Fantasy Premier League Predictor

- Iosif Pintirishis
- Andreas Papadopoulos
- Fivos Lympouras
- Constantinos Constantinou

Fall Semester 23/24
DSC 510 – Introduction to Data Science

Content

- Introduction
- Problem Formulation
- Data Description
- Exploratory Data Analysis
- Data Preprocessing
- Evaluation Methodology
- Best Model & Predictions
- Conclusion

CONSTANTINOU

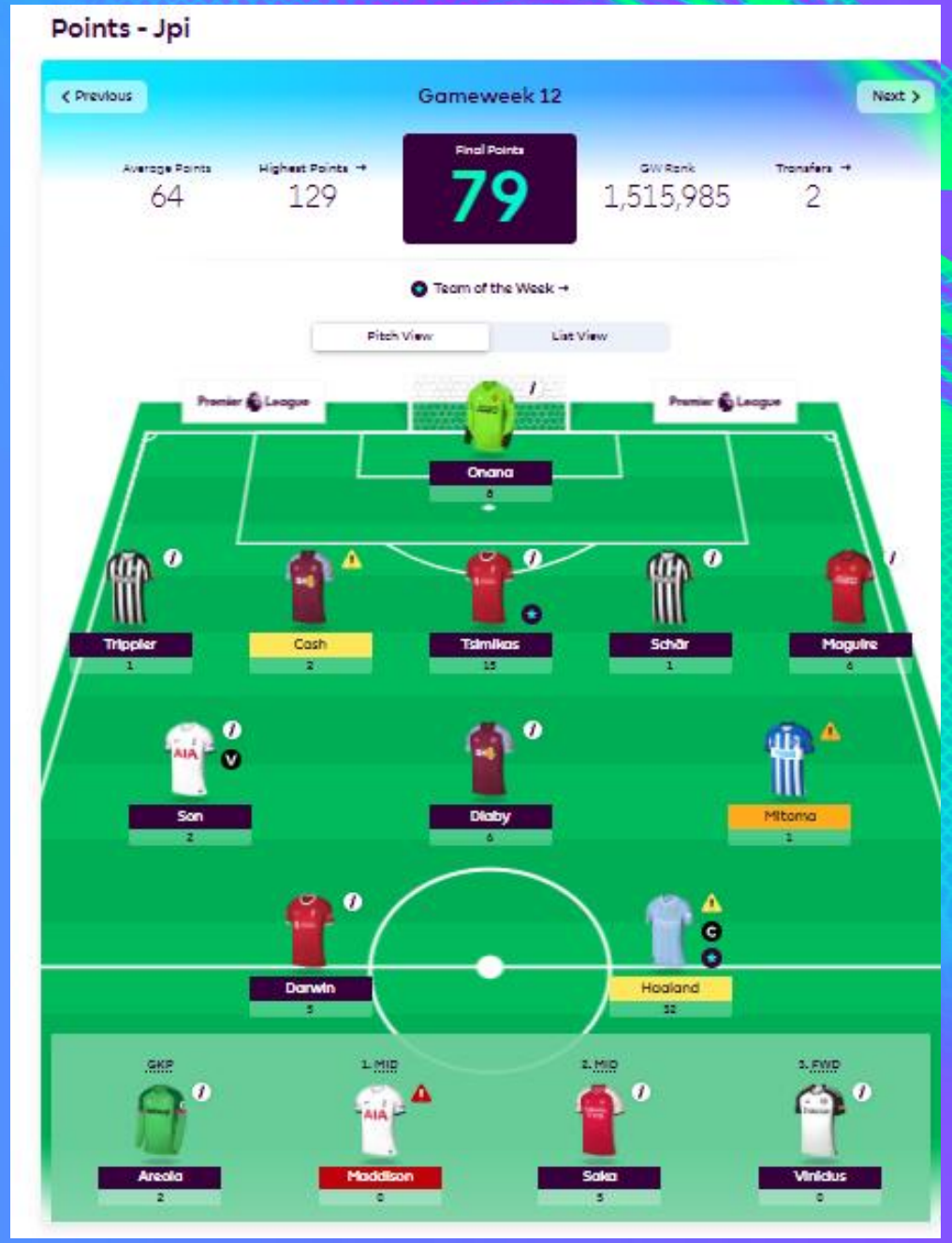
LYMPOURAS

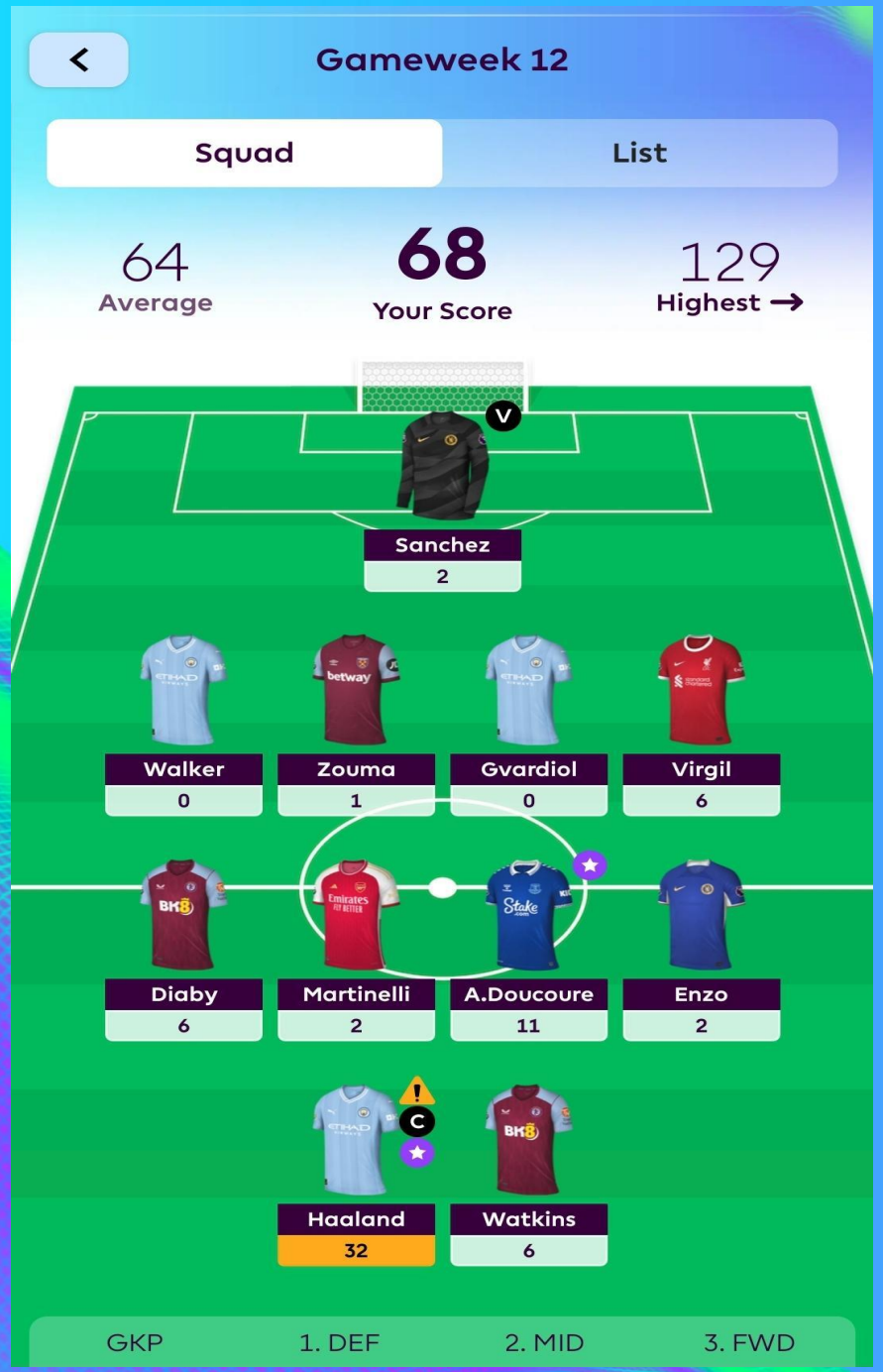
PAPADOPOULOS

PINTIRISHIS

Introduction

- A Fantasy manager of Premier League players
- Budget of £100.0m to spend on 15 players
- One free transfer per week





Problem Formulation

- Winner of your FPL League
- Select the Best XI
- Right Transfer (= Key Factor)
- Predict the next game week points

