

Predicting NBA's Most Valuable Player

STAT 6021 Team 5

Andrew Kromkowski, Shannon Mitchell, Kaustav Saha, Yifeng Song

07 Dec 2015

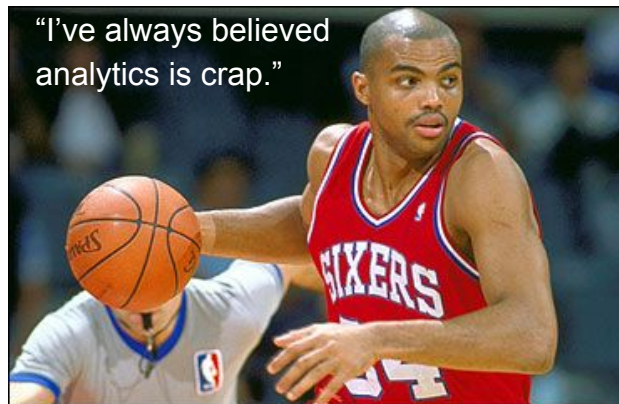
Advanced Analytics in Sports

- **Sabermetrics:** The application of statistical analysis to baseball records
- Advanced analytics have since expanded to all major sports.
- APBRMetrics is the basketball-related offspring of Sabermetrics



Critique of Advanced Analytics

- Many highly notable basketball players and coaches have criticized the increased dependency on advanced analytics in basketball
- Basketball is a highly complex game and statistics are not a suitable substitute for human evaluation



MVP Winner	PER Ranking
Stephen Curry	3
Kevin Durant	1
LeBron James	1
LeBron James	1
Derrick Rose	9
LeBron James	1
LeBron James	1
Kobe Bryant	7
Dirk Nowitzki	2
Steve Nash	14
Steve Nash	18
Kevin Garnett	1
Tim Duncan	4 (Tied)
Average	4.88

Objective

Goal: Develop a quantitative metric to model basketball player performance and predict NBA Most Valuable Player (MVP)

NBA Winners		
Season	Lg	Player
2014-15	NBA	Stephen Curry
2013-14	NBA	Kevin Durant
2012-13	NBA	LeBron James
2011-12	NBA	LeBron James
2010-11	NBA	Derrick Rose
2009-10	NBA	LeBron James
2008-09	NBA	LeBron James
2007-08	NBA	Kobe Bryant
2006-07	NBA	Dirk Nowitzki
2005-06	NBA	Steve Nash
2004-05	NBA	Steve Nash
2003-04	NBA	Kevin Garnett
2002-03	NBA	Tim Duncan
2001-02	NBA	Tim Duncan
2000-01	NBA	Allen Iverson

