English Premier League 24/25 Season Player Analysis and Scouting Report – By Kevan Tay

Summary

In any top tier football league, players are valued as assets and tagged with a price ranging from thousands to millions of dollars. These player valuations are determined through a multifaceted process influenced by both on-field performance and off-field considerations. Factors such as player performance, position, age and marketing factors like media hype, fan appeal can all greatly influence a player's valuation.

As a football scout or a club manager for a newly promoted team, performance on the pitch should be the most important consideration to drive team success. Hence, we will look to **identify the most value-for-money players in the English Premier League 24/25 season** for our recruitment. We will assess and rank the players based on how much impact they have made on the pitch in comparison to their market value.

Data Wrangling

Datasets:

1. English Premier League - Player Stats - 24/25
    a. Kaggle dataset containing player statistics for **562** players with **53** statistical features
    b. Contains statistics used to measure offensive performance, defensive contribution, passing & creativity, possession and ball movement, and disciplinary records.
    c. Source: https://www.kaggle.com/datasets/aesika/english-premier-league-player-stats-2425
    d. Wrangling Steps:
        i. Inspect both dataset for NULLs and duplicates
        ii. Change Percentages columns from string to float

2. Football Data from Transfermarkt
    a. A collection of CSVs scraped from reputable football statistics site Transfermarkt containing:
        i. 60,000+ games from many seasons on all major competitions
        ii. 400+ clubs from those competitions
        iii. 30,000+ players from those clubs
        iv. 400,000+ player market valuations historical records
        v. 1,200,000+ player appearance records from all games

b. Using players.csv and player_valuations.csv to obtain player valuations.
c. Source: https://www.kaggle.com/datasets/davidcariboo/player-scores/data
d. Wrangling Steps:
    i. Inspect players.csv and player_valuations.csv datasets for NULLs and duplicates
    ii. Change columns to their respective formats:
    iii. Date columns from string to DateTime
    iv. Calculate Age column using DOB
    v. Calculate the closest valuation of the player at the start of the EPL 24/25 season
    vi. Merge the two datasets to obtain a final_player_val_df containing Player Name, Player_Value_in_EUR and their Age.

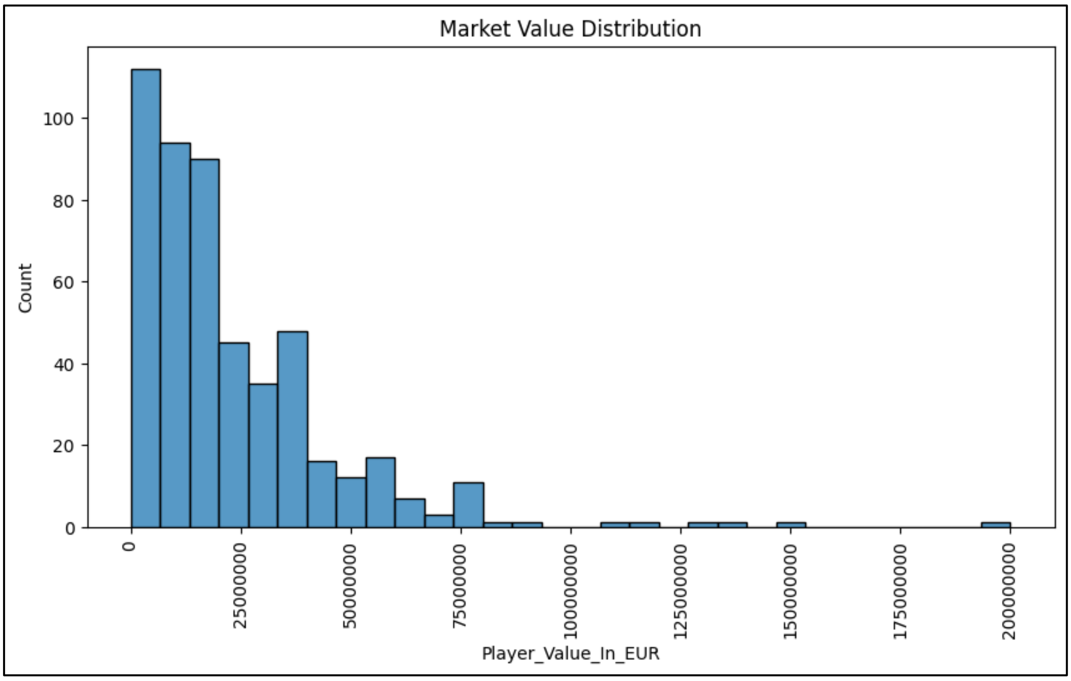
3. EPL Player Stats 24/25 with Player Valuation
    a. Merge the two datasets above using Player Name to obtain our final dataset.
    b. Used fuzzy mapping to match player names that might have mismatched due to naming convention:
        i. E.g. Heung Min Son vs Son Heung Min

Exploratory Data Analysis

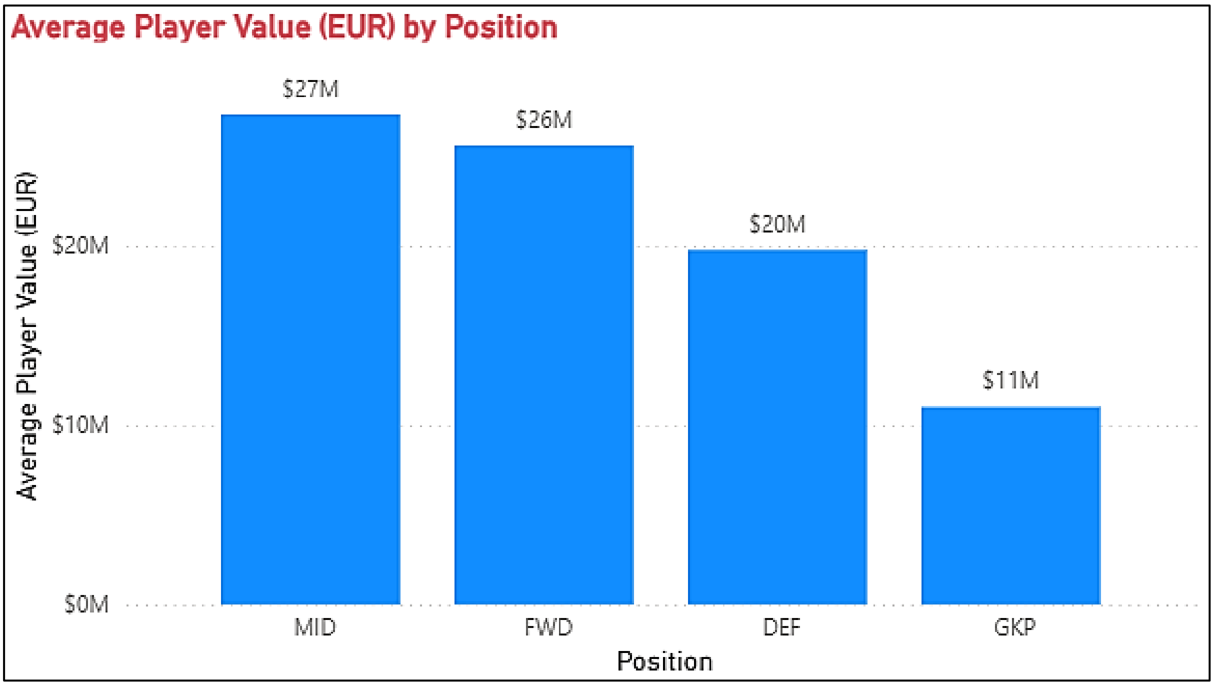
As player valuations are constantly changing throughout the season, we will use the closest valuations prior to the start of the EPL 2024/25 season on 2024-08-16. The focal point of our analysis is to determine which players strongly outperformed/underperformed compared to their valuations during the start of the season.

