Football analysis

Italo Alberto Ferrante, Antonio Laudante,
Antonio Nappa, Celestino Santagata

November 17, 2020

Contents

Introduction

Over the past two decades, the influence of data analysis has grown in every area of our lives: in companies of all kinds, but also in healthcare, the media and sports. Until a few years ago, football was deemed immune to this trend.
Now, those who first adopted data analytics in the main European championships began to benefit from an important competitive advantage: Liverpool, AZ Alkmaar and Brentford are just some of the success stories. Indeed, nowadays there is a lot of research around football, we are surrounded by numbers, but there are few we value, those that give us a significant parameter, they are called **advanced metrics** since the first time that Bill James employed some in baseball analysis.
Expected goals and expected assists have become some of the most used and understood in the world of football:

- *Expected goals* (xG) measures the quality of a shot based on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance. Adding up a player or team's expected goals can give us an indication of how many goals a player or team should have scored on average, given the shots they have taken.

- *Expected assists* (xA) measures the likelihood that a given pass will become a goal assist. It considers several factors including the type of pass, pass end-point and length of pass. Adding up a player or team's expected assists gives us an indication of how many assists a player of team should have had based on their build up and attacking play.

How much do statistics matter in determining success in football?
It is very important to study the connection between sports, in this case football, and the use of advanced statistical analysis: how it has developed and how it can be used as a tool for both coaches and team managers to improve their teams, selection process, player development, etc. As we said before, football was not the pioneer in this area, actually came into it quite late, but rather Baseball who has preoccupied the imagination of statisticians and mathematicians for decades and with the huge success of **Sabermetrics** approach to building teams

in baseball (pioneered in Oakland Athletics, later adopted by majority of baseball franchises). Teams have had great success with the approach of using advanced statistical analysis to build teams in Baseball that suddenly teams with much lower budgets started performing at the high level (baseball has no salary cap, like all other professional sports in the USA).
"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). Statistics are a good way to tell how a player is doing in a sport and what he is best at, they can also help to determine where the player and his team need to improve; but it is equally important to learn how to make conscious use of it, not to let oneself be guided only by numbers.

Chapter 1

Dataset creation

The creation of a dataset suitable for our analysis required a considerable effort, as many resources are not accessible except through payment or subscription. Many sites try to sell their evaluation systems by providing aggregated data and preventing a transparent study of the indices used, for example OPTAsports and Wyscout. In this context we have identified several sites, among which only FBref.com has proved suitable for our objectives.

1.1 Sports Reference

Sports reference, LLC is an American company which operates several sports-related websites, including Sports-Reference.com, Baseball-Reference.com for baseball, Basketball-Reference.com for basketball, Hockey-Reference.com for ice hockey, Pro-Football-Reference.com for American football, and FBref.com for association football (soccer). Between 2008 and 2020, Sports Reference also provided pages for Olympic Games and its competitors.

FBref
FBref.com is a website devoted to tracking statistics for football teams and players from around the world, it was launched in June 2018 with league coverage for six nations: England, France, Spain, Italy, Germany, and the USA.
FBref is not yet a Baseball-Reference for football, but the long-term vision for the site includes all global football competitions stats, as well as transfers, individual results, future fixtures, a suite of analytical tools, and much more, all while ensuring FBref is as fast and easy-to-use as the other sites of the network.

On FBref data are available both by team and by player (through the link *"View Player Stats"*), and are divided into 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 (fig.1.1). Our discussion excluded goalkeepers, so we did not consider the tables relating to this position; moreover, we grouped the tables according to four characteristics: *attack*, *defense*, *passing_types*, *possession*.
Records are collected for seasons 2017-2018, 2018-2019 and 2019-2020 (SerieA_stats) and stored in csv files. *Attack* file contains features from tables *Squad Standard Stats*, *Squad Shooting* and *Squad Goal and Shot Creation*; tables *Squad Defensive Actions* and *Squad Miscellaneous Stats* have been included in *Defense*, *Squad Passing* and *Squad Pass Types* in *Passing_types* and, finally, *Squad Possession* in *Possession*[1] (fig.1.2).
Each file is then used to perform a *PCA*, Principal Component Analysis, in order to reduce space dimensionality.

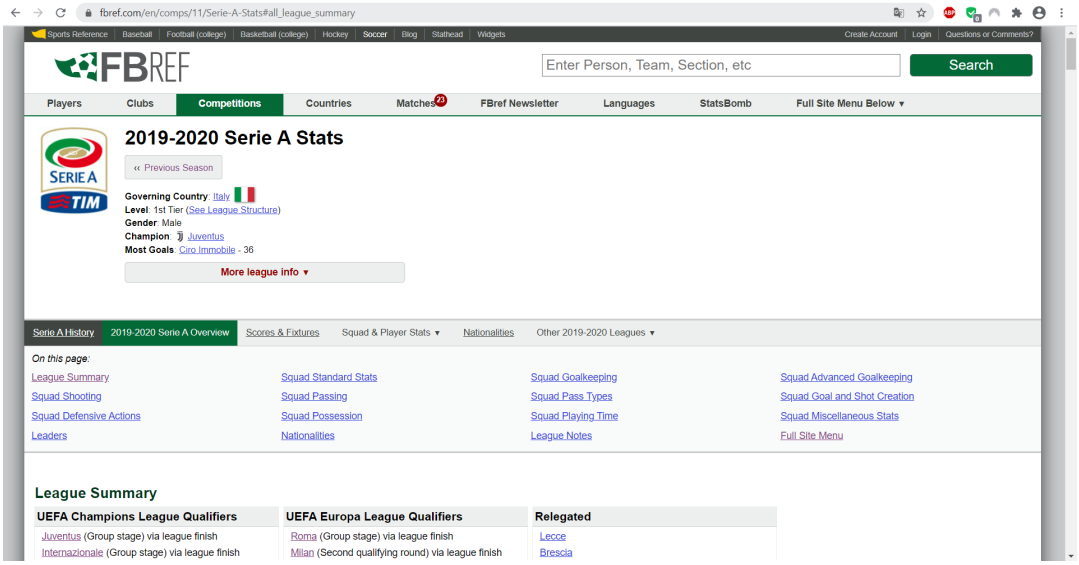


Figure 1.1: FBref site overview.

Even though we found an API for scraping FBref website, we did not use it because of two main problems with collecting data via web scrapers: firstly, the hosts of websites do not like that information is automatically extracted and, secondly, the web scraping process itself is unstable. Website hosts do not like that information is extracted via a scraper because it means additional non-human web traffic which is costly and they fear that their data could be sold or illegitimately used after extraction. Many website have ways of detecting web scrapers and

[1]Note that not all features were included and the table *Squad Playing Time* was neglected.

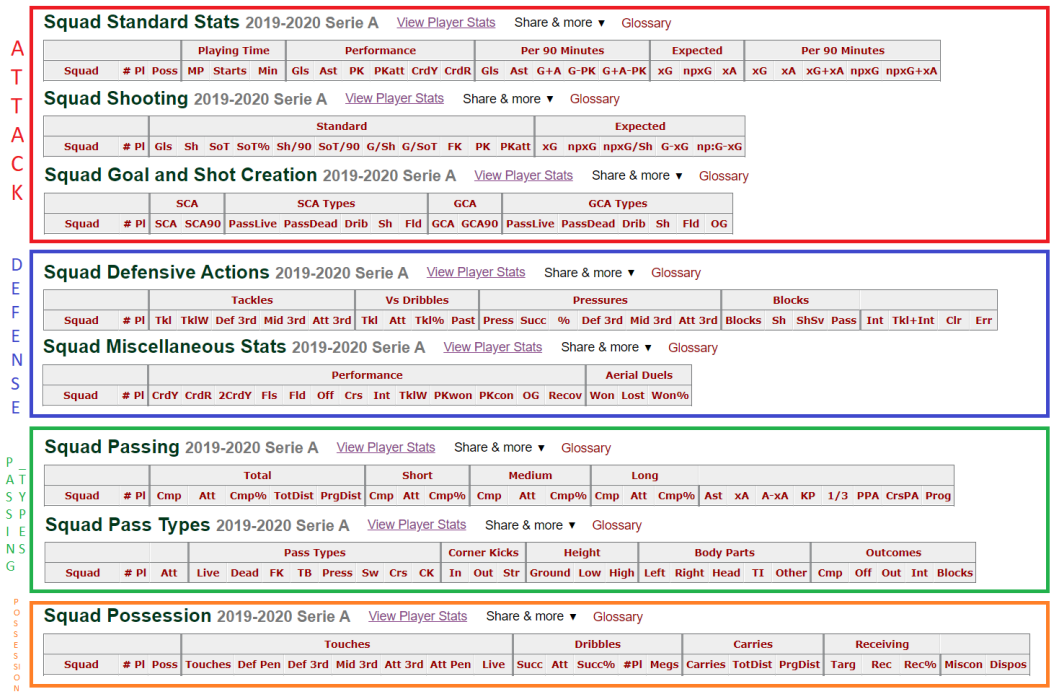


Figure 1.2: FBref tables features, subdivision in groups.

blocking their IP addresses in order to prevent them to visit the website. To circumvent being detected and blocked, random elements in the form of variable clicking times can be integrated into the scripts to mimic human behaviour. In addition, proxy servers can be used to alter the IP address from time to time to create the impression that a new user is visiting. These two measures successfully prevent the scraper of being blocked, but web scraping is inherently unstable because of varying Internet connection speed (with occasional blackouts), other programs that interact with the web browser and/or servers that do not respond. Moreover, this API takes in the url of a webpage and css selector id of the content to be scrapped and returns the outer html text of the scrapped content, which needs to be converted into a csv file; so, we preferred to import them individually as csv files using the option *"Get table as CSV (for Excel)"*, and then we corrected encoding errors.

1.2 Features analysis

Here we report the summary tables with the main descriptive statistical indices and the box-plots, in this way we can give a quick overview of individual features distributions.

Teams

ATTACK

	season	Squad	# Pl	Gls	Ast	Gls_90	Ast_90	xG	xA	xG_90	xA_90	Off	PK	PKatt	Sh	SoT	FK	SoT%
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	50.13	32.02	1.35	0.86	50.45	33.31	1.36	0.90	71.77	5.45	6.98	504.15	161.05	19.45	31.75
std			3.57	15.85	11.46	0.43	0.31	11.33	8.79	0.31	0.24	15.72	2.85	3.19	96.36	38.98	6.54	2.98
min			22	25	14	0.66	0.37	33.3	21.2	0.88	0.56	40	0	0	353	92	8	24.3
25%			25	37.75	23	0.99	0.66	40.68	26.28	1.12	0.71	61.25	4	5	426.25	131.75	14.75	29.9
50%			29	48	31.5	1.32	0.83	48.35	32.65	1.31	0.86	73.5	5	6.5	493.5	151.5	20	31.45
75%			30	57.75	38.25	1.58	1.01	60.03	38.95	1.59	1.1	83.25	7	8	584.25	190.75	23	34
max			41	94	68	2.61	1.89	82.7	53.6	2.3	1.49	99	13	16	702	250	40	37.9

	SCA	SCA90	SCA_Passl	SCA_Passl	SCA_Drib	SCA_Sh	SCA_Fld	GCA	GCA90	GCA_Passl	GCA_Passl	GCA_Drib	GCA_Sh	GCA_Fld	GCA_OG
	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60
	781.77	21.11	584.75	67.35	47.82	40.03	40.55	81.23	2.20	55.17	5.32	5.58	7.18	6.72	1.27
	170.61	4.60	147.09	13.53	13.54	10.57	9.20	27.90	0.76	22.36	2.40	3.45	2.73	2.83	1.18
	504	13.26	329	34	22	18	18	35	0.92	20	0	0	3	1	0
	648	17.71	468.75	57	36.75	31	33.75	59.5	1.57	37.25	4	3	5	4	0
	749.5	20.21	550.5	67.5	45.5	39	40	78	2.125	52.5	5	5	7	7	1
	935.25	25.11	704.75	77	57	48	47	96.75	2.61	65	7	7	9	8.25	2
	1145	30.13	951	103	80	65	65	150	4.17	112	11	14	16	13	4

DEFENSE

	season	Squad	# Pl	Tkl	TklW	Tkl_Def3rd	Tkl_Mid3rd	Tkl_Att3rd	Tkl.1	Att	Tkl%	Past	Press	Succ	%	Press_Def3	Press_Mid	Press_Att3
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	627.68	399.42	307.40	239.15	81.13	213.17	622.65	34.20	409.48	5802.03	1582.25	27.34	1947.80	2558.40	1295.83
std			3.57	62.46	40.19	41.98	32.27	15.32	29.82	61.04	2.78	41.21	503.64	143.85	2.04	328.00	264.67	218.67
min			22	502	312	216	181	51	161	486	27.9	325	4704	1261	22.9	1191	2043	912
25%			25	582.75	371.75	283.25	214.75	69.75	192.75	574.75	31.95	373.75	5504.75	1481	26.1	1764.25	2354.25	1141.5
50%			29	616	389	306	237.5	81.5	210	630.5	34.3	419	5790.5	1565.5	27.45	1898	2551.5	1268.5
75%			30	673	426.25	337.5	262	89	232.5	672	35.63	438.5	6158.5	1687.75	28.5	2147	2758	1404.5
max			41	782	496	379	325	118	286	726	41.5	493	7284	1985	32.2	2930	3369	2088

	Sh	ShSv	Pass	Int	Clr	Err	Fls	Fld	PKcon	CrdY	CrdR	OG	Recov
	60	60	60	60	60	60	60	60	60	60	60	60	60
	137.05	2.28	448.67	411.07	820.88	9.37	492.35	472.75	6.43	82.95	4.58	1.58	3288.47
	26.26	1.62	44.42	67.61	142.59	3.20	54.53	55.67	2.71	14.55	1.95	1.29	354.45
	77	0	372	314	471	3	348	359	1	45	1	0	2666
	120	1	415	359	731.5	7	463	443	4	72	3	1	2956
	132.5	2	437.5	393.5	846.5	9.5	496.5	474.5	6.5	85.5	4.5	1	3283.5
	157.5	3	473.5	458	912.5	12	524.25	500	8	94	6	2	3582.25
	191	7	559	589	1132	19	623	653	13	112	9	5	3940

Figure 1.3: Teams tables summaries (1).

Box-Plots allow us to quickly compare the distributions, but first we need to standardize the variables.
It is quite evident in fig.1.7 that the most marked differences among SerieA teams emerge in *Passing_types* features, indeed here almost every feature has a strongly skewed distribution.