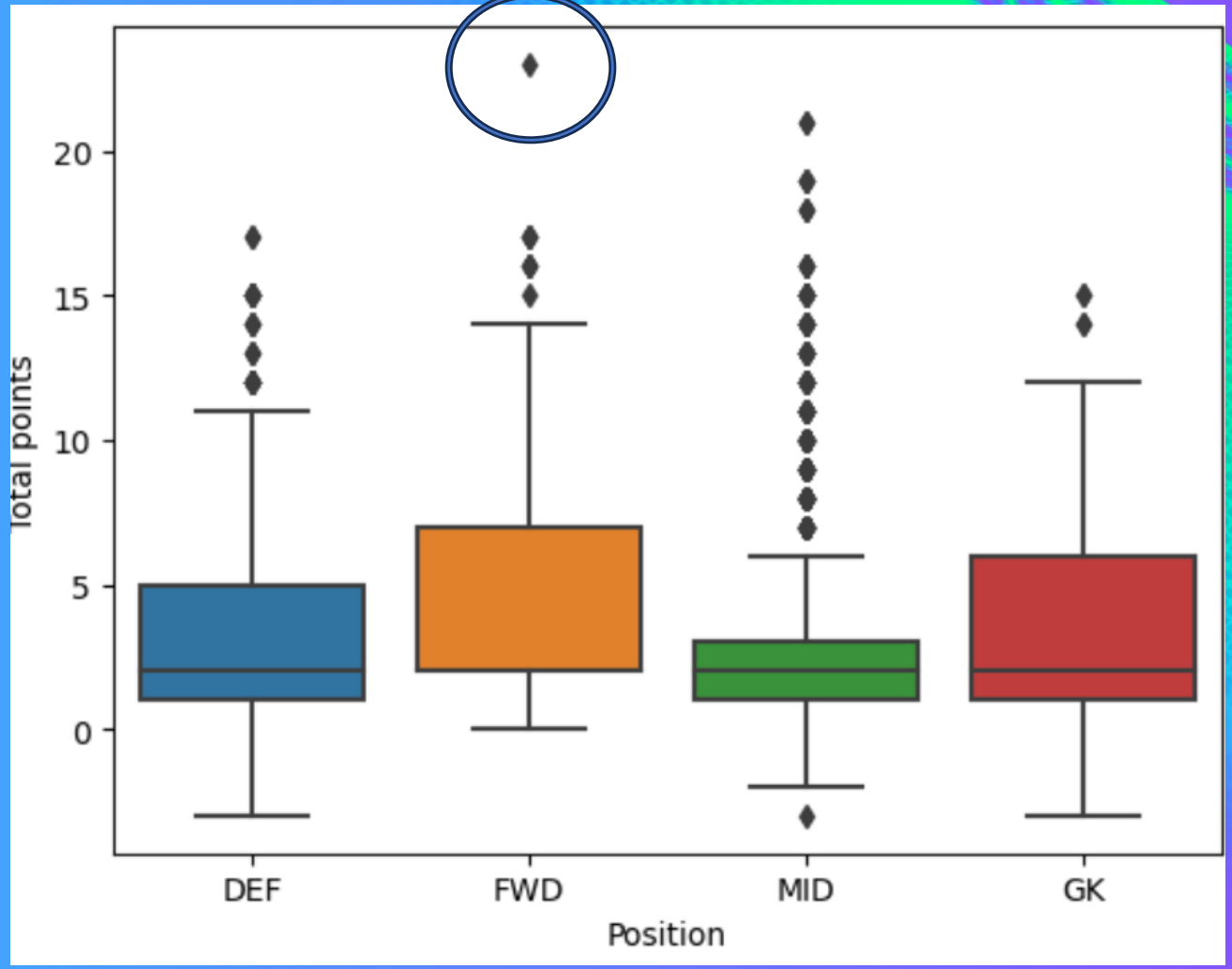
Data Description

- Premier League 2022-23
- Merged 38 Gameweeks
- 26505 rows & 40 columns
- Collected from GitHub

	name	position	team	xP	assists	bonus	bps	clean_sheets	creativity	element	...	team_a_score	team_h_score	threat
0	Nathan Redmond	MID	Southampton	0.0	0	0	0	0	0.0	403	...	0	4	0.0
1	Junior Stanislas	MID	Bournemouth	-0.1	0	0	0	0	0.0	58	...	1	2	0.0
2	Armando Broja	FWD	Chelsea	3.5	0	0	27	0	0.3	150	...	0	3	17.0
3	Fabian Schär	DEF	Newcastle	3.3	0	0	10	0	0.5	366	...	1	5	2.0
4	Jonny Evans	DEF	Leicester	2.5	0	0	15	0	1.5	249	...	1	2	33.0

Exploratory Data Analysis

- All positions have the same median except FWD
- FWD & GK have approximately the same IQR
- MID has a lot of outliers
- Would it be better to separate them?