Player Market Value Distribution

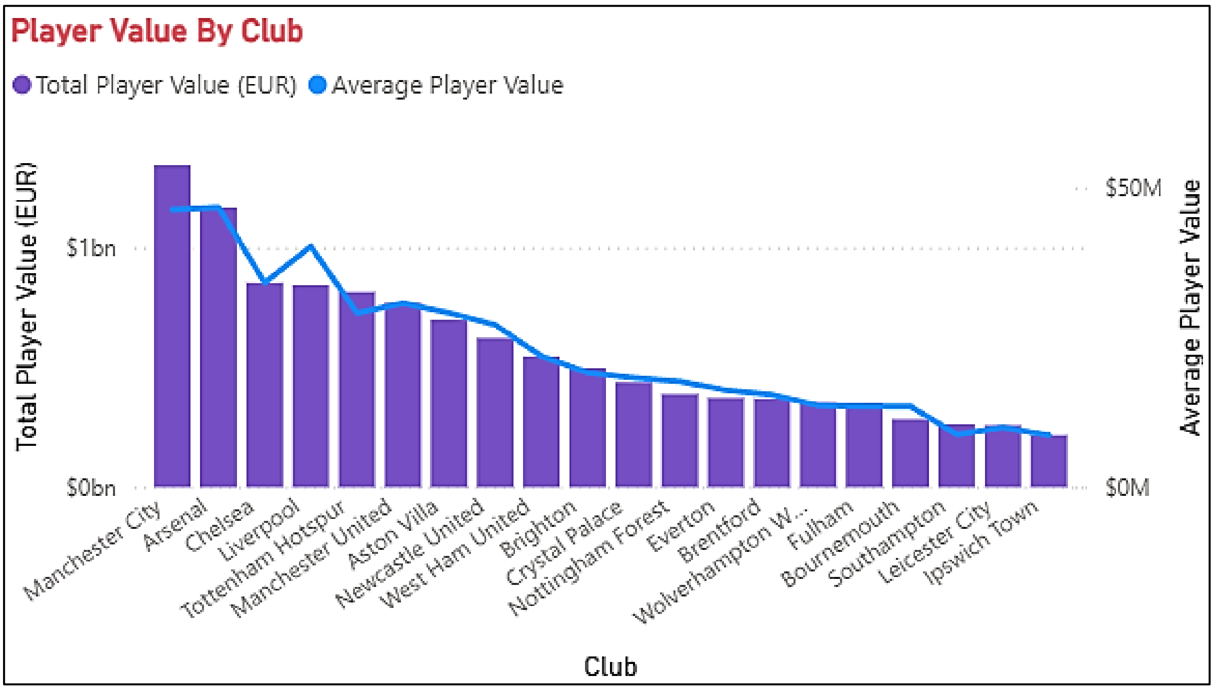
From a league wide perspective, majority of players are valued between 0 – €50,000,000, with a small number of top players such as Erling Haaland forming the outlier with a market valuation of €200,000,000.



We also observed that player positions have a noticeable trend on player valuations. Attacking positions such as Forwards and Midfielders having a higher valuation compared to defensive positions such as Defenders and Goalkeepers. Contextually, this is a common sight in the football industry due to attacking players being determining factor to wins (i.e. Goals). They also more appealing to fans with their attacking flair and are generally more marketable assets for teams.



Top and established teams in the league also have the highest average player value within the team. Arsenal, Manchester City, Liverpool, Chelsea, Manchester United and Tottenham Hotspur (often referred to as the Top 6 in the EPL) have the highest average player value and total player value. Meanwhile, newly promoted teams like Southampton, Leicester City and Ipswich Town have the lowest overall squad value.

