WHO SCORES?

PREDICTING THE NUMBER OF GOALS SCORED IN A PREMIER LEAGUE SEASON

CLAYTON YOUNG

Why Premier League?

Viewership and influence

- In 2019, viewership rose in 11 per cent to 1.35 billion[1]
- Broadcast in 212 territories to 643 million homes and a potential TV audience of 4.7 billion people[2]
- 10% of the world's population support Manchester United, including 100 million people in China – more than are members of the Communist Party [3]
- Premier League striker, Didier Drogba, ends civil war in Ivory Coast[4]

1. https://www.premierleague.com/news/1280062
2. https://www.thetimes.co.uk/article/history-and-time-are-key-to-power-of-football-says-premier-league-chief-3d3zf5kb35m
3. https://www.britishcouncil.org/research-policy-insight/insight-articles/playing-game-soft-power-sport
4. https://www.bbc.com/sport/football/52072592

Goals?

- Predicting number of non-penalty goals in a season.
  - Assumption: non-penalty goals better measure of player performance.

Data

Rows: 2091
Columns: 81

WIKIPEDIA
The Free Encyclopedia

YOURDICTIONARY

Africa Country Abbreviations

Africa is a diverse area with a rich history. From the Egyptians to the Nenet tribe, you country abbreviations below.

2-Letter	3-Letter	Country Name
DZ	DZA	Algeria
AO	AGO	Angola
BJ	BEN	Benin

FBREF

GDP (US$ million) by country

	Country or territory	Region	IMF[1]		UN[12]		World Bank[13]	
			Estimate	Year	Estimate	Year	Estimate	Year
1	United States *(more)*	Americas	22,675,271	2021	21,433,226	2020	20,936,600	2020
2	China *(more)*	Asia	16,642,318	[n 2]2021	14,342,933	[n 3]2020	14,722,731	2020
3	Japan *(more)*	Asia	5,378,136	2021	5,082,465	2020	5,064,873	2019
4	Germany *(more)*	Europe	4,319,286	2021	3,861,123	2020	3,806,060	2020
5	United Kingdom *(more)*	Europe	3,124,650	2021	2,826,441	2020	2,707,744	2020
6	India *(more)*	Asia	3,049,704	2021	2,891,582	2020	2,622,984	2020
7	France *(more)*	Europe	2,938,271	2021	2,715,518	2020	2,603,004	2020
8	Italy *(more)*	Europe	2,106,287	2021	2,003,576	2020	1,886,445	2020
9	Canada *(more)*	Americas	1,883,487	2021	1,741,496	2020	1,643,408	2020
10	South Korea *(more)*	Asia	1,806,707	2021	1,646,539	2020	1,630,525	2020
11	Russia *(more)*	Europe	1,710,734	2021	1,692,930	2020	1,483,498	2020
12	Brazil *(more)*	Americas	1,491,772	2021	1,847,795	2020	1,444,733	2020
13	Australia *(more)*	Oceania	1,617,543	2021	1,380,207	2020	1,330,901	2020
14	Spain *(more)*	Europe	1,461,552	2021	1,393,490	2020	1,281,199	2020
15	Mexico *(more)*	Americas	1,192,480	2021	1,256,440	2020	1,076,163	2020
16	Indonesia *(more)*	Asia	1,158,783	2021	1,119,190	2020	1,058,424	2020