Erling Haaland

- 35 Goals
- 8 Assists
- 7,961,047 Captained by
- Overall, **259** Points

Erling Haaland

Goals: 3

Assists: 2

Bonus points: 3

Total points: 23

Price: £12.1m

Ownership: 82.1%

Top 100k ownership: 99.5%

Top 100k captaincy: 72.6%

fantasyfootballfix.com

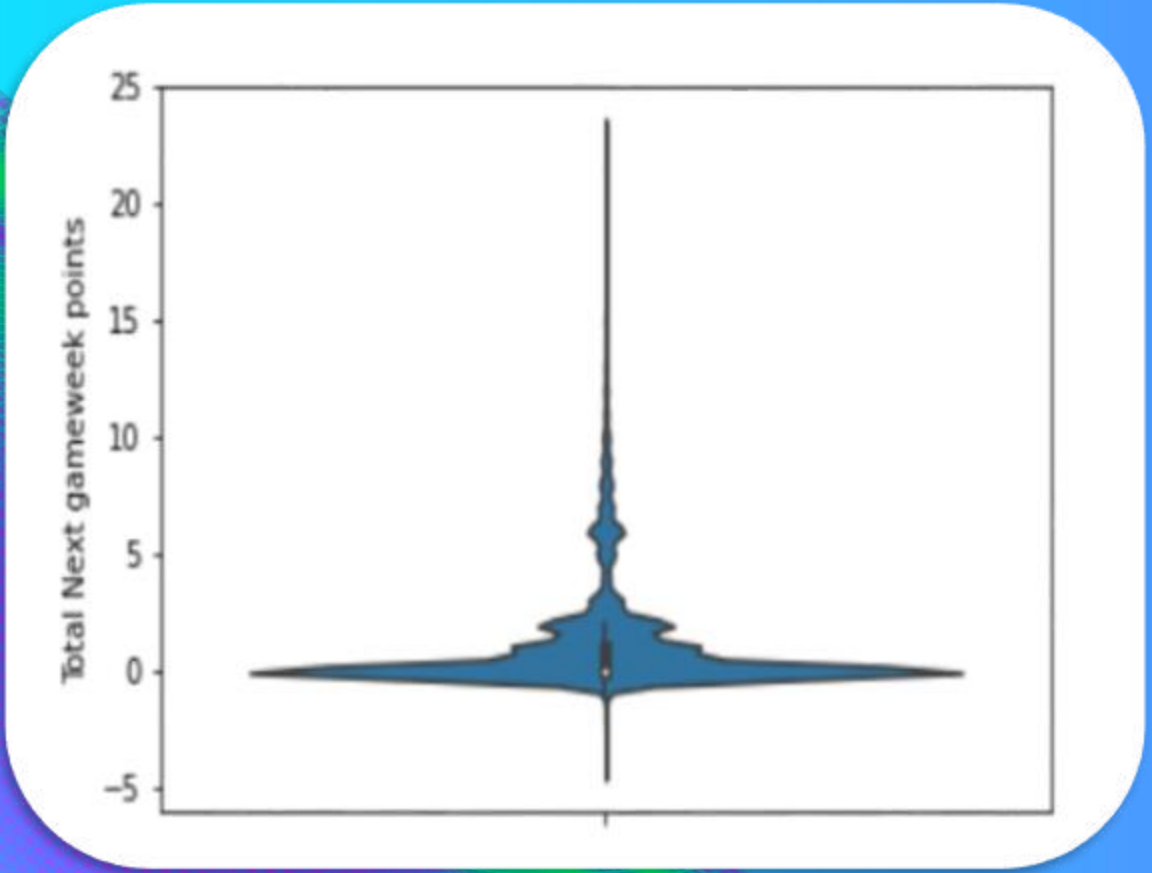
Exploratory Data Analysis

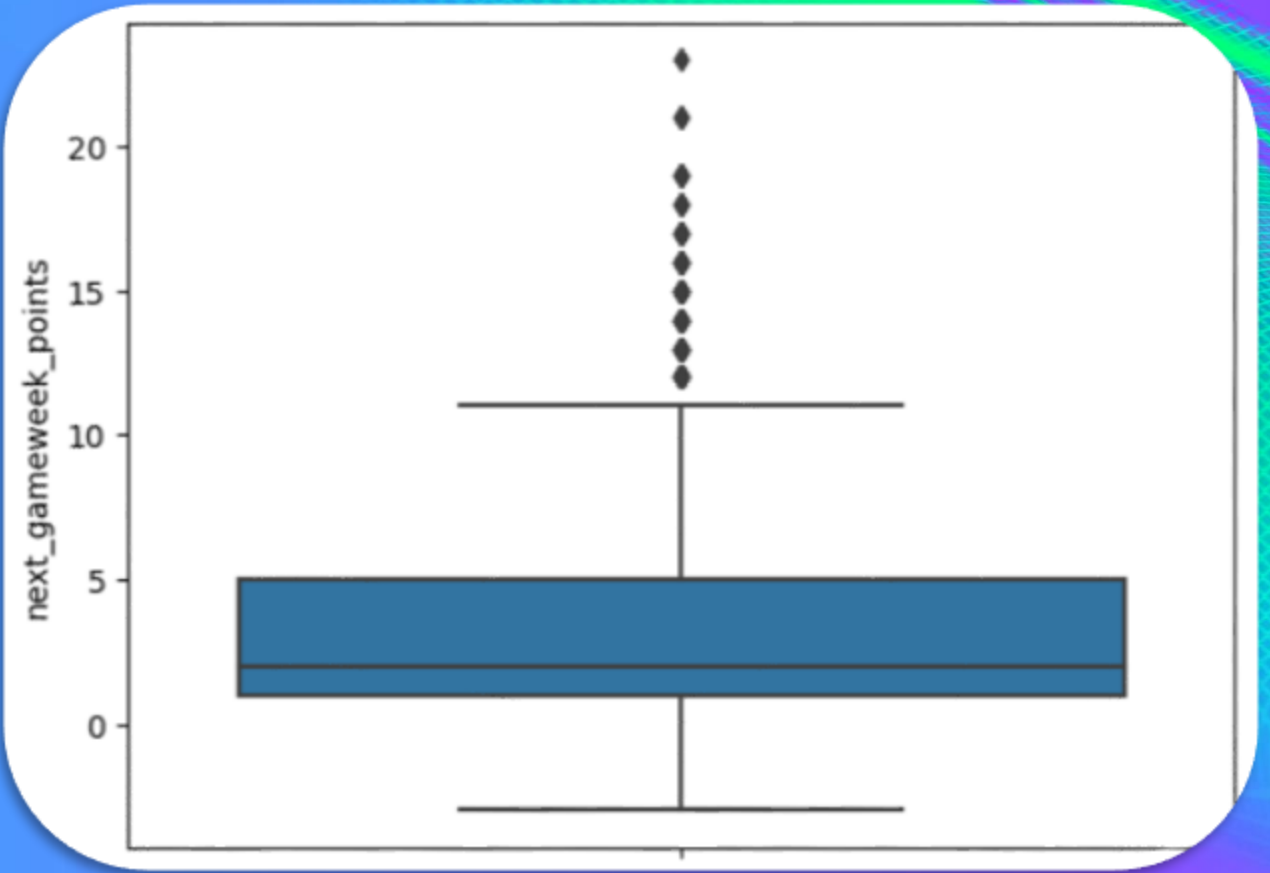
- Q1, Q2, Q3 is zero in almost everything
- Very low mean for the goals_scored, assists and total_points

Reason: Many players play rarely

	goals_scored	assists	total_points
count	26505.000000	26505.000000	26505.000000
mean	0.039162	0.034975	1.196906
std	0.215718	0.197954	2.355236
