

S.C.O.R.E.

*Soccer Club Optimization of
Recruitment Expenses*

Cody Tipton, Rafael Magaldi



Background and Objective

Soccer Transfer Landscape

- **Dynamic Transfer Market:** The soccer transfer market is highly competitive and volatile, with player valuations frequently fluctuating. Clubs have access to two transfer windows per season, when thousands of transfers take place.
- **Financial Implications:** Accurate player valuations are crucial for clubs to make informed financial decisions, avoiding overpayment and ensuring sustainable investments that pay off in the field.

Goal

- Develop a predictive model to accurately estimate player values
- Target: **Market Value**
- Features: **Player Statistics**

#	↓	Player	Age	Market value	↓	Nat.	Joined	Fee	↓
1		 Julián Alvarez Centre-Forward	24	€90.00m			 Atlético Madrid LaLiga	€75.00m	
2		 Dominic Solanke Centre-Forward	26	€40.00m			 Tottenham Premier League	€64.30m	
3		 Leny Yoro Centre-Back	18	€50.00m			 Man Utd Premier League	€62.00m	
4		 Pedro Neto Right Winger	24	€55.00m			 Chelsea Premier League	€60.00m	
5		 Moussa Diaby Right Winger	25	€55.00m			 Al-Ittihad Saudi Pro League	€60.00m	
6		 João Neves Defensive Midfield	19	€55.00m			 Paris SG Ligue 1	€59.92m	
7		 Amadou Onana Defensive Midfield	22	€50.00m			 Aston Villa Premier League	€59.35m	
8		 Dani Olmo Attacking Midfield	26	€60.00m			 Barcelona LaLiga	€55.00m	
9		 Teun Koopmeiners Attacking Midfield	26	€50.00m			 Juventus Serie A	€54.70m	
10		 Michael Olise Right Winger	22	€55.00m			 Bayern Munich Bundesliga	€53.00m	

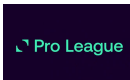
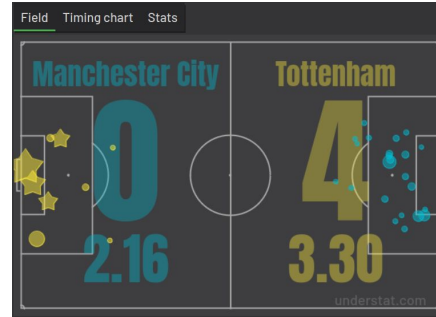
Data

Sources

- **Transfermarkt** - Kaggle
- **Understat** - Web scraping
- **Sofascore** - Web scraping + API

Combined Dataset

- Over **10k players** from across the top 10 strongest soccer leagues in Europe, including the English Premier League, La Liga, Bundesliga, etc...
- Stats and personal information for each player in every league match across many seasons



KPIs and Stakeholders

Key Performance Indicators

- Identify key features that impact market value
- Predict market value accurately
- Create a model that allows actionable insight

Stakeholders

- Soccer clubs looking to negotiate players during transfer windows
- Fantasy Soccer players who seek accurate player valuations to optimize their fantasy teams



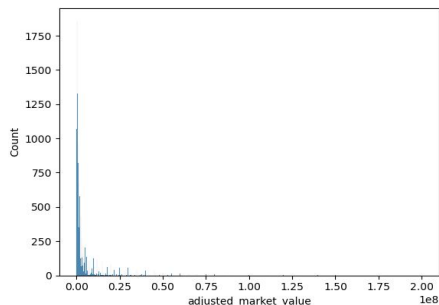
Feature Engineering and Target Preprocessing

Player Statistics and Information

- In-game stats aggregated, including:
 - Expected goals
 - Saves
 - Minutes played
 - Accurate passes
 - etc..
- Personal information:
 - Name
 - Date of birth
 - Preferred foot
 - Height
 - Position
 - etc..