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

POSSESSION

	season	Squad	# Pl	Poss	Touches	DefPen	Tch_Def3r	Tch_Mid3r	Tch_Att3r	AttPen	Live	Succ	Att	Succ%	#Npdribbl	Megs	Carries	TotDist
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	50.01	22614.53	2387.13	7226.68	10941.45	5929.23	854.33	20858.32	379.83	623.87	60.88	409.63	20.65	14942.37	87637.57
std			3.57	6.00	3070.84	229.42	508.34	1996.64	1230.47	183.03	3105.65	59.80	93.09	3.10	63.85	6.70	2934.83	14112.04
min			22	37.7	17431	1901	6043	7447	3864	552	15569	261	449	52.2	284	7	10116	63840
25%			25	45.05	20094.5	2218.5	6936.5	9355.75	4868.5	716.25	18365.75	345.25	551.75	59.075	369	15.75	12409.75	76805
50%			29	50.4	22000.5	2386	7208	10480	5765.5	827	20229.5	380.5	630.5	61.6	410	20	14474	87539.5
75%			30	54.88	24773.25	2535.5	7545.75	12389.5	6631	996	23091	426.25	694.5	62.9	456.25	25.25	17154.75	99766.25
max			41	63.2	32977	3000	8464	18407	9349	1275	31365	492	812	67	523	36	23717	116757

	Rec	Rec%	Miscon	Dispos
	60	60	60	60
	14662.35	85.57	437.30	414.88
	3142.88	2.81	51.89	54.56
	9975	79.5	345	307
	12238.5	83.7	398.75	376.75
	14243.5	85.5	431	412.5
	16895.75	87.55	471.25	451.25
	24985	90.5	558	554

Figure 1.4: Teams tables summaries (2).

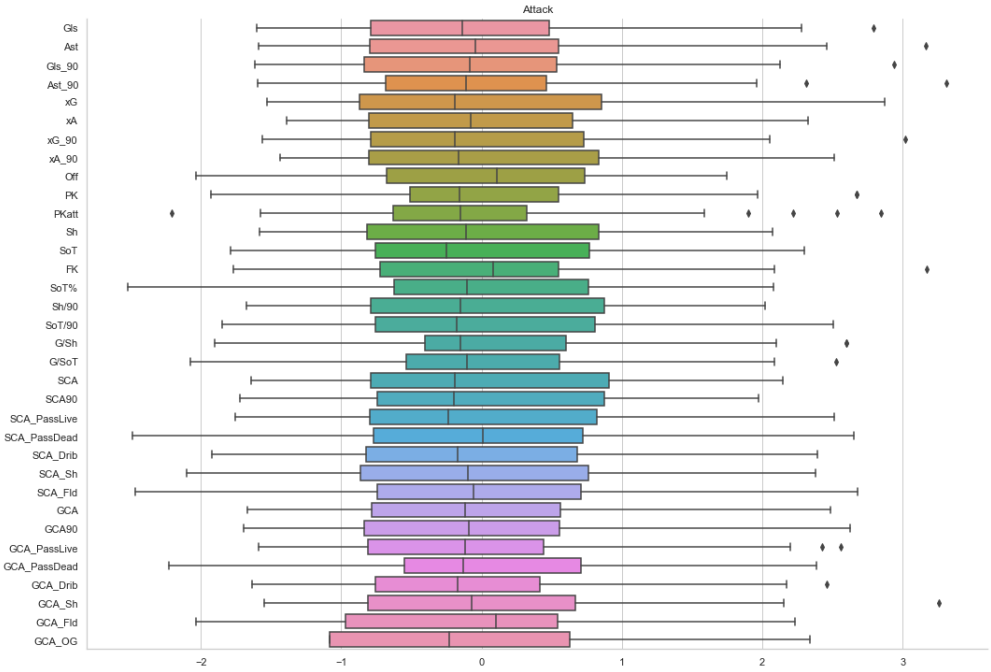


Figure 1.5: Teams Attack features Box-plots.

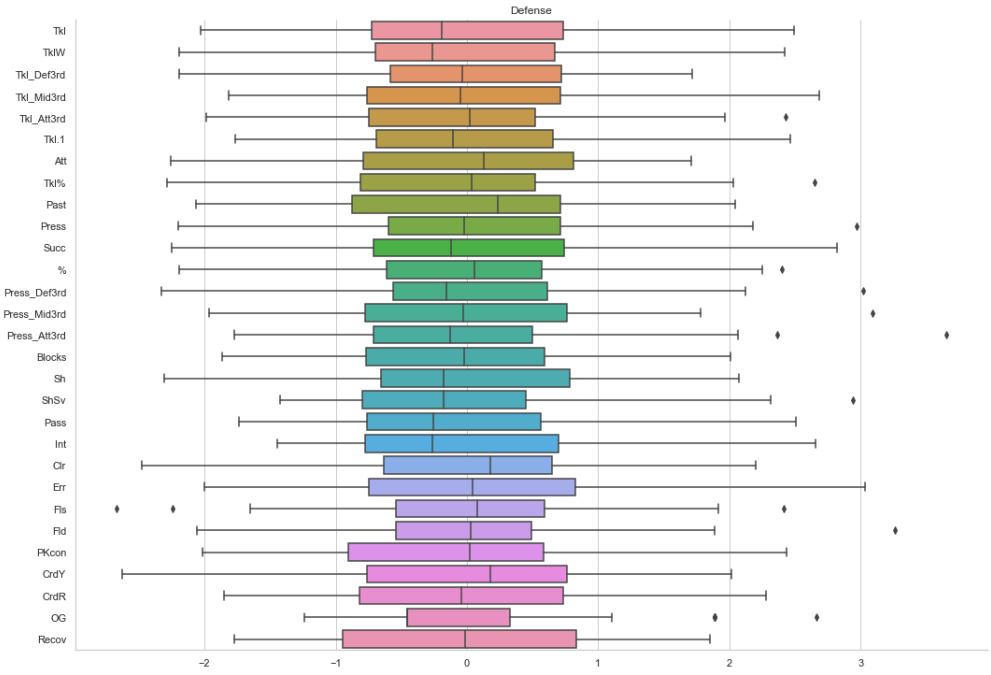


Figure 1.6: Teams Defense features Box-plots.

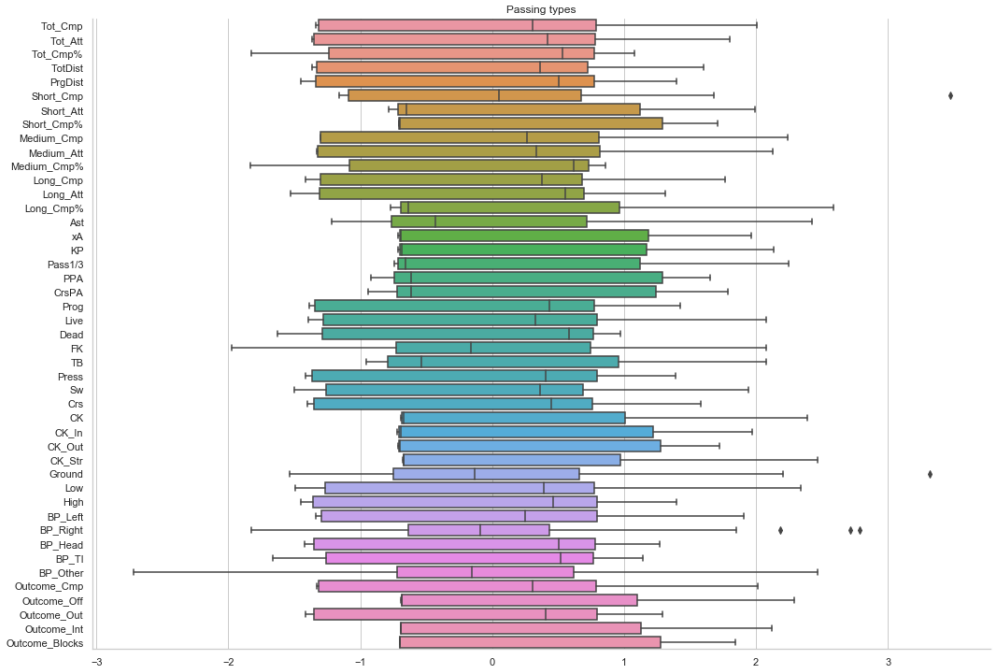


Figure 1.7: Teams Passing Types features Box-plots.

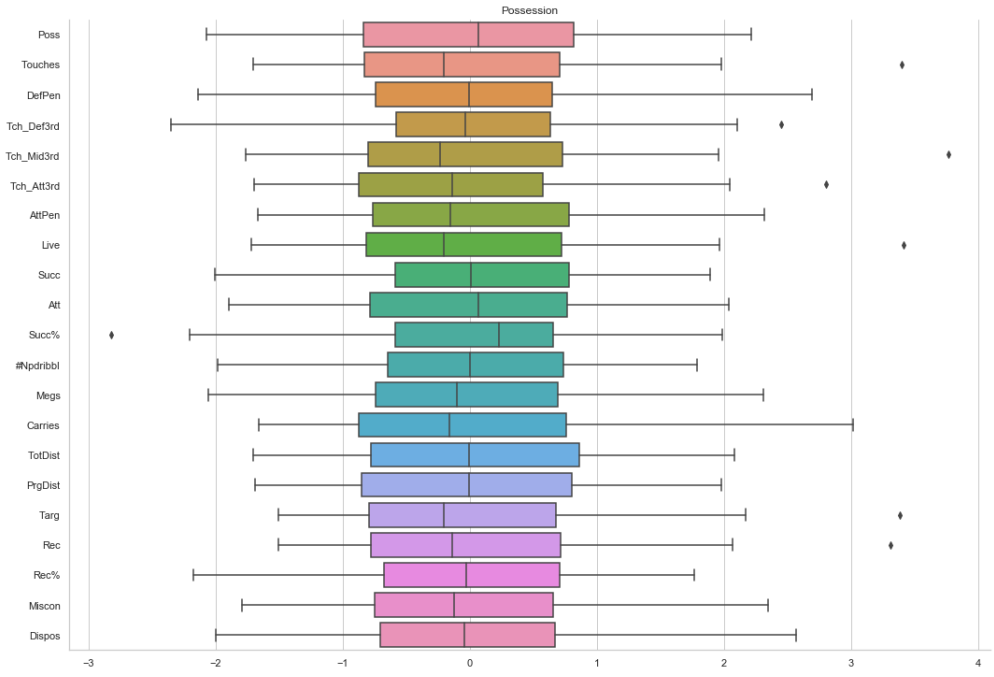


Figure 1.8: Teams Possession features Box-plots.

Players

ATTACK

	season	Player	Nation	Pos	Squad	Age	90s	MP	Starts	Min	Gls	Ast	Gls_90	Ast_90	xG	xA	xG_90	xA_90	Off	PK	PKatt	Sh
count	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
unique	3	852	74	10	25																	
top	2019/20	Alessandro	it ITA	DF	Genoa																	
freq	538	5	647	536	91																	
mean						25.44	14.22	18.60	14.26	1279.63	1.93	1.23	0.11	0.08	1.97	1.28	0.14	0.09	2.75	0.21	0.27	19.38
std						4.69	10.68	11.13	11.17	961.18	3.50	1.92	0.20	0.14	3.14	1.72	0.19	0.14	5.48	0.95	1.12	24.89
min						0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
25%						22	4.4	9	4	399	0	0	0	0	0.2	0.1	0.03	0.02	0	0	0	3
50%						25	12.7	19	12	1141	1	0	0.03	0	0.8	0.6	0.07	0.06	1	0	0	11
75%						29	23.2	29	24	2087	2	2	0.16	0.11	2.4	1.8	0.19	0.12	3	0	0	25
max						39	38	38	38	3420	34	16	3.33	2.9	26.7	12.4	2.54	3.42	44	14	15	186

	SoT	FK	SoT%	Sh/90	SoT/90	G/Sh	G/SoT	SCA	SCA90	SCA_Passl	SCA_Passl	SCA_Drib	SCA_Sh	SCA_Fld	GCA	GCA90	GCA_Passl	GCA_Passl	GCA_Drib	GCA_Sh	GCA_Fld	GCA_OG
	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
	6.19	0.75	25.83	1.43	0.41	0.07	0.20	29.98	2.08	22.43	2.55	1.84	1.54	1.56	3.12	0.19	2.11	0.20	0.22	0.28	0.26	0.05
	9.00	2.46	20.85	1.69	0.53	0.11	0.26	33.30	1.97	24.16	6.16	3.23	2.40	2.44	4.20	0.25	2.91	0.70	0.63	0.64	0.64	0.22
	0	0	0	0	0	-0.08	-0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	6.9	0.43	0.04	0	0	5	0.79	4	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	26.7	0.97	0.23	0.03	0.09	19	1.84	14	0	0	1	1	2	0.13	1	0	0	0	0	0
	8	0	36.85	2.1	0.61	0.1	0.33	43	2.87	33	2	2	2	2	4	0.29	3	0	0	0	0	0
	65	27	100	30	6.92	1	1	208	33.75	168	63	28	25	20	29	3.21	22	8	7	6	5	2

DEFENSE

	season	Player	Nation	Pos	Squad	Age	90s	MP	Starts	Min	Tkl	TklW	Tkl_Def3rc	Tkl_Mid3rc	Tkl_Att3rd	Tkl.1	Att	Tkl%	Past	Press	Succ	%
count	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
unique	3	852	74	10	25																	
top	2019/20	Alessandro	it ITA	DF	Genoa																	
freq	538	5	647	536	91																	
mean						25.44	672.61	17.19	15.66	850.13	450.26	18.23	13.06	10.15	5.15	6.28	18.70	28.29	21.17	149.89	118.52	38.80
std						4.69	932.53	11.13	11.43	972.81	831.12	18.29	13.35	10.69	6.83	7.25	20.12	21.76	19.15	174.56	145.81	33.97
min						0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25%						22	8.3	7	5	20	11	4	2	2	1	1	3	11	5	14	25	23.5
50%						25	22.9	16.4	15	412	33	13	9	7	3	4	12	27	17	71	69	28.9
75%						29	1988.5	27	26	1576	352.5	27	20	15	7	9	27	42.1	33.3	247	140	37.5
max						39	2001	38	38	3420	3420	124	92	66	61	47	133	129	100	925	939	264

	Press_Def	Press_Mid	Press_Att3	Blocks	Sh	ShSv	Pass	Int	Clr	Err	Fls	Fld	PKcon	CrdY	CrdR	OG	Recov
	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
	57.05	89.89	67.07	32.22	11.02	1.85	11.53	16.39	25.96	10.80	12.43	18.35	6.31	2.16	1.22	0.09	79.17
	58.64	83.13	78.04	41.91	15.43	4.98	14.88	16.42	36.26	28.84	15.09	17.14	13.97	2.78	2.27	0.31	101.71
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	21	22.5	10	6	1	0	0	3	3	0	0	4	0	0	0	0	0
	30.3	69	39	19	4	0	4	11	13	0	6	14	0	1	0	0	33
	81	135	100	41.5	15	1	19	25	32	3	21	27	4	4	1	0	131.5
	397	489	579	361	96	44	75	91	266	243	76	101	111	15	15	2	520

Figure 1.9: Players tables summaries (1).

