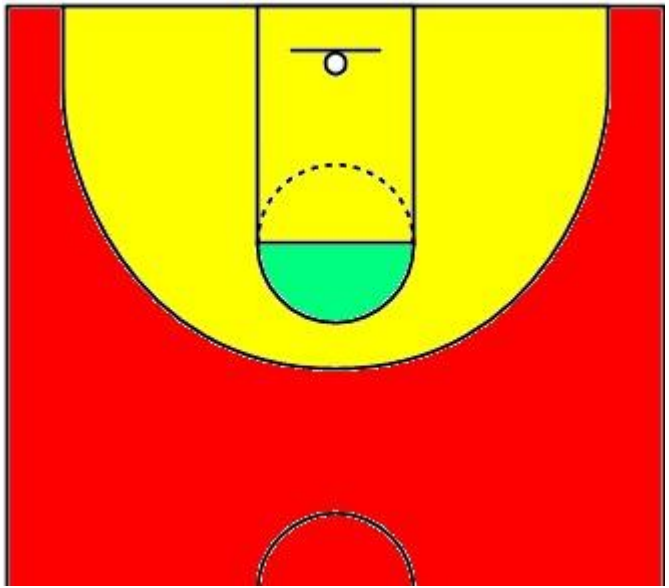
NBA player statline predictions

NCAA data - https://www.sports-reference.com/
NBA data - https://www.basketball-reference.com/
Draft data - https://data.world/rodikurucsmvp/nba-draft-ageperformance-relationship-data

Erwin Hessing
20 May 2021

Basketball

G -- Games
MP -- Minutes Played Per Game
2P -- 2-Point Field Goals Per Game
2PA -- 2-Point Field Goal Attempts Per Game
2PP -- 2-Point Field Goal Percentage
3P -- 3-Point Field Goals Per Game
3PA -- 3-Point Field Goal Attempts Per Game
3PP -- 3-Point Field Goal Percentage
FT -- Free Throws Per Game
FTA -- Free Throw Attempts Per Game
FTP -- Free Throw Percentage
ORB -- Offensive Rebounds Per Game
DRB -- Defensive Rebounds Per Game
TRB -- Total Rebounds Per Game
AST -- Assists Per Game
STL -- Steals Per Game
BLK -- Blocks Per Game
TOV -- Turnovers Per Game
PTS -- Points Per Game



The problem

Teams need to decide what players to pick with their draft pick, their information is limited and draft picks are very valuable, especially lottery draft picks.