min	0.000000	0.000000	-4.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000
75%	0.000000	0.000000	1.000000
max	3.000000	3.000000	23.000000

Data Preprocessing





Data Preprocessing - Encoding

position_GK	position_DEF	position_MID	position_FWD
1.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0

next_gameweek_home_away_True	next_gameweek_home_away_False
1.0	0.0
0.0	1.0

One hot encoding for columns

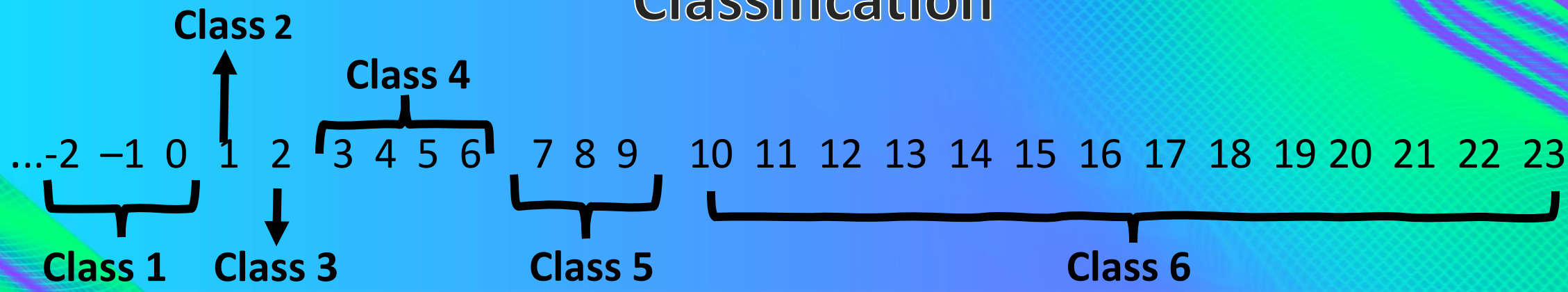
- Position
- Was_home

Evaluation Methoddology

Random Forest Regression

Evaluation Methodology

Classification



---------- CLASSIFICATION MODEL PERFORMANCE EVALUATION ----------

[93]:

	Model	Cross Val Score	Test Accuracy	Average_Accuracy	Precision	Recall	F1 Score
5	AdaBoostClassifier	0.352163	0.3428	0.347481	0.392001	0.243961	0.208615
1	SVC	0.346310	0.3418	0.344055	0.161906	0.220056	0.157236
3	RandomForestClassifier	0.297455	0.3052	0.301328	0.257092	0.255983	0.252142
2	DecisionTreeClassifier	0.290585	0.3032	0.296893	0.249590	0.256242	0.246491
4	SGDClassifier	0.243511	0.3418	0.292656	0.178640	0.226260	0.183049
0	GaussianNB	0.214249	0.2228	0.218525	0.138928	0.244341	0.154758

Evaluation Methodology

Classification for **Defenders** only

*---------- CLASSIFICATION MODEL PERFORMANCE EVALUATION ----------

	Model	Cross Val Score	Test Accuracy	Average_Accuracy	Precision	Recall	F1 Score
3	RandomForestClassifier	0.386565	0.4007	0.393633	0.336321	0.333281	0.329838
2	DecisionTreeClassifier	0.386579	0.3941	0.390339	0.323051	0.327529	0.322425
5	AdaBoostClassifier	0.258543	0.3388	0.298672	0.249384	0.244671	0.237207
1	SVC	0.305884	0.2899	0.297892	0.118371	0.187061	0.118444
0	GaussianNB	0.290311	0.2834	0.286855	0.277633	0.227315	0.208001
4	SGDClassifier	0.229368	0.0945	0.161934	0.117518	0.222390	0.081564

Evaluation Methodology

Rolling Statistics

- **Previous: Last game week**
- **Recent: Mean of the last four game weeks**
- **Seasonal: Mean of the last 16 game weeks**

- **Apply to the columns:**
- clean sheets
- expected goals conceded
- goals conceded
- goals scored
- assists

Evaluation Methoddology

Classification using Rolling Statistics with Defenders only

CLASSIFICATION MODEL PERFORMANCE EVALUATION

[47]:

	Model	Cross Val Score	Test Accuracy	Average_Accuracy	Precision	Recall	F1 Score
2	DecisionTreeClassifier	0.380137	0.4295	0.404818	0.348620	0.349182	0.345340
3	RandomForestClassifier	0.376156	0.4295	0.402828	0.437167	0.379999	0.388292
5	AdaBoostClassifier	0.310464	0.2853	0.297882	0.198221	0.203492	0.189788
0	GaussianNB	0.294777	0.2978	0.296289	0.256863	0.201546	0.178702
1	SVC	0.263337	0.2790	0.271168	0.046499	0.166667	0.072712
4	SGDClassifier	0.207696	0.2790	0.243348	0.138624	0.190880	0.130730

Evaluation Methodology

PCA and SVD not optimal for Classification

Evaluation Methoddology

Binary Classification

- **Class 1**: 4 points or below (0)
- **Class 2**: above 4 points (1)

