Who is the Next Lebron James?

Predicting NBA Player Impact with Advanced NBA Statistics

Mateo Martinez

Data Science Capstone | November 2020

Springboard

Introduction

There has been a rise in the last ten years in interest in fantasy sports and online sports betting.

Often casual sports fans and professional analysts implement beliefs such as the Hot Hand fallacy to inform future decisions.

In 2011 the NBA began collecting Advanced Analytics, tracking hundreds of new individual player metrics using advanced camera technologies.

Goal

To apply machine learning algorithms to this newly collected data to determine if new insights can be found and better predict NBA player performance.

Business Application/ Stakeholders

This project could benefit NBA talent scouts, managers, and coaching staffs looking to forecast player performance to inform resigning current players, acquiring new players, or making trade decisions.

Additionally, individuals invested in fantasy sports or online sports betting could benefit as well by being able to find value players to add to their rosters.

Data

- The data scraped from www.nba.com/stats/ from using Selenium and BeautifulSoup.

	PLAYER	TEAM	AGE	GP	W	L	MIN	OFFRTG	DEFRTG	NETRTG	AST%	AST/TO	AST RATIO	OREB%	DREB%	REB%	TO RATIO	EFG%	TS%	USG%	PACE	PIE
4	Giannis Antetokounmpo	MIL	25	63	51	12	30.4	112.8	97.4	15.4	32.8	1.54	17.0	6.8	30.7	19.6	11.0	58.9	61.3	36.3	107.47	23.9
5	LeBron James	LAL	35	67	50	17	34.6	112.1	103.6	8.5	47.7	2.62	28.5	2.8	19.1	11.0	10.9	55.0	57.7	30.8	101.50	19.8
6	Joel Embiid	PHI	26	51	32	19	29.5	107.2	102.0	5.2	17.8	0.96	11.7	9.1	27.9	18.7	12.2	51.2	59.0	31.5	102.07	19.4
7	Luka Doncic	DAL	21	61	36	25	33.6	116.7	111.4	5.3	45.4	2.07	23.6	3.6	22.4	13.2	11.4	53.1	58.5	35.5	101.24	19.4
9	Kawhi Leonard	LAC	29	57	41	16	32.4	116.9	104.7	12.2	25.5	1.88	16.2	2.8	17.5	10.3	8.6	52.4	58.9	32.7	100.54	19.1

- **Explanatory Variables**: 118 unique player statistics from the 2017-2018 season.

- **Target Variable**: the Player Impact Estimate (PIE) in the 2018-2019 season, which is a calculation of what percent of game events (points, rebounds, etc.) did that player contribute to the team. This is an effective metric at comparing player value across all player positions.

Explanatory Features (full names and formulas in glossary):

Descriptive Stats: Team, College, Country

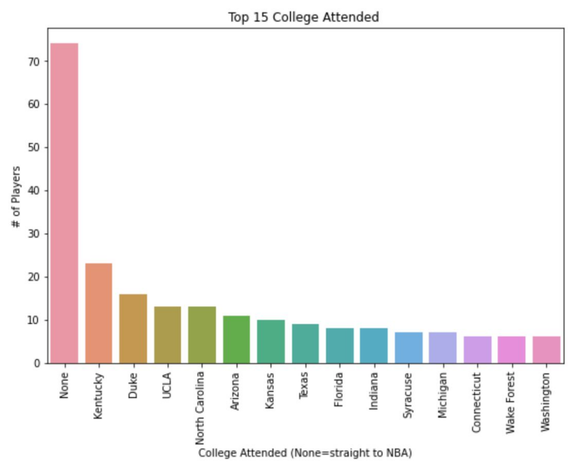
Player Attributes: Age, Height, Weight, Draft Number

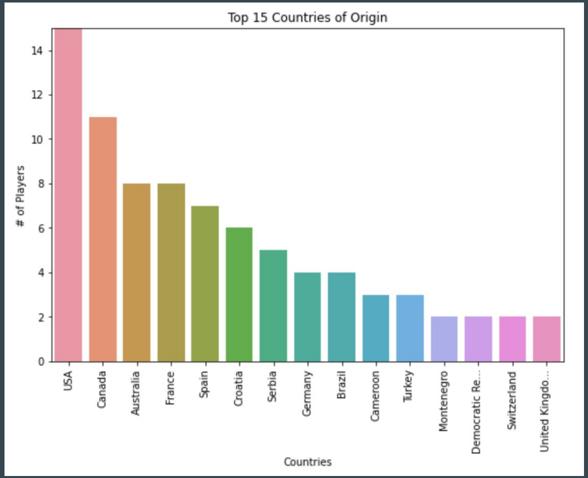
Advanced Stats: PassesMade, AST_PTS_Created, AdjustedREB_Chance%, AVG_REBDistance, %_Loose_BallsRecovered_OFF, Contested3PT_Shots,FG%_20_24ft, OPP_FG%_15_19ft