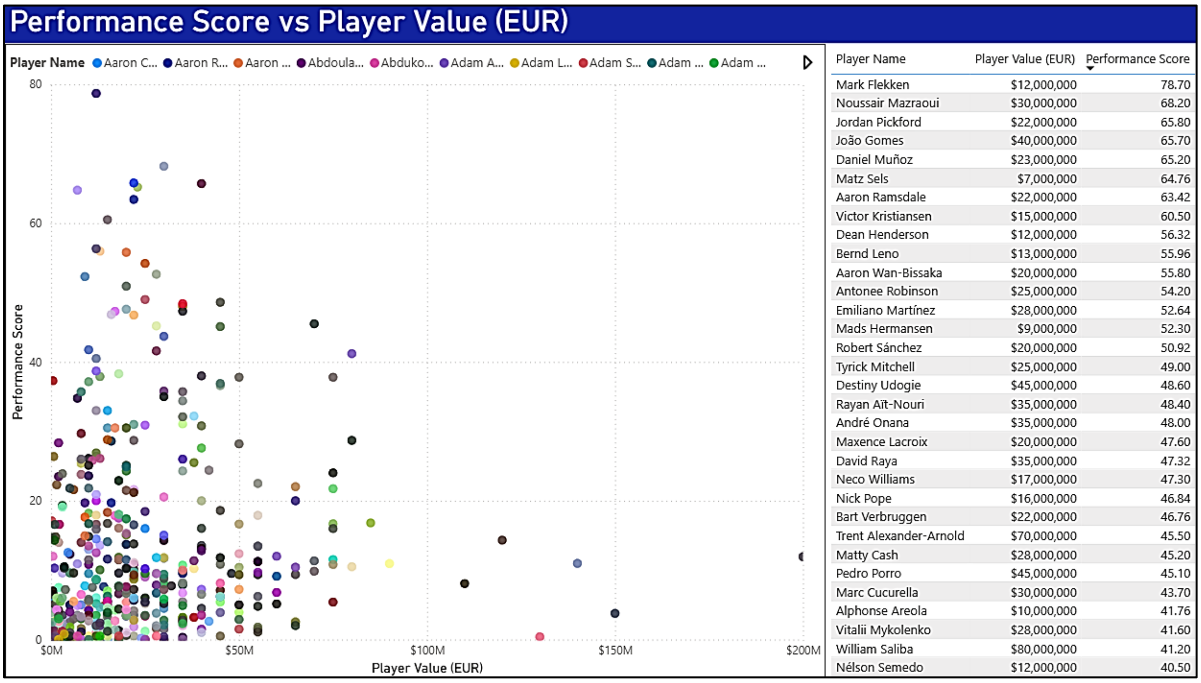


Player Performance

From a football perspective, players of different positions should be evaluated on different statistical feature. To evaluate each player on a common ground, we would create a calculated **Performance Score** using a weighted score of each feature:

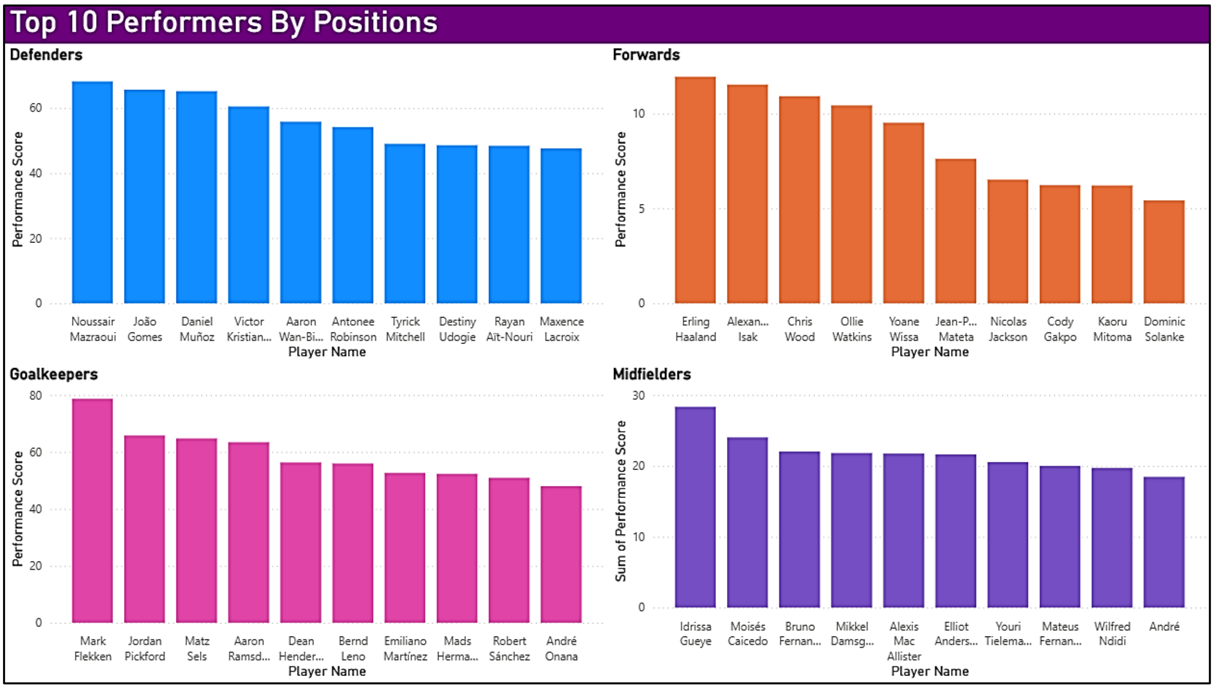
- **Weightage per feature**: 0.5 – 0.3 – 0.2
- **Forwards**: Goals - Assists - Conversion %
- **Midfielders**: Assists - fThird Passes % - Tackles
- **Defenders**: Tackles - Interceptions - Clean Sheets
- **Goalkeepers**: Saves - Clean Sheets – Goals Prevented

In terms of overall performance score, Mark Flekken, Noussair Mazraoui and Jordan Pickford forms the top 3 performers for the 24/25 season. When looking at the overall distribution of performance against the player value, we can see that they are high performers who are greatly outperforming their market value. In comparison, highly valued players (>€100,000,000) like Bukayo Saka, Declan Rice and falls in the lower end of the spectrum.



Player Name	Player Value (EUR)	Performance Score
Mark Flekken	$12,000,000	78.70
Noussair Mazraoui	$30,000,000	68.20
Jordan Pickford	$22,000,000	65.80
João Gomes	$40,000,000	65.70
Daniel Muñoz	$23,000,000	65.20
Matz Sels	$7,000,000	64.76
Aaron Ramsdale	$22,000,000	63.42
Victor Kristiansen	$15,000,000	60.50
Dean Henderson	$12,000,000	56.32
Bernd Leno	$13,000,000	55.96
Aaron Wan-Bissaka	$20,000,000	55.80
Antonee Robinson	$25,000,000	54.20
Emiliano Martínez	$28,000,000	52.64
Mads Hermansen	$9,000,000	52.30
Robert Sánchez	$20,000,000	50.92
Tyrick Mitchell	$25,000,000	49.00
Destiny Udogie	$45,000,000	48.60
Rayan Aït-Nouri	$35,000,000	48.40
André Onana	$35,000,000	48.00
Maxence Lacroix	$20,000,000	47.60
David Raya	$35,000,000	47.32
Neco Williams	$17,000,000	47.30
Nick Pope	$16,000,000	46.84
Bart Verbruggen	$22,000,000	46.76
Trent Alexander-Arnold	$70,000,000	45.50
Matty Cash	$28,000,000	45.20
Pedro Porro	$45,000,000	45.10
Marc Cucurella	$30,000,000	43.70
Alphonse Areola	$10,000,000	41.76
Vitalii Mykolenko	$28,000,000	41.60
William Saliba	$80,000,000	41.20
Nélson Semedo	$12,000,000	40.50

However, the performance score could be skewed due to the units of the features used for each position. Analysing the players within the same position would be able to provide a fairer comparison.