Feature Selection

- Random Forest Classifier
- Decision Tree Classifier

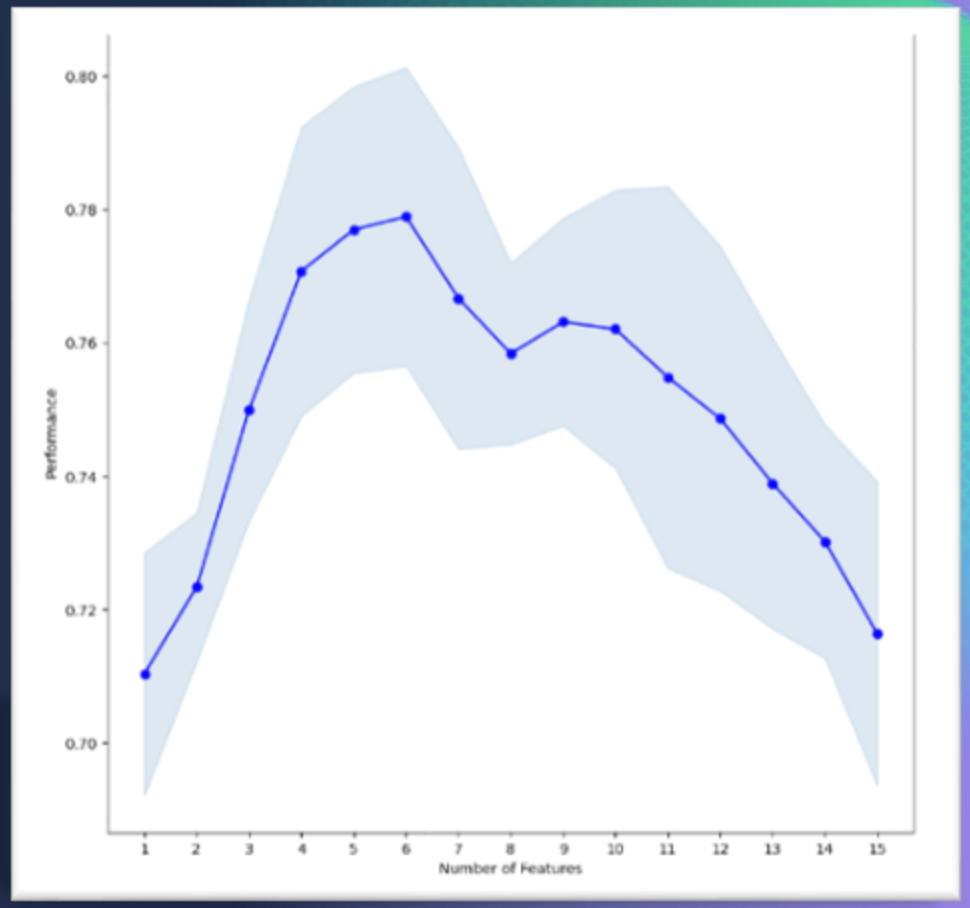
Evaluation Methodology

Feature Selection using Forward Sequential Method

Random Forest Classifier

Best features:

- round
- recent_clean_sheets
- prev_clean_sheets
- prev_expected_goals_conceded
- prev_goals_conceded
- prev_assists



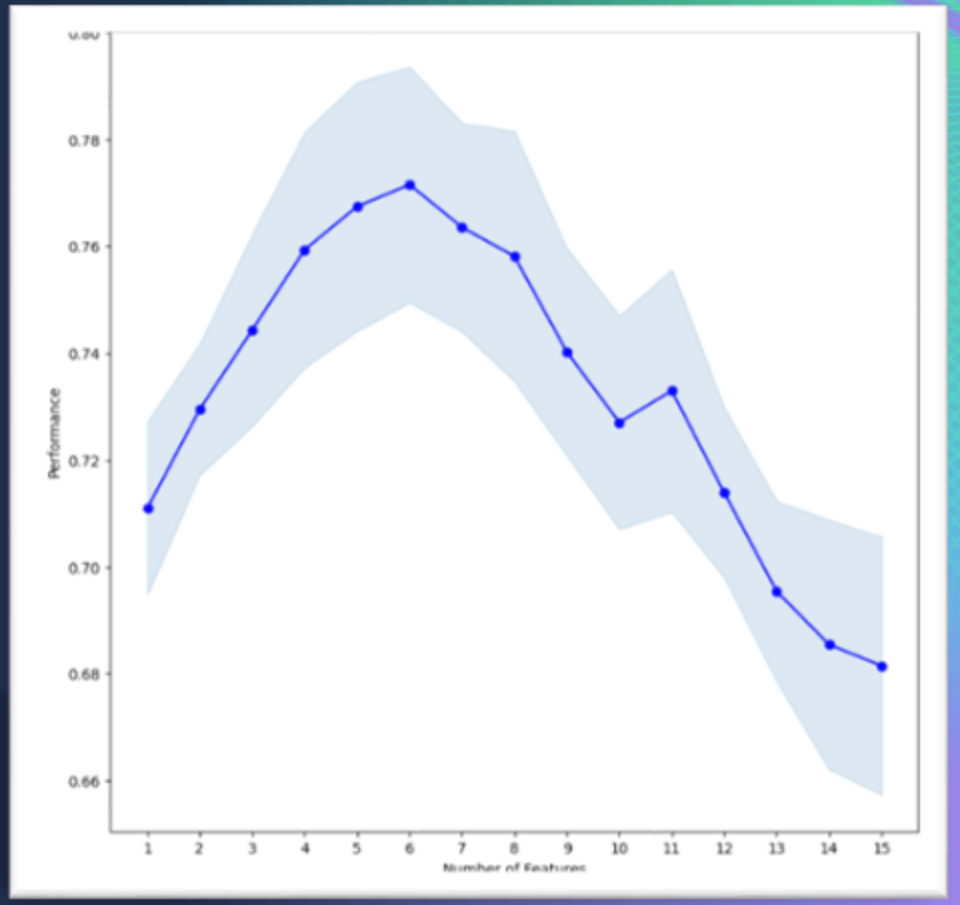
Evaluation Methodology

Feature Selection using Forward Sequential Method

Decision Tree Classifier

Best features:

- round
- recent_clean_sheets
- prev_clean_sheets
- prev_expected_goals_conceded
- prev_goals_conceded
- prev_goals_scored.