							Playing Time				Performance						
Rk	Player	Nation	Pos	Squad	Age	Born	MP	Starts	Min	90s	Gls	Ast	G-PK	PK	PKatt	CrdY	CrdR
1	Max Aarons	ENG	DF	Norwich City	21-253	2000	4	4	360	4.0	0	0	0	0	0	1	0
2	Che Adams	SCO	FW	Southampton	25-063	1996	3	3	250	2.8	0	1	0	0	0	0	0
3	Rayan Aït Nouri	FRA	FW	Wolves	20-100	2001	1	0	7	0.1	0	0	0	0	0	0	0
4	Kristoffer Ajer	NOR	DF	Brentford	23-150	1998	4	4	340	3.8	0	0	0	0	0	1	0
5	Nathan Aké	NED	DF	Manchester City	26-208	1995	1	1	90	1.0	0	0	0	0	0	0	0
6	Marc Albrighton	ENG	MF,FW	Leicester City	31-300	1989	2	2	180	2.0	1	0	1	0	0	1	0
7	Thiago Alcántara	ESP	MF	Liverpool	30-156	1991	3	1	116	1.3	0	1	0	0	0	0	0
8	Trent Alexander-Arnold	ENG	DF	Liverpool	22-342	1998	4	4	360	4.0	0	2	0	0	0	0	0
9	Alisson	BRA	GK	Liverpool	28-347	1992	4	4	360	4.0	0	0	0	0	0	0	0
10	Allan	BRA	MF	Everton	30-249	1991	4	4	360	4.0	0	1	0	0	0	0	0
11	Dele Alli	ENG	MF	Tottenham	25-156	1996	4	4	360	4.0	1	0	0	1	1	1	0
12	Miguel Almirón	PAR	MF	Newcastle Utd	27-216	1994	4	4	360	4.0	0	0	0	0	0	0	0
13	Marcos Alonso	ESP	DF	Chelsea	30-260	1990	4	4	355	3.9	1	0	1	0	0	1	0
14	Steven Alzate	COL	MF	Brighton	23-006	1998	1	1	71	0.8	0	0	0	0	0	0	0
15	Daniel Amartey	GHA	DF	Leicester City	26-267	1994	3	3	270	3.0	0	0	0	0	0	0	0
16	Joachim Andersen	DEN	DF	Crystal Palace	25-106	1996	4	3	304	3.4	0	0	0	0	0	1	0
17	Michail Antonio	JAM	FW	West Ham	31-170	1990	4	4	356	4.0	4	3	4	0	1	2	1
18	Adam Armstrong	ENG	FW	Southampton	24-216	1997	4	4	343	3.8	1	0	1	0	0	0	0
19	Pierre-Emerick Aubameyang	GAB	FW	Arsenal	32-088	1989	3	2	178	2.0	1	0	1	0	0	0	0
20	Jordan Ayew	GHA	FW,MF	Crystal Palace	30-003	1991	4	3	288	3.2	0	0	0	0	0	0	0