Data Source



															Totals										Shooting								
Player	Season	Age	Tm	Lg	G	GS	MP	FG	FGA	2P	2PA	3P	3PA	FT	FTA	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	FG%	2P%	3P%	eFG%	FT%	TS%	WS		
James Harden	2014-15	25	HOU	NBA	81	81	2981	647	1470	439	915	208	555	715	824	75	384	459	565	154	60	321	208	2217	.440	.480	.375	.511	.868	.605	16.4		
Chris Paul	2014-15	29	LAC	NBA	82	82	2857	568	1170	429	821	139	349	289	321	52	324	376	838	156	15	190	203	1564	.485	.523	.398	.545	.900	.596	16.1		
Stephen Curry	2014-15	26	GSW	NBA	80	80	2613	653	1341	367	695	286	646	308	337	56	285	341	619	163	16	249	158	1900	.487	.528	.443	.594	.914	.638	15.7		
Anthony Davis	2014-15	21	NOP	NBA	68	68	2455	642	1199	641	1187	1	12	371	461	173	523	696	149	100	200	95	141	1656	.535	.540	.083	.536	.805	.591	14.0		
DeAndre Jordan	2014-15	26	LAC	NBA	82	82	2820	379	534	378	530	1	4	187	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Jimmy Butler	2014-15	25	CHI	NBA	65	65	2513	421	912	348	719	73	193	386	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Damian Lillard	2014-15	24	POR	NBA	82	82	2925	590	1360	394	788	196	572	344	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Russell Westbrook	2014-15	26	OKC	NBA	67	67	2302	627	1471	541	1183	86	288	546	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Pau Gasol	2014-15	34	CHI	NBA	78	78	2681	570	1153	558	1127	12	26	294	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Kyrie Irving	2014-15	22	CLE	NBA	75	75	2730	578	1235	421	857	157	378	315	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
LeBron James	2014-15	30	CLE	NBA	69	69	2493	624	1279	504	940	120	339	375	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Tyson Chandler	2014-15	32	DAL	NBA	75	75	2286	293	440	293	440	0	0	185	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Marc Gasol	2014-15	30	MEM	NBA	81	81	2687	530	1072	527	1055	3	17	350	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Tim Duncan	2014-15	38	SAS	NBA	77	77	2227	419	819	417	812	2	7	230	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Rudy Gobert	2014-15	22	UTA	NBA	82	37	2158	258	427	258	425	0	2	170	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Blake Griffin	2014-15	25	LAC	NBA	67	67	2356	574	1144	564	1119	10	25	311	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Klay Thompson	2014-15	24	GSW	NBA	77	77	2455	602	1299	363	754	239	545	225	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Gordon Hayward	2014-15	24	UTA	NBA	76	76	2618	484	1087	364	757	120	330	375	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Al Horford	2014-15	28	ATL	NBA	76	76	2318	519	965	508	929	11	36	107	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		
Kevin Love	2014-15	26	CLE	NBA	75	75	2532	413	952	269	560	144	392	258	131	203	222	1235	51	81	122	100	245	816	.310	.312	.258	.311	.397	.638	12.8		

Most Valuable Player

				Voting			
Rank	Player	Age	Tm	First	Pts Won	Pts Max	Share
1	Stephen Curry	26	GSW	100.0	1198.0	1300	0.922
2	James Harden	25	HOU	25.0	936.0	1300	0.720
3	LeBron James	30	CLE	5.0	552.0	1300	0.425
4	Russell Westbrook	26	OKC	0.0	352.0	1300	0.271
5	Anthony Davis	21	NOP	0.0	203.0	1300	0.156
6	Chris Paul	29	LAC	0.0	124.0	1300	0.095
7	LaMarcus Aldridge	29	POR	0.0	6.0	1300	0.005
9T	Marc Gasol	30	MEM	0.0	3.0	1300	0.002
9T	Blake Griffin	25	LAC	0.0	3.0	1300	0.002
12T	Tim Duncan	38	SAS	0.0	1.0	1300	0.001
12T	Kawhi Leonard	23	SAS	0.0	1.0	1300	0.001
12T	Klay Thompson	24	GSW	0.0	1.0	1300	0.001

Data Pre-processing

REMOVE separator rows from player stats

SUBSET seasons 1956-2015

MERGE player stats and MVP votes

IMPUTE missing values using 'Amelia'

FINAL DF: 18026 rows x 36 columns

amelia {Amelia}

R Documentation

AMELIA: Multiple Imputation of Incomplete Multivariate Data

Description

Runs the bootstrap EM algorithm on incomplete data and creates imputed datasets.

Usage

```
## Default S3 method:
amelia(x, m = 5, p2s = 1, frontend = FALSE, idvars = NULL,
       ts = NULL, cs = NULL, polytime = NULL, splintime = NULL, intercs = FALSE,
       lags = NULL, leads = NULL, startvals = 0, tolerance = 0.0001,
       logs = NULL, sqrts = NULL, lgstc = NULL, noms = NULL, ords = NULL,
       incheck = TRUE, collect = FALSE, arglist = NULL, empri = NULL,
       priors = NULL, autopri = 0.05, emburn = c(0,0), bounds = NULL,
       max.resample = 100, overimp = NULL, boot.type = "ordinary",
       parallel = c("no", "multicore", "snow"),
       ncpus = getOption("amelia.ncpus", 1L), cl = NULL, ...)
```


Predictive Modeling

VARIABLE SELECTION - Lasso/Forward-Backward Subset Selection / Multi-Collinearity through VIFs (FGA, FG, TRB, PtsWon, PtsMax, Share removed)

MODELS - 4 Models :- Model A had 16 variables with interaction terms. Model B had 4 variables. Minus_Baseline and Baseline Models.(Win_Share -- Score based on the contribution of a player in the winnings of a team).

