Disclaimer:

We made it fun because we know our audience.

They play with the ball
We play with the data

Cumika Anastasija
Bubenchikov Kirill
Burmistrov Pavel

Data history

RSU/Dobele futsal team

Their coach collected data manually

How did our data look like?

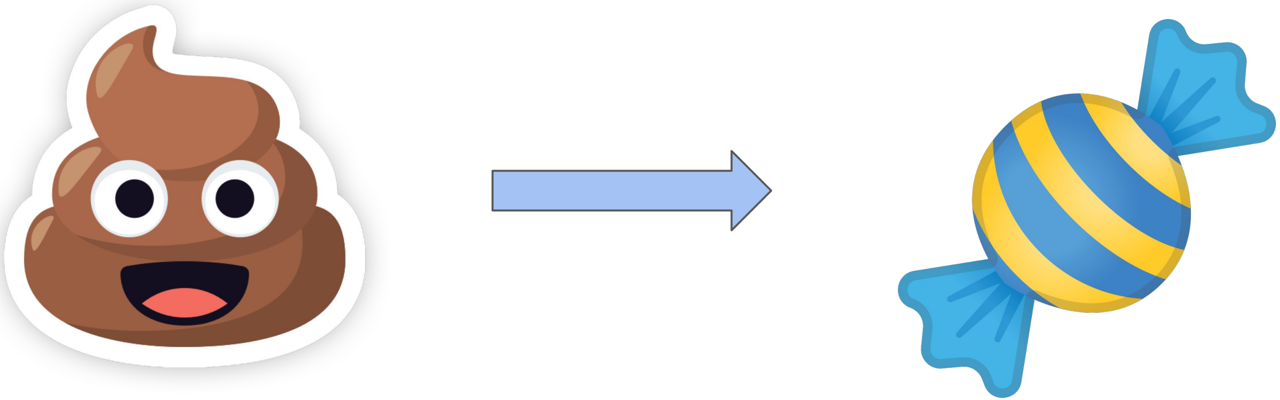
Итоговые показатели за матч.

Общие результаты команды	Удары по воротам соперников								Взятия ворот соперников								Удары по нашим воротам								Взятия ворот						+/-		
		по дистанционному удалению			по точности исполнения				по манере исполнения					по дистанционному удалению				по дистанционному удалению			по точности исполнения				по манере исполнения			по дистанционному удалению					
	всего ударов по воротам	в пределах 6 м.	6-10 метров	с дальней дистанции	в створ	в каркас ворот	мимо ворот	блокированные удары	голы с игры	голы со штрафных	голы с угловых	первые голевые пасы	вторые голевые пасы	в пределах 6 м.	6-10 метров	с дальней дистанции	всего ударов по воротам	в пределах 6 м.	6-10 метров	с дальней дистанции	в створ	в каркас ворот	мимо ворот	блокрованные удары	голы с игры	голы со штрафных	голы с угловых	в пределах 6 м.	6-10 метров	с дальней дистанции	+	-	+/-
Talsi (4:2)	39	5	15	19	15	1	12	11	4	0	0	3	1	2	2	0	48	4	20	24	18	0	18	12	2	0	0	1	1	0	4	-2	2
Liepāja (3:4)	37	10	14	13	15	1	10	11	3	0	0	1	1	3	0	0	57	12	22	23	26	0	17	14	4	0	0	1	2	1	3	-4	-1
Tukums (8:3)	28	3	18	7	14	2	6	6	7	1	0	6	2	1	5	2	62	7	34	21	18	5	24	15	2	0	1	0	2	1	8	-3	5
Nica (3:4)	43	4	22	17	9	1	15	18	1	1	1	2	1	1	1	1	51	6	27	18	23	2	10	16	3	0	1	0	2	2	3	-4	-1
Nikers (0:4)	37	4	16	17	11	1	11	14	0	0	0	0	0	0	0	0	29	9	14	6	17	1	6	5	4	0	0	3	1	0	0	-4	-4
Talsi (5:2)	50	12	22	16	17	4	15	14	5	0	0	4	1	4	1	0	37	4	17	16	13	1	13	10	1	1	0	1	0	1	5	-2	3
Liepāja (1:1)	57	7	28	22	15	2	20	20	1	0	0	0	0	1	0	0	36	5	20	11	21	1	9	5	1	0	0	0	0	1	1	-1	0
Nica (5:1)	48	8	18	22	23	1	14	10	3	1	1	3	1	3	1	1	33	3	15	15	8	0	13	12	1	0	0	1	0	0	5	-1	4
UPTK (9:3)	45	14	23	8	22	3	7	13	8	0	1	6	3	3	5	1	69	16	26	27	32	2	19	16	1	1	1	0	2	1	9	-3	6
UPTK (5:3)	45	8	21	16	19	2	12	12	5	0	0	5	3	0	2	3	39	4	13	22	14	1	14	10	3	0	0	2	1	0	5	-3	2
Nikers (7:3)	37	5	17	15	20	1	10	6	6	1	0	5	5	2	5	0	42	5	8	29	15	1	15	11	3	0	0	1	2	0	7	-3	4
LDZ (1:7)	24	3	10	11	10	0	10	4	1	0	0	1	1	1	0	0	58	4	32	22	23	0	19	16	4	1	2	1	6	0	1	-7	-6
LDZ (4:6)	42	7	20	15	16	1	12	13	3	0	1	4	3	3	1	0	45	12	22	11	20	1	15	9	5	0	1	2	4	0	4	-6	-2
Liepāja (2:1)	43	11	18	14	19	3	15	6	2	0	0	2	1	1	0	1	38	2	25	11	15	1	15	7	1	0	0	0	1	0	2	-1	1
ВСЕГО	575	101	262	212	225	23	169	158	49	4	4	42	23	25	23	9	644	93	295	256	263	16	207	158	35	3	6	13	24	7	57	-44	13
В среднем		575			575				57			42	23	57				644			644				44			44			0,9		

~x40

Data preprocessing

1. From initial data given we made a data suitable for data science analysis and ML
2. Created a target variable - (goals minus - goals plus)/goals plus
3. Scaled all the counted features to the minutes the players spent on the field (# per minute on the field).