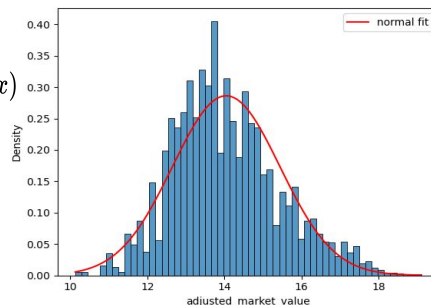
Market Value

- Used last known market value for each player
- Adjusted for inflation
- Applied $\ln(1+x)$ to correct the skewness



High Skewness and Kurtosis
Causes negative predictions in LR

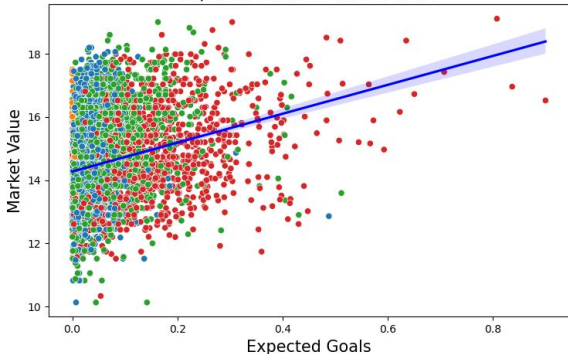
$\ln(1+x)$



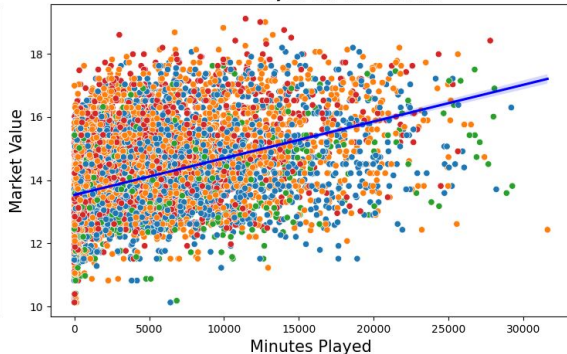
Approximately normally distributed
Diminishes outliers