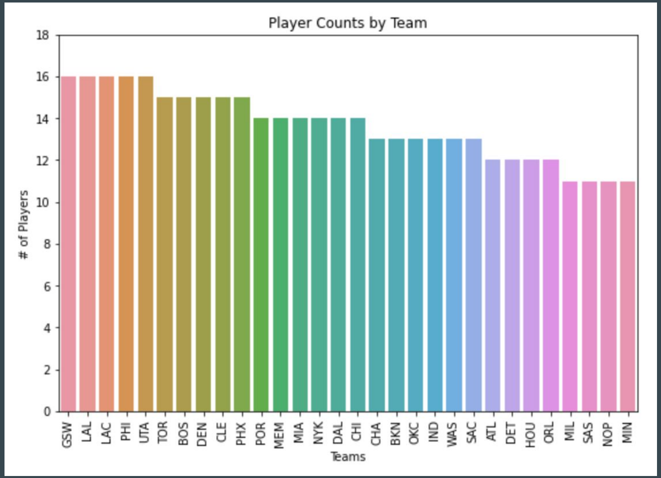
```
['Player', 'TEAM', 'College', 'Country', 'MIN_2017', 'MIN_2018', 'AGE', 'Height', 'Weight', 'Draft_Number', 'GP',
'W', 'L', 'PTS', 'FGM', 'FGA', 'FG%', '3PM', '3PA', '3P%', 'FTM', 'FTA', 'FT%', 'OREB', 'DREB', 'REB', 'AST', 'TOV',
'STL', 'BLK', 'PF', 'FP', 'DD2', 'TD3', '+/-', 'Box_Outs', 'OFF_Box_Outs', 'DEF_Box_Outs', 'Team_RebOn_Box_Outs', 'Pl
ayer_RebOn_Box_Outs', '%_Box_Outs_Off', '%_Box_Outs_Def', '%_Team_RebWhen_Box_Out', '%_Player_RebWhen_Box_Out', 'Cont
estedREB', 'ContestedREB%', 'REBChances', 'REBChance%', 'DeferredREB_Chances', 'AdjustedREB_Chance%', 'AVG_REBDistanc
e', 'PassesMade', 'PassesReceived', 'SecondaryAST', 'PotentialAST', 'AST_PTS_Created', 'ASTAdj', 'AST_ToPass%', 'AST_
ToPass%_Adj', 'ScreenAssists', 'ScreenAssists_PTS', 'Deflections', 'OFF_Loose_BallsRecovered', 'DEF_Loose_BallsRecove
red', 'Loose_BallsRecovered', '%_Loose_BallsRecovered_OFF', '%_Loose_BallsRecovered_DEF', 'ChargesDrawn', 'Contested2
PT_Shots', 'Contested3PT_Shots', 'ContestedShots', 'FGM_und_5ft', 'FGA_und_5ft', 'FG%_und_5ft', 'FGM_5_9ft', 'FGA_5_9
ft', 'FG%_5_9ft', 'FGM_10_14ft', 'FGA_10_14ft', 'FG%_10_14ft', 'FGM_15_19ft', 'FGA_15_19ft', 'FG%_15_19ft', 'FGM_20_2
4ft', 'FGA_20_24ft', 'FG%_20_24ft', 'FGM_25_29ft', 'FGA_25_29ft', 'FG%_25_29ft', 'OPP_FGM_und_5ft', 'OPP_FGA_und_5f
t', 'OPP_FG%_und_5ft', 'OPP_FGM_5_9ft', 'OPP_FGA_5_9ft', 'OPP_FG%_5_9ft', 'OPP_FGM_10_14ft', 'OPP_FGA_10_14ft', 'OPP_
FG%_10_14ft', 'OPP_FGM_15_19ft', 'OPP_FGA_15_19ft', 'OPP_FG%_15_19ft', 'OPP_FGM_20_24ft', 'OPP_FGA_20_24ft', 'OPP_FG%
_20_24ft', 'OPP_FGM_25_29ft', 'OPP_FGA_25_29ft', 'OPP_FG%_25_29ft', 'DEFRTG', 'NETRTG', 'AST%', 'OREB%', 'DREB%', 'RE
B%', 'eFG%', 'TS%', 'USG%', 'PACE', 'PIE']
```

