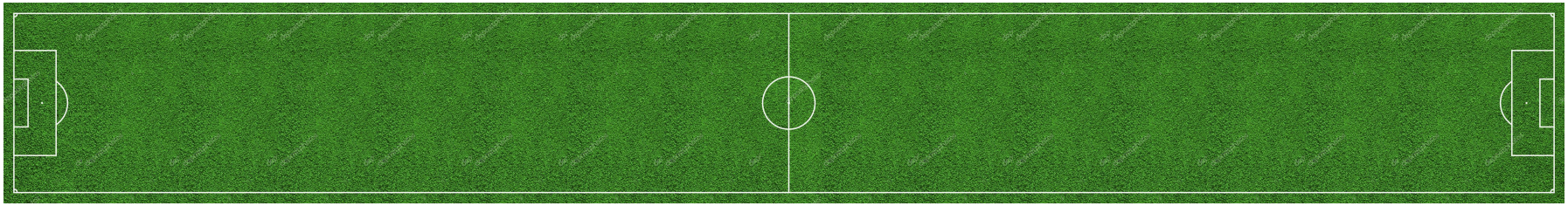


Balling or Bawling

Ranking Football Players Using Data

A project by
Deniz Adıgüzel
Francesco Ragaini
Sanvi Kirloskar
Niels den Besten

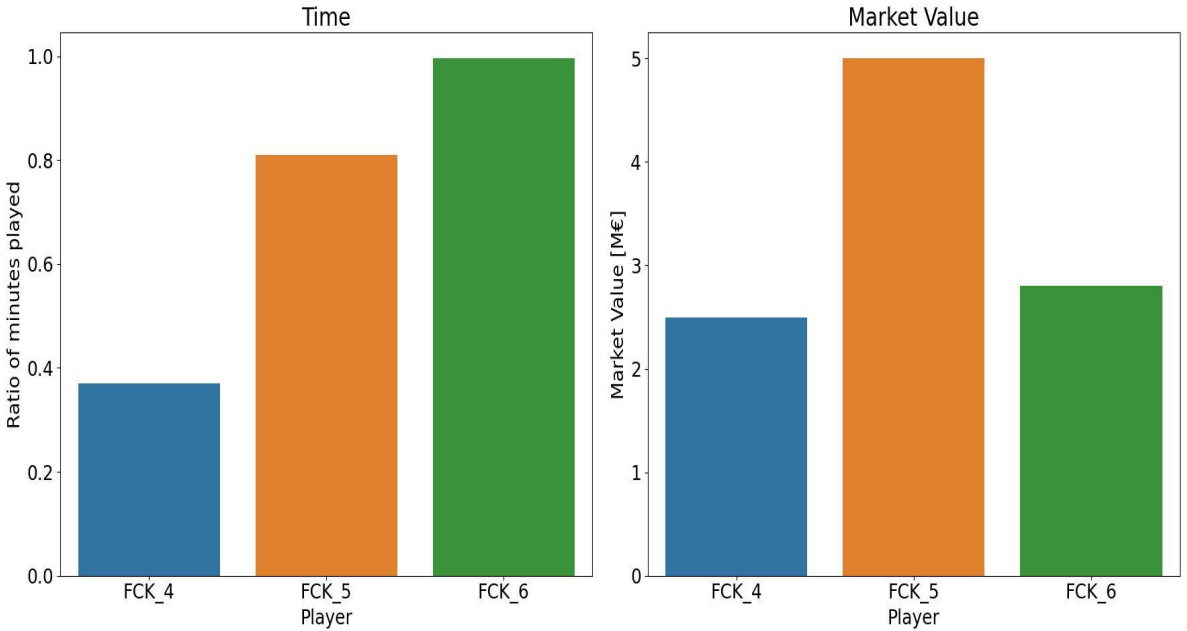
Case study: Best first season in FCK defence

Contenders:

- #4 Munashe Garananga
- #5 Gabriel Pereira
- #6 Pantelis Hatzidiakos*

*on loan

What does the coach see?



Data Structure - Games

Time	Ball	Home-team	Away-team
	x,y,z,v,in-play	x,y,z,v,jersey-number	

25 Hz 90 min game

~135000 datapoints per game

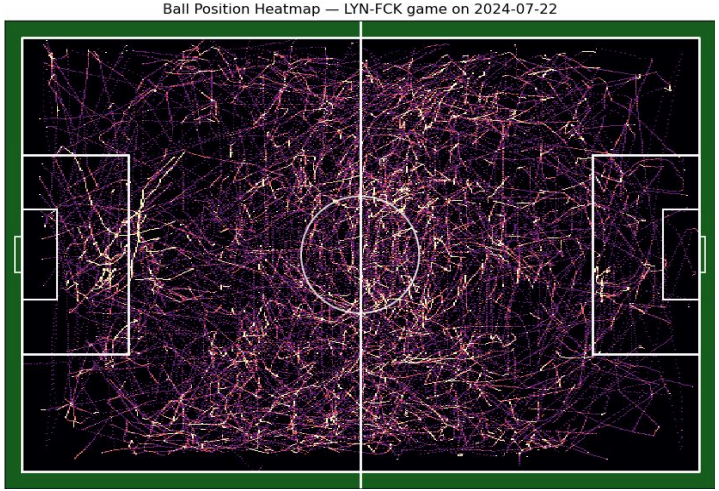
All games from Danish football league season 24-25

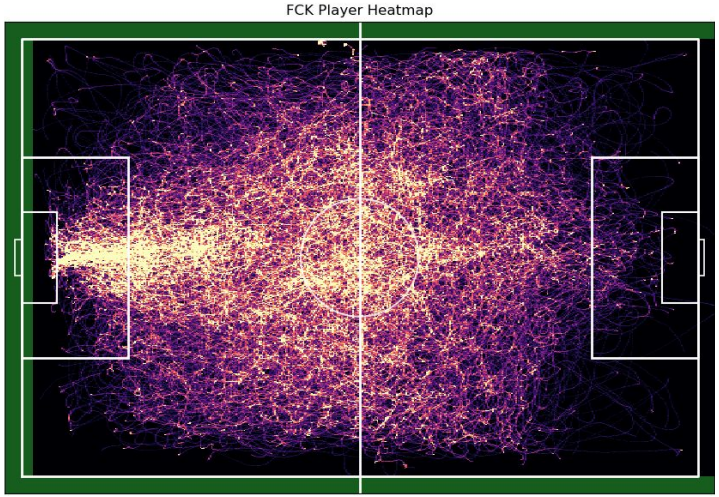
-includes teams: `AAB, AGF, BIF, FCK, FCM, FCN, LYN, RFC, SIF, SJE, VB, VFF`

-12 x 30 Games => 360 games in total, 180 unique games

-Cards, Lineups, Subs, Passes, XG-data included as well

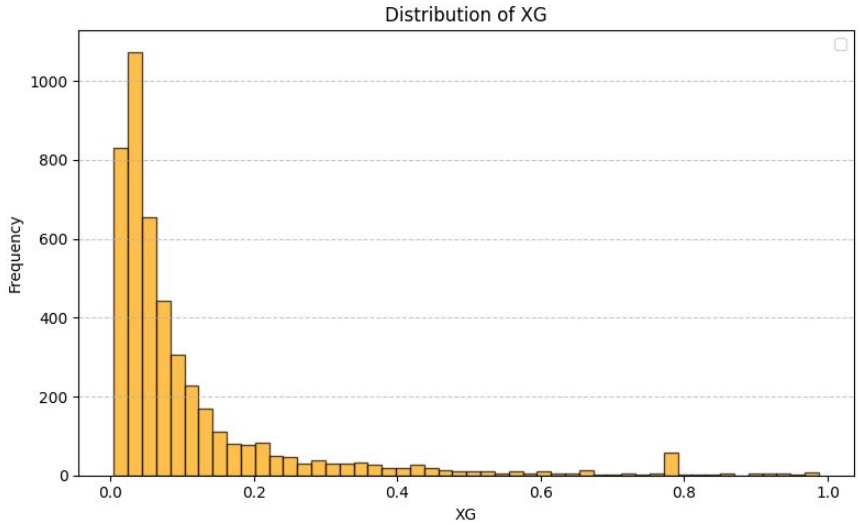
~60GB raw data





Data Structure - XG

The expected goal (XG) is the likelihood of a shot being scored.