How did our ready data look like in Pandas?

```
In [830]: data.head(n=7)
Out[830]:
```

	Name	Ball_intercept_defensive	Ball_Intercept_midfield	Ball_intercept_attacking	Ball_interccept_total	Lost_ball_dribbling	Lost_ball_recieveing_ground	Lo
0	Edgars Andrejevs	140	3	0	143	0	0	
1	Emils Dobrajs	21	1	0	22	0	0	
2	Kristaps Stankevic	30	44	2	76	20	22	
3	Kristaps Balcuns	17	22	1	40	12	13	
4	Elgar Ludborzs	21	32	2	55	9	7	
5	Pavels Zagrebins	28	51	5	84	22	33	
6	Mark Puhalskis	17	22	4	43	33	11	

```
In [831]: data.shape
Out[831]: (16, 72)
```

What do we want to do ~~today~~ with this data?

1) Find the interesting correlations between statistics

Why?

To find what actions on the field are correlated and we didn't expect them.

How can help?

To see on what aspects of play coach should work on to increase the performance of the players.

What do we want to do today?

2) Predict players efficiency factor of the player based on the statistical data

Why?

When coach is choosing/looking for a new player he wants to find the most efficient one (not always the goal scorer) based on the statistics given or collected from the previous games.

How can help?

With machine learning coach could discover great players that might not seem as appealing to others and build better team for less money.

WDWWDT?

3) Classification of the players based on their statistical data (play[illegible]ers/ holding players, individualists/team players)

Why?

For coach it is important to know type the player is so the team can perform the needed taktics at the needed time

How can help?

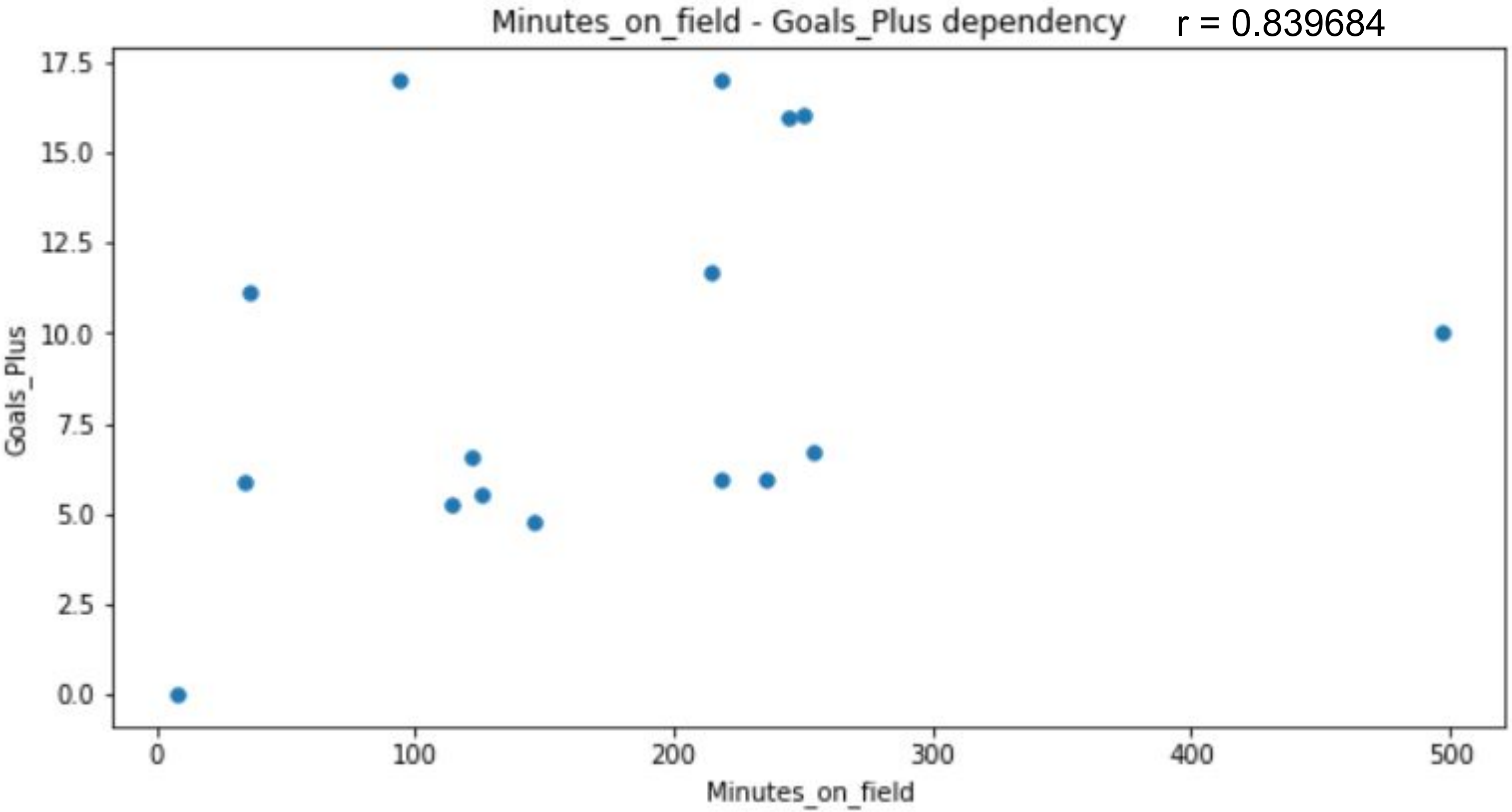
When coach has a clear classification of players in his team he can increase the chance of success by the correct choice of the players for the specific moment or opponent.