As we can see in figs.1.11-1.14, there is a significant presence of outliers: just as the attackers excel in offensive characteristics, so the defenders excel in the defensive ones. The same reasoning applies to other fields and to other positions, and is perfectly in line with the analysis.

PASSING_TYPES

	season	Player	Nation	Pos	Squad	Age	90s	MP	Starts	Min	Tot_Cmp	Tot_Att	Tot_Cmp%	TotDist	PrgDist	Short_Cmp	Short_Att	Short_Cmp	Medium_C	Medium_A	Medium_C
count	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
unique	3	852	74	10	25																
top	2019/20	Alessandro	it ITA	DF	Genoa																
freq	538	5	647	536	91																
mean						25.44	14.22	18.60	14.26	1279.63	533.32	662.04	77.94	10063.48	3137.96	8.44	22.84	34.73	407.61	466.72	84.39
std						4.69	10.68	11.13	11.17	961.18	488.15	583.24	11.77	9627.00	3248.12	8.42	20.51	21.76	371.67	413.10	12.42
min						0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
25%						22	4.4	9	4	399	129	168.5	73.4	2209.5	527.5	2	6	22.2	105	124.5	81.7
50%						25	12.7	19	12	1141	392	503	79.5	6920	1961	6	18	34.4	305	362	86.6
75%						29	23.2	29	24	2087	840	1043.5	84.5	15818.5	4972.5	13	35	47.5	626	718.5	90.6
max						39	38	38	38	3420	2864	3229	100	51015	19423	55	104	100	2479	2682	100

	Long_Cmp	Long_Att	Long_Cmp	Ast	xA	KP	Pass1/3	PPA	CrsPA	Prog	Live	Dead	FK	TB	Press	Sw	Crs	CK	CK_In	CK_Out	CK_Str
	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
	117.26	172.49	63.41	1.23	1.28	14.21	42.16	12.31	3.48	62.62	607.80	54.24	15.62	1.17	105.59	20.49	28.12	7.46	1.73	1.67	0.52
	126.73	173.44	19.37	1.92	1.72	17.10	46.84	15.33	5.70	64.39	536.89	70.86	21.79	2.10	89.19	23.78	43.39	20.22	5.81	5.89	1.82
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	20	32	54.35	0	0.1	2	7.5	2	0	12	155	8	1	0	28	3	2	0	0	0	0
	73	113	65.9	0	0.6	8	27	7	1	41	464	26	6	0	87	13	10	0	0	0	0
	177	273.5	75.25	2	1.8	21	63	18	4	95	969.5	70	22	2	165	29	34	3	0	0	0
	831	971	100	16	12.4	119	442	117	49	347	3153	473	179	17	473	161	327	190	69	73	23

	Ground	Low	High	BP_Left	BP_Right	BP_Head	BP_TI	BP_Other	Outcome_	Outcome_	Outcome_	Outcome_	Outcome_
	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
	440.21	96.73	125.10	197.71	392.87	27.53	27.81	3.90	533.32	2.18	11.61	10.46	17.17
	407.07	87.52	121.76	298.58	440.17	26.78	56.96	5.01	488.15	2.65	10.94	10.19	16.98
	0	0	0	0	0	0	0	0	0	0	0	0	0
	109.5	25	27	29	55.5	6	1	1	129	0	3	2	4
	323	75	87	87	209	19	4	2	392	1	8	7	12
	680.5	147.5	193	217	602.5	43	18	6	840	3	18	16	25.5
	2479	594	637	2026	2900	163	366	62	2864	21	73	62	107

POSSESSION

	season	Player	Nation	Pos	Squad	Age	90s	MP	Starts	Min	Touches	DefPen	Tch_Def3r	Tch_Mid3r	Tch_Att3r	AttPen	Live	Succ	Att	Succ%
count	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
unique	3	852	74	10	25															
top	2019/20	Alessandro	it ITA	DF	Genoa															
freq	538	5	647	536	91															
mean						25.44	14.22	18.60	14.26	1279.63	821.82	49.17	230.27	420.50	228.04	32.88	769.43	14.58	23.94	56.51
std						4.69	10.68	11.13	11.17	961.18	686.86	66.35	259.61	375.18	244.45	43.88	642.42	17.38	28.61	27.34
min						0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
25%						22	4.4	9	4	399	224	6	36	108.5	38	5	211	2	4	48.4
50%						25	12.7	19	12	1141	660	24	128	328	143	17	617	8	13	60
75%						29	23.2	29	24	2087	1309.5	64	344.5	650.5	354	44	1230	20	34	72.4
max						39	38	38	38	3420	3466	469	1428	2395	1673	320	3393	140	221	100

	#Npdribbl	Megs	Carries	TotDist	PrgDist	Targ	Rec	Rec%	Miscon	Dispos
	1571	1571	1571	1571	1571	1571	1571	1571	1571	1571
	15.72	0.79	552.73	3267.86	1812.12	637.86	546.19	83.69	16.79	15.95
	18.65	1.42	475.22	2896.57	1707.90	528.02	470.49	16.45	19.92	17.83
	0	0	0	0	0	0	0	0	0	0
	3	0	154	827	422.5	183	153.5	77.25	3	3
	9	0	440	2541	1372	528	438	88.5	9	9
	22	1	864	5074	2738	980	855.5	95.8	25	24
	151	11	2546	15631	10101	3096	2833	100	138	105

Figure 1.10: Players tables summaries (2).

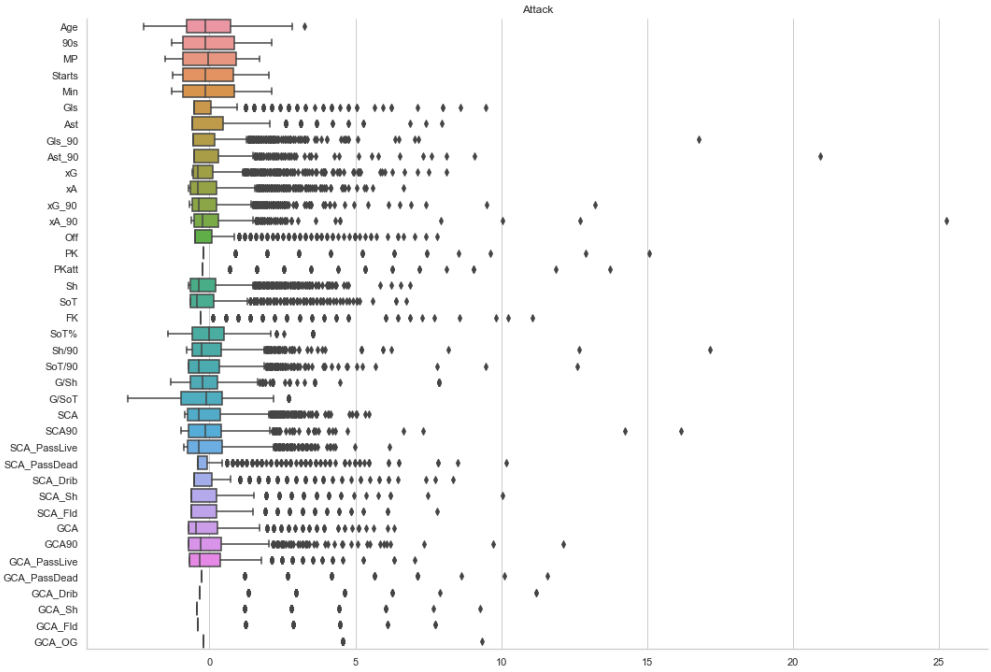


Figure 1.11: Players Attack features Box-plots.

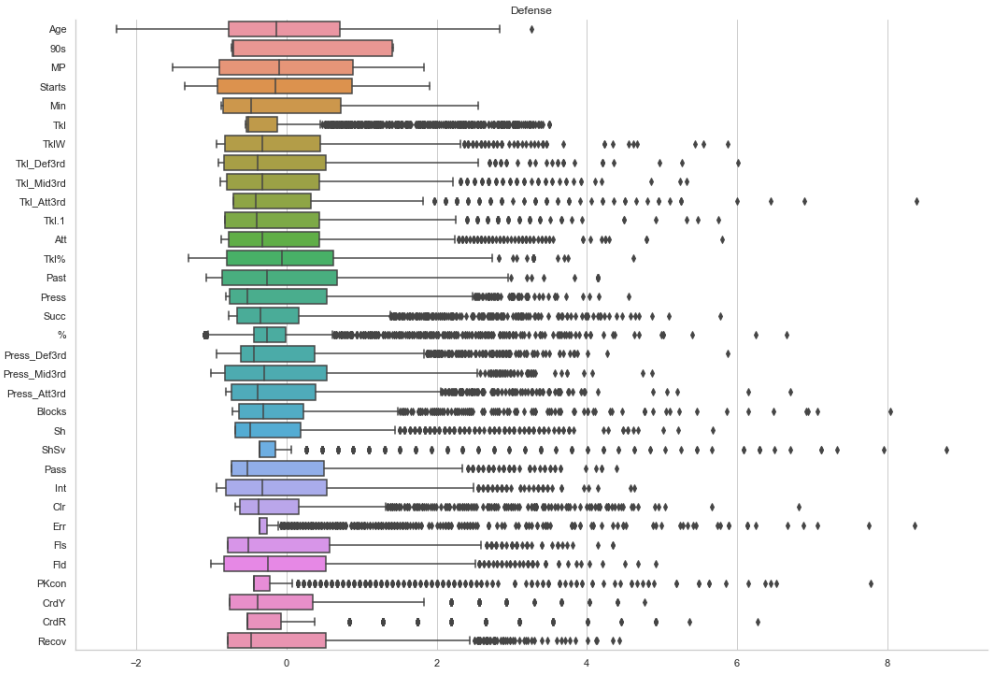


Figure 1.12: Players Defense features Box-plots.

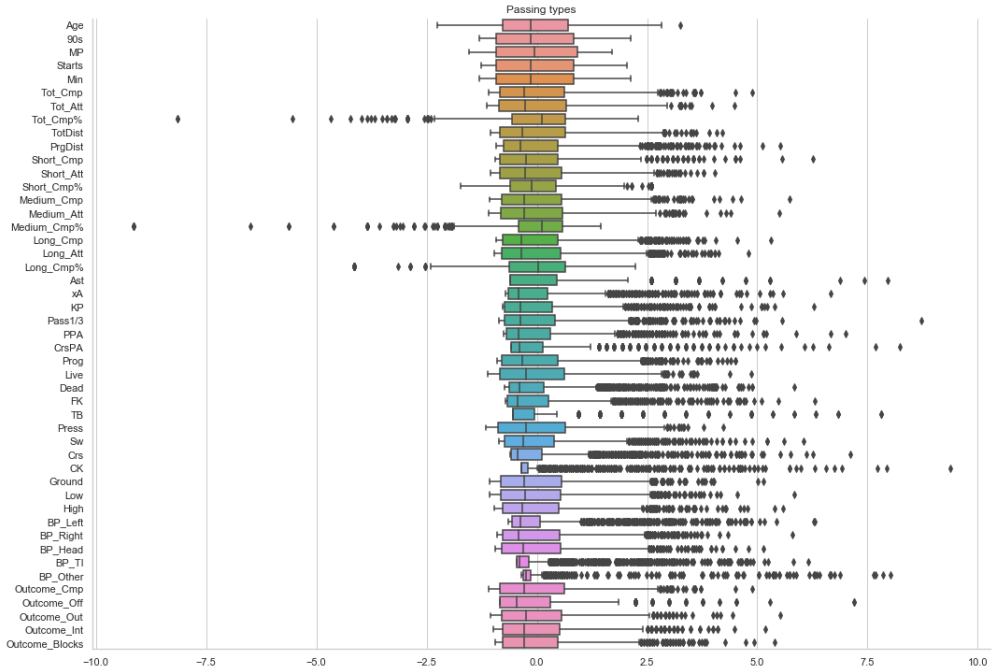


Figure 1.13: Players Passing Types features Box-plots.

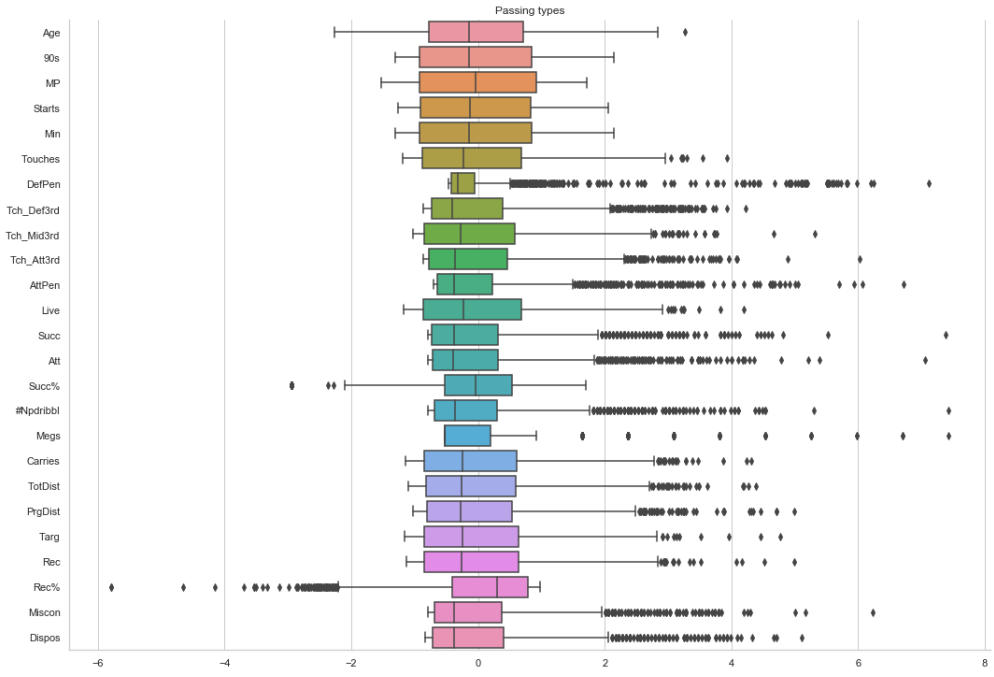


Figure 1.14: Players Possession features Box-plots.

