Big Data Analytics Symposium - Summer 2020

Analytics Project:

Traditional vs. Advanced Statistics and the NBA Regular Season MVP Award

Team:

Alex Spence

Data Science, NYU GSAS

ajs811@nyu.edu

Mike Urciuoli

Computer Science, NYU GSAS

mlu216@nyu.edu

Traditional vs. Advanced Statistics and MVP

Agenda

- Abstract
- Motivation
- Goodness
- Data Sources
- Data Samples
- Design Diagram
- Code Challenges
- Results
- Obstacles
- Summary
- Future Works
- Acknowledgements
- References

Traditional vs. Advanced Statistics and MVP

Abstract

- Traditional NBA player statistics have been collected since the onset of the NBA in 1946.
- Advanced stats started being collected in the 1990s.
- Which statistics are most correlated with the MVP award?
  - The seasons used for comparison were 2002-03 to 2016-17.
  - Tools used were Hadoop MapReduce and Apache Impala
  - "Estimated wins added", "value added" and "player efficiency rating" (all advanced statistics) were determined to be most correlated.
- Able to predict MVP using most correlated statistics for 10/16 seasons
  - More reliable during later seasons

Traditional vs. Advanced Statistics and MVP

Motivation

- Who are the users?
  - NBA franchises
  - NBA players
  - Media
- Who will benefit?
  - The above
  - Also fans
- Why is this analytic important?
  - Provides an objective assessment of player performance as it relates to the regular season MVP
  - Determines which statistics are most informative

Traditional vs. Advanced Statistics and MVP
Goodness

- What steps were taken to assess the 'goodness' of the analytic?
  - Simple statistics and rankings of typical player statistics are performed per season.
  - Data readily available from sites like ESPN.com and NBA.com are used to spot check calculations
  - Domain knowledge allows personal assessment of goodness
  - Actual MVP winners are predicted using this analytic

Traditional vs. Advanced Statistics and MVP

Data Sources

- Advanced Player Statistics
  - 2002-03 to 2017-18 seasons
  - NBA advanced player statistics organized by season
  - Ex) “Player efficiency rating”, “estimated wins added”
  - Size: 0.6 MB
- Traditional Player Statistics
  - 1949-50 to 2016-2017 seasons
  - NBA traditional player statistics organized by season
  - Ex) Points scored, rebounds, assists
  - Size: 5.1 MB
- Regular Season MVPs
  - 1949-50 to 2018-19 seasons
  - NBA MVPs organized by season
  - Size of data: 4 KB

Traditional vs. Advanced Statistics and MVP

Data Sample - Advanced Player Statistics

season_player	team	per	va	ewa	ts_perc	usg	orr
2003_Tracy McGrady	ORL	31.01	452.00	15.10	0.56	32.60	5.00
2003_Shaquille O'Neal	LAL	29.43	356.20	11.90	0.60	27.80	11.00
2003_Kobe Bryant	LAL	27.07	420.60	14.00	0.55	31.10	3.00
2003_Kevin Garnett	MIN	26.46	370.80	12.40	0.55	25.00	9.00
2003_Tim Duncan	SA	26.46	355.00	11.80	0.56	25.70	10.00
2003_Dirk Nowitzki	DAL	25.94	336.00	11.20	0.58	24.80	3.00
2003_Steve Nash	DAL	23.51	253.10	8.40	0.58	24.00	3.00
2003_Paul Pierce	BOS	23.49	300.00	10.00	0.53	30.10	4.00
2003_Sam Cassell	MIL	23.08	243.40	8.10	0.56	25.20	2.00
2003_Jason Kidd	NJ	22.62	259.30	8.60	0.53	26.00	4.00
2003_Vince Carter	TOR	22.48	131.50	4.40	0.53	26.30	4.00
2003_Ray Allen	MIL/SEA	22.19	251.10	8.40	0.56	25.40	3.80
2003_Allen Iverson	PHI	22.14	302.80	10.10	0.50	30.90	2.00
2003_Gary Payton	MIL/SEA	21.94	262.00	8.70	0.50	24.20	2.90
2003_Karl Malone	UTAH	21.71	223.80	7.50	0.53	26.40	5.00
2003_Chris Webber	SAC	21.48	212.90	7.10	0.49	28.00	7.00
2003_Michael Redd	MIL	21.33	187.10	6.20	0.59	20.30	5.00
2003_Jermaine O'Neal	IND	21.16	206.50	6.90	0.54	23.80	8.00
2003_Steve Francis	HOU	21.12	250.60	8.40	0.54	25.20	6.00
2003_Chauncey Billups	DET	21.10	175.30	5.80	0.58	22.80	2.00