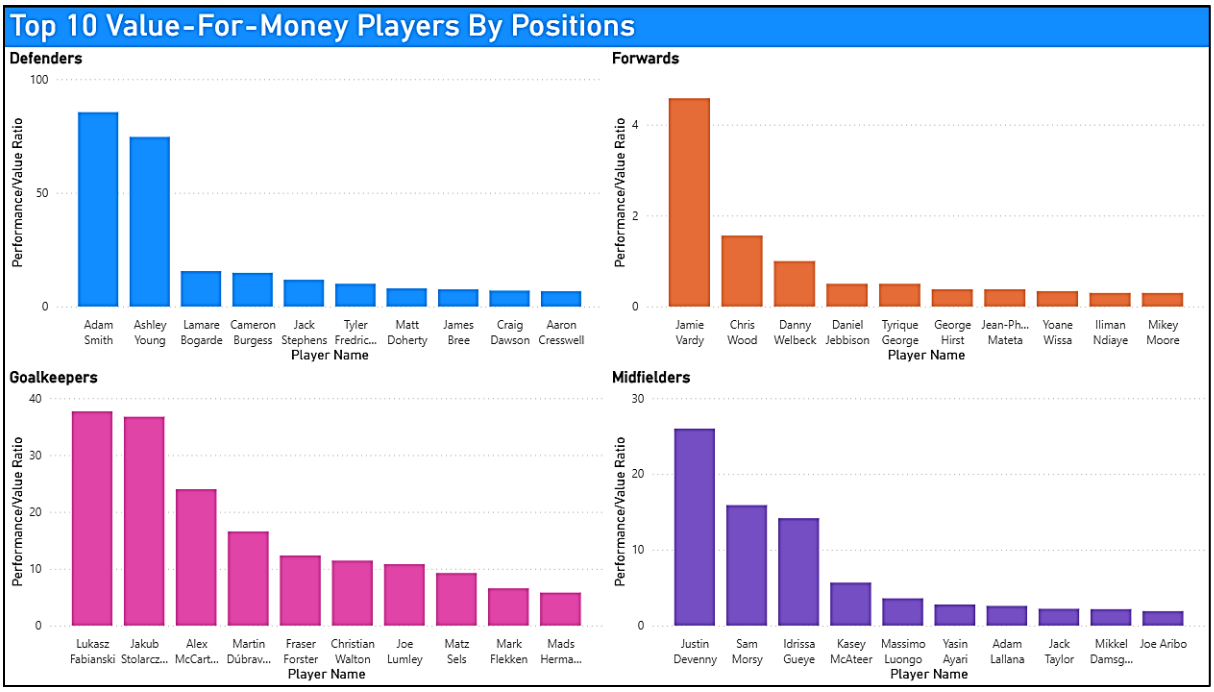


Top Forwards	Erling Haaland	Alexander Isak	Chris Wood
Top Midfielders	Idrissa Gueye	Moises Caicedo	Bruno Fernandes
Top Defenders	Noussair Mazraoui	Joao Gomes	Daniel Munoz
Top Goalkeepers	Mark Flekken	Jordan Pickford	Matz Sels

Value-For-Money Player Analysis

As the manager/team scout for a newly promoted team, the budget allocated for player recruitment is often very small in comparison to other established teams. Hence, we want to identify players that are able perform on the pitch at a high-level while being reasonable/cheaply priced. We will do this by looking at a ratio between their performance score against their player value (per million euros).

Performance/Value Ratio = *Performance Score ÷ Market Value/1000000*



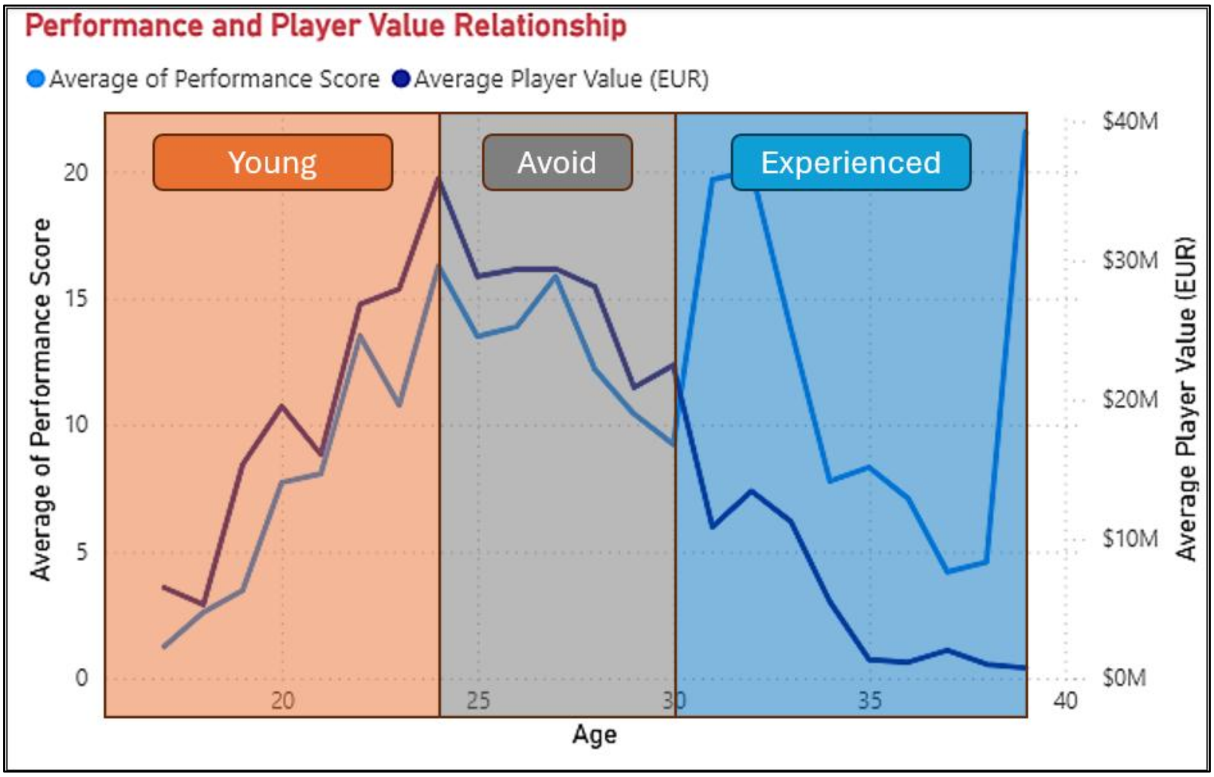
Top Forwards	Jamie Vardy	Chris Wood	Danny Welbeck
Top Midfielders	Justin Devenny	Sam Morsy	Idrissa Gueye
Top Defenders	Adam Smith	Ashley Young	Lamare Bogarde
Top Goalkeepers	Lukasz Fabianski	Jakub Stolarczyk	Alex McCarthy

From a football club perspective, players are considered valuable assets much like commodities such as gold or silver. Player recruitment is also considered as an investment. Like every investment, it is important to consider that player valuations may fluctuate based on an array of factors like player performance, position, marketability and crucially **age**.