1.3 Other resources

Football datasets:

- European Football Data Analysis (link-to-kaggle)
- Datahub Football data (link): football datasets by league, World Cup, Stadium Datasets, Football Apps, etc.
- Football Data: Expected Goals and Other Metrics (link-to-kaggle, source: understat.com)
- CIES Football Observatory and InStat Performance Index (link)
- Serie A Player Statistics: whoscored.com rating system

Interesting projects:

- Football Player's Performance and Market Value (pdf)
- Is Football players' performance influenced by the quality of opposition? Application of the Golden Index formula in Club Atlético de Madrid 2016/2017 (link)

PDF reading and parsing tools useful for websites reports:

- Excalibur (link, github)
- Apache Tika (tika-python)
- pdfreader (documentation)
- PyPDF2 (documentation)

Chapter 2

Team performance - SPI

References: SerieA_SPI, how-it-works.
Dataset: SPIdataset.

In this part of our project we focus on the interpretation of *Soccer Power Index*, SPI.
This index, as used by football analysts, is the reinterpretation of ESPN's Soccer Power Index, a rating system originally devised by FiveThirtyEight editor-in-chief Nate Silver in 2009 for rating international soccer teams. SPI states that every team has an offensive rating that represents the number of goals it would be expected to score against an average team on a neutral field, and a defensive rating that represents the number of goals it would be expected to concede. These ratings, in turn, produce an overall SPI rating, which represents the percentage of available points — a win is worth 3 points, a tie worth 1 point, and a loss worth 0 points — the team would be expected to take if that match were played over and over again. It is interesting because we can, for example, project the result of a match or simulate whole seasons to arrive at the probability each team will win the league, qualify for the Champions League or be relegated to a lower division.
Before a season begins, a team's SPI ratings are based on two factors: its ratings at the end of the previous season and its market value as calculated by Transfermarkt (a site that assigns a monetary value to each player, based on what they would fetch in a transfer). Market values are used to infer each team's preseason SPI rating (fig.2.1).
In SPI, three metrics are used to evaluate a team's performance after each match: *adjusted goals*, *shot-based expected goals* and *non-shot expected goals*.
The first, *adjusted goals*, accounts for the conditions under which each goal was scored. For adjusted goals, the value of goals scored is reduced when a team has more players on the field, as well as goals scored late in a match when a team is already leading. After downweighting these goals, the value of all other goals is

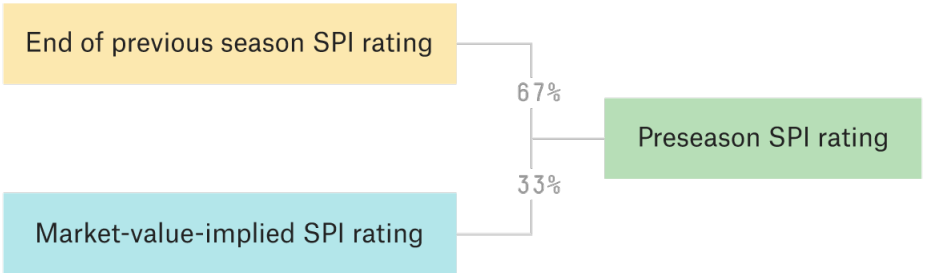


Figure 2.1: Preseason SPI rating.

increased to make the total number of adjusted goals generally add up to the total number of actual goals scored over time.
Shot-based expected goals are an estimate of how many goals a team "should" have scored, given the shots they took in that match. Each shot is assigned a probability of scoring based on its distance and angle from the goal, as well as the part of the body the shot was taken with, with an adjustment for which specific player took the shot. These individual shot probabilities are added together to produce a team's shot-based expected goals for that match, which may be bigger or smaller than the number of goals it actually scored.
Non-shot expected goals are an estimate of how many goals a team "should" have scored based on non-shooting actions they took around the opposing team's goal: passes, interceptions, take-ons and tackles. For example, we know that intercepting the ball at the opposing team's penalty spot results in a goal about 9% of the time, and a completed pass that is received at the center of the six-yard box leads to a goal about 14% of the time. We add these individual actions up across an entire match to arrive at a team's non-shot expected goals. Just as for shot-based expected goals, there is an adjustment for each action based on the success rates of the player or players taking the action (both the passer and the receiver, in the case of a pass).
Anyway, football is a tricky sport to model because there are so few goals scored in each match. The final scoreline will fairly often disagree with many people's impressions of the quality of each team's play, and the low-scoring nature of the sport will sometimes lead to prolonged periods of luck, where a team may be getting good results despite playing poorly (or vice versa).
To get a better idea of how SPI can be calculated, we tried to fit different models using FBref dataset features. Given the high number of features p compared to the number of records n, we decided to immediately introduce a method of dimensionality reduction, the *Principal Component Analysis* (PCA).
We tried to understand if and how SPI values fit with our principal components and we achieved significant results. To do this, we built a matrix in which the columns were the principal components gained with PCA and the rows were the

various teams of each season (so there were included teams that played more than a season in Serie A, marked by the year). We added to this matrix the column with the SPI values. Later, we observed the correlation between each feature and SPI, see fig.2.2 and tab.2.1.

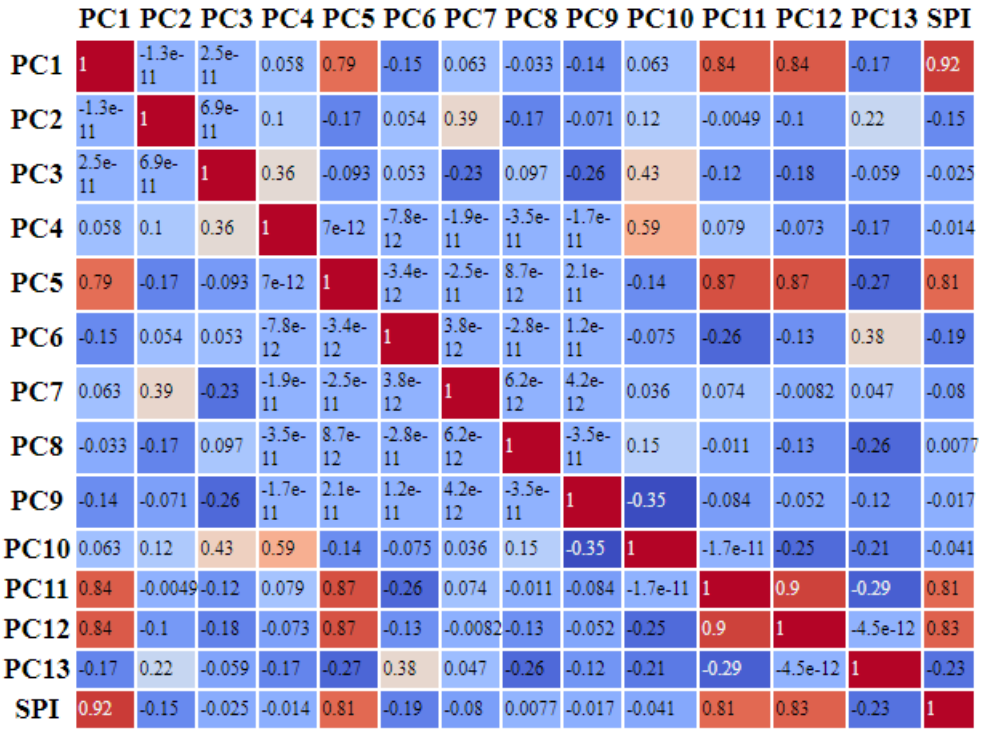


Figure 2.2: SPI PCA correlation matrix.

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

Table 2.1: Correlation values among SPI index and PCs.

So, we begin with the explanation of Principal Component Analysis.

2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) allows us to summarize and to visualize the information in a data set containing individuals/observations described by multiple inter-correlated quantitative variables. Each variable could be considered as a different dimension; if you have more than 3 variables in your data sets, it could be very difficult to visualize a multi-dimensional hyperspace.
PCA is used to extract the important information from a multivariate data table and to express this information as a set of few new variables called *principal components*. These new variables correspond to a linear combination of the originals. The number of principal components is less than or equal to the number of original variables.
The information in a given dataset corresponds to the total variation it contains. The goal of PCA is to identify directions (or principal components) along which the variation in the data is maximal. In other words, PCA reduces the dimensionality of a multivariate data to a small number of principal components, that can be visualized graphically, with minimal loss of information.
Moreover, the Principal Components determined have some properties: they are orthogonal, normalized, not correlated with each other.

After the initial phase of variables standardization, we ran the PCA using the *scikit-learn* module **sklearn.decomposition.PCA**: giving the number n of principal components and the data to fit, the module returns the calculated principal components and the *explained_variance_ratio_* of each of them. We did the same thing through the diagonalization of the covariance matrix, just to verify the results and to obtain the explicit value of eigenvalues.
In order to decide which eigenvector(s) can be dropped without losing too much information for the construction of lower-dimensional subspace, we need to inspect the corresponding eigenvalues: the eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones which can be dropped. In order to do so, the common approach is to rank the eigenvalues from highest to lowest and to choose the top k eigenvectors.
"How many principal components are we going to choose for our new feature subspace?" A useful measure is the so-called *explained variance*, which can be calculated from the eigenvalues. The explained variance tells us how much information (variance) can be attributed to each of the principal components. Usually the selection criteria are:

1. Explained variance: a minimum threshold of explained variance is set.

2. Eigenvalue-one: since the standardized original variables have unit variance, those greater than 1 are chosen because they express PCs which synthesize more information than the single original variables.

3. Scree-Test: the PCs whose eigenvalues precede the maximum jump of explained variability are considered.

This represents the main characteristic of Principal Component Analysis: the "trade off" between loss of information and simplification of the problem.

We first analyzed Teams tables and, in particular, we performed the PCA by dividing the features into the four fields identified previously: *attack*, *defense*, *passing_types*, *possession*. The main reason is that, by separating the fields of action, a better interpretation of the resulting space can be obtained and the dimensionality reduction does not drastically affect the analysis.
Below are the graphs with the explained variance percentage and the cumulative explained variance (the threshold is set at 75%, see fig.2.3).

In the case of *passing_types*, following the selection criteria listed before, the first principal component alone would be enough, indeed itself exceeds the threshold of 75%; anyway, we considered correct to add an "extra" criterion: to take into account at least two principal components. The reason is due to the fact that, facing the dimensionality reduction of a space with a minimum of twenty features, reducing everything to only one PC can cause several difficulties in the interpretation of the PC itself; so, considering at least one more PC can make the analysis easier.
The retrieval of loading factors and the study of the correlation circles allowed us to provide an interpretation of the principal components obtained:

- *Attack* table - 3 PCs: 'Capacità Finalizzazione', 'Capacità Realizzazione', 'Tiro da palla inattiva'

- *Defense* table - 6 PCs: '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* table - 2 PCs: 'Passaggi propositivi media gittata (precisione nei passaggi e pochi tocchi in area)' , 'Palla contesa a media altezza (gioco medio-alto / palla non giocata a terra)'

- *Possession* table - 2 PCs: 'Controllo palla nella metà campo avversaria', 'Efficacia del possesso'

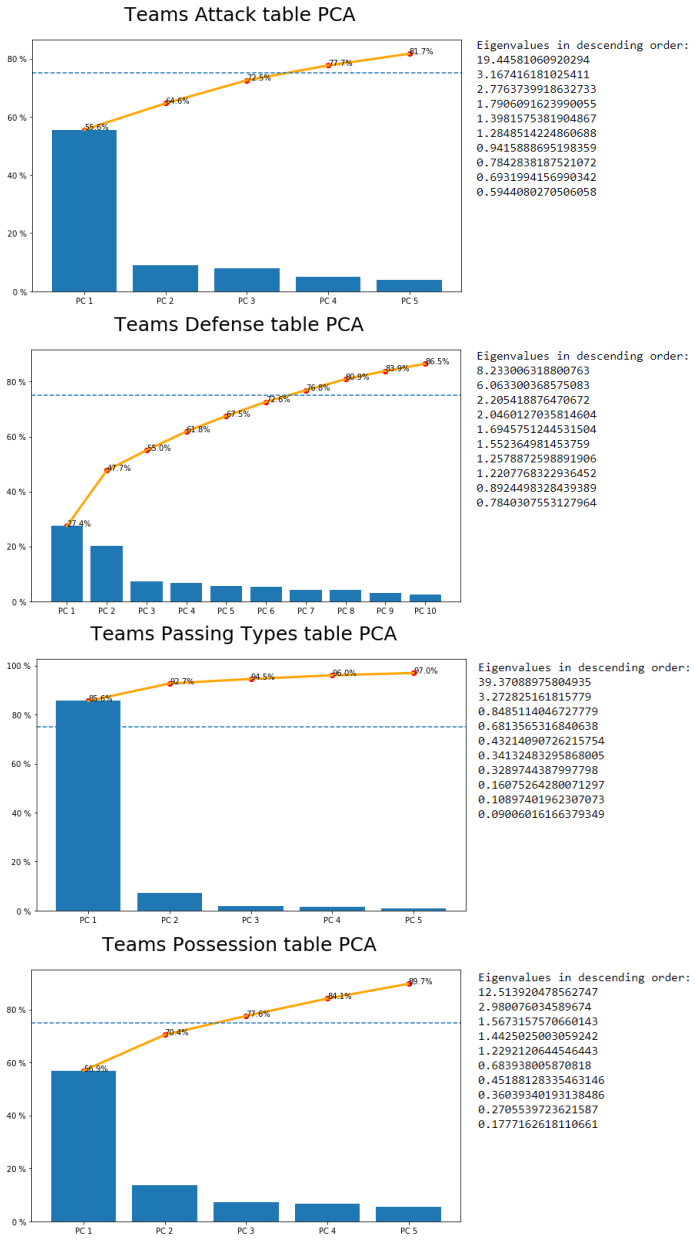


Figure 2.3: PCA explained variance percentage and cumulative explained variance for Teams tables.

Summaries (2.4) and box-plots (2.5) of the principal components obtained follow.

	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

Figure 2.4: Teams Principal Components summaries.