Evaluation Methoddology

Binary classification with outliers

------------------------------ CLASSIFICATION MODEL PERFORMANCE EVALUATION ----------------------

25]:

	Model	Cross Val Score	Test Accuracy	Average_Accuracy	Precision	Recall	Avg Precision Recall	F1 Score
0	XGBClassifier	0.753955	0.7962	0.775078	0.783429	0.796238	0.653557	0.782559
6	RandomForestClassifier	0.739850	0.8056	0.772725	0.795063	0.805643	0.662219	0.795560
5	DecisionTreeClassifier	0.739087	0.7837	0.761394	0.777090	0.783699	0.490184	0.779694
8	AdaBoostClassifier	0.704528	0.7398	0.722164	0.702604	0.739812	0.403286	0.659460
4	SVC	0.709246	0.7335	0.721373	0.538084	0.733542	0.259205	0.620792
1	LogisticRegression	0.707677	0.7335	0.720589	0.672394	0.733542	0.325786	0.626508
2	KNeighborsClassifier	0.697546	0.7304	0.713973	0.713417	0.730408	0.424562	0.719437
3	GaussianNB	0.700603	0.7022	0.701401	0.611095	0.702194	0.346295	0.632189
7	SGDClassifier	0.636380	0.7210	0.678690	0.581091	0.721003	0.291080	0.620047

Evaluation Methodology

Binary classification without outliers

------------------------------ CLASSIFICATION MODEL PERFORMANCE EVALUATION ----------------------

	Model	Cross Val Score	Test Accuracy	Average_Accuracy	Precision	Recall	Avg Precision Recall	F1 Score
0	XGBClassifier	0.760076	0.8089	0.784488	0.799372	0.808917	0.626189	0.792651
6	RandomForestClassifier	0.758508	0.7930	0.775754	0.780055	0.792994	0.541322	0.780234
5	DecisionTreeClassifier	0.742540	0.7739	0.758220	0.760617	0.773885	0.446842	0.763953
4	SVC	0.724311	0.7325	0.728406	0.536533	0.732484	0.257219	0.619380
1	LogisticRegression	0.726698	0.7229	0.724799	0.534643	0.722930	0.295407	0.614691
2	KNeighborsClassifier	0.725086	0.7102	0.717643	0.672996	0.710191	0.377376	0.682304
8	AdaBoostClassifier	0.720337	0.7102	0.715268	0.562876	0.710191	0.329844	0.613646
3	GaussianNB	0.717949	0.6943	0.706125	0.548938	0.694268	0.314385	0.605267
7	SGDClassifier	0.556178	0.7293	0.642739	0.663105	0.729299	0.296669	0.643831

Evaluation Methodology

Grid Search CV for Decision Tree and XGBoost Classifiers

- **Decision Tree Classifier:**

Final accuracy score on the testing data: 0.7743
Final Weighted F1 score on the testing data: 0.7696
The hyperparameters of the optimized model were
max_depth =30 ,
max_features = 0 . 5

- **XGBoost Tree Classifier:**

Final accuracy score on the testing data: 0.7994
Final Weighted F1 score on the testing data: 0.7844
The hyperparameters of the optimized model were
col_sample_by_tree =0.8
enable_categorical=False ,
eval_metric =’ mlogloss ’ ,
learning_rate =0.2 , max_depth =7,
min_ child_weight =1, missing=nan ,
n_estimators =400

Best model & Predictions

- **Random Forest Classifier**

Unoptimized model

——

Accuracy score on testing data: 0.8213

Weighted F1 score on testing data: 0.8093

Optimized Model

——

Final accuracy score on the testing data: 0.8245

Final Weighted F1 score on the testing data: 0.8121

The hyperparameters of the optimized model were,

criterion= entropy

max depth =20

n_estimators =400

Best features used:

- round
- recent_clean_sheets
- prev_clean_sheets
- prev_expected_goals_conceded
- prev_goals_conceded
- prev_assists

Best model & Predictions

We tried to predict with our model, the points of Van Dijk in

round 31

Using the parameters with values:

- round=31
- recent_clean_sheets = 0.25
- prev_clean_sheets = 0
- prev_expected_goals_conceded = 1.41
- prev_goals_conceded = 2
- prev_assists = 0

We successfully predicted that he will be classified in Class 1.

This means he scored 4 points or less in round 31.

Best model & Predictions

- Prediction for Harry Maguire

- FPL 2023-24

- Gameweek 13

Conclusion

- Better predictions scores to specific player position
- Importance of recent form – Rolling statistics
- Binary classification

Thank You, Questions?