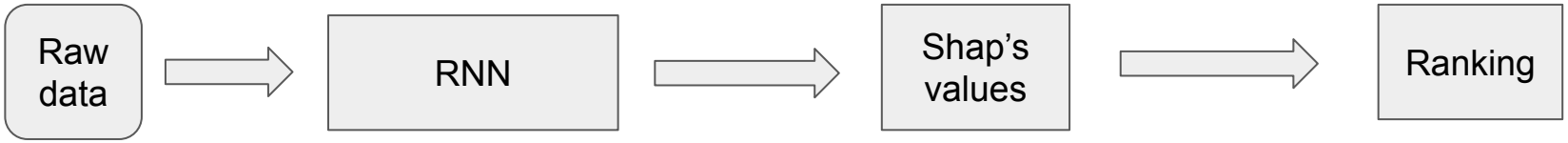


XG-value, halftime, min, sec, shooter, team, x,y, goal, penalty

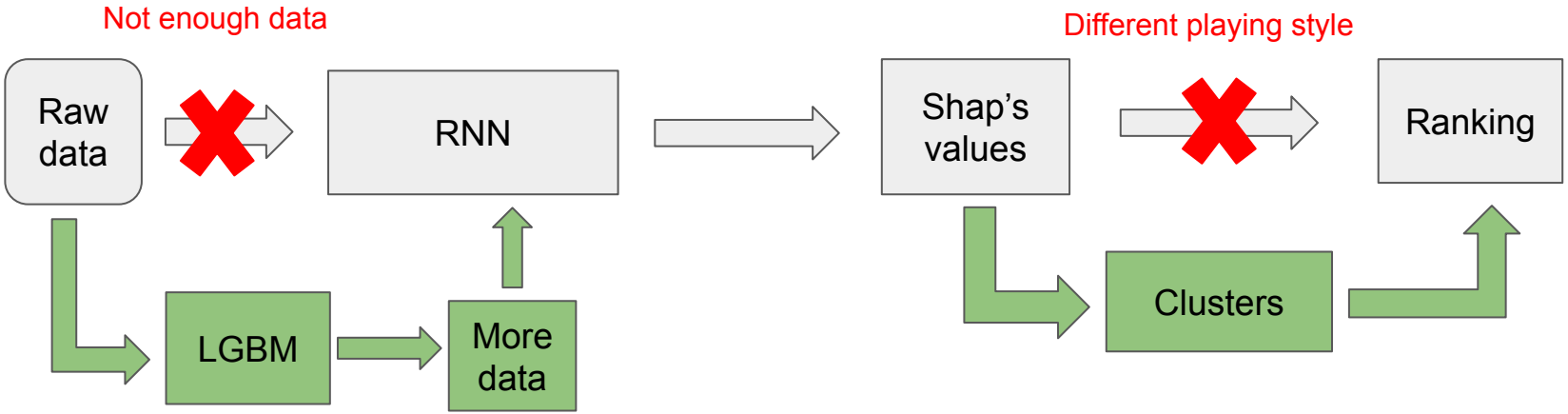
~25 datapoints per game (12 per team)