Exploratory Data Analysis

Trends with Categorical Variables

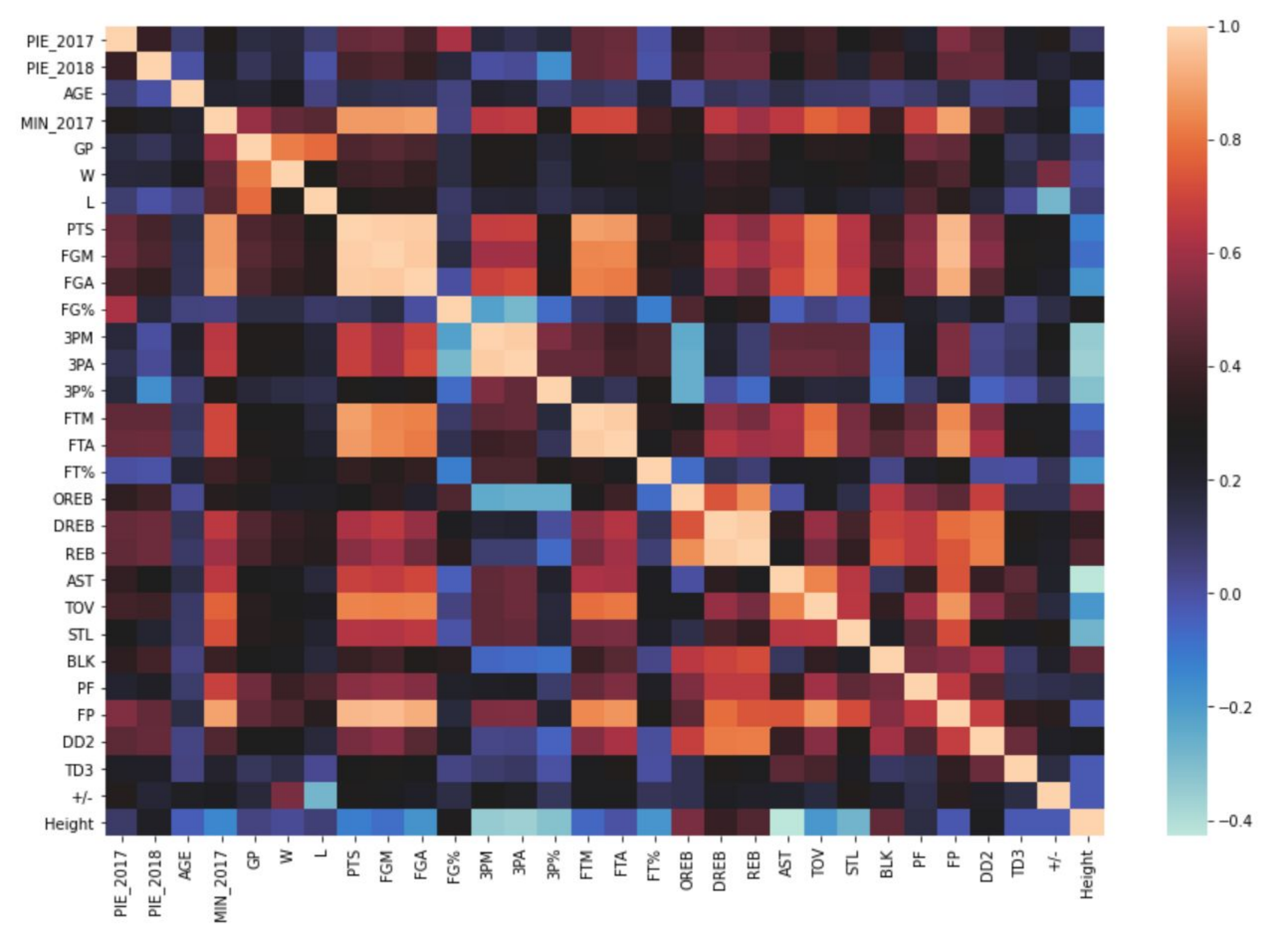






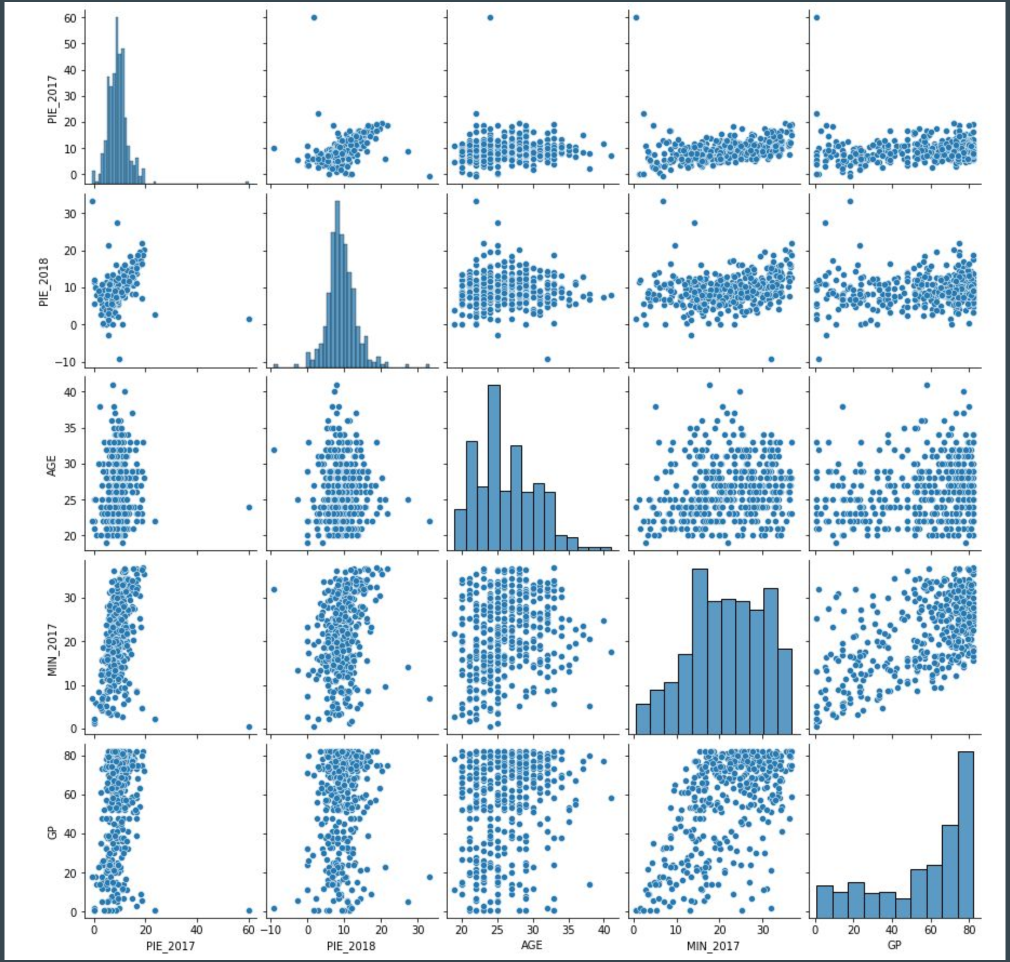
EDA

Correlation Heat Map of first 30 numeric features:



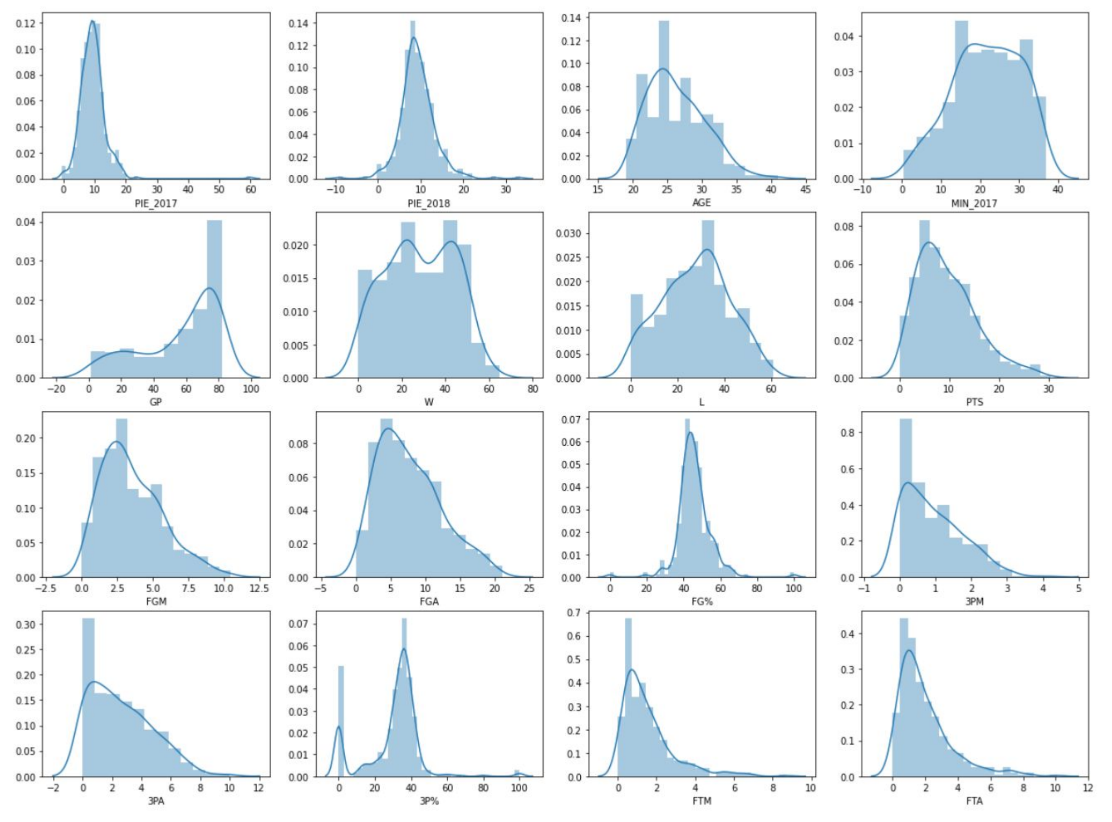
EDA

Scatter Plot Matrix of First 5 Numeric Features:



EDA

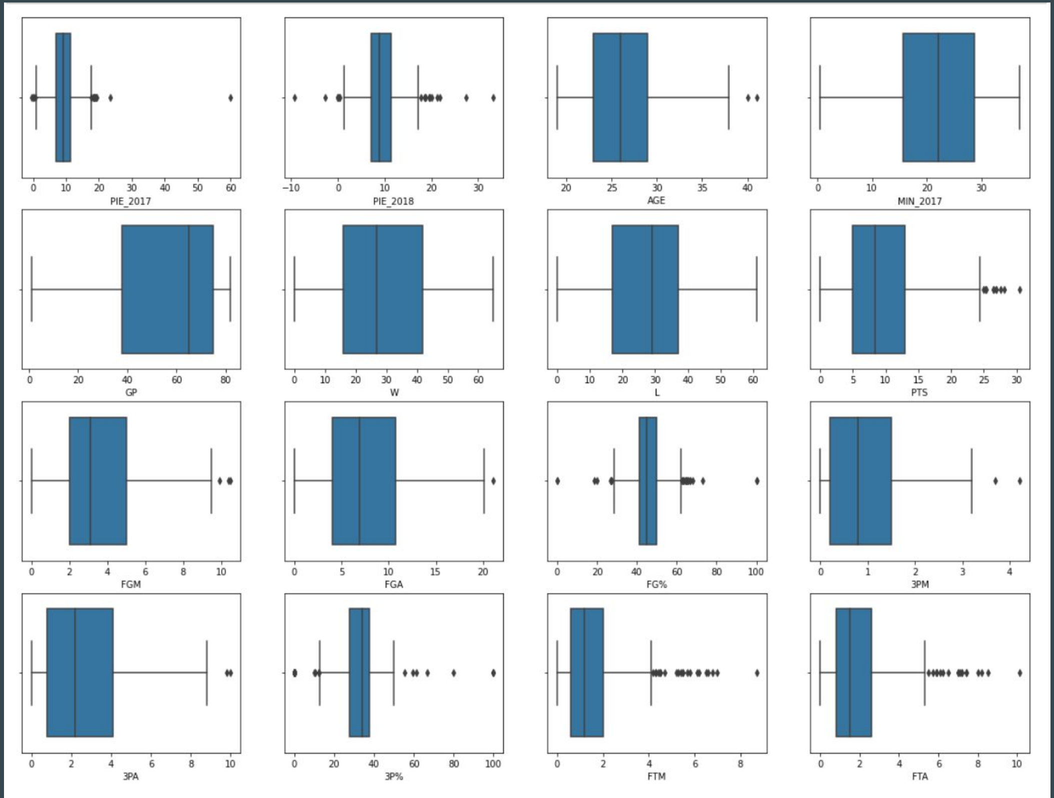
Examining distributions of first 16 numeric features.



