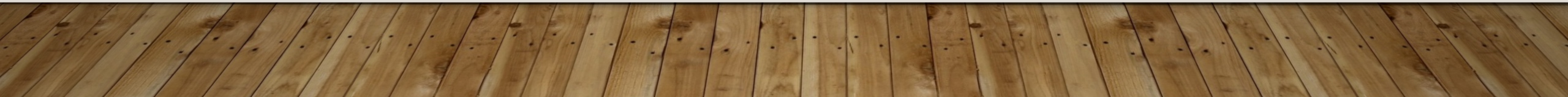
PREDICTING NBA HOME GAME ATTENDANCE

W/ LINEAR REGRESSION

PROJECT #2

RUDY WANG

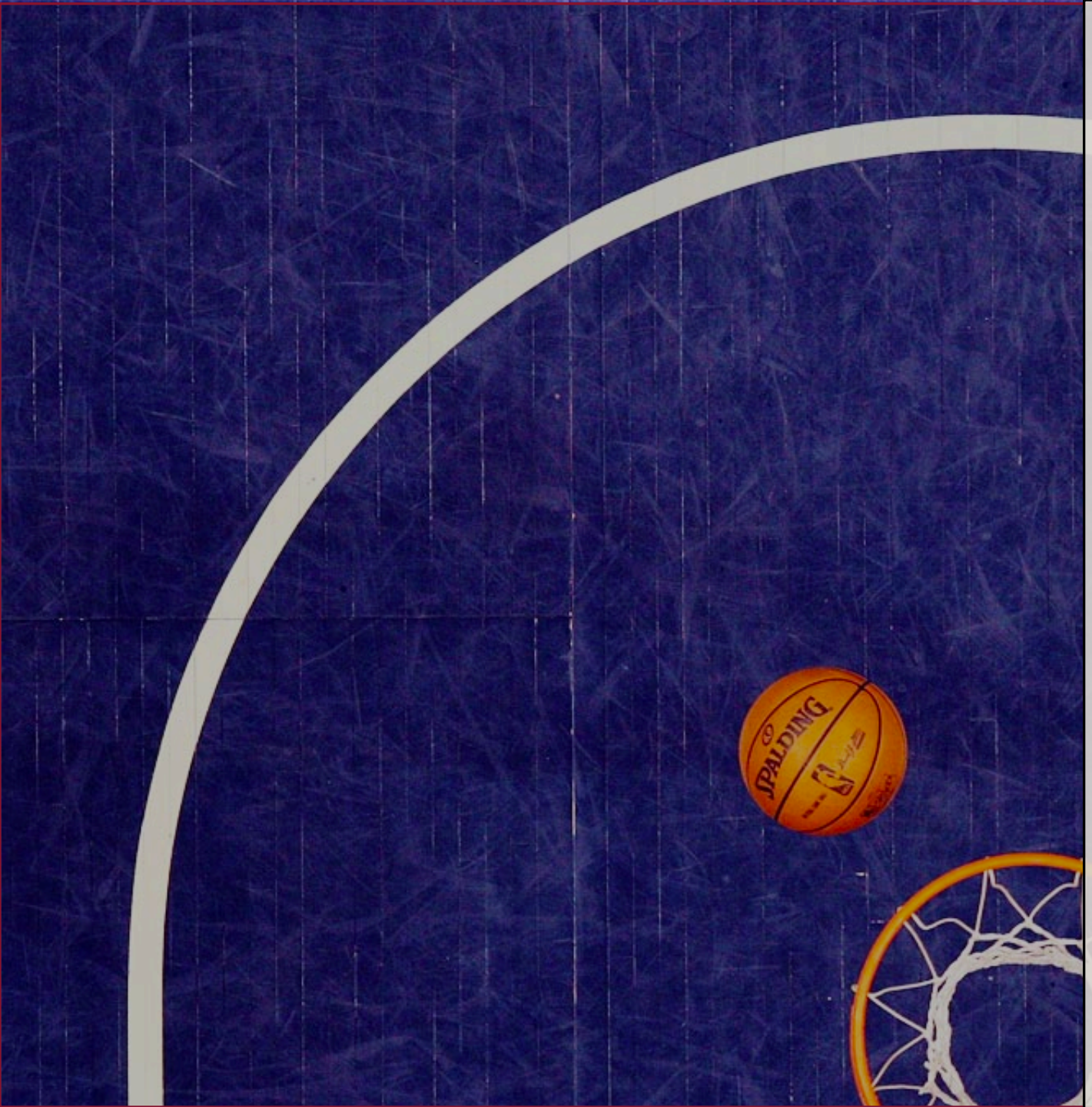
JULY 17, 2020



WHY IS ATTENDANCE IMPORTANT?