Exploratory Data Analysis

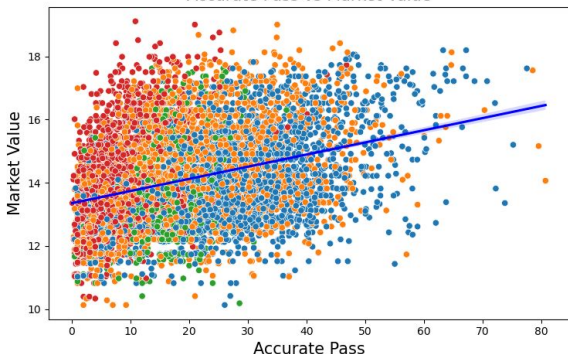
Expected Goals vs Market Value



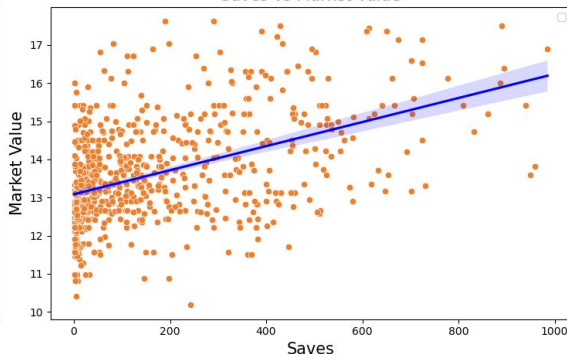
Minutes Played vs Market Value



Accurate Pass vs Market Value

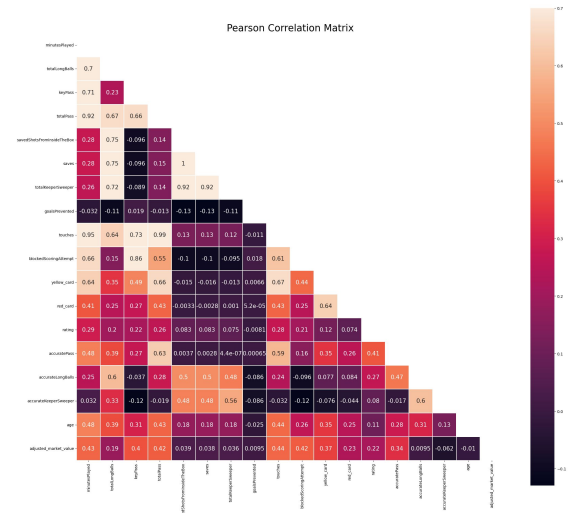


Saves vs Market Value



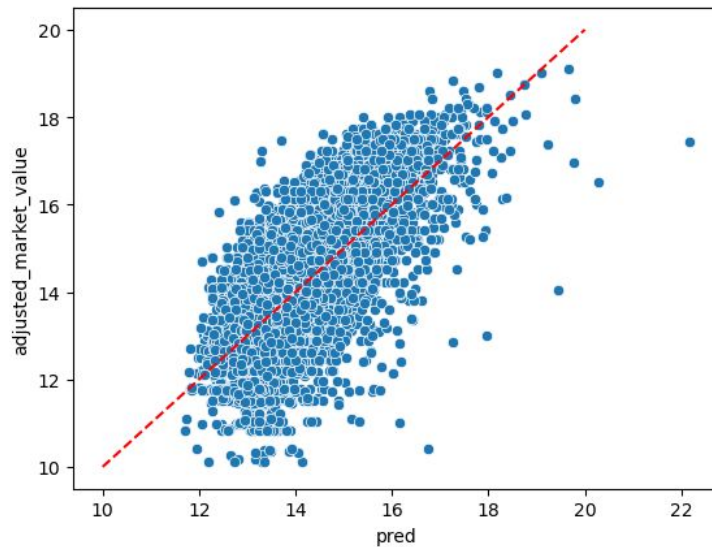
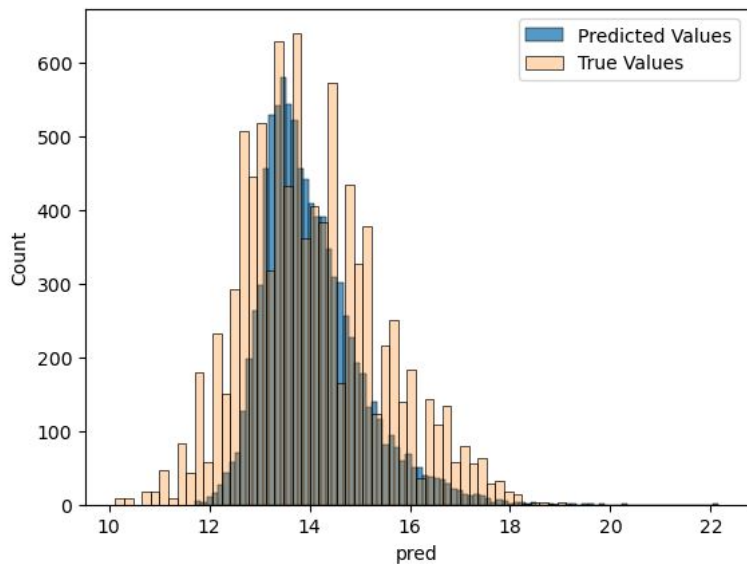
Positions
 ● D
 ● G
 ● M
 ● F