DATASETS - Pre-1980/Post-1980 (Player Voted/Media Voted)

OUTLIERS - Detect and remove Influential & Leverage Points using DFFITS, DFBETAS, Cook's D, COVRATIO.

ANALYSIS & EVALUATION - Cross-Validation (leave-one-out), Model Adequacy Checking (Test for Significance), Residual Analysis

Leave One (Season) Out Cross Validation

TRAIN model on all seasons except season XXXX

PREDICT season XXXX vote shares

ITERATE over all seasons

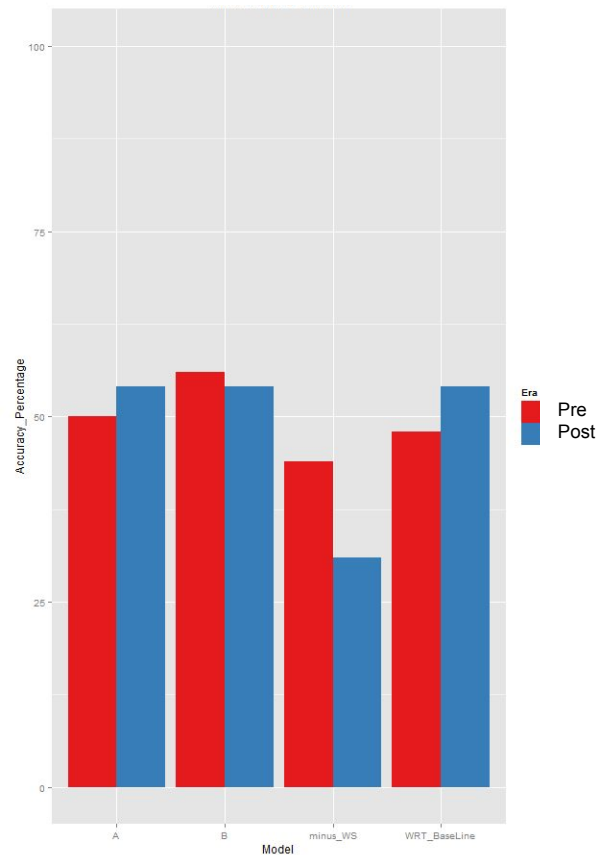
COMPARE the top 3 players according to predicted vote shares to actual MVP winners for each season

Results: Top 3 predicted players

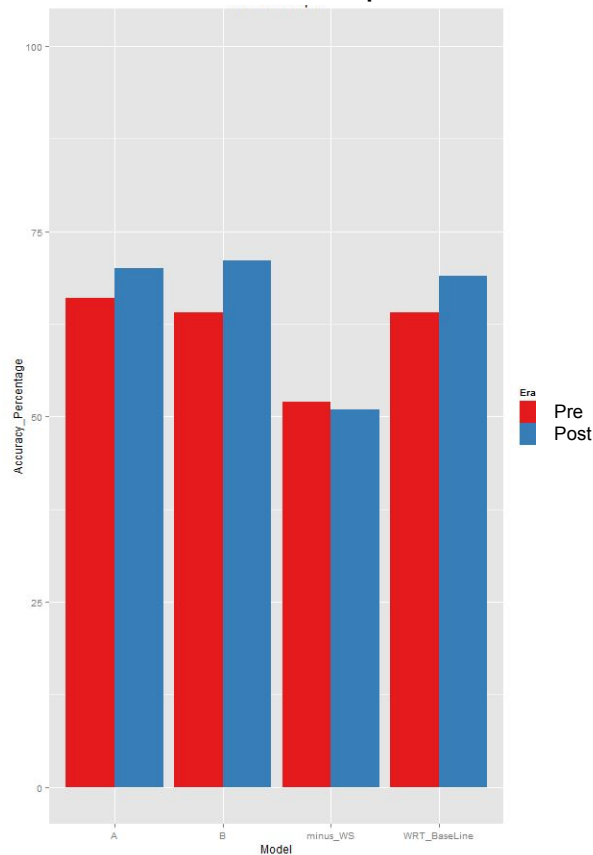
Model	Predictors	Correct MVP Predicted	MVP within top 2	MVP within top 3
A	16 variables including quadratic and interaction terms	56%	70%	80%
B	Minutes Played^2 Win Shares^2 Blocks Personal Fouls	57%	70%	80%
Minus Baseline	Minutes Played Defensive Rebounds Personal Fouls^2 Free Throws 2 Pt Field Goals 3 Pt Field Goals Assists Blocks	50%	67%	83%
Baseline	Win Shares	53%	68%	77%

