Football Analysis

An in-depth analysis of performance indices in the world of football

Index

Introduction

- How much do statistics matter in determining success in sports?
  - Baseball: **SABERMETRICS**, Bill James
  - American Football
  - Basketball
  - Ice Hockey
  - Football

Introduction

SPORTS REFERENCE

Baseball-Reference.com

Baseball statistics from 1871 to the present for major league players, teams, and leagues. Complete postseason and managerial data is included as well. Minor league stats back to 1877, and box scores, gamelogs and splits back to 1904. Updated daily.

Basketball-Reference.com

Professional basketball statistics for every season from Abdul-Jabbar to Zaslofsky. Find statistics for your favorite player, team, or league. The site also includes sections for coaches, awards, leaders, and the playoffs. Also has sections on G-League, WNBA, international leagues and Olympics. Updated daily.

College Football @ Sports-Reference.com

College Football game results for every major school back to Princeton vs Rutgers in 1869. Player Stats for recent players schools, or conferences back to 1956. The site also includes sections for coaches, awards, leaders, and bowl games. Use the Play Index to create your own custom searches.

College Basketball @ Sports-Reference.com

College Basketball Stats and Game Result. Game results for every game featuring two major schools since 1949. Player Stats from Stephen Curry to Bill Walton. Stats for schools and conferences. The site also includes sections for coaches, awards, leaders, and the postseason. Use the Play Index to create your own custom searches.

Pro-Football-Reference.com

Football statistics and game results for all of NFL, AFL, AAFC & APFA history. The site includes player, Coach, and team stats, games scores for each team, Pro Bowl selections for each season, extensive leaderboards, and every draft pick ever selected. Our play index allows for custom searches through every NFL box score since 1960, every TD in NFL history and every play since 1994. Additionally, you can search through every play in Super Bowl history. Updated the morning after every NFL game.

Hockey-Reference.com

Hockey statistics from the creators of *Total Hockey* to our website. Find statistics for your favorite player from Aalto to Zyuzin, team, or league.The site also includes sections for coaches, drafts, awards, leaders, and the playoffs. Updated daily.

FBref.com

Football/Soccer statistics from all over the world. Basic statistics and information from the world of football with new countries being added regularly. Sign up to our e-mail list for updates on our progress.

Sports Reference Blog

All of the latest news and announcements for the Sports Reference sites

Introduction

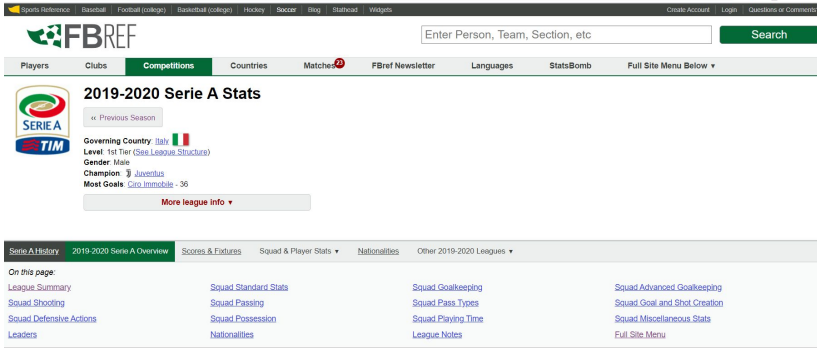
"Many stats matter. End of story. Basketball, Football, Baseball, Hockey, Soccer, you pick the sport, there are statistics which really matter and cut to the core of the game. Understanding them is critically important."

"Many stats don’t matter. There are a lot of stats people obsess over that actually don’t say much."

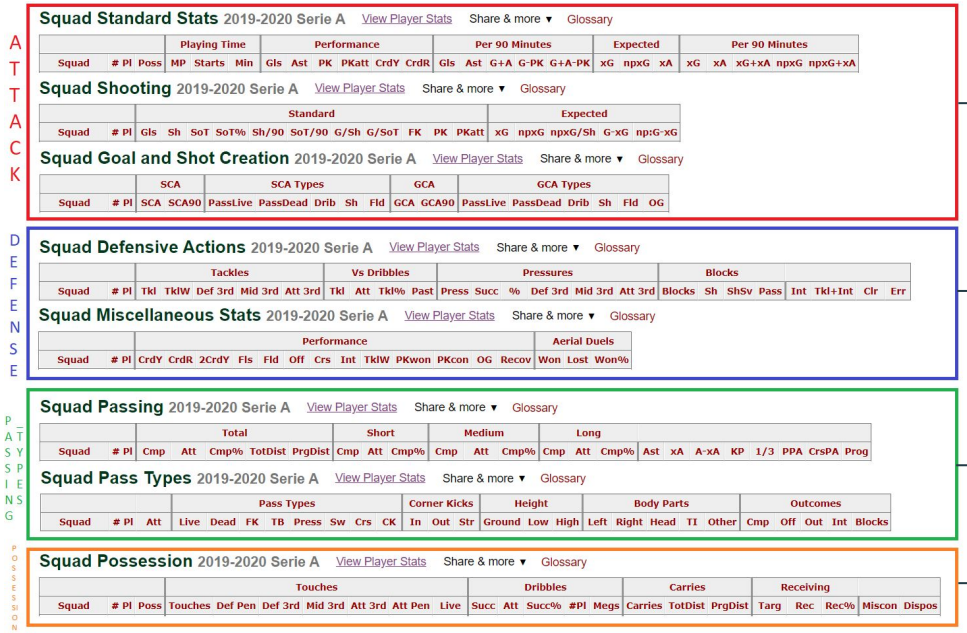
Forbes

Dataset creation

- *FBref.com* tables: League Summary, Squad Standard Stats, Squad Goalkeeping, Squad Advanced Goalkeeping, Squad Shooting, Squad Passing, Squad Pass Types, Squad Goal and Shot Creation, Squad Defensive Actions, Squad Possession, Squad Playing Time, Squad Miscellaneous Stats
- Seasons 17/18, 18/19, 19/20
- CSV files retrieval (API, website's solutions)
- 4 reference tables:
  - *Attack*
  - *Defense*
  - *Passing_types*
  - *Possession*
- Merging with SPI and FIFA OVR values



Dataset creation



Too many features, we need a dimensionality reduction technique

PCA

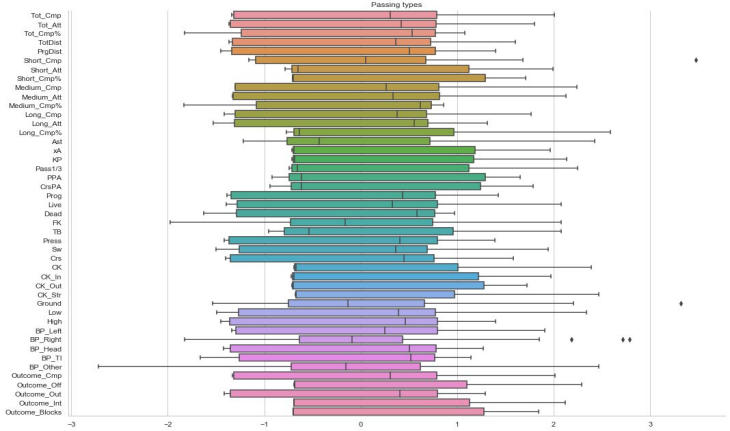
Features analysis

PASSING_TYPES

	season	Squad	# Pl	Tot_Cmp	Tot_Att	Tot_Cmp%	TotDist	PrgDist	Short_Cmp	Short_Att	Short_Cmp	Medium_C	Medium_A	Medium_C	Long_Cmp	Long_Att	Long_Cmp	Ast
count	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60
unique	3	25																
top	2017/18	SPAL																
freq	20	3																
mean			28.27	10085.42	12827.30	65.56	199557.73	71564.25	166.12	1487.27	1645.21	7510.92	8609.28	68.87	2494.30	3968.65	146.20	48.75
std			3.57	7496.21	9056.68	19.22	142179.22	43763.41	70.87	1336.63	2304.33	5761.76	6529.52	25.92	1595.10	2144.82	121.92	28.78
min			22	168	541	30.7	7310	8638	84.6	449	27.7	59.3	15	21.8	253	722	52.9	14
25%			25	284.75	712	42	12040.5	13456.25	89.75	538.5	32.6	72.725	43	41.05	432	1188.25	62.575	27
50%			29	12384	16585	75.7	250532.5	93363	169.5	626	38.55	9033.5	10759.5	84.75	3087.5	5146.5	69.5	36.5
75%			30	15946.25	19829.75	80.25	302127.3	105071.8	213.75	2976	4591	12122.5	13889.5	87.75	3570.5	5450.5	262.75	69.25
max			41	24985	29004	86.1	425699	132297	410	4128	5547	20299	22359	90.9	5288	6763	459	118