Correlations

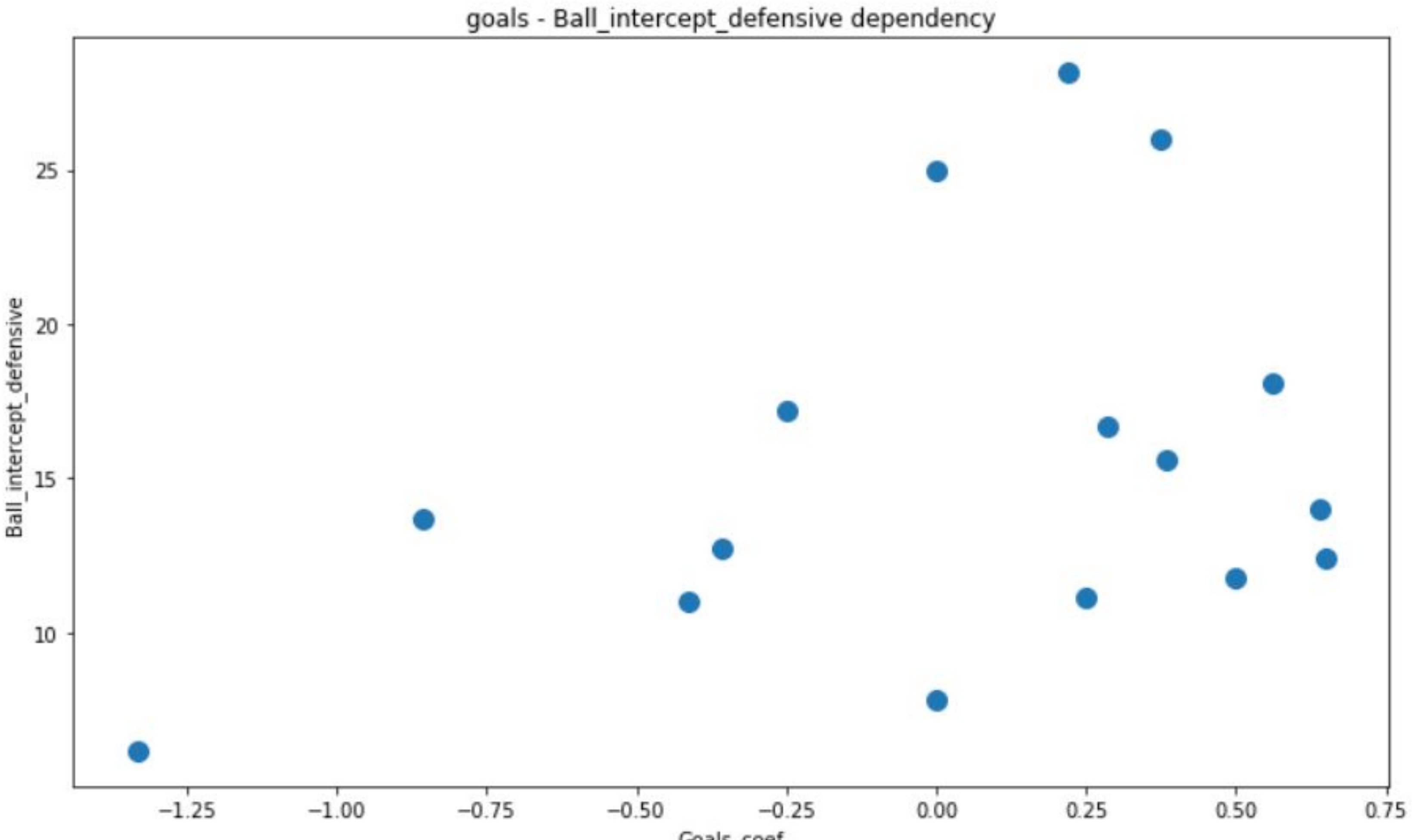
Working with the initial data, we investigated the possible dependencies in the dataset.

An independent woman - does not correlate with anyone.

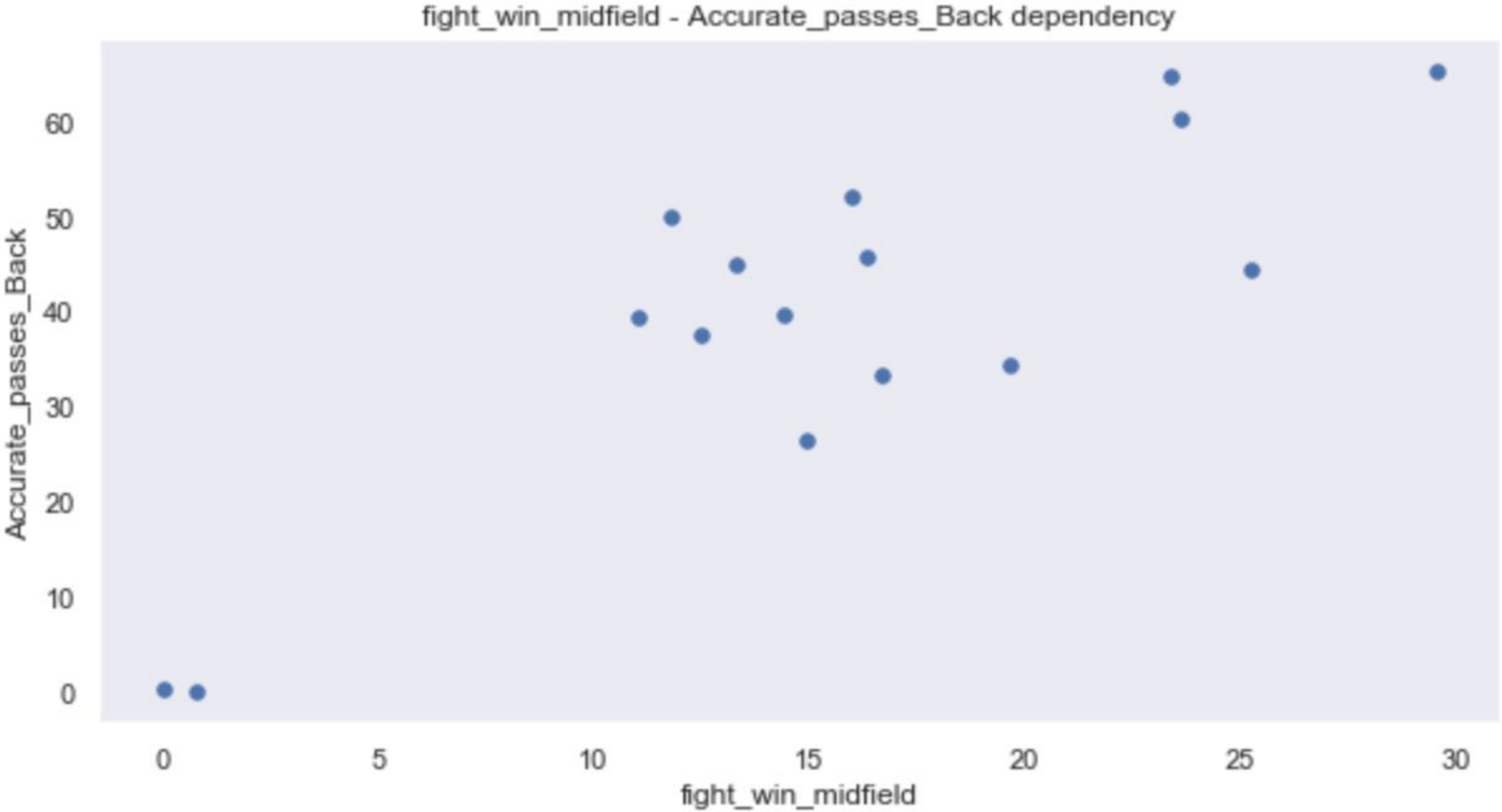
More Time on the field more goals scores on average! Good choice coach!



