SC 1015 Mini Project: Football Player Analysis

By Muthu and Lennard
(Lab FCE3, Team 9)

01 - Introduction

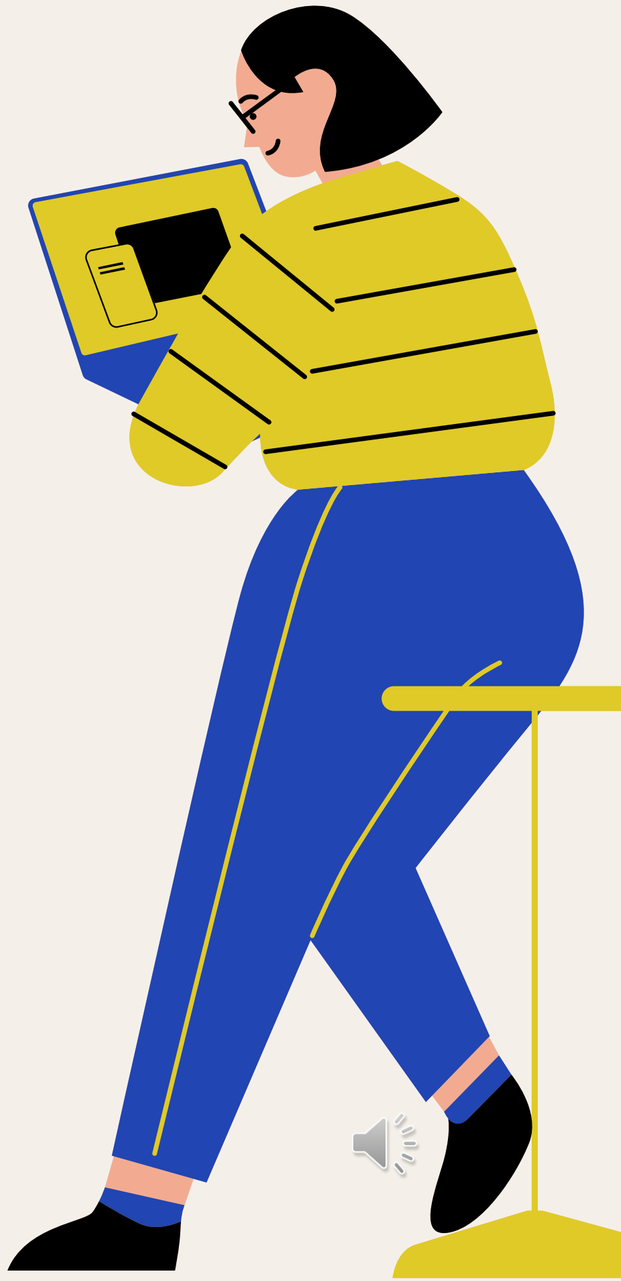
02 - Exploratory Data Analysis

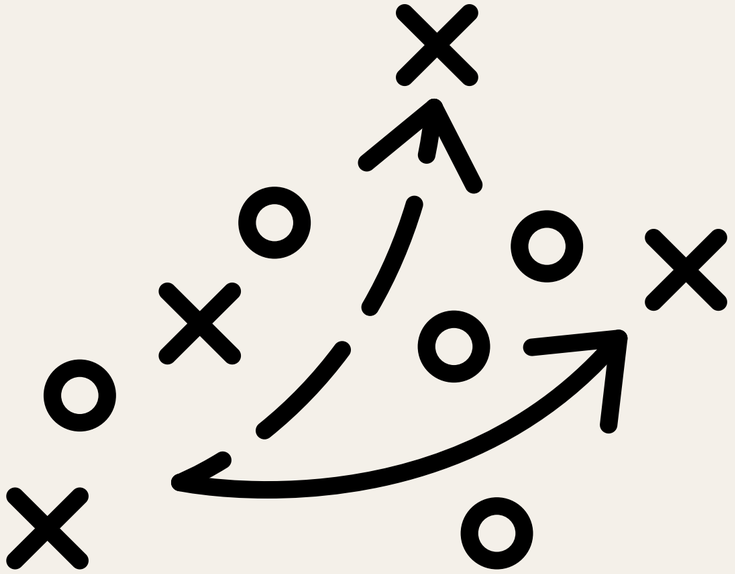
03 - Machine Learning

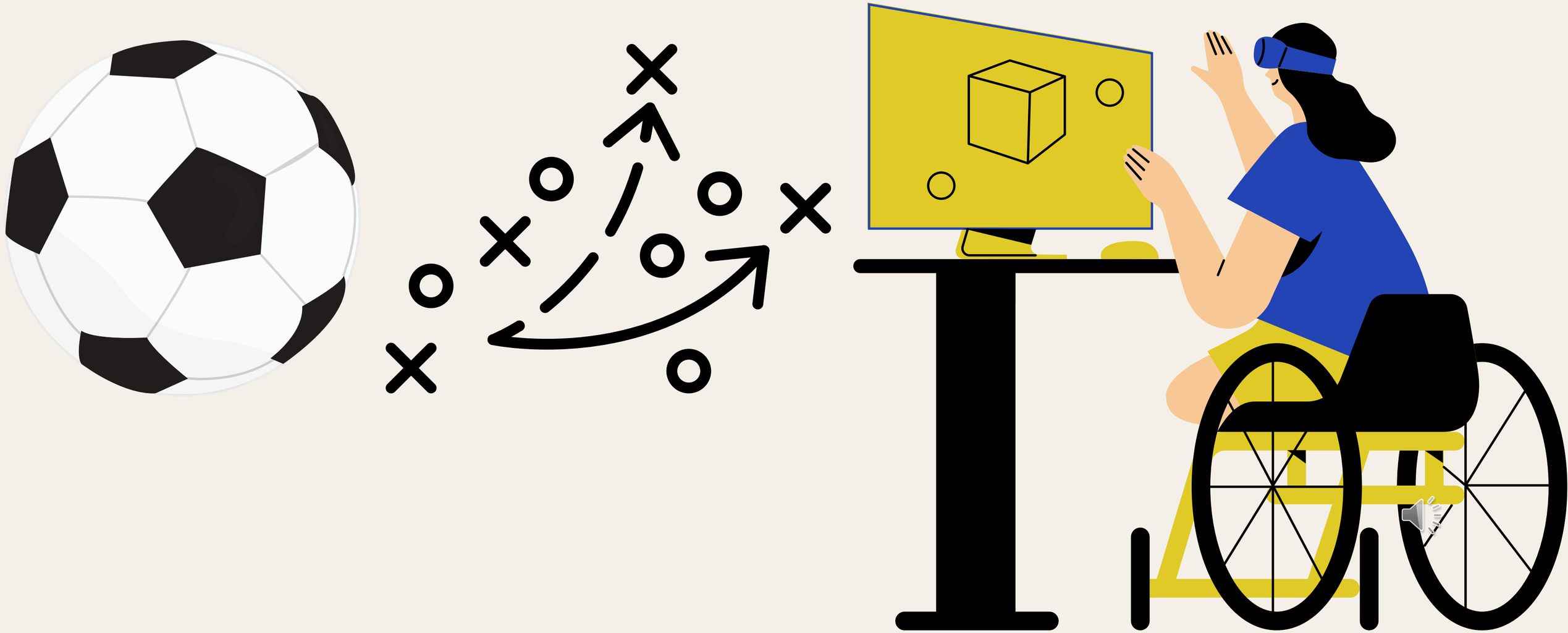
04 - Conclusion

Data

Visualization







01 - Introduction

Analyzing data enables informed decision-making

Our Task: Predicting the overall impact of a player in each relevant position based on a few Key Performance Indicators (target variables)

About our Dataset

2021-2022 Football Players Dataset from Kaggle

https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats

About our Dataset

2021-2022 Football Players Dataset from Kaggle

https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats

2900+ Rows (Player Stats)

143 Columns (Variables)

About our Dataset

2021-2022 Football Players Dataset from Kaggle

https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats

2900+ Rows (Player Stats)

143 Columns (Variables)

Statistics are based on per 90 minutes (usual duration of a football match)

Data Cleaning

Found Key Variables for each position

Reducing 143 variables to 10-15 variables for each position, which were considered as Key Performance Indicators (KPIs)

Data Cleaning

Data had a few values with question marks (?) which was removed

2 NaN/Null Values which were in variables that were not considered as KPIs

02 - EDA

- 10-15 variables important but can’t predict all

Data

Visualization

02 - EDA

- 10-15 variables important but can’t predict all
    - Used Correlation Matrix and Pair plots to decide on which variable to predict