Traditional vs. Advanced Statistics and MVP

Data Sample - Traditional Player Statistics

season_player	pts	pos	fg_perc	thp_perc	ft_perc	orb	ast	stl	blk	tov
2003_Brian Scalabrine	180	PF	0.40	0.36	0.83	40	46	16	18	46
2003_Brian Shaw	250	PG	0.39	0.35	0.67	20	103	32	13	54
2003_Bruce Bowen	583	SF	0.47	0.44	0.40	59	113	66	42	72
2003_Jahidi White	67	C	0.47	NaN	0.68	37	2	1	12	9
2003_Kenny Thomas	1334	PF	0.47	NaN	0.75	370	224	124	56	232
2003_Loren Woods	80	C	0.38	0.33	0.78	27	19	10	13	23
2003_Predrag Savovic	83	SG	0.31	0.15	0.72	9	22	14	1	21
2003_Raef LaFrentz	639	C	0.52	0.41	0.68	125	54	35	91	46
2003_Ricky Davis	1626	SG	0.41	0.36	0.75	97	436	125	36	277
2003_Ryan Bowen	223	SG	0.49	0.29	0.66	78	54	65	29	43
2003_Wesley Person	727	SG	0.46	0.43	0.81	24	112	42	19	56
2004_Adrian Griffin	11	SG	0.28	0.50	0.00	1	10	7	2	3
2004_Andrei Kirilenko	1284	PF	0.44	0.34	0.79	226	244	150	215	215
2004_Andrew DeClercq	230	C	0.48	NaN	0.81	131	44	47	32	54
2004_Ansu Sesay	200	SF	0.46	0.29	0.70	43	19	19	20	23
2004_Bruce Bowen	565	SF	0.42	0.36	0.58	45	113	84	33	90
2004_Chris Bosh	861	C	0.46	0.36	0.70	191	78	59	106	107
2004_Chris Crawford	569	PF	0.45	0.39	0.87	58	45	37	20	55
2004_Desmond Ferguson	13	SF	0.42	0.38	0.00	0	1	0	0	1
2004_Joel Przybilla	98	C	0.36	NaN	0.42	64	14	10	34	36

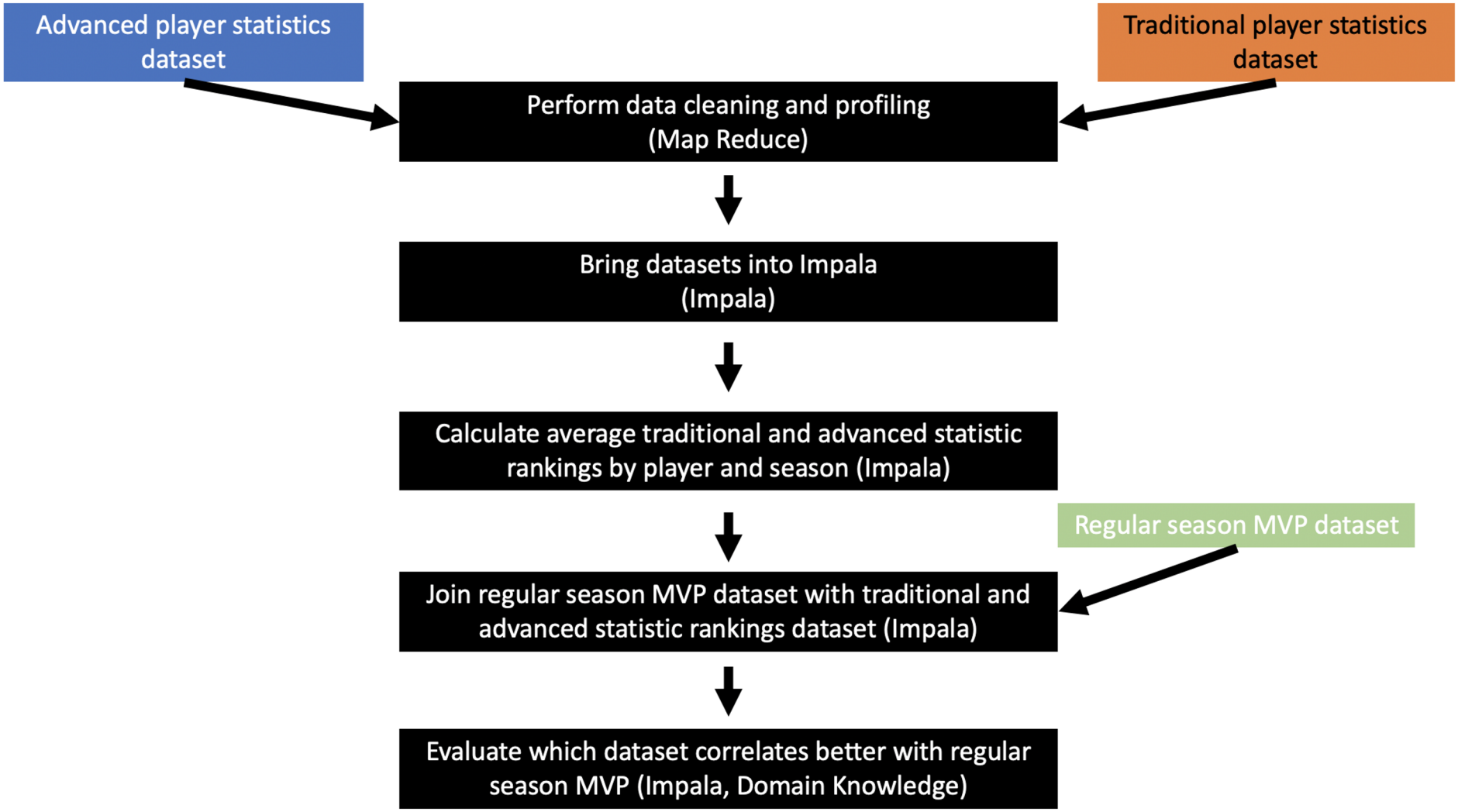
Traditional vs. Advanced Statistics and MVP