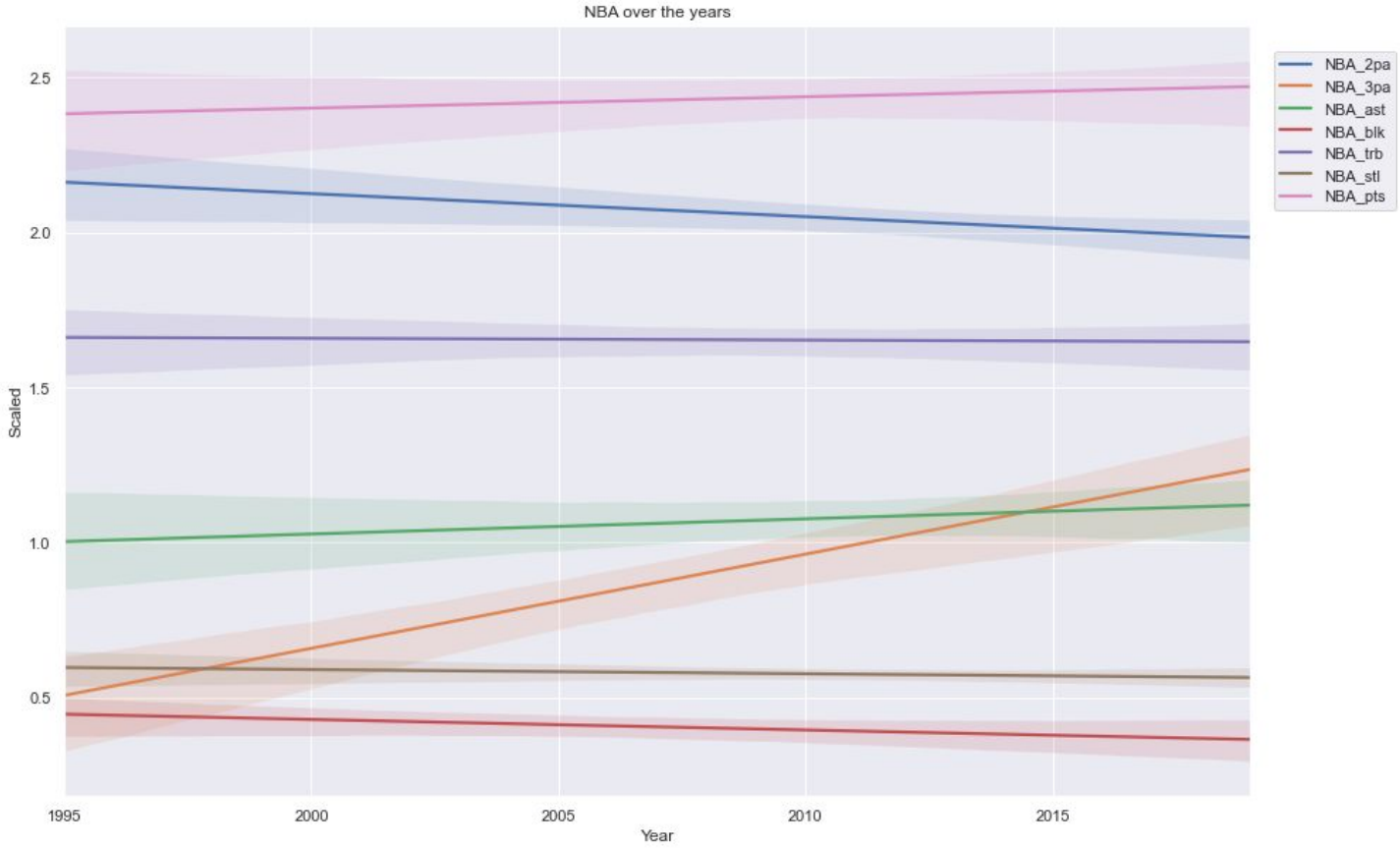
Predict the scoreline

