3-Point Percentage Prediction

Brandon Veber; veber001@umn.edu

I. Introduction

The goal of this project is to predict 3-point percentages for all NBA players for the 2014-15 season using only data available before the start of the season. The best approach for this project is to formalize it as a regression problem, in which past data will be used to predict future values for each player in the 0-100% range. This regression analysis will be performed for two separate groups:

A. All NBA players with at least 2 years of NBA experience (veterans)
B. All NBA players with 1 or fewer year of NBA experience (novices)

This distinction is made because of the inherent lack of data for the players in group B (novices). Rookies implicitly have no NBA data, and most players play so few minutes in their first season that they don't accrue enough meaningful NBA statistics.

In order to create the computer software required to perform these goals, the problem is further subdivided into three stages. As shown in Figure 1, the first stage is data collection. Next, important statistical features are extracted from the data. Finally, the features are fed into a supervised learning algorithm to generate useful predictions. Each of these stages will be described in sufficient detail below.

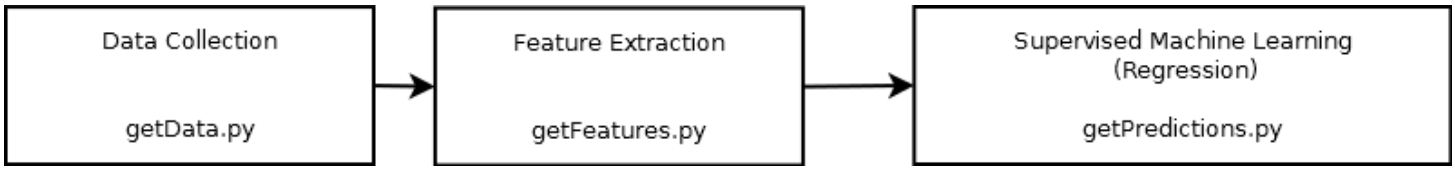

Figure 1: Software Flow Diagram

II. Data Collection

In order to determine a useful strategy, it is necessary to analyze data from previous NBA seasons. Therefore, total end-of-season stats for all active players in each season during the past 15 seasons (1999-00 through 2013-14) was collected. Also, every season of career data (not limited by the 1999-00 start date) for all players who were active during that span was collected. This data was gathered from www.basketball-reference.com using custom Python modules found in the getData.py file.

III. Feature Extraction

The regression analysis is done by generating two unique sets of features for each group (veterans and novices) using the custom modules in the getFeatures.py file. This feature extraction stage is critical in insuring the analyses and

predictions are useful. It is important to include a set of relevant features that is concise and does not overly complicate the problem. Low-value features will not offer any increase in predictive performance, and will only add noise to the data. Therefore, it is imperative to keep the feature size limited to truly beneficial information. Shown below are all the chosen features and their justifications.

A. Veterans

For all individual veterans, a total of 7 features are extracted for analysis:

- NBA Career 3-point percentage (3P%)
- Previous NBA season 3P%
- NBA Career 3-point attempts (3PA)
- Previous NBA season 3PA
- NBA Career Free Throw percentage (FT%)
- Previous NBA season FT%
- Years of NBA experience

These features are used because they represent essential categories that can help define a player's 3-point accuracy. The career 3P% is an obvious choice, and the previous season 3P% is included to hopefully capture any recent changes in performance (i.e. a player hired a new shooting coach; moved to a team with a point guard who is better at finding open shooters; recently suffered an injury, and get less lift on their shot, etc.). The 3PA features are valuable in determining the players expected volume of 3-point attempts. A player who takes more "threes" is more likely to be a good 3-point shooter, otherwise the coach would reduce his shots. Next, the FT% features are included because they contain important information regarding a players shooting stroke without being influenced by complex and difficult to quantify factors such as shot selection, shot location and overall team offense. Finally, the total years of experience is used because a player's 3-point prowess should be more predictable with more years of experience.

These features are generated for all veterans, for each season of interest, when applicable[1]. In order to assure reliable analysis, it is critical that these features are calculated using only data available before the start of the season of interest. For example, to determine features for the 1999-00 season, only career data from the 1998-99 season and earlier are used. Table 1 below show the total number of players (samples) that meet the minimum statistical requirements for feature generation from each season.

