Predicting Highest Scoring Players for Fantasy Premier League

Anna Li, Hisham Aziz, Iishaan Shekhar, Katherine Voss-Robinson

We are motivated by the goal of optimizing a starting 11 for a Fantasy Premier League (FPL) game week

- **Research Question:** How can we predict FPL player performance per game week?

- FPL is extremely popular, with over 11 million players worldwide

- Creating a high-scoring team is challenging and can be lucrative
  - Additionally, this topic comes with a wealth of data
  - This problem is ripe for an ML solution

Speaker: Hisham
- When presenting, hit home that it's an interesting, challenging question
- English Premier League

We will use an FPL library containing game week-level data from 2016 through the present to build our model(s)

Data source

Link: https://github.com/vaastav/Fantasy-Premier-League/tree/master/data

Columns

In total, we have 37 columns (split between **continuous** and **categorical** variables) and ~140k rows (one for each player and game week in the data)

['season_x', 'name', 'position', 'team_x', 'assists', 'bonus', 'bps', 'clean_sheets', 'creativity', 'element', 'fixture', 'goals_conceded', 'goals_scored', 'ict_index', 'influence', 'kickoff_time', 'minutes', 'opponent_team', 'opp_team_name', 'own_goals', 'penalties_missed', 'penalties_saved', 'red_cards', 'round', 'saves', 'selected', 'team_a_score', 'team_h_score', 'threat', 'total_points', 'transfers_balance', 'transfers_in', 'transfers_out', 'value', 'was_home', 'yellow_cards', 'GW']

Speaker: Hisham

Very importantly, one that is updated every week as the current season unfolds

After excluding records missing team name and for players who played 0 minutes, we are left with ~32k rows

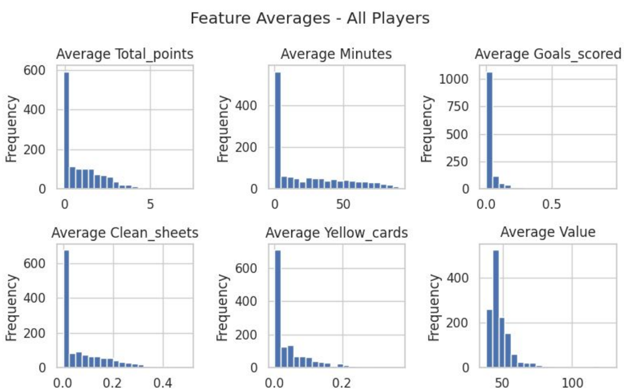
Season	Overall Records	Played >0 Minutes	Record Contains Team Name	Distinct Players with > 0 Minutes
2016-17	8,567	5,139	-	-
2017-18	11,285	6,584	-	-
2018-19	21,790	10,480	-	-
2019-20	22,560	10,614	-	-
2020-21	24,365	10,393	10,393	524
2021-22	25,447	10,485	10,485	537
2022-23	26,505	11,345	11,345	554

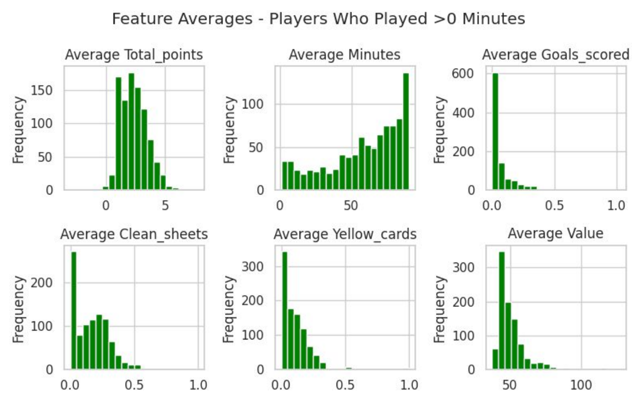
Speaker: Hisham

- Notable observations:
  - Records are sparse in 2016-2017 - we may exclude if we can't figure out why
  - We suspect we can get team name from an alternate source, so we're not counting out these rows yet

Continuous Variables

Notably, the distributions of our data change dramatically when filtering out players who played 0 minutes