xA	KP	Pass1/3	PPA	CrsPA	Prog	Live	Dead	FK	TB	Press	Sw	Crs	CK	CK_In	CK_Out	CK_Str	Ground
60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60
523.25	5990.65	5928.38	799.88	237.25	1138.87	12279.92	1402.13	559.12	82.40	1964.30	411.82	511.42	3908.87	881.93	1127.90	1676.93	11642.42
707.67	8158.26	6988.35	681.04	207.87	818.21	7330.78	631.84	110.71	78.44	1395.27	262.37	365.54	5509.68	1218.40	1575.36	2502.04	2889.71
21.2	235	763	178	44	14	2157	384	342	8	7	21	4	128	9	10	1	7248
30.075	333.75	982.25	299	88.5	48.25	2977.25	596	479	20.75	73.5	84.75	22	176.5	28.75	24.75	10	9496.25
38.6	450.5	1351	386	110.5	1493	14682.5	1766.5	541.5	40.5	2525.5	507	675	229	47.5	35	15	11263
1355.5	15483	13717.75	1670.75	493.75	1769.5	18064.5	1882	640.75	156.75	3066.5	591.25	787.5	9418.75	2360.25	3117	4093.75	13543
1899	23274	21495	1914	605	2294	27360	2010	787	244	3884	917	1084	16954	3262	3819	7798	21142

Low	High	BP_Left	BP_Right	BP_Head	BP_TI	BP_Other	Outcome_	Outcome_	Outcome_	Outcome_I	Outcome_
60	60	60	60	60	60	60	60	60	60	60	60
1917.90	2922.50	3733.97	12145.87	551.28	594.70	263.62	10140.48	4707.48	266.11	91245.95	30467.68
1029.59	1654.96	2685.64	2976.13	368.12	234.04	40.25	7422.81	6807.73	136.29	132211.32	43101.31
394	541	167	6753	33	209	155	372	32	74.7	193	404
621.5	695	289	10260.5	58	301.75	234.75	457.5	64.75	83.375	264	447.75
2314	3684	4398.5	11874.5	736	715	257.5	12384	74	321	323	484
2711	4223	5847.75	13418.75	836	773.25	288.25	15946.25	12143.5	374	238678.5	84930
4307	5216	8817	20370	1014	859	362	24985	20158	441	369600	109387



Soccer Power Index interpretation

Proposed by: FiveThirtyEight

Inspiration: ESPN

Metrics to evaluate a team's performance after each match:

- Adjusted goals;
- Shot-based expected goals;
- Non-shot expected goals.

This section goal: better understanding SPI

PCA

“You should not add a comma to what can be said in a few words” (Gregory David Roberts)

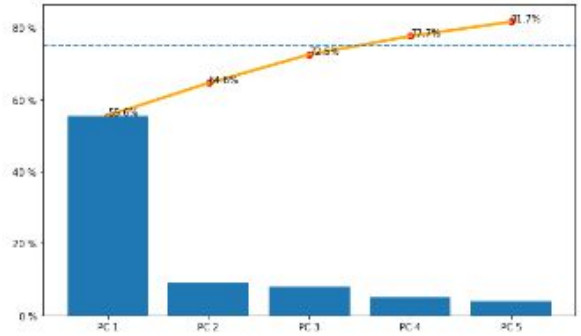
Loss of information \ Simplification of the problem

scikit-learn
sklearn.decomposition.PCA

Analysis of PCs (1)

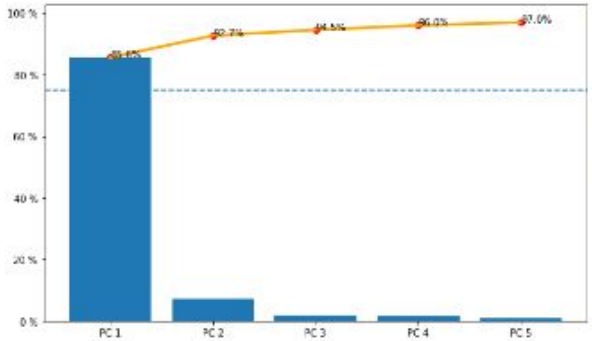
We divided the initial features in four fields: *attack, defense, passing types, possession*

Teams Attack table PCA



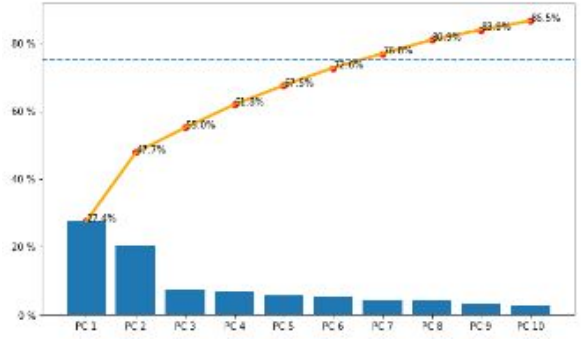
Eigenvalues in descending order:
19.44581060920294
3.167416181025411
2.7763739918632733
1.7906091623990055
1.3981575381904867
1.2848514224866688
0.9415888695198359
0.7842838187521072
0.6931994156990342
0.5944080270506058

Teams Passing Types table PCA



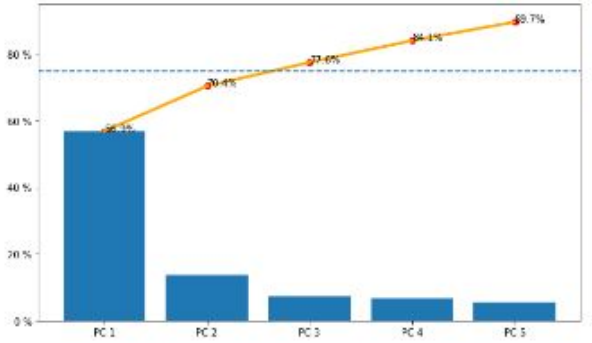
Eigenvalues in descending order:
39.37088975804935
3.272825161815779
0.8405114046727779
0.6813565316840638
0.43214090726215754
0.3413248329586805
0.3289744387997798
0.16075264280071297
0.10897401962307073
0.09006016166179349

Teams Defense table PCA



Eigenvalues in descending order:
8.233006318800763
6.863380368575083
2.205418876470672
2.0460127035814604
1.6945751244531504
1.552364981453759
1.2578872598891906
1.2207768322936452
0.8924498328439389
0.7840307553127964

Teams Possession table PCA



Eigenvalues in descending order:
12.513920478562747
2.9800760345896674
1.5673157570660143
1.4425025003059242
1.2292120644546443
0.6839380057870818
0.45188128335463146
0.36039340193138486
0.2705539723621587
0.1777162618110661

Brief focus on correlation

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	SPI
PC1	1	-1.3e-11	2.5e-11	0.058	0.79	-0.15	0.063	-0.033	-0.14	0.063	0.84	0.84	-0.17	0.92
PC2	-1.3e-11	1	6.9e-11	0.1	-0.17	0.054	0.39	-0.17	-0.071	0.12	-0.0049	-0.1	0.22	-0.15
PC3	2.5e-11	6.9e-11	1	0.36	-0.093	0.053	-0.23	0.097	-0.26	0.43	-0.12	-0.18	-0.059	-0.025
PC4	0.058	0.1	0.36	1	7e-12	-7.8e-12	-1.9e-11	-3.5e-11	-1.7e-11	0.59	0.079	-0.073	-0.17	-0.014
PC5	0.79	-0.17	-0.093	7e-12	1	-3.4e-12	-2.5e-11	8.7e-12	2.1e-11	-0.14	0.87	0.87	-0.27	0.81
PC6	-0.15	0.054	0.053	-7.8e-12	-3.4e-12	1	3.8e-12	-2.8e-11	1.2e-11	-0.075	-0.26	-0.13	0.38	-0.19
PC7	0.063	0.39	-0.23	-1.9e-11	-2.5e-11	3.8e-12	1	6.2e-12	4.2e-12	0.036	0.074	-0.0082	0.047	-0.08
PC8	-0.033	-0.17	0.097	-3.5e-11	8.7e-12	-2.8e-11	6.2e-12	1	-3.5e-11	0.15	-0.011	-0.13	-0.26	0.0077
PC9	-0.14	-0.071	-0.26	-1.7e-11	2.1e-11	1.2e-11	4.2e-12	-3.5e-11	1	-0.35	-0.084	-0.052	-0.12	-0.017
PC10	0.063	0.12	0.43	0.59	-0.14	-0.075	0.036	0.15	-0.35	1	-1.7e-11	-0.25	-0.21	-0.041
PC11	0.84	-0.0049	-0.12	0.079	0.87	-0.26	0.074	-0.011	-0.084	-1.7e-11	1	0.9	-0.29	0.81
PC12	0.84	-0.1	-0.18	-0.073	0.87	-0.13	-0.0082	-0.13	-0.052	-0.25	0.9	1	-4.5e-12	0.83
PC13	-0.17	0.22	-0.059	-0.17	-0.27	0.38	0.047	-0.26	-0.12	-0.21	-0.29	-4.5e-12	1	-0.23
SPI	0.92	-0.15	-0.025	-0.014	0.81	-0.19	-0.08	0.0077	-0.017	-0.041	0.81	0.83	-0.23	1