Pearson Correlation Matrix



Baseline Model

Linear Regression with All Features

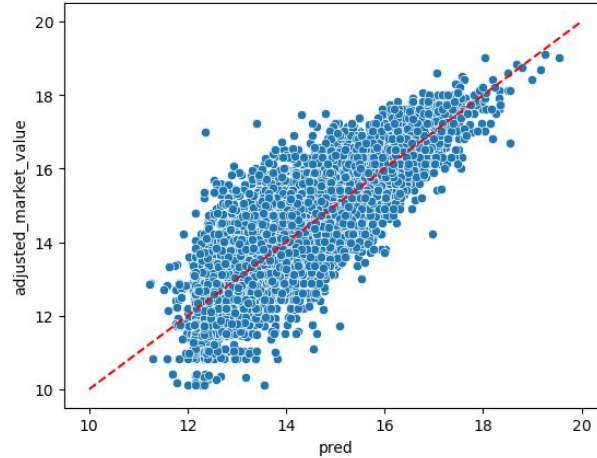
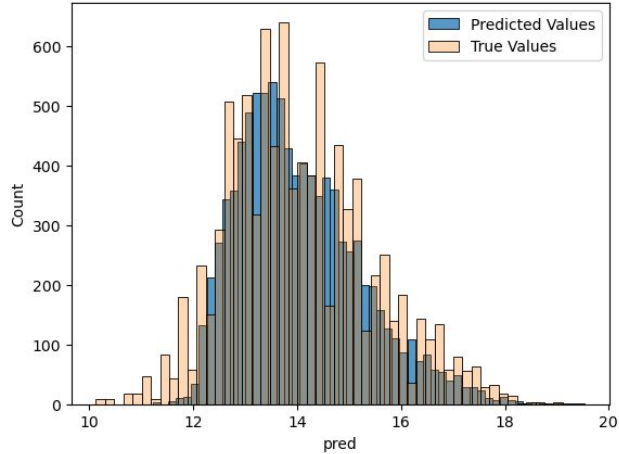


RMSE:

- Train: 0.986
- Test: 1.017

Better Model

Gradient Boosting with All Features

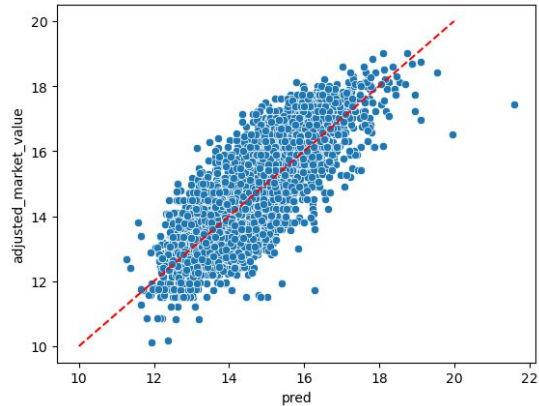
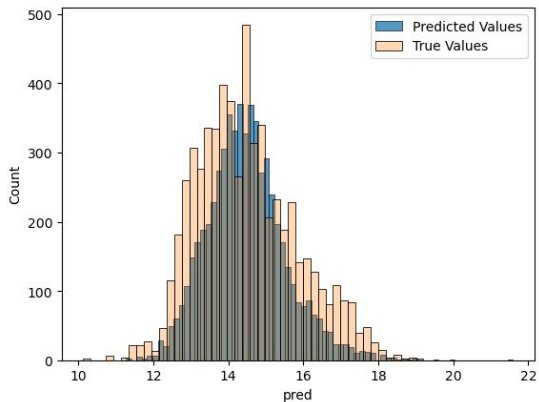


RMSE

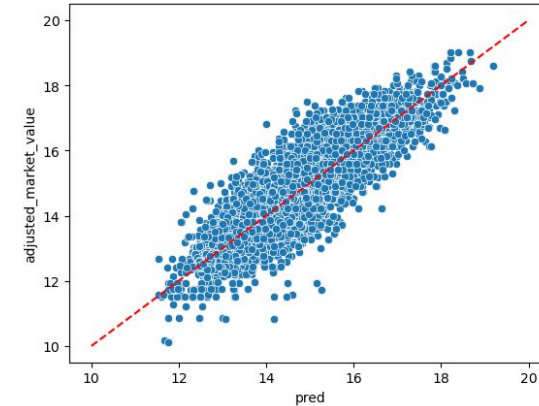
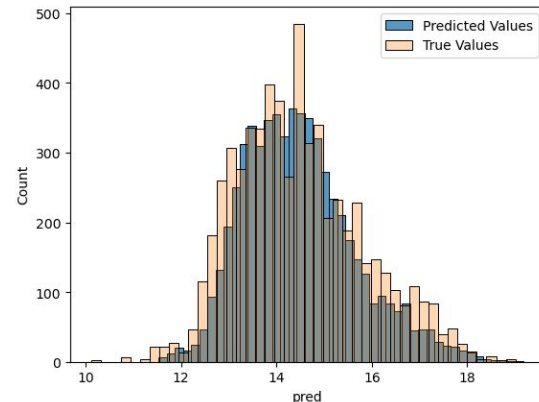
Pos	Train	Test
G	0.787	0.947
D	0.758	0.837
M	0.782	0.861
F	0.804	0.853
All	0.780	0.859