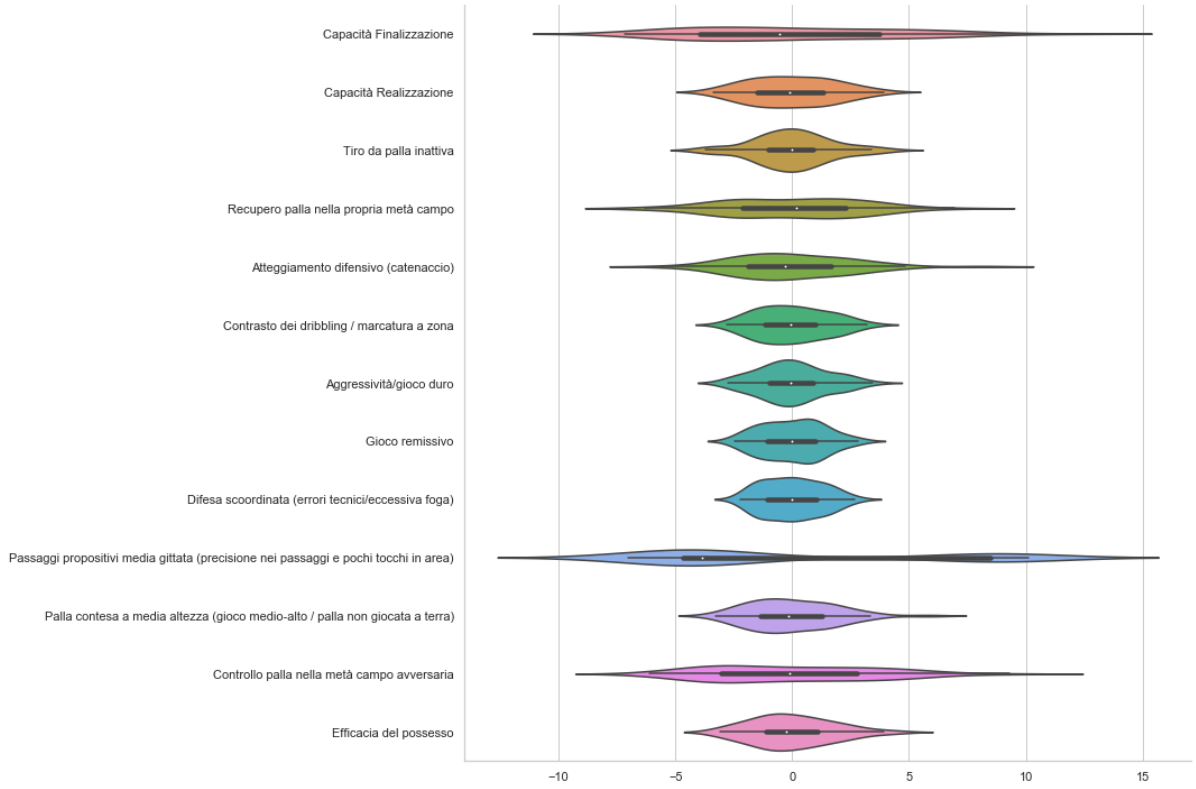


Figure 2.5: Teams Principal Components Box-plots.

All the distributions seem to be symmetric, except for ‘Capacità Finalizzazione’, ‘Atteggiamento difensivo (catenaccio)’, ‘Passaggi propositivi media gittata (precisione nei passaggi e pochi tocchi in area)’ and ‘Efficacia del possesso’, which present a positive asymmetry ($median < mean$); in particular, ‘Capacità Finalizzazione’ and ‘Passaggi propositivi media gittata (precisione nei passaggi e pochi tocchi in area)’ show a quite marked bimodal behaviour.

2.2 Regression Models

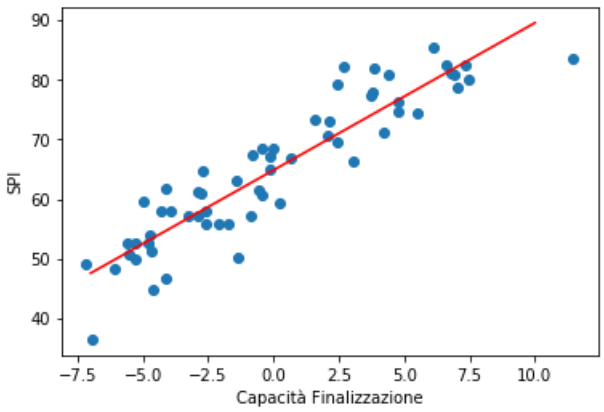
At this point we created four different regression models and we tried to figure out how well our principal components fit with SPI.

Simple Linear Regression

Simple Linear Regression is an approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. Mathematically, we can write this linear relationship as:

$$Y \approx \beta_0 + \beta_1(X). \tag{2.1}$$

To implement it, we imported **LinearRegression** from *sklearn.linear_model* and fitted a linear regression first, using the most correlated feature: 'Capacità finalizzazione' ($\rho = 0.92$).
We found values for intercept and coefficient equal to 64.85 and 2.47, and a score of 0.84 (fig.2.6).




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

Figure 2.6: SPI linear model.

Multiple Linear Regression

Including all features, we created a multiple regression model because instead of fitting a separate linear regression model for each predictor, a better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors. We can do this by giving each predictor a separate slope coefficient in a single model.
In general, the multiple linear regression model takes the form:

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + ... + \beta_p(X_\mathrm{p}) + \epsilon \tag{2.2}$$

We interpret β_j as the average effect of Y of a one unit increase in X_j, holding all other predictors fixed.
Let's see our results.
Score grew from 0.84 to 0.90, and we also calculated an R^2_{adj} of 0.87. We have to say that the intercept is substantially the same ($\sim$64.85), but now we have a number of coefficients equal to the number of components.
Coefficients estimates and the relative p-values are shown in fig.2.7.

```
=========== 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
```

Figure 2.7: SPI multiple linear regression model summary.

As we can see, most of p-values are higher than 0.05, except for intercept, 'Capacità Finalizzazione', 'Aggressività/gioco duro', 'Difesa scoordinata' ones. This makes us guess that SPI is highly represented by the feature that we, in PCA, identified as 'Capacità Finalizzazione', and that this is consistent with the high correlation between them ($\rho = 0.92$).
We followed the same logical steps to implement also Lasso and Ridge Regression.

Lasso

To advance in the analysis we introduced *Shrinkage Methods*: *Lasso* and *Ridge*; we start with the Lasso. Lasso regression shrinks the regression coefficients by imposing a penalty on their size: it consists in the minimization of the residual sum of squares to which a penalty factor is added. Here α is a tuning parameter that controls the amount of shrinkage: the larger the value of α, the greater the

amount of shrinkage.
From *sklearn.linear_model*, we imported **Lasso** and gained an R^2_{adj} of 0.87 with an α of 0.1. As we did previously, we report coefficients estimates and the relative p-values below (fig.2.8).

```
========== SUMMARY ==========
Residuals:
    Min      1Q  Median      3Q     Max
-8.1916 -2.7421 -0.5346  2.442  8.4112


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                      64.854500    0.491104  132.0587  0.000000
Capacità Finalizzazione                          2.138345    0.220419    9.7013  0.000000
Capacità Realizzazione                          -0.399401    0.328882   -1.2144  0.229426
Tiro da palla inattiva                           0.063552    0.350951    0.1811  0.856921
Recupero palla nella propria metà campo         -0.221101    0.220122   -1.0044  0.319264
Atteggiamento difensivo (catenaccio)             0.673822    0.521239    1.2927  0.201143
Contrasto dei dribbling / marcatura a zona      -0.385928    0.406684   -0.9490  0.346511
Aggressività/gioco duro                         -0.811227    0.383531   -2.1152  0.038650
Gioco remissivo                                  0.153199    0.405112    0.3782  0.706666
Difesa scoordinata (errori tecnici/eccessiva foga)  0.840214    0.458620    1.8320  0.071993
Passaggi propositivi media gittata (precisione ...  -0.003543    0.104664   -0.0339  0.973109
Palla contesa a media altezza (gioco medio-alto...  -0.000000    1.248553   -0.0000  1.000000
Controllo palla nella metà campo avversaria      0.098840    0.617934    0.1600  0.873465
Efficacia del possesso                          -0.102882    0.442552   -0.2325  0.816975
---
R-squared:  0.89765,    Adjusted R-squared:  0.86872
F-statistic: 31.03 on 13 features
```

Figure 2.8: SPI Lasso regression model summary, α = 0.1.

As we can see, only the intercept, “Capacità Finalizzazione” and “Aggressività/-gioco duro” have got p-values lower than 0.05.
Anyway, unlike least squares, which generates only one set of coefficient estimates, Lasso regression produces a different set of coefficient estimates for each value of α. Selecting a good value for α is critical, to do so we used *Cross-Validation*.

– *K-fold Cross-Validation*

We followed two “paths” to perform a 5-fold Cross-Validation:

1. Using the module **LassoCV** in *sklearn.linear_model*: this module only requires to insert the value k (the list of α values can be passed explicitly with the parameter *alphas*, if *None* alphas are set automatically), then returns the best α value and the corresponding score: α = 1.42, score = 0.87.

2. Using the module **cross_val_score** in *sklearn.model_selection* and combining it with the already seen **Lasso** in *sklearn.linear_model*: in this case we iterated over a specified set of alpha values and retrieved the results to calculate the final score: α = 1.42, score = 0.82.

LassoCV is the best choice, as it is the reference module in this area and in the literature it has excellent results; however, the introduction of *cross_val_score* allowed us to have some useful graphical feedback: as we can see in fig.2.9, the test_mse obtained with *LassoCV* (on the left) and the score obtained with *cross_val_score* (on the right) both lead to the same alpha value.

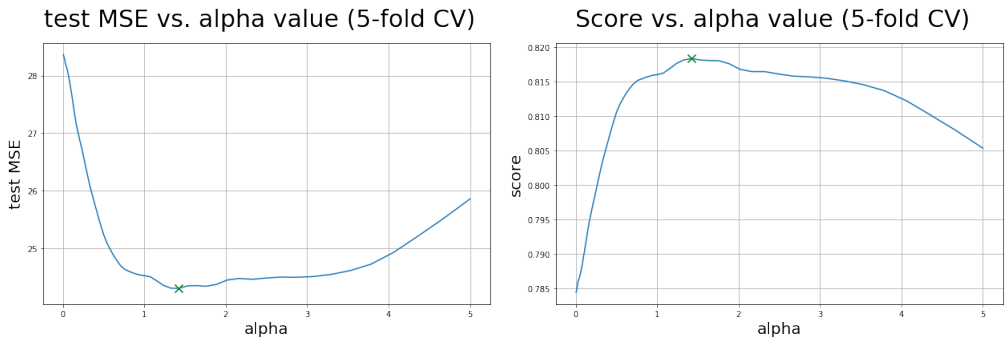


Figure 2.9: SPI Lasso regression model α selection.

Finally, we re-ran the Lasso regression with the best α value and report the corresponding summary, fig.2.10. The score value is 0.84.

```
========== 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
```

Figure 2.10: SPI Lasso regression model summary, α = 1.42.

Ridge

To conclude this section, we built a Ridge regression.
In Ridge regression, we add a penalty term which is equal to the square of the

coefficient. We also add a coefficient α to control that penalty term. Ridge regression decreases the complexity and multi-collinearity of a model, but does not reduce the number of variables since it never leads to a coefficient been zero rather only minimizes it (hence, this model is not good for feature reduction).
First of all, we imported **Ridge** from *sklearn.linear_model* and fixed α equal to 0.1, getting an R^2_{adj} equal to 0.87. Coefficients estimates and the relative p-values follow (fig.2.11).

```
=========== 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
```

Figure 2.11: SPI Ridge regression model summary, α = 0.1.

Almost every p-value is higher than 0.05, except for the intercept, “Capacità Finalizzazione”, “Aggressività/gioco duro” and “Difesa scoordinata”.
Then, to select the best value for α, we made a 5-fold Cross-Validation.

– *K-fold Cross-Validation*

The reasoning is similar to that made for Lasso regression. Since the method *mse_path_* used in *LassoCV* is not available with **RidgeCV**, we report as graphic result only the *cross_val_score* one, fig.2.12.
The best α value was determined, α = 106, and the corresponding summary reported, fig.2.13. The final score is 0.85.

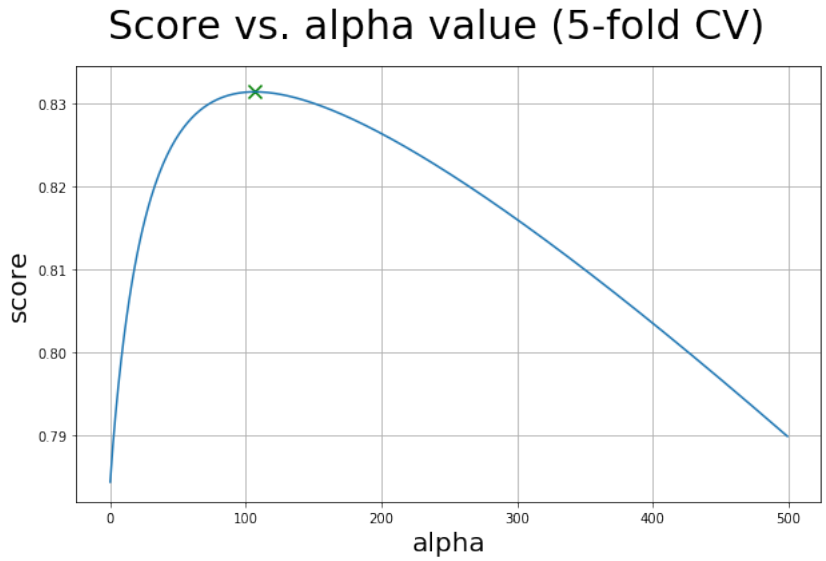


Figure 2.12: SPI Ridge regression model α selection.

```
========== 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
```

Figure 2.13: SPI Ridge regression model summary, α = 106.

Chapter 3

Single player evaluation - FIFA indices

References: Player Rating Guide, SoFIFA.
Dataset: FIFAdataset.

Ratings are usually used as basis for comparison between two or more players. A player's rating is represented by a number, from 1 to 99. It represents the player's potential relatively to a determined set of attributes, according to his position in the field, and that also takes into account his popularity and the environment he is in. So, it is not just a simple number that measures the player's quality and does not tell how good a player is at all. Overall rating is not even some kind of average or reflex of all the attributes. It tells nothing about the player's technical, mental and physical attributes. Sometimes, players with lower ratings can be better choices in a particular situation.