Table 1. Samples generated per season for veterans

Season	1999-00	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14
# Samples	212	213	213	194	197	208	210	210	228	222	222	228	243	230	233

B. Novices

For all novices, a total of 4 features are extracted for analysis:

- College Career 3P%
- College Career 3PA
- College Career FT%
- Number of college games played

1. All players who attempted less than 10 3-point attempts per season are excluded from the analysis and simply classified as 'low-volume' 3-point shooters.

The justification for these features is the same as described in the veteran group, with the notable lack of the previous season's statistics. The reason for this is because many NBA players only had 1 or 2 years of college experience, so the difference between the previous season and their career average is negligible. Table 2 below shows the total number of players (samples) that meet the minimum requirements[2] for feature generation from each season.

Table 2. Samples generated per season for novices

Season	1999-00	2000-01	2001-02	2002-03	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14
# Samples	78	78	72	74	74	75	76	88	72	78	81	81	96	95	101

IV. Supervised Machine Learning

Three supervised machine learning techniques are used during model selection testing:

- Random Forest (an ensemble of Decision Trees)
- K-Nearest-Neighbors
- Support Vector Machine

In addition, two simple and realistic heuristics were chosen as benchmark methods for comparison. This is important, because it shows if the more complex machine learning techniques provide value over simple and straightforward solutions. The two heuristics are:

- Predict all players based on their career average 3P% (NBA average for veterans, College average for novices)
- Predict all players based on their 3P% from the previous season (only for veterans)[3]

Next, a set of metrics must be defined in order to empirically determine the best method. First is Mean Absolute Error (MAE). This metric is conceptually easy to understand, because an MAE of 10 implies that a prediction of 40% actually means 40±10%. Second is the commonly used metric Mean Squared Error (MSE). This is the primary metric for evaluation in this report.

Finally, in order to evaluate the effectiveness of the predictive models it is essential to perform useful cross-validation experiments on available past data. In this case, predictions for the past 5 seasons (2009-10 through 2013-14) are simulated by using only data available before the start of that season. This way the true predictive capability of the model is tested. For example, to simulate predictions for the 2011-12 season, only data from the 2010-11 season and earlier are used. In this particular experiment, you can see in Table 3 the total number of training and test samples for each group. The number of training samples is simply all samples from the 1999-00 season through the 2010-11 and the test samples are all samples from the 2011-12 season. This testing procedure is contained in the getPredictions.py file and the experimental results will be presented in the next section.

Table 3. Number of Training and Test Samples for 2011-12 season experiment

	# Training Samples	# Test Samples
Veterans	2,347	243
Novices	927	96

2. Only players with NCAA basketball experience that attempted at least one 3-pointer and one free throw were included in the analysis. Players not meeting this minimum statistical requirements are classified as 'low-volume' 3-point shooters. Players before the 2006 season who were drafted directly from high school and all international players are excluded from the analysis due to a lack of available data.

3. This benchmark is not used for novices because so many of them have only one or two seasons of college experience, so the difference between their career average, and previous season average is negligible

V. Experimental Results

A. Veterans

The results for all 5 techniques are shown in Table 4 below.

Table 4: Experimental Results (Veterans)

	Test Scores									
	Previous Season 3P%		Career 3P%		Random Forest		K-Nearest Neighbors		Support Vector Machine	
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
2009-10	7.05	110.85	5.82	72.52	5.31	59.48	5.47	65.79	5.44	65.32
2010-11	8.30	162.68	7.09	115.95	6.63	104.27	6.66	104.81	6.64	108.11
2011-12	6.85	94.24	5.58	58.11	5.13	51.28	5.08	51.03	4.99	48.07
2012-13	6.71	96.68	5.95	78.36	5.53	68.69	5.60	70.06	5.40	68.60
2013-14	6.87	110.08	6.26	86.88	5.74	75.16	5.86	76.59	5.68	73.78
Average	**7.16**	**114.91**	**6.14**	**82.36**	**5.67**	**71.68**	**5.73**	**73.66**	**5.63**	**72.78**

The Random Forest outperformed all other methods, offering on average a 13% improvement (in MSE) over the career 3P% method, and a 38% improvement over the previous season 3P% method (in MSE).