SPI PCA correlation matrix

Feature	Correlation value
PC1: Capacità Finalizzazione	0.92
PC2: Capacità Realizzazione	-0.15
PC3: Tiro da palla inattiva	-0.02
PC4: Recupero palla nella propria metà campo	-0.01
PC5: Atteggiamento difensivo (catenaccio)	0.81
PC6: Contrasto dei dribbling / marcatura a zona	-0.19
PC7: Aggressività/gioco duro	-0.08
PC8: Gioco remissivo	0.01
PC9: Difesa scoordinata (errori tecnici/eccessiva foga)	-0.02
PC10: Passaggi propositivi media gittata (precisione nei passaggi e pochi tocchi in area)	-0.04
PC11: Palla contesa a media altezza (gioco medio-alto / palla non giocata a terra)	0.81
PC12: Controllo palla nella metà campo avversaria	0.83
PC13: Efficacia del possesso	-0.23
SPI	1.00

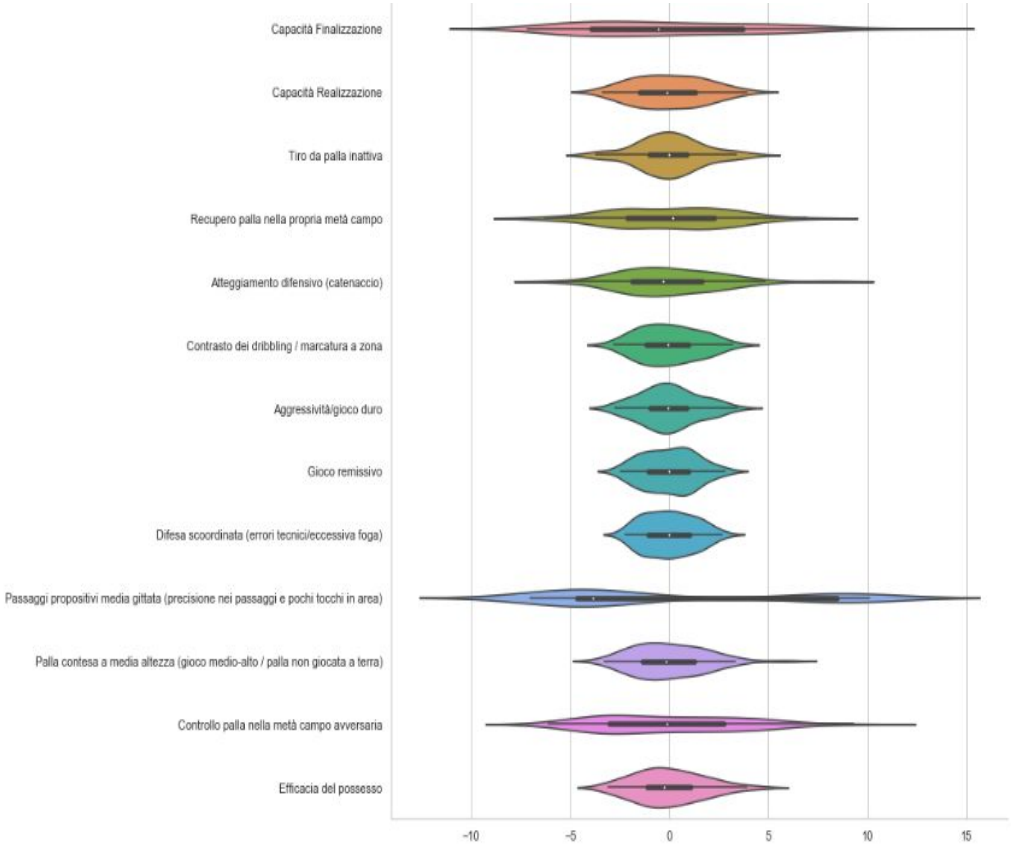
Correlation values among SPI index and PCs

Analysis of PCs (2)

Attack	Capacità Finalizzazione, Capacità Realizzazione, Tiro da palla inattiva
Defense	Recupero palla nella propria metà campo, Atteggiamento difensivo (catenaccio), Contrasto dei dribbling / marcatura a zona, Aggressività / gioco duro, Gioco remissivo, Difesa scoordinata (errori tecnici/eccessiva foga)
Passing Types	Passaggi propositivi media gittata (precisione passaggi e pochi tocchi in area), Palla contesa a media altezza (gioco medio-alto / palla non giocata a terra)
Possession	Controllo palla nella metà campo avversaria, Efficacia del possesso

	Capacità Finalizzazione	Capacità Realizzazione	Tiro da palla inattiva	Recupero palla nella propria	Atteggiamento difensivo	Contrasto dei dribbling/	Aggressività /gioco duro	Gioco remissivo	Difesa scoordinata (errori	Passaggi propositivi media	Palla contesa a media	Controllo palla nella metà	Efficacia del possesso
count	60	60	60	60	60	60	60	60	60	60	60	60	60
unique													
top													
freq													
mean	-5.00E-11	2.78E-17	-8.33E-11	3.33E-11	3.33E-11	-1.67E-11	3.33E-11	-5.00E-11	-6.67E-11	-1.67E-11	-5.00E-11	3.33E-11	5.00E-11
std	4.45	1.79	1.68	2.89	2.48	1.50	1.44	1.31	1.26	6.33	1.82	3.57	1.74
min	-7.18	-3.37	-3.73	-6.31	-5.63	-2.82	-2.77	-2.46	-2.21	-7.03	-3.26	-6.13	-3.09
25%	-3.97	-1.49	-1.01	-2.11	-1.89	-1.15	-0.99	-1.07	-1.05	-4.68	-1.38	-3.04	-1.11
50%	-0.52	-0.10	-0.04	0.16	-0.29	-0.08	-0.05	-0.01	0.01	-3.83	-0.16	-0.13	-0.24
75%	3.72	1.34	0.89	2.27	1.67	1.02	0.92	0.98	1.06	8.48	1.27	2.75	1.08
max	11.47	3.91	4.12	6.95	8.14	3.21	3.42	2.82	2.70	10.10	5.83	9.30	4.48

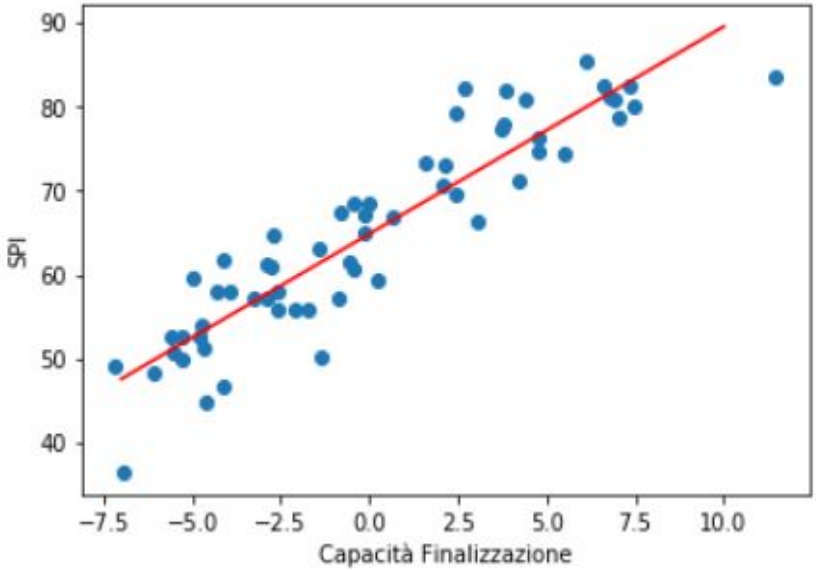
Teams Principal Components summary



Teams Principal Components Box-plots

Simple Linear Regression

Importing **LinearRegression** from *sklearn.linear_model* and fitting a linear regression with the most correlated feature, i.e. 'Capacità finalizzazione' ($\rho = 0.92$)




```
Coefficients:
 [2.47374139]
Intercept:
 64.85450000012369
Score:
 0.8416782208178843
```