Comparing 1956-1980 vs. 1981-2015

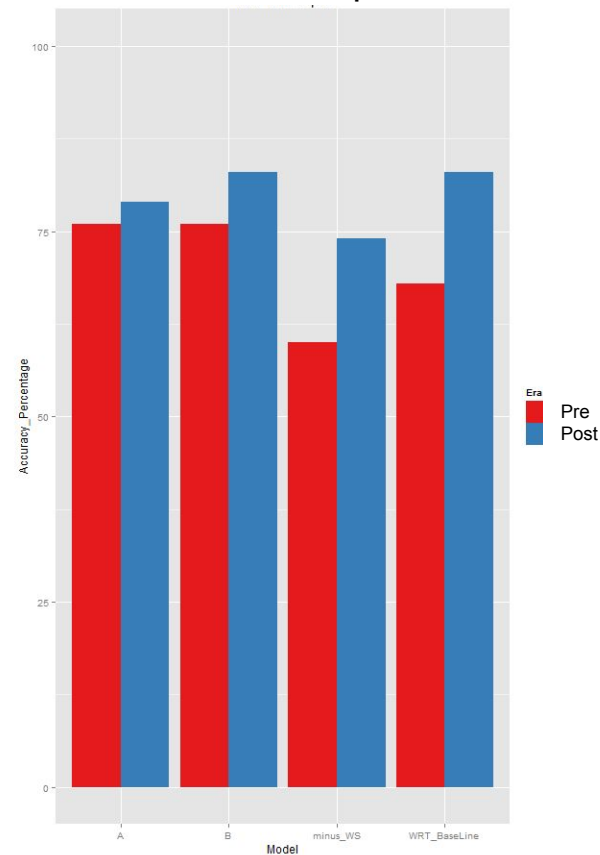
Correct MVP Predicted



MVP within top 2



MVP within top 3



Queries

Top 5 player seasons:

1964 Wilt Chamberlain

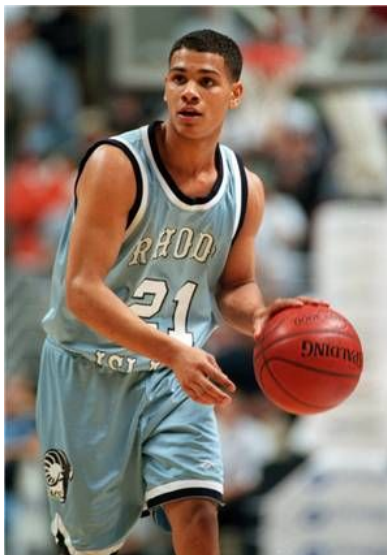
1972 Kareem Abdul-Jabbar

1962 Wilt Chamberlain

1988 Michael Jordan

1964 Oscar Robertson

Lowest single season:



Biggest Snub:

2011 MVP: Derrick Rose
Score= **1.276**

Best Player: LeBron James
Score= **1.8877**



Interesting Hurdles

Merging tables: people with same name

Missing values: more missing values in 1999 than other seasons

Response imbalance: 17155/18026 with vote share of 0

Conclusions

MVP can be correctly predicted in 50-60% of the previous seasons

Basketball-reference has engineered “Win Shares” which has a good predictive power, but it can be improved using additional factors

Possible future work: correlate team revenue with predicted MVP vote shares, could help owners and managers make roster decisions

Questions?