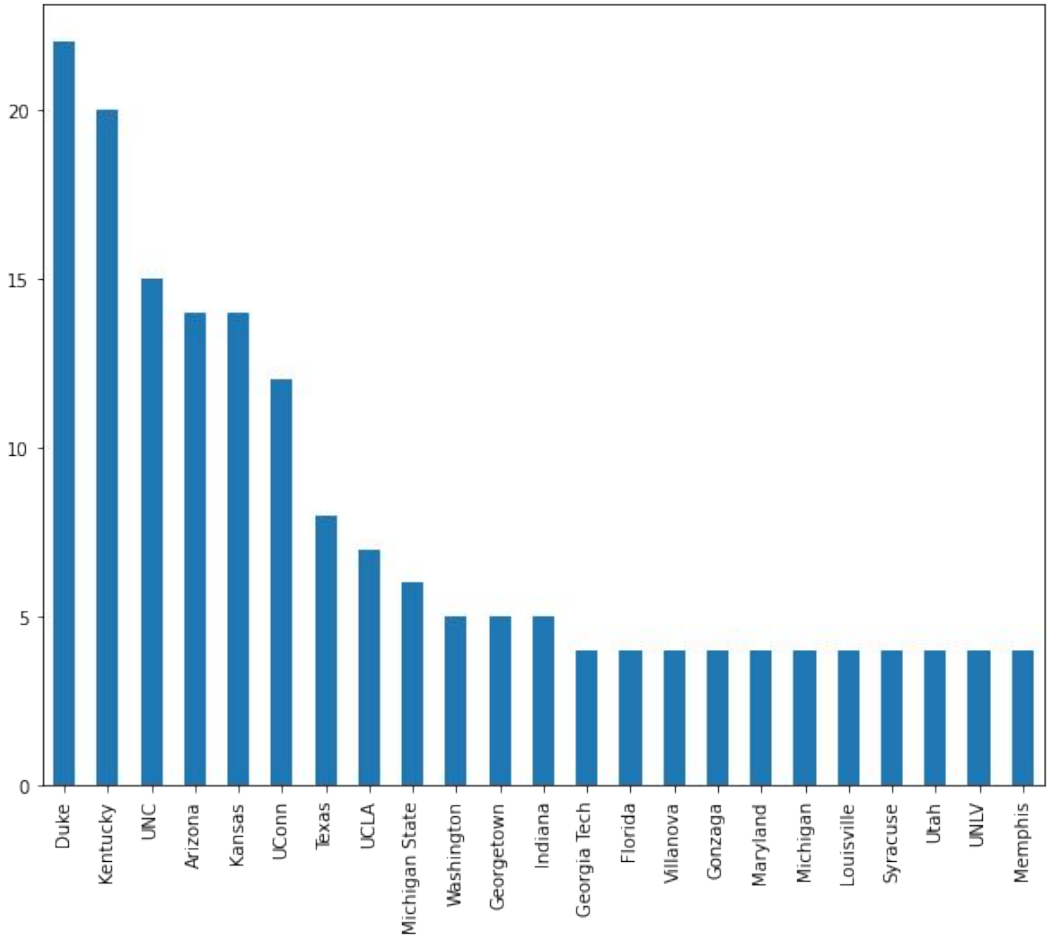
- Points
- Assists
- Rebounds
- Blocks
- Steals