Speaker: Katherine

- This is an illustrative subset of our features
- It highlights how much the volume of the “no minutes” players overwhelms the rest
  - We will either remove rows that didn’t play or downsample them so that they don’t overwhelm our model
- Also note that the variables are on different scales; we’ll need to normalize
  - Here, value is in hundreds of millions

Continuous Variables

Many of our numerical features are on different scales; we will standardize them before building our model

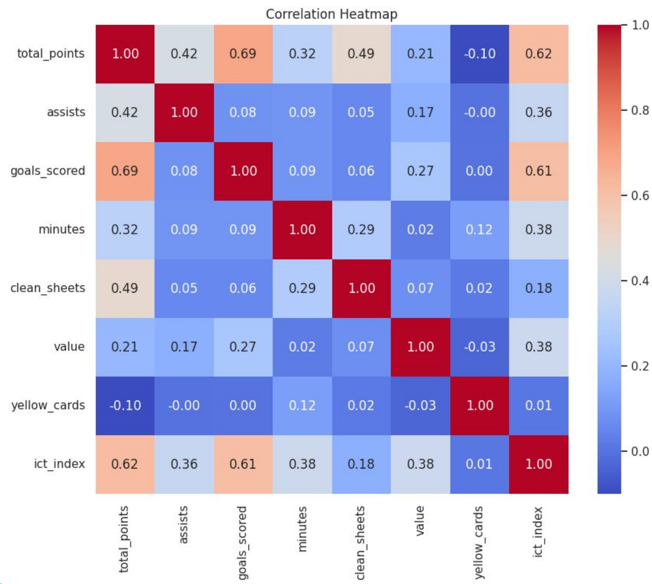
	Clean Sheets	Yellow Cards	Assists	Goals Scored	ICT Index	Minutes	Value (in $100M)
Mean	0.22	0.12	0.086	0.095	3.58	69.76	54.82
Standard Deviation	0.42	0.32	0.30	0.33	3.45	29.94	14.25
Min	0	0	0	0	0	1	37
Max	1	1	4	4	32.8	90	133

Speaker: Katherine

- Again, these are an illustrative subset of variables
    - Data has been filtered for play time, which is why minutes has a minimum of 1

Continuous Variables

After filtering out players who did not play, several of our numerical features are highly correlated with total points



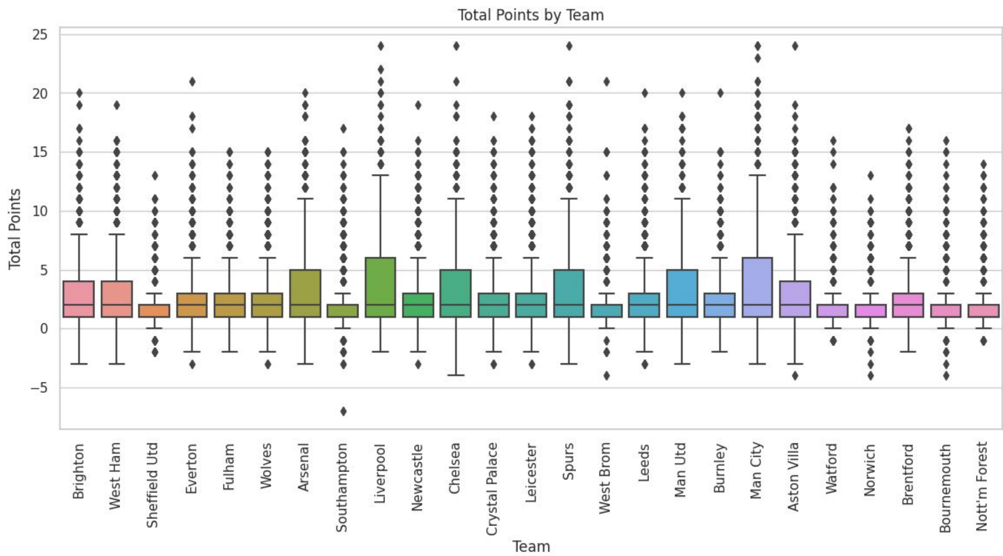
- To feed these into our model, we will need to transform them into lagged rolling averages
  - E.g., feed in “average assists over the prior 3 game weeks” rather than just “assists”
  - Any player who did not play at all in the past 3 weeks will get a “0” for all variables
- We will add a feature for “opposition difficulty”