Data

Visualization

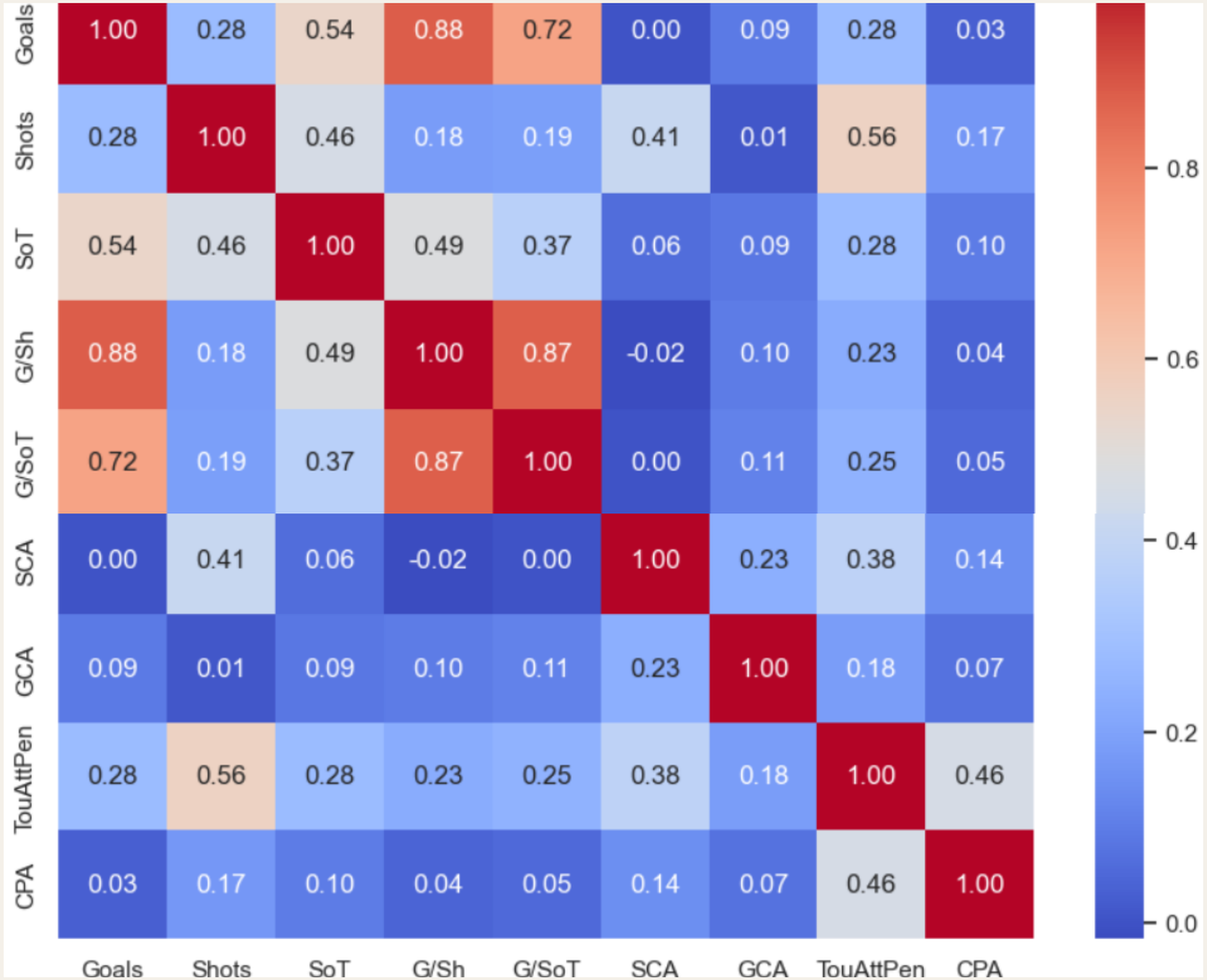
02 - EDA

- 10-15 predictor variables important but can't predict all
- Used Correlation Matrix and Pair plots to decide on which variable to predict
- Based on highest correlation and strong linear relationship, picked few target variables

Data

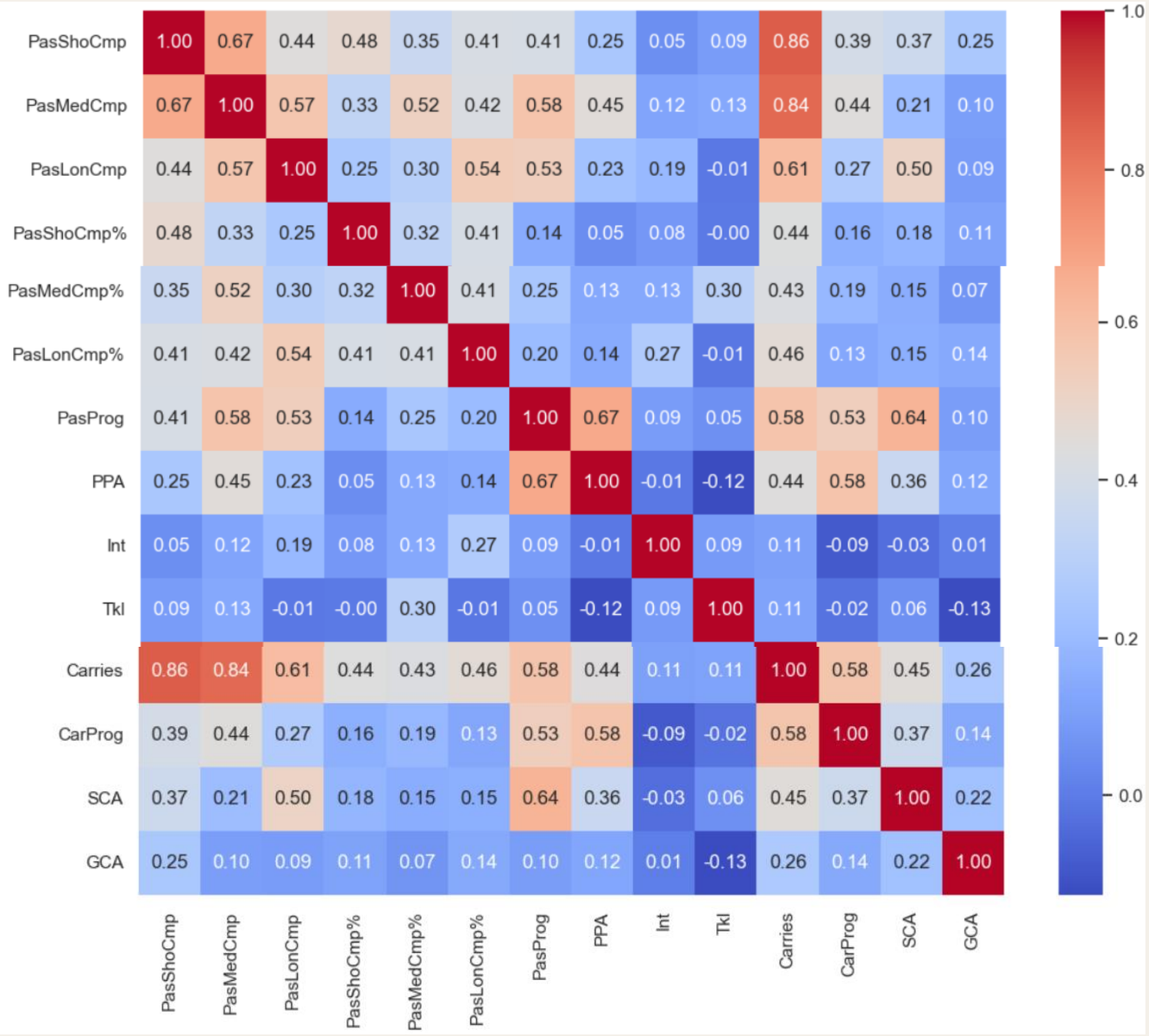
Visualization

Correlation Matrix for Forwards



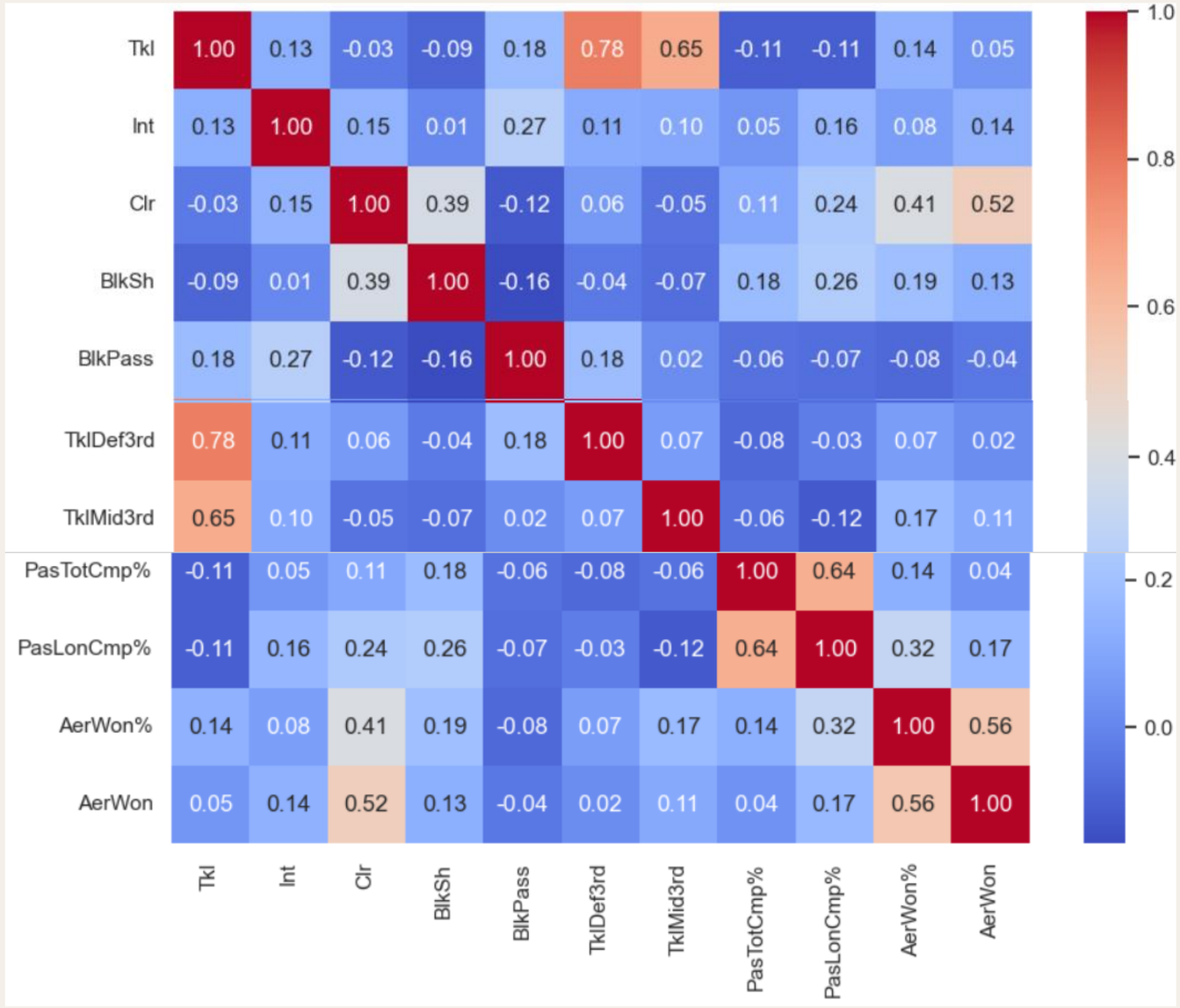
- Goals/Shot(G/Sh)
- Goals/ Shot on Target(G/SoT)