Data Sample - Regular Season MVPs

season_player	player	team
2019_Giannis Antetokounmpo	Giannis Antetokounmpo	Milwaukee Bucks
2018_James Harden	James Harden	Houston Rockets
2017_Russell Westbrook	Russell Westbrook	Oklahoma City Thunder
2016_Stephen Curry	Stephen Curry	Golden State Warriors
2015_Stephen Curry	Stephen Curry	Golden State Warriors
2014_Kevin Durant	Kevin Durant	Oklahoma City Thunder
2013_LeBron James	LeBron James	Miami Heat
2012_LeBron James	LeBron James	Miami Heat
2011_Derrick Rose	Derrick Rose	Chicago Bulls
2010_LeBron James	LeBron James	Cleveland Cavaliers
2009_LeBron James	LeBron James	Cleveland Cavaliers
2008_Kobe Bryant	Kobe Bryant	Los Angeles Lakers
2007_Dirk Nowitzki	Dirk Nowitzki	Dallas Mavericks
2006_Steve Nash	Steve Nash	Phoenix Suns
2005_Steve Nash	Steve Nash	Phoenix Suns
2004_Kevin Garnett	Kevin Garnett	Minnesota Timberwolves
2003_Tim Duncan	Tim Duncan	San Antonio Spurs
2002_Tim Duncan	Tim Duncan	San Antonio Spurs
2001_Allen Iverson	Allen Iverson	Philadelphia 76ers
2000_Shaquille O'Neal	Shaquille O'Neal	Los Angeles Lakers

Traditional vs. Advanced Statistics and MVP

Design Diagram



Traditional vs. Advanced Statistics and MVP

Code Challenge 1

- MapReduce (Cleaning)
  - Ignoring the header row of data
  - Converting YYY1-YYY2 to YYY2

```
//If it's the first line (i.e. header row, don't output anything)
Long line_idx = key.get();
Long zero = new Long(0);
if (line_idx.equals(zero)) {
    String placeholder = "nothing";
}
```

```
//Get season and convert to YYYY
int season = Integer.parseInt(items[16].substring(0,4)) + 1;
```

- Impala (Developing Analytic)
  - Develop rank code

```
Create table fg_perc_rank as
select season_player, rank() over(partition by season order by fg_perc desc) as
fg_perc_rank from trad;
```