A visual representation, histograms of prediction error for the 2009-10 season are shown in Figure 2 below. It can be seen that both the career 3P% method and random forest have a similar amount of accurate guesses (<50 Squared Error), however the career 3P% method has a slight tendency to make more large mispredictions (>100 Squared Error).

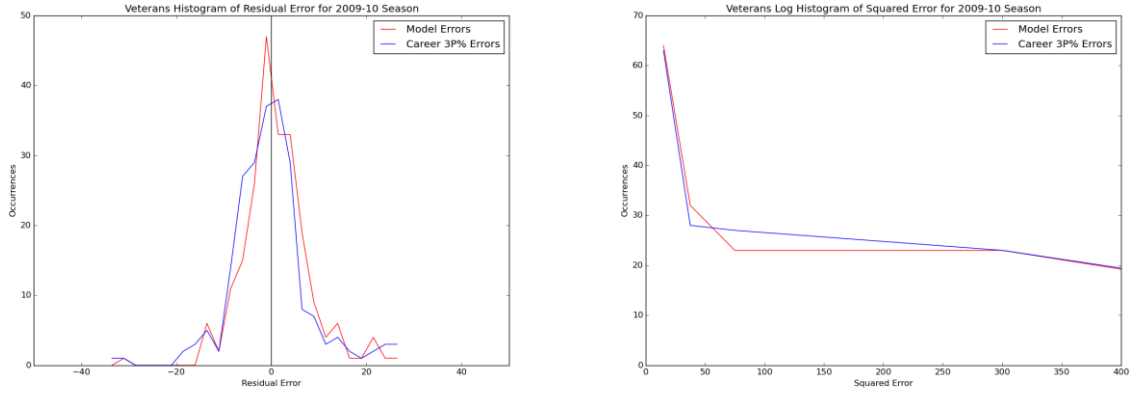

Figure 2: Histogram of Errors Random K-Nearest-Neighbors, and Career 3P% method; Veterans, 2009-10

Residual Error = prediction – actual

Squared Error = $(\text{prediction} - \text{actual})^2$

B. Novices

The results for all 4 techniques are shown in Table 5 below.

Table 5: Experimental Results (Novices)

	Test Scores							
	College 3P%		Random Forest		K-Nearest Neighbors		Support Vector Machine	
	MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
2009-10	11.40	261.14	10.68	192.74	10.88	200.45	10.40	191.52
2010-11	11.06	246.73	9.57	142.46	10.00	153.17	9.84	151.57
2011-12	12.27	271.77	10.32	167.25	10.11	165.90	10.11	170.11
2012-13	12.24	300.22	11.17	203.40	11.13	201.88	10.96	206.59
2013-14	12.32	407.33	11.36	309.73	11.21	286.55	10.71	284.37
Totals	**11.86**	**297.44**	**10.62**	**203.12**	**10.67**	**201.59**	**10.4**	**200.83**

The SVM technique shows the best performance, offering on average a 32% improvement over the college career 3P% method.

The histograms of the prediction error for the 2011-12 season are shown in Figure 3 below. It can be seen that the college career 3P% method tends to overestimate the players actual 3P% (>0% Residual Error), while the SVM model tends to underestimate (<0% Residual Error). Also, similar to the veterans, the career 3P% method produces more very large mispredictions (>400 MSE).

