlete Y's statistics as a baseline for Athlete X's future performance.

In the above example, though, the sample size is small. Perhaps Athlete Y is an anomaly – suppose he or she was severely injured later in his or her career – and in actuality, most athletes will progress differently. Thus,

by using a much larger data set, we can apply simple linear regression on subsets of the data. For example, to fit an athlete’s 200m performance, we can use athletes with times within a certain qualifying range (meeting predetermined heuristics) and look at their progressions in order to calculate an expected value for the original athlete.

“Runner’s Log and Predicting Performance Analytics,” a paper by Alexander White of the Worcester Polytechnic Institute, attempted to do something similar to the above: use existing data to predict future results of track and field athletes. White’s paper took pairs of performances with the similarity measure being within 0.05 meters/second of the input speed and then averaged the output pairs in order to produce a predicted output for a new race distance. When compared against existing models such as Purdy Points, Riegel, and the VO2 Max Model, this simple average beat them all, resulting in an average error of 3.14% for men and a 3.57% average error for women. For context, the Riegel model had an average error of 7.32% and 6.80% for men and women, respectively [Whi07].

There seems to be significant value in utilizing a data-centric model in order to predict athletic performances in track and field. However, White’s model only incorporated 186,687 running performances (a later version employed almost 325,000 performances). It also only used athletes from specific states, mainly the New England region, and gathered data from a single source, DirectAthletics. Our project will synthesize data from multiple sources, take data from a wider breadth, and use well-defined heuristics in order to increase the quantity and quality of the data set and attempt to



reduce error. However, it will investigate predictions for already-competed distances instead of extrapolating for new ones.

#### **1.2.2.2 Evaluation of Qualitative Concepts**

In the context of data modeling, it is generally sufficient to develop a set of criteria or threshold as a definition of a qualitative concept. For example, given the qualitative notion of a “clutch” basketball player, the requisite criteria might be as follows:

1. a clutch instance begins with five minutes remaining in a game if
  - (a) the score differential is at most ten points
  - (b) the player is in the game for 70 percent of the remaining game time
2. The clutch instance is deemed positive when occurs when the player
  - (a) successfully makes at least seventy percent of his or her shots
  - (b) scores at least six points
  - (c) commits at most one turnover
3. A clutch player is a player who
  - (a) has a positive clutch instance in at least 60 percent of all possible clutch instances

It then becomes fairly straightforward to develop an algorithm that examines the available data in search of possible clutch instances that fit the

criteria above. The player’s achieved clutch instances can then be counted and an overall determination of clutchness can be accomplished.

However, it is imperative to develop criteria that are in line with the qualitative definition of a given term. For example, if we defined “clutchness” to be simply scoring twenty points in a game, then it would be difficult to get any outside approval of the study. Clutchness has a timing factor (must be at the end of games) and an importance factor (must be high-stakes). Any study performed without these two criteria would not be getting an accurate reading on the essence of clutchness.

Momentum is another notion that fits this general model. It is commonly believed to have a real effect in individual and team performance in sports. This is especially true in basketball, in which the notion of a “hot-hand” is thought to exist by players and fans alike. We define hot-hand to be the idea that a player’s likelihood of making his/her next shot is higher following a made shot than following a missed shot.

The hot-hand has spurred a lot of research and experiments in the fields of both data analysis and psychology. In one of the foundational papers in this field, Gilovich et al. [GVT85] tracked the performance of basketball players at Cornell University in controlled shooting experiments. They also incorporated a betting component to their experiment, where the participants could either bet high or bet low immediately preceding shot attempts. In this way, they could quantify a perceived momentum effect, even if the actual shooting data did not back up any such actual effect.

Gilovich et al. used conditional probabilities to determine how well a player shot given previous performance — after either 1, 2, or 3 makes or 1, 2, or 3 misses. Furthermore, they tracked the number of runs — defined as each streak of consecutive makes or misses — that a player had, and determined how different this number was from its expected value given the number of shots taken.

They concluded that while the players did believe in the hot-hand, as evidenced by their betting throughout the experiment, there was corroborating evidence in the actual shooting data. This falls in line with the consensus among research in this field — that the hot-hand is a cognitive illusion that has no real effect outside of human perception.

However, a basketball game is much more complex than just shooting. There are other types of plays that can have profound effects on, at the very least, within which team the perceived momentum lies. Such events that positively affect momentum include 3 point shots (and 2 point shots to a lesser extent), steals and turnovers for the other team, while events that negatively affect momentum include missed shots and turnovers [BBJ99].

Mace et al. [MLSN92] performed a study which had trained observers watch several basketball games and track three types of events: reinforcers, adversities, and response to adversities. We may define reinforcers as positive events, adversities as negative events, and response to adversities as the result of the first offensive possession after an adversity. Mace et al. determined that a team's response to an adversity generally increased as the

rate of reinforcement increased in the time (specifically 3 min) preceding the adversity. This shows that, although the sample size was small, momentum had a notable effect performance, at least when taking other events into account beyond just shooting.

However, more recently, the hot-hand effect has been found to exist even in an isolated shooting environment [MS14]. While this isn't enough to render irrelevant decades of other research, it does show that there is clearly more work to be done on this topic and that new methods of quantification are worthy of exploration. This project will deal only with actual NBA game data and try to determine how players and teams react to positive and negative stimuli in games. It will attempt to answer questions such as: to what extent does momentum affect the performance of players or teams? Is there any correlation between players and teams that bounce back effectively from negative momentum and the success of those players and teams?

We will look at these questions in two different ways: once in the context of only shooting data and another in the context of all positive and negative events— shooting, steals, turnovers, etc.

### 1.3 Contributions

This project achieved significant results in each of its three components — predictive modeling, recommendation, and statistical analysis. Through data-driven modeling, we achieved both a quality prediction system and recommendation system in track and field. Through the k-means clustering algorithm, we were able to achieve accurate predictions while gleaning a number of insights regarding track and field performances, regarding areas such as gender, event, and race frequency. Most notably, the prediction system achieved an average percent of error of 1.38% in the mens 400 meter dash in outdoor seasons. The recommendation system also performed admirably with k-means clustering, yielding an average user satisfaction rating of 7.79/10 as well as a high innovation rating, with 89.5% of the survey participants stating that they have never seen another recommendation system. This begs the question of when a tool will occupy this space, and to what extent this recommendation system is a step in the right direction.

In terms of statistical analysis in basketball, while some of the results were promising, there was not enough corroborating data to make any strong claims about the effects of momentum on in-game shooting. However, analysis on certain players supports the notion that shooting efficiency may actually drop after making previous shots. Yet, it is possible this result came about due to a flaw in the model - not accounting for shot distance or difficulty. This can now be quantified by metrics using player-tracking data such as closest defender distance and closest defender height.

Regardless, there is room for further development. With track and field, higher-quality results could be seen with a larger data set. Other data sources such as TFRRS (Track and Field Results Reporting System) would also be an ideal source, given permission. In addition, cross-country would be a feasible sport to extrapolate towards. Finally, there are more sophisticated machine learning techniques that could prove useful.

For basketball, other factors such as time passed between shots and substitutions could be potential catalysts for momentum. Adding a machine learning element of results clustering in order to group basketball players with similar shot streaks and patterns could also prove fruitful, as well as increasing the size of the dataset by scraping and analyzing additional NBA seasons.

Altogether, we have produced work in three areas - predictions, recommendations, and statistical analyses - utilizing solely freely available data, and our success indicates that there is a wealth of potential for future work.

## 2 System Overview

Our system involved several different phases to get from the data collection stage to being ready for analysis. The basketball data and track and field data were extracted from BasketballReference.com and Athletic.net, respectively. Then, before it could be fed into our analysis, the data had to be parsed, cleaned, normalized, and stored in a database. Each phase is described in detail below.

### 2.1 Scraping

We used the Node.js request module to retrieve intact HTML files, which were then stored in raw-HTML tables for both Athletic.net and BasketballReference.com so that incorrect parsing would not necessitate re-scraping from the original websites. For track and field, athletes were conveniently accessible through integer indexing in the URLs. For basketball, it was imperative to iterate through web pages, one per day in the season, and extract links to play-by-plays, which were scraped in turn.