Roadmap -XG

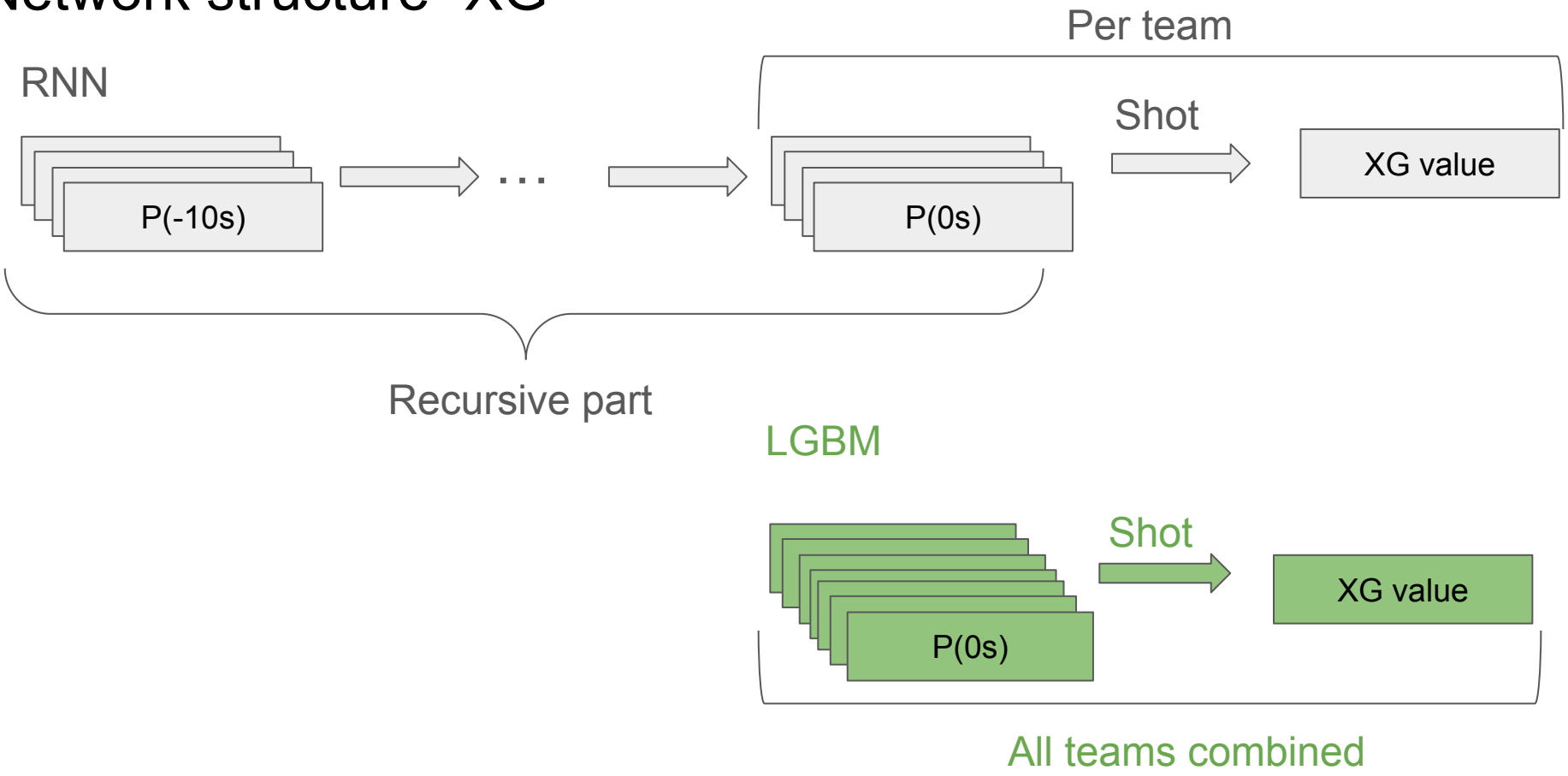
Envisioned Approach



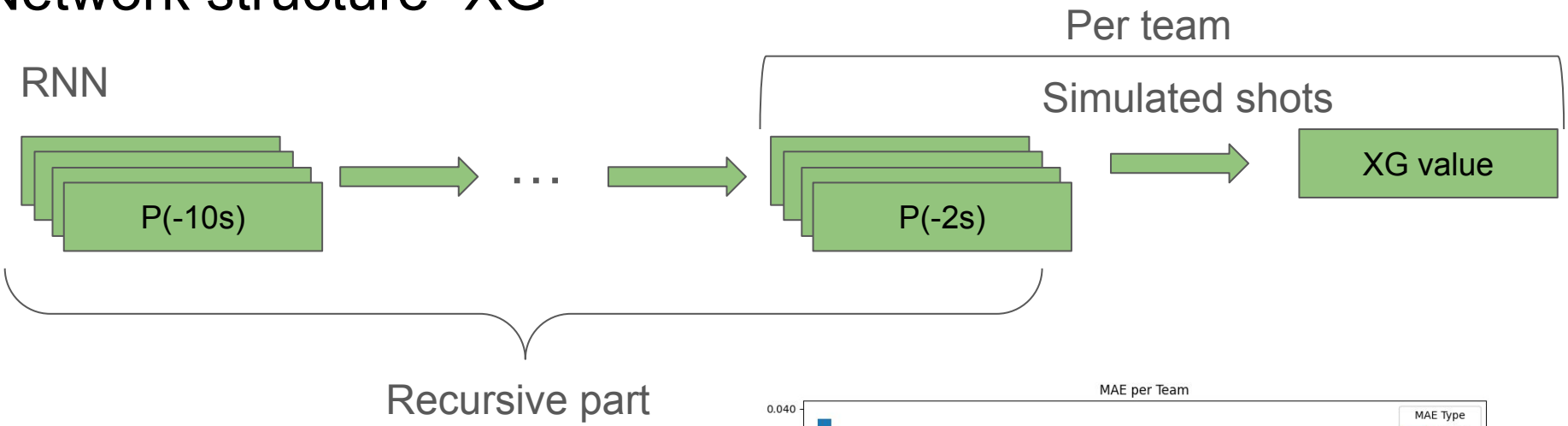
Realized Approach



Network structure -XG

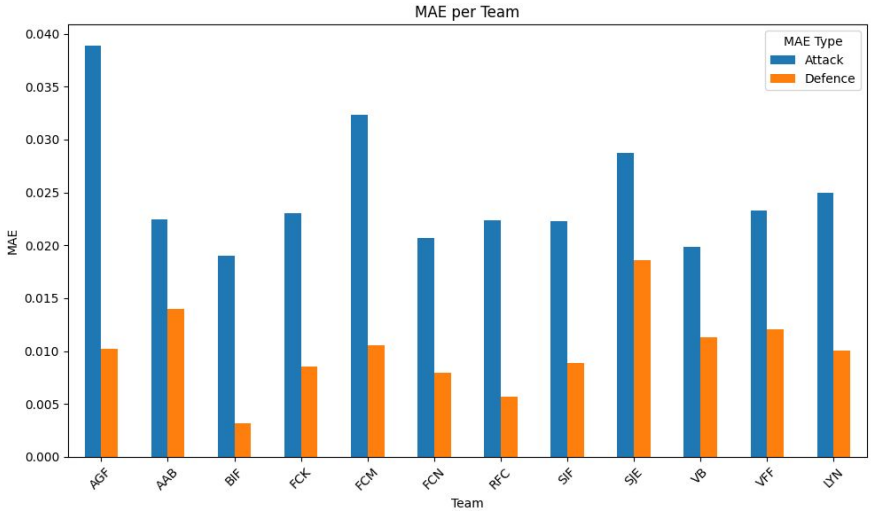


Network structure -XG

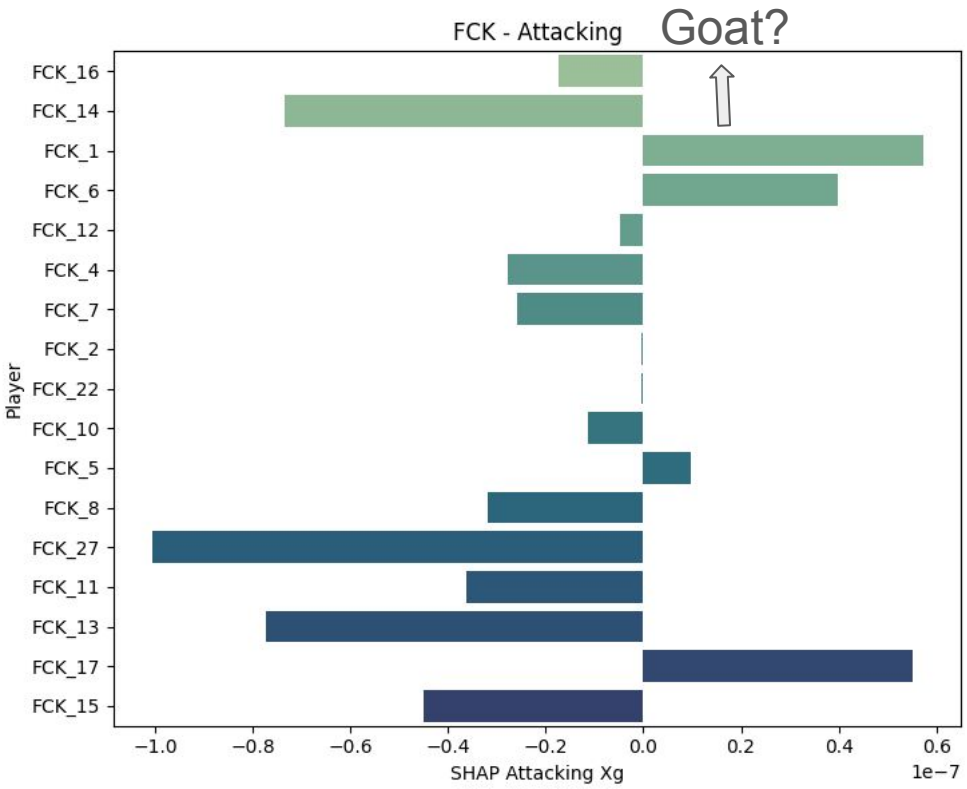


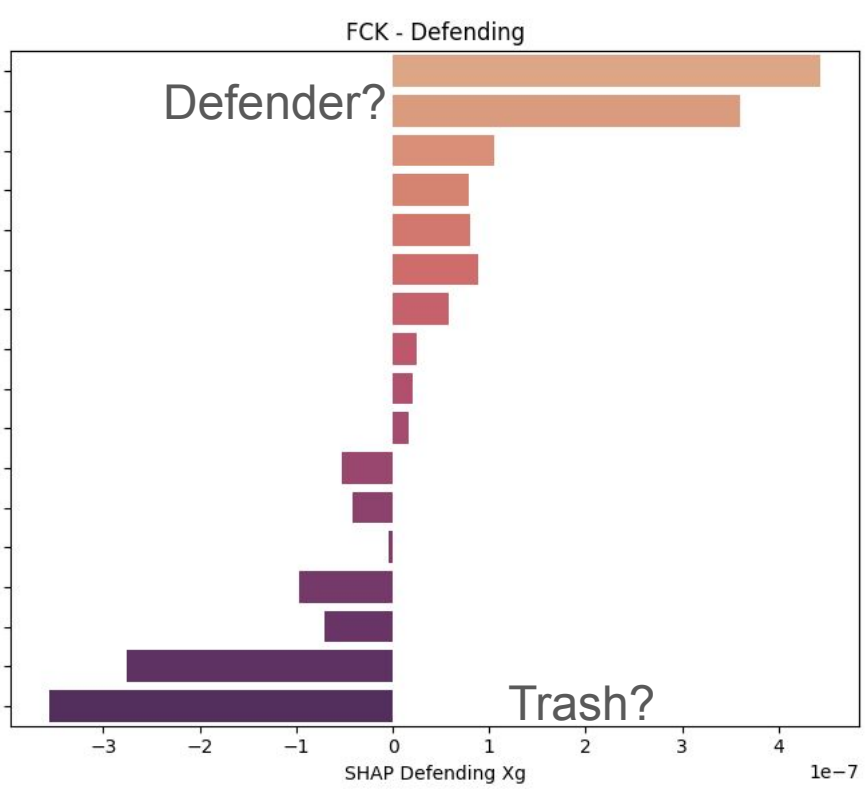
Key points:

- Free from LGBM
- Two RNN per team (offensive and defensive)
- Defensive and offensive contribution
- Just keys seconds



FCK results


FCK - Attacking
Goat?
FCK_16
FCK_14
FCK_1
FCK_6
FCK_12
FCK_4
FCK_7
FCK_2
FCK_22
FCK_10
FCK_5
FCK_8
FCK_27
FCK_11
FCK_13
FCK_17
FCK_15
Player