- Is it a really an NBA game if there are no fans?
- *Attendance is ticket sales – not body count.
- A measure of popularity and bragging rights – loudest fans in the house, etc.
- Higher Attendance… More revenue a team can generate through food, drinks, merchandise.
- Business sense: finding the sweet spot for either increasing/decreasing more capacity, increase/decrease ticket prices, or getting 'popular' players on the team.

$$\% \, Filled = \frac{Average \; Home \; Game \; Attendance}{Total \; Arena \; Capacity}$$



METHODS/TOOLS

Data Sources:

- Basketball-Reference (2000-2019 Individual Team Stats, 2000-2019, 2000-2019 NBA All Star Stats)
  - 40+ Features
  - 565 Observations
- HISPANOSNBA (Total NBA Championships)
  - 20+ Observations
- Wikipedia – Stadium Capacity Information
  - 30+ Observations

Python Tools:

- BeautifulSoup
- Selenium
- Pandas/Seaborn/Matplotlib
- Scikit-Learn
- Requests

WEB(B) SCRAPING

Basketball-Reference (Selenium):

- Rotate different years, different teams
- Individual Team Stats (2 Tables)
- All Star Stats (2 Tables)

Team and Opponent Stats Ranks are per game (except for MP, which are total) and sorted descending (except for TOV and PF); opponents ranked are flippe

	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
Team	82	19830	2771	6455	.429	293	907	.323	2478	5548	.447	1552	2138	.726	1178	2203	3381	1694	649	441	1258	1617	7387
Team/G		241.8	33.8	78.7	.429	3.6	11.1	.323	30.2	67.7	.447	18.9	26.1	.726	14.4	26.9	41.2	20.7	7.9	5.4	15.3	19.7	90.1
Lg Rank		10	28	16	26	22	18	24	23	13	26	16	16	20	6	25	17	22	18	10	15	1	27
Year/Year		-0.1%	-2.4%	-0.6%	-.008	-48.3%	-45.4%	-.018	9.1%	14.7%	-.023	5.3%	8.3%	-.020	10.0%	-0.8%	2.7%	0.3%	-6.5%	21.5%	0.6%	-1.6%	-4.3%
Opponent	82	19830	2930	6460	.454	313	967	.324	2617	5493	.476	1302	1769	.736	1089	2268	3357	1740	747	388	1259	1866	7475
Opponent/G		241.8	35.7	78.8	.454	3.8	11.8	.324	31.9	67.0	.476	15.9	21.6	.736	13.3	27.7	40.9	21.2	9.1	4.7	15.4	22.8	91.2
Lg Rank		10	15	13	17	5	7	5	19	15	18	1	1	15	17	11	12	8	24	13	13	10	7
Year/Year		-0.1%	-1.2%	0.2%	-.006	-31.2%	-27.1%	-.019	4.2%	7.2%	-.014	-4.3%	-4.9%	+.004	0.8%	-5.4%	-3.5%	-3.2%	6.1%	-7.0%	-4.2%	5.5%	-3.5%

Team Misc

											Advanced		Offense Four Factors				Defense Four Factors					
	W	L	PW	PL	MOV	SOS	SRS	ORtg	DRtg	Pace	FTr	3PAr	eFG%	TOV%	ORB%	FT/FGA	eFG%	TOV%	DRB%	FT/FGA	Arena	Attendance
Team	41	41	38	44	-1.07	0.54	-0.53	102.6	103.8	87.1	.331	.141	.452	14.5	34.2	.240	.478	14.8	66.9	.202	Orlando Arena	667,322
Lg Rank	17	12	19	19	19	4	18	22	12	29	14	17	27	14	3	18	14	10	20	3		18

West Share & more ▾ Glossary

		Totals											
Starters	Tm	MP	FG	FGA	3P	3PA	FT	FTA	ORB	DRB	TRB	AST	STL
Kevin Garnett	MIN	35:00	10	19	0	1	4	4	3	7	10	5	1
Jason Kidd	PHO	34:00	4	9	3	6	0	0	0	5	5	14	4
Tim Duncan	SAS	33:00	12	14	0	0	0	0	7	7	14	4	1
Kobe Bryant	LAL	28:00	7	16	1	4	0	0	1	0	1	3	2
Shaquille O'Neal	LAL	25:00	11	20	0	0	0	2	4	5	9	3	0
		Totals											