Note: many are not normally distributed. To optimize model performance, we will use the Yeo-Johnson transformation to normalize the distributions before standardizing them.

EDA

Examining boxplots of first 16 numeric features, with whiskers = 1.5*IQR.



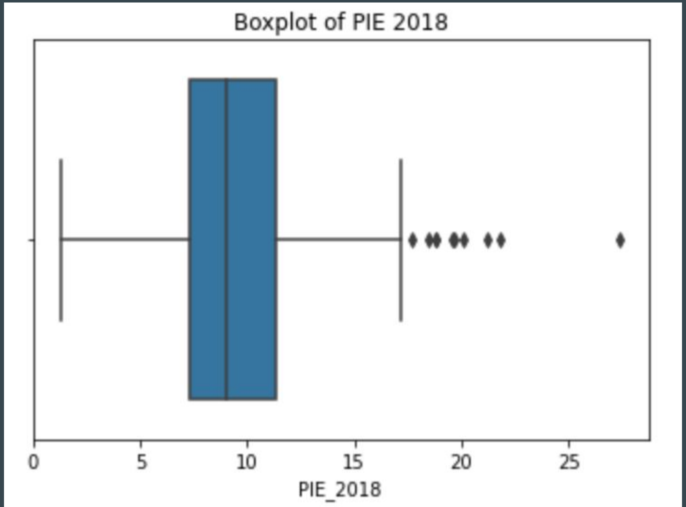
Note: there are a minimal number of outliers. Some of these outliers with be filtered out by excluding players with essentially zero playing time in the 2018-2019 season.

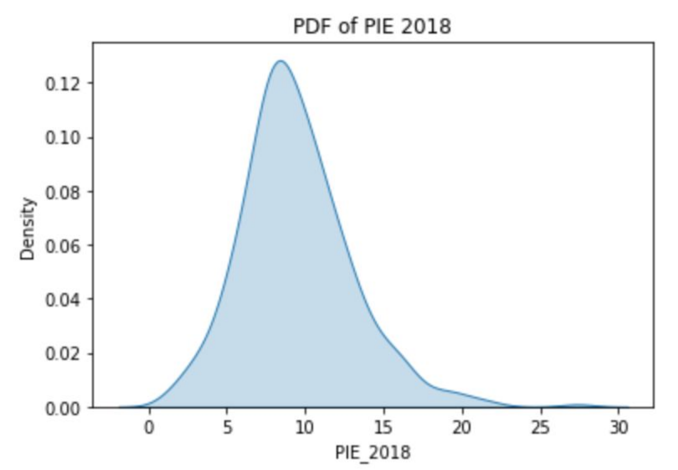
EDA: Closer Look at Target Variable

Looking at the summary statistics for PIE_2018, there is a relatively small interquartile range around the mean PIE value of 9.44. The boxplot also reflects this displaying a small but notable number of outliers beyond the plot's whiskers. The PDF show that by visual analysis the PIE scores are normally distributed.

```
df.PIE_2018.describe()

count    399.000000
mean       9.444862
std        3.549127
min        1.300000
25%        7.300000
50%        9.000000
75%       11.300000
max       27.400000
Name: PIE_2018, dtype: float64
```





Pre-Processing

- Normalized and standardized the data.

- created an additional feature with K-Nearest Neighbors to create 5 players clusters.

- Converted categorical features into dummy variables.

- Split training and testing data (60/40)

- Used a 5 fold cross validation when tuning hyper-parameters

Modeling Overview

Type: Multivariate Linear Regression

Scoring Metric: RMSE

Algorithms: OLS, Lasso, Ridge, Elastic Net, Random Forest, XGBoost

Hyper-parameter Tuning: Grid Search, Random Grid Search, Bayesian Optimization

Tools/Libraries: Statsmodels, Scikit-Learn, xgBoost, BayesOpt

Model Comparison

- XGBoost and Random Forest had the lowest RMSE, but their results are less interpretable

- The most accurate model was Elastic Net and with additional hyperparameter tuning it had a comparable RMSE with Random Forest.

	RMSE	R-Squared	Hyperparameters
XGBoost	2.20	0.53	Default
Random_Forest	2.26	0.50	Default
Elastic_Net	2.28	0.49	Bayesian Optimization
Elastic_Net	2.36	0.45	Grid Search
Elastic_Net	2.36	0.45	Grid Search
Lasso	2.36	0.45	Grid Search
Ridge	2.36	0.45	Grid Search
Elastic_Net	2.38	0.44	Random Grid Search
SmOLS	3.02	0.10	Default

Model Predictions: What Went Wrong?