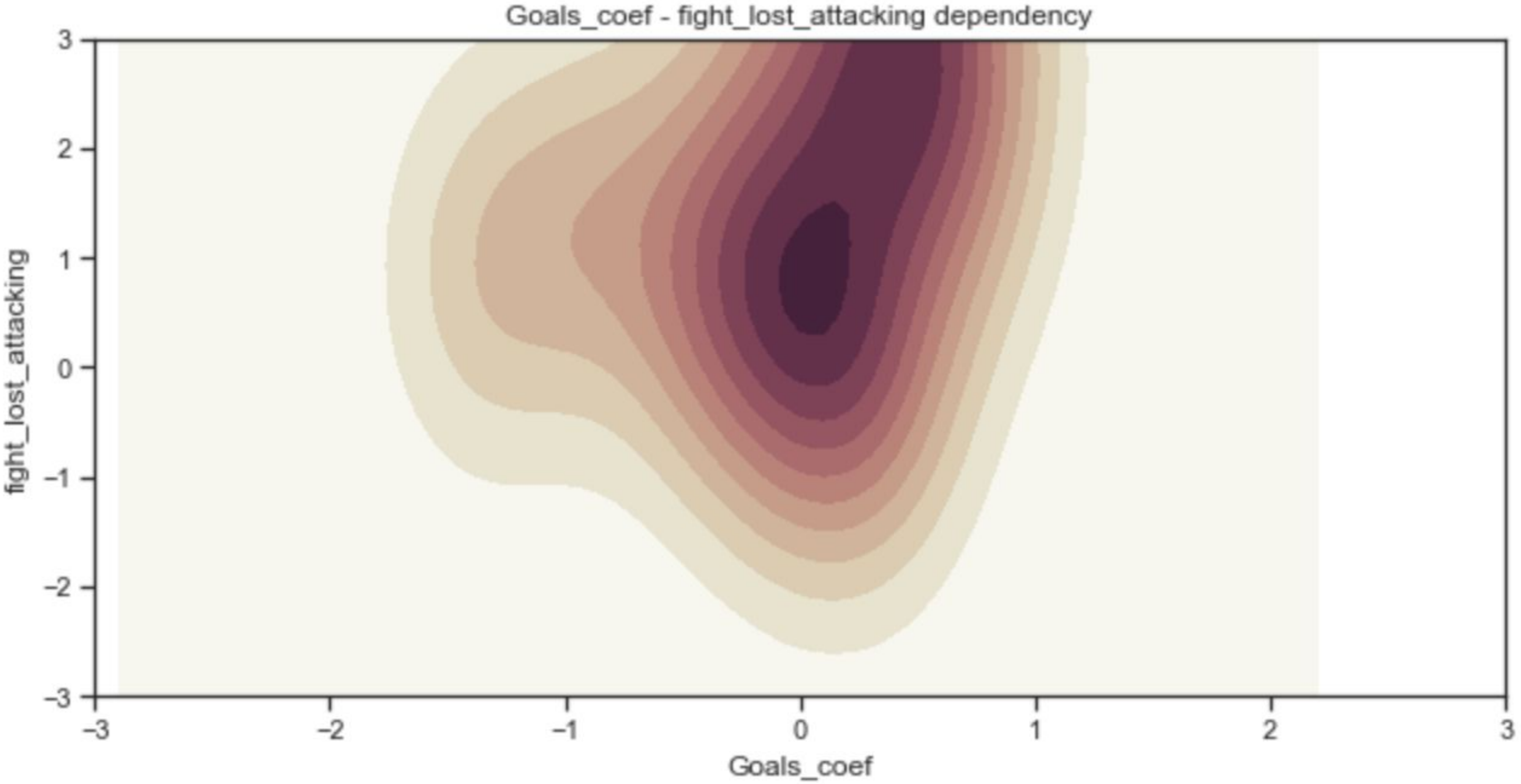
BEING ACTIVE IS GOOOD



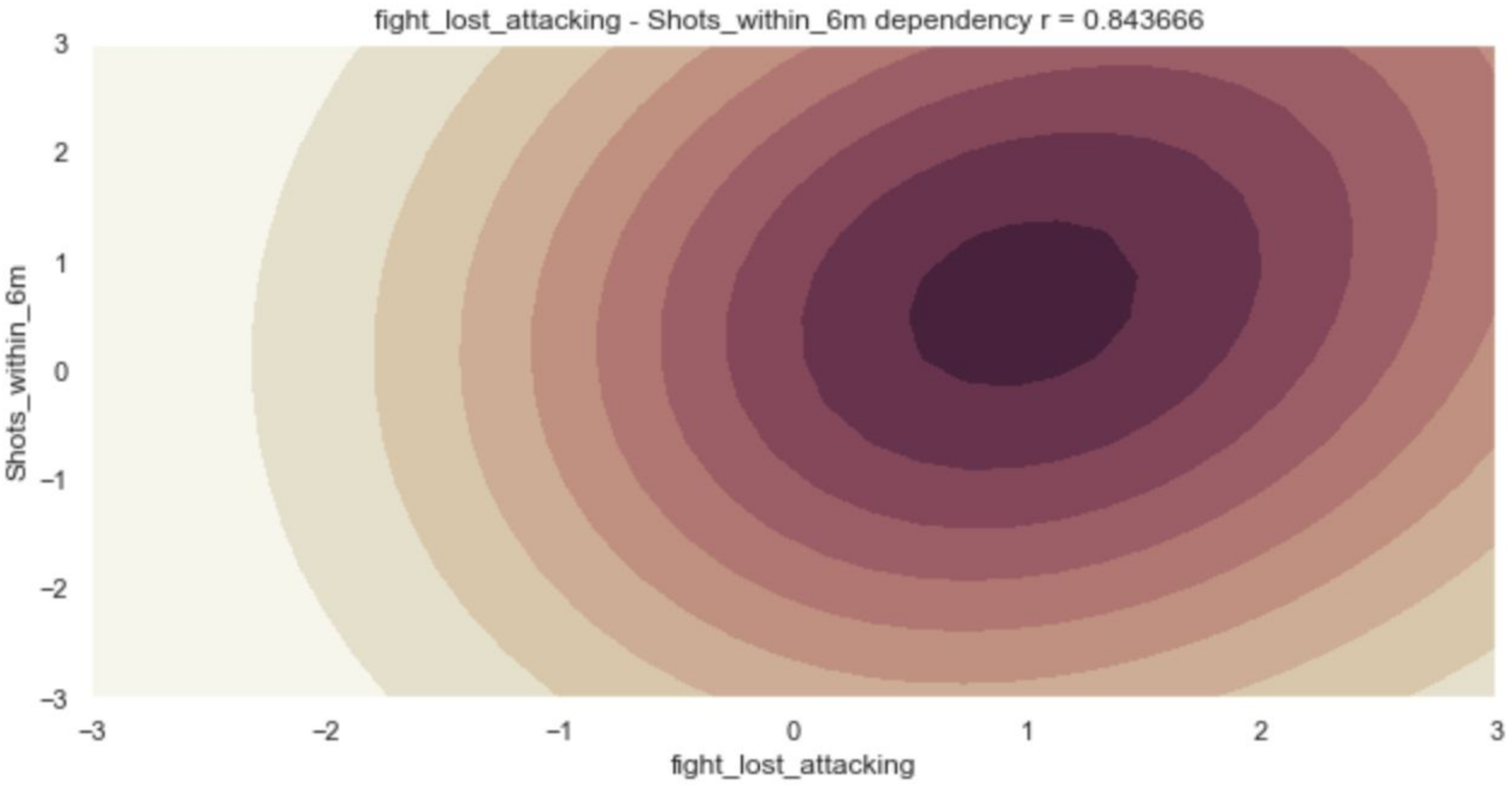
Win ball pass it back so team keeps it!



Just the opposite of what we expected….



Active players engage in close