$$Y \approx \beta_0 + \beta_1(X)$$

Multiple Linear Regression

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + ... + \beta_p(X_p) + \epsilon$$

```
========== SUMMARY ==========
Residuals:
    Min      1Q  Median      3Q   Max
-7.9182 -2.5793 -0.6083  2.6794  7.44


Coefficients:
                                                  Estimate  Std. Error   t value   p value
_intercept                                       64.854500    0.487490  133.0377  0.000000
Capacità Finalizzazione                           2.047121    0.218797    9.3562  0.000000
Capacità Realizzazione                           -0.222374    0.326461   -0.6812  0.498433
Tiro da palla inattiva                            0.092132    0.348368    0.2645  0.792342
Recupero palla nella propria metà campo          -0.262204    0.218502   -1.2000  0.234932
Atteggiamento difensivo (catenaccio)              0.819467    0.517403    1.5838  0.118584
Contrasto dei dribbling / marcatura a zona       -0.521419    0.403692   -1.2916  0.201523
Aggressività/gioco duro                          -0.791760    0.380709   -2.0797  0.041907
Gioco remissivo                                   0.277357    0.402131    0.6897  0.493075
Difesa scoordinata (errori tecnici/eccessiva foga)  0.925413    0.455245    2.0328  0.046583
Passaggi propositivi media gittata (precisione ...  0.077497    0.103894    0.7459  0.458672
Palla contesa a media altezza (gioco medio-alto...  -1.191480    1.239365   -0.9614  0.340293
Controllo palla nella metà campo avversaria       0.710940    0.613387    1.1590  0.251109
Efficacia del possesso                           -0.397232    0.439295   -0.9042  0.369542
---
R-squared:  0.89915,    Adjusted R-squared:  0.87064
F-statistic: 31.55 on 13 features
```

Lasso (1)

Using **Lasso** from *sklearn.linear_model*

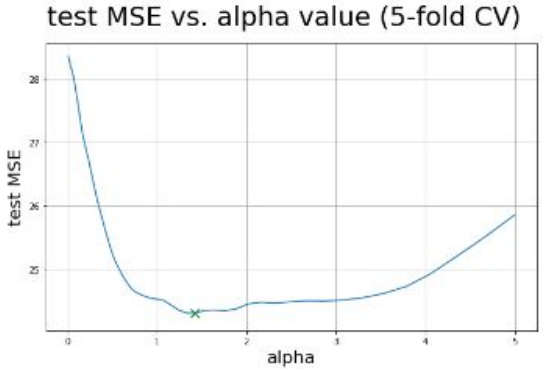
```
========== SUMMARY ==========
Residuals:
    Min      1Q  Median      3Q     Max
-7.9182 -2.5793 -0.6083  2.6794  7.44


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                     64.854500    0.487490  133.0377  0.000000
Capacità Finalizzazione                         2.047121    0.218797    9.3562  0.000000
Capacità Realizzazione                         -0.222374    0.326461   -0.6812  0.498433
Tiro da palla inattiva                          0.092132    0.348368    0.2645  0.792342
Recupero palla nella propria metà campo        -0.262204    0.218502   -1.2000  0.234932
Atteggiamento difensivo (catenaccio)            0.819467    0.517403    1.5838  0.118584
Contrasto dei dribbling / marcatura a zona     -0.521419    0.403692   -1.2916  0.201523
Aggressività/gioco duro                        -0.791760    0.380709   -2.0797  0.041907
Gioco remissivo                                 0.277357    0.402131    0.6897  0.493075
Difesa scoordinata (errori tecnici/eccessiva foga)  0.925413    0.455245    2.0328  0.046583
Passaggi propositivi media gittata (precisione ...  0.077497    0.103894    0.7459  0.458672
Palla contesa a media altezza (gioco medio-alto...  -1.191480    1.239365   -0.9614  0.340293
Controllo palla nella metà campo avversaria     0.710940    0.613387    1.1590  0.251109
Efficacia del possesso                         -0.397232    0.439295   -0.9042  0.369542
---
R-squared:  0.89915,    Adjusted R-squared:  0.87064
F-statistic: 31.55 on 13 features
```

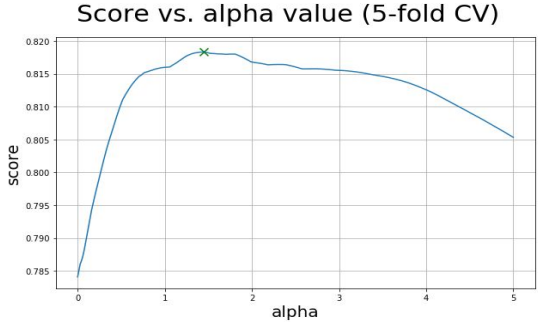
In this case we choose α arbitrarily (α = 0.1), but select a good value for α is critical, to do that we used *Cross Validation*

Lasso (2)

using **LassoCV** in *sklearn.linear_module*: α = 1.42, score = 0.87

using **cross_val_score** in *sklearn.model_selection:* α = 1.42, score = 0.82






```
========== SUMMARY ==========
Residuals:
    Min      1Q  Median      3Q     Max
-8.7678 -2.7146 -0.3193  2.0716  12.152


Coefficients:
```

	Estimate	Std. Error	t value	p value
_intercept	64.854500	0.543371	119.3558	0.000000
Capacità Finalizzazione	2.157990	0.243878	8.8486	0.000000
Capacità Realizzazione	-0.329387	0.363884	-0.9052	0.369043
Tiro da palla inattiva	0.000000	0.388302	0.0000	1.000000
Recupero palla nella propria metà campo	-0.000000	0.243549	-0.0000	1.000000
Atteggiamento difensivo (catenaccio)	0.399690	0.576713	0.6930	0.490998
Contrasto dei dribbling / marcatura a zona	-0.000000	0.449967	-0.0000	1.000000
Aggressività/gioco duro	-0.211625	0.424350	-0.4987	0.619843
Gioco remissivo	0.000000	0.448228	0.0000	1.000000
Difesa scoordinata (errori tecnici/eccessiva foga)	0.000000	0.507431	0.0000	1.000000
Passaggi propositivi media gittata (precisione ...	-0.086136	0.115803	-0.7438	0.459939
Palla contesa a media altezza (gioco medio-alto...	0.000000	1.381435	0.0000	1.000000
Controllo palla nella metà campo avversaria	0.113087	0.683700	0.1654	0.869191
Efficacia del possesso	-0.024525	0.489652	-0.0501	0.960222

```
---
R-squared:  0.87470,    Adjusted R-squared:  0.83929
F-statistic: 24.70 on 13 features
```

Ridge (1)

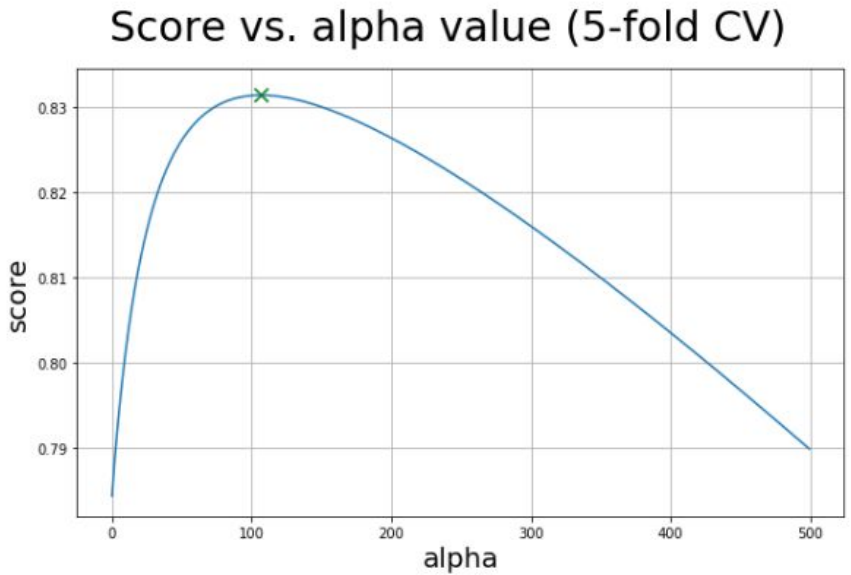
Using **Ridge** from *sklearn.linear_model*

```
========== SUMMARY ==========
Residuals:
    Min      1Q  Median      3Q     Max
-7.9198 -2.5793 -0.6076  2.6773  7.457


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                      64.854500    0.487491  133.0374  0.000000
Capacità Finalizzazione                          2.047935    0.218798    9.3599  0.000000
Capacità Realizzazione                          -0.225557    0.326462   -0.6909  0.492330
Tiro da palla inattiva                           0.092788    0.348369    0.2663  0.790899
Recupero palla nella propria metà campo         -0.261860    0.218502   -1.1984  0.235540
Atteggiamento difensivo (catenaccio)             0.816450    0.517404    1.5780  0.119919
Contrasto dei dribbling / marcatura a zona      -0.519150    0.403692   -1.2860  0.203466
Aggressività/gioco duro                         -0.791724    0.380710   -2.0796  0.041917
Gioco remissivo                                  0.275871    0.402132    0.6860  0.495387
Difesa scoordinata (errori tecnici/eccessiva foga)  0.924268    0.455246    2.0303  0.046846
Passaggi propositivi media gittata (precisione ...  0.076171    0.103894    0.7332  0.466361
Palla contesa a media altezza (gioco medio-alto...  -1.170922    1.239367   -0.9448  0.348628
Controllo palla nella metà campo avversaria      0.701662    0.613388    1.1439  0.257280
Efficacia del possesso                          -0.392823    0.439296   -0.8942  0.374843
---
R-squared:  0.89915,    Adjusted R-squared:  0.87064
F-statistic: 31.55 on 13 features
```