Play-By-Play			
Jump to: <a href="#">1st</a>   <a href="#">2nd</a>   <a href="#">3rd</a>   <a href="#">4th</a>   <a href="#">scoring play</a>   <a href="#">tie</a>   <a href="#">lead change</a>			
1st Quarter			
Time	Cleveland	Score	LA Lakers
12:00.0	Start of 1st quarter		
12:00.0	Jump ball: <a href="#">R. Sacre</a> vs. <a href="#">A. Varejao</a> ( <a href="#">P. Gasol</a> gains possession)		
11:44.0		0-0	Turnover by <a href="#">W. Johnson</a> (bad pass)
11:36.0		0-0	Shooting block foul by <a href="#">R. Sacre</a> (drawn by <a href="#">L. Deng</a> )
11:36.0	<a href="#">Deng</a> makes free throw, 1 of 2	+1	1-0
11:36.0	<a href="#">Deng</a> makes free throw, 2 of 2	+1	2-0
11:28.0		2-0	Turnover by <a href="#">K. Marshall</a> (bad pass; steal by <a href="#">L. Deng</a> )
11:25.0	<a href="#">Deng</a> makes 2-pt shot from 2 ft	+2	4-0
11:13.0		4-2	+2 <a href="#">P. Gasol</a> makes 2-pt shot from 10 ft (assist by <a href="#">K. Marshall</a> )
10:54.0	<a href="#">L. Deng</a> misses 2-pt shot from 1 ft	4-2	
10:53.0		4-2	Defensive rebound by <a href="#">P. Gasol</a>
10:47.0		4-2	<a href="#">W. Johnson</a> misses 3-pt shot from 23 ft
10:44.0	Defensive rebound by <a href="#">C. Miles</a>	4-2	
10:35.0	<a href="#">Varejao</a> makes 2-pt shot from 7 ft	+2	6-2
10:24.0		6-4	+2 <a href="#">P. Gasol</a> makes 2-pt shot from 20 ft (assist by <a href="#">K. Marshall</a> )
10:09.0	<a href="#">A. Varejao</a> misses 2-pt shot from 16 ft	6-4	
10:07.0		6-4	Defensive rebound by <a href="#">K. Marshall</a>
10:02.0		6-6	+2 <a href="#">W. Johnson</a> makes 2-pt shot from 2 ft (assist by <a href="#">K. Marshall</a> )
9:36.0	<a href="#">L. Deng</a> misses 2-pt shot from 18 ft (block by <a href="#">W. Johnson</a> )	6-6	
9:35.0	Offensive rebound by <a href="#">A. Varejao</a>	6-6	
9:35.0	Turnover by Team (shot clock)	6-6	
9:20.0		6-8	+2 <a href="#">P. Gasol</a> makes 2-pt shot from 11 ft
9:06.0	Offensive foul by <a href="#">C. Miles</a> (drawn by <a href="#">R. Sacre</a> )	6-8	
9:06.0	Turnover by <a href="#">C. Miles</a> (offensive foul)	6-8	
8:51.0		6-8	<a href="#">K. Marshall</a> misses 3-pt shot from 27 ft
8:50.0	Defensive rebound by Team	6-8	
8:50.0	Cleveland full timeout	6-8	
8:35.0	<a href="#">Deng</a> makes 2-pt shot from 2 ft (assist by <a href="#">A. Varejao</a> )	+2	8-8
8:23.0	Violation by <a href="#">A. Varejao</a> (kicked ball)	8-8	
8:18.0		8-8	Turnover by <a href="#">P. Gasol</a> (bad pass; steal by <a href="#">C. Miles</a> )
8:10.0	<a href="#">C. Miles</a> misses 3-pt shot from 24 ft	8-8	
8:09.0	Offensive rebound by <a href="#">A. Varejao</a>	8-8	

Figure 2.1: A sample play-by-play, courtesy of basketball-reference.com

## 2.2 Parsing

Parsing involved taking the data from the raw html tables and using the Node.js cheerio module to extract desired fields. For example, attributes such as gender, name, season type, level, and actual performances were taken from individual track and field profiles. A table of athlete names and characteristics was created, alongside another table for specifically for track and field event marks. For basketball, as shown in Figure 2.2, play-by-plays were parsed into database-readable formats, along with game titles, event types, team names, and player names.



```

▼ <tr>
  <td class="align_right">9:20.0</td>
  <td>&nbsp;</td>
  <td>&nbsp;</td>
  <td class="align_center background_yellow">6-8</td>
  <td class="align_right background_lime">+2</td>
  ▼ <td class="background_lime">
    <a href="/players/g/gasolpa01.html">P. Gasol</a>
    "makes 2-pt shot from 11 ft"
  </td>
</tr>

```

Figure 2.2: The source code for one row of a play-by-play page

## 2.3 Cleaning

Cleaning the data involved removing extraneous whitespace from the parsed attributes and also removing dirty data from the parsed-HTML tables. Although BasketballReference.com contained relatively clean data, the manual data entry format of Athletic.net yielded substantial human error. Gibberish track and field marks such as “ajjfdfsd” and clearly incorrect marks such as 99:99.99 for a 100 meter dash were removed from the database. Furthermore, due to the wide variation in mark submission by coaches, the parsing occasionally returned an empty string — the error handling result — and cleaning the data involved removing these entries as well.

## 2.4 Normalization

Finally, the data was normalized in preparation for feeding into the algorithms and statistical analyses. For track and field, this involved converting all marks into inches or seconds, depending on the event. Dimension ta-

Farwell HS		2013 Outdoor Season - 9th Grade	
<b>200 Meters</b>			
5	28.60a	Jv F May 6	Farwell JV Invite
<b>400 Meters</b>			
4	65.38a	V F Apr 24	JPC Houghton Lake @ Fa...
	SCR	V F May 7	JPC Roscommon-Meridian...
<b>3200 Meters</b>			
5	81:05.00	V F May 13	Harrison 9/10 Meet
<b>Discus - 1.6kg</b>			
4	71' 0.00	Jv F May 6	Farwell JV Invite
6	81' 2.00	V F May 7	JPC Roscommon-Meridian...

4x200 Relay	
2	1:52.00 FS F Apr 27 Gladwin vs. Rosco and ... Austen Weaver Lucas Buccilli Justin Norburry Damon Scheidt
5	1:52.80 FS F May 4 white cloud jr high in... Damon Scheidt Austen Weaver Justin Norburry Lucas Buccilli
999999	FS F May 8 Clare @ Farell Austen Weaver Kyle Hurley Damon Scheidt Justin Norburry
4	1:50.90 FS F May 15 Jack Pine Conference M... Austen Weaver Damon Scheidt Lucas Buccilli Justin Norburry
5	1:52.31a V F May 30 MEGASTAR Lucas Buccilli Justin Norburry Austen Weaver Damon Scheidt

(a) An incorrect 3200m time      (b) A mistyped 4x200m mark

Figure 2.3: Dirty marks due to human or system error

bles were also created to improve query speed, holding values such as the distinct events and mark types — automatic, converted, or wind-aided, to name several. For basketball, this meant matching players to teams given their first initial and last name. Once the data was normalized, the data collection process was complete.

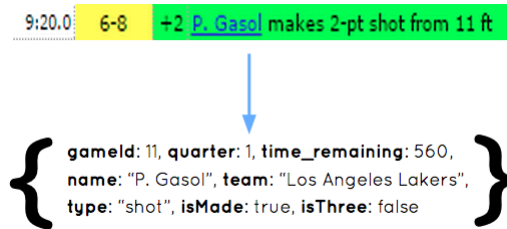


Figure 2.4: The normalization process for one play-by-play event

## 2.5 Archiving

Data was organized in a PostgreSQL database. We employed the Node.js pg module to manage the insertion and deletion of data, as well as table creation and updating. For the raw-HTML table, the URLs were stored as the primary keys, with an HTML object associated with each URL stored as a text object since all text types in PostgreSQL are saved the same way under the hood. For the parsed-HTML table, objects were established that encapsulated relevant data — this process was accomplished separately for track and field and basketball. Steps such as cleaning and normalization were more trivial and involved direct manipulation of the parsed-HTML tables in lieu of creating new, cleaned-HTML or normalized-HTML tables.

type	gid	pname	team	quarter	time_remaining	is_made	is_three
shot	11	P. Gasol	Los Angeles Lakers	1	560	t	f
shot	11	L. Deng	Cleveland Cavaliers	1	515	t	f
shot	11	N. Young	Los Angeles Lakers	1	406	f	t

Table 2.1: 3 sample rows from events table in database

## 2.6 Data Recovery

The final step in the data collection process was to perpetually store backup copies of the database in case of hard drive failure. This was accomplished via the built-in PostgreSQL pg\_dump command, which creates consistent backups even when the database is being accessed. For track and field, due to the relatively large database size of 150 gigabytes, backups were

kept on an external hard drive and replaced on a three-spot rotational basis. For basketball, backup database copies were stored on the Cloud without purchasing additional storage space.

### 3 Track and Field

It is surprising that while the field of sports data analytics has enjoyed a rise to the forefront in recent years, track and field, a sport that lends itself to statistical analysis by its very nature, has remained relatively unchanged. Perhaps the reason lies in the simplicity of the sport. In basketball, measures such as points per game, assists, rebounds, and blocks are combined to evaluate a player's assets. However, in track and field, there is no confusion — the athlete with the best record in his or her event should be valued highest.

Still, there are multifold ways to leverage the existing structured data of track and field performances to create practical tools. With a track and field athlete and his or her slate of performances, there are three areas to consider.

1. What factor(s) are responsible for the athlete's past/present performances?
2. Given past performances, how will this athlete perform in the future?
3. Who are athletes that have performed comparably to this athlete?

That is, based on times and marks, it is fairly straightforward to ascertain a given athlete's current performance level. Predictions and parallels, however, are more difficult. Therefore, this paper explores how an abundance of track and field performance data can be utilized in an attempt to

garner both predictions — by means of athletes that have progressed similarly — and comparisons, at the profile level. This involves examination of athlete marks and times by event, over the span of a season, and again over the course of an entire career. The first question, exploring what factors may have contributed to an athlete’s performances, is not covered at this time due to the lack of athlete metadata. The second and third questions, though, are explored and wrapped up in a user interface to underline the feasibility of widespread usage of this tool and others like it.

