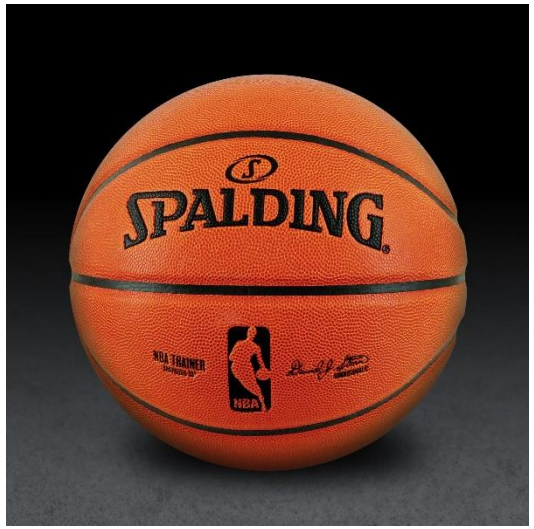
Predicting NBA achievement from inaugural performance



By: Nick Smirnov

August 26th, 2015

GA Data Science 2015

Problem Statement

Can we measure initial performance of a NBA player and use that performance to determine his success and the accent to being an elite NBA player?

My Hypothesis

A player's performance early in his career will determine his success later in his career. A player who exhibits success within his first 3 seasons is likely to continue his progress into becoming an elite player.

Description of my data set

My data set consist of conventional box-score & advanced NBA player metrics from players since 1978.

Number of players: 2,454

Number of total seasons: 13,849

Conventional Statistics:

i.e. Counting Stats

Name	Pos	Min	FG	3Pt	FT	Off	Reb	Ast	TO	Stl	Blk	PF	Pts
Ernest Aguilar 3	SF	17.4	0-5	0-1	0-2	1	4	0	3	1	0	0	0
Albert Warwick B Ps	PG	33.8	4-10	0-2	4-4	1	1	6	2	1	0	0	12
Thomas Reyes	SG	28.7	6-10	0-0	1-1	1	2	0	2	3	0	1	13
Edward Hassett	PF	29.5	8-15	0-2	6-6	1	9	1	2	0	2	5	22
Curtis Meade	FC	28.7	8-12	0-0	1-2	1	7	0	5	1	2	0	17
Matthew Matos B Ps	PG	24.2	0-2	0-0	0-2	1	3	5	3	2	0	0	0
Chad Jordan	GF	25.3	2-4	0-0	0-0	0	5	2	2	0	0	2	4
Stewart Kowalski	SF	14.0	3-6	0-2	2-2	0	3	1	1	0	0	0	8
Craig Temple Po R	FC	12.5	2-5	0-0	3-5	2	4	0	2	0	0	0	7
Frederick Engel	PG	17.8	1-6	0-0	0-0	1	1	3	2	1	0	0	2
Michael Wilson	PF	3.8	2-3	0-0	0-0	0	0	0	0	0	0	0	4

Advanced Statistics:

Stats which can't be measured easily on the floor

PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%

STL%	BLK%	TOV%	USG%		OWS	DWS	WS	WS/48

Why 1978?

SEATTLE (104)

	Min.	FG	FT	OR	DR	TR	A	PF	TP
J. Johnson	24	4-6	1-2	1	1	2	0	2	9
Shelton	32	6-11	0-0	3	5	8	1	4	12
Sikma	44	7-11	7-8	2	8	10	1	5	21
D. Johnson	44	10-14	3-4	3	4	7	6	4	23
Williams	36	6-18	4-4	1	4	5	5	3	16
Silas	30	2-5	6-8	8	1	9	1	3	10
Brown	16	7-11	1-2	0	2	2	2	2	15
Awtrey	2	0-0	0-0	0	0	0	0	0	0
Totals		42-78	22-30	18	25	43	16	23	106

PHOENIX (105)

	Min.	FG	FT	OR	DR	TR	A	PF	TP
Davis	37	10-18	6-6	0	2	2	7	2	26
Robinson	24	2-6	0-0	1	5	6	1	3	4
Kramer	32	8-10	3-4	5	3	8	2	6	19
Buse	37	3-8	0-0	1	1	2	0	2	6
Westphal	39	12-24	5-5	1	0	1	8	4	29
Scott	11	1-3	0-0	0	1	1	0	0	2
Heard	16	1-5	0-0	0	2	2	1	4	2
Forrest	14	3-3	0-0	4	1	5	2	1	6
Bratz	15	3-8	3-4	0	0	0	1	0	9
McClain	5	1-2	0-1	0	0	0	1	2	2
Totals		44-87	17-20	12	15	27	23	24	105