TRAINING VS TESTING

Divided the datasets into 2 parts:

Training/Validation Data:

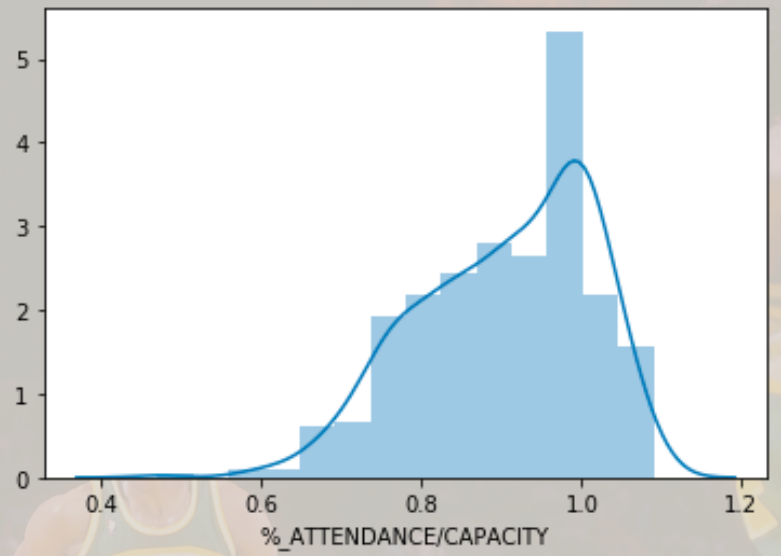
- NBA Season 2000 – 2015* (80%)
- 445 Rows

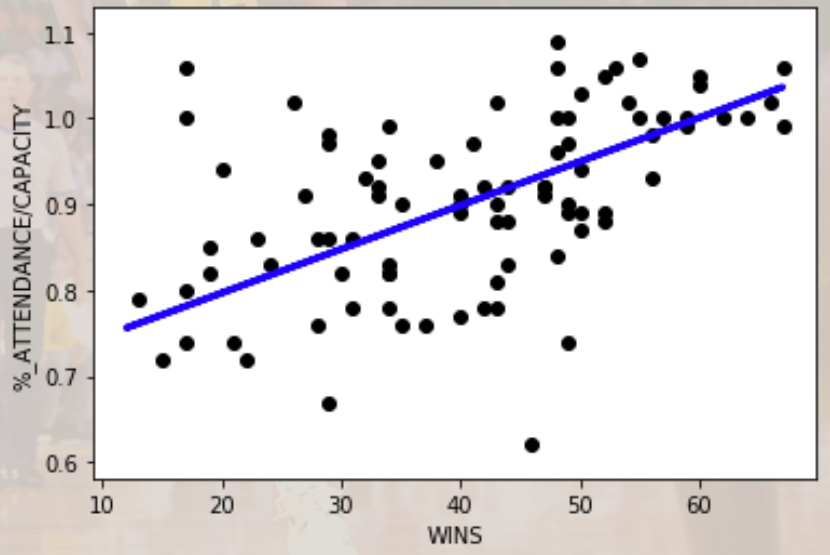
Testing Data:

- NBA Seasons 2016 – 2019 (20%)
- 120 Rows

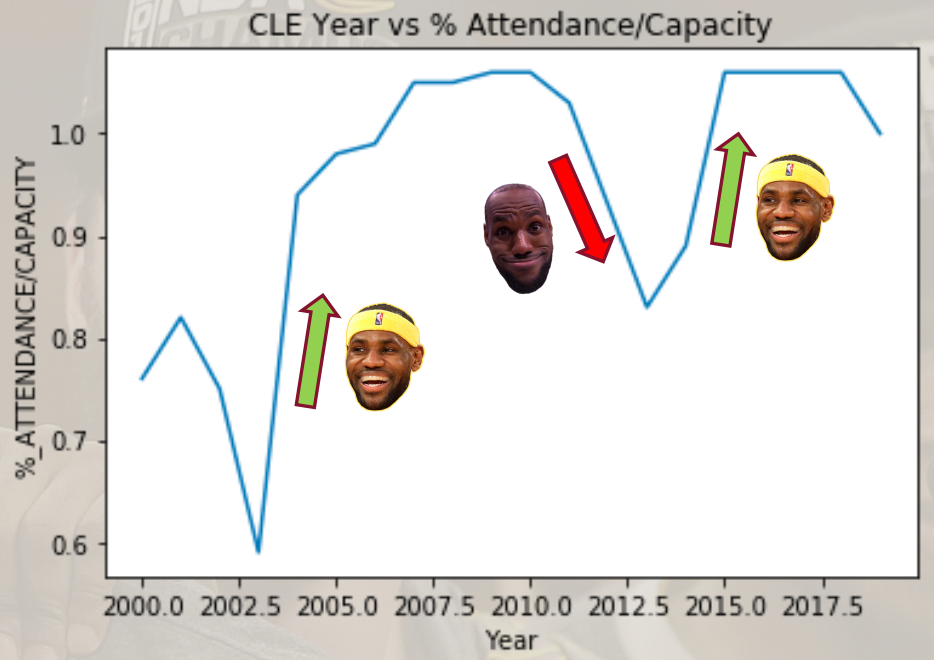
*Excluded 2012 season, as it was a shorted season due to a lockout.