1000 Minutes Played Models

Linear Regression



Gradient Boosting



RMSE

LR

GB

Pos	Train	Test	Train	Test
G	0.887	0.811	0.735	0.785
D	0.788	0.857	0.659	0.822
M	0.827	0.841	0.708	0.810
F	0.876	0.871	0.684	0.765
All	0.828	0.841	0.687	0.775

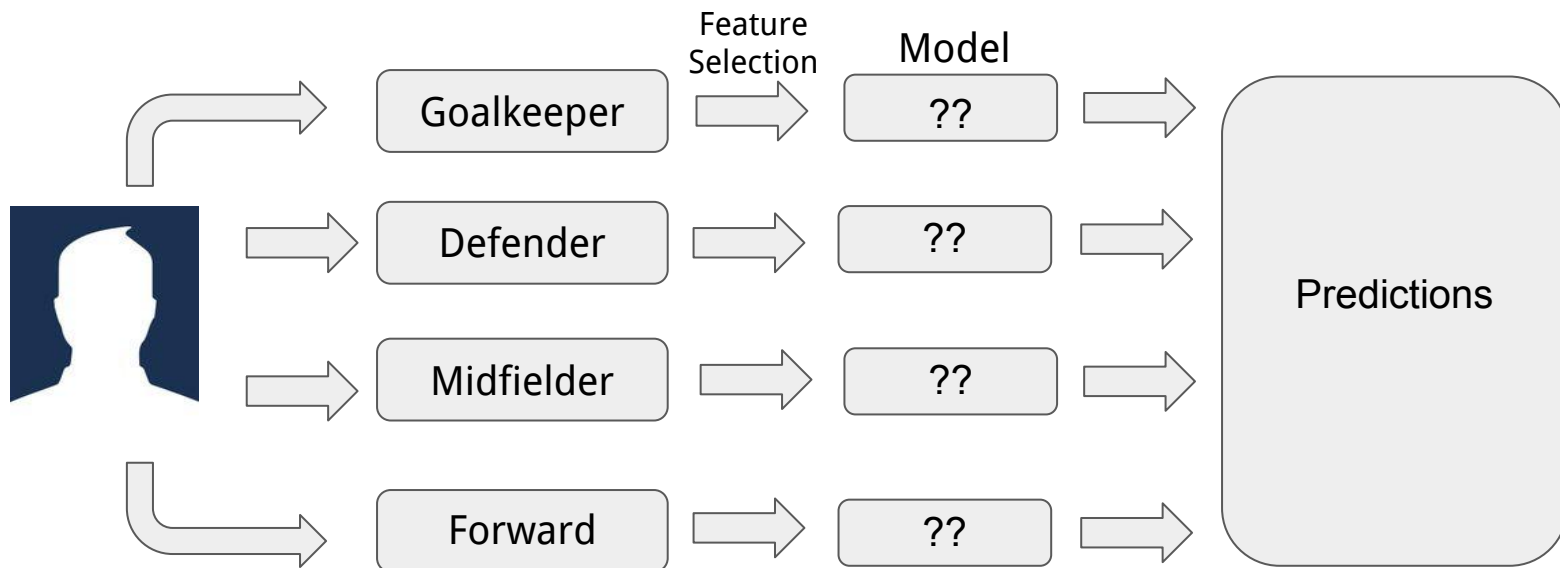
Ensemble Models

Position-Specific Features

- Divided players into positions
- Domain knowledge for choosing relevant features