intermissions a lot even if they do not win



Predicting efficiency of players (ML)

1) Choosing model and metrics
2) 1st trial
3) 2nd trial
4) 3rd trial

   .

   .

   .

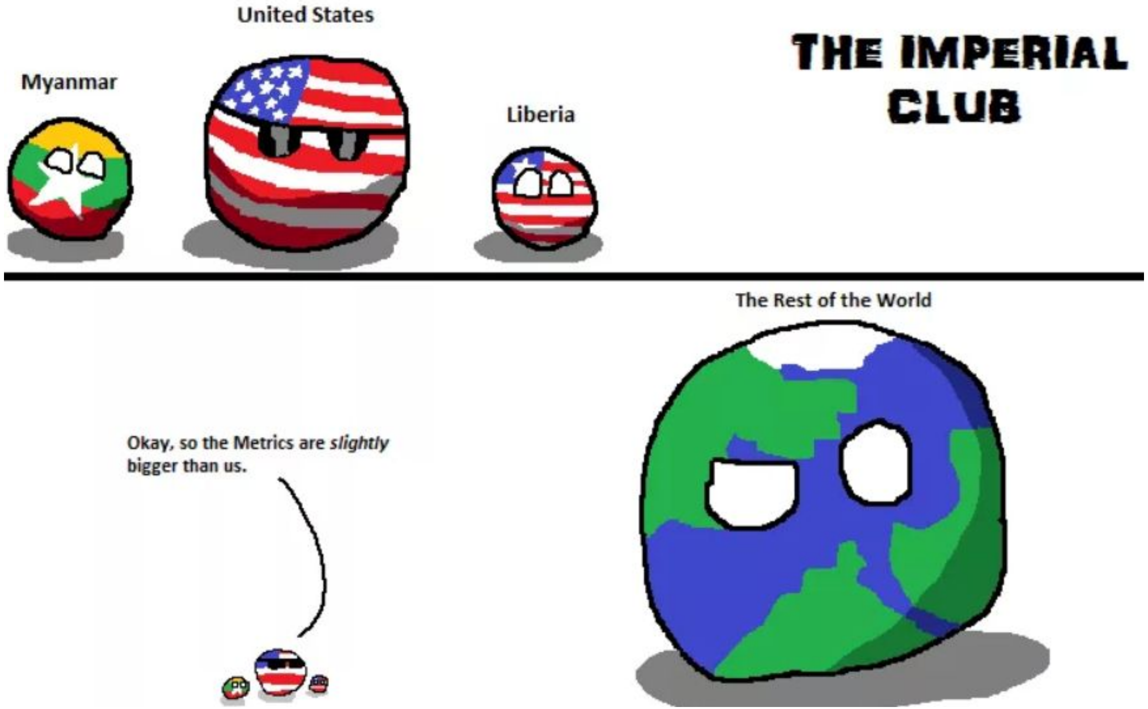
5) PROFIT

BAM!