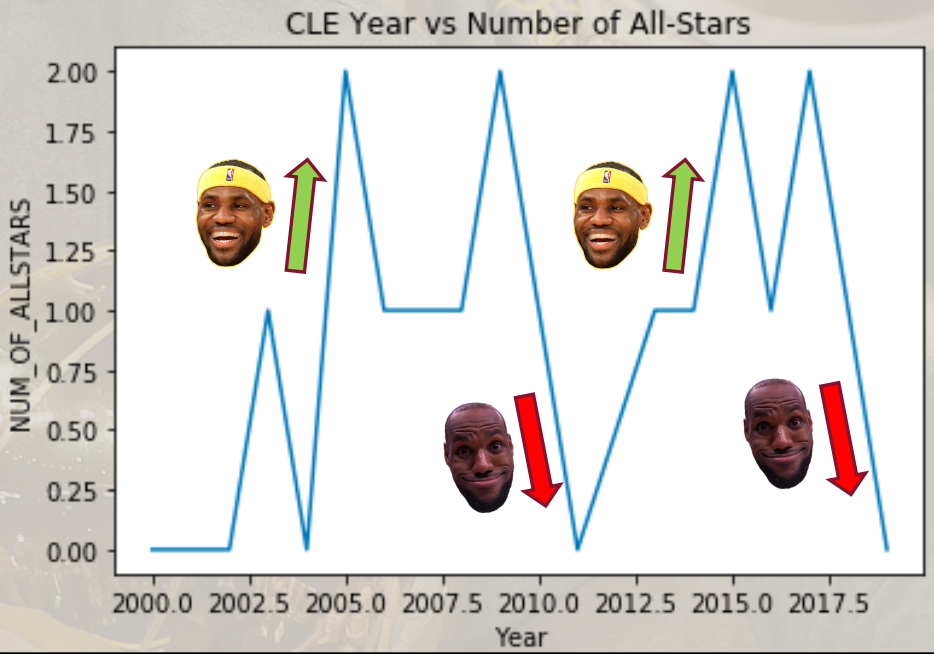
EDA


5
4
3
2
1
0
0.4
0.6
0.8
1.0
1.2
%_ATTENDANCE/CAPACITY


1.1
1.0
0.9
0.8
0.7
0.6
%_ATTENDANCE/CAPACITY
10
20
30
40
50
60
WINS

FEATURE ENGINEERING





FEATURE ENGINEERING II

Arena/Team Name Changes:

- NJN FROM 2010-2012, BRK 2013-2019
- CHH 2000-2002, NOH (2003-2005, 2008-2013), NOK 2006-2007, NOP 2014-2019
- CHA FROM 2004-2014, CHO 2015-2019
- VAN FROM 2000-2001, MEM 2001-2020
- SEA FROM 2000-2008, OKC 2008-2019

Constructed a Feature to take into consideration of Arena Changes.