Variables of Interest

Existing

- Age
- Matches played
- Starts
- Min
- 90s

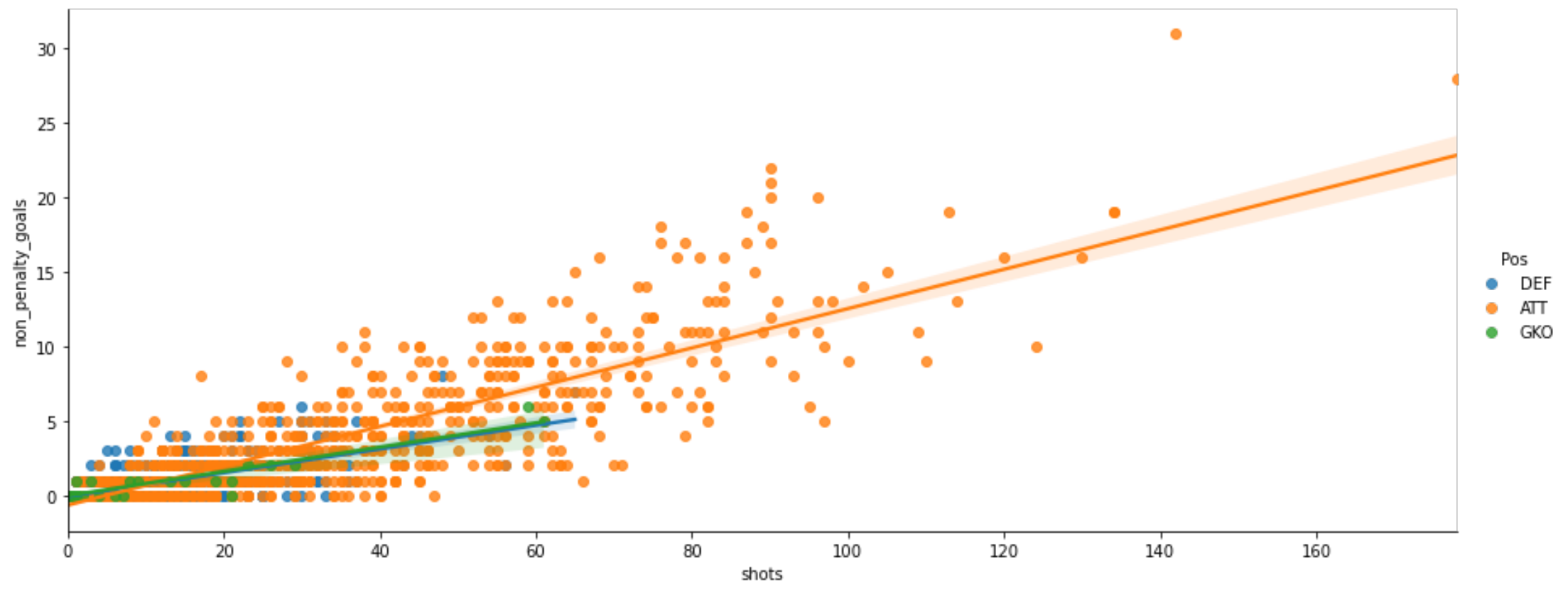
Collapsed

- Position
- Team
- Continent

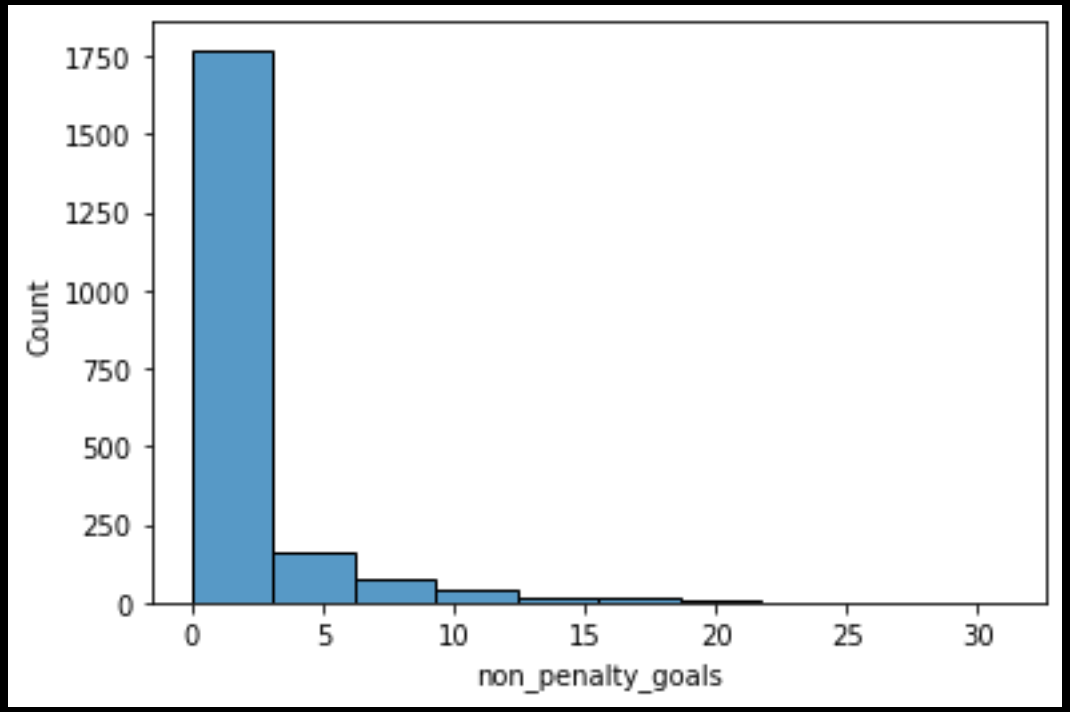
Interaction

- position*shots
- GDP*continent

Predictors


non_penalty_goals
shots
Pos
DEF
ATT
GKO
0
5
10
15
20
25
30
0
20
40
60
80
100
120
140
160

Distribution of non-penalty goals



First Models

Scaled

- Lasso R^2: 0.726 +- 0.042
- Ridge R^2: 0.727 +- 0.040
- ElasticNet R^2: 0.700 +- 0.087