Traditional vs. Advanced Statistics and MVP

Code Challenge 2

- Impala (Developing Analytic)
  - Staying organized with long queries with many columns and joins!

```
//Create joined traditional stats ranking table

Create table trad_rank as
select age_rank.season, age_rank.season_player, age_rank.age_rank,
fg_rank.fg_rank, fga_rank.fga_rank, fg_perc_rank.fg_perc_rank,
thp_rank.thp_rank, thpa_rank.thpa_rank, thp_perc_rank.thp_perc_rank,
twop_rank.twop_rank, twopa_rank.twopa_rank, twop_perc_rank.twop_perc_rank,
ft_rank.ft_rank, fta_rank.fta_rank, ft_perc_rank.ft_perc_rank,
orb_rank.orb_rank, drb_rank.drb_rank, trb_rank.trb_rank,
ast_rank.ast_rank, stl_rank.stl_rank, blk_rank.blk_rank,
tov_rank.tov_rank, pf_rank.pf_rank, pts_rank.pts_rank

from age_rank inner join
fg_rank inner join fga_rank inner join fg_perc_rank inner join
thp_rank inner join thpa_rank inner join  thp_perc_rank inner join
twop_rank inner join twopa_rank inner join  twop_perc_rank inner join
ft_rank inner join fta_rank inner join  ft_perc_rank inner join
orb_rank inner join drb_rank inner join  trb_rank inner join
ast_rank inner join stl_rank inner join  blk_rank inner join
tov_rank inner join pf_rank inner join pts_rank

on age_rank.season_player = fg_rank.season_player and
fg_rank.season_player = fga_rank.season_player and
fga_rank.season_player = fg_perc_rank.season_player and
fg_perc_rank.season_player = thp_rank.season_player and
thp_rank.season_player = thpa_rank.season_player and
thpa_rank.season_player = thp_perc_rank.season_player and
thp_perc_rank.season_player = twop_rank.season_player and
twop_rank.season_player = twopa_rank.season_player and
twopa_rank.season_player = twop_perc_rank.season_player and
twop_perc_rank.season_player = ft_rank.season_player and
ft_rank.season_player = fta_rank.season_player and
fta_rank.season_player = ft_perc_rank.season_player and
ft_perc_rank.season_player = orb_rank.season_player and
orb_rank.season_player = drb_rank.season_player and
drb_rank.season_player = trb_rank.season_player and
trb_rank.season_player = ast_rank.season_player and
ast_rank.season_player = stl_rank.season_player and
stl_rank.season_player = blk_rank.season_player and
blk_rank.season_player = tov_rank.season_player and
tov_rank.season_player = pf_rank.season_player and
pf_rank.season_player = pts_rank.season_player;
```

Traditional vs. Advanced Statistics and MVP

Code Challenge 3

- MapReduce (Cleaning)
  - In the traditional stats dataset, players who were transferred from one team to another during the same season appeared more than once, with their stats split between the different teams
  - There were also instances of multiple players with the same name in one season
- Solution
  - Concatenate year, player, and age during map to send unique player records to reduce phase

ID	Year	Player	Pos	Age	Team	G	GS
0	1950	Curly Armstrong	G-F	31	FTW	63	null
1	1950	Cliff Barker	SG	29	INO	49	null
2	1950	Leo Barnhorst	SF	25	CHS	67	null
3	1950	Ed Bartels	F	24	TOT	15	null
4	1950	Ed Bartels	F	24	DNN	13	null
5	1950	Ed Bartels	F	24	NYK	2	null
6	1950	Ralph Beard	G	22	INO	60	null
7	1950	Gene Berce	G-F	23	TRI	3	null
8	1950	Charlie Black	F-C	28	TOT	65	null
9	1950	Charlie Black	F-C	28	FTW	36	null
10	1950	Charlie Black	F-C	28	AND	29	null
11	1950	Nelson Bobb	PG	25	PHW	57	null
12	1950	Jake Bornheimer	F-C	22	PHW	60	null
13	1950	Vince Boryla	SF	22	NYK	59	null
14	1950	Don Boven	F-G	24	WAT	62	null
15	1950	Harry Boykoff	C	27	WAT	61	null
16	1950	Joe Bradley	G	21	CHS	46	null
17	1950	Bob Brannum	PF	24	SHE	59	null
18	1950	Carl Braun	G-F	22	NYK	67	null
19	1950	Frankie Brian	G	26	AND	64	null

Traditional vs. Advanced Statistics and MVP

Results - Definitions

Advanced Statistics

- **EWA**: estimated wins added
- **VA**: value added
- **PER**: player efficiency rating
- **USG**: usage rating
- **TS%**: true shooting %

Traditional Statistics

- **PTS**: points
- **FG**: field goals made
- **TOV**: turnovers
- **FT**: free throws made
- **AST**: assists

Traditional vs. Advanced Statistics and MVP

Results

- Most informative advanced statistics by average MVP ranks
  - EWA (3.8) / VA (3.8) / PER (4.1) / USG (10.6) / TS% (35.9)
- Most informative traditional statistics by average MVP ranks
  - PTS (12.6) / FG (12.67) / TOV (15.2) / FT (17.1) / AST (17.6)
- Composite advanced statistics are the most informative overall
  - EWA, VA, PER