Figure 3.1: Card Sample

3.1 FIFA Player Ratings Calculations

Each player has 36 attributes associated with him, but, analysing so many numbers can be a bit of a complex process. Therefore, six numbers were introduced as basic attributes to make it all simpler (fig.3.1). These six numbers, related to as many features (pace, shooting, passing, dribbling, defending and physicality) are an average representation of the 36 attributes. Relatively to the player's overall rating, it is not based on the six basic attributes but on another selection of attributes. A player's overall rating is the result of two parcels added up together: one is the weighted average of a certain selection of attributes, the other refers to International Reputation. Here is what the FIFA 19 player ratings calculation formula looks like:

$$OVR = ATT + IR$$

where:

- *OVR* is the Player Overall Rating (1 to 99)
- *ATT* is the Weighted Average of Attributes (1 to 99)
- *IR* is International Reputation (0 to 3)

Table 3.1 represents the ponderations that should be considered for the calculation of the *ATT* variable, for each one of the positions. The weighted average of attributes is not calculated the same way for all players. It depends on their position on the pitch, so the only attributes that will be taken into account are the ones considered to be important for the position. *International Reputation*, also known as International Recognition, is an attribute that affects the player's rating according to his club's local and international prestige (fig.3.2). It is based essentially on the popularity, history and results of them both. Basically, *IR* was created in order to adjust the players rating relatively to everything that does not actually have to do with his technical, physical and mental capacities. As with weak foot and skills, players are evaluated relatively to International Reputation on a scale of 1 to 5 stars; so, if there are more stars, that means the player has a better reputation.

3.1.1 Understanding FIFA overall index

A player who has a higher score than another is not necessarily better. Ratings are way too subjective, they cannot determine a player's quality with accuracy. Some examples, useful for understanding the phenomenon, are provided below. Say we have two players for the same position: Thiago Alcântara and Radja Nainggolan.

DEFENDERS			WINGERS	
CB	**RB/LB**	**RWB/LWB**	**RM/LM**	**RW/LW**
15% Marking	13% Sliding Tackle	11% Standing Tackle	14% Crossing	16% Crossing
15% Standing Tackle	12% Standing Tackle	10% Sliding Tackle	14% Dribbling	12% Attack Positioning
15% Sliding Tackle	12% Interceptions	10% Crossing	12% Short Passing	11% Dribbling
10% Heading	10% Marking	10% Short Passing	12% Ball Control	11% Ball Control
10% Strength	08% Stamina	10% Ball Control	08% Long Passing	10% Shot Power
08% Aggression	08% Reactions	10% Interceptions	08% Vision	10% Long Shots
08% Interceptions	07% Crossing	09% Marking	07% Reactions	10% Reactions
05% Short Passing	07% Heading	08% Stamina	07% Attack Positioning	06% Short Passing
05% Ball Control	07% Ball Control	08% Reactions	05% Stamina	05% Heading
05% Reactions	06% Short Passing	07% Dribbling	05% Acceleration	05% Vision
04% Jumping	05% Sprint Speed	04% Sprint Speed	05% Sprint Speed	04% Acceleration
	05% Aggression	03% Agility	03% Agility	04% Sprint Speed
CENTRE MIDFIELDERS			**STRIKERS**	
CDM	**CM**	**CAM**	**RF/CF/LF**	**ST**
13% Short Passing	15% Short Passing	16% Short Passing	12% Finishing	20% Finishing
12% Interceptions	13% Long Passing	16% Vision	12% Attack Positioning	12% Attack Positioning
11% Long Passing	12% Vision	13% Ball Control	11% Dribbling	10% Heading
10% Marking	10% Ball Control	12% Attack Positioning	11% Ball Control	10% Shot Power
10% Standing Tackle	09% Dribbling	11% Dribbling	10% Shot Power	10% Reactions
09% Ball Control	08% Reactions	08% Reactions	10% Long Shots	08% Dribbling
09% Reactions	08% Interceptions	06% Long Shots	10% Reactions	08% Ball Control
08% Vision	08% Attack Positioning	05% Finishing	06% Short Passing	05% Volley
06% Stamina	06% Standing Tackle	05% Shot Power	05% Heading	05% Long Shots
06% Strength	06% Stamina	04% Acceleration	05% Vision	05% Acceleration
05% Aggression	05% Long Shots	04% Agility	04% Acceleration	04% Sprint Speed
			04% Sprint Speed	03% Strength

Table 3.1: Attributes calculation table per position.

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

Figure 3.2: International reputation

The first one has an overall rating of 86, whilst the second has one point less. Comparisons will always have some kind of subjectivity, but for most of the FIFA community the Belgian midfielder is better, hence his superior price. The reason for that is simple: Nainggolan is a much more polyvalent player. When EA try to transform their capacities into one number, they do it accordingly to the card's position, *CM*. That means they do not take into account a lot of extremely important attributes for a good *CDM*, such as aggression, strength and marking. Besides, there are attributes that are very appreciated by the community but weigh little on

the rating calculation, which is stamina's case, having Radja beating Thiago by miles.
Overall ratings tend to underrate the most versatile players because they focus mainly on a single position and, also, ignore skills, weak foot, height, work rates or even some attributes like balance, curve, free kick accuracy, penalties and flaws the player might have. Figure 3.3 shows what attributes are most often considered for the overall rating.

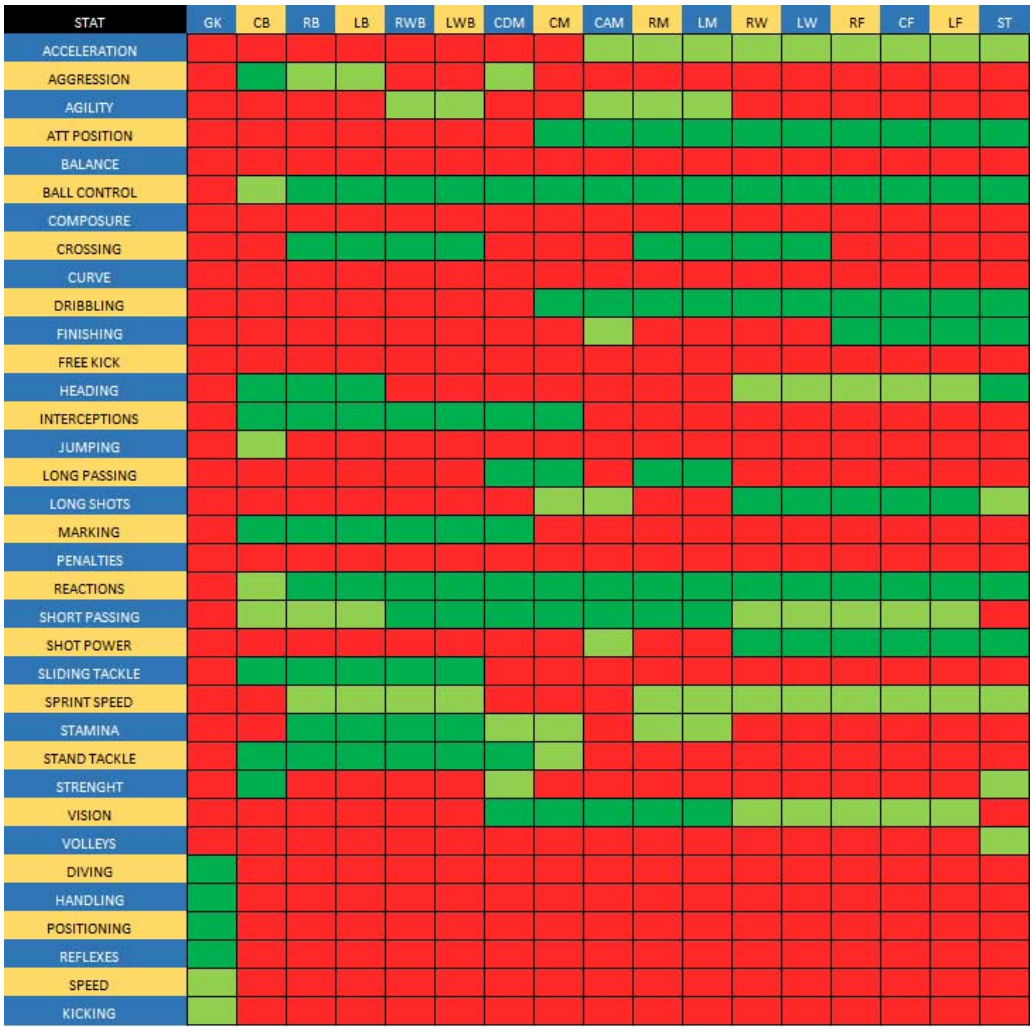


Figure 3.3: Attributes matrix, in this figure are shown the attributes considered relevant for each position.

To analyze even more in detail and definitively clarify, next will be compared two players playing in the same league (fig.3.4): Vincent Kompany and Davinson Sánchez. Kompany has a rating of 85, Sánchez 84. Although Kompany has a

higher score, it is a common opinion of the FIFA community that Sánchez is pretty superior. How can this be explained? International reputation is the answer.

Figure 3.4: Players' indices comparison.

Sánchez has only two stars of reputation which give him no rating bonus, whilst Kompany has four of them which end up giving him two extra points of rating. This means that Sánchez has indeed a base rating superior to Kompany's ($84 > 83$) and that he has everything needed to be the be the best of the two.

3.2 Statistical Learning Models

Following the line of thought traced during the analysis of the SPI, we have decided to proceed with the PCA in this case too. The main reason is related to the fact that a reduction in space from more than one hundred features to about ten allows for easier interpretation of the models and better management of machine resources, indeed, trying to run the same models using the initial features, we realized that convergence problems and (probable) calculation errors due to inefficiencies in machine precision (ϵ) are quickly incurred.

3.2.1 PCA

Retracing the steps detailed in the previous chapter, we introduce the PCA in players' tables too. The criteria followed for the selection of the components are the same.

The graphs relating to the explained variance percentage and the cumulative explained variance (the threshold is still fixed at 75%) can be seen in fig.3.5.

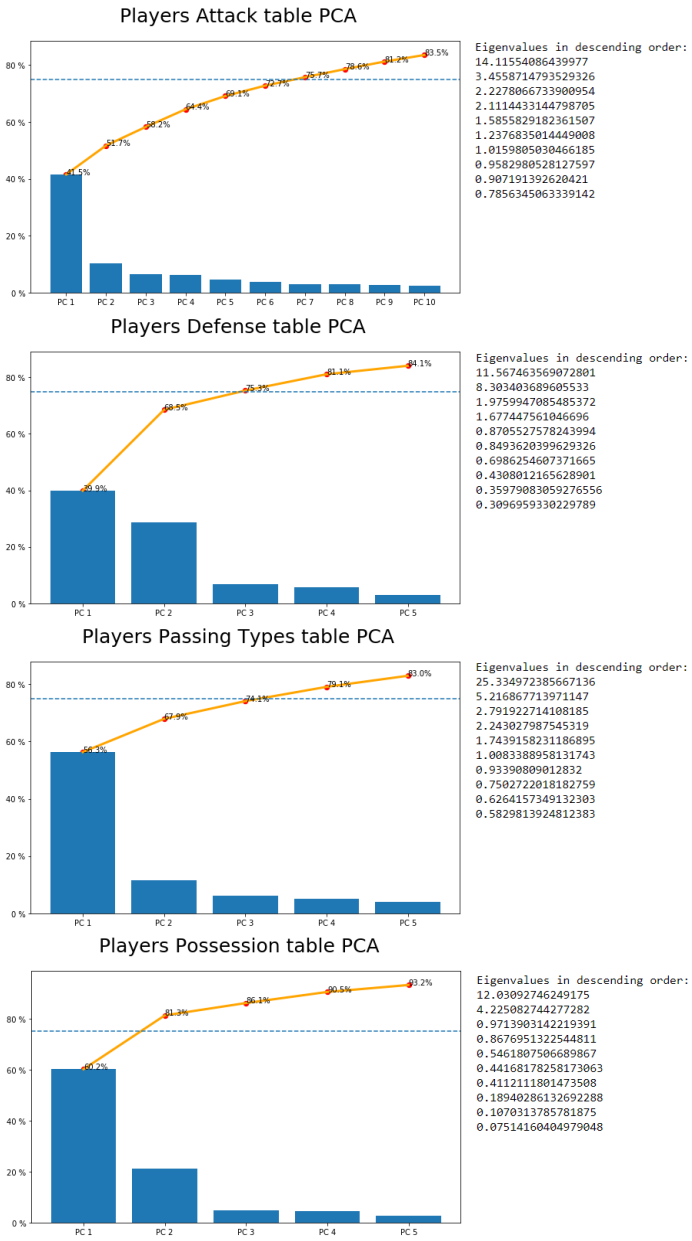


Figure 3.5: PCA explained variance percentage and cumulative explained variance for Players tables.

The retrieval of loading factors and the study of the correlation circles allowed us to provide an interpretation of the principal components obtained:

- *Attack* table - 6 PCs: 'Incisività offensiva', 'Individualismo (palla in movimento)', 'Capacità di finalizzazione', 'Inefficacia del tiro', 'Gioco offensivo con palla in movimento', 'Altruismo'
- *Defense* table - 2 PCs: 'Supremazia atletica / giocate aggressive', 'Pressing efficace in zona difensiva'
- *Passing Types* table - 3 PCs: 'Passaggio propositivo e nell'ultimo 1/3 di campo', 'Gioco con palla in movimento', 'Inefficacia del passaggio'
- *Possession* table - 2 PCs: 'Alta propositività nella metà campo avversaria', 'Avanzamento e dribbling aggressivo/superfluo'[1]

Summaries (3.6) and box-plots (3.7) of the principal components obtained follow.