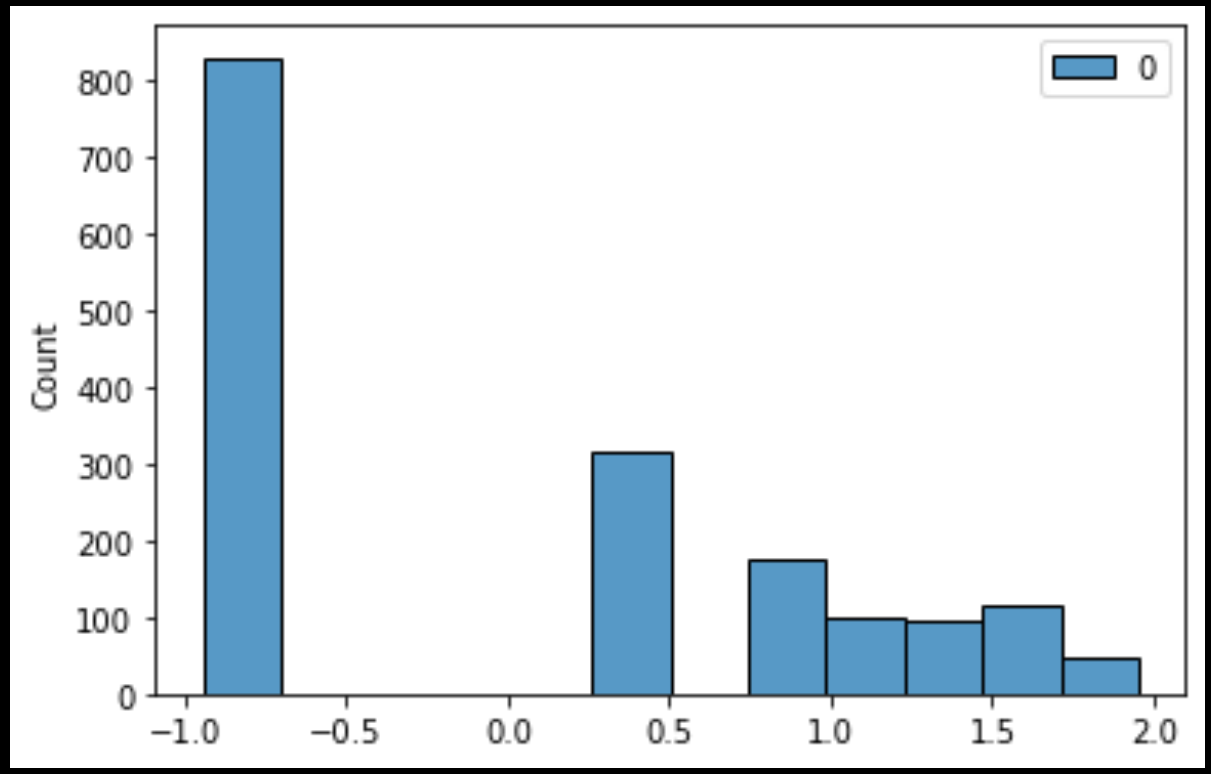
Not Scaled

- Polynomial Features R^2: -15.125 +- 31.535
- Simple R^2: 0.750 +- 0.051
- Lasso R^2: 0.749 +- 0.051
- Ridge R^2: 0.750 +- 0.051
- ElasticNet R^2: 0.750 +- 0.051

	variables	vif
0	shots	3.599829
1	tackles_won	4.666362
2	Pos_DEF	3.385871
3	Pos_GKO	2.093084
4	Squad_Top	1.510735
5	Continent_Europe	14.780894
6	Continent_South America	3.930148
7	Continent_other	2.059496
8	DefXshots	2.385224
9	GKOXshots	1.129904
10	gpdXeurope	1973.533991
11	gpdXSA	61.844581
12	gpdXo	1867.991593
13	Estimate.1	3889.110788
14	Age	14.871670
15	Min	9.602529