In the following sections, the design decisions of the project, the user interface, the implementation, and the desired use cases are each considered.

### **3.1 Design and Interface**

The track and field project addresses the two notions of generalization and simplicity and their entwinement and tradeoffs. It is evident in both the user interface and the algorithmic implementation itself.

For instance, the user interface follows modern design principles and shows a flat, minimalist display. The default interface has two buttons, leading the user to either a predictor or a comparator. If the user selects the predictor, he or she need only input his or her gender, select between a collegiate or high school level, choose between an indoor/outdoor season type, and pick an event, along with his or her top eight marks in said event. A simplified version of the predictor that solely accounts for athlete personal records (PRs) is shown in action in Figure 3.1 below:

### Track & Field Mark Predictor

Gender:  
 Male       Female

Season:  
 Indoor       Outdoor

Event:  
 200 Meters       1600 Meters

Level:  
 High School       Collegiate

Season #1 PR:

Season #2 PR:

Season #3 PR:

Figure 3.1: Predictor user interface where the user can select desired criteria and input PRs

### Predicted Mark: 22.91

75%: 22.72 to 23.04  
85%: 22.61 to 23.14  
95%: 22.45 to 23.25

\*estimates above based on model

[BACK](#)

Figure 3.2: A sample output from a user submission to the predictor

The result is a predicted mark and ranges with probabilities according to our model as shown in Figure 3.2. For example, the men's 200m dash might show a range of 22.51 - 23.04 with a likelihood of 90%. Alternatively, if the user chooses the comparator, he or she then searches by name to display his or her track and field profile from the database. The athletic profiles of similar athletes are shown on the interface as well, with color and font highlights representing degrees of similarity. Yellow might mean that the athletes have run one event together with comparable marks, while orange represents athletes with more similar events or more closely related marks.

The above functionality allows the user to experiment with the predictor and comparator while needing relatively few clicks to do so. In this sense, it is a simple tool with explicit instructions fitting neatly on screen at one time. Moreover, the predictor setup is significantly generalizable as well. As long as the user can recall his or her top eight marks in a given event over the course of his or her first three seasons, the predictor will function normally. There is, however, a tradeoff when considering the comparator. Here, it is necessary to have an Athletic.net profile in the database, as the composite history of an athlete is used to locate similar ones. While it would have been possible to allow the user to type in his or her entire athlete profile, this extension was not included because it would have greatly increased the complexity of the tool. Future versions may draw from other websites' track and field profiles or have the option to "create a profile".

Internally, the design principle of keeping the tool simple while keeping it generalizable is also at play. The same algorithmic framework of k-means



clustering is used for both predicting and recommending similar athletes. In addition, similar feature sets are used as input for the algorithm. That is, using the top eight marks in an event for the predictor is not event-specific, gender-specific, level-specific, or season-type-specific. This provides consistency for the user and a much cleaner algorithmic implementation. The tradeoff here is that using similar feature sets for every event might slightly lower the quality of the results. Yet, given the usefulness of the output in its current state and the increased elegance of the design process, the simplicity remains.

In the same vein, athletes are compared using a generalized notion of similarity that will be further discussed in the implementation section. Here, although quality is still a factor, time plays a significant role as well. Determining which athletes to include in the clustering process can be a time-consuming process if it is not kept relatively simple. Additionally, athletes have run anywhere from one to four years in a wide capacity of events, with notably varying marks. Finding matches for any athlete profile is therefore an undertaking that must inherently be generalized, to an extent.

## **3.2 Implementation**

The predictor and recommender portions of this recommendation system both utilize k-means clustering, albeit in quite different manners. In order to calculate predictions, athletes that meet predetermined criteria are clustered together in a specific event by performances of their first three years. The cluster that most closely aligns with the input athlete is then further

scrutinized: the fourth-year personal records of those athletes are averaged in order to formulate a prediction for the athlete at hand. On the contrary, there is no labeled data for the recommendation system in terms of predefined athlete similarity. Athlete profiles are measured solely on the basis of their content across multiple events and years. This requires significant post-processing to further refine search results.

### 3.2.1 Prediction

The primary machine learning algorithm behind the predictor is the k-means clustering algorithm. It was chosen for its ability to handle diverse feature sets and its potential to lump together athletes with atypical progressions - for example, someone who improved for his or her first two seasons, but worsened in the third season. Furthermore, while the athletes are labeled in the sense that they have fourth-year personal records that are used to predict the person record of an input athlete, they are not labeled insofar as the athlete profile is concerned. K-means clustering allows for the exploration of the effect of differing features on an athlete's fourth-year performance by means of grouping them together and discerning previously hidden relationships.

For a user querying the predictor on the user interface, it is necessary to first run the k-means clustering algorithm on the training set for the input criteria - gender, event, level, and season type. The user's feature set's euclidean distance from each centroid is then measured, with the mean of the closest cluster's fourth-year personal records being returned as a prediction

for the input athlete. Then, the percent error from the testing set is used to return a series of ranges to the user, with confidence intervals based on the percentage of athletes in the testing set that fit within some range of percent error.

The first major question to address was which athletes to extract from the database, given a gender, event, level, and season type. This issue is similar to determining which athletes to include in the training, validation, and test sets. Due to the large quantity of athletes — around seven million profiles — the IDs of desired athletes were stored in intermediary tables on a gender, event, level, and season type basis. Athletes were chosen that:

1. Ran all four years in the specific event, gender, level, season type
2. Ran at least four times per year in the specific event, gender, level, season type

The first requirement ensured that athletes in the data set would have three years to include as characteristics of the feature set, and a fourth-year event personal record to use for the prediction. The second requirement set a reasonable standard of accuracy for the feature set data. For example, an athlete's personal record is much more likely to be reflective of his or her capabilities if he or she has run multiple times, rather than just once. Collectively, these requirements substantially reduced the quantity of the data set. In the men's high school outdoor 200 meters, for instance, there were 1.2 million athletes that ran the 200 meters at least once, but only 7000 that ran it all four years, and merely 2000 that ran at least four times

in each year. Setting the threshold at four instances per year allowed for a reasonable size for the training, validation, and testing sets, split at 60-20-20 percentages, but in order to evaluate the effect of the threshold, the clustering analysis was also carried out on lower thresholds as well, increasing the size of the data set but reducing the accuracy as the meaningfulness of a personal record was diminished. These requirements are depicted below in Figure 3.3:

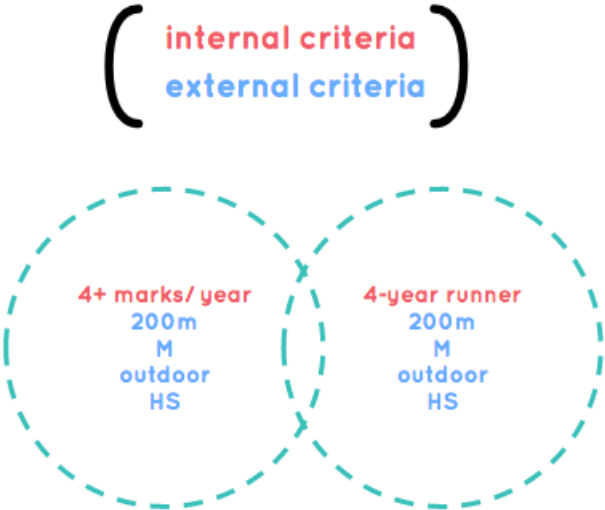


Figure 3.3: The intersection between the two circles depicts the marks extracted from the database

Once the data set was established, it was necessary to realize the features that should be included in the feature set and the optimal  $k$  clusters for the  $k$ -means clustering algorithm. This was accomplished in a prototypical manner by running the analysis on the validation set and using the percent

error results as a measure when evaluating the algorithmic performance. After trying a number of features, including a mean mark for each of the first three seasons, median mark per season, variance per season, standard deviation per season, and personal record per season, it was discovered that the lowest percent error was achieved through including the athlete's top eight marks to-date, irrespective of season. This was confirmed on an event-by-event basis and recalibrated slightly for events with higher margins of error. Generally speaking, this entailed incorporating more or fewer marks into the feature set. The k value for the k-means clustering was discerned after the finalization of the feature set, largely due to the fact that the feature set impacted the margin of error to a much greater extent. Here, the k-means clustering algorithm was run through both events and genders in intervals of two to find an optimal k value. Each possible k value was then used ten times and its average was taken due to the intrinsic variance present in k-means clustering. While it would have been a relatively trivial process to choose a k value dependent on the event and gender at hand, this process actually found that there was more generalizable relationship between the optimal k value and the number of elements in the data set. Setting k to  $(\# \text{ of elements in the data set}) / 5$  yielded clusters with five elements on average whose fourth-year personal records formulated a strong prediction for the athlete in question. Hence, k was established to satisfy this ratio.