Model and metrics selection

Lasso because many features

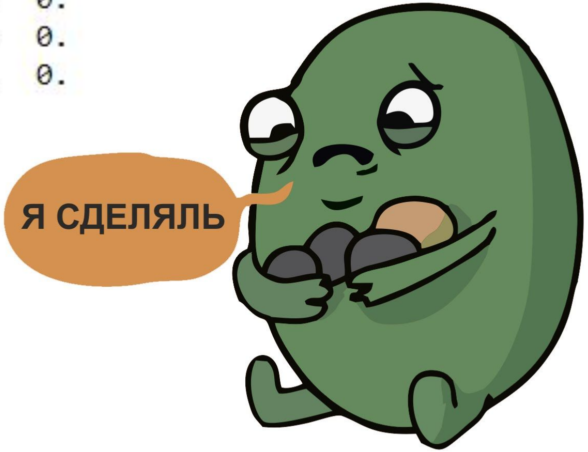
MSE and **R2** - good metrics for estimating regressions



1st trial

We found alpha with the help of GridSearchCV, we used it for Lasso.
Also we had LeaveOneOut().
We got overfitting. That’s all you have to know...

```
1.2954154975029346e-30 0.3286483440947148
[ 0.   0.   0.   0.   0.   0.  -0.   0.  -0.   0.   0.   0.   0.   0.   0.   0.  -0.   0.
  0.   0.   0.  -0.  -0.  -0.  -0.  -0.   0.   0.  -0.  -0.  -0.   0.  -0.   0.  -0.   0.
  0.  -0.   0.   0.   0.  -0.  -0.  -0.   0.  -0.   0.  -0.  -0.   0.   0.   0.   0.   0.
  0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.  -0.  -0.]
```



2nd trial

1. We made our table smaller by dropping out some columns.
2. We used GridSearchCV
3. We used Lasso
4. We looked at the results
5. Guess what..?


OVERFITTING
OVERFITTING EVERYWHERE
LIGHTYEAR
risovach.ru

3rd trial

Successful success!

DOUBLE

BAM!


BUT!!!
meme-arsenal.ru

Out[803]:

Name	Lost_ball_total	Total shots	first_assist	Shots_blocked.1	Goals_coef	passes_total	Accurate_passes_total_percent	fight_win_total_percent
Edgars Andrejevs	0.000000	0.000000	0.000000	0.000000	0.220000	1.006036	0.866000	1.000000
Emils Dobrajs	0.000000	0.000000	0.000000	0.000000	0.285714	0.547619	0.898551	1.000000
Kristaps Stankevic	0.237288	0.059322	0.008475	0.059322	-0.357143	2.364407	0.856631	0.500000
Kristaps Balcuns	0.137615	0.146789	0.004587	0.041284	0.000000	2.027523	0.825792	0.558442
Elgar Ludborzs	0.155738	0.131148	0.000000	0.065574	-0.250000	2.549180	0.729904	0.540541
Pavels Zagrebins	0.295276	0.208661	0.003937	0.074803	-0.411765	1.625984	0.920097	0.636364
Mark Puhalskis	0.500000	0.223404	0.042553	0.053191	0.562500	2.648936	0.771084	0.508772
Arturs Mahitarjans	0.412000	0.488000	0.016000	0.072000	0.375000	2.620000	0.842748	0.664865
Alberts Mahitarjans	0.274590	0.295082	0.036885	0.090164	0.384615	2.758197	0.827637	0.565476
Edgar Strautins	0.348624	0.605505	0.032110	0.073394	0.648649	2.128440	0.808190	0.620690
Shota Giorgadze	0.210526	0.078947	0.017544	0.061404	-1.333333	1.859649	0.750000	0.530612
Zanis Pinka	0.266355	0.266355	0.028037	0.112150	0.640000	3.294393	0.792908	0.475410
Andrejs Kravcenkovs	0.222222	0.166667	0.027778	0.083333	0.250000	1.416667	0.725490	0.480000
Raimonds Pavulins	0.375000	0.375000	0.000000	0.000000	0.000000	2.375000	0.947368	0.666667
Nikolas Petriga	0.294118	0.176471	0.029412	0.029412	0.500000	1.823529	0.822581	0.538462
Aleksandrs Radcenko	0.184932	0.205479	0.020548	0.095890	-0.857143	3.349315	0.803681	0.521127

In [804]:
```
data.shape
```

Out[804]: (16, 8)

What happened in the 3rd trial?

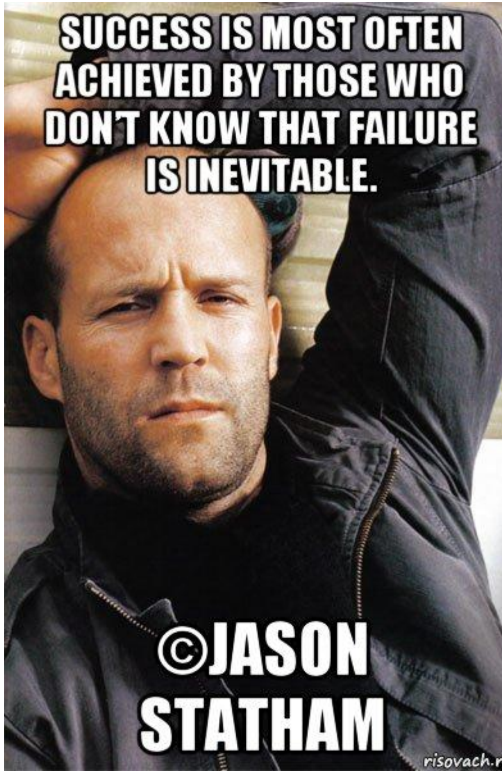
+

- We won the battle with overfitting (MSE_train = 0.1603… instead of MSE_train = (10) ^ (- 100500)

-

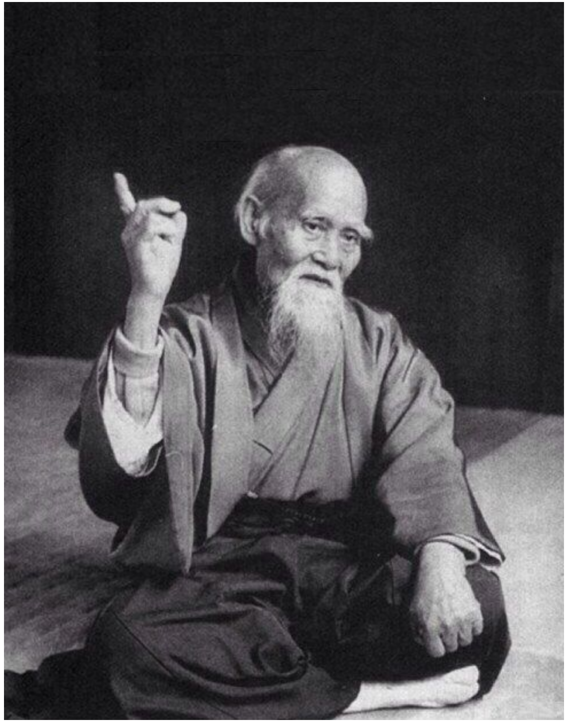
- Such a small amount of data does not allow to train the model well
- In this regard, big errors arise

```
0.16034687389307029 1.3636641958718072
[0. 0. 0. 0. 0. 0. 0.]
```



Lessons **we** learned from it:

- **Experience with live data**, not with Kaggle abstract compilation
- A dataset **may not always be ideal**
- **Many features and few samples are not the best decision** when you decide to do a project on DS.
- Do not rejoice **until you made fit on test**



Classifications

Individuals and team players

Attacking and holding players.