Correlation Matrix for Midfielders



- Short Pass Completion Percentage (PasShoCmp%)
- Medium Pass Completion Percentage (PasMedCmp%)
- Long Pass Completion Percentage (PasLonCmp%)

Correlation Matrix for Defenders



- Tackles (Tkl)
- Interceptions (Int)
- Clearances (Clr)

02 - EDA

Forwards (FW)

- *10 predictor variables*
- *2 target variables*

Midfielder (MF)

- *15 predictor variables*
- *3 target variables*

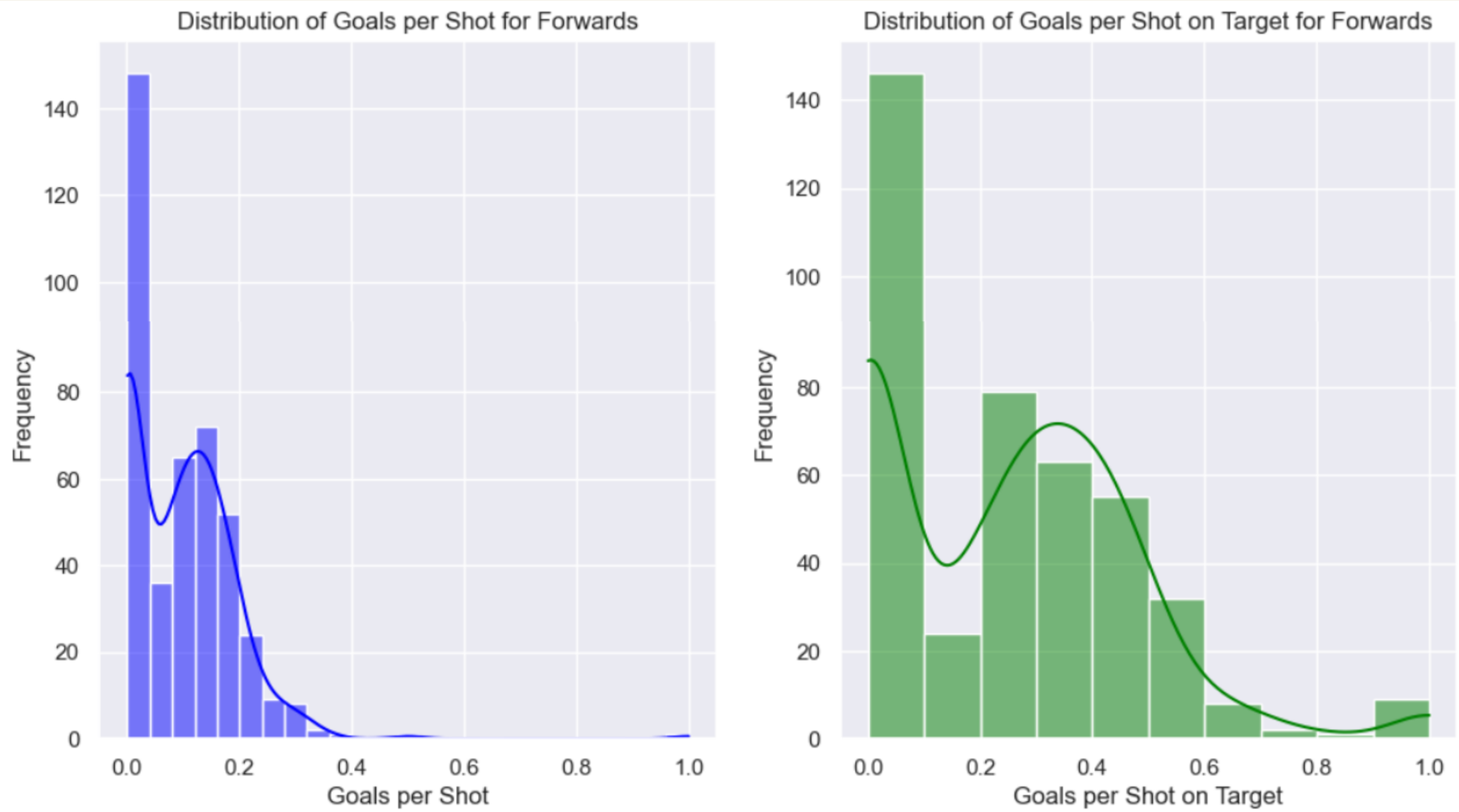
Defender (DF)

- *12 predictor variables*
- *3 target variables*

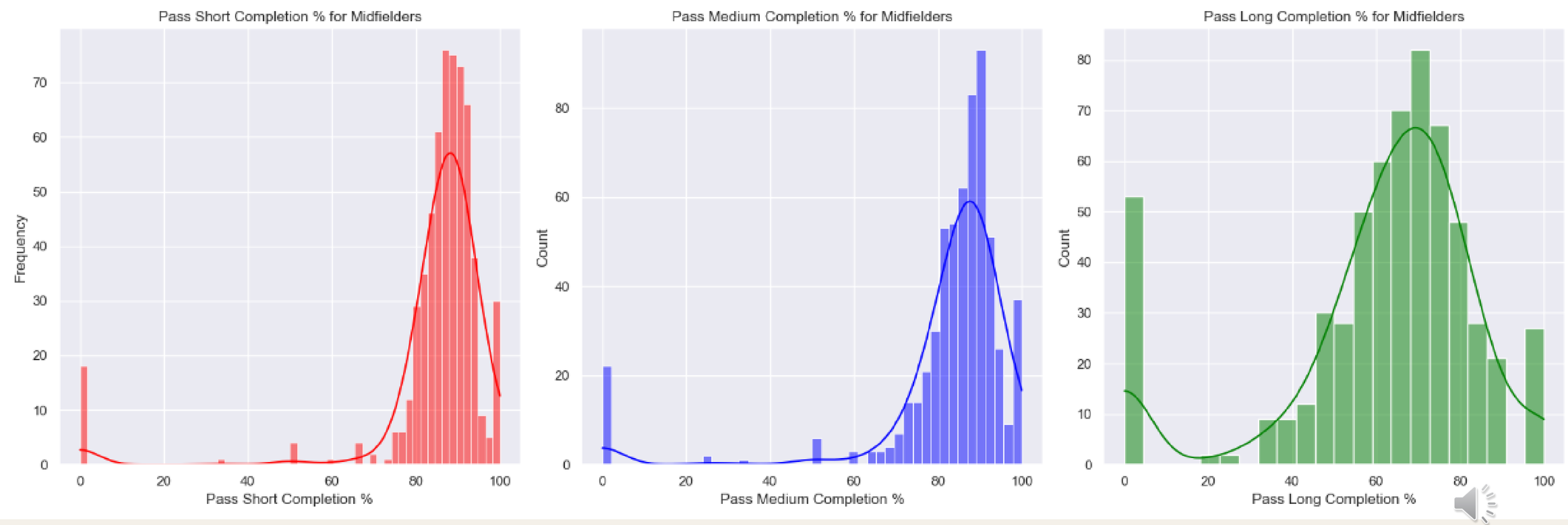
Wanted to look at the Data spread of the target variables which is going to be used for machine learning