Clubs must also balance between a short-term investment for immediate success on the pitch or a long-term investment on potential young players that can be later sold to other teams for a higher price at their prime.

- For **short-term** investments:
    - Consider experienced players that are past their peak valuations due to depreciation with their age but at the same time able to perform at a high level.

- For **long-term** investments:
  - Consider young, inexperienced players below the age of 24 that can perform at a decent level with potential for growth.
  - The club can get a return on their investment when the player is sold to other clubs after development.



Player Recommendations

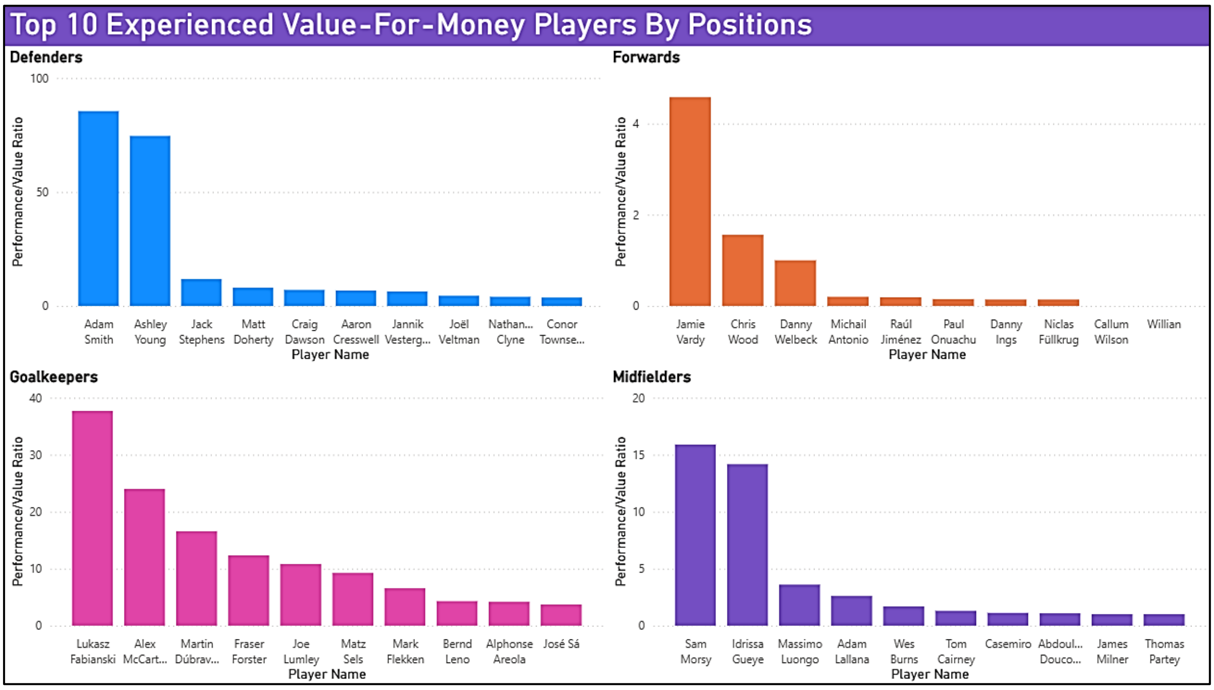
For **short-term** investments, I would recommend that the club approach the below players based on their positions.

- **Forwards**: Jamie Vardy, Chris Wood, Danny Welbeck
  - All 3 players are seasoned EPL forwards with plenty of experience under their belt.
  - Jamie Vardy is the best performer but considering his age of 38, we might only be able to secure his service for 1-2 years.
  - Else, consider to secure Chris Wood and Danny Welbeck instead for a slightly long stint (e.g. 3 years).

- **Midfielders**: Idrissa Gueye
  - One of the top performing midfielders in the league for the season.
  - Valued at only €2,000,000, you are guaranteed a high performing player that is a bargain for value.

- **Defenders**: Adam Smith

    - 32-year-old defender valued at only €200,000.
    - Able to bring aggression to the defence with his clearances and aerial duelling prowess.
    - Decent on the ball for a defender with 79% passing accuracy and ability to contribute to the offense with his fThird passing.
- **Goalkeepers**: Alex McCarthy
    - Cheap backup keeper option at 35 years old and valued at €500,000.



For **long-term** investments, I would recommend that the club approach the below players based on their positions:

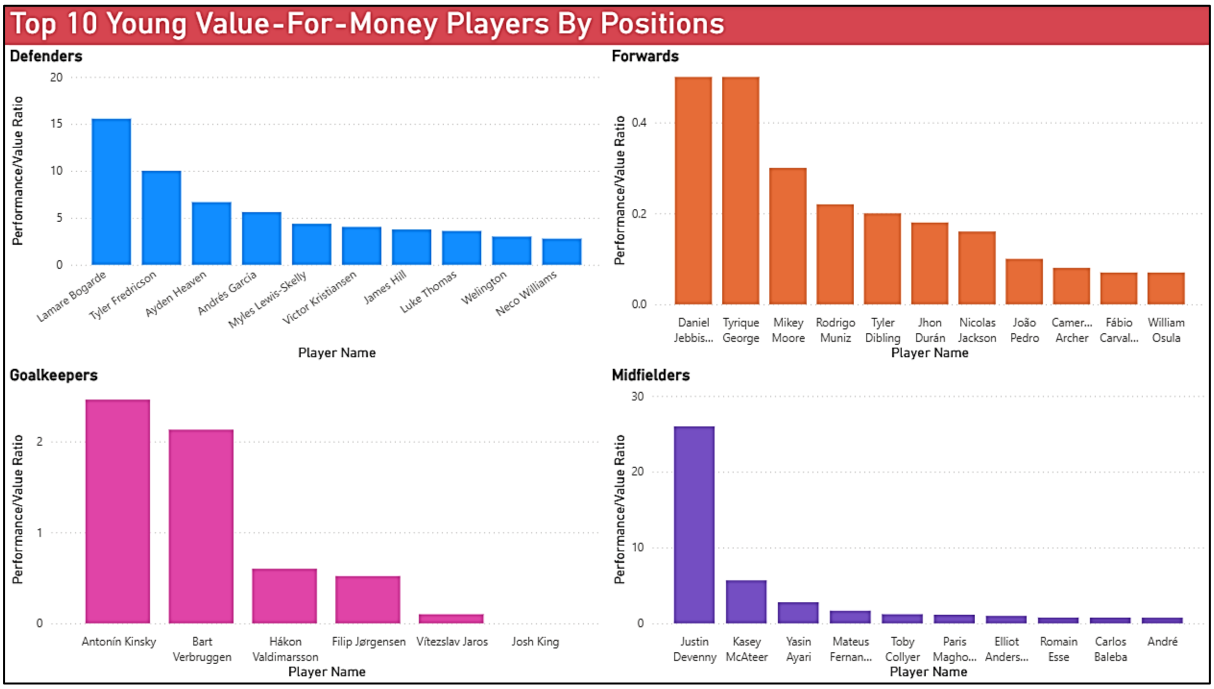
- **Forwards**: Daniel Jebbison, Tyrique George
    - Both are young talents aged 21 and 18 respectively that are valued at €1,000,000 which could be cheap investments for the future.
    - Both players are averaging ~10 mins of game time per appearance
        - Consider convincing their current clubs to send them on loan for development/more game time opportunities.