Model and Metrics Selection

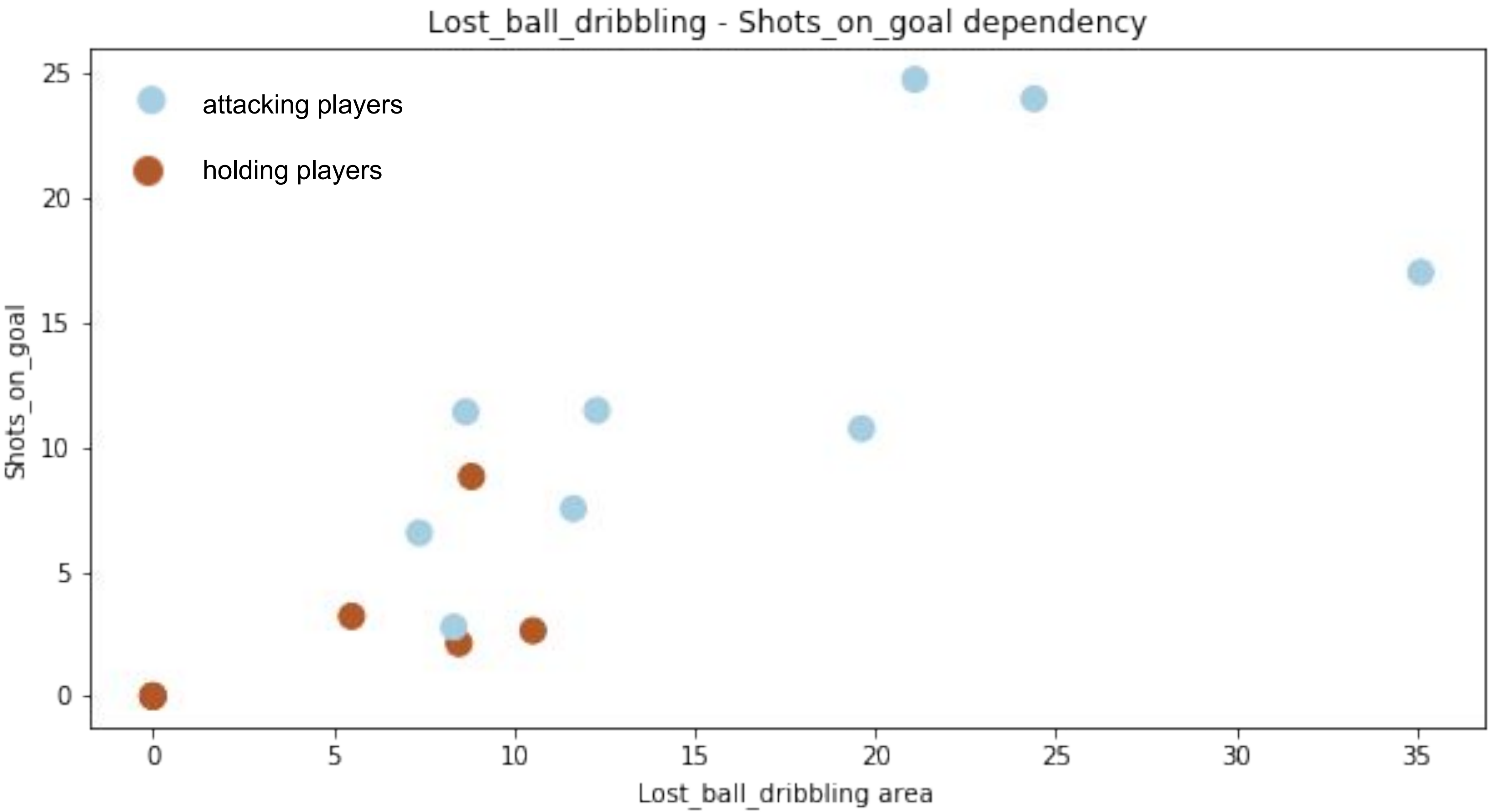
Model

Clustering : KMeans

Metrics

V_measure_score
mutual_info_score

Attacking and holding players



Attacking and holding players.

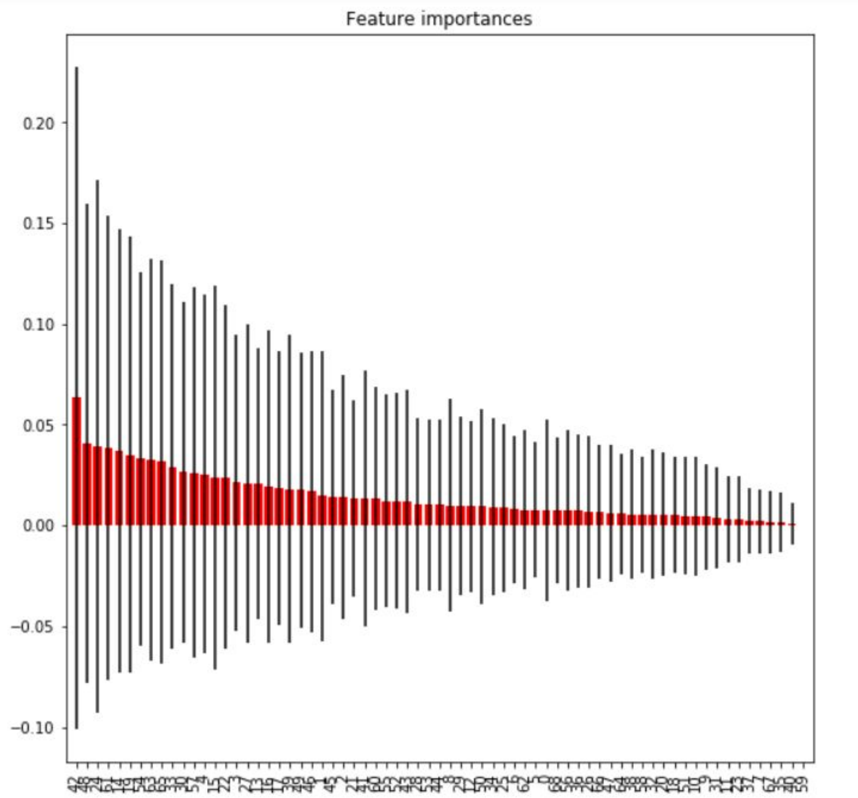
Name	0	real_lb_AttDef
Edgars Andrejevs	1	1
Emils Dobrajs	1	1
Kristaps Stankevic	1	1
Kristaps Balcuns	1	1
Elgar Ludborzs	0	0
Pavels Zagrebins	1	1
Mark Puhalskis	0	0
Arturs Mahitarjans	0	0
Alberts Mahitarjans	0	0
Edgar Strautins	0	0
Shota Giorgadze	1	1
Zanis Pinka	0	0
Andrejs Kravcenkovs	1	0
Raimonds Pavulins	0	1
Nikolas Petriga	1	1
Aleksandrs Radcenko	0	0