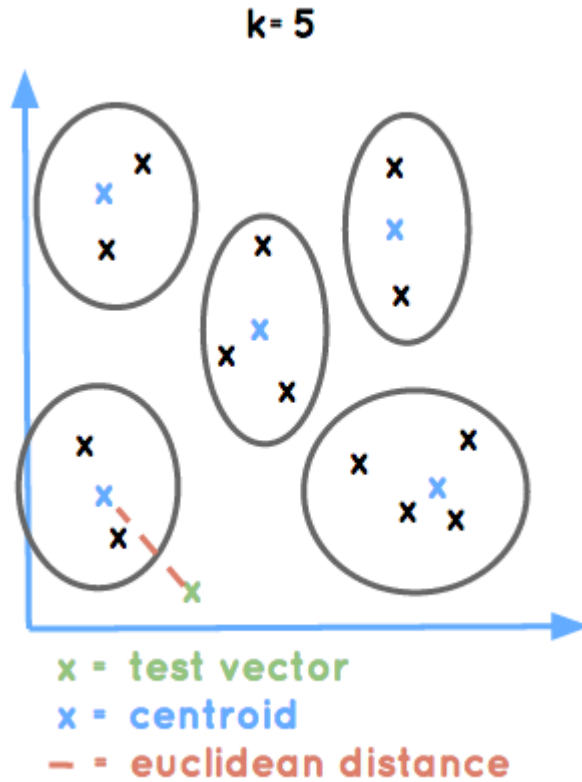


Figure 3.4: An 2D simplification of the k-means clustering algorithm

Together, the data set criteria, the feature set features, and the  $k$  value for k-means clustering were responsible for the results of the algorithm. However, it is worthwhile to mention that there was additional preprocessing that was done prior to the data set extraction. About one percent of the total marks were simply erroneous and needed to be removed from the database entirely. The most likely hypothesis for this is the manual entry of meet results by coaches from around the country. Some fields, such as event type,

are selected via a drop-down menu and therefore have a finite number of possible states, but others, such as the actual mark, are typed into a blank field and could take on any value. Notable instances include a 4x400m relay time of 999999 and a 1600 meters time of 81:02.

Yet, the more important preprocessing regarded marks that were indeed valid but caused noise within the dataset and should be excluded. For example, while most track meets possess an FAT (Fully Automatic Timer) system and hence their marks are punctuated with an 'a', a small percentage of meets use solely hand timers and convert the runner's time by adding .24, making up for the delay between starting a watch and the starter's pistol actually firing. This is usually a sufficient equalizer, but there are instances where a runner has an outlying 'converted' time that far betters any automatic time they've run. In these cases, the converted time was likely a result of human error and should not be included in any analyses. Therefore, Wolfram Alpha's algorithm for determining outliers was used for removing any outlying converted times. In cases of athletes with only converted times, the outlier algorithm was not used, since it is much more difficult to distinguish between 'real' outliers and 'fake' outliers without an FAT correlator. Through measures such as these, extensive preprocessing was done to ensure the validity of the data set and to remove dirty marks.

### **3.2.2 Recommendation**

For recommendation, the k-means clustering algorithm was employed once again. Since the data is unlabeled for this process, meaning that there is

no predetermined similarity measure between athletes, the choice of k-means clustering as an unsupervised learning algorithm makes intuitive sense. Relationships between clustered athletes can surface and it is even possible that the clustered athletes are related in ways that reach beyond the feature set itself. K-means clustering is a storytelling algorithm that can extract similarities without explaining exactly what they are.

There are two underlying questions when considering k-means clustering and recommendation: which athletes should be included in the clustering process, and what the feature set should be for each athlete. The questions, although seemingly distinct, interplay to formulate a singular dilemma: how can one create a cohesive feature set while clustering athletes that may differ in numerous ways?

Hence, determining athletes for the clustering process is not as easy a task as it was for prediction, as the scope is now athlete-wide instead of event-wide. For instance, an athlete may have run the 400 meters and competed in the javelin throw several times each year. Should a similar athlete have competed in both events, or just one of them? Qualifying similar athletes as only those who competed in both events would drastically reduce the data set and hence the likelihood of finding significant matches. However, allowing 400 meters-only runners and javelin throw-only athletes into the data set raises an issue as well. How can a single feature set be created for clustering when some athletes in the data set competed in disparate events? Not only that, but it's reasonable to assume that a similar athlete will follow an analogous progression to the athlete in question. Therefore, should



a similar athlete have run the same events in the same grade? This quickly becomes a very complex problem. How can data sets of similar athletes be found that retain a sizeable quantity while yielding feasible feature sets?

To resolve this, the notions of simplicity and generalizability again come into play. A solution is needed that allows any athlete to input his or her profile and perform some sort of clustering with a number of similar athletes in a reasonable amount of time. The problem is circumvented as follows:

1. Select events and the number of times each event was competed in on a per-grade basis with specific gender, level, school type <sup>1</sup>
2. Keep events on a per-grade basis that were competed in at least 4 times
3. Take the top 8 events on a per-grade basis by number of times competed
4. Similar athletes are those who competed in all 8 of the events (and possibly others) on a per-grade basis at least 2 times

Table 4.1 below shows a sample of what would be taken from the database.

---

<sup>1</sup>Note that relays and splits were excluded from this selection since relay times are more indicative of team ability rather than individual skill.

Table 3.1: Sample Athlete Match Requirements

Count	Event	Grade
14	200m	12th
11	200m	11th
10	100m	12th
10	400m	10th
10	400m	9th
9	200m	10th
7	100m	10th
6	400m	11th

Events and grades that the algorithm will use for matching similar athletes.

This solution has substantial benefits, the foremost being its generality. It allows the athlete to have competed in any number of events for any number of years without requiring a similar athlete to have tried all of the events. In addition, it is not event-specific but still allows for mapping comparable progressions, since similar athletes must have run the events in the same grade as the input athlete. It also allows similar athletes to have competed in other events that the input athlete did not, and permits input athletes who have run relatively infrequently.

This does, however, make the assumption that an athlete’s profile can be summarized by merely the top 8 events on a per-grade basis, as shown in Table 3.1. For instance, if an athlete ran the 200 meters ten times in grades 10 and 12 but only seven times in grade 11 and it did not make the list

of top 8 events specifically for grade 11, by the above definition, a similar athlete could well not have run the 200 meters in his or her junior year. Still, the similar athlete must have run the 200 meters at least twice in his or her sophomore and senior seasons, and therefore it is likely that he or she ran the 200 meters in his or her junior season as well. It thus seems like a reasonable hypothesis that using the top 8 events could be sufficient criteria for similarity. <sup>2</sup>

Finally, this solution resolves the issue of potentially differing feature sets by only using the top 8 events on a per-grade basis as the foundation for the feature set. Since similar athletes are guaranteed to have run those events in the specific grades, any calculations or measures done to determine features should work in all cases.

Since the data is no longer labeled - it is not possible to produce a margin of error - the features from the optimal feature set for the predictor are used for the comparator as well. Thus, the top two marks, on an event and grade basis, are used as features for clustering the athlete profiles. In the same vein, the optimal k value for clustering was taken to be the  $\frac{\text{elements\_in\_the\_data\_set}}{5}$ .

The ability of this solution to simultaneously find similar athletes and extract feature sets therefore allows a user to input his or her profile and quickly receive results - athletes in the closest cluster - on average, five athletes - by euclidean distance on the feature set. As in the predictor, the recommender undergoes the same data preprocessing before running the

---

<sup>2</sup>However, if the events are too restrictive and result in an unfeasible table size, the constraints are lessened.

k-means clustering algorithm.

### **3.3 Results**

Results across both the prediction and recommendation systems shared a similar issue: a lack of data. Despite the large volume of athlete profiles, only a small fraction of them were robust enough to warrant inclusion in the analyses. Still, the predictor was able to receive quality results in terms of average percent error, especially for the running events on the men side. In that same vein, although few athletes met the matching criteria of running certain events some number of times over the course of several seasons, the recommendations were surprisingly accurate. This sentiment was validated through recommendation evaluations by collegiate varsity track and field athletes.

#### **3.3.1 Prediction**

The prediction algorithm was run independently on data sets for ten events, by gender - the 100 meters, the 200 meters, the 400 meters, the 800 meters, the 1600 meters, the 3200 meters, the long jump, the shot put, the javelin throw, and the pole vault. These events were chosen to include a sampling of the different types of track and field events while allowing for specific analyses, such as the effect of race distance on prediction margin of error. Below is the average percent error, by event, when comparing the prediction to the actual personal record achieved, for each data set.