Histograms and Boxplots to look at data spread

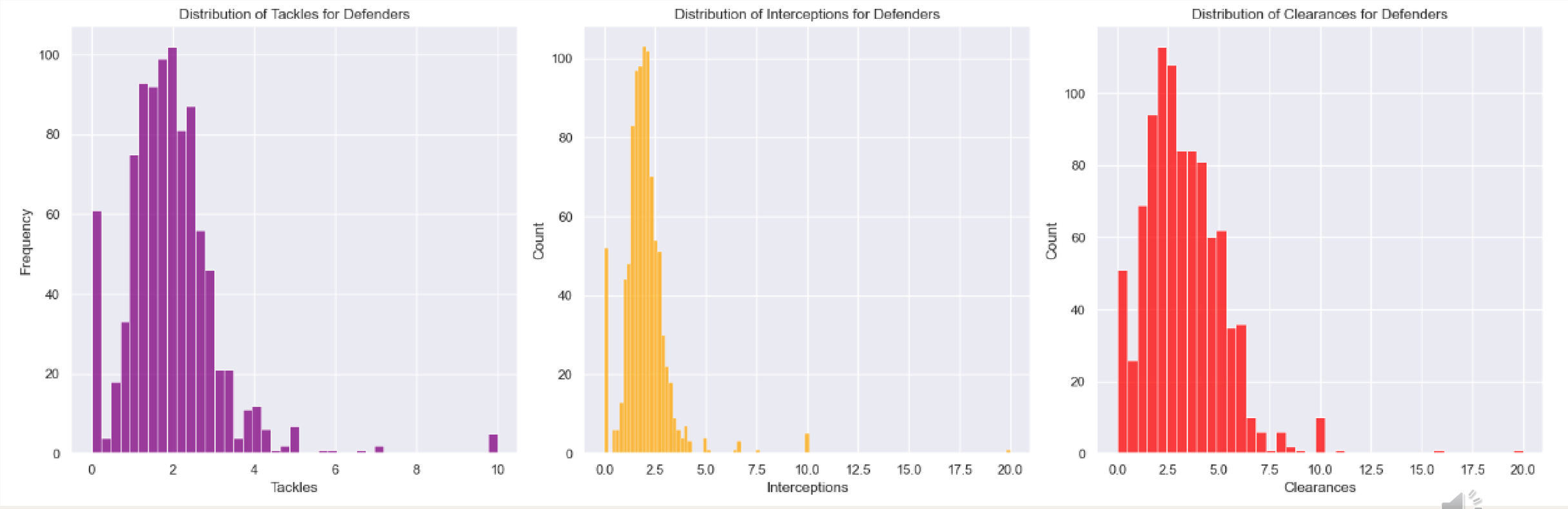
Histogram for Forwards



Histogram for Midfielders



Histogram for Defenders

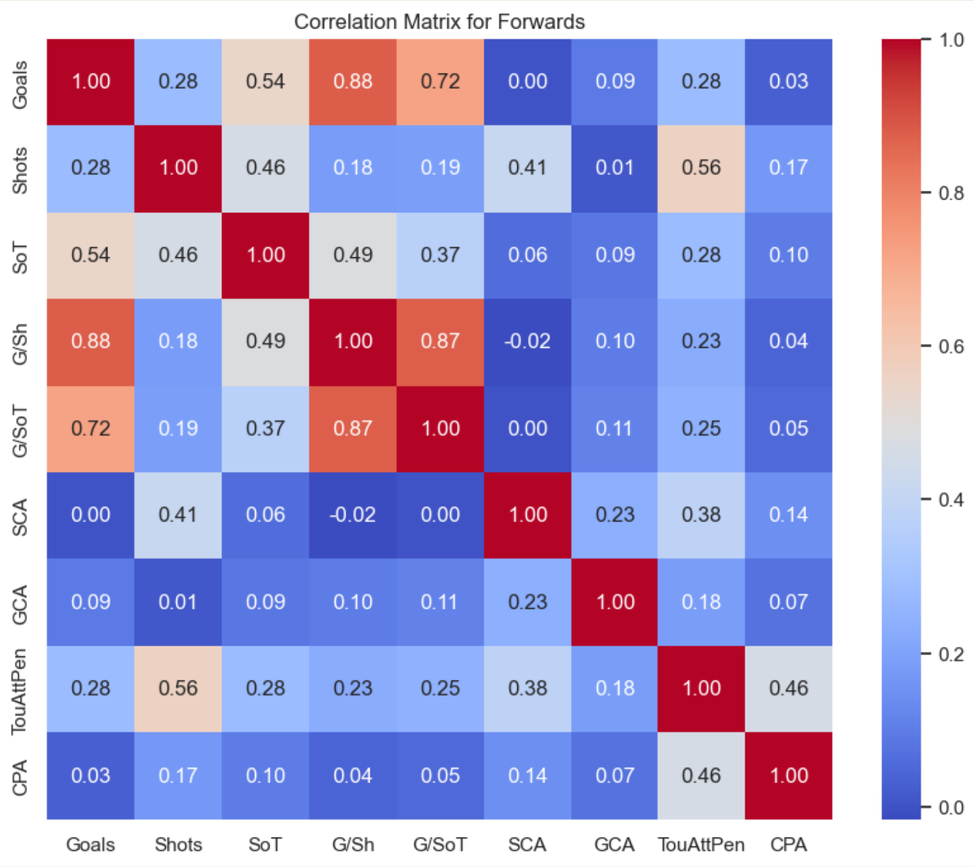


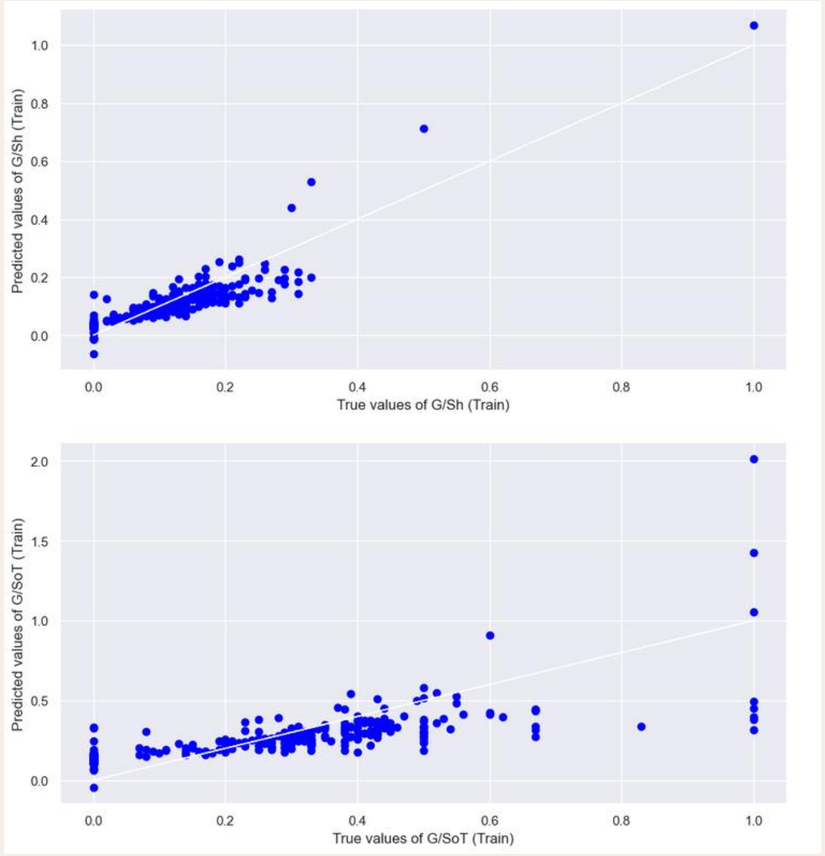
03 - Machine Learning

- *Linear regression*
- *Random forest*
- *Gradient boosting*

Linear Regression

Linear regression is used to predict the value of a variable based on other variables. It works best when the linear relationship of these variables are stronger.





Linear Regression Results

Forwards:

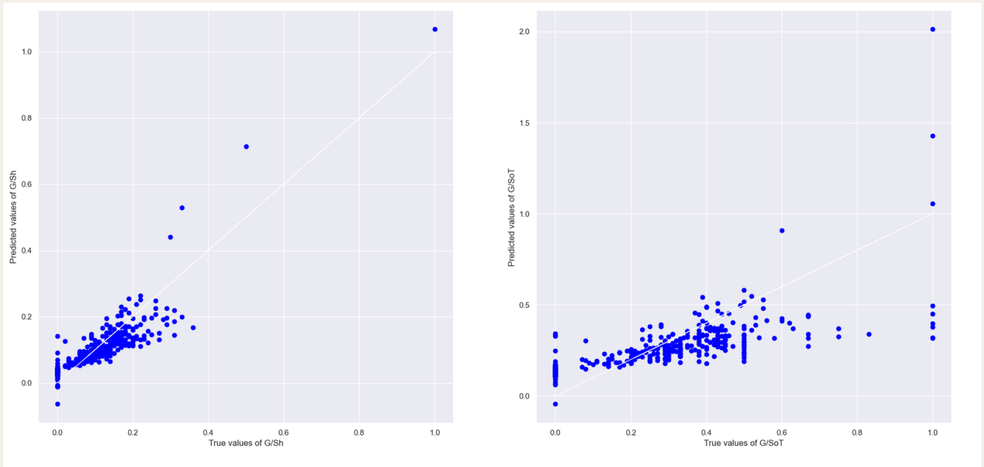
G/Sh Mean Squared Error: 0.002086
G/Sh R-squared: 0.774078
G/SoT Mean Squared Error: 0.026153
G/SoT R-squared: 0.486711

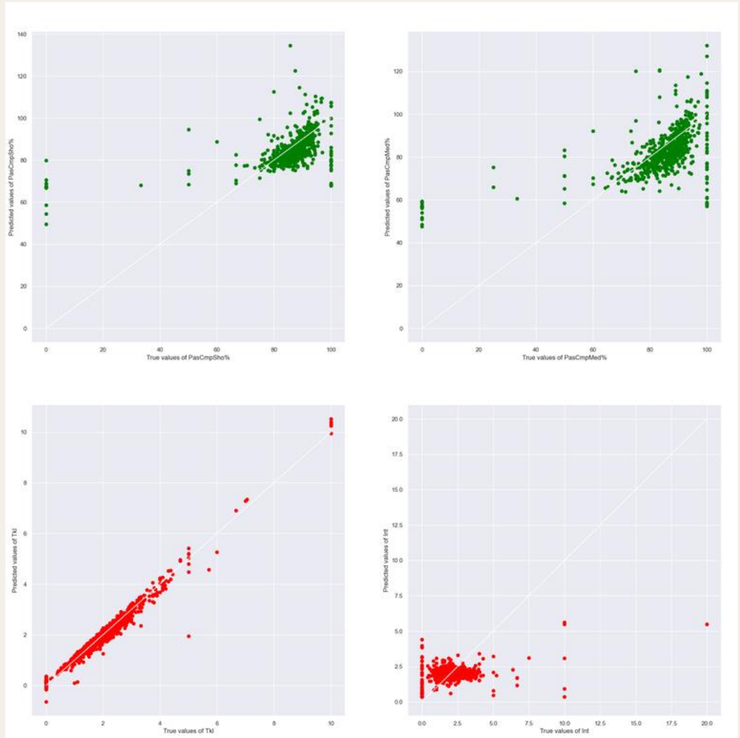
MidFielders:

Sho Mean Squared Error: 219.414044
Sho R-squared: 0.0.784481
Med Mean Squared Error: 226.161899
Med R-squared: 0.848058
Lon Mean Squared Error: 332.815054
Lon R-squared: 0.407103

Defenders:

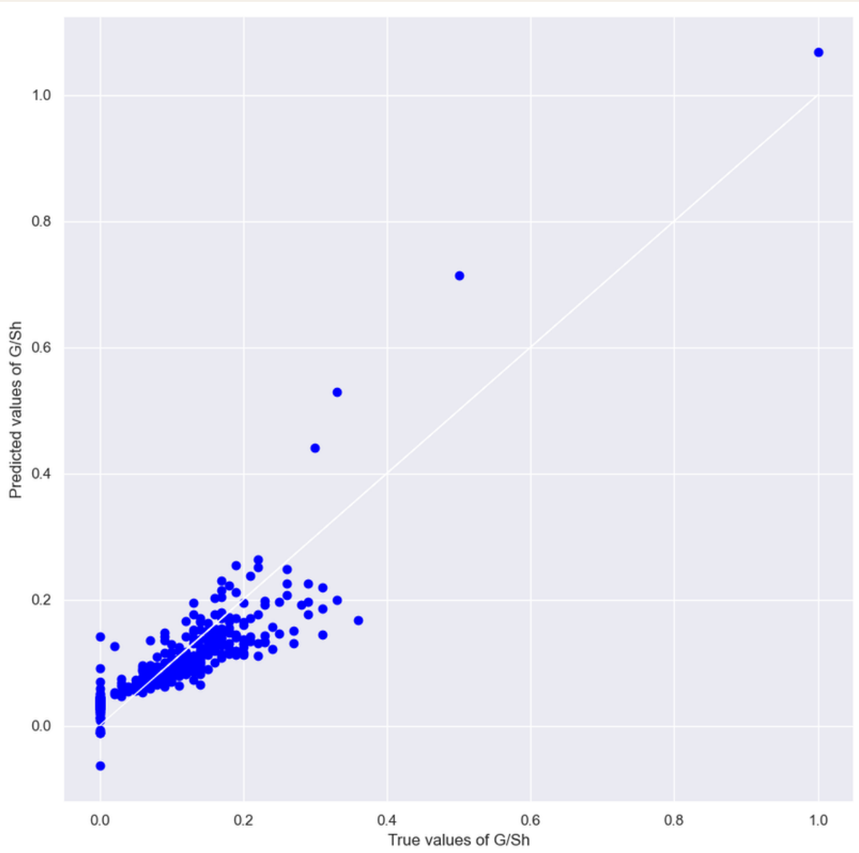
Tkl Mean Squared Error: 0.038136
Tkl R-squared: 0.826358
Int Mean Squared Error: 2.349181
Int R-squared: -0.582759
Clr Mean Squared Error: 6.93258
Clr R-squared: -0.790921

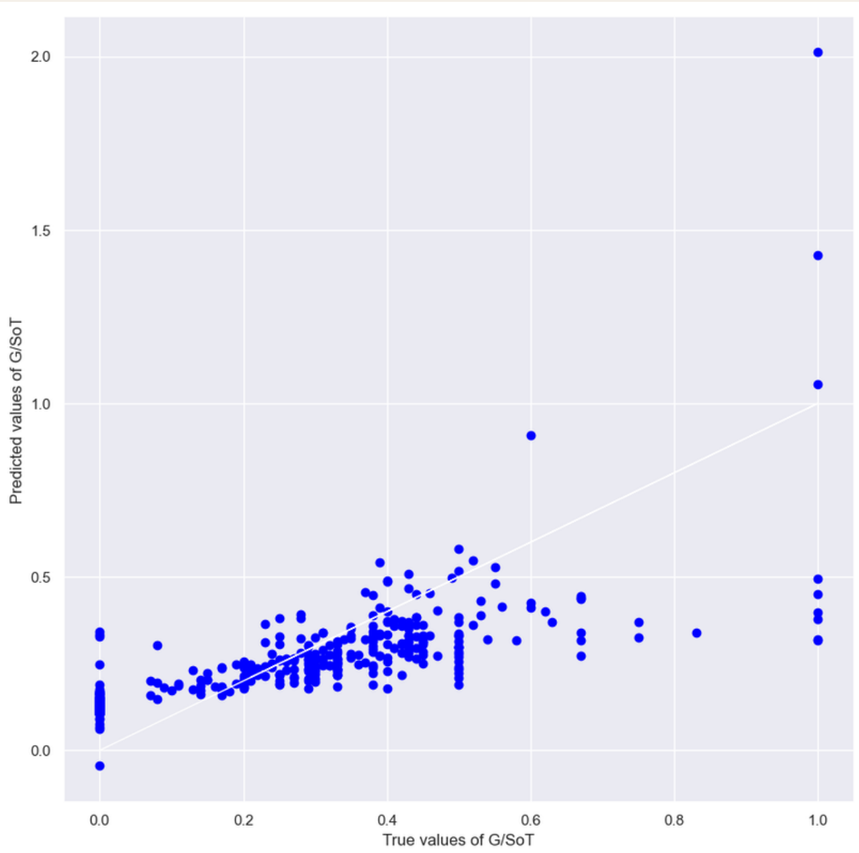




Linear Regression Results

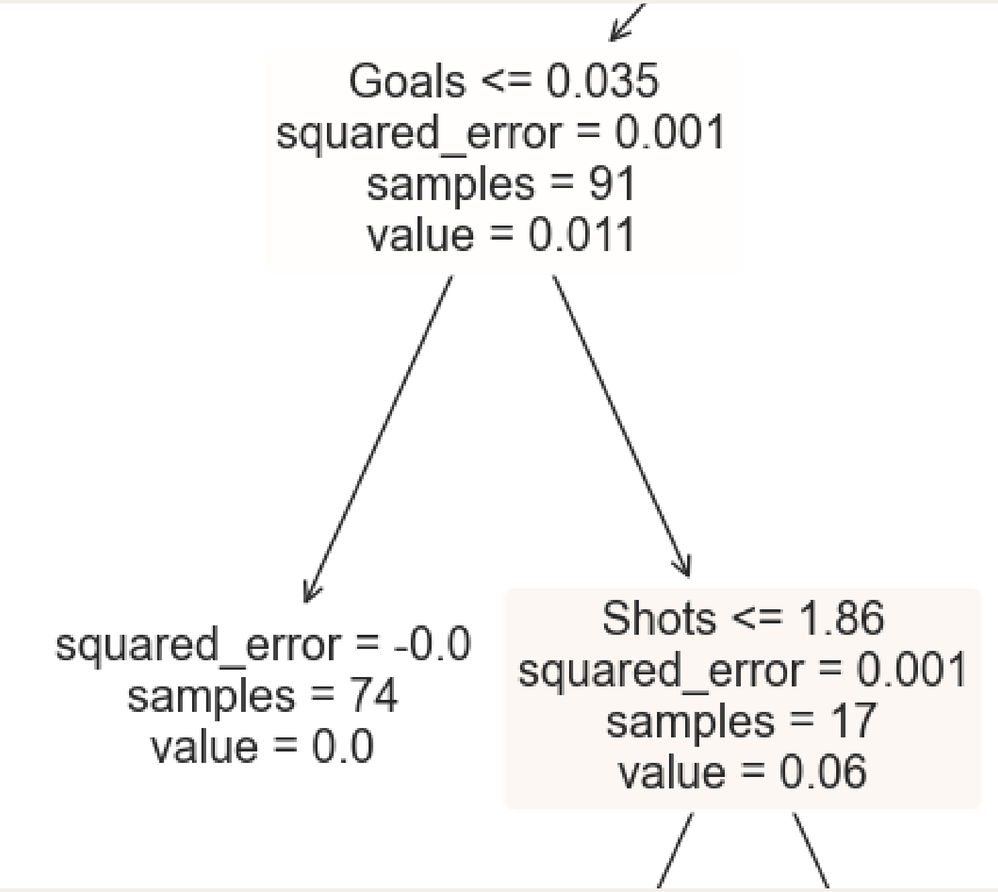
Forwards:
G/Sh Mean Squared Error: 0.002086
G/Sh R-squared: 0.774078
G/SoT Mean Squared Error: 0.026153
G/SoT R-squared: 0.486711