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
SHAP Attacking Xg
1e−7

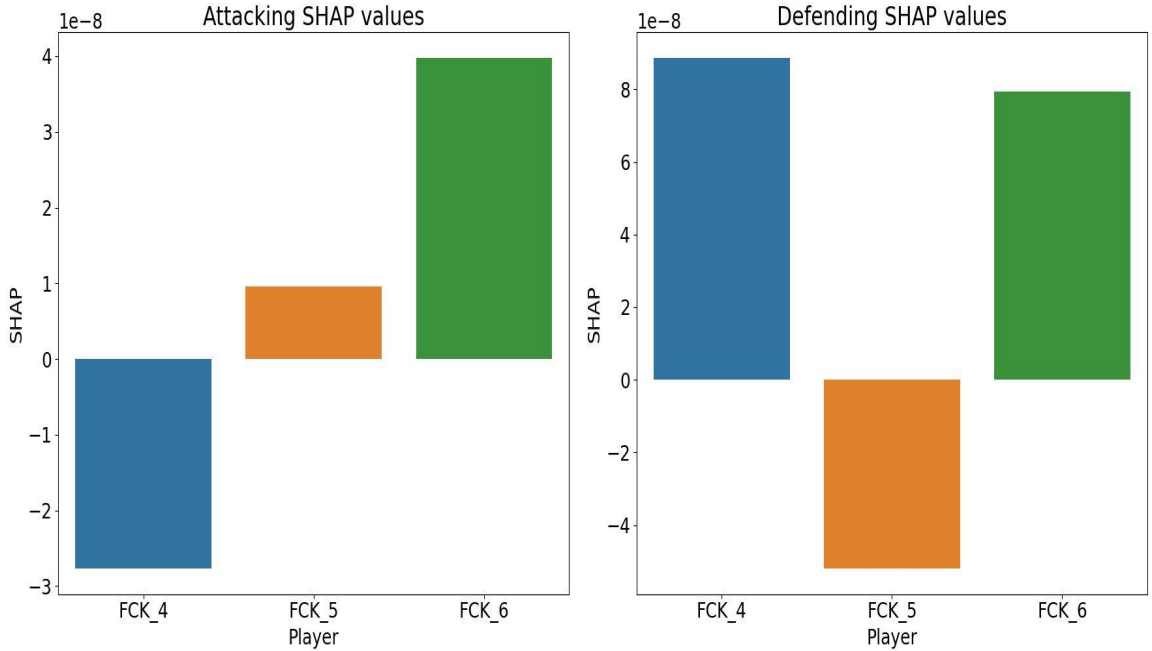

FCK - Defending
Defender?
Trash?
−3
−2
−1
0
1
2
3
4
SHAP Defending Xg
1e−7

Case study: Best first season in FCK defence

Contenders:

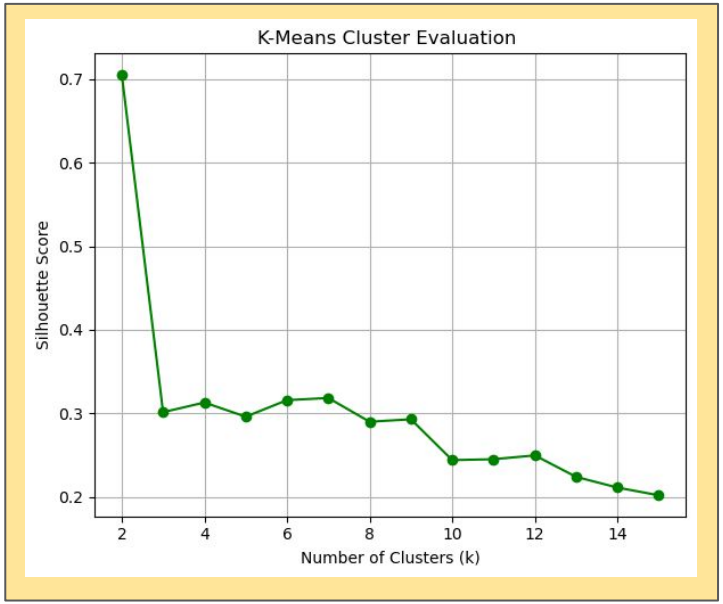
- #4 Munashe Garananga
- #5 Gabriel Pereira
- #6 Pantelis Hatzidiakos*

What do the RNNs see?



Players in Clusters

- Clustering players according to the (x, y, v) ± stds
- Players positions are normalized with respect to their team's average position per game!
- K-means algorithm was used to cluster.