MODEL RESULTS

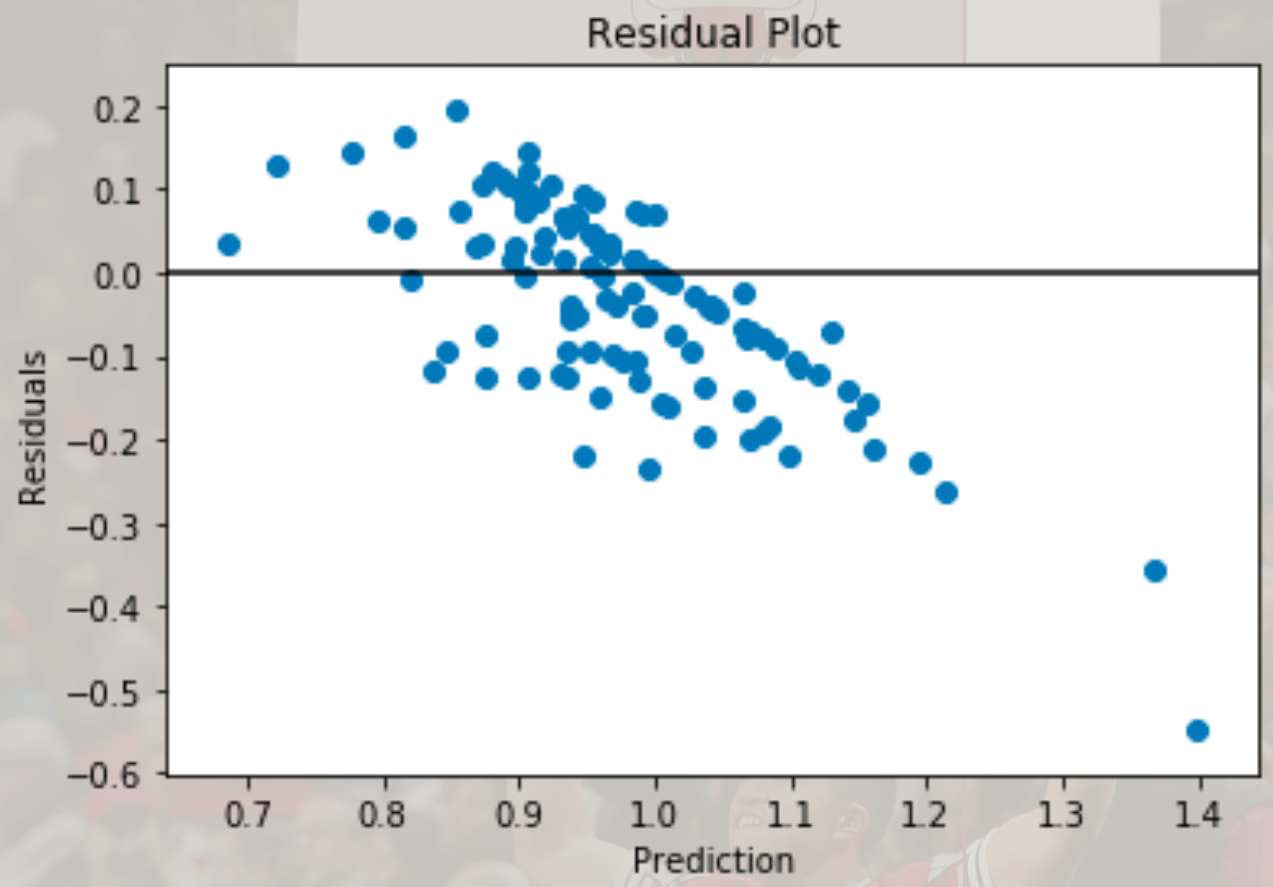
Model	R-Squared (Training)	R-Squared (Validation)	R-Squared (Testing)
Only Features	0.34	.11	-0.33
Features + Polynomials	.70	.23	-1.18
Features + Polynomials w/ LassoCV	.61	.43	-.06

UNDERSTANDING RESULTS

Features	Coefficients
Arena Change	HUGE NUM.
Everything Else	~0

Mean Absolute Error	0.09

RESIDUALS VS PREDICTION



FUTURE PLANS?

THE BIG TAKEAWAY:

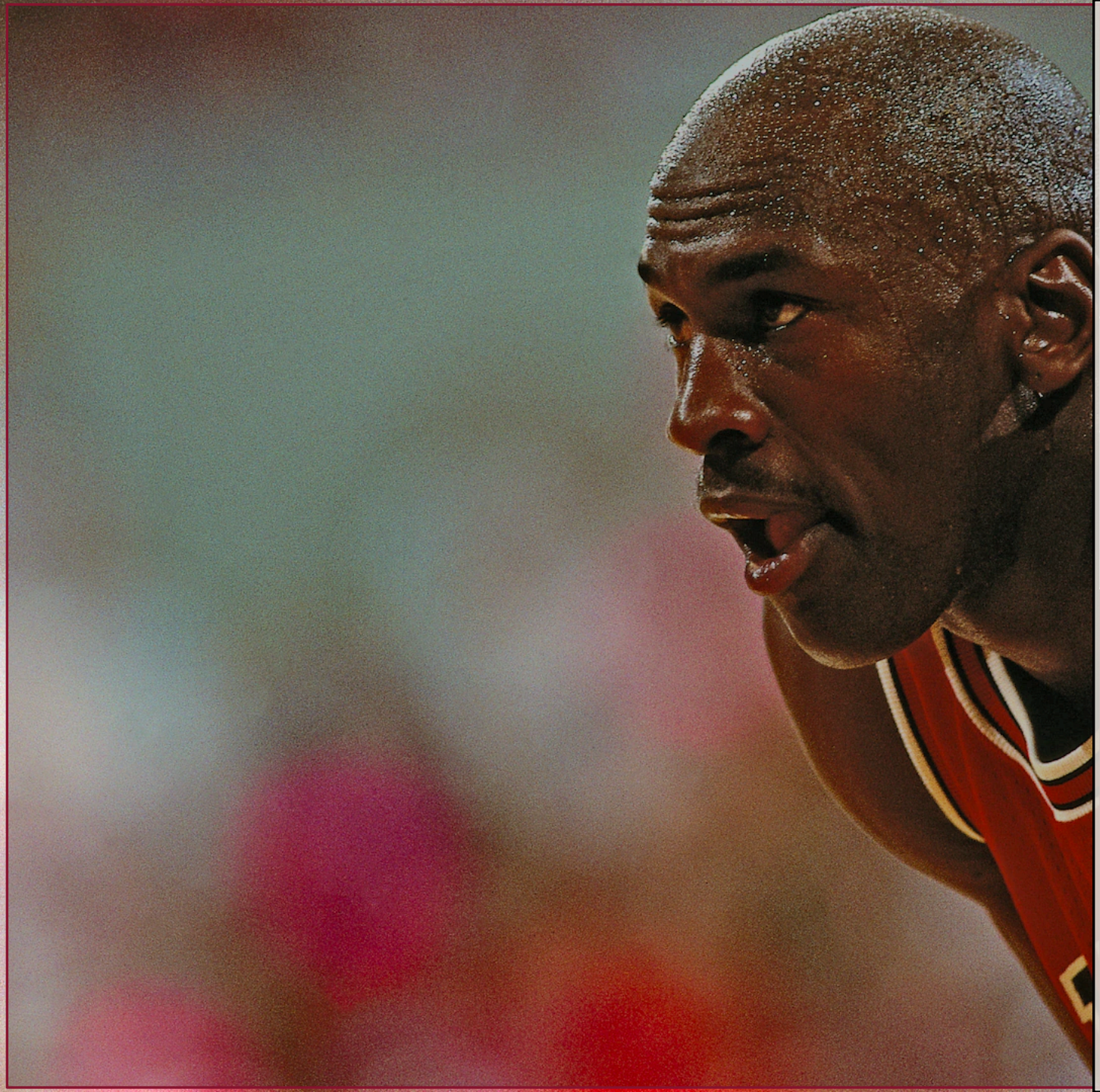
ATTENDANCE IS MORE THAN JUST THE GAME STATS.

- Ticket Price Averages
- Social Media Popularity
- Time Series Analysis
- Population, Income, Demographics, etc.
- Superstar Legacies

QUESTIONS?

SEPT 2020: UPDATED MODEL

- Added in U.S. Cities Populations as a feature from 2000-2019 using sources:
  - https://www.macrotrends.net/cities
  - https://www.biggestuscities.com/
  - https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html
- Added in Superstar Legacy as a feature. Superstar legacy is defined as:
  - Top PER (Player Efficiency Rating) of all time in the NBA (https://www.basketball-reference.com/leaders/per_career.html)
  - Played between 1995 and 2019. 1995 was selected as the cutoff time for superstar lingering effects.
  - Players were known as the 'face of the franchise' – as determined by domain knowledge.
- Large model improvement – means social factors and past superstar legacies outside of game stats play a LARGE role in determining attendance for NBA games.



UPDATE 2: FINAL MODEL RESULTS

Model	R-Squared (Training)	R-Squared (Validation)	R-Squared (Testing)
Less Features	.52	.33	.15
Less Features + Polynomials	.60	60	.52
Less Features + Polynomials w/ LassoCV	.59	.53	-.10

- When using only the wins, capacity, number of all stars and superstar legacy features, we improved our R-Squared (testing) to around 52%.
- This reaffirms my conclusion: game attendance is more than just the immediate in-game, advanced stats. It's rooted in social factors such as population combined with superstar dynasties that carried over to fans.

QUESTIONS??