Seattle	27	28	22	29	— 106
Phoenix	33	17	35	20	— 105

Shooting: Seattle .538, Phoenix .506
Team rebounds: Seattle 10, Phoenix 10.
Turnovers: Seattle 21, Phoenix 15.
Steals: Seattle 8, Phoenix 8.
Blocked shots: Seattle 6, Phoenix 7.
Officials: Jack Madden, John Vanak, Jim Capers.
Attendance: 12,660.

GAME OF FRIDAY, APRIL 19
AT PHILADELPHIA

BOSTON (100)

	FG	FT	Pts
Embry	0	0- 3	0
Havlicek	7	7- 7	21
Howell	8	1- 3	17
S. Jones	9	4- 4	22
Nelson	5	0- 0	10
Russell	4	4-10	12
Seigfried	7	4- 4	18
Thacker	0	0- 0	0
Totals	40	20-31	100

PHILADELPHIA (96)

	FG	FT	Pts
Chamb'lain	4	6-15	14
Green	1	2- 6	4
Greer	8	6- 6	22
Jackson	7	1- 1	15
W. Jones	8	2- 2	18
Melchionni	0	0- 0	0
Walker	8	3- 6	19
Totals	38	20-36	96

Boston	26	20	27	27—100
Philadelphia	21	19	29	27— 96

Fouled out—S. Jones,* Howell.* Total fouls—Philadelphia 23,* Boston 28.* A—15,202.*

A modern box score

San Antonio															
Starters		**Min**	**FG**	**3Pt**	**FT**	**+/-**	**Off**	**Reb**	**Ast**	**TO**	**Stl**	**BS**	**BA**	**Pf**	**Pts**
G. Hill	G	35:40	8-15	0-1	6-9	+12	0	2	5	2	1	0	0	3	22
K. Bogans	G	16:42	2-3	1-2	4-4	-7	0	0	0	0	0	0	0	2	9
T. Ratliff	C	8:37	1-2	0-0	0-0	-2	1	1	0	0	0	1	0	3	2
R. Jefferson	F	37:17	9-16	1-3	5-8	+1	2	8	7	1	0	0	2	1	24
M. Finley	F	14:30	2-4	2-2	1-2	-5	1	3	0	1	1	0	0	0	7
Bench		**Min**	**FG**	**3Pt**	**FT**	**+/-**	**Off**	**Reb**	**Ast**	**TO**	**Stl**	**BS**	**BA**	**Pf**	**Pts**
M. Bonner		34:00	7-16	4-8	0-0	+10	1	4	2	0	0	0	0	3	18
M. Ginobili		32:14	8-15	6-8	14-16	+2	1	4	8	1	1	4	1	0	36
A. McDyess		23:44	2-5	0-0	3-5	+16	5	10	3	0	0	0	0	5	7
D. Blair		18:59	3-7	0-0	0-0	-4	4	7	1	1	0	0	1	5	6
R. Mason		17:23	0-4	0-4	0-0	-10	1	2	2	0	0	1	0	0	0
M. Hairston		0:54	0-0	0-0	0-0	-7	0	0	0	0	0	0	0	0	0
Totals			42-87	14-28	33-44		16	41	28	6	3	6	4	22	131
Percentages:			.483	.500	.750		**Team Rebounds:** 10								

How my data was obtained

Basketball-reference stores data from the every NBA/ABA season (since 1946).

Beautiful Soup was utilized to extract data from each HTML page.

I extracted Debut, Traditional Per Game, and Advanced individual basketball statistics from Basketball-reference.

My Feature Selection choice

I needed to have some kind of singular value for my dependent variable to evaluate player performance.

I decided to use the following two metrics:

WinShares $(WS)_1$

PER (Player Efficiency Rating)

1: http://www.basketball-reference.com/about/ws.html

PER: Player Efficiency Rating

PER strives to measure a player's per-minute performance, while adjust for pace of play[2].