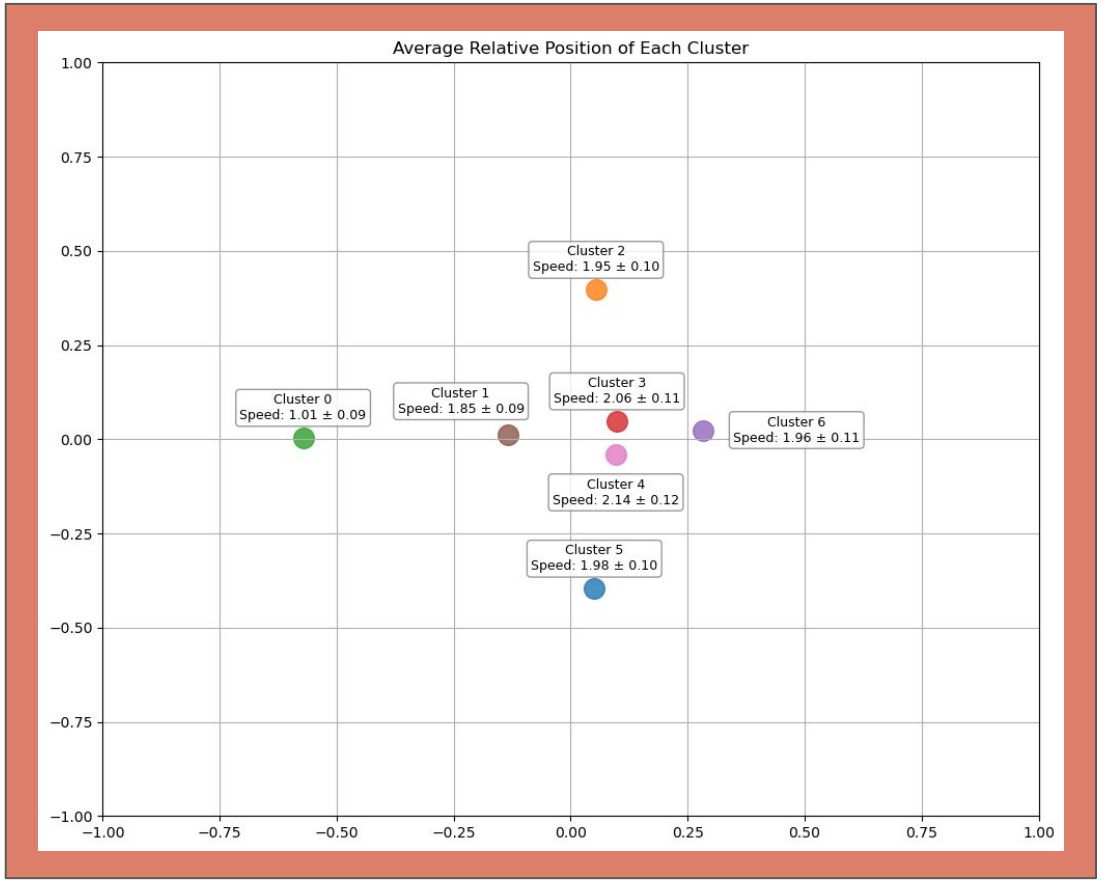


Clustering the players

Clusters can be labeled as:

0 - Goalkeeper
1 - Defence
2 - Left wing
3 - Left Midfielder
4 - Right Midfielder
5 - Right wing
6 - Attack

What about positions like right back, left back and others?



Regressing the players

- ➔ **Idea:** Train on defense, midfielder, attack clusters (5,3,6,4) to later predict player behaviour in the left and right wing positions.
- ➔ **Approach:** Train a regression model by assigning role scores (0 = defense, 0.5 = midfield, 1 = attack) using player features (x, y, v,) ± stds

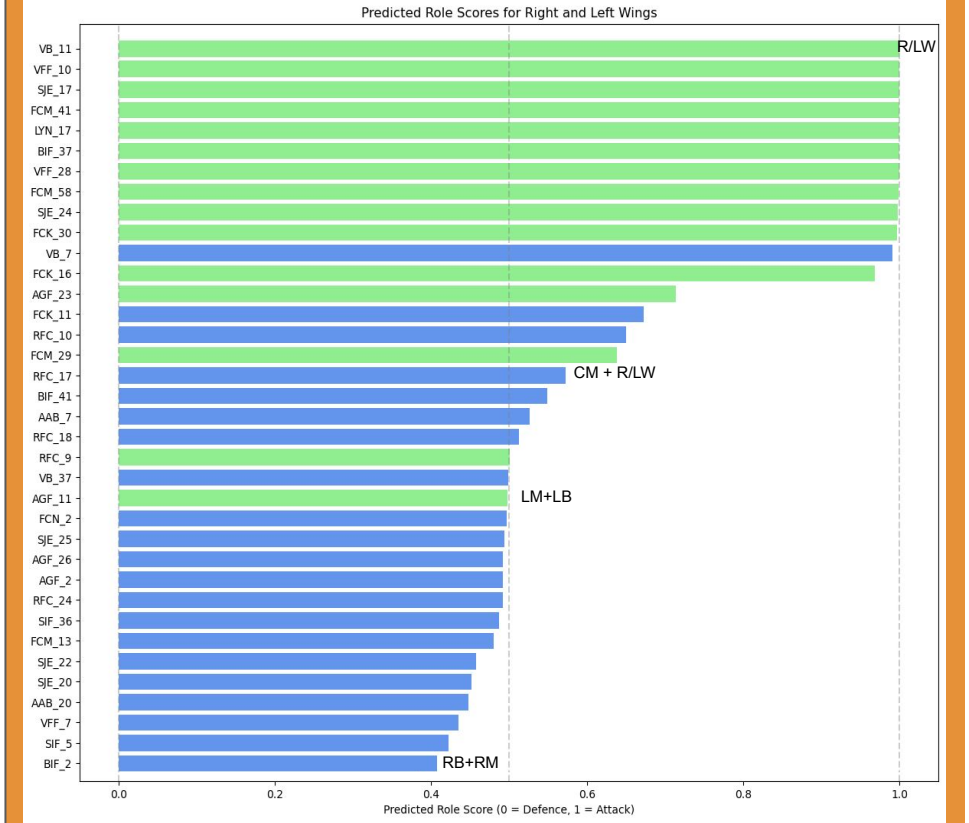
Algorithm: Tensorflow sequential
Key HP values: Nhidden1=32, Nhidden2=16, LearningRate=0.01
HP optimisation: Adam optimizer, Sigmoid output, MSE loss

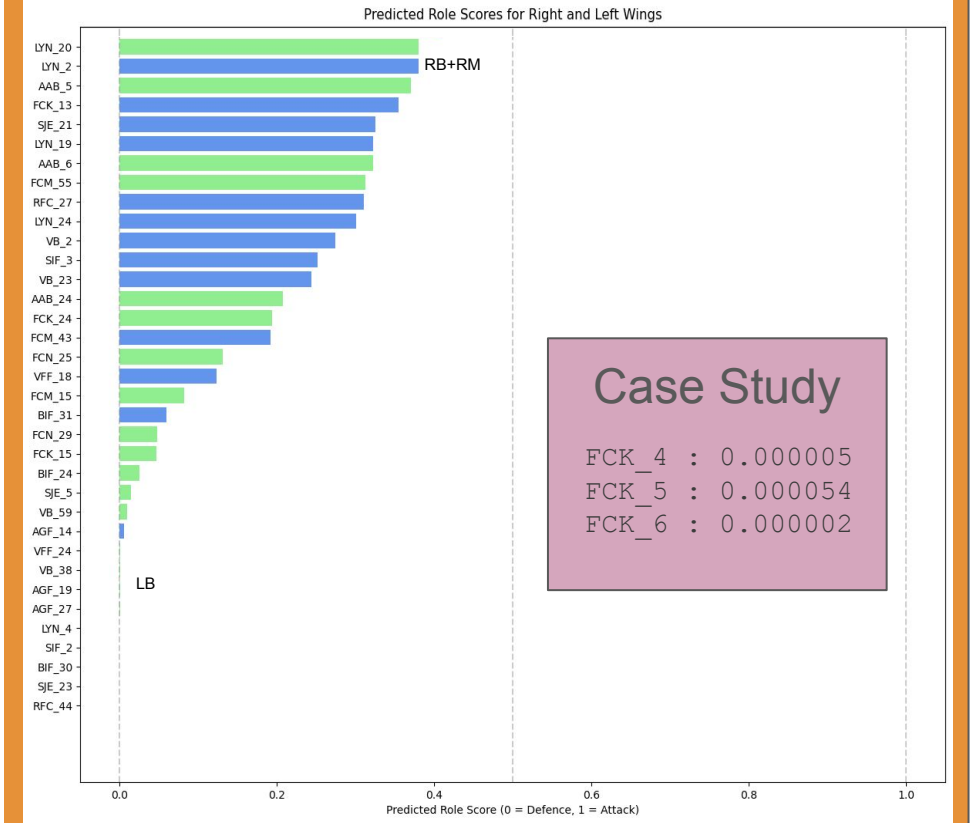
Performance:

```
Test MAE   : 0.0373
Test RMSE  : 0.1032
Test R²    : 0.9041
```