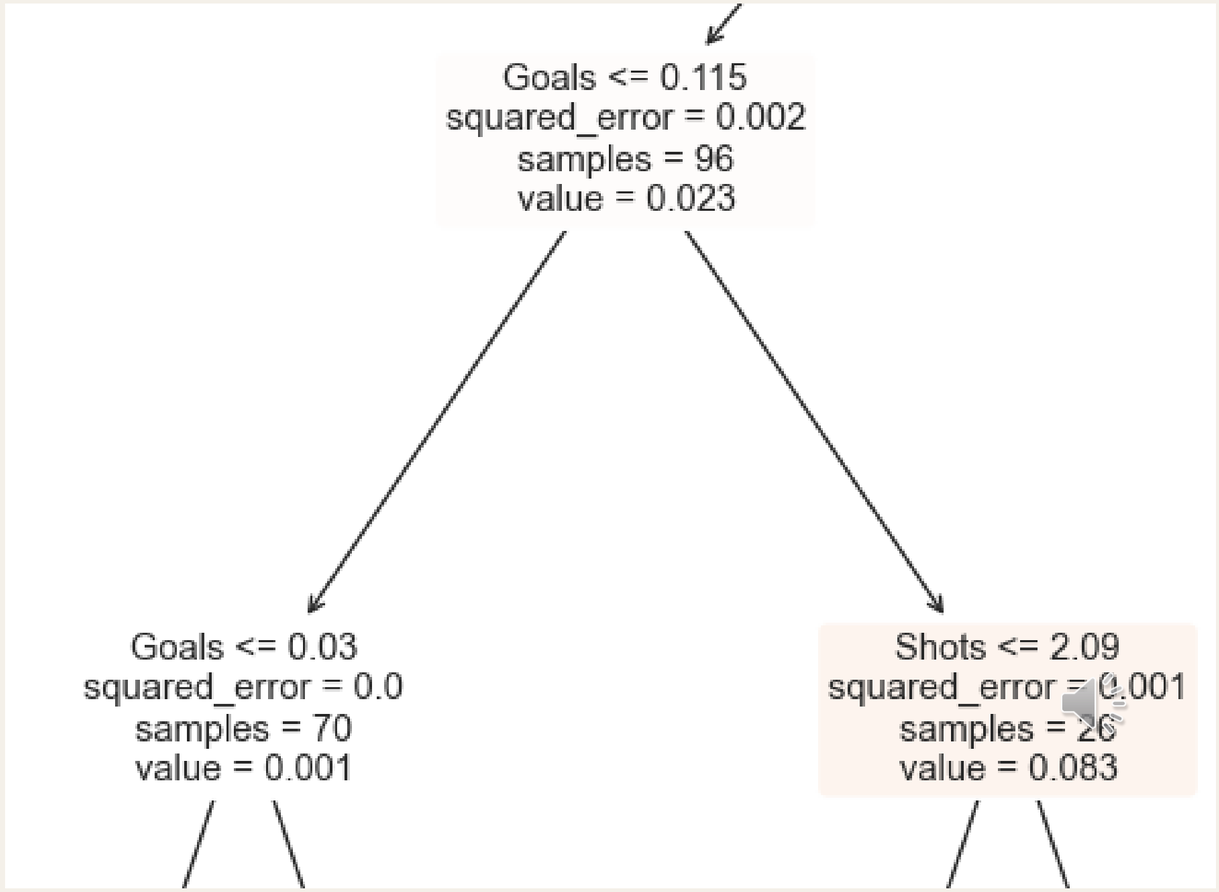




Random Forest

Random forest is an ensemble learning method that combines multiple instances of differently generated trees to form an average prediction.





Random Forest Results

Forwards:

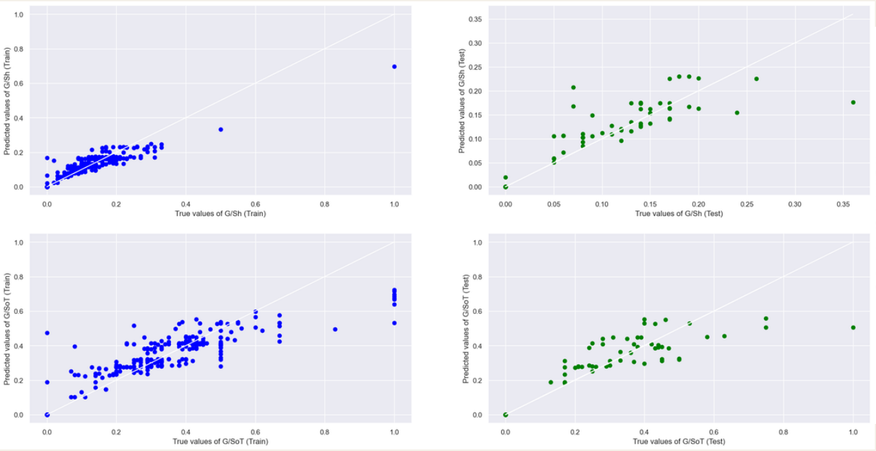
G/Sh Out-of-Bag Score: 0.828358
G/Sh Mean Squared Error: 0.006970
G/Sh R-squared: 0.561233
G/SoT Out-of-Bag Score: 0.787785
G/SoT Mean Squared Error: 0.044262
G/SoT R-squared: 0.104444

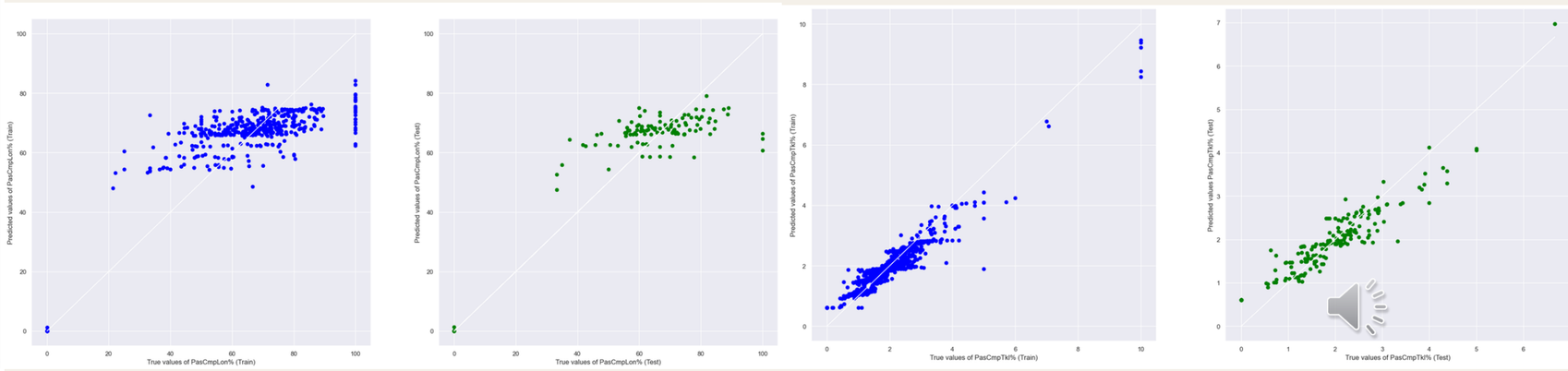
MidFielders:

Sho Out-of-Bag Score: 0.844804
Sho Mean Squared Error: 34.625104
Sho R-squared: 0.784481
Med Out-of-Bag Score: 0.812667
Med Mean Squared Error: 45.473705
Med R-squared: 0.848058
Lon Out-of-Bag Score: 0.731318
Lon Mean Squared Error: 164.383955
Lon R-squared: 0.644032

Defenders:

Tkl Out-of-Bag Score: 0.865155
Tkl Mean Squared Error: 0.160949
Tkl R-squared: 0.826358
Int Out-of-Bag Score: -0.092419
Int Mean Squared Error: 2.267593
Int R-squared: 0.234085
Clr Out-of-Bag Score: 0.274755
Clr Mean Squared Error: 1.878818
Clr R-squared: 0.419339