- **Midfielders**: Justin Devenney
    - Young 21 year old midfielder from Crystal Palace FC.
    - Averaging ~20 mins per appearance (Played in 23 out of 38 games).
    - For a player value of €150,000, any growth and development almost guarantee a positive Return-on-Investment (ROI).

- **Defenders**: Lamare Bogarde, Tyler Fredricson
    - Cheap and young defenders who are valued at <€500,000.

- Performing at a high performance/value comparable to senior players
- Any growth and development almost guarantee a positive ROI.

- **Goalkeepers**: Bart Verbruggen
  - While Antonin Kinsky has an overall high performance/value ratio, Bart Verbruggen is an experience starter in the league with 36 appearances and 3240 minutes played this season.
  - At only 22 years old and currently valued at €22,000,000, he could prove to be a valuable asset as he approaches his prime.

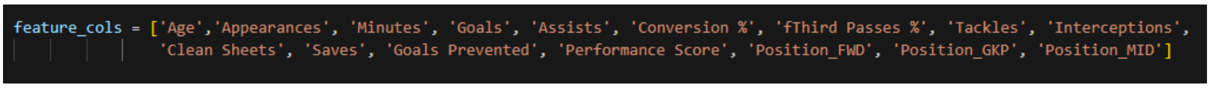


Player Value Prediction Model

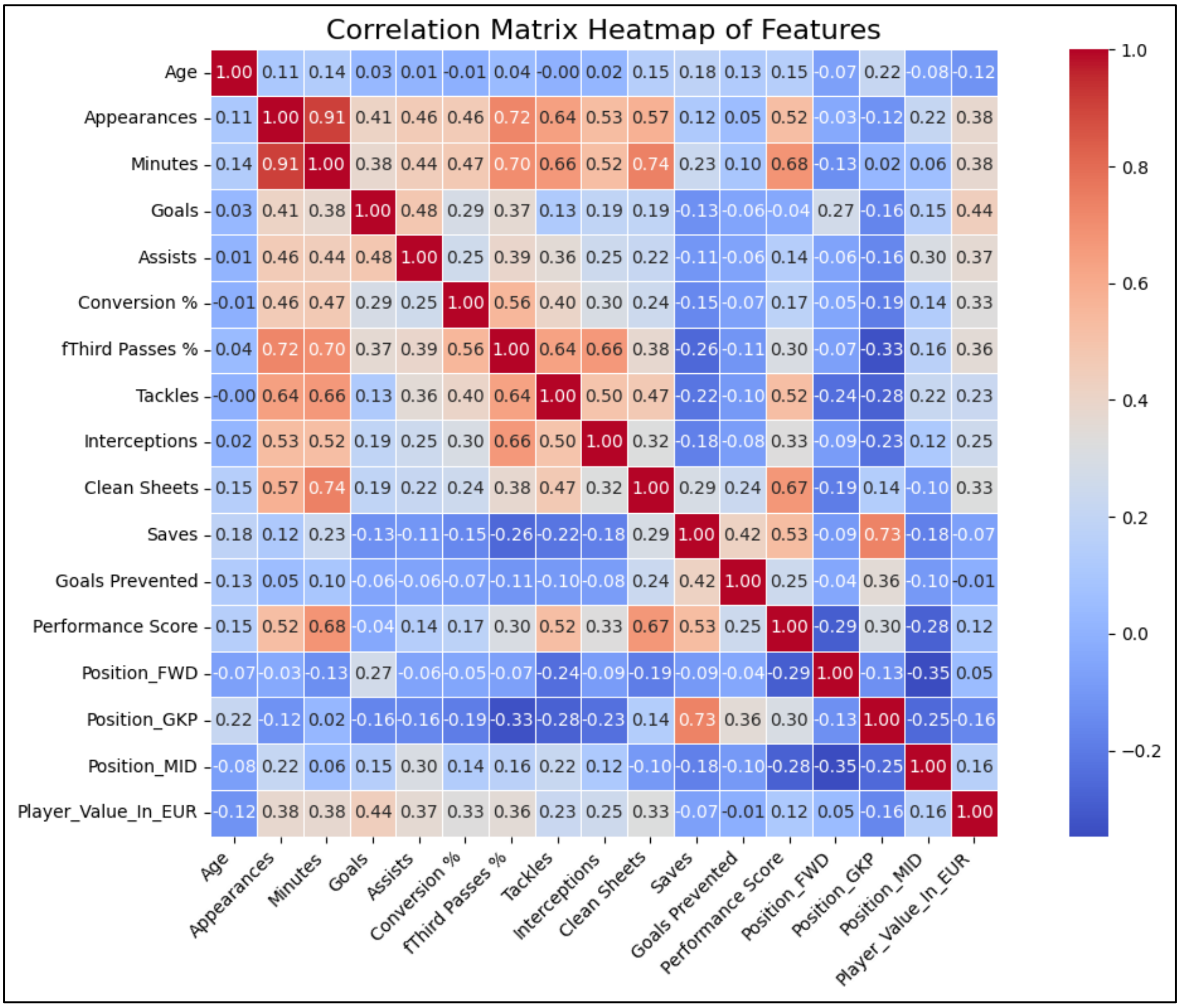
To help the club identify players that are outperforming their current market value, we would build a Player Value Prediction Model to predict the player's valuation based on their statistics for a selected number of features.