Regressing the players

Blue: Right wing, Green: Left wing



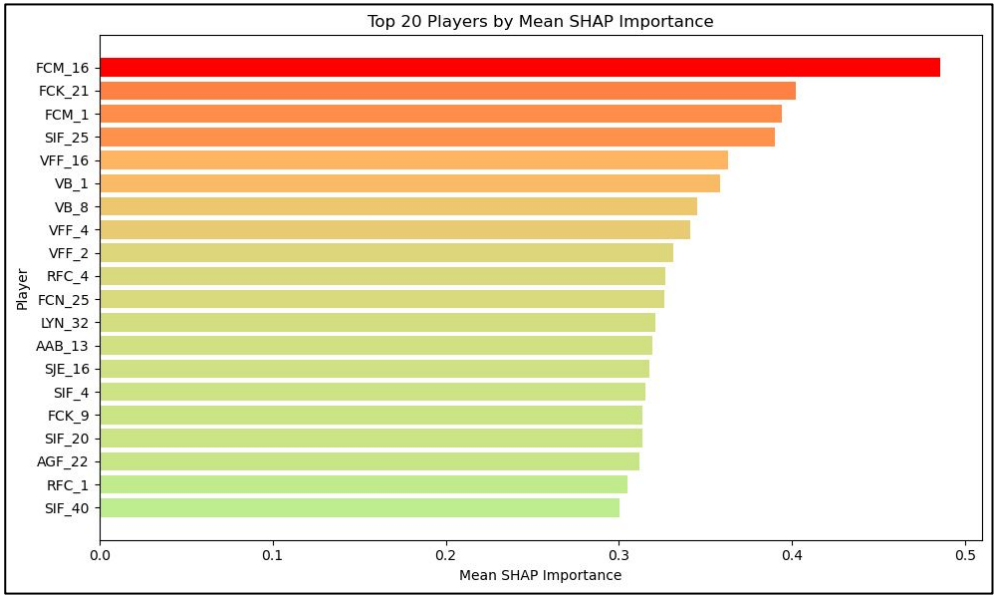


Case Study

```
FCK_4 : 0.000005
FCK_5 : 0.000054
FCK_6 : 0.000002
```

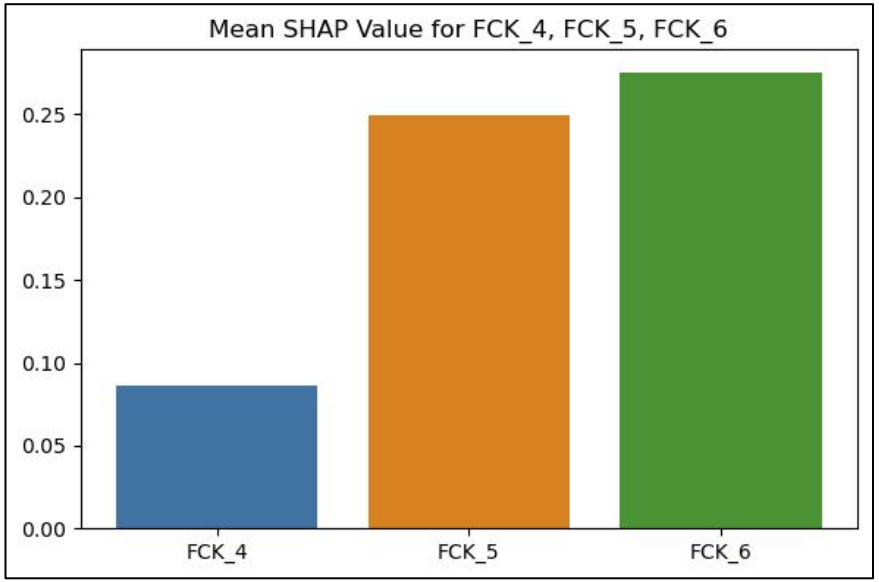
The Possession Route

- Used RNN to find ball
- Found distances of players from ball
- Found SHAP values of each player



The SHAP values tell us how important each player was to positioning the ball

Case Study



- #4 Munashe Garananga
- #5 Gabriel Pereira
- #6 Pantelis Hatzidiakos*

In conclusion: The FCK coach should buy Player 6

Balling!

Position Tracking

Question: Can we create an algorithm capable of predicting the future?

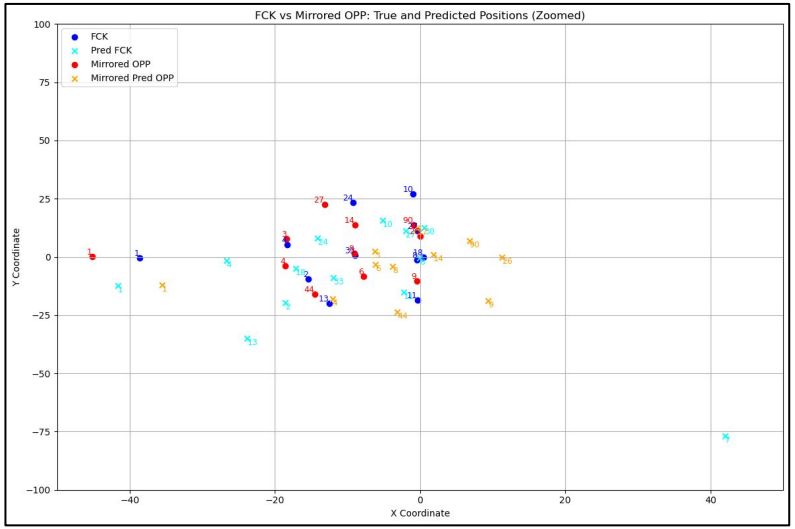
2 Methods:

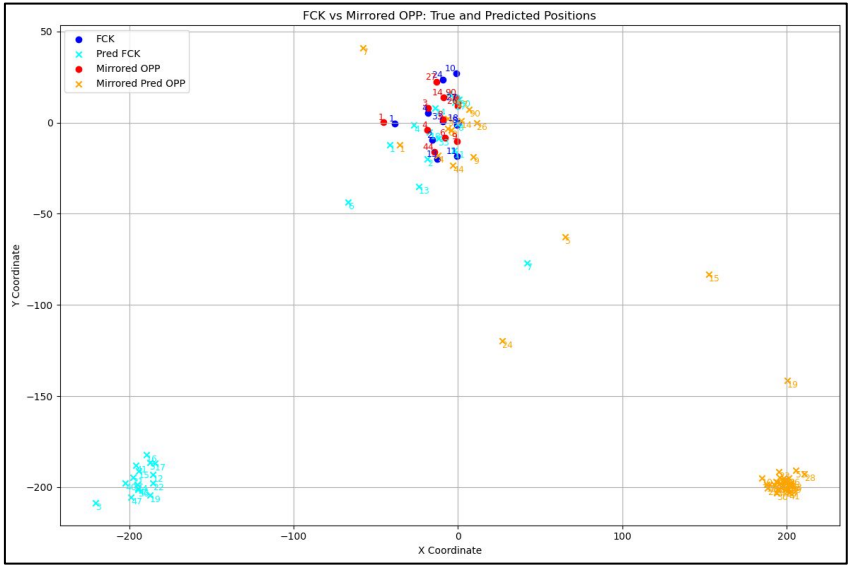
- RNN (GRU) to determine starting positions of each player
- Seq2Seq to predict the future positions of all players