Traditional vs. Advanced Statistics and MVP
Results

- Advanced statistics seem to be valued less in earlier years
  - 2003 to 2010 average MVP EWA (5.75 rank / 593 value)
  - 2009 to 2017 average MVP EWA (1.67 rank / 810 value)
  - 2012 to 2018 MVP EWA rank is 1 for 5/7 seasons
- 2005 and 2006 has the lowest ranked MVP EWA
  - Steve Nash (18 rank, 15 rank)
- Using average rank of EWA, VA, and PER accurately predicts the MVP in 10/16 seasons from 2003 to 2018

Traditional vs. Advanced Statistics and MVP

Results

season_player
2003_Tim Duncan
2004_Kevin Garnett
2005_Steve Nash
2006_Steve Nash
2007_Dirk Nowitzki
2008_Kobe Bryant
2009_LeBron James
2010_LeBron James
2011_Derrick Rose
2012_LeBron James
2013_LeBron James
2014_Kevin Durant
2015_Stephen Curry
2016_Stephen Curry
2017_Russell Westbrook
2018_James Harden

season_player	predictor
2003_Tracy McGrady	1
2004_Kevin Garnett	1
2005_Kevin Garnett	1.666666666666667
2006_LeBron James	1.333333333333333
2007_Dirk Nowitzki	2
2008_LeBron James	1
2009_LeBron James	1
2010_LeBron James	1
2011_LeBron James	1
2012_LeBron James	1
2013_LeBron James	1
2014_Kevin Durant	1
2015_Anthony Davis	1.666666666666667
2016_Stephen Curry	1
2017_Russell Westbrook	1
2018_LeBron James	1.666666666666667
2018_James Harden	1.666666666666667

Traditional vs. Advanced Statistics and MVP

Obstacles

- More players provided in traditional than advanced dataset per year
  - Solution: treat datasets independently
  - Join on the MVP dataset for final average rank calculations
- Distinguishing between string and integer fields in the profile reducer
  - Solution: Use try and except clauses
- Connecting Impala tables to Tableau

Traditional vs. Advanced Statistics and MVP

Summary

- Hadoop MapReduce and Apache Impala used to compare traditional and advanced statistics and evaluate the correlation to the NBA regular season MVP award.
- During 2002-03 to 2016-17 seasons:
  - Most correlated advanced statistic
    - EWA, VA, and PER
  - Most correlated traditional statistics
    - PTS, FG, and TOV
- Overall EWA, VA, and PER are most useful
  - Able to predict MVP in 10/16 seasons when average of values is taken
  - Better predictor in most recent years

Traditional vs. Advanced Statistics and MVP

Future Works

- Include more years using NBA API or Basketball-Reference API
- Develop machine learning model with the most informative statistics to predict current regular season MVP
- Utilize Tableau for visualizations
- Develop analytics for other awards
  - Defensive player of the year
  - Rookie of the Year
  - Etc.

Traditional vs. Advanced Statistics and MVP

Acknowledgements

- Thank you to Prof. McIntosh for answering our analysis questions!
- Thank you to Kaggle for compiling the data sets!
- Thank you to HPC for providing the technology to make this project possible!
- Thank you to the NBA for providing a medium for professional basketball!

Traditional vs. Advanced Statistics and MVP

References

1. A Neural Network Model of the NBA Most Valued Player Selection Prediction (Chen, Dai, Zhang)
2. Key Game Indicators in NBA Players' Performance Profiles (Dehesa, Vaquera, Goncalves, Mateus, Gomez-Ruano, Sampaio)
3. Predicting Per Game Performance Through Per Minute Performance in Basketball (Martinez)
4. Game-to-Game Prediction of NBA Players' Points in Relation to Their Season Average (Zovak, Šarčević, Pintar)
5. Revisiting the Correlation of Basketball Stats and Match Outcome Prediction (Li)
6. Sports analytics— Evaluation of basketball players and team performance (Sarlis, Tjortjis)

Traditional vs. Advanced Statistics and MVP

Questions?

Big Data Analytics Symposium - Summer 2020

Analytics Project:

Traditional vs. A
and the NBA I
Award

THANK YOU!

Team:

Alex Spence
Data Science, NYU GSAS
ajs811@nyu.edu

Mike Urciuoli
Computer Science, NYU GSAS
mlu216@nyu.edu