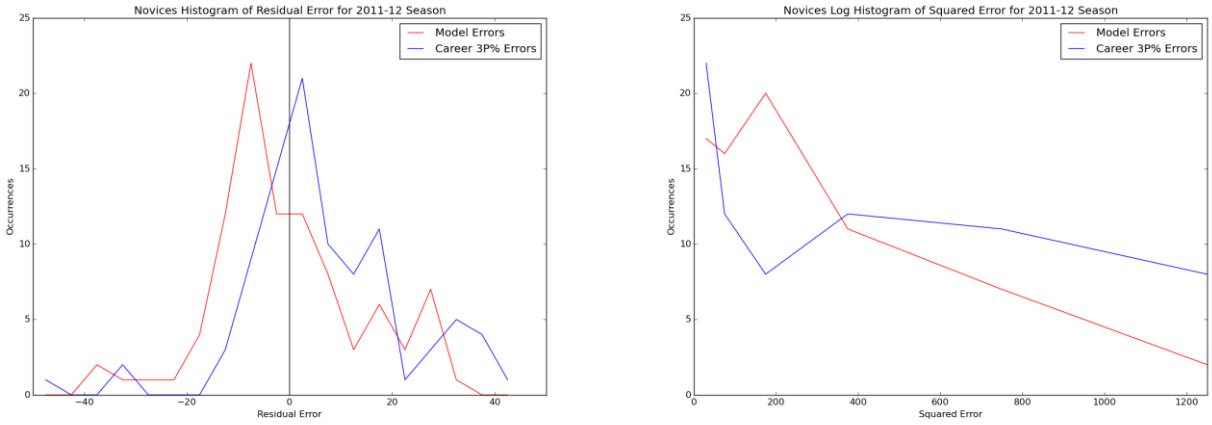

Figure 3: Histogram of Errors SVM method and career 3P% method; Novices, 2011-12

Residual Error = prediction – actual

Squared Error = $(\text{prediction} - \text{actual})^2$

VI. Discussion

The two ideal predictive algorithms are the Random Forest for veterans and the Support Vector Machine for novices. It is not surprising that these algorithms performances changed from group to group. The basic performance for novice prediction, is understandably worse that veteran prediction. This is due to the smaller sample size (the number of available samples for veterans is ~ 2.5x greater than for novices), the inherent difference in NBA stats vs non NBA basketball stats, and the expected difficulty in trying to predict young talent. Also, the differences in dimensionality most likely accounts for the relative change in performance between models for the two groups. Finally, the consistency, season by season for the two ideal algorithms in outperforming the simple heuristics is promising. This consistency makes it very unlikely that these models will ever be outperformed by simple methods. Therefore, the upcoming predictions for veterans in the 2014-15 season are made using the Random Forest, and predictions for novices are made using the Support Vector Machine.

This project represents a positive first step in tackling the problem of 3-point percentage prediction, but could benefit from a more detailed and complex analysis. For instance, it seems easier to predict certain player's 3P% than others (e.g. a player who has had a long career, with many 3-point attempts vs. a third year player who hasn't played significant minutes). Therefore, it might be possible to determine a simple heuristic that can identify easily predictable players and make better predictions for this subset. Moreover, this problem was intentionally limited by simple individual box-score statistics. It is important to start with these in order to keep the analysis simple. However, it might prove useful to include extra features. For example, data regarding shot location and distance to the nearest defender could provide further insight into 3P%. This type of shot log and shot chart data exists and can be obtained from basketball-reference and NBA.com. Furthermore, team related features would add a new and potentially powerful concept into the analysis. Quantifying the talent surrounding each player, and determining their role in the offense is possible with advanced stats like PER, Usage Percentage, and Offensive Ratings. It is also possible to simply note when a player changes teams, and when new players are added to his existing team. It seems likely that at least some of these ideas would add value to the predictive models.

Moreover, these models are unable to make predictions about novices without NCAA experience. This may not have been an issue 10 years ago, but there is an ever increasing amount of talented young European players and it is important to make predictions for them too. It might be difficult to find enough data, or to be able to relate statistics from the Euro style game to the NBA game, but it would be a very helpful tool.

Finally, it would be useful to generate a confidence rating for each prediction. This would be possible by finding the distribution of errors for players with different numbers of 3PA's. The theory is that after a player has taken enough shots (>500) in their career, their shooting percentage is less random and thus easier to predict. Intuitively, these high-volume shooters should have a higher confidence rating than one who has shot far fewer 3-pointers.

References:

[1] *Predictions Are Hard, Especially About Three Point Shooting* "Counting the Baskets." Web. 04 Jan. 2015.

[2] *Adjusting (and Projecting) Three Point Shooting Statistics*. "82Games.com; Web. 04 Jan. 2015

[3] Oliver, Dean. *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington, D.C.: Brassey's, 2004. Print.

[4] "Basketball-Reference.com Web. 04 Jan. 2015.