In this case we choose α arbitrarily (α = 0.1), but select a good value for α is critical, to do that we used *Cross Validation*

Ridge (2)

using **cross_val_score** in *sklearn.model_selection:* α = 106, score = 0.85




```
=========== SUMMARY ===========
Residuals:
    Min      1Q  Median      3Q     Max
-8.1331 -2.8128 -0.4165  2.4513  11.2692


Coefficients:
                                                 Estimate  Std. Error  t value  p value
_intercept                                     64.854500    0.520341  124.6386 0.000000
Capacità Finalizzazione                         1.650898    0.233542    7.0690 0.000000
Capacità Realizzazione                         -0.321265    0.348461   -0.9220 0.360307
Tiro da palla inattiva                          0.121614    0.371844    0.3271 0.744782
Recupero palla nella propria metà campo        -0.168454    0.233226   -0.7223 0.472977
Atteggiamento difensivo (catenaccio)            0.515573    0.552270    0.9336 0.354339
Contrasto dei dribbling / marcatura a zona     -0.204931    0.430896   -0.4756 0.636120
Aggressività/gioco duro                        -0.420413    0.406364   -1.0346 0.305092
Gioco remissivo                                 0.129923    0.429230    0.3027 0.763191
Difesa scoordinata (errori tecnici/eccessiva foga)  0.356157  0.485923  0.7329 0.466492
Passaggi propositivi media gittata (precisione ...  0.006701  0.110895  0.0604 0.952021
Palla contesa a media altezza (gioco medio-alto...  0.161121  1.322883  0.1218 0.903475
Controllo palla nella metà campo avversaria     0.576440    0.654722    0.8804 0.382196
Efficacia del possesso                         -0.294598    0.468899   -0.6283 0.532247
---
R-squared:  0.88510,    Adjusted R-squared:  0.85262
F-statistic: 27.26 on 13 features
```

Single Player Evaluation : FIFA Indexes to compare players

Ratings from 1 to 99 = OVR

36 Attributes, weighted mean of attributes 1 to 99= ATT

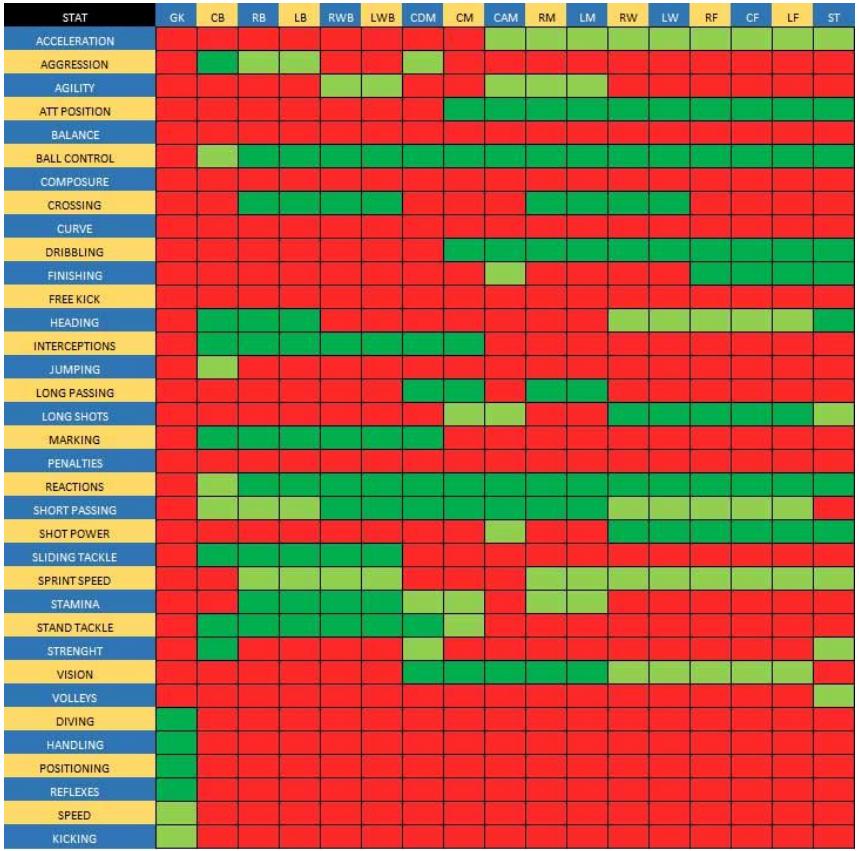
Additional Value = **International Reputation (1 to 3) =IR**

OVR = ATT + IR

Reference: sofifa.com

A ‘justified’ paradox with in-game evaluation

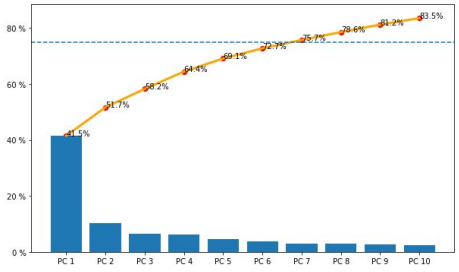
RATING BASE	★	★★	★★★	★★★★	★★★★★
01 – 28	0	0	0	0	0
29 – 33	0	0	0	0	+1
34 – 49	0	0	0	+1	+1
50 – 66	0	0	+1	+1	+2
67 – 74	0	0	+1	+2	+2
75 – 99	0	0	+1	+2	+3



Statistical learning models

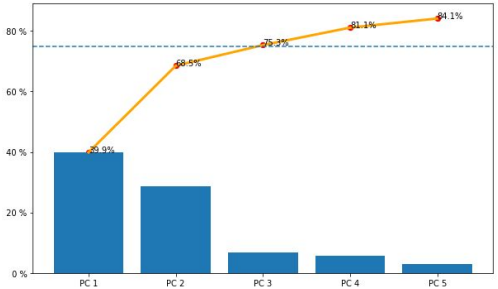
PCA once again chosen as the most effective way to improve pivotal features from a starting point of nearly a hundred variables

Players Attack table PCA



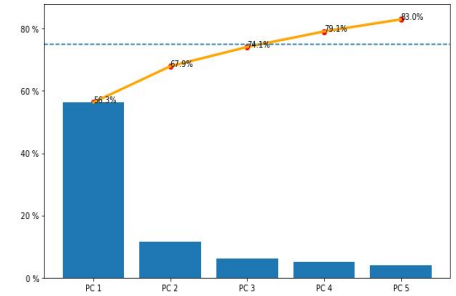
Eigenvalues in descending order:
14.11554086439977
3.4558714793529326
2.2278066733900954
2.1114433144798705
1.5855829182361507
1.2376835014449008
1.0159805030466185
0.9582980528127597
0.907191392620421
0.7856345063339142

Players Defense table PCA



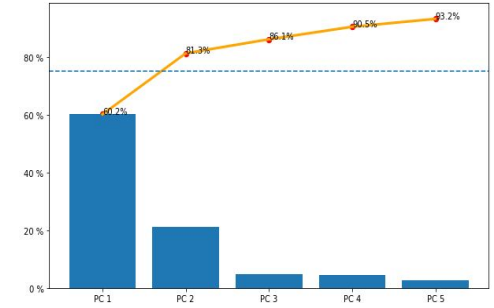
Eigenvalues in descending order:
11.567463569072801
8.303403689605533
1.9759947085485372
1.677447561046696
0.8705527578243994
0.8493620399629326
0.6986254607371665
0.4308012165628901
0.35979083059276556
0.3096959330229789

Players Passing Types table PCA



Eigenvalues in descending order:
25.334972385667136
5.216867713971147
2.791922714108185
2.243027987545319
1.7439158231186895
1.0083388958131743
0.93390809012832
0.7502722018182759
0.6264157349132303
0.5829813924812383

Players Possession table PCA



Eigenvalues in descending order:
12.03092746249175
4.225082744277282
0.9713903142219391
0.8676951322544811
0.5461807506689867
0.44168178258173063
0.4112111801473508
0.18940286132692288
0.1070313785781875
0.07514160404979048