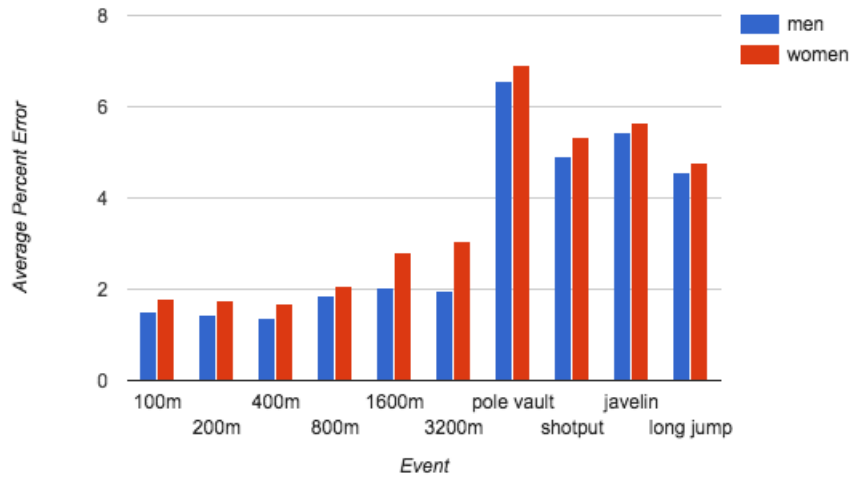


Figure 3.5: The average percent error for a variety of events separated by gender

It is evident that the greater the race distance, the higher the margin of error. This could possibly be attributed to the increased linearity of sprint race progressions. Furthermore, the field events all have higher margins of error than the track events. This is demonstrated in Figure 3.5. Potential reasons for this include smaller data sets and the ability of field athletes to “scratch”. For instance, long jumpers scratch their jumps if they step over the jump board when pushing off, and it is therefore not unheard of for jumpers to have a personal record that clearly exceeds the other marks.

In addition, the women have higher margins of error in almost all track and field events. One possible explanation for this is the fact that women

tend to have higher performance variability than men in both track events and field events. Figure 3.6 shows a margin of error histogram for the men's 1600 meters.

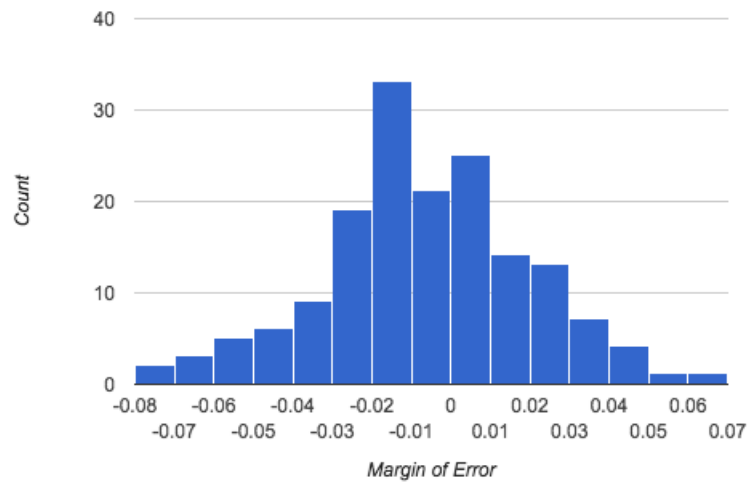


Figure 3.6: A histogram depicting the margin of error spread for the men's 1600m

Figure 3.6 represents the model that was used to calculate the ranges and confidence intervals alongside the mark predictions. Figure 3.7 is a graph displaying the effect of minimum marks per year on the clustering analysis.

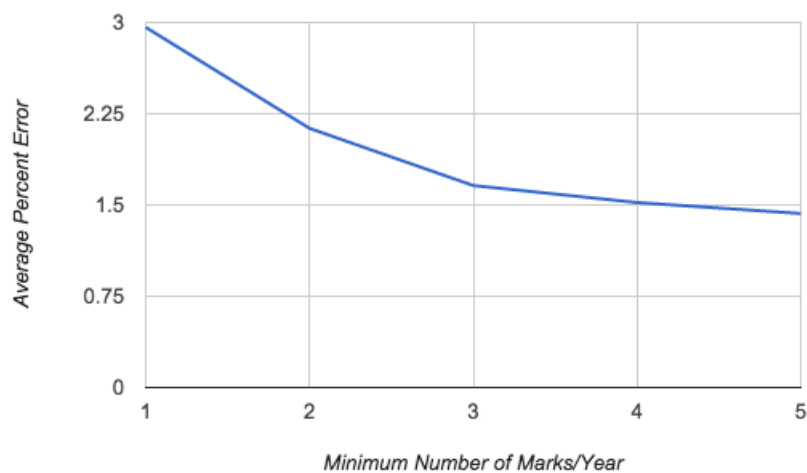


Figure 3.7: The effect of marks per year as a criteria for athlete selection in the dataset

As expected, the greater the number of times an athlete has run an event each year, the better the k-means clustering algorithm performs.

Finally, a baseline comparison test was performed utilizing a least squares regression on only the athlete's individual performances to formulate a prediction for the fourth-year personal record. The results of the k-means clustering algorithm against the baseline test over a number of events are shown in Figure 3.8.

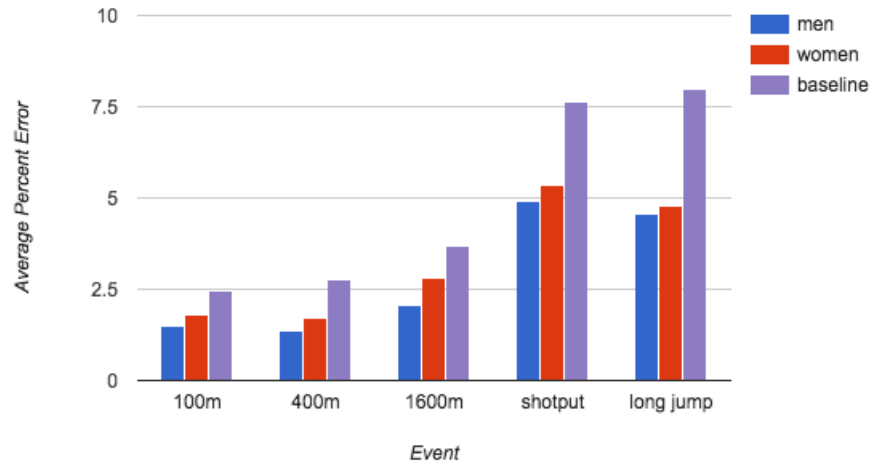


Figure 3.8: The predictor algorithm compared to a simple linear regression

The k-means clustering algorithm outperforms the baseline comparison in all events.

### 3.3.2 User Study of Recommendation

Evaluation of the recommendation system was challenging due to its novelty and lack of actual similarity labels. It appears as if there is no existing system that takes track and field athlete profiles and finds comparable athletes. In other words, there is no baseline system to stack up against. Hence, evaluation was accomplished entirely through questionnaires filled out by varsity track and field athletes with profiles on Athletic.net. The questions asked were:



1. How satisfied were you with your recommendations (on a 1-10 scale)?
2. How similar do you believe the recommended athlete profiles were to your profile, taking into consideration your entire profile?
3. How similar do you believe the recommended athlete profiles were to your profile, taking into consideration only your most frequently competed event? If you did not answer 'never' to the previous question, how frequently would you visit this website?
4. If this recommender were made available for free online, how likely would you be to use it?
5. If this recommender were made available for free online, how often would you use it?
6. How does this recommender system compare to other recommendation systems that you've seen previously?
7. If this recommendation system were packaged as a tool that found high school rivals and notified you of their recent performances, how likely do you think athletes would be to use it?
8. Please list any desired additional capabilities or functionalities for the recommendation tool here.
9. Please list any additional comments about the recommendation tool here.

Figures 3.9-3.13 below display the user feedback for the recommendation system.

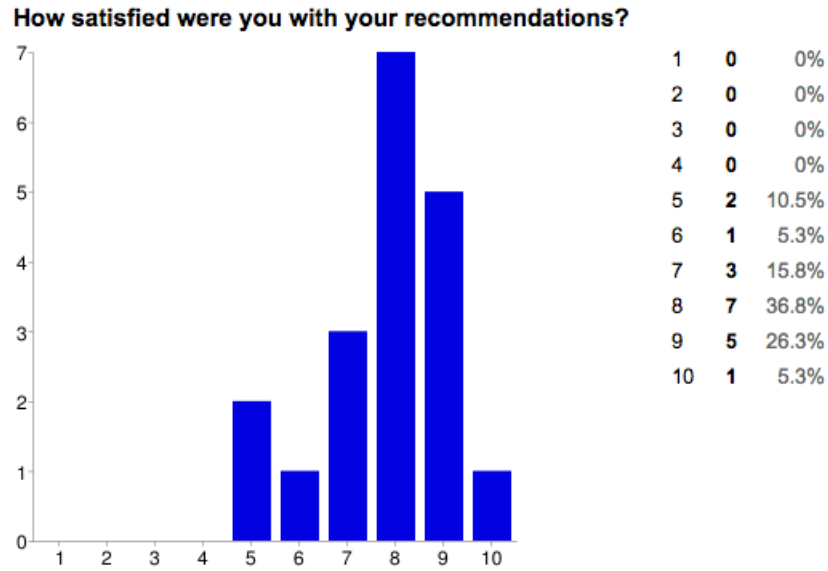


Figure 3.9: Graph showing user satisfaction with the recommendations.