<u>10 LEAST Accurate Predictions</u>

Themes between these players:

- Changed teams

- No regression towards mean

- Large change in PIE score

Player	PIE_2017	PIE_2018	predictions	pred_error	AGE	MIN_2017	true_change_in_PIE	diff_team
Wade Baldwin IV	9.4	2.1	9.142897	7.042897	22	11.5	-7.3	False
James Harden	19.4	20.1	13.754321	6.345679	28	35.4	0.7	False
Andrew Harrison	8.3	2.5	8.453300	5.953300	23	23.7	-5.8	True
Nikola Vucevic	13.9	18.5	13.312251	5.187749	27	29.5	4.6	False
Anthony Davis	18.8	19.7	14.630221	5.069779	25	36.4	0.9	False
Paul George	12.0	16.1	11.068824	5.031176	28	36.6	4.1	False
Mike Conley	10.9	15.1	10.366519	4.733481	30	31.1	4.2	False
Tyrone Wallace	6.0	4.3	8.926866	4.626866	24	28.4	-1.7	False
Jeremy Lin	7.1	10.0	5.511644	4.488356	29	25.2	2.9	True
Walt Lemon Jr.	3.0	10.9	6.474474	4.425526	25	7.0	7.9	True

Model Predictions: What Went Right?

10 MOST Accurate Predictions

Themes between these players:

- Did not changed teams

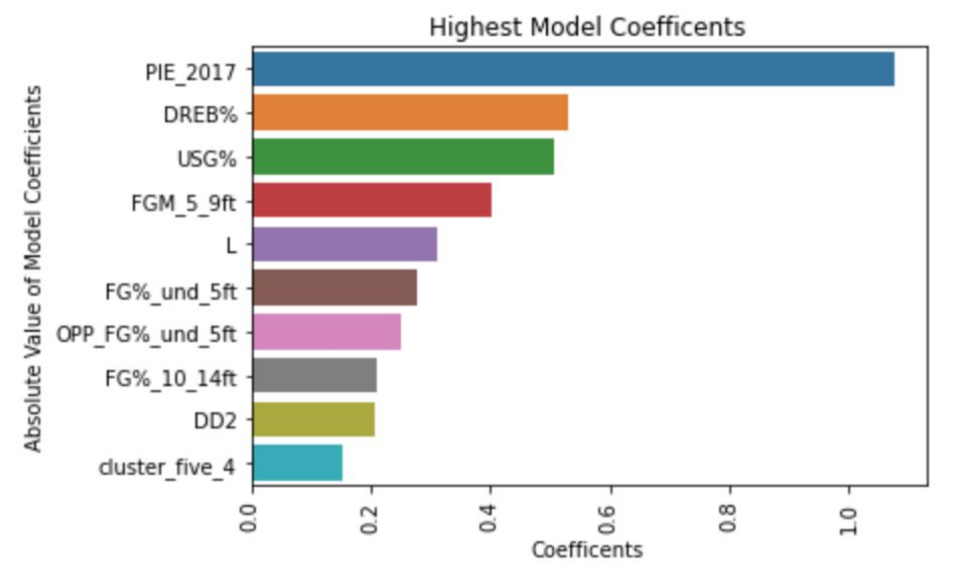
- Small change in PIE

- General regression towards the mean.

Player	PIE_2017	PIE_2018	predictions	pred_error	AGE	MIN_2017	true_change_in_PIE	diff_team
Davis Bertans	9.2	8.2	8.193908	0.006092	25	14.1	-1.0	False
Kyle Lowry	13.6	11.5	11.488014	0.011986	32	32.2	-2.1	False
JaMychal Green	10.1	10.3	10.324258	0.024258	28	28.0	0.2	True
Jon Leuer	7.0	9.1	9.052195	0.047805	29	17.1	2.1	False
Ante Zizic	13.2	10.3	10.354570	0.054570	21	6.7	-2.9	False
Doug McDermott	6.9	7.5	7.415245	0.084755	26	21.8	0.6	True
Terry Rozier	11.8	10.3	10.432317	0.132317	24	25.9	-1.5	False
Tyler Dorsey	7.0	7.2	7.061712	0.138288	22	17.4	0.2	True
Udonis Haslem	2.1	6.7	6.849273	0.149273	38	5.1	4.6	False
Fred VanVleet	11.1	9.5	9.292208	0.207792	24	20.0	-1.6	False

Analysis: Most Impactful Coefficients

	Absolute Value of Model Coefficients	+/- Corr.
PIE_2017	1.074437	+
DREB%	0.531392	+
USG%	0.507683	+
FGM_5_9ft	0.402081	+
L	0.312532	-
FG%_und_5ft	0.278564	+
OPP_FG%_und_5ft	0.249301	-
FG%_10_14ft	0.210631	-
DD2	0.207227	+
cluster_five_4	0.153731	+