Testing Models

- Tested a range of models on each position
- Performed extensive **hyperparameter tuning** on each model



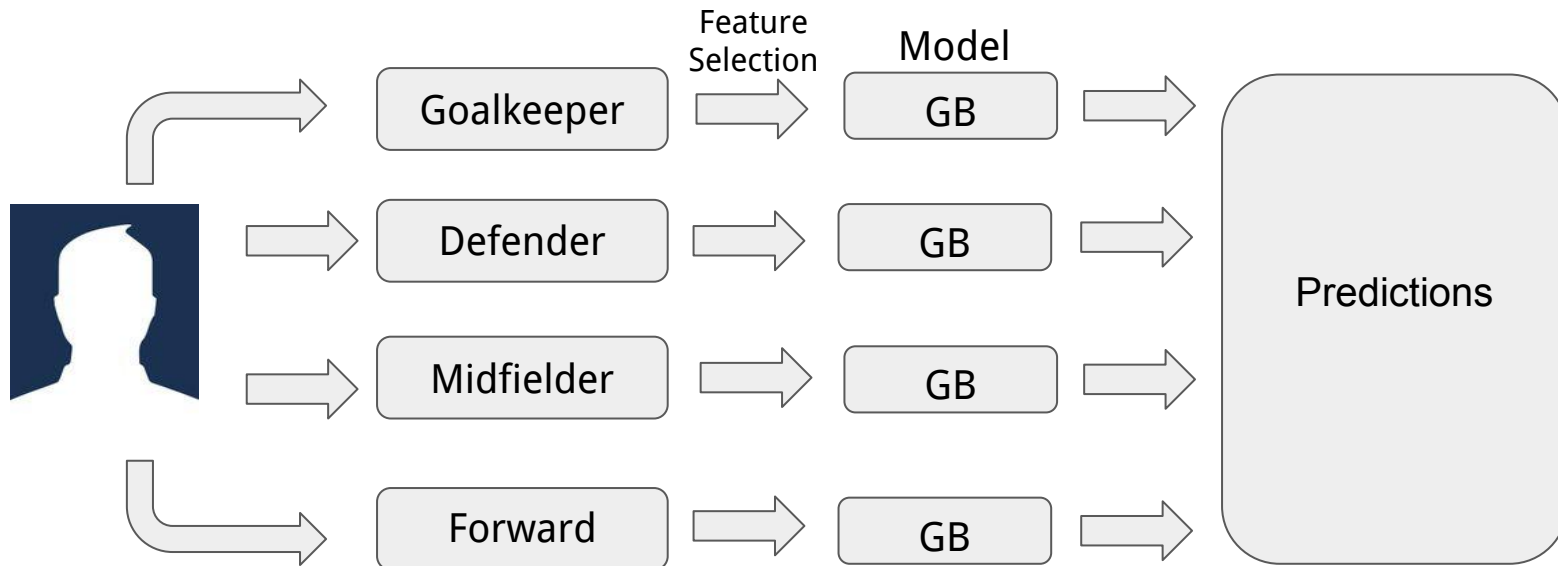
Ensemble Models

Position-Specific Features

- Divided players into positions
- Domain knowledge for choosing relevant features