I’ve made some assumptions

- Predicting 2nd year rather than rookie season
- Data from 1995 to 2019
- Lottery Draft picks (top 14)

The game has changed in the last 25 years





Interesting info:

The number of lottery draft picks by college in the last ~25 years

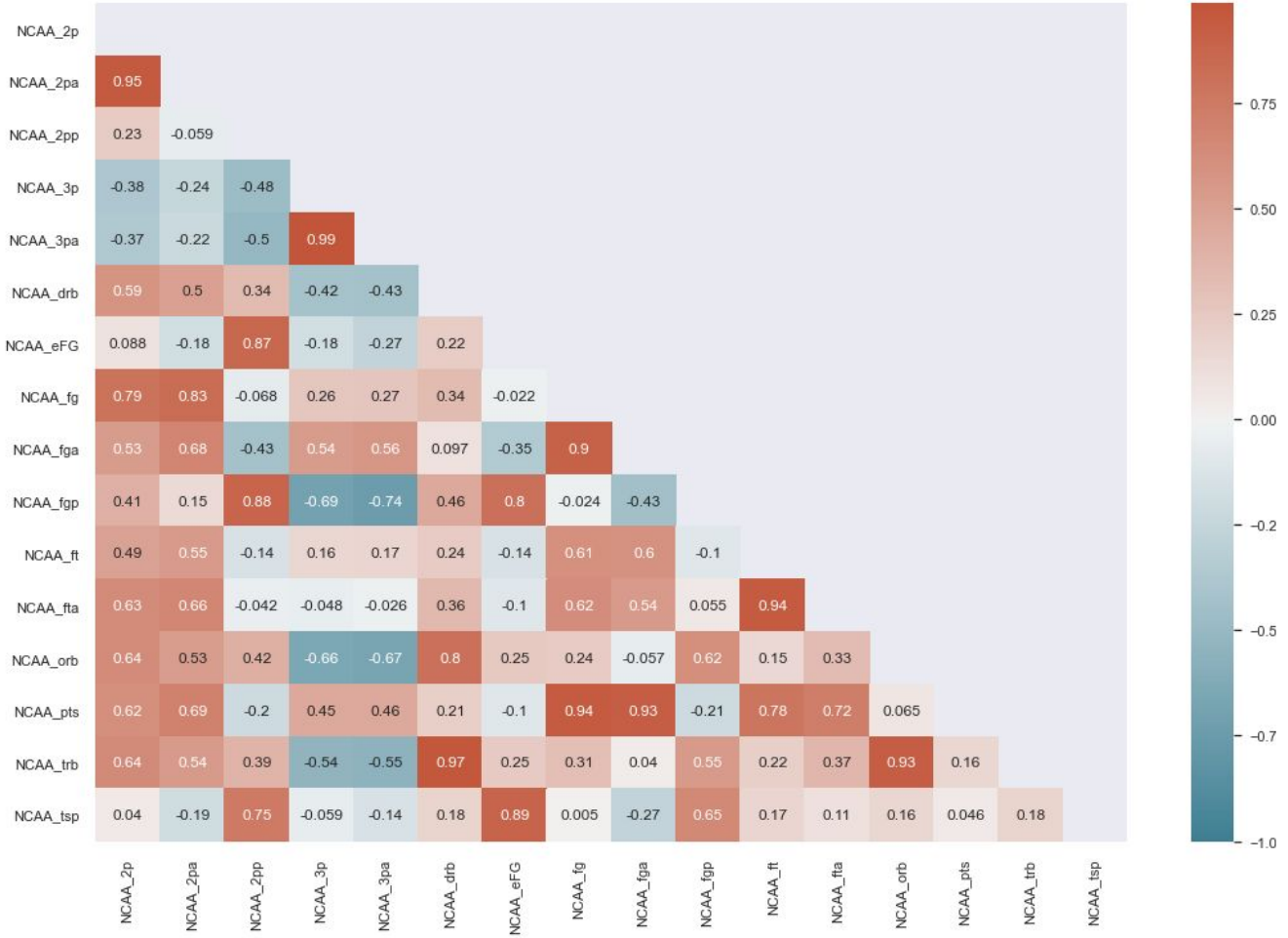
Data

- Removed any player that did not play in NCAA or NBA in their second year
- Removed any player that did not play at least 15 games in either NCAA/NBA
- Rebounds were not split between offensive and defensive prior to 2000

Train / Validate:
Approximately 265 players data was used

Test:
Data not seen by the model - 2019 Draft Class

Correlation

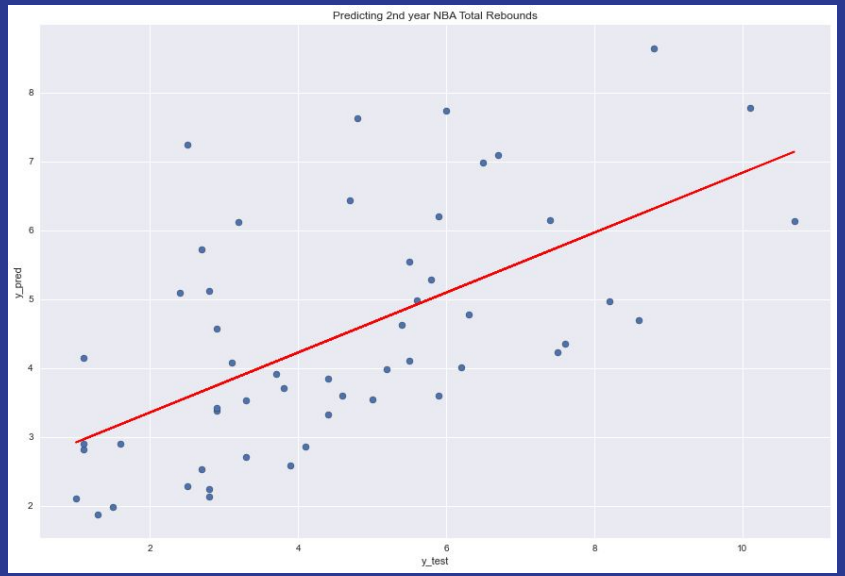


Modelling

Feature Importance
- Correlation
- Kbest
- Info Gain

Models
- Linear Regression
- Random Forest
- K-Nearest Neighbours (KNN)
- XGBoost (Did not end up using)

Scoring Models (Cross Validation)
- R-Squared (R2)
- Root Mean Square Error (RMSE)