BoxCox transformation

- Transformed:
  - Goals (shown)
  - Shots
  - Tackles

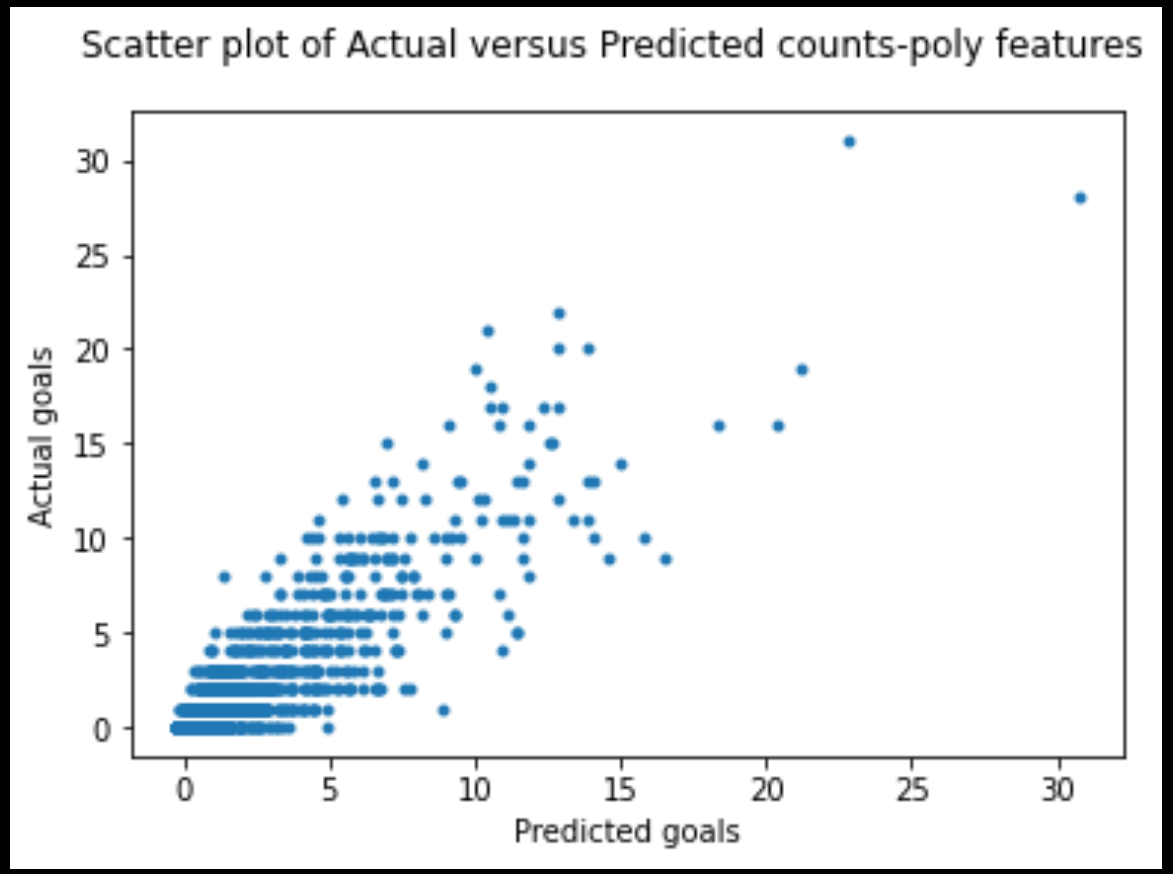


Simple R^2: 0.663 +- 0.026
Lasso R^2: 0.658 +- 0.026
Ridge R^2: 0.656 +- 0.026
Elastic Net R^2: 0.657 +- 0.026

Final Model

- Polynomial Features R^2: 0.754 +- 0.045
- Cook's d (max): 0.227
- Durbin-Watson: 2.005

	variables	vif
0	shots	1.282513
1	Pos_DEF	2.114895
2	Pos_GKO	1.095816
3	Squad_Top	1.362244
4	DefXshots	2.076611
5	GKOXshots	1.073879