Feature Selection

Looking at the correlation heatmap for these selected features we identified to evaluate player performance, we identify the below features to be used as our first set of feature columns to train our model.



```
feature_cols = ['Age','Appearances', 'Minutes', 'Goals', 'Assists', 'Conversion %', 'fThird Passes %', 'Tackles', 'Interceptions',
                'Clean Sheets', 'Saves', 'Goals Prevented', 'Performance Score', 'Position_FWD', 'Position_GKP', 'Position_MID']
```



Correlation Matrix Heatmap of Features

	Age	Appearances	Minutes	Goals	Assists	Conversion %	fThird Passes %	Tackles	Interceptions	Clean Sheets	Saves	Goals Prevented	Performance Score	Position_FWD	Position_GKP	Position_MID	Player_Value_In_EUR
Age	1.00	0.11	0.14	0.03	0.01	-0.01	0.04	-0.00	0.02	0.15	0.18	0.13	0.15	-0.07	0.22	-0.08	-0.12
Appearances	0.11	1.00	0.91	0.41	0.46	0.46	0.72	0.64	0.53	0.57	0.12	0.05	0.52	-0.03	-0.12	0.22	0.38
Minutes	0.14	0.91	1.00	0.38	0.44	0.47	0.70	0.66	0.52	0.74	0.23	0.10	0.68	-0.13	0.02	0.06	0.38
Goals	0.03	0.41	0.38	1.00	0.48	0.29	0.37	0.13	0.19	0.19	-0.13	-0.06	-0.04	0.27	-0.16	0.15	0.44
Assists	0.01	0.46	0.44	0.48	1.00	0.25	0.39	0.36	0.25	0.22	-0.11	-0.06	0.14	-0.06	-0.16	0.30	0.37
Conversion %	-0.01	0.46	0.47	0.29	0.25	1.00	0.56	0.40	0.30	0.24	-0.15	-0.07	0.17	-0.05	-0.19	0.14	0.33
fThird Passes %	0.04	0.72	0.70	0.37	0.39	0.56	1.00	0.64	0.66	0.38	-0.26	-0.11	0.30	-0.07	-0.33	0.16	0.36
Tackles	-0.00	0.64	0.66	0.13	0.36	0.40	0.64	1.00	0.50	0.47	-0.22	-0.10	0.52	-0.24	-0.28	0.22	0.23
Interceptions	0.02	0.53	0.52	0.19	0.25	0.30	0.66	0.50	1.00	0.32	-0.18	-0.08	0.33	-0.09	-0.23	0.12	0.25
Clean Sheets	0.15	0.57	0.74	0.19	0.22	0.24	0.38	0.47	0.32	1.00	0.29	0.24	0.67	-0.19	0.14	-0.10	0.33
Saves	0.18	0.12	0.23	-0.13	-0.11	-0.15	-0.26	-0.22	-0.18	0.29	1.00	0.42	0.53	-0.09	0.73	-0.18	-0.07
Goals Prevented	0.13	0.05	0.10	-0.06	-0.06	-0.07	-0.11	-0.10	-0.08	0.24	0.42	1.00	0.25	-0.04	0.36	-0.10	-0.01
Performance Score	0.15	0.52	0.68	-0.04	0.14	0.17	0.30	0.52	0.33	0.67	0.53	0.25	1.00	-0.29	0.30	-0.28	0.12
Position_FWD	-0.07	-0.03	-0.13	0.27	-0.06	-0.05	-0.07	-0.24	-0.09	-0.19	-0.09	-0.04	-0.29	1.00	-0.13	-0.35	0.05
Position_GKP	0.22	-0.12	0.02	-0.16	-0.16	-0.19	-0.33	-0.28	-0.23	0.14	0.73	0.36	0.30	-0.13	1.00	-0.25	-0.16
Position_MID	-0.08	0.22	0.06	0.15	0.30	0.14	0.16	0.22	0.12	-0.10	-0.18	-0.10	-0.28	-0.35	-0.25	1.00	0.16
Player_Value_In_EUR	-0.12	0.38	0.38	0.44	0.37	0.33	0.36	0.23	0.25	0.33	-0.07	-0.01	0.12	0.05	-0.16	0.16	1.00

Model Training and Features Tuning

Working with the initial set of feature columns, we explored different regression models and sets of features to optimize the RMSE (Average magnitude of the prediction error) and R^2 (Variance of the dependent variable). Ideally, we would want to obtain a low RMSE score and a high R2 value.

Feature Set	Features
feature_cols	'Age', 'Appearances', 'Minutes', 'Goals', 'Assists', 'Conversion %', 'fThird Passes %', 'Tackles', 'Interceptions', 'Clean Sheets', 'Saves', 'Goals Prevented', 'Performance Score', 'Position_FWD', 'Position_GKP', 'Position_MID'
feature_cols_2	'Age', 'Appearances', 'Minutes', 'Goals', 'Conversion %', 'fThird Passes %', 'Assists', 'Tackles', 'Interceptions', 'Clean Sheets', 'Performance Score', 'Position_FWD', 'Position_GKP', 'Position_MID'
feature_cols_3	'Age', 'Appearances', 'Minutes', 'Goals', 'Assists', 'Conversion %', 'Touches', 'Passes%', 'Carries', 'Progressive Carries', 'Carries Ended with Goal', 'Carries Ended with Assist', 'Carries Ended with Shot', 'Shots', 'Shots On Target', 'Big Chances Missed', 'Hit Woodwork', 'Offsides', 'Crosses', 'Successful Crosses', 'Crosses %', 'fThird Passes', 'Successful fThird Passes', 'fThird Passes %', 'Through Balls', 'Possession Won', 'Clearances', 'Interceptions', 'Blocks', 'Tackles', 'Ground Duels', 'gDuels %', 'Aerial Duels', 'aDuels %', 'Saves', 'Saves %', 'Penalties

	Saved', 'Goals Conceded', 'xGoT Conceded', 'Goals Prevented', 'Punches', 'High Claims', 'Fouls', 'Yellow Cards', 'Red Cards', 'Position_FWD', 'Position_GKP', 'Position_MID'
feature_cols_3_refined	'Age', 'Appearances', 'Goals', 'Assists', 'Conversion %', 'Passes%', 'Carries', 'Carries Ended with Goal', 'Carries Ended with Assist', 'Carries Ended with Shot', 'Shots', 'Big Chances Missed', 'Hit Woodwork', 'Offsides', 'Crosses', 'Crosses %', 'Through Balls', 'Clearances', 'Interceptions', 'Blocks', 'Tackles', 'Aerial Duels', 'Saves', 'Penalties Saved', 'Goals Prevented', 'Punches', 'Fouls', 'Yellow Cards', 'Red Cards', 'Position_FWD', 'Position_GKP', 'Position_MID'