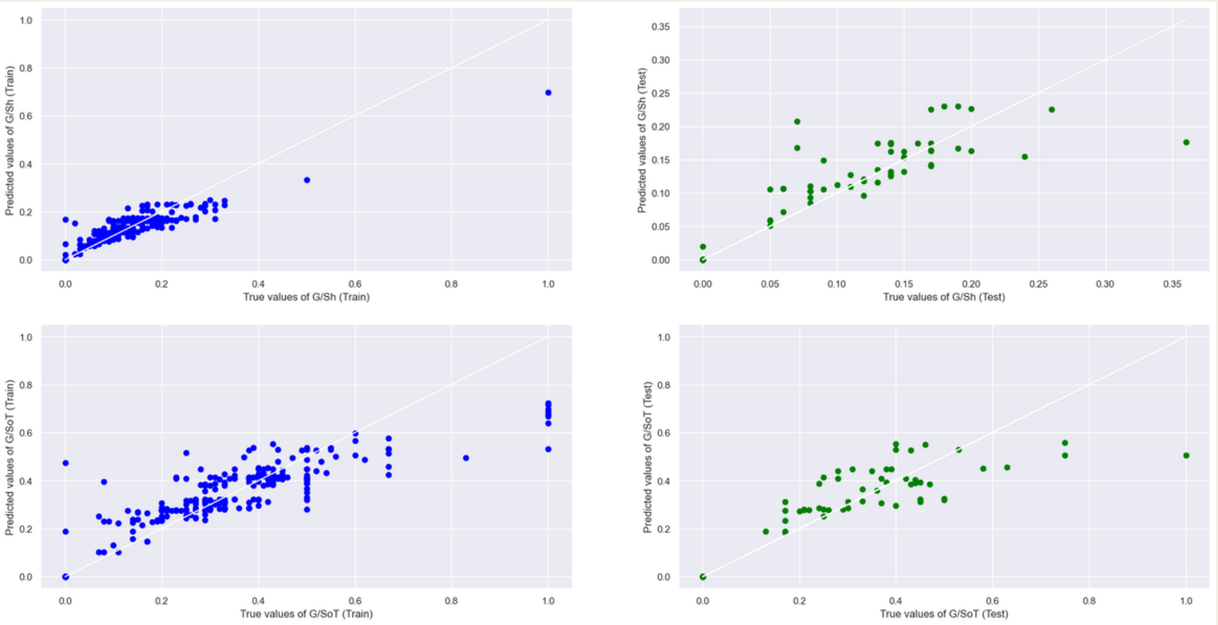
Random Forest Results

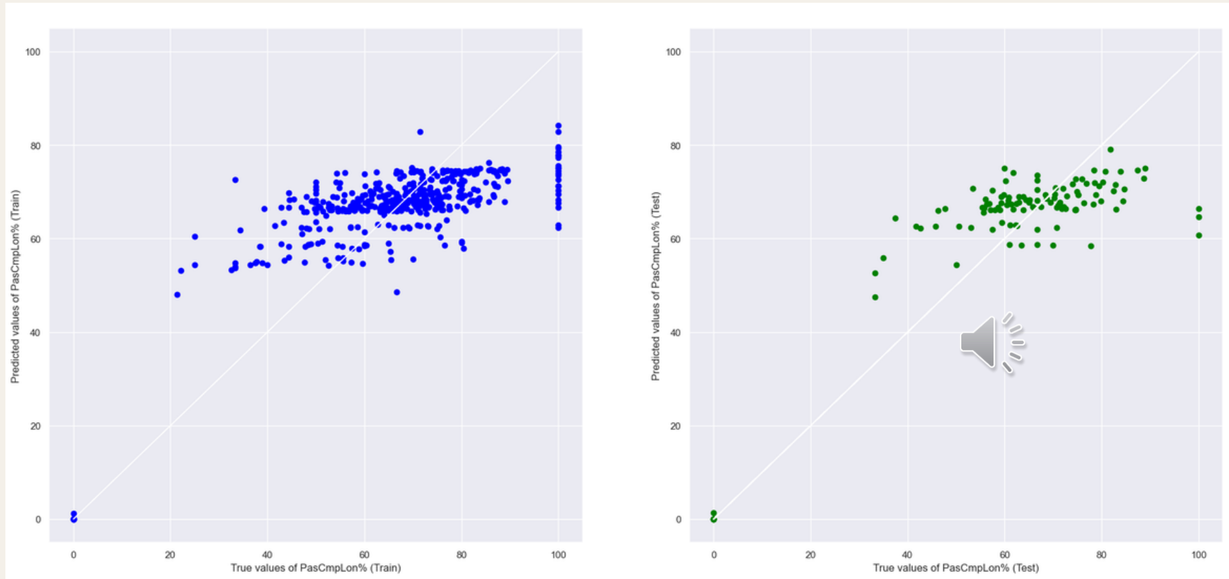
Forwards:

G/Sh Out-of-Bag Score: 0.828358
G/Sh Mean Squared Error: 0.006970
G/Sh R-squared: 0.561233
G/SoT Out-of-Bag Score: 0.787785
G/SoT Mean Squared Error: 0.044262
G/SoT R-squared: 0.104444

MidFielders:

Sho Out-of-Bag Score: 0.844804
Sho Mean Squared Error: 34.625104
Sho R-squared: 0.784481
Med Out-of-Bag Score: 0.812667
Med Mean Squared Error: 45.473705
Med R-squared: 0.848058
Lon Out-of-Bag Score: 0.731318
Lon Mean Squared Error: 164.383955
Lon R-squared: 0.644032





Gradient Boosting

Gradient boosting is an ensemble learning method that goes through multiple iterations to optimize the model. It uses its previous iterations to produce a more optimized model by minimising prediction errors.

```
FW_gradient_GSh.train_score_.round(4)
```

Out[34]:

```
array([0.0084, 0.007 , 0.0059, 0.005 , 0.0043, 0.0037, 0.0032, 0.0028,
       0.0024, 0.0021, 0.0019, 0.0017, 0.0015, 0.0013, 0.0012, 0.0011,
       0.001 , 0.0009, 0.0008, 0.0008, 0.0007, 0.0007, 0.0006, 0.0006,
       0.0006, 0.0005, 0.0005, 0.0005, 0.0004, 0.0004, 0.0004, 0.0004,
       0.0004, 0.0004, 0.0004, 0.0004, 0.0003, 0.0003, 0.0003, 0.0003,
       0.0003, 0.0003, 0.0003, 0.0003, 0.0003, 0.0003, 0.0003, 0.0003,
       0.0003, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002,
       0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002,
       0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0002, 0.0001,
       0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001,
       0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001,
       0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001, 0.0001,
       0.0001, 0.0001, 0.0001, 0.0001])
```

Gradient Boosting Results

Forwards:

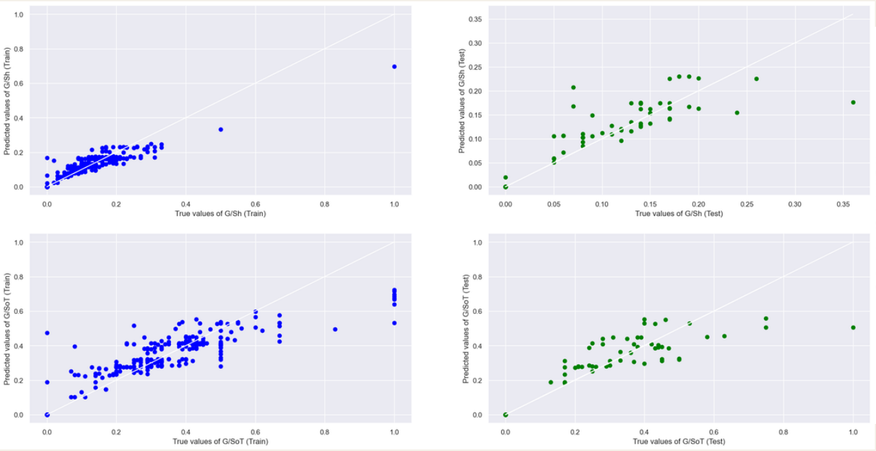
G/Sh Mean Squared Error: 0.005960
G/Sh R-squared: 0.624774
G/SoT Mean Squared Error: 0.043430
G/SoT R-squared: 0.121271

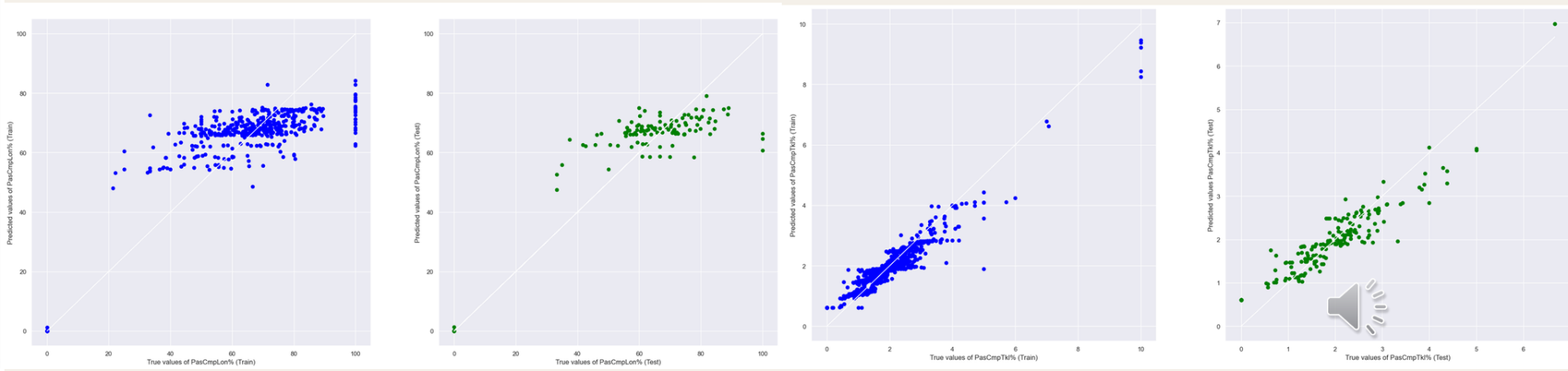
MidFielders:

Sho Mean Squared Error: 46.675189
Sho R-squared: 0.709478
Med Mean Squared Error: 179.863372
Med R-squared: 0.399024
Lon Mean Squared Error: 1206.27
Lon R-squared: -1.612145

Defenders:

Tkl Mean Squared Error: 0.039165
Tkl R-squared: 0.957752
Int Mean Squared Error: 1.967345
Int R-squared: 0.335501
Clr Mean Squared Error: 2.135259
Clr R-squared: 0.340084