KMeans
n=2

```
Accuracy 0.875
v_score 0.4564355568004039
mutual info 0.3163770193035087
```

Try to find best features with ML

ExtraTreesClassifier




```
InAccurate_passes_Medium (6-10)
InAccurate_passes_Total
Inaccurate_passes_MidfieldZone
Second_assist
fight_win_attacking
fight_lost_total
Shots_on_goal
Goal_medium_distance
Shots_blocked.1
InAccurate_passes_across
```

KMeans:

```
Accuracy 0.75
v_score 0.3437110184854506
mutual info 0.2157615543388356
```

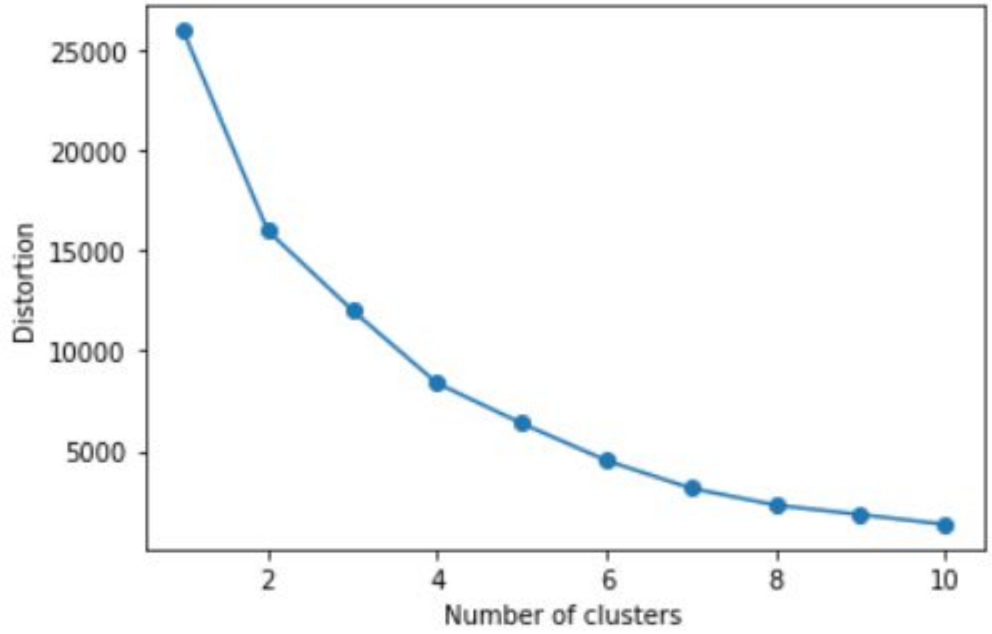
Attacking and holding players

Name	0	three_clust
Edgars Andrejevs	2	2
Emils Dobrajs	2	2
Kristaps Stankevic	1	1
Kristaps Balcuns	1	1
Elgar Ludborzs	0	0
Pavels Zagrebins	1	1
Mark Puhalskis	0	0
Arturs Mahitarjans	0	0
Alberts Mahitarjans	0	0
Edgar Strautins	0	0
Shota Giorgadze	1	1
Zanis Pinka	0	0
Andrejs Kravcenkovs	1	0
Raimonds Pavulins	0	1
Nikolas Petriga	1	1
Aleksandrs Radcenko	0	0

KMeans
n=3

```
Accuracy 0.875
v_score 0.633233990452569
mutual info 0.6169692189162668
```

Attacking and holding players



No clear elbow, so I stay with n=2 or n=3

Individuals and team players

Name	0	real_lb_passer
Edgars Andrejevs	1	1
Emils Dobrajs	1	1
Kristaps Stankevic	0	1
Kristaps Balcuns	0	1
Elgar Ludborzs	0	0
Pavels Zagrebins	1	1
Mark Puhalskis	0	0
Arturs Mahitarjans	0	0
Alberts Mahitarjans	0	0
Edgar Strautins	0	0
Shota Giorgadze	1	1
Zanis Pinka	0	0
Andrejs Kravcenkovs	1	1
Raimonds Pavulins	0	1
Nikolas Petriga	1	1
Aleksandrs Radcenko	0	0

KMeans
n=2

```
Accuracy 0.8125
v_score 0.4507077010176628З
mutual info 0.30352401849214955
```

Individuals and team players

GaussianMixture: covariance_type='spherical’, n=2

```
Accuracy 0.8125
v_score 0.313046932243041
mutual info 0.215761554338835
```

ExtraTreesClassifier choosen features
KMeans n=2

```
Accuracy 0.9375
v_score 0.720990991431914
mutual info 0.496929126648239
```

Conclusions

- Our data set is small.
- Unsupervised clusterization worked pretty well.
- Also this the coach said that this data set could be improved with giving each opponent team weight based on their position in the league table. And based on that weight we give weigh to the statistics collected.
- Everyone likes kitties (even if they are not in the dataset)


Thank you very much
Good luck
God bless you and be
happy!