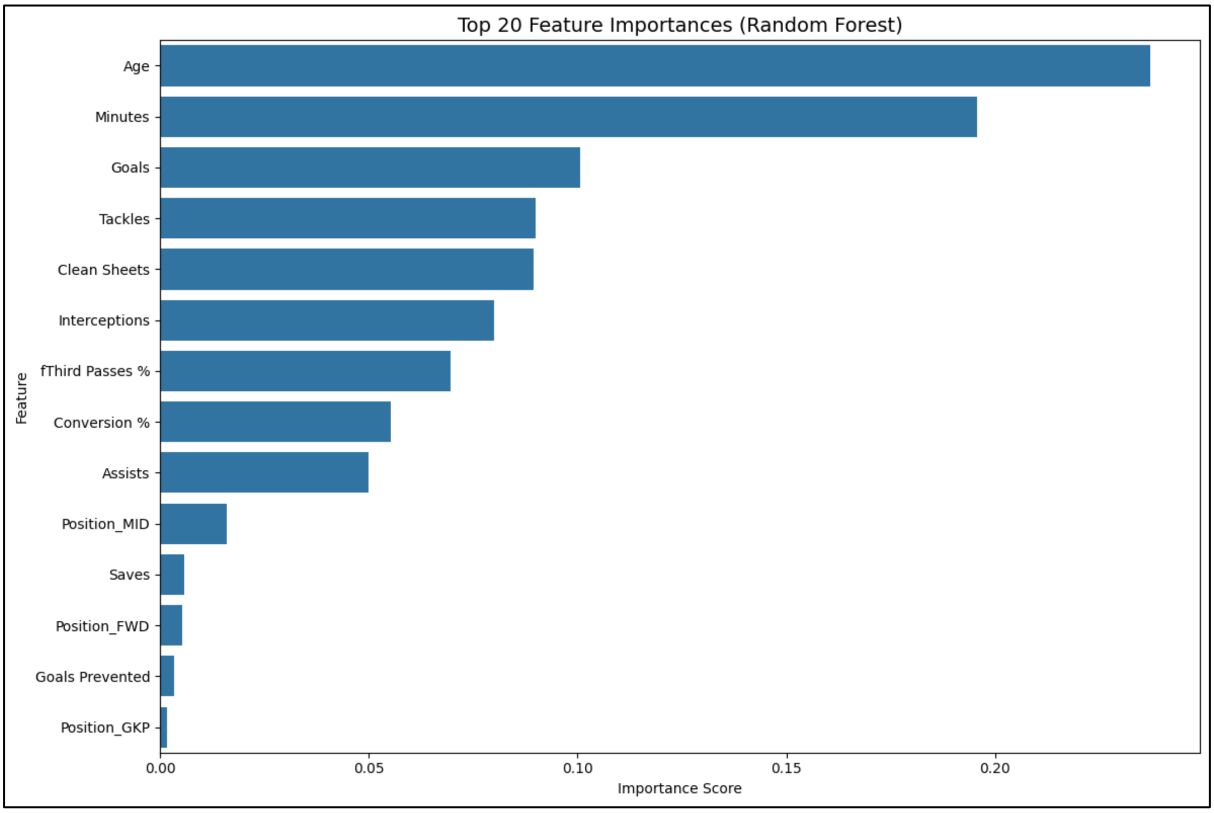
Model	Feature Set	RMSE	R^2	Remarks
DecisionTreeRegressor	feature_cols	23737968	-0.394	-
DecisionTreeRegressor	feature_cols_2	24670514	-0.505	-
RandomForestRegressor	feature_cols	16115671	0.358	-
RandomForestRegressor	feature_cols	15333680	0.419	With GridSearchCV hyperparameter tuning
RandomForestRegressor	feature_cols_2	15978510	0.369	-
RandomForestRegressor	feature_cols_2	15303500	0.421	With GridSearchCV hyperparameter tuning
Removed outliers from dataset				
RandomForestRegressor	feature_cols	13428963	0.306	
RandomForestRegressor	**feature_cols**	**13415802**	**0.307**	**GridSearchCV hyperparameter tuning**
LinearRegression	feature_cols	14226678	0.221	
LinearRegression	feature_cols_2	14198151	0.224	
PolynomialRegression	feature_cols	14226678	0.221	
PolynomialRegression	feature_cols_2	14198151	0.224	
XGBoost	feature_cols	14208326	0.223	
XGBoost	feature_cols	13953906	0.251	GridSearchCV hyperparameter tuning
RandomForestRegressor	feature_cols_3_refined	13737990	0.274	
RandomForestRegressor	feature_cols_3_refined	13430079	0.306	GridSearchCV hyperparameter tuning
KNNRegressor	feature_cols_3_refined	13777584	0.269	
KNNRegressor	feature_cols_3_refined	14063634	0.239	GridSearchCV hyperparameter tuning

Final Model and Features

Since RandomForestRegressor with feature_cols is the best performing, we will stick to this combination. However, we can further tune the feature_cols by removing feature that have similar meanings with other features such as - Appearances, Performance Score. Ultimately, we were able to achieve an RMSE of ~€13 million and R^2 of 0.34.

Feature Set	Features
feature_cols_4	'Age', 'Minutes', 'Goals', 'Assists', 'Conversion %', 'fThird Passes %', 'Tackles', 'Interceptions', 'Clean Sheets', 'Saves', 'Goals Prevented', 'Position_FWD', 'Position_GKP', 'Position_MID'

Model	Feature Set	RMSE	R^2	Remarks
RandomForestRegressor	**feature_cols_4**	**13098958**	**0.340**	**FINAL**
RandomForestRegressor	feature_cols_4	13162410	0.333	GridSearchCV hyperparameter tuning



Conclusion

With an RMSE score of ~€13 million, we would expect our prediction for the player's value to have a rather large variance based on this dataset. This could be possibly due to a small sample size of players since we are analysing based on the English Premier League alone. Moving forward, we can further enhance the model by expanding the dataset to other leagues or include past historical data into consideration. However, the model can still be used for scouts and managers to identify underrated players who are outperforming their valuations.