Here's the formulate to calculate uPER (unweighted PER):

$$uPER = \frac{1}{min} \times \left(3P - \frac{PF \times lgFT}{lgPF} + \left[\frac{FT}{2} \times \left(2 - \frac{tmAST}{3 \times tmFG}\right)\right] + \left[FG \times \left(2 - \frac{factor \times tmAST}{tmFG}\right)\right] + \frac{2 \times AST}{3} + VOP \times \left[DRBP \times (2 \times ORB + BLK - 0.2464 \times [FTA - FT] - [FGA - FG] - TRB) + \frac{0.44 \times lgFTA \times PF}{lgPF} - (TO + ORB) + STL + TRB - 0.1936(FTA - FT)\right]\right)$$

$$factor = \frac{2}{3} - \left[\left(0.5 \times \frac{lgAST}{lgFG}\right) \div \left(2 \times \frac{lgFG}{lgFT}\right)\right]$$

$$VOP = \frac{lgPTS}{lgFGA - lgORB + lgTO + 0.44 \times lgFTA}$$

$$DRBP = \frac{lgTRB - lgORB}{lgTRB}$$

Here's the formulate to calculate PER after determining uPER:

$$PER = \left(uPER \times \frac{lgPace}{tmPace}\right) \times \frac{15}{lguPER}$$

2: Wikipedia: https://en.wikipedia.org/wiki/Player_efficiency_rating

PER: Player Efficiency Rating

What do the ratings correspond to?

PER	Category
35.0+	A Year for the Ages
30.0	Runaway MVP Candidate
27.5	Strong MVP Candidate
25.0	Weak MVP Candidate
22.5	Bona fide All-Star
20.0	Borderline All-Star
18.0	Solid 2nd option
16.5	3rd Banana
15.0	Average
13.0	In Rotation

WS: Win Shares[3]

WS = Offensive WS + Defensive WS

Offensive WS = Points produced - 0.92 * (1) * (Offensive possessions)

Defensive WS = (player minutes played / team minutes played) * (team defensive possessions) * (1.08 * (league points per possession) - ((Defensive Rating) / 100))