Use of correlation with effective features shortage

	Incisività offensiva	Individualismo (palla in movimento)	Capacità di finalizzazione	Inefficacia del tiro	Gioco offensivo con palla in movimento	Altruismo	Supremazia atletica / giocate aggressive	Pressing efficace in zona difensiva	Passaggio propositivo e nell'ultimo 1/3 di campo	Gioco con palla in movimento	Inefficacia del passaggio	Alta propositività nella metà campo avversaria	Avanzamento e dribbling aggressivo/superfluo	shooting	passing	dribbling	defending	Overall_index
Incisività offensiva	1	-0.00079	-0.014	0.0085	-0.0019	0.00094	0.36	0.044	0.53	0.54	0.09	0.7	0.52	0.56	0.4	0.52	-0.35	0.44
Individualismo (palla in movimento)	-0.00079	1	0.0072	0.0017	-0.00034	0.012	-0.25	-0.078	-0.43	-0.27	-0.15	-0.27	0.16	0.054	-0.31	-0.14	-0.26	-0.035
Capacità di finalizzazione	-0.014	0.0072	1	-0.098	-0.034	0.035	-0.32	-0.074	-0.28	0.16	0.082	-0.26	0.19	0.23	0.11	0.2	-0.31	-0.071
Inefficacia del tiro	0.0085	0.0017	-0.098	1	-0.022	0.034	-0.28	-0.071	-0.28	0.15	-0.022	-0.23	0.24	0.17	-0.016	0.06	-0.28	-0.14
Gioco offensivo con palla in movimento	-0.0019	-0.00034	-0.034	-0.022	1	0.014	-0.052	0.0042	0.071	0.19	0.34	-0.12	-0.15	-0.0023	0.042	-0.025	0.0013	-0.016
Altruismo	0.00094	0.012	0.035	0.034	0.014	1	-0.062	0.014	-0.048	-0.07	-0.023	-0.092	-0.095	-0.08	-0.024	-0.043	0.1	0.026
Supremazia atletica / giocate aggressive	0.36	-0.25	-0.32	-0.28	-0.052	-0.062	1	-0.009	0.82	-0.17	-0.11	0.79	-0.29	-0.016	0.2	0.11	0.33	0.32
Pressing efficace in zona difensiva	0.044	-0.078	-0.074	-0.071	0.0042	0.014	-0.009	1	0.2	-0.03	0.044	0.16	-0.12	-0.022	0.067	0.013	0.1	0.053
Passaggio propositivo e nell'ultimo 1/3 di campo	0.53	-0.43	-0.28	-0.28	0.071	-0.048	0.82	0.2	1	0.0098	-0.0078	0.91	-0.28	0.073	0.33	0.23	0.31	0.46
Gioco con palla in movimento	0.54	-0.27	0.16	0.15	0.19	-0.07	-0.17	-0.03	0.0098	1	0.014	0.082	0.74	0.5	0.35	0.45	-0.51	-0.037
Inefficacia del passaggio	0.09	-0.15	0.082	-0.022	0.34	-0.023	-0.11	0.044	-0.0078	0.014	1	-0.052	-0.085	0.041	0.096	0.028	0.026	0.16
Alta propositività nella metà campo avversaria	0.7	-0.27	-0.26	-0.23	-0.12	-0.092	0.79	0.16	0.91	0.082	-0.052	1	0.0053	0.21	0.35	0.33	0.12	0.48
Avanzamento e dribbling aggressivo/superfluo	0.52	0.16	0.19	0.24	-0.15	-0.095	-0.29	-0.12	-0.28	0.74	-0.085	0.0053	1	0.61	0.26	0.48	-0.69	-0.053
shooting	0.56	0.054	0.23	0.17	-0.0023	-0.08	-0.016	-0.022	0.073	0.5	0.041	0.21	0.61	1	0.68	0.79	-0.51	0.36
passing	0.4	-0.31	0.11	-0.016	0.042	-0.024	0.2	0.067	0.33	0.35	0.096	0.35	0.26	0.68	1	0.86	-0.0004	0.51
dribbling	0.52	-0.14	0.2	0.06	-0.025	-0.043	0.11	0.013	0.23	0.45	0.028	0.33	0.48	0.79	0.86	1	-0.29	0.51
defending	-0.35	-0.26	-0.31	-0.28	0.0013	0.1	0.33	0.1	0.31	-0.51	0.026	0.12	-0.69	-0.51	-0.0004	-0.29	1	0.27
Overall_index	0.44	-0.035	-0.071	-0.14	-0.016	0.026	0.32	0.053	0.46	-0.037	0.16	0.48	-0.053	0.36	0.51	0.51	0.27	1

PCA Results

The thirteen principal components used for an effective comparison establish features

Feature	Correlation value
PC1: Incisività offensiva	0.44
PC2: Individualismo (palla in movimento)	-0.03
PC3: Capacità di finalizzazione	-0.07
PC4: Inefficacia del tiro	-0.14
PC5: Gioco offensivo con palla in movimento	-0.02
PC6: Altruismo	0.03
PC7: Supremazia atletica / giocate aggressive	0.32
PC8: Pressing efficace in zona difensiva	0.05
PC9: Passaggio propositivo e nell'ultimo 1/3 di campo	0.46
PC10: Gioco con palla in movimento	-0.04
PC11: Inefficacia del passaggio	0.16
PC12: Alta propositività nella metà campo avversaria	0.48
PC13: Avanzamento e dribbling aggressivo/superfluo	-0.05
OVR	1.00

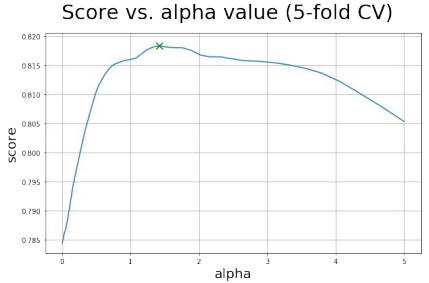
Studying the Overall index: linear methods

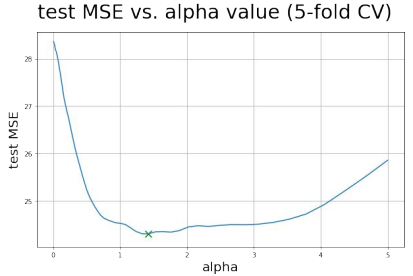
P-values confirm that all features are relevant, except for “Individualismo (palla in movimento)”, “Capacità di finalizzazione”, “Altruismo” and “Avanzamento e dribbling aggressivo/superfluo”. Lasso regression with α set equal to 0.1, 5-fold Cross-Validation (as seen with the two “paths” of *LassoCV* and *cross_val_score*), selecting an optimal alpha value of 0.04. The small difference with respect to the value 0.1 does not lead to significant improvements , then imported Ridge from *sklearn.linear_model*, setting alpha equal to 0.1

Multiple Linear Regression

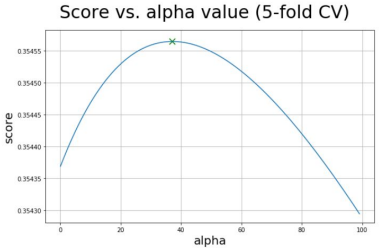
```
=========== SUMMARY ===========
Residuals:
     Min       1Q   Median       3Q      Max
 64.1561  71.2371  72.9911  75.5821  91.8181


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                      73.715091    0.124276 593.1581  0.000000
Incisività offensiva                             0.819583    0.081104  10.1053  0.000000
Individualismo (palla in movimento)              0.115740    0.113377   1.0208  0.307487
Capacità di finalizzazione                       0.005031    0.099058   0.0508  0.959499
Inefficacia del tiro                            -0.200485    0.099168  -2.0217  0.043384
Gioco offensivo con palla in movimento          -0.371233    0.130809  -2.8380  0.004600
Altruismo                                        0.025050    0.118496   0.2114  0.832605
Supremazia atletica / giocate aggressive        -0.483173    0.070756  -6.8287  0.000000
Pressing efficace in zona difensiva             -0.164137    0.045300  -3.6233  0.000300
Passaggio propositivo e nell'ultimo 1/3 di campo 0.942810    0.099934   9.4343  0.000000
Gioco con palla in movimento                    -1.020693    0.127258  -8.0206  0.000000
Inefficacia del passaggio                        0.455493    0.089205   5.1061  0.000000
Alta propositività nella metà campo avversaria  -0.583772    0.157378  -3.7094  0.000215
Avanzamento e dribbling aggressivo/superfluo     0.301290    0.168861   1.7842  0.074581
---
R-squared:  0.36796,    Adjusted R-squared:  0.36256
F-statistic: 68.12 on 13 features
```