Points

	corr	KBest Score	info_gain
NCAA_pts	0.233244	15.131141	0.064270
NCAA_stl	0.233162	15.119864	0.069452
NCAA_fga	0.225860	14.137557	0.105027
position_C	0.210654	12.212541	0.011295
NCAA_fg	0.197890	10.718944	0.005900
NCAA_3pa	0.191257	9.985566	0.048523
NCAA_3p	0.181897	8.999506	0.009371
NCAA_ft	0.178814	8.687071	0.000000
NCAA_ast	0.167351	7.577932	0.020296
NCAA_g	0.164848	7.346613	0.080983
position_G-F	0.164689	7.332032	0.000000
NCAA_fta	0.142392	5.442821	0.099895
NCAA_fgp	0.139217	5.198081	0.000000
NCAA_ftp	0.137521	5.069745	0.000000
NCAA_mp	0.135064	4.886871	0.044861
position_G	0.130305	4.542702	0.000000
NCAA_3pp	0.121802	3.960562	0.000000
NCAA_blk	0.120884	3.900185	0.086723
position_F	0.107179	3.056280	0.000000

Model - Points	How did it do?
Linear Regression [pts, stl, ft, 3p, blocks, positions]	R2 = 0.13 (0.09) RMSE = 5.48
Random Forest	R2 = 0.07 RMSE = 5.26
KNN - 45	R2 = 0.08 RMSE = 5.49
Linear Regression - Scaled [non correlated features]	R2 = 0.22 (0.15) RMSE = 5.41
Linear Regression [non correlated features]	R2 = 0.22 (0.14) RMSE = 5.42
Linear Regression [All features]	R2 = 0.26 (0.14) RMSE = 5.47

Assists

	corr	KBest Score	info_gain
NCAA_ast	0.748698	335.472735	0.420863
position_G	0.594137	143.490551	0.194268
NCAA_stl	0.518438	96.671879	0.155438
NCAA_orb	0.500732	88.009329	0.176389
NCAA_trb	0.431497	60.171087	0.151214
NCAA_blk	0.408500	52.677811	0.113386
NCAA_tov	0.390352	47.278655	0.081941
NCAA_3pa	0.390132	47.215700	0.111442
NCAA_fgp	0.380995	44.659036	0.117616

Model - Assists	How did it do?
Linear Regression [ast, stl, blk, trb, positions]	R2 = 0.58 (0.56) RMSE = 1.34
Random Forest	R2 = 0.39 RMSE = 1.81
KNN - 11	R2 = 0.55 RMSE = 1.44
Linear Regression [non correlated features]	R2 = 0.62 (0.58) RMSE = 1.34
Linear Regression - scaled [non correlated features]	R2 = 0.59 (0.55) RMSE = 1.38
Linear Regression [All Features]	R2 = 0.63 (0.57) RMSE = 1.37

Steals

	corr	KBest Score	info_gain
NCAA_stl	0.649592	191.993599	0.349713
NCAA_ast	0.442724	64.116236	0.112396
position_G	0.330014	32.143847	0.000000
NCAA_fgp	0.317634	29.511855	0.083450
NCAA_3pa	0.293812	24.848698	0.158349
position_C	0.289169	23.998516	0.060653
NCAA_orb	0.271225	20.883372	0.152353
NCAA_3p	0.267024	20.192115	0.110312
NCAA_blk	0.257110	18.616404	0.093470
NCAA_trb	0.223216	13.791233	0.034398
NCAA_tov	0.222383	13.683121	0.061296

Model - Steals	How did it do?
Linear Regression [stl, ast, fgp, 3pa, positions]	R2 = 0.46 (0.43) RMSE = 0.33
Random Forest - 1000	R2 = 0.29 RMSE = 0.36
KNN - 3	R2 = 0.51 Rmse = 0.41
Linear Regression [non correlated features]	R2 = 0.51 (0.47) RMSE = 0.32
Linear Regression - Scaled [All Features]	R2 = 0.55 (0.48) RMSE = 0.34