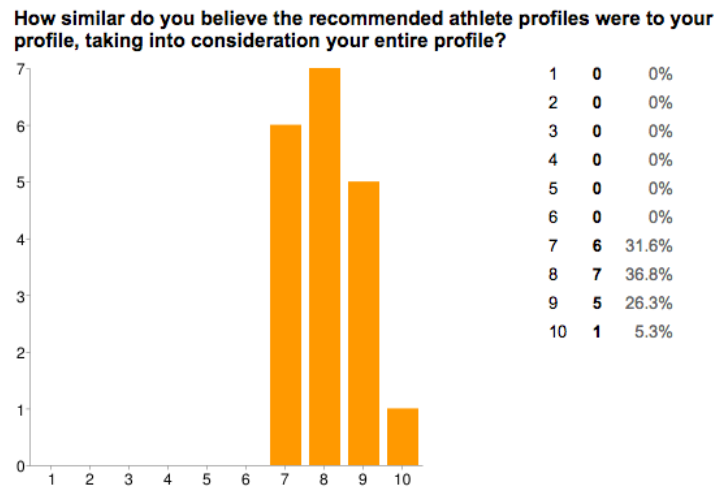


Figure 3.10: Graph showing user satisfaction in relation to entire profile.

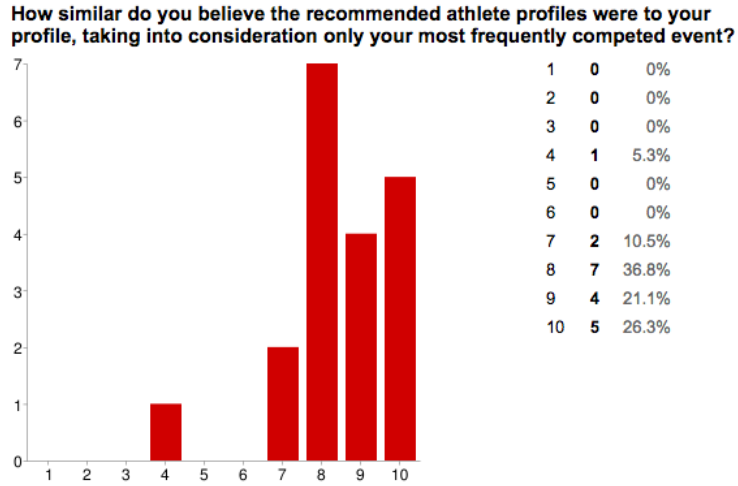


Figure 3.11: Graph showing user satisfaction in relation to most frequent event.

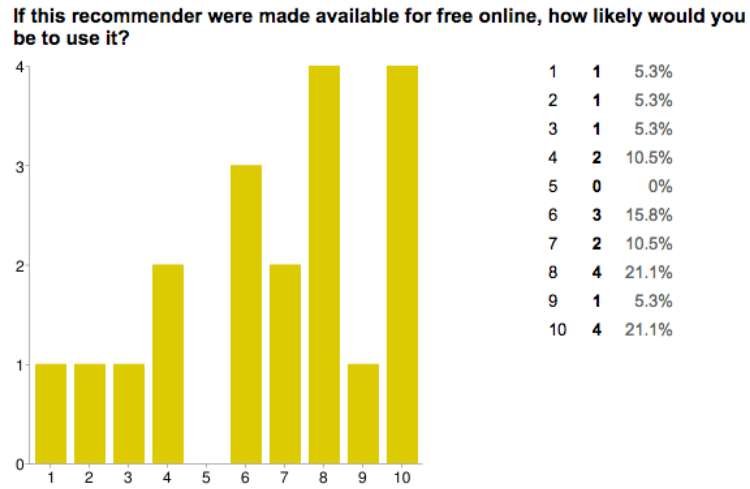


Figure 3.12: Graph showing user likelihood of using recommendation system.

**If this recommender were made available for free online, how often would you use it?**

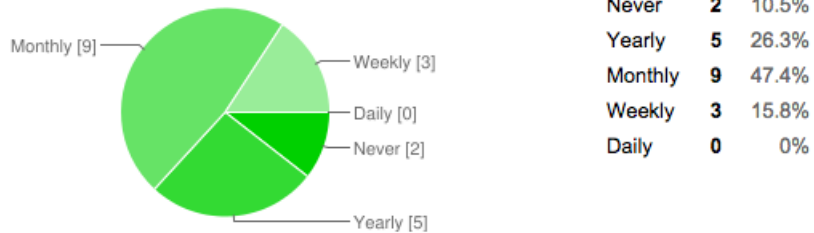


Figure 3.13: Pie chart showing anticipated user activity with recommendation system.

**If this recommendation system was packaged as a tool that found high school rivals and notified you of their recent performances, how likely do you think athletes would be to use it?**

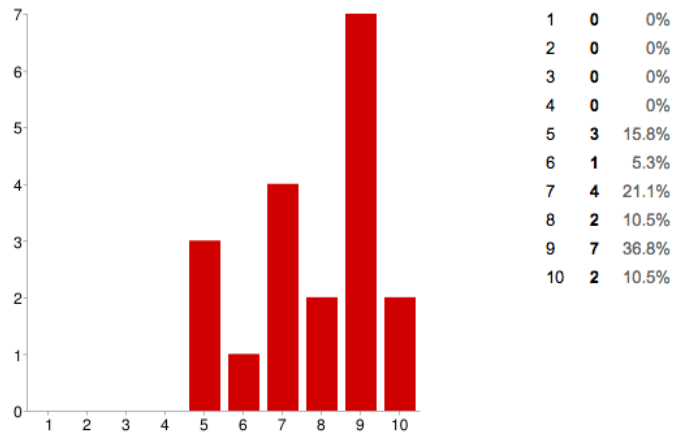


Figure 3.14: Graph showing user interest in a rival-matching tool.

Overall, evaluators of the recommendation system were pleased with their recommendations. Each of the evaluators were given somewhere between 1 and 5 matches depending on table sizes and degrees of similarity. Some participants expressed interest in using the recommendation tool on a

regular basis, while others were not as intrigued by the idea. This correlated with satisfaction rates and perceived similarity between athlete profiles on both an event and profile basis. The final question of the survey prompted the user to consider an extension of the tool that would allow him or her to view recent performances of similar athletes and treat them as rivals, receiving notifications after recent races. Evaluators were interested in this notion and, as such, this would be an area worth pursuing and developing further.

### **3.4 Discussion**

Through data-driven modeling, it was possible to achieve both a quality prediction system and recommendation system.

For the prediction system, while the analysis of this algorithm has lent itself to a number of insights regarding track and field performances as a whole, the ultimate aspiration is to reduce their margin of error across all events in both genders. To this end, the first step would be to collect significantly more data, as there were simply not enough athletes with a sufficient number of performances to effectively cluster them.

The recommendation system would also benefit from a larger data set, but based on the evaluations, it still performed reasonably well, yielding an average of 7.79/10 satisfaction rate with evaluators expressing an interest to use the tool several times per month. Regardless of its shortcomings, considering the fact that there isn't currently an online tool occupying this

space, it is certainly a step in the right direction. At the very least, the reduced simplicity of the athlete selection for clustering yielded surprisingly effective clusters.

## 4 Basketball

Previous research [GVT85] has studied momentum strictly from the perspective of shooting, focusing on the potential existence of the hot-hand. This involved holding isolated shooting experiments to evaluate the effect of streaks on field goal percentage, with the premise that a player develops a hot-hand solely from previous made shots. Slightly more recently, though, there have been studies ([BBJ99] and [MLSN92]) that tracked sequences of perceived momentum in actual basketball games. A high-level overview of their results suggest that there are many different positive and negative events over the course of a basketball game that affect momentum. Neither study, however, tracked how these momentum-inducing events affect shooting.

The fundamental question arising from these studies is whether the hot-hand can be developed by these momentum-inducing events, rather than strictly by made shots. That is, what can be learned about the hot-hand theory by factoring in, along with made shots, other events that influence momentum within the context of a game?

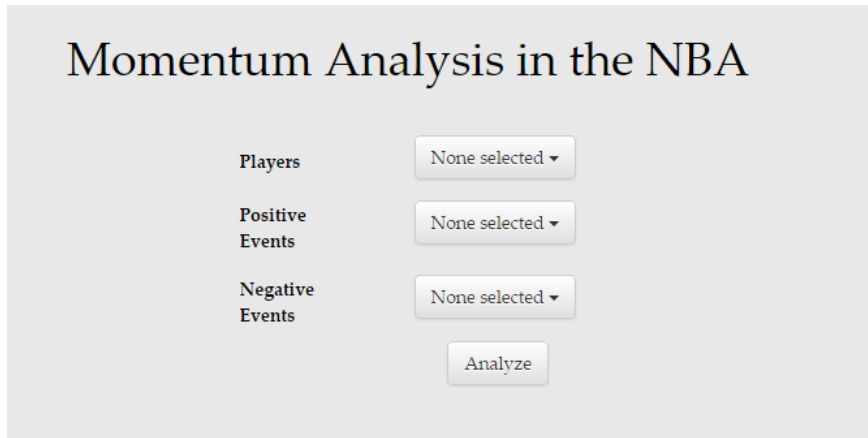
The interface for this analysis — described in more detail in section 4.2 below — gives the ability to visualize the effect of any combination of these momentum-inducing events on in-game shooting and, more specifically, the hot-hand theory.