Lasso Regression





Ridge Regression



PolyFit & GAM: beyond linearity

- **GAM** is used to blend additive models with linear models
- Predictor depends linearly on unknown smooth functions of some predictor variables
- Backfitting Algorithm and the boosting+bagging turn to improve in case of high dimensional settings
- Polynomial fitting as an effective way to draw through data points
- the non-linear fits can potentially make more accurate predictions for the response Y

Moving beyond linearity: polynomial fit

- importing **PolynomialFeatures** from *sklearn.preprocessing*
- we set the *degree parameter* to 3 and ran **LinearRegression** using the transformed dataset

  → $R^2_{adj} = 0.40$

- *fit_transform*, the method of **PolynomialFeatures**, includes all features up to the chosen maximum degree (3)
- we decided to modify the dataset by adding only the most correlated feature up to the third order, "*Alta propositività nella metà campo avversaria*" (***ρ***=0.48)

  → $R^2_{adj} = 0.39$

Moving beyond linearity: GAM

- importing **LinearGAM** from *pygam*
- splines order fixed to 3, the standard settings provide $\boldsymbol{\lambda}$ factors equal to 0.6 and number of splines equal to 20
- it could be necessary to optimize regarding two parameters, *n_splines* and $\boldsymbol{\lambda}$-factors vector: with *gridsearch* we were able to optimize regarding splines number by achieving *n_splines* equal to 6. Unfortunately, we did not succeed in optimizing factors because of computational resources that are not accessible to us are required.

LinearGAM

Distribution:	NormalDist	Effective DoF:	137.8209
Link Function:	IdentityLink	Log Likelihood:	-6166.4431
Number of Samples:	1535	AIC:	12610.5281
		AICc:	12638.3526
		GCV:	25.8395
		Scale:	21.7001
		Pseudo R-Squared:	0.4727

Feature Function	Lambda	Rank	EDoF	P > x	Sig. Code
s(0)	[0.6]	20	14.9	1.18e-05	***
s(1)	[0.6]	20	13.1	7.74e-01	
s(2)	[0.6]	20	11.0	8.87e-01	
s(3)	[0.6]	20	9.1	7.19e-01	
s(4)	[0.6]	20	11.2	9.73e-01	
s(5)	[0.6]	20	5.9	4.31e-01	
s(6)	[0.6]	20	12.3	4.61e-01	
s(7)	[0.6]	20	11.8	1.21e-03	**
s(8)	[0.6]	20	12.0	1.23e-06	***
s(9)	[0.6]	20	10.0	2.81e-04	***
s(10)	[0.6]	20	8.2	8.02e-01	
s(11)	[0.6]	20	9.2	1.36e-01	
s(12)	[0.6]	20	9.2	4.75e-03	**
intercept		1	0.0	1.11e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

LinearGAM

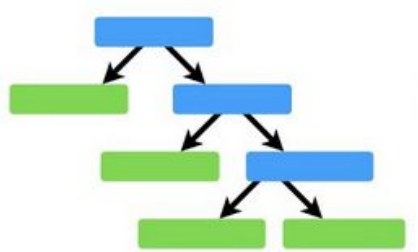
Distribution:	NormalDist	Effective DoF:	32.4466
Link Function:	IdentityLink	Log Likelihood:	-6176.5012
Number of Samples:	1535	AIC:	12419.8955
		AICc:	12421.4311
		GCV:	22.6717
		Scale:	21.8107
		Pseudo R-Squared:	0.4301

Feature Function	Lambda	Rank	EDoF	P > x	Sig. Code
s(0)	[0.6]	6	4.3	1.01e-10	***
s(1)	[0.6]	6	2.9	3.47e-01	
s(2)	[0.6]	6	2.7	2.90e-02	*
s(3)	[0.6]	6	2.4	9.59e-01	
s(4)	[0.6]	6	2.4	9.02e-01	
s(5)	[0.6]	6	1.4	7.46e-01	
s(6)	[0.6]	6	2.8	1.22e-05	***
s(7)	[0.6]	6	2.5	3.71e-06	***
s(8)	[0.6]	6	2.6	4.75e-13	***
s(9)	[0.6]	6	2.4	7.85e-08	***
s(10)	[0.6]	6	2.0	4.30e-03	**
s(11)	[0.6]	6	2.2	4.75e-04	***
s(12)	[0.6]	6	1.9	7.21e-04	***
intercept		1	0.0	1.11e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Decision tree regression

- Powerful algorithms for regression tasks
- Based on CART: high interpretability and verifiability of results



- ✚ ‘White-box’ models: high interpretability of results
- ✚ Hyperparameters tuning
- ▬ High instability and variability
- ▬ Precarious interpretations

Ensemble methods

Ensemble methods make predictions based on a number of different models and combine several weak learners into a strong learner to achieve higher flexibility

Popular ensemble methods:

- Bagging trains individual models in a parallel way and is useful to reduce trees high variance
- Boosting trains in a sequential way and each individual model learns from mistakes made by the previous one

The general idea is to combine several weak learners to obtain a more robust one

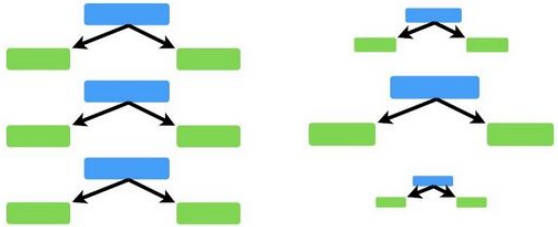
Random Forest

Model that uses bagging as an ensemble method and decision trees as individual models, it performs bootstrap on data by randomly choosing subsets for each iteration

- ➕ Overfitting reduction
- ➕ Robustness due to its resistance to noise and outliers
- ➕ Fast and simple to implement
- ➖ Heavy optimization due to the large number of parameters
- ➖ Complex optimization can lead to overfitting

AdaBoost

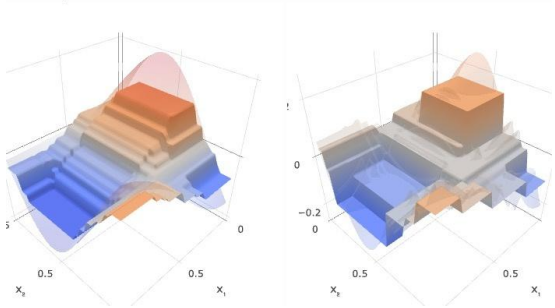
In AdaBoost (Adaptive Boosting) predictors learn from the previous made mistakes. In particular, the AdaBoost algorithm involves using very short (one-level) decision trees, stumps.



Compared to Random forest, it has fewer parameters to work on and in general it has lower performance, especially in datasets with the presence of noise

Gradient Boosting & XGBoost

Gradient Boosting tries to adapt, always in a sequential manner, the new predictor to the residual errors committed by the previous one, thus creating a prediction and the contribution of the weaker predictor to the more robust one is computed using a gradient descent optimization process.



This method is very expensive and therefore a faster and scalable version has been proposed: **XGBoost** which is very well suited to parallel and distributed systems.

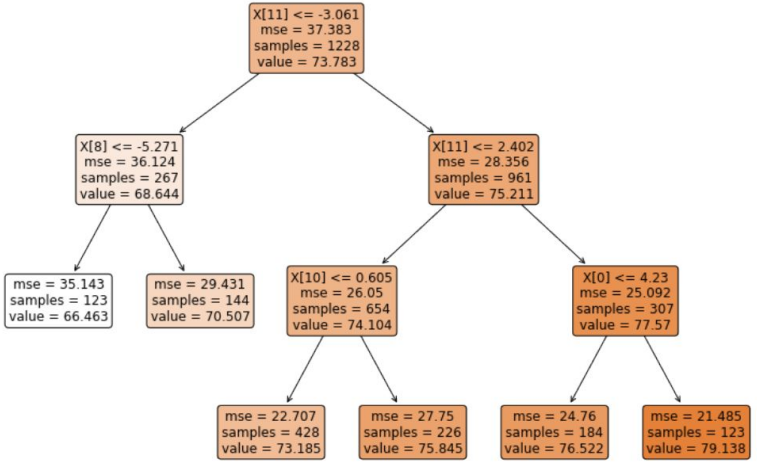
Models implementation

Decision tree regressor with grid search and cross-validation for hyperparameters tuning

Best estimator based on tuning:

- `max_depth = 3`
- `max_features = 4`
- `min_sample_split = 10%`
- `min _sample_leaf = 10%`

$R^2 = 0.23$



Models implementation - results

Analysis was performed with all the methods described above, obtaining the best result with gradient boosting. For each of the methods, Grid search with cross validation was carried out for parameter tuning.