Speaker: Katherine

- These variables are a subset of our numerical variables chosen based on domain knowledge for the sake of this presentation (we will analyze all of them)
- Notable strong correlations: goals_scored, clean_sheets, ict_index (all unsurprising, as is the fact that yellow cards is negative)
- Data dictionary for less obvious terms:
  - Clean Sheets - when a team concedes zero goals in a match
  - Value - FPL dollar value for a given player
  - ICT Index - statistical index generating a single score for three key areas: influence, creativity, and threat
    - Influence - evaluates the degree to which a player has made an impact on a single match or throughout the season
    - Creativity - assesses player performance in terms of producing goalscoring opportunities for others. It can be used as a guide to identify the players most likely to supply assists
    - Threat - examines a player's threat on goal. It gauges the individuals most likely to score goals

Categorical Variables

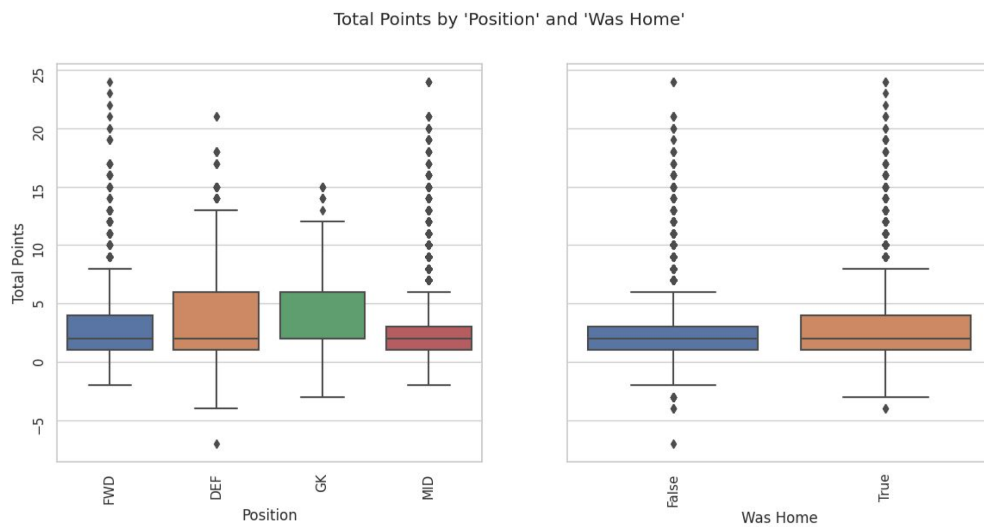
We expect team to be an important predictor variable (both Played For and Against)



Speaker: Anna

Categorical Variables

We also expect the “position” and “was home” variables to be important predictors



Speaker: Anna

We plan to evaluate and choose between several types of predictive algorithms

- We will aim to use regression to predict points for a given player in a given game week
- Target variable: total_points
- Prediction algorithm options:
  - Linear regression
  - Random forest of regressive decision trees
    - Additional optimized versions using ensemble methods (e.g., boosting, bagging, and stacking)
  - Feed forward neural networks

Speaker: Iishaan

- Link

We will evaluate our results using test data, measurements of loss, and a t-test to choose our final model

- We will choose between two methodologies to split our data:
  - **Simple holdout** - use the 2016-17 through 2020-21 seasons for train, the 2021-22 season for validation, and the 2022-23 season for test (a roughly 60/20/20 split)
  - **K-fold cross-validation** - ignoring the order of data, split the rows in our data set randomly across seasons in a 60/20/20 split

- We will compare our model to our baseline using root mean squared error (RMSE) and mean absolute percentage error (MAPE)

- We will conduct a t-test on the above metrics for each of our possible models and select the best among them, using simplicity as a tie break, and, as new data streams from the current premier league season (23-24), we will test the top model's performance against this 'unseen' data.

Speaker: Iishaan

- Potentially ask - would looking at R^2 be valuable?