## 4.1 Design and Interface

The goal of the interface design was to present a very simplistic and minimalistic display, while still allowing the user to provide input into 3 aspects of the analysis:

1. Players or Teams
2. Positive Events
3. Negative Events

A page is presented with 3 multiselect widgets: one for each of the inputs listed above. I used David Stutz's JQuery multiselect plugin based on Twitter Bootstrap for its sleek appearance as well as for added functionality such as option groups to group players by team and filtering by name. The multiselect widgets are centered and rendered in a vertical layout to ensure ease of use and a clean presentation.



Momentum Analysis in the NBA

Players None selected ▾

Positive Events None selected ▾

Negative Events None selected ▾

Analyze

Figure 4.1: The interface for the basketball momentum analysis



## 4.2 Implementation

Once the input (players/teams, positive events, and negative events) is selected, the analysis process consists of a pipeline of 3 phases.

The first step is using the input provided by the user to construct and execute a database query to retrieve the necessary rows from the PostgreSQL database. This process is performed by the Query Generator module, designed to handle any combination of events.<sup>1</sup>

The query is constructed by following this algorithm for each positive or negative event:

1. Map the name of the inputted positive/negative event to its event type (i.e. made free throw  $\rightarrow$  ft). Since the analysis involves looking at how shooting is affected by momentum, all shots (with the exception of free throws) are included by default.
2. Determine if there is any extraneous information that can be removed and add where clause in query if necessary. For example, if offensive rebounds are deemed to be positive events but defensive rebounds are not, then the analysis is not dependent on any information regarding defensive rebounds. Therefore, this partial query can read:

```
“SELECT * FROM events WHERE type=rebound AND is_defensive  
= FALSE”
```

---

<sup>1</sup>The schema for events (i.e. shots, rebounds, etc) in the database includes the fields pname — which is the name of the player who performed the action, henceforth referred to as the acting agent — and team — which is the team of the acting agent

3. Include the player or team names that will be analyzed in where clause.

If the analysis involves players, this will be a sequence of `pname=$1`  
OR `pname=$2` OR `team=$1` OR `team=$2` OR `$i` in the case of  
teams. The `$i` syntax simply is a placeholder denoting the  $i^{th}$  argument  
in the provided values array.

These queries are combined via UNION ALL statements, and the re-  
turned rows are ordered by the player/team name, game, quarter and time  
remaining to maintain chronological integrity.

As explained generally in section 1.2.2.2, this analysis is highly depen-  
dent on a quantitative model as a means of tangibly measuring the effect  
momentum has on shooting. Therefore, we define momentum to be a streak  
of consecutive positive or negative events performed by a player or team. We  
further define 6 different states of momentum (ordered from most negative  
to most positive):

1. 3+ negative events
  2. 2 negative events
  3. 1 negative event
  4. 1 positive event
  5. 2 positive events
  6. 3+ positive events
- (0. catch-all bin)

For example, if a player is the acting agent of 3 consecutive negative events, he is said to be in State 1. However, as soon as this player performs a positive event, he jumps to Stage 4. To avoid streaks continuing across games, the state is reset to 0 at the beginning of each contest, and shots that come before any streak begins are kept in state 0, but otherwise discarded.

Using the results of the database query constructed in the previous phase, the next phase analyzes the set of events for each player or team. After each event, the current streak is updated based on the state transitions described in the figure above. If the event is a shot, it is binned based on the value of the current state of momentum. These six bins, as well as an additional bin that serves as a catch-all, are built up with the data from the entire season. Once these bins are populated with the proper shots, we compute the statistics relevant to the analysis.

That is, for each bin  $i$  with  $1 \leq i \leq 6$  (bin  $i$  refers to all the shots taken while in momentum state  $i$ ), we compute:

$$\frac{\text{made\_shots\_in\_bin\_}i}{\text{total\_shots\_in\_bin\_}i} - \frac{\text{made\_shots\_in\_bin\_}0}{\text{total\_shots\_in\_bin\_}0}$$

which is equivalent to the player or team's FG% on current streak - overall FG%.

This formula normalizes over the player's or team's overall field goal percentage in order to isolate the direct change in shooting caused by the momentum state.

### 4.3 Results

The results are presented in graph format, with the notches on the x-axis corresponding to momentum state 1, 2, 3, 0, 4, 5, and 6 respectively. Momentum state 0 (catch-all state) is displayed in the middle as it is the medium between the positive states (1, 2, and 3) and negative states (4, 5, and 6).

The interface provides  $x$  positive events to choose from and  $y$  negative events to choose from, producing  $2^{10} * 2^5 = 32,768$  possible combinations of events. This is far too many combinations, as analyzing each involves iterating through an entire season's worth of play-by-play events for every player and team. Therefore, for simplicity as well as feasibility issues, we choose to further analyze the results for the following positive and negative events:

positive events = {"made shot"}  
negative events = {"missed shot"}

Figure 4.2 shows the results for an analysis of LeBron James (Miami Heat), Stephen Curry (Golden State Warriors), James Harden (Houston Rockets), Paul George (Indiana Pacers), and Nick Young (Los Angeles Lakers).

These players were chosen for closer analysis for several reasons. Firstly, they are prominent and well-recognized NBA players. Secondly, they are all high volume shooters and scorers; this is important to make sure there

is enough data to produce meaningful results. Finally, they are all streaky shooters, especially from the perimeter. While LeBron James and James Harden drive to the basket for easy shots with some regularity, they also tend to score on long jump shots, especially when they have scored recently.

Generally speaking, the names chosen represent players that shoot from the outside with regularity, and have the authority within the offense to take quick shots in succession when they are on streaks. When players don't have this authority, their distribution of shots over time typically remains fairly constant, which lessens the potential impact of momentum, or the hot-hand.

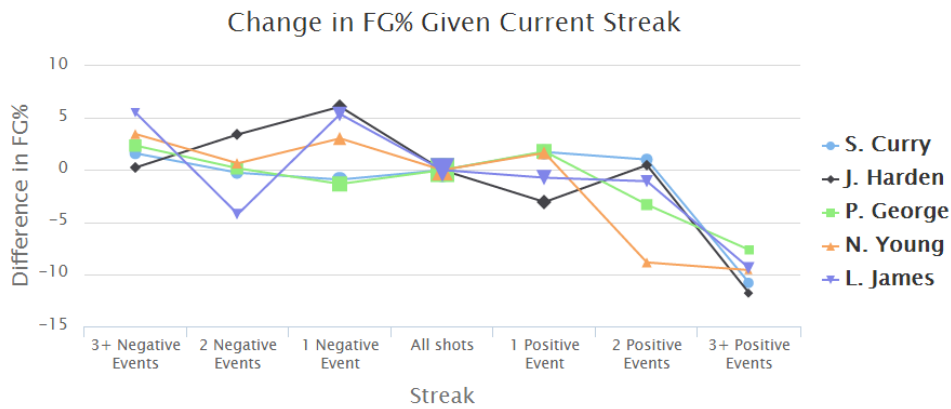


Figure 4.2: Example output of analysis for 5 high-profile NBA players. Y-axis shows the player's field goal percentage on a given streak compared to his overall field goal percentage.

The results for the linear regressions of each plot of points is shown in Table 4.1. The r values show a moderate negative correlation for Stephen

Table 4.1: Regression Analysis

Player	Regression Line	R	$R^2$
Stephen Curry	$y = 2.352 - 1.145x$	-0.5618	0.3156
James Harden	$y = 4.819 - 1.8284x$	-0.6933	0.4807
Paul George	$y = 2.485 - 1.209x$	-0.7594	0.5767
Nick Young	$y = 5.002 - 2.128x$	-0.8354	0.6979
LeBron James	$y = 4.109 - 1.589x$	-0.6566	0.4311

Regression equations and r values for data from Figure 4.2

Curry, James Harden, and LeBron James, and a strong negative correlation for Paul George and Nick Young. This demonstrates that, amongst these players, shooting percentage drops as their momentum (streak of made or missed shots) improves. Examining the slopes of the regression lines, as the momentum state changes from negative to positive, shooting percentage drops by 1.58% on average. Perhaps most striking though, are the endpoints in the plot for each player in Figure 4.2 - momentum states 1 and 6 (3 or more consecutive missed shots and 3 or more consecutive made shots, respectively). Each player shot better than his overall field goal percentage after missing 3 or more shots (on average 2.62% better), but shot dramatically worse after making 3 or more shots (on average 9.589% worse). These results are in opposition to the expectation of the hot-hand theory.

This, however, was not a general trend. To investigate the results for the remaining NBA players, we constructed two graphs (Figure 4.3 and Figure 4.4). The first plots player points per game and change in field goal percentage while in momentum state 1 (3 or more consecutive missed shots), while the second is the same except using momentum state 6 (3 or

more consecutive made shots). Note that the points in red are the players analyzed in Figure 4.2.