	Incisività offensiva	Individualismo (palla in movimen	Capacità di finalizzazione	Inefficacia del tiro	Gioco offensivo con palla in	Altruismo	Supremazia atletica / giocate aggressiv	Pressing efficace in zona difensiva	Passaggio propositivo e nell'ultim	Gioco con palla in movimen	Inefficacia del passaggio	Alta propositività nella metà	Avanzamento e dribbling aggressiv
count	1535	1535	1535	1535	1535	1535	1535	1535	1535	1535	1535	1535	1535
unique													
top													
freq													
mean	6.30E-02	3.62E-03	-1.11E-02	-3.60E-02	-1.13E-03	5.28E-03	8.78E-02	1.53E-02	1.27E-01	-1.82E-02	9.99E-03	9.06E-02	-1.09E-02
std	3.77	1.87	1.40	1.41	1.27	1.10	3.39	2.91	5.02	2.30	1.67	3.45	2.08
min	-3.49	-10.04	-7.53	-7.84	-6.65	-4.21	-4.39	-7.90	-6.40	-8.68	-6.57	-4.51	-7.35
25%	-2.51	-0.78	-0.80	-0.73	-0.47	-0.51	-2.72	-1.46	-4.11	-1.11	-0.61	-2.83	-1.08
50%	-1.20	-0.20	-0.26	0.21	-0.14	0.01	-0.60	-0.13	-1.15	0.18	0.19	-0.59	0.19
75%	1.39	0.81	0.55	0.68	0.35	0.37	2.27	1.99	3.34	0.90	0.82	2.33	1.13
max	23.61	10.25	13.06	8.12	10.62	14.94	15.17	8.31	19.89	12.99	11.12	14.52	7.65

Figure 3.6: Players Principal Components summaries.

[1]Coherently with what has been said in the case of Teams Passing Types, we chose two PCs, even though only the first one would have been enough.

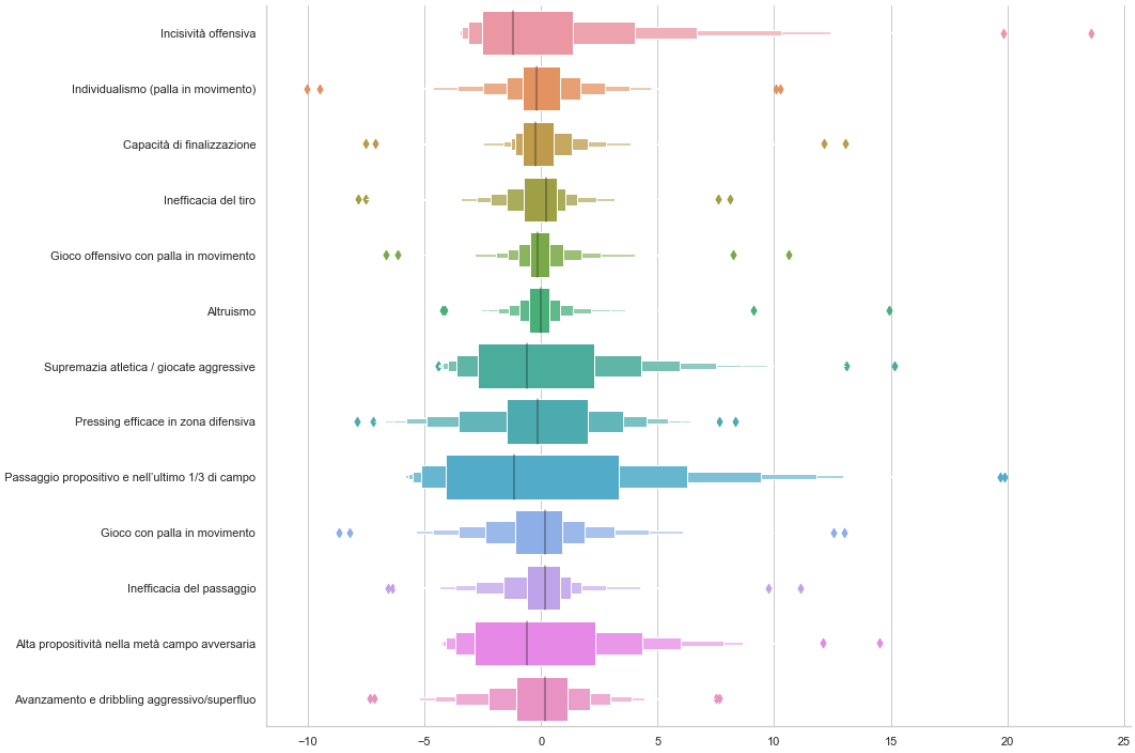


Figure 3.7: Players Principal Components Box-plots.

As in the case of Teams, most of the distributions seem to be symmetric, except for 'Incisività offensiva', 'Supremazia atletica / giocate aggressive', 'Passaggio propositivo e nell'ultimo 1/3 di campo' and 'Alta propositività nella metà campo avversaria', which present a positive asymmetry ($median < mean$).
Later, we observed the correlation between each feature and Overall_index, see fig.3.8 and tab.3.2.

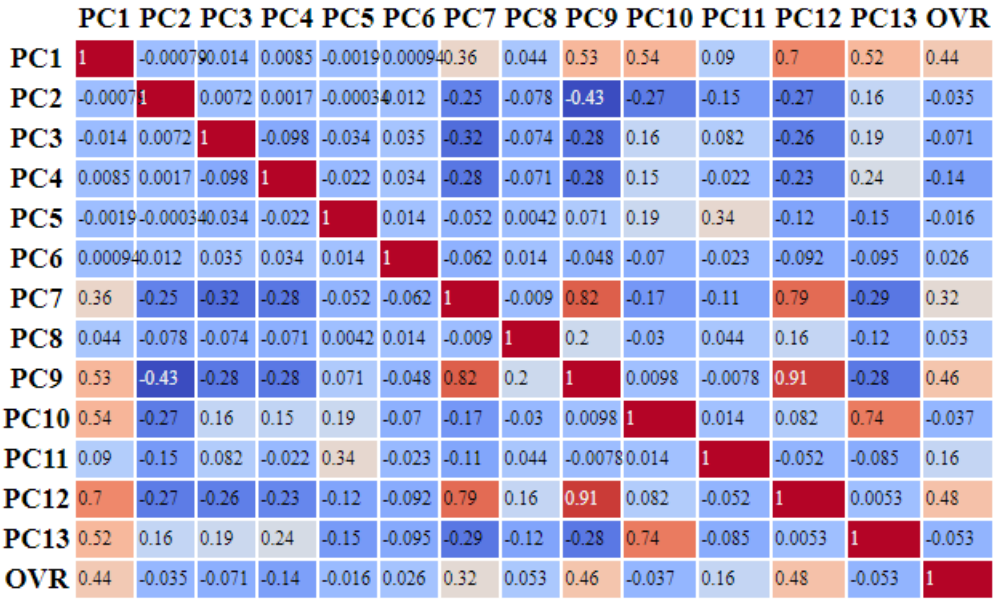


Figure 3.8: FIFA PCA correlation matrix.

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

Table 3.2: Correlation values among FIFA Overall_index and PCs.

3.2.2 Linear Models

As we have done with the *SPI*, here we report the linear models used to study the *Overall_index*.

Multiple Linear Regression

Using the features retrieved with the PCA, we ran a multiple linear regression. Summary is reported in fig.3.9; anyway, the R^2_{adj} is equal to 0.36, quite low.

```
=========== 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
```

Figure 3.9: FIFA Overall_index multiple linear regression model summary.

P-values confirm that all features are relevant, except for "Individualismo (palla in movimento)", "Capacità di finalizzazione", "Altruismo" and "Avanzamento e dribbling aggressivo/superfluo".
Trying to improve the value of R^2_{adj} we deepened the analysis and introduced *Lasso* and *Ridge*.

Lasso

Following the same steps introduced in the previous chapter, we ran a Lasso regression with α set equal to 0.1, keeping on an R^2_{adj} of 0.36 (fig.3.10). After that we performed a 5-fold Cross-Validation (as seen with the two "paths" of *LassoCV* and *cross_val_score*), selecting an optimal alpha value of 0.04 (fig.3.11).
The small difference with respect to the value 0.1 does not lead to significant improvements, as we can see in fig.3.12

```
=========== SUMMARY ===========
Residuals:
     Min       1Q   Median       3Q      Max
 65.1824  71.2601  73.0489  75.6519  92.2561


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                      73.719869    0.124664  591.3474  0.000000
Incisività offensiva                             0.806619    0.081358    9.9145  0.000000
Individualismo (palla in movimento)              0.000000    0.113731    0.0000  1.000000
Capacità di finalizzazione                       0.000000    0.099368    0.0000  1.000000
Inefficacia del tiro                            -0.149664    0.099478   -1.5045  0.132660
Gioco offensivo con palla in movimento          -0.128227    0.131218   -0.9772  0.328624
Altruismo                                        0.000000    0.118867    0.0000  1.000000
Supremazia atletica / giocate aggressive        -0.428787    0.070977   -6.0412  0.000000
Pressing efficace in zona difensiva             -0.133361    0.045442   -2.9347  0.003388
Passaggio propositivo e nell'ultimo 1/3 di campo  0.562815    0.100246    5.6143  0.000000
Gioco con palla in movimento                    -0.877262    0.127656   -6.8721  0.000000
Inefficacia del passaggio                        0.357264    0.089484    3.9925  0.000068
Alta propositività nella metà campo avversaria  -0.118863    0.157870   -0.7529  0.451614
Avanzamento e dribbling aggressivo/superfluo    -0.000000    0.169389   -0.0000  1.000000
---
R-squared:  0.36400,    Adjusted R-squared:  0.35857
F-statistic: 66.96 on 13 features
```

Figure 3.10: FIFA Overall_index Lasso regression model summary, $\alpha = 0.1$.

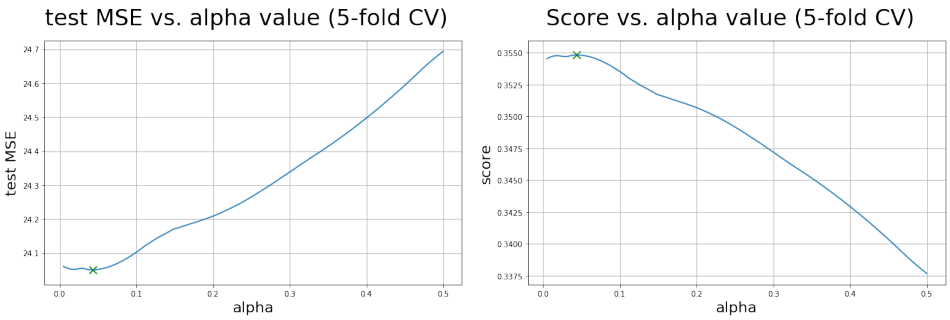


Figure 3.11: FIFA Overall_index Lasso regression model α selection.

```
=========== SUMMARY ===========
Residuals:
     Min       1Q   Median       3Q      Max
 65.0076  71.2697  73.0244  75.5838  92.2266


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                      73.718264    0.124414  592.5218  0.000000
Incisività offensiva                             0.839303    0.081195   10.3369  0.000000
Individualismo (palla in movimento)              0.060064    0.113503    0.5292  0.596754
Capacità di finalizzazione                       0.000000    0.099169    0.0000  1.000000
Inefficacia del tiro                            -0.176075    0.099279   -1.7735  0.076337
Gioco offensivo con palla in movimento          -0.273268    0.130956   -2.0867  0.037078
Altruismo                                        0.000000    0.118629    0.0000  1.000000
Supremazia atletica / giocate aggressive        -0.455878    0.070835   -6.4358  0.000000
Pressing efficace in zona difensiva             -0.150570    0.045351   -3.3201  0.000921
Passaggio propositivo e nell'ultimo 1/3 di campo  0.694968    0.100046    6.9465  0.000000
Gioco con palla in movimento                    -0.885841    0.127400   -6.9532  0.000000
Inefficacia del passaggio                        0.398864    0.089305    4.4663  0.000009
Alta propositività nella metà campo avversaria  -0.299558    0.157554   -1.9013  0.057449
Avanzamento e dribbling aggressivo/superfluo     0.016810    0.169050    0.0994  0.920804
---
R-squared:  0.36655,    Adjusted R-squared:  0.36114
F-statistic: 67.70 on 13 features
```

Figure 3.12: FIFA Overall_index Lasso regression model summary, $\alpha = 0.04$.

Ridge

In addition to Lasso, we also ran Ridge Regression. After importing **Ridge** from *sklearn.linear_model*, and setting alpha equal to 0.1, we obtained an R^2_{adj} of 0.36. Coefficients estimates and the relative p-values are reported in fig.3.13.

```
========== SUMMARY ==========
Residuals:
     Min       1Q  Median       3Q      Max
 64.1569  71.2371  72.991  75.5821  91.8179


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                      73.715093    0.124276  593.1581  0.000000
Incisività offensiva                             0.819537    0.081104   10.1047  0.000000
Individualismo (palla in movimento)              0.115772    0.113377    1.0211  0.307356
Capacità di finalizzazione                       0.005053    0.099058    0.0510  0.959322
Inefficacia del tiro                            -0.200453    0.099168   -2.0214  0.043417
Gioco offensivo con palla in movimento          -0.371198    0.130809   -2.8377  0.004604
Altruismo                                        0.025071    0.118496    0.2116  0.832467
Supremazia atletica / giocate aggressive        -0.483151    0.070756   -6.8284  0.000000
Pressing efficace in zona difensiva             -0.164133    0.045300   -3.6232  0.000300
Passaggio propositivo e nell'ultimo 1/3 di campo  0.942606    0.099934    9.4323  0.000000
Gioco con palla in movimento                    -1.020489    0.127258   -8.0190  0.000000
Inefficacia del passaggio                        0.455486    0.089205    5.1061  0.000000
Alta propositività nella metà campo avversaria  -0.583478    0.157378   -3.7075  0.000217
Avanzamento e dribbling aggressivo/superfluo     0.301023    0.168861    1.7827  0.074838
---
R-squared:  0.36796,    Adjusted R-squared:  0.36256
F-statistic: 68.12 on 13 features
```

Figure 3.13: FIFA Overall_index Ridge regression model summary, $\alpha = 0.1$.

As we can see, most of the p-values are acceptable, except for "Capacità di finalizzazione", "Inefficacia del tiro", "Altruismo", "Pressing efficace in zona difensiva", "Avanzamento e dribbling aggressivo/superfluo".
In addition, we repeated the steps for the 5-fold Cross-Validation, with both *RidgeCV* and *cross_val_score*, resulting in an optimal α of 37 (fig.3.14). The corresponding summary is reported in fig.3.15. The final score is 0.36.