RNN: 3x GRU (64) layers
Optimiser: ‘adam’
Loss: ‘mse’

Seq2Seq: GRU(64), RepeatVector, GRU(64)
Optimiser: ‘adam’
Loss:’mse’

Position Tracking

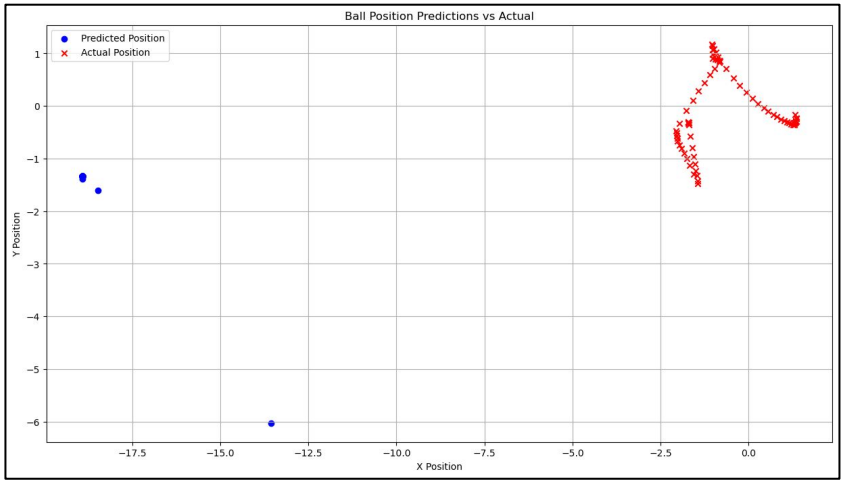




Comparison of the predicted vs actual positions of players at the beginning of the test game - GRU

Average RMSE: 72.82

Autoencoder Result



Relative RMSE: ~5000

Bawling

Why do the results look like this?

- Lack of games per team
- Substitutions
- Different lineups per game
- Ineffective masking

Starting lineup for test game (RFC/FCK)

Starting lineup for one of the training games (BIF/FCK)

Summary

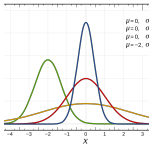
The data can be applied to increase team's performances.

Data is dirty, it takes a lot of work to make them ready to be used.

Dealing with humans is difficult and it is so to label them

Future works

More data will lead into better predictions for positions



Using a foundation model for LGBM.

New data to cluster and to regress, in particular for midfield

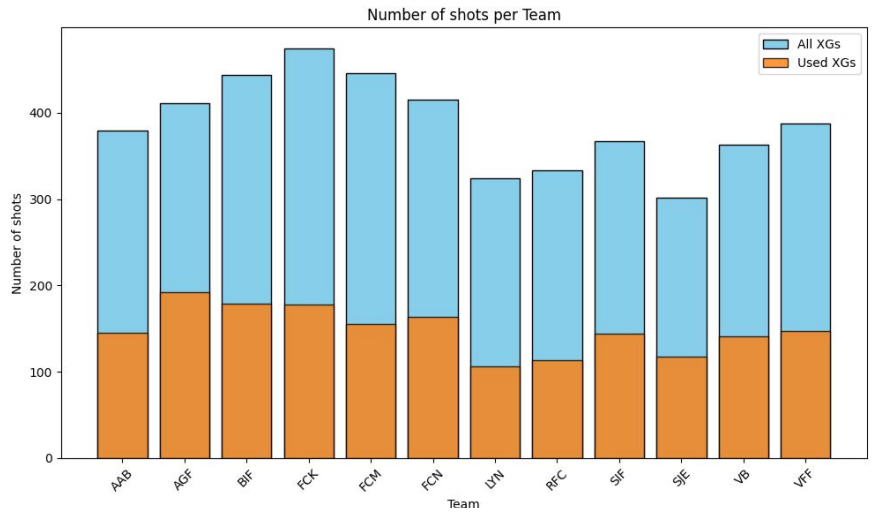
Appendix

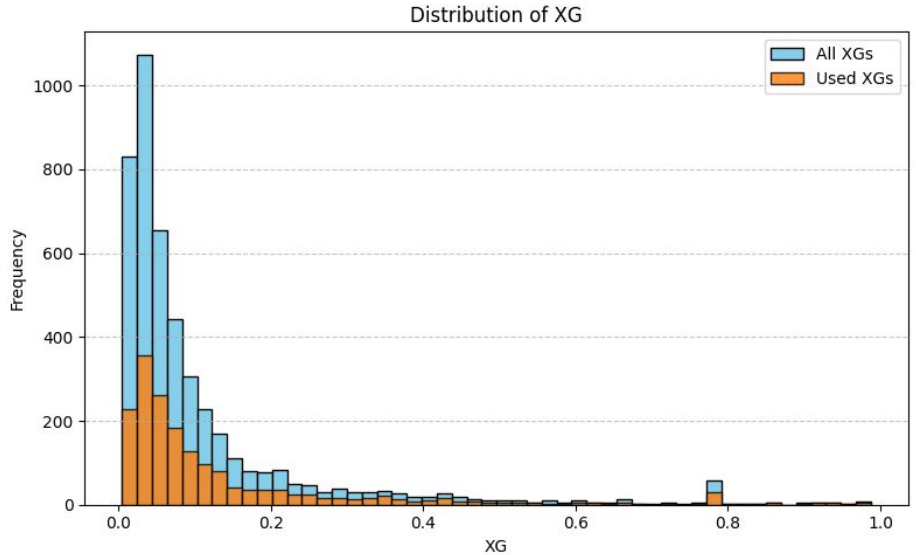
Motivation/Goal

- Ranking players in order of their importance to the game
- Predicting player roles (e.g., goalkeeper, midfielder) based on their position and speed
- Comparing actual on-field behaviour with assigned roles to evaluate if player fit their positions
- Being able to predict the positions of the ball and any player at any point in the game

Lack of data for RNN

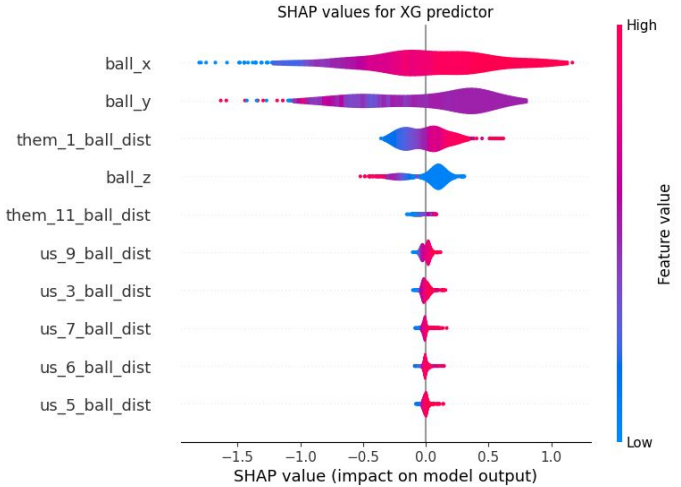
- The data available for this project was 180 games in the Danish Superliga 2024/25 season.
- This translates to 4631 non-NaN XGs.
- To use these values, the approximate frame of shooting had to be found, but the dataset proved very inaccurate.
- XG used if:
  1) Ball is in play
  2) Shooter is in dataset
  3) Shooter is within 2m of ball
  4) Time of frame is within 2 seconds of shot
- These requirements leave 1781 XGs.
- This is too little data, especially as the RNN is trained per team.