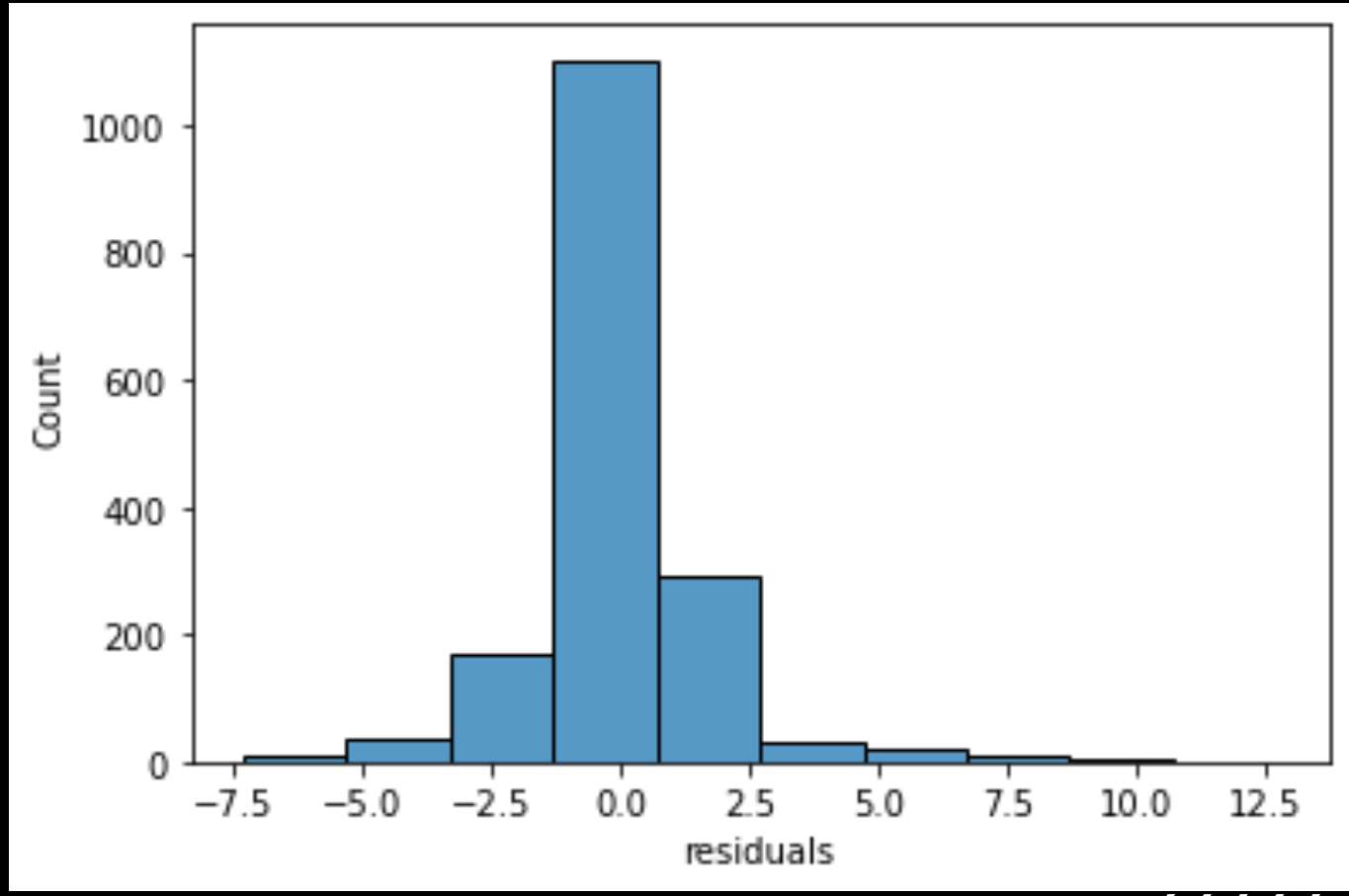


Model Fit Test

- R^2: 0.788
- MSE: 2.020
- MAE: 0.871

Performance:

- random.randint(0, 20)
  - 16, 8
- lm_poly.predict(X_poly_test[[rand_int]])
  - 1.6, 2.6
- y_test[[randint]]
  - 0, 1



Coefficients & Intercept

- Intercept: -0.205
- Shots: 0.000
- Pos_DEF: 0.089
- Pos_GKO: 0.099
- Squad_Top: 0.101
- DefXshots: -0.021
- GKOXshots: -0.008