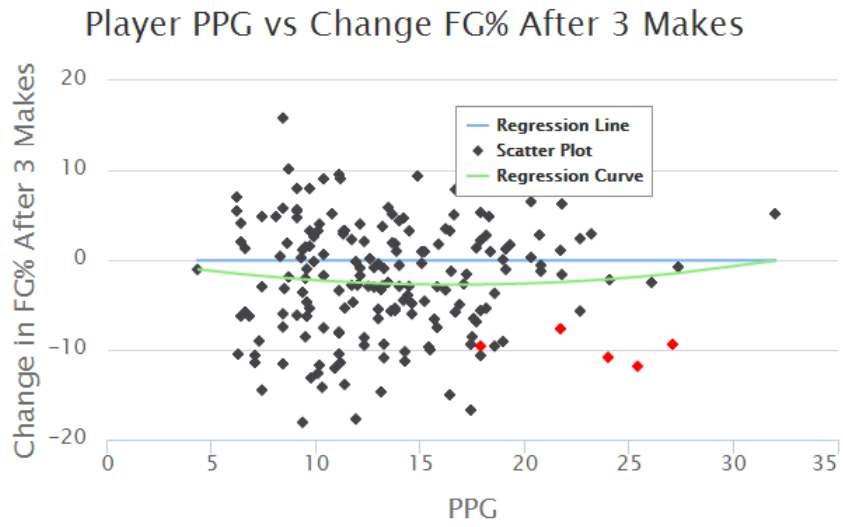


Figure 4.3: Plot of player points per game and change in shooting percentage after 3 consecutive made field goals.

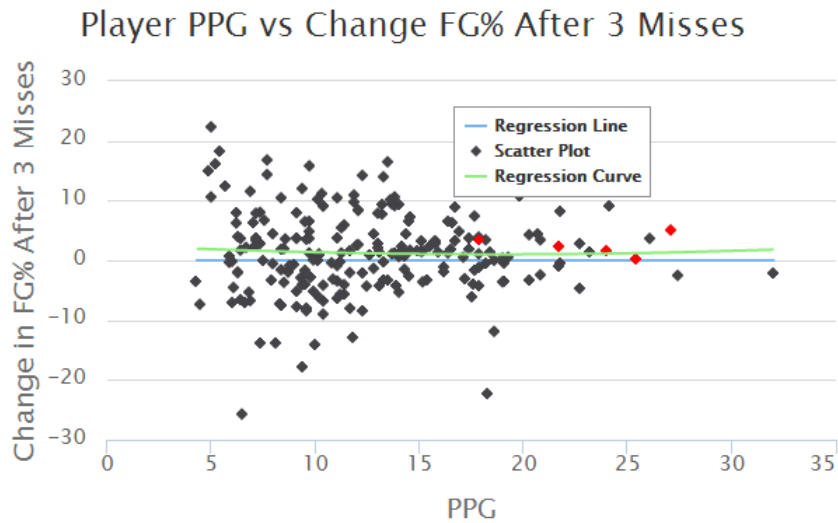


Figure 4.4: Analogous to Figure 4.3 but 3 consecutive *missed* field goals

From the regression lines in Figures 4.3 and 4.4 (rendered in blue), there is no discernable correlation between points per game and shooting while on positive or negative shooting streaks.

## 4.4 Discussion

While some of the results did show promise, there is not enough corroborating data to make any strong claims about the effects of momentum on in-game shooting. Analyzing high profile NBA players like LeBron James, Stephen Curry, James Harden, Paul George, and Nick Young, along with others who showed similar results such as DeMar DeRozan, Brandon Jennings, and Kobe Bryant (albeit on a limited sample size - Bryant only played



6 games and attempted 73 shots), revealed that shooting efficiency may actually drop after making previous shots. However, more likely, this is simply the result of a flaw in the model. The model does not currently control for shot distance or shot difficulty in the analysis. With recently available player tracking data, shot difficulty can now be quantified using metrics such as closest defender distance and closest defender height. Adding this information into the model would answer the question of whether these players actually shot worse when they made shots, or if the drop in field goal percentage was due to them taking more difficult shots.

Due to this factor, we hypothesized there would be some correlation between player points per game and their change in shooting percentage after hot or cold shooting streaks. Players with the ball in their hands a lot (typically high scorers) would be more likely to have a quick trigger if they felt they were hot, potentially resulting in more difficult shots and therefore, a lower field goal percentage. However, despite finding several players that fit this trend, there were also players like Kawhi Leonard, Chris Paul, and Dirk Nowitzki whose field goal percentage remained fairly constant regardless of the shooting streak they were on. Therefore, it might be more reasonable to say that this analysis, while it might not illuminate any general trend amongst all players, does uncover different types or classes of players with respect to how they shoot in response to positive or negative momentum.

## 5 Conclusion

Although this project achieved significant results in each of its three components — predictive modeling, recommendation, and statistical analysis — there are certainly areas for improvement and further development.

In terms of track and field, as previously mentioned, results could likely be seriously improved given a larger dataset to work with. TFRRS (Track and Field Results Reporting System) has a much greater, cleaner dataset online and is the results host for all collegiate track and field runners in the United States. With permission to scrape or otherwise obtain their data, an analysis could be done on their dataset as well. Furthermore, Athletic.net also has a significant database of cross-country athlete profiles and performances. It would be relatively simple to formulate a similar predictive model and recommendation system for cross-country runners with only some minor adjustments on both the data collection and algorithmic implementation sides. Finally, it is possible that there are other machine learning techniques that would perform better than k-means clustering, especially on the prediction side. Perhaps experimenting with an n-dimensional regression algorithm would yield improved results.

In basketball, it might also be relevant to look at specifically the time that has passed between made or missed shots, not just the number of consecutively made or missed ones. For instance, how is a streak affected by time between shots? How can we define precisely what a streak is when factoring in time? In the same vein, it would likely be significant to account

for substitutions — a player sitting on the bench could possibly be analogous to the ending of momentum. Other areas of future work include clustering the results to ascertain similar basketball players and reasons behind the clusters, as well as analyzing multiple seasons to increase the dataset size.

Taking a step back and looking at the bigger picture, the field of sports data analytics is rising to the forefront of the industry. Teams from sports worldwide are hiring data analysts to constantly remodel and reevaluate player and team performance. We have shown that it is possible to make predictions, recommendations, and perform statistical analysis given data that is already available online, and that there is remarkable potential for subsequent work.

## Bibliography

- [BBJ99] Kevin Burke, Michelle Burke, and Barry Joyner. Perceptions of momentum in college and high school basketball: An exploratory, case study investigation. *Journal of Sport Behavior*, September 1999.
- [Dav06] Thomas Davenport. Competing on analytics. *Harvard Business Review*, January 2006.
- [Dav14] Thomas Davenport. Analytics in sports: The new science of winning. *International Institute for Analytics*, February 2014.
- [Dea67] Michael Deakin. Estimating bounds on athletic performance. *The Mathematical Gazette*, 1967.
- [DG79] Jack Daniels and Jimmy Gilbert. *Oxygen Power: Performance Tables for Distance Runners*. self published, 1979.
- [Exp14] Draft express, November 2014.
- [GVT85] Thomas Gilovich, Robert Vallone, and Amos Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 1985.
- [Hea06] Timothy Heazlewood. Prediction versus reality: The use of mathematical models to predict elite performance in swimming and athletics at the olympic games. *Journal of Sports Science of Medicine*, 2006.

- [Hop05] Will Hopkins. Competitive performance of elite track-and-field athletes: Variability and smallest worthwhile enhancements. *Sportscience*, 2005.
- [Liu93] Yuanlong Liu. Evaluation and prediction of world records and ultimate performance in track and field. *UBC Retrospective Theses Digitization Project*, 1993.
- [Liu10] Yuanlong Liu. Track and field performance data and prediction models: Promises and fallacies. *Olympic Special Issue*, 2010.
- [MLSN92] F. Charles Mace, Joseph Lalle, Michael Shea, and John Nevin. Behavioral momentum in college basketball. *Journal of Applied Behavioral Analysis*, 1992.
- [MS14] Joshua Miller and Adam Sanjurjo. A cold shower for the hot hand fallacy. *Institute for Economic Research*, July 2014.
- [PTH03] David Pyne, Cassie Trewin, and William Hopkins. Progression and variability of competitive performance of olympic swimmers. *Journal of Sports Sciences*, 2003.
- [Pur70] J.G. Purdy. Computer generated track scoring tables. *Medicine and Science in Sports*, 2, 1970.
- [Pur74] Gerry Purdy. Computer analysis of champion athletic performance. *American Alliance for Health, Physical Education, and Recreation*, 1974.

- [Ras06] Carl Edward Rasmussen. Gaussian processes in machine learning. 2006.
- [Ref14] Similarity scores, November 2014.
- [Sch08] Jurgen Schiffer. The 400 metres. *New Studies in Athletics*, 2008.
- [Sol14] Howard Solomon. The amount of data we're creating is out of this world, April 2014.
- [Whi07] Alexander White. Runner's log and predictive performance analytics. 2007.
- [Wik14] Quantification (science), November 2014.
- [Wil14] Jane Williams. Beyond moneyball: Data-mining the premier league. *Knowledge Arabia*, 2014.