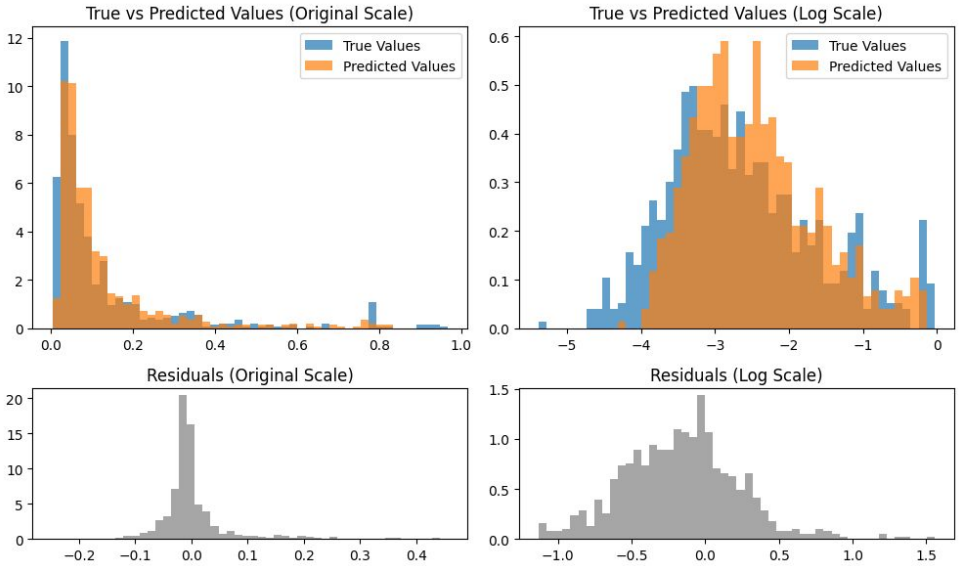




LightGBM

- To increase the available data, we trained our own XG model, using LightGBM.
- Optimization included feature augmentation (excluding speeds, xy->euclidian distance to ball), and predicting the log(XG) instead of XG directly.
- This algorithm was then used on moments during the game where a shot was not taken, but could have been taken.





FCK SHAPS

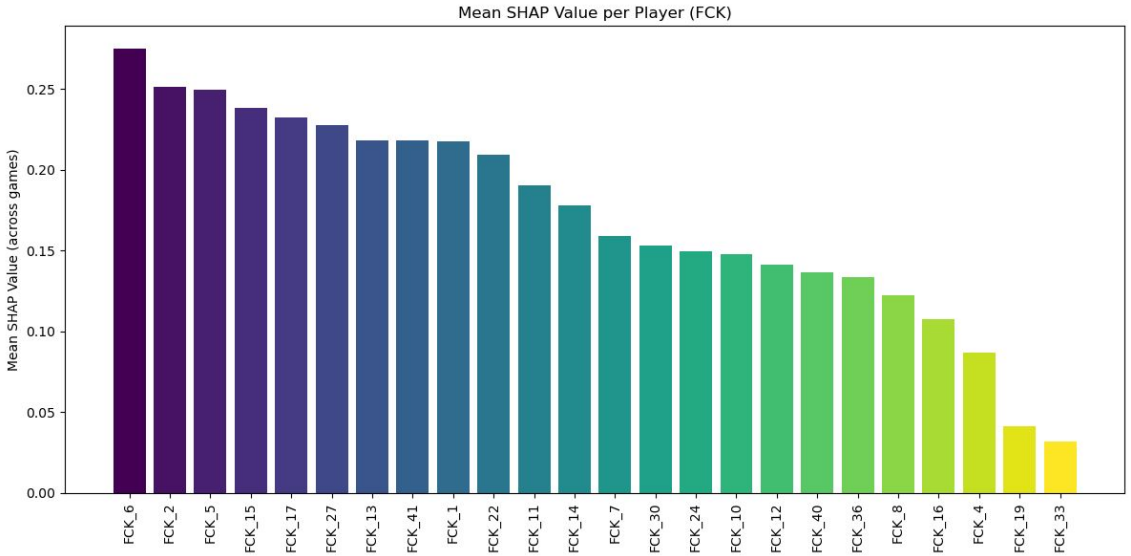
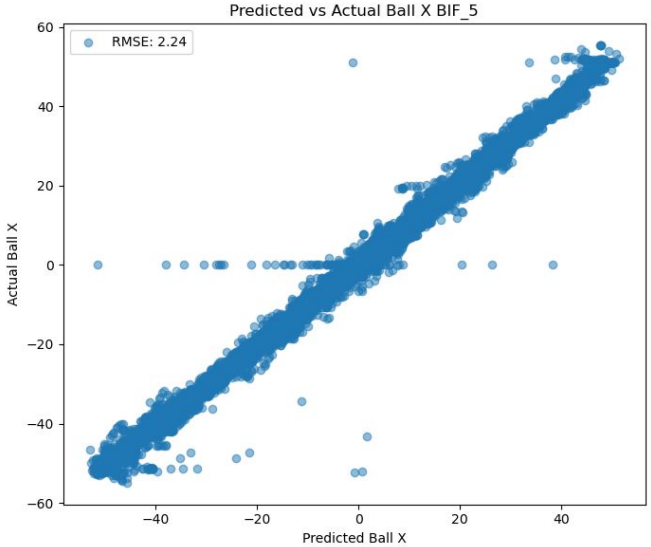
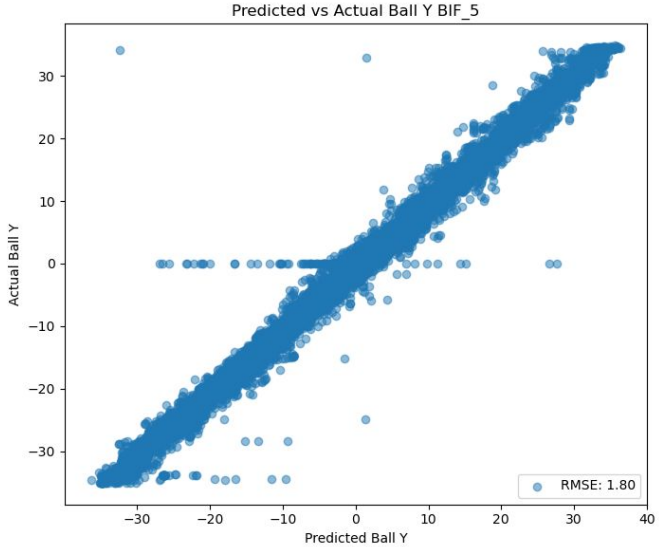


Table ranks all players from team FCK in order how how they have influenced the position of the ball in all their games - primarily through possession but could be by controlling movement of a different player with the ball etc.

The possession route seems to think goalkeepers have the biggest effect on the ball positioning in general

Ball Tracking





These graphs show how well our model predicted the position at the ball in the next timestep

Project Statement

We believe all group members to have contributed equally to the project.

The presentation is done by Sanvi, Francesco and Deniz, as Niels could not be there due to having the Early Universe exam at the same time.