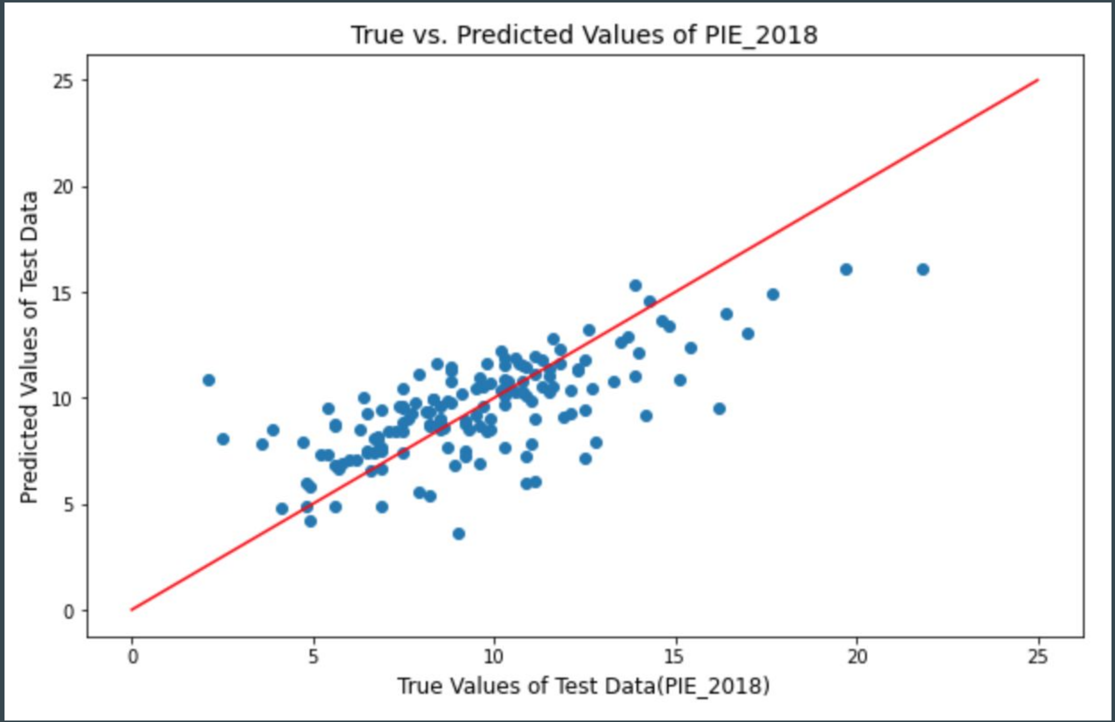
- DREB%, USG%, FGM_5_ft are surprisingly most impactful positive coefficients
- FG%_10_14ft is surprisingly negatively correlated with Player Impact

Analysis: Residuals

- The model underestimated higher true values and overestimated lower true values.

- The model was more accurate in predicting the general trend of a Regression Towards the mean.

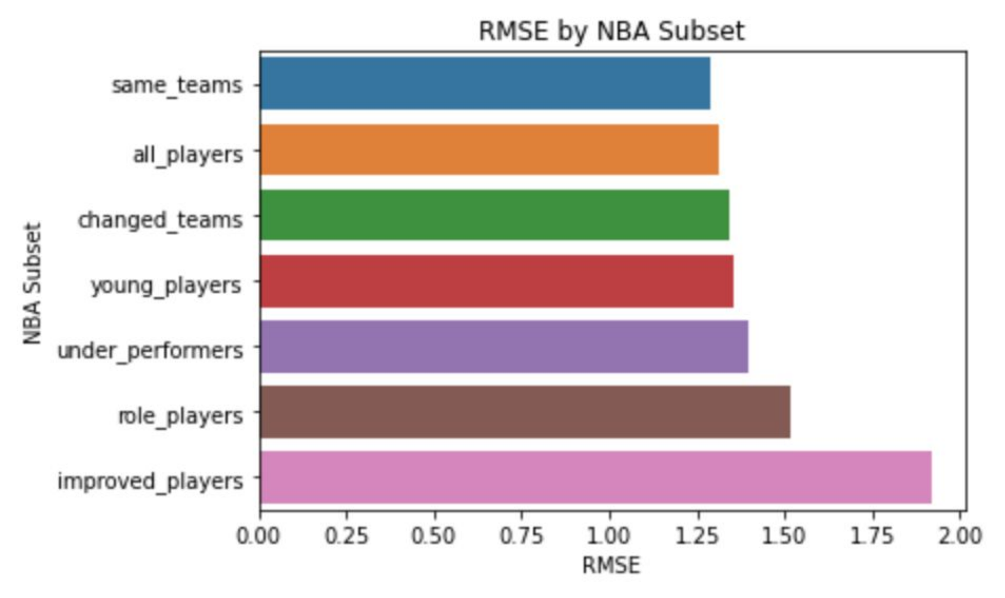
- The model was less accurate at capturing outlier player progress between seasons.



Analysis: On Which Subsets was More/Less Accurate?

-Most accurate predictions with players who did not change teams.

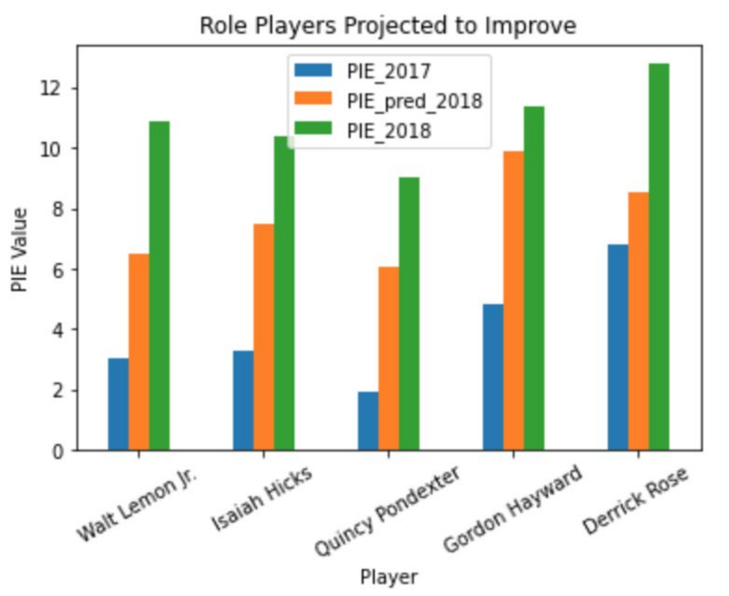
-Least accurate predictions on most improved players, less able to identify outliers.