Blocks

	corr	KBest Score	info_gain
NCAA_blk	0.745142	328.325851	0.530247
NCAA_trb	0.519064	96.991863	0.215260
NCAA_orb	0.516322	95.598286	0.284865
NCAA_3pa	0.490751	83.433851	0.130090
NCAA_3p	0.482450	79.786084	0.144435
NCAA_drb	0.478440	78.073155	0.198381
position_F-C	0.469512	74.370378	0.025901
NCAA_fgp	0.463206	71.844122	0.173671
position_G	0.452482	67.709507	0.169671
NCAA_ast	0.397479	49.347595	0.151985

Model - Blocks	How did it do?
Linear Regression [blk, trb, 3pa, fga, positions]	R2 = 0.59 (0.58) RMSE = 0.36
Random Forest - 1000	R2 = 0.54 RMSE = 0.45
KNN - 3	R2 = 0.67 RMSE = 0.46
Linear Regression [non correlated features]	R2 = 0.61 (0.57) RMSE = 0.38
Linear Regression - Scaled [All Features]	R2 = 0.65 (0.59) RMSE = 0.38

Rebounds

	corr	KBest Score	info_gain
NCAA_trb	0.627443	170.767831	0.280918
NCAA_orb	0.626120	169.584412	0.287138
NCAA_drb	0.577641	131.698719	0.254527
NCAA_3pa	0.476676	77.329932	0.228914
NCAA_blk	0.468049	73.777815	0.249807
NCAA_3p	0.465725	72.844653	0.138440
position_G	0.453381	68.048497	0.148071
NCAA_fgp	0.451004	67.155029	0.110727
position_F-C	0.406562	52.080427	0.133663
NCAA_2p	0.366748	40.872079	0.000000
NCAA_ftp	0.356796	38.364699	0.039262

Model - Rebounds	How did it do?
Linear Regression [trb, 3pa, blk, positions]	R2 = 0.45 (0.43) RMSE = 1.88
Random Forest - 1000	R2 = 0.44 RMSE= 1.76
KNN - 19	R2 = 0.43 RMSE = 1.98
Linear Regression [Non correlated features]	R2 = 0.52 (0.47) RMSE = 1.96
Linear Regression - Scaled [All Features]	R2 = 0.56 (0.49) RMSE = 1.92

I have some models, how good are they really?

	name	draft_pick	NBA_g	NBA_pts	pred_pts	NBA_ast	pred_ast	NBA_blk	pred_blk	NBA_stl	pred_stl	NBA_trb	pred_trb
0	Zion Williamson	1	61	27.0	23.23	3.7	3.777472	0.6	0.60	0.9	1.55	7.2	8.61
1	Ja Morant	2	58	19.4	21.60	7.3	7.414830	0.2	0.33	0.9	1.55	3.9	8.03
2	RJ Barrett	3	66	17.6	20.16	3.0	3.861372	0.3	0.60	0.7	1.19	5.7	7.21
3	De'Andre Hunter	4	20	16.0	12.45	2.1	1.426095	0.5	1.33	0.9	0.72	5.1	5.47
4	Jarrett Culver	6	34	5.3	15.64	0.7	3.301485	0.3	1.07	0.5	1.29	3.1	5.24
5	Coby White	7	63	14.9	16.93	4.7	4.097885	0.2	0.33	0.5	1.16	4.0	4.46
6	Jaxson Hayes	8	54	6.6	4.81	0.5	-0.296258	0.5	1.03	0.4	0.19	4.0	1.26
7	Rui Hachimura	9	52	13.8	11.52	1.5	1.082253	0.1	1.20	0.8	0.67	5.5	4.58
8	Cam Reddish	10	26	11.2	12.81	1.3	1.631893	0.3	0.33	1.3	0.88	4.0	2.50
9	Cameron Johnson	11	60	9.6	13.04	1.4	1.638386	0.3	0.23	0.6	0.84	3.3	6.68
10	Tyler Herro	13	48	14.9	14.83	3.4	2.667412	0.4	0.43	0.7	1.14	4.9	3.83

What’s next?

- I would look at the field goal percentages
- Would have been keen to see if less but more recent data made a difference
- Seen whether all 1st round draft picks would have made a difference