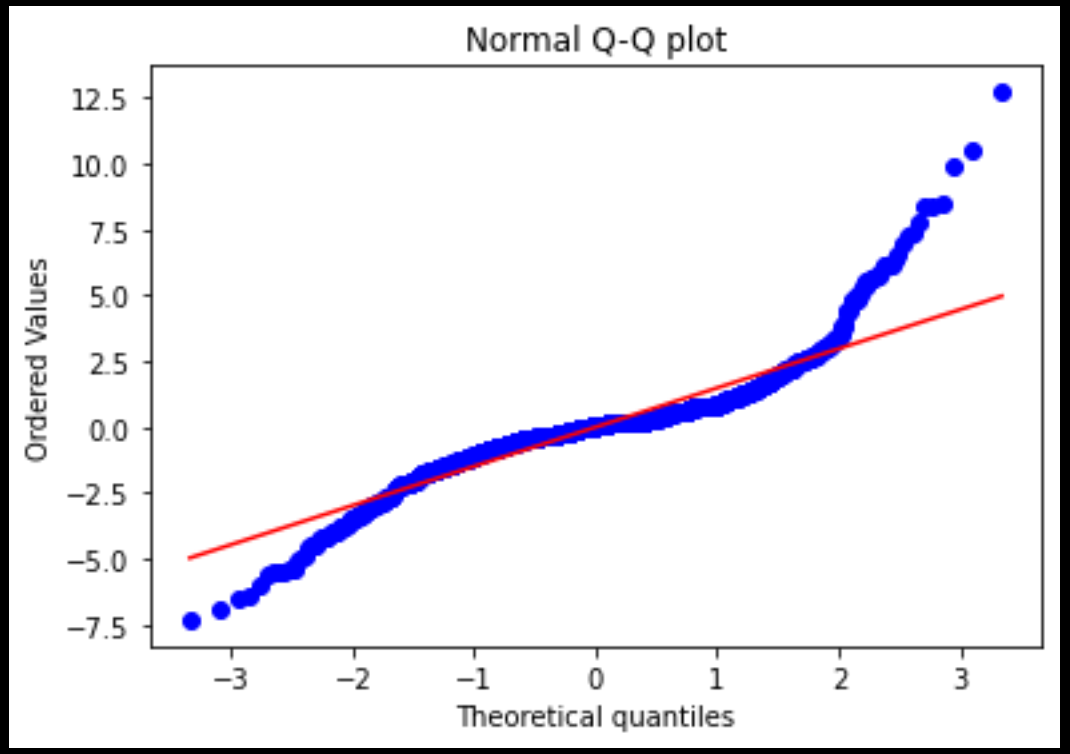
First round supplementary

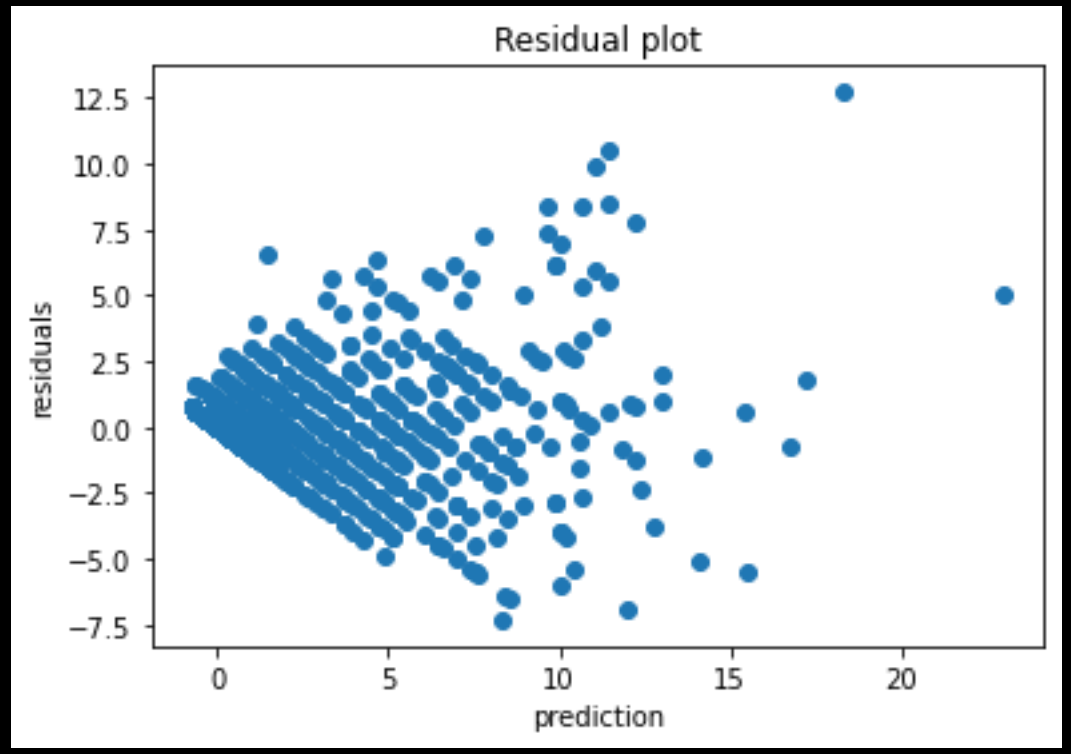
- basic r^2: 0.8075057399285094
- ridge r^2: 0.8078898169222304
- lasso r^2: 0.8092520529861036
- eNet r^2: 0.8074703871326342
- poly r^2: 0.7761784037632165
- basic mse: 1.8318209079250967
- ridge mse: 1.8281659404102744
- lasso mse: 1.8152025798279599
- eNet mse: 1.8321573334923515
- poly mse: 2.1299392484712985
- basic mae: 0.9082069848828243
- ridge mae: 0.9049915432396918
- lasso mae: 0.8905999674002204
- eNet mae: 0.9080380681447343
- poly mae: 0.9885802204892361

BoxCox w/min variables

- Simple mean cv r^2: 0.653 +- 0.027
- Lasso mean cv r^2: 0.646 +- 0.027
- Ridge mean cv r^2: 0.647 +- 0.028
- eNet mean cv r^2: 0.646 +- 0.027
- poly mean cv r^2: 0.670 +- 0.026

Residuals and QQ plot





Actual first model*

OLS Regression Results

Dep. Variable:	Gls	R-squared:	0.338
Model:	OLS	Adj. R-squared:	0.330
Method:	Least Squares	F-statistic:	43.02
Date:	Sun, 05 Sep 2021	Prob (F-statistic):	2.10e-95
Time:	18:53:27	Log-Likelihood:	-2917.2
No. Observations:	1193	AIC:	5864.
Df Residuals:	1178	BIC:	5941.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.6440	0.297	-2.166	0.031	-1.227	-0.061