Business Implications/Recommendations

1. Identify undervalued role/bench players projected to show moderate improvement (10-20%) the following season.

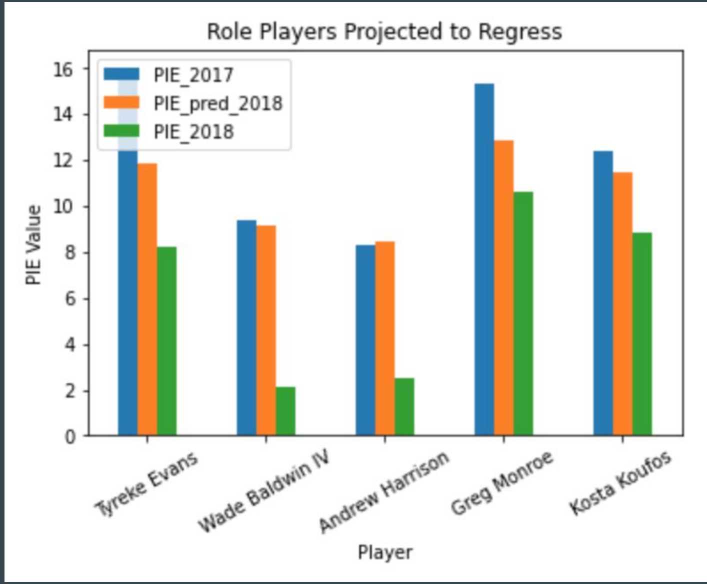
Player	PIE_2017	PIE_pred_2018	PIE_2018
Walt Lemon Jr.	3.0	6.474474	10.9
Isaiah Hicks	3.3	7.461030	10.4
Quincy Pondexter	1.9	6.034350	9.0
Gordon Hayward	4.8	9.861101	11.4
Derrick Rose	6.8	8.504182	12.8



Business Implications/Recommendations

2. Informing decision to not resign or offer less money to role/bench players that are projected to regress the following season.

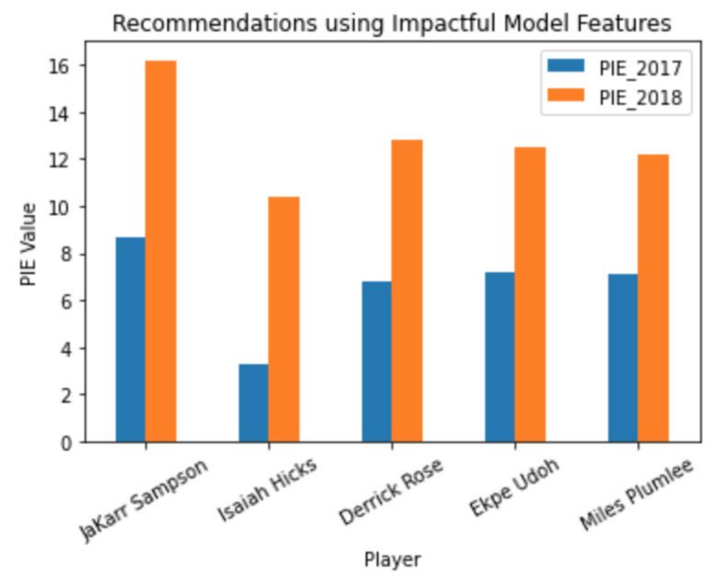
Player	PIE_2017	PIE_pred_2018	PIE_2018
Tyreke Evans	16.0	11.836909	8.2
Wade Baldwin IV	9.4	9.142897	2.1
Andrew Harrison	8.3	8.453300	2.5
Greg Monroe	15.3	12.884193	10.6
Kosta Koufos	12.4	11.480949	8.8



Business Implications/Recommendations

3. Utilize features that were most impactful coefficients of the regression model to identify growth players.

Player	DREB%	USG%	FGM_5_9ft	FG%_und_5ft	PIE_2017	PIE_2018
JaKarr Sampson	16.1	12.4	0.3	61.7	8.7	16.2
Isaiah Hicks	11.8	17.5	0.1	56.8	3.3	10.4
Derrick Rose	5.3	24.5	0.2	62.5	6.8	12.8
Ekpe Udoh	10.5	8.9	0.2	65.3	7.2	12.5
Miles Plumlee	16.0	12.7	0.2	64.7	7.1	12.2



Next Steps

- Add additional seasons into the training set to improve model performance and hopefully the additional data would better capture the unique progression and growth of superstar players.

- Build this regression model into a Flask app where a user could select a target feature and a player of their choice and see a prediction for that feature in the upcoming season.

Mateo Martinez
mateomartinez510@gmail.com
linkedin.com/in/mateomartinez510
github.com/mateomartinez510