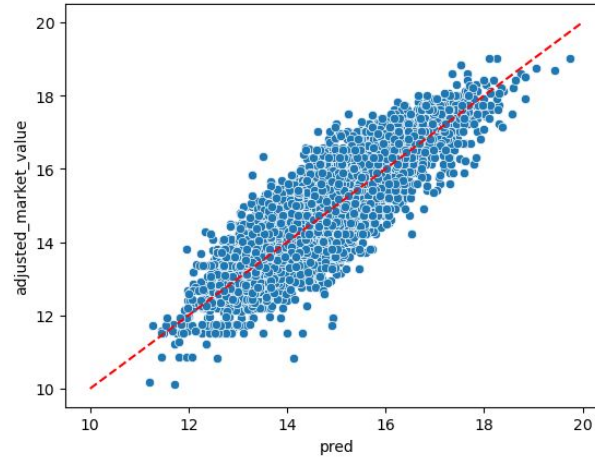
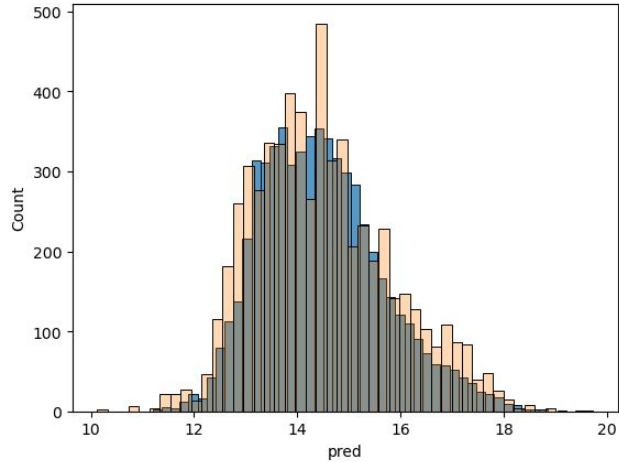
Testing Models

- Tested a range of models on each position
- Performed extensive **hyperparameter tuning** on each model



Ensemble Models - Results

HP-tuned Gradient Boosting



RMSE

Pos	Train	Test
G	0.505	0.980
D	0.653	0.833
M	0.709	0.828
F	0.661	0.822
All	0.666	0.842

Model Comparisons

GB - All Features



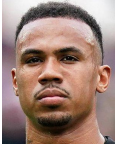
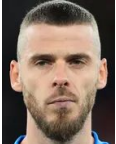
LR - 1000 mins

GB - 1000 mins

GB - HPtuned

Pos	Train	Test	% Improvement	Train	Test	% Improvement	Train	Test	% Improvement	Train	Test	% Improvement
G	0.787	0.947	6.88	0.887	0.811	20.26	0.735	0.785	22.81	0.505	0.98	3.63
D	0.758	0.837	17.70	0.788	0.857	15.73	0.659	0.822	19.17	0.653	0.833	18.09
M	0.782	0.861	15.34	0.827	0.841	17.31	0.708	0.81	20.35	0.709	0.828	18.58
F	0.804	0.853	16.12	0.876	0.871	14.36	0.684	0.765	24.78	0.661	0.822	19.17
All	0.78	0.859	15.54	0.828	0.841	17.31	0.687	0.775	23.79	0.666	0.842	17.20

Some Predictions

Player	True Value	Predicted Value
Randal Kolo Muani (F) 	€40.00m	€30.70m
Kees Smit (M) 	€800k	€4.85m
Gabriel Magalhães (D) 	€75.00m	€32.51m
David de Gea (G) 	€5.00m	€3.80m

Conclusions

Modeling

- Key ways we tried improving upon the baseline:
 - Position-specific features
 - Minutes played threshold
 - Range of models and hyperparameter tuning
- Best model (GB with 1000 minutes cutoff) improved upon the baseline by ~24%

Main Difficulties

- Outliers with incredibly high market values, or minimal minutes played, can reduce the generalization power of the models
- Too much overfitting when using models other than linear regression

Future Goals

Model Improvements

- Gather data from more competitions:
 - Leagues in other countries
 - National and international cups
- Create synthetic data for training
- Include information about "decisiveness" and "star power" of players:
 - Titles won
 - Individual awards
 - Popularity (jerseys sold, online following, etc)
- More hyperparameter tuning
- Create different models for players in each country
- Give more weight to most recent statistics
- Test different modeling approaches, like time series