index	7.731e-05	4.5e-05	1.718	0.086	-1.1e-05	0.000
Age	0.0149	0.020	0.762	0.446	-0.023	0.053
MP	0.0940	0.027	3.522	0.000	0.042	0.146
Starts	0.1335	0.088	1.513	0.130	-0.040	0.307
Min	0.0724	0.033	2.165	0.031	0.007	0.138
90s	-6.6072	3.001	-2.201	0.028	-12.496	-0.719
Continent_Africa	-0.1349	0.288	-0.468	0.640	-0.700	0.431
Continent_Asia	-1.0026	0.500	-2.004	0.045	-1.984	-0.021
Continent_Europe	-0.4695	0.207	-2.265	0.024	-0.876	-0.063
Continent_North America	-0.1167	0.428	-0.273	0.785	-0.956	0.722
Continent_Oceania	0.9252	0.810	1.142	0.254	-0.664	2.515
Continent_South America	0.1545	0.317	0.487	0.627	-0.468	0.777
Pos_ATT	1.4250	0.165	8.611	0.000	1.100	1.750
Pos_DEF	-0.8336	0.157	-5.298	0.000	-1.142	-0.525
Pos_GKO	-1.2354	0.254	-4.859	0.000	-1.734	-0.737
Squad_Not	-0.8769	0.175	-5.010	0.000	-1.220	-0.533
Squad_Top	0.2329	0.173	1.349	0.178	-0.106	0.572

Omnibus:	778.821	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11550.428
Skew:	2.811	Prob(JB):	0.00
Kurtosis:	17.169	Cond. No.	9.82e+19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.76e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.