OVERALL INDEX						
	Regression tree	**Bagging**	**Random forest**	**AdaBoost**	**Gradient Boosting**	**XGBoost**
MSE	29.08	26.96	24.58	26.4	24.32	24.88
RMSE	5.39	5.19	4.96	5.14	4.93	4.99
MAE	4.32	4.13	3.94	4.13	3.88	3.94
MAPE	5.94	5.66	5.42	5.66	5.36	5.43
Accuracy(%)	94.06	94.34	94.58	94.34	94.64	94.57
R^2	0.23	0.29	0.35	0.30	0.36	0.34

Important: without grid search, models tend to overfit

Overall Index decomposition

Decomposition of OVR considering the hexagon of skills to reach better results and adapting hexagon to a rhombus, based on the features subdivision.
The correspondence between couples of skills are:

- attack - shooting
- defense - defending
- passing types - passing
- possession - dribbling

	Incisività offensiva	Individualismo (palla in movimento)	Capacità di finalizzazione	Inefficacia del tiro	Gioco offensivo con palla in movimento	Altruismo	shooting
Incisività offensiva	1.00	-0.00079	-0.014	0.0085	-0.0019	0.00094	0.56
Individualismo (palla in movimento)	-0.00079	1.00	0.0072	0.0017	-0.00034	0.012	0.054
Capacità di finalizzazione	-0.014	0.0072	1.00	-0.098	-0.034	0.035	0.23
Inefficacia del tiro	0.0085	0.0017	-0.098	1.00	-0.022	0.034	0.17
Gioco offensivo con palla in movimento	-0.0019	-0.00034	-0.034	-0.022	1.00	0.014	-0.0023
Altruismo	0.00094	0.012	0.035	0.034	0.014	1.00	-0.080
shooting	0.56	0.054	0.23	0.17	-0.0023	-0.080	1.00

	Supremazia atletica / giocate aggressive	Pressing efficace in zona difensiva	defending
Supremazia atletica / giocate aggressive	1.00	-0.0090	0.33
Pressing efficace in zona difensiva	-0.0090	1.00	0.1
defending	0.33	0.1	1.00

	Passaggio propositivo e nell'ultimo 1/3 di campo	Gioco con palla in movimento	Inefficacia del passaggio	passing
Passaggio propositivo e nell'ultimo 1/3 di campo	1.00	0.0098	-0.0078	0.33
Gioco con palla in movimento	0.0098	1.00	0.014	0.35
Inefficacia del passaggio	-0.0078	0.014	1.00	0.096
passing	0.33	0.35	0.096	1.00

	Alta propositività nella metà campo avversaria	Avanzamento e dribbling aggressivo/superfluo	dribbling
Alta propositività nella metà campo avversaria	1.00	0.0053	0.33
Avanzamento e dribbling aggressivo/superfluo	0.0053	1.00	0.48
dribbling	0.33	0.48	1.00

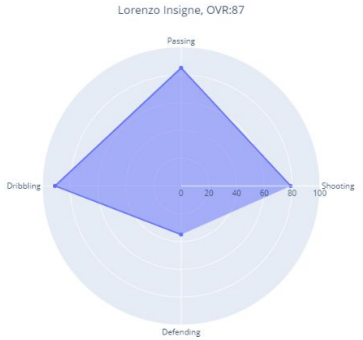
OVR decomposition - results

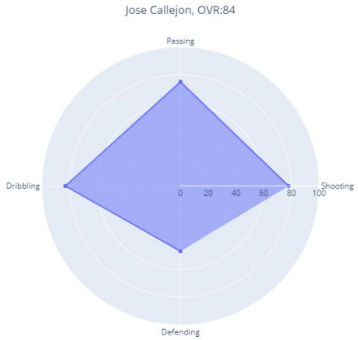
In terms of attack, passing and possession, were achieved slight improvements and the introduction of non-linear methods led to an improvement of a tenth of a point in R2, while for the defense the result was completely disappointing.

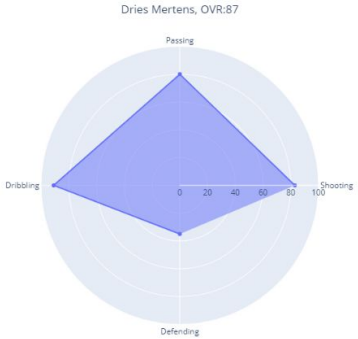
	Attack	**Defense**	**Passing_types**	**Possession**
Multiple Linear Regression	$R^2_{adj} = 0.42$	$R^2_{adj} = 0.12$	$R^2_{adj} = 0.24$	$R^2_{adj} = 0.34$
Lasso, 5-fold CV	$\alpha = 0.024$ $R^2_{adj} = 0.42$	$\alpha = 0.34$ $R^2_{adj} = 0.12$	$\alpha = 0.05$ $R^2_{adj} = 0.24$	$\alpha = 0.02$ $R^2_{adj} = 0.34$
Ridge, 5-fold CV	$\alpha = 67$ $R^2_{adj} = 0.42$	$\alpha = 140$ $R^2_{adj} = 0.12$	$\alpha = 135$ $R^2_{adj} = 0.24$	$\alpha = 18$ $R^2_{adj} = 0.34$
Polynomial Regression, up to 3rd order	$R^2_{adj} = 0.52$	$R^2_{adj} = 0.13$	$R^2_{adj} = 0.38$	$R^2_{adj} = 0.47$
GAM	$n\_splines = 10$ $R^2_{adj} = 0.52$	$n\_splines = 10$ $R^2_{adj} = 0.14$	$n\_splines = 11$ $R^2_{adj} = 0.37$	$n\_splines = 12$ $R^2_{adj} = 0.47$

OVR decomposition - results

ATTACK						
	Regression tree	**Bagging**	**Random Forest**	**AdaBoost**	**Gradient Boosting**	**XGBoost**
MSE	143.76	127.21	110.05	121.9	109.21	108.29
RMSE	11.99	11.28	10.49	11.04	10.45	10.41
MAE	9.69	8.85	8.30	9.98	8.19	8.23
MAPE	20.41	18.81	17.61	18.76	17.56	17.34
Accuracy (%)	79.59	81.19	82.39	81.24	82.44	82.66
R^2	0.40	0.47	0.54	0.49	0.55	0.55
DEFENSE						
	Regression tree	**Bagging**	**Random Forest**	**AdaBoost**	**Gradient Boosting**	**XGBoost**
MSE	304.62	303.39	307.7	307.23	305.48	302.22
RMSE	17.45	17.42	17.54	17.53	17.48	17.38
MAE	14.83	14.72	14.79	14.73	14.74	14.55
MAPE	33.75	33.73	33.88	33.9	33.79	33.49
Accuracy (%)	66.25	66.27	66.12	66.1	66.21	66.51
R^2	0.11	0.11	0.10	0.10	0.10	0.11
PASSING_TYPES						
	Regression tree	**Bagging**	**Random Forest**	**AdaBoost**	**Gradient Boosting**	**XGBoost**
MSE	104.33	89.13	80.86	94.91	80.54	79
RMSE	10.21	9.44	8.99	9.74	8.97	8.89
MAE	8.24	7.52	7.22	8.04	7.29	7.12
MAPE	14.00	12.87	12.29	13.59	12.4	12.12
Accuracy (%)	86	87.13	87.71	86.41	87.6	87.88
R^2	0.22	0.34	0.40	0.29	0.40	0.41
POSSESSION						
	Regression tree	**Bagging**	**Random Forest**	**AdaBoost**	**Gradient Boosting**	**XGBoost**
MSE	75.95	55.51	52.3	60.94	52.6	53.44
RMSE	8.71	7.45	7.23	7.81	7.25	7.31
MAE	6.78	5.78	5.65	6.09	5.64	5.72
MAPE	10.69	9.02	8.77	9.49	8.84	8.88
Accuracy (%)	89.31	90.98	91.23	90.51	91.16	91.12
R^2	0.26	0.46	0.49	0.40	0.49	0.48







Skills Rhombus

In summary...

- The purpose of this project was to study how the performance of Serie A teams and players are rated by the best known indices:
  - SPI, Soccer Power Index, to assess the performance of teams
  - FIFA Overall Index for player skills
- The large amount of features led the study towards a PCA (R^2 = 0.84 just with a simple linear regression)
- Critical issues found with OVR which led to the introduction of non-linear method:
  - polynomial regression (up to the third order)
  - GAM (based on third order splines)
  - regression trees and ensemble methods
- Decomposition of the Overall index into four subindices

Conclusions

Football is a sport where collective and individual performances are closely interconnected and very difficult to untangle.

A good evaluation system should be able to make clear some connection between the weight of the individual player and the overall performance of the team.

We believe that the development of these indices will become increasingly important as more data are made free and open, to allow the study of more complex interactions between features to explain numerical relationships that are currently infeasible.