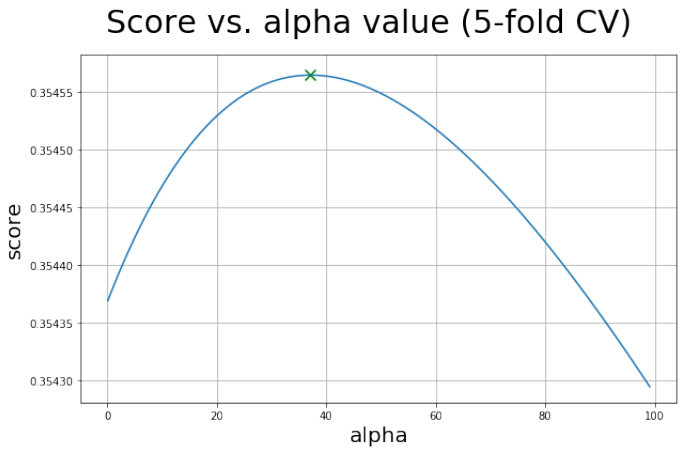


Figure 3.14: FIFA Overall_index Ridge regression model α selection.

```
========== SUMMARY ==========
Residuals:
    Min       1Q  Median       3Q      Max
 64.426  71.2501  73.014  75.5816  91.7687


Coefficients:
                                                 Estimate  Std. Error   t value   p value
_intercept                                      73.715885    0.124289  593.1011  0.000000
Incisività offensiva                             0.803956    0.081113    9.9116  0.000000
Individualismo (palla in movimento)              0.125866    0.113389    1.1100  0.267156
Capacità di finalizzazione                       0.012347    0.099069    0.1246  0.900832
Inefficacia del tiro                            -0.189941    0.099178   -1.9151  0.055660
Gioco offensivo con palla in movimento          -0.358499    0.130823   -2.7403  0.006209
Altruismo                                        0.032050    0.118509    0.2704  0.786854
Supremazia atletica / giocate aggressive        -0.475539    0.070763   -6.7201  0.000000
Pressing efficace in zona difensiva             -0.162611    0.045305   -3.5892  0.000342
Passaggio propositivo e nell'ultimo 1/3 di campo  0.876004    0.099945    8.7649  0.000000
Gioco con palla in movimento                    -0.954060    0.127272   -7.4962  0.000000
Inefficacia del passaggio                        0.453032    0.089215    5.0780  0.000000
Alta propositività nella metà campo avversaria  -0.487594    0.157395   -3.0979  0.001984
Avanzamento e dribbling aggressivo/superfluo     0.215589    0.168879    1.2766  0.201940
---
R-squared:  0.36783,    Adjusted R-squared:  0.36242
F-statistic: 68.08 on 13 features
```

Figure 3.15: FIFA Overall_index Ridge regression model summary, α = 37.

3.2.3 Non-linear Models

Now we move beyond linearity.

Polynomial fit

To try to achieve higher results we implemented a polynomial fit. First, we imported **PolynomialFeatures** from *sklearn.preprocessing*; *PolynomialFeatures* allows you to enter the maximum degree and transforms the dataset through *fit_transform* including all features up to the chosen maximum degree. We set the *degree* parameter to 3 and ran a *LinearRegression* using the transformed dataset. The R^2_{adj} reached is equal to 0.40.
However, we did not like this restriction, so we decided to modify the dataset by adding only the most correlated features up to the third order; and, after that, adding only the most correlated one: "Alta propositività nella metà campo avversaria" (ρ = 0.48). The main result is that, by only considering the most correlated feature up to the third order, we reached a R^2_{adj} of 0.39.

Generalized Additive Model (GAM)

At this point we tried to implement a Generalized Additive Model importing **LinearGAM** from *pygam*.
Generalized Additive Models provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while

maintaining additivity. Just like linear models, GAMs can be applied with both quantitative and qualitative response. A GAM is a model written in the form:

$$y_i = \beta_0 + f_1(x_{i1}) + ... + f_p(x_{ip}) + \epsilon_i \tag{3.1}$$

It is called *additive* because we calculate a separate f_j for each X_j, and then add together all of their contributions.
Before we move on, let us summarize the advantages and limitations of a GAM, as listed in *"Introduction to Statistical Learning"* of Hastie and Tibshirani:

- GAMs allow us to fit a non-linear f_j to each X_j, so that we can automatically model non-linear relationships that standard linear regression will miss;
- the non-linear fits can potentially make more accurate predictions for the response Y;
- because the model is additive, we can still examine the effect of each X_j on Y while holding all of the other variables fixed;
- the smoothness of the function f_j for the variable X_j can be summarized via degrees of freedom;
- the main limitation of GAMs is that the model is restricted to be additive.

In the first run, we fixed the *splines order* to 3, the standard settings provide λ factors equal to 0.6 and number of splines equal to 20. Results are reported in fig.3.16.

The first run of *LinearGAM* provides *GCV* = 25.84 (*Generalized Cross-Validation score*) and Pseudo-R^2 = 0.47; after that, it could be necessary to optimize regarding two parameters, *n_splines* and λ-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 λ factors because of computational resources that are not accessible to us are required.
The *n_splines* optimization leads to *GCV* = 22.67 and Pseudo-R^2 = 0.43 (fig.3.17). Even though the in-sample Pseudo-R^2 value is lower, we can expect our model to generalize better because the *GCV* error is lower.

```
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
```

Figure 3.16: Linear GAM view.

```
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
```

Figure 3.17: Linear GAM view, after *n_splines* optimization.

3.2.4 Regression Trees and Ensemble Methods

Decision Trees are powerful algorithms that can perform regression tasks. One major problem with trees is their high variance. Often a small change in the data can result in a very different series of splits, making interpretation precarious. The major reason for this instability is the hierarchical nature of the process: the effect of an error in the top split is propagated down to all of the splits below it. Trees are also the fundamental component of *Random Forest*, one of the most powerful machine learning algorithm. Ensemble methods make predictions based on a number of different models to achieve higher flexibility. Two most popular ensemble methods are bagging and boosting. *Bagging* trains a bunch of individual models in a parallel way and is useful to reduce trees high variance. *Boosting* trains a models in a sequential way and each individual model learns from mistakes made by the previous one. It can combine several weak learners into a strong learner. Boosting is one of the most powerful learning idea introduced in the last twenty years.

Random Forest

Random forest, one of the most powerful machine learning algorithm, is an ensemble model using bagging as the ensemble method and decision tree as the individual model. The Random Forest algorithm introduces extra randomness when growing trees; instead of searching for the very best feature when splitting a node, it searches for the best feature among a random subset of features. The algorithm results in greater tree diversity, which (again) trades a higher bias for a lower variance, generally yielding an overall better model.

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*, as weak learners that are added sequentially to the ensemble. Each subsequent model attempts to correct the predictions made by the model before it in the sequence. This is achieved by weighing the training dataset to put more focus on training examples on which prior models made prediction errors.

Gradient Boosting and XGBoost

Gradient Boosting, just like AdaBoost, works by sequentially adding predictors to an ensemble, each one correcting its predecessor. This method tries to fit the new predictor to the residual errors made by the previous one and makes a new prediction by simply adding up the predictions (of all trees). *XGBoost* (Extreme

Gradient Boosting) is an improvement of Gradient Boosting and aims to be extremely fast, scalable, and portable. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. The same code runs on major distributed environments (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

Models implementation

First we implemented a simple decision tree regression model and performed hyperparameters tuning via grid search. GridSearchCV is a Python library function that helps to loop through predefined hyperparameters and fit a model on training set. In addition, it is possible to specify cross-validation for each set of hyperparameters. The model has returned (through the function *grid_search.best_estimator_*) the best estimator providing the tree in fig.3.18.

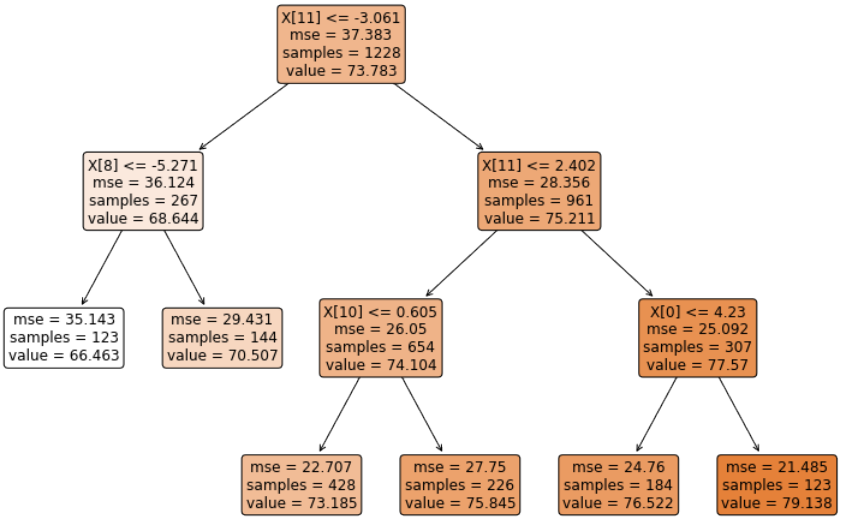


Figure 3.18: OVR Regression tree.

The best estimator shows the effect of pruning too, indeed the leaves relative to the further left second order splits were removed. The main result consists in having an accuracy of 94.06%.
In general, tree based methods results are reported in table 3.3. The most accurate one is the Random Forest algorithm, but the difference is very small, just an increment of 0.5 in the accuracy.

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

Table 3.3

3.3 Overall index decomposition

Finally, we decided to deepen the study of the Overall_index considering the hexagon of skills. In particular, we adapted FIFA skills hexagon to a rhombus, based on the features subdivision described at the beginning of the report; the correspondence between couples of skills are: *attack-shooting*, *defense-defending*, *passing_types-passing*, *possession-dribbling*; perhaps the latter is the weakest association, but we did not found the direct counterpart of ball possession.
Correlation matrices among our PCs and the selected FIFA skills are shown in fig.3.19, while results are reported in tabs.3.4-3.5.

	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

Figure 3.19: Correlation matrices among the PCs obtained and the selected FIFA subindices.

We hoped to reach better results with this subdivision, as it was more correct also on a conceptual level, but this did not happen. With attack, passing_types and possession we obtain slight improvements, with defense, instead, the result is completely disappointing. The introduction of non-linear methods leads to an improvement of about a tenth in R^2_{adj} for attack, passing_types and possession, while it has no adding value for defense, proof of the fact that no method seems to grasp the relationship binding this macro-skill with the identified principal components. So, although we moved beyond linearity, applying tree based ensemble methods too, we did not succeed in finding a satisfying result.

	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	α = 0.024 R^2_{adj} = 0.42	α = 0.34 R^2_{adj} = 0.12	α = 0.05 R^2_{adj} = 0.24	α = 0.02 R^2_{adj} = 0.34
Ridge, 5-fold CV	α = 67 R^2_{adj} = 0.42	α = 140 R^2_{adj} = 0.12	α = 135 R^2_{adj} = 0.24	α = 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

Table 3.4: Summary of the results obtained for each subindex.

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

Table 3.5: Summary of the tree methods applied on each subindex.

3.3.1 Skills Rhombus script

Finally, here we report a useful and short script written to show the rhombus of skills, figs.3.20-3.21.

```
import pandas as pd
import plotly.graph_objects as go

directory = 'C:\\Users\\Admin\\Documents\\University\\DataScience\\Statistical_data_analysis\\pro
file = 'radar20.csv'
radar_df = pd.read_csv(f'{directory}/{file}')

print(radar_df.shape)
radar_df.head(5)
```

```
(516, 6)
```

	Player	shooting	passing	dribbling	defending	Overall_index
0	Francesco Acerbi	50	61	63	86	83
1	Bobby Adekanye	56	52	64	30	60
2	Claud Adjapong	51	64	75	71	71
3	Lucien Agoume	54	59	64	46	63
4	Kevin Agudelo	58	61	68	49	63

```
variables = ['Shooting', 'Passing', 'Dribbling', 'Defending']

fig = go.Figure()

player = list(radar_df.iloc[312, :])

fig.add_trace(go.Scatterpolar(r=player[1:], theta=variables, fill='toself'))
fig.update_layout(polar=dict(radialaxis=dict(visible=True, range=[0, 100])), showlegend=False)
#fig.update_traces(textposition='bottom left')
fig.update_layout(
    height=600,
    title_text = f'{player[0]}, OVR:{player[-1]}',
    title_x=0.5
)
fig.show()
```

Figure 3.20

Figure 3.21: Skills rhombuses of S.S.C.NAPOLI's offensive trident.

Conclusions

The purpose of this project is to study how the performances of SerieA teams and players are rated by the most popular indices. In particular, because of data availability and transparency problems, we considered *SPI*, Soccer Power Index, to evaluate teams performances and *FIFA Overall index* for players skills. To do this we developed several scripts, starting from dataset creation up to the setting of non-linear statistical models.
Due to the size of the feature space, we performed a Principal Component Analysis as a preliminary step. Results are very satisfactory in the case of the SPI, indeed we got an R^2 of 0.84 with a Simple Linear Regression, using the most correlated feature (the principal component 'Capacità finalizzazione', ρ = 0.92); Multiple Linear Regression, Lasso and Ridge Regressions perform well, but without substantial differences.
In the case of the FIFA Overall index, instead, we encountered a few difficulties, so we considered it necessary to introduce more complex statistical models. Going beyond linearity, we applied a polynomial regression model (up to the third order) and a GAM, Generalized Additive Model, based on third order splines; after that, we considered Regression Trees too, with the algorithms of Bagging, Random Forest, AdaBoost, Gradient Boosting and XGBoost. Comparisons are reported in different tables, anyway the relevant results are given by GAM (with an improvement in the score from 0.36 (Multiple Linear Regression) to 0.43) and Gradient Boosting (reaching an accuracy of 94.64%).
Finally, as a further step of analysis, we considered the decomposition of the Overall index into four subindices: attack, defense, passing_types, possession. Usually the index is divided into six main skills, but we adapted the division to our case study. This point of the analysis did not lead to significant results: attack, passing_types and possession show a very slight improvement in the value of R^2_{adj}, while there is an important lack of interpretative capability in every statistical model considered with defense.

So, resuming what has been said along all the report, football is a team sport and therefore team and individual performance are tightly connected and impossible

to disentangle. Nevertheless, the aim of a good player performance rating should be to disentangle as much as possible in order to be fair with regard to all players.
Individual vs. Team Performance - A good rating system in football should be able to find a balance between the influence of the individual and the team performance on the rating. It should identify good performances of players in bad teams and bad performances of players in well performing teams. This feature of a good rating system is hard to satisfy. Even though players might play many accurate crosses, they will receive a poor rating because their teammates could not turn any crosses into goals and consequently none of the players will receive any (rating) points for assists or goals.
The identification of an index of this type would represent the turning point in statistics in the football field, it would represent the fundamental tool for building new theories and giving an in-depth reading of the dynamics of this sport. As data relating to player contributions become more sophisticated and more representative of the game, we expect further development of the index. Such data could allow complex interactions between player actions to be modelled in a way that is not currently considered in performance indices construction.