3: Basketball-reference: http://www.basketball-reference.com/about/ws.html

Pre-processing before Analysis

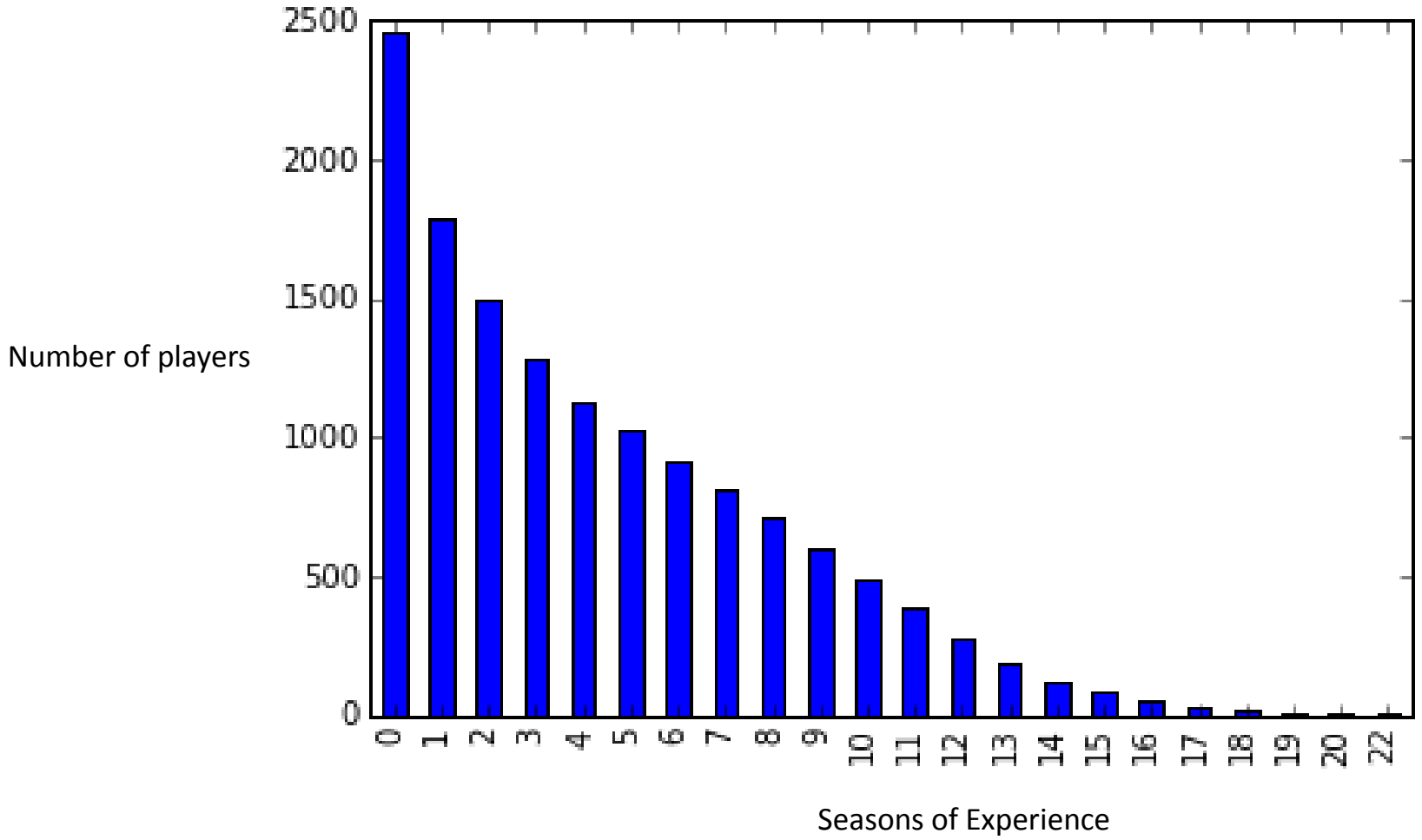
Removing Nulls

Integrating players across multiple seasons

Resolving when different players have the same name

Calculating Experience, Mean PER, and Mean WS

Playing Professional Basketball is hard!!!!



My Original Hypothesis

A player's performance early in his career will determine his success later in his career. A player who exhibits success within his first 3 seasons is likely to continue his progress into becoming an elite player.

My revised Hypothesis

A player's performance early in his career will determine his success later in his career. A player who exhibits success within his **first season** is likely to continue his progress into becoming an elite player.

Model selection

The best models that I could think to use are:

- Linear regression
- Logarithmic regression
- Decision Tree
- Naïve Bayes

Linear

PER:

MAE	RMSE
2.372	3.458

Feature analysis:

Dimension	Coef
TS%	24.72
EFG%	19.61
FGA	3.10
TRB	1.43
FT%	1.40
PTS	1.23
TRB%	1.18

Dimension	Pvalues
MP	6.15E-132
AGE	6.39E-124
AST	4.59E-109
PF	2.91E-47
FT%	1.52E-45
BLK	1.52E-38
STL	7.90E-36

WS:

MAE	RMSE
1.001	1.363

Feature analysis:

Dimension	Coef
FG%	2.325
x3P	0.705
STL	0.668
x2P	0.641
TRB	0.610
TRB	0.517

Dimension	Pvalues
age	0.00E+00
ast	1.52E-163
tov	3.94E-74
stl	3.71E-47
pf	1.04E-41
blk	1.37E-39

Logistic

PER:

	Predicted: No	Predicted: Yes
Actual: No	598	1
Actual: Yes	12	3

2.8% of population qualified

WS:

	Predicted: No	Predicted: Yes
Actual: No	592	5
Actual: Yes	12	5

3.3% of population qualified

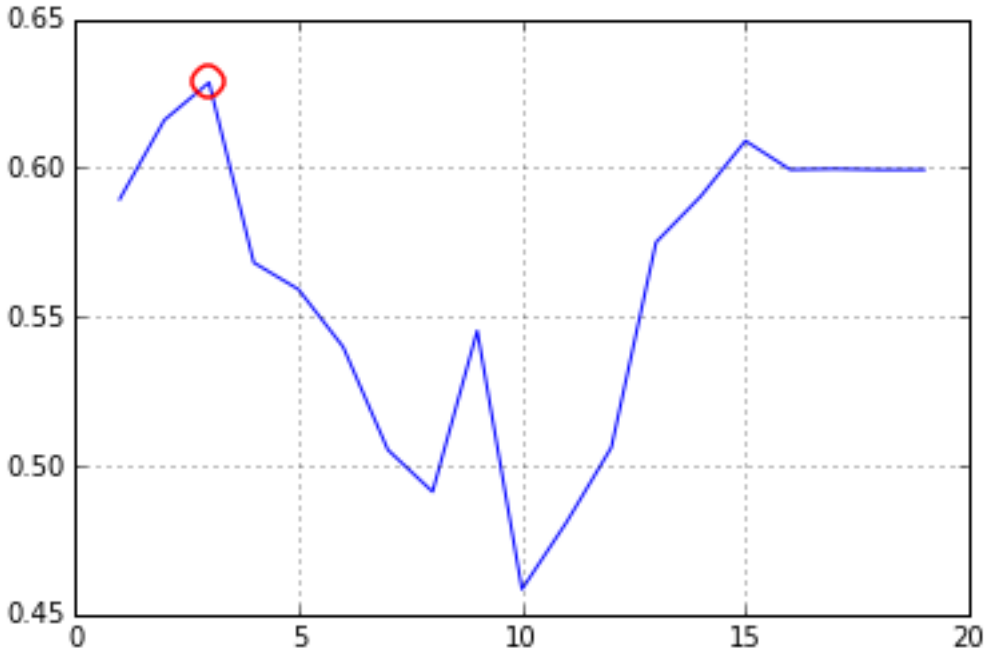
Decision Tree

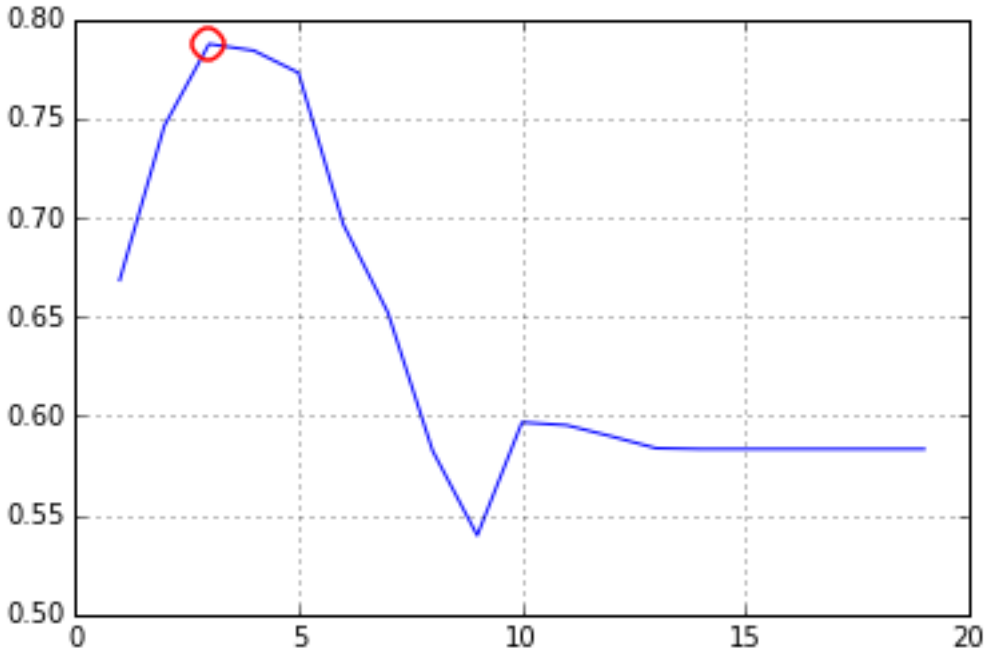
PER:

Dimension	Coef
FT	0.492
TS%	0.268
BLK%	0.239

WS:

Dimension	Coef
G	0.138
MP	0.264
TRB	0. 598





Model Cross Validation

PER:

Model	Accuracy	ROC_AUC	Specific*	Sensitive*
Log	97.6%	0.90	99.8%	20%
Decision Tree	96.6%	0.62	97.5	0
Naïve Bayes	86.8%	0.77	N/A	N/A
KNN	97.2%	0.67	N/A	N/A

WS:

Model	Accuracy	ROC_AUC	Specific*	Sensitive*
Log	96.7%	0.92	99.1%	29.4%
Decision Tree	96.6%	0.73	97.5	25%
Modified Tree	96.3%	0.86	N/A	N/A
Naïve Bayes	82.1%	0.88	N/A	N/A
KNN	96.5	0.73	N/A	N/A

Challenges:

- Reduced group of players
- Dealing with similarly named players
- Calculating PER, mean WS, and mean PER
- Extremely Low Success Rate

Successes:

- Finding strong feature correlations
- Create my own NBA RDBS


WHO'S AWESOME
YOU'RE AWESOME