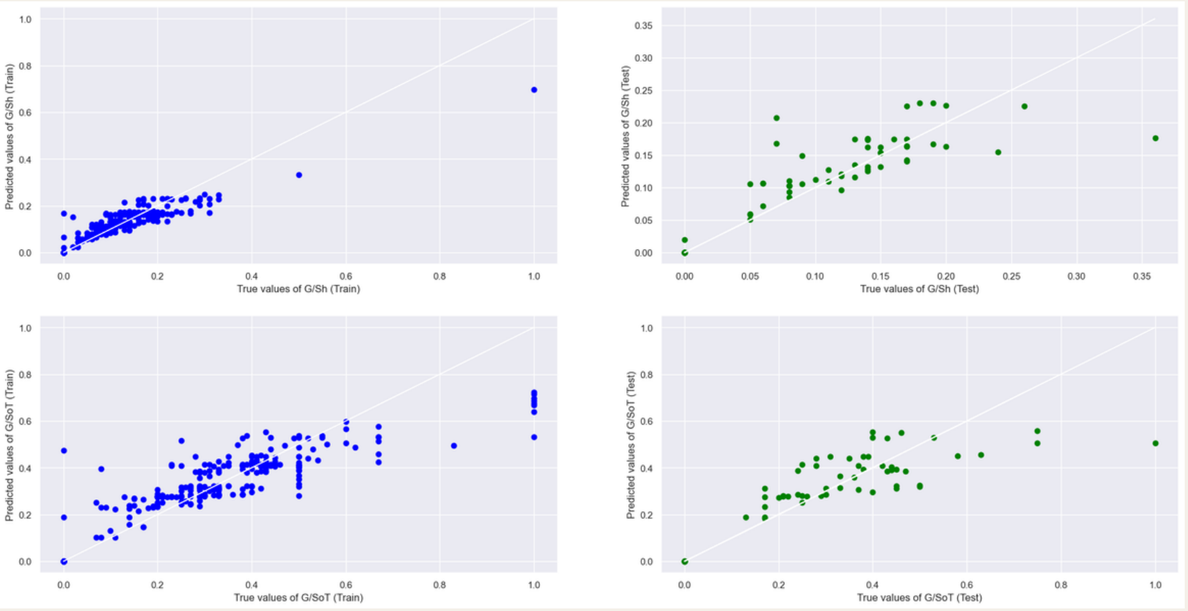
Gradient Boosting Results

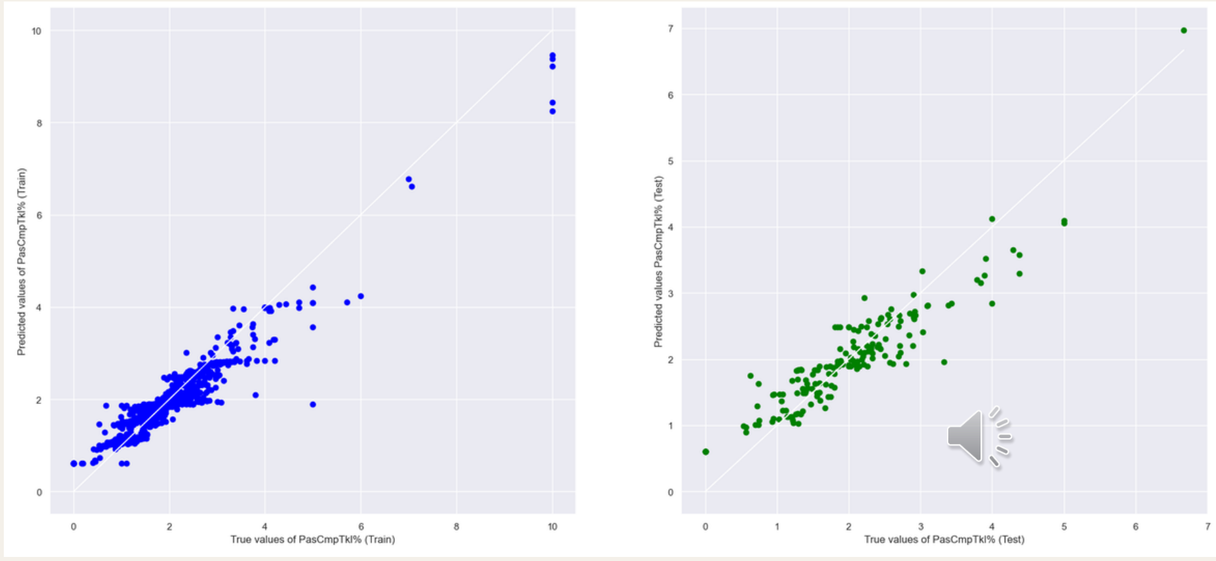
Forwards:

G/Sh Mean Squared Error: 0.005960
G/Sh R-squared: 0.624774
G/SoT Mean Squared Error: 0.043430
G/SoT R-squared: 0.121271

Defenders:

Tkl Mean Squared Error: 0.039165
Tkl R-squared: 0.957752
Int Mean Squared Error: 1.967345
Int R-squared: 0.335501
Clr Mean Squared Error: 2.135259
Clr R-squared: 0.340084





Result analysis

Linear regression is the least effective model, only capable of effectively predicting the stats of Forwards.
Random Forest is effective in predicting the stats of Forwards and Midfielders
Gradient Boosting is effective in predicting the stats of Forwards and Defenders.

Variable	R2_LR	MSE_LR	R2_RF	MSE_RF	OOB_RF	R2_GB	MSE_GB
G/Sh	0.774078	0.00208653	0.561233	0.00697013	0.828358	0.624774	0.00596073
G/SoT	0.486711	0.0261533	0.104444	0.044262	0.787785	0.121271	0.0434303
PasCmpSho%	0.186138	219.414	0.784481	34.6251	0.844804	0.709478	46.6751
PasCmpMed%	0.336065	226.161	0.848058	45.4737	0.812667	0.399024	179.863
PasCmpLon%	0.407103	332.815	0.644032	164.383	0.731318	-1.61214	1206.27
Tkl	0.971423	0.0381361	0.826358	0.160949	0.865155	0.957752	0.03916
Int	-0.582759	2.34918	0.234085	2.26759	-0.0924197	0.335501	1.96734
Clr	-0.790921	6.93258	0.419339	1.87881	0.274755	0.340084	2.13525

Result analysis

Variable	R2_LR	MSE_LR	R2_RF	MSE_RF	OOB_RF	R2_GB	MSE_GB
G/Sh	0.774078	0.00208653	0.561233	0.00697013	0.828358	0.624774	0.00596073
G/SoT	0.486711	0.0261533	0.104444	0.044262	0.787785	0.121271	0.0434303
PasCmpSho%	0.186138	219.414	0.784481	34.6251	0.844804	0.709478	46.6751
PasCmpMed%	0.336065	226.161	0.848058	45.4737	0.812667	0.399024	179.863
PasCmpLon%	0.407103	332.815	0.644032	164.383	0.731318	-1.61214	1206.27
Tkl	0.971423	0.0381361	0.826358	0.160949	0.865155	0.957752	0.03916
Int	-0.582759	2.34918	0.234085	2.26759	-0.0924197	0.335501	1.96734
Clr	-0.790921	6.93258	0.419339	1.87881	0.274755	0.340084	2.13525

Learning outcomes

Different machine learning techniques

- Random Forest
- Gradient boosting

Outcome

- Teams can predict and identify top players
- Create a targeted training regime
- Preplanning against teams with strong players

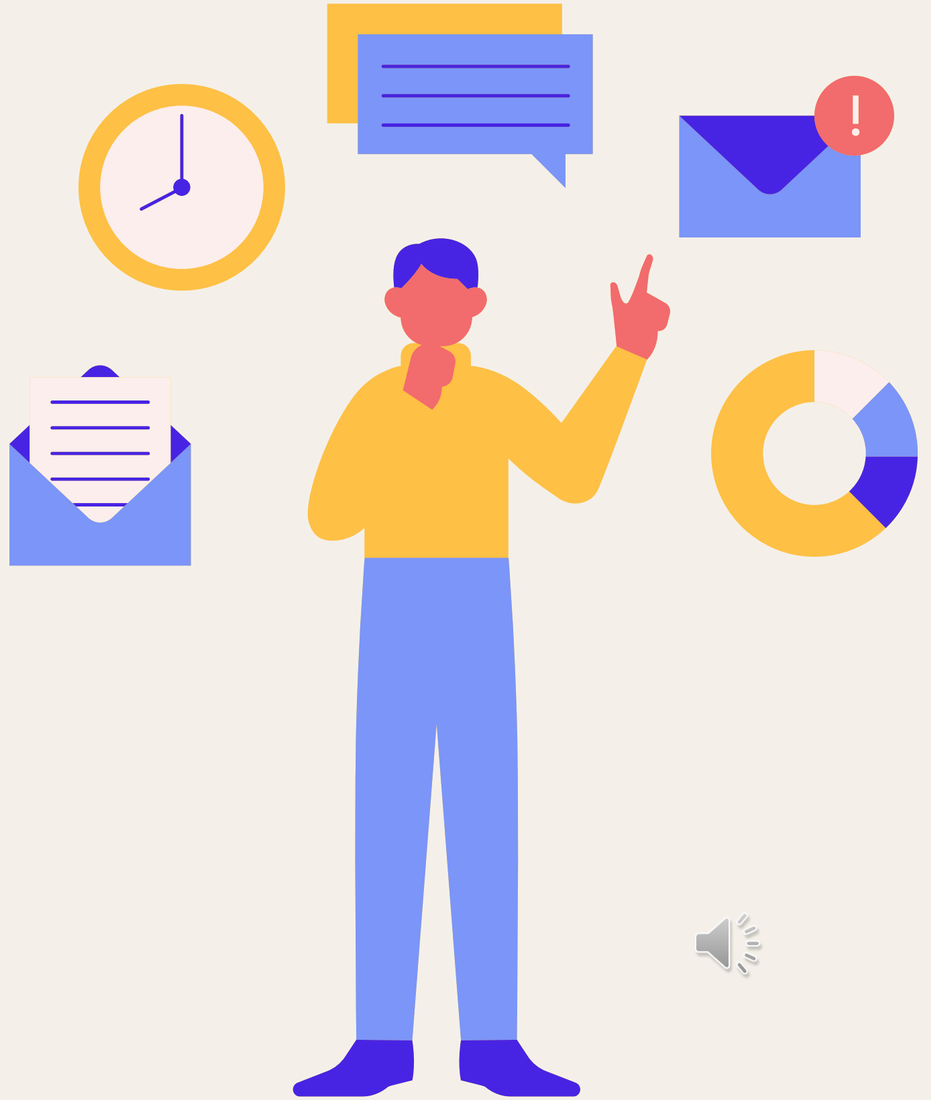
Data-driven Insights

Stat-based prediction

- Not only focused on goals
- e.g. G/Sh shows how accurate they are

Identify Weaknesses

- Identify which stats brings down the prediction
- Players can work on those weaknesses to become better



Thanks