Adam Corren

Using Machine Learning to Mitigate Human Bias: Fantasy Football

PROJECT REVIEW

MARCH 2023

What are Fantasy Sports?

Fantasy sports are online games where participants create teams of real-life sports players. These players earn points based on the actual performance of their real-world counterparts. Fantasy sports players compete against each other in private and public leagues, and the player with the highest overall score at the end of an annual season wins. The fantasy sports industry is experiencing rapid growth, with a projected Compound Annual Growth Rate (CAGR) of 12-13% and an expected value of nearly $13 billion by 2028. The industry is currently the fastest-growing betting market globally. While this project focuses on the UK Fantasy Premier League, there are three primary sports that make up the fantasy sports world.

Cricket

Cricket is the second most popular sport globally, with an estimated 2.5 billion fans worldwide. It is particularly popular in India, the world's second most populated country, where it is by far the most followed sport. Fantasy cricket has gained immense popularity in recent years, with an estimated 150 million people currently participating in the game.

American Football

In 2022, an estimated 30-40 million individuals participated in Fantasy American football, with over 80% of all players residing in the United States. Unlike many other countries, fantasy sports leagues in the US are closely associated with betting. Participants are typically required to submit an entry fee to join a league, and winners collect earnings from these submission fees. The passage of the Unlawful Internet Gambling Enforcement Act in 2006 prohibited various forms of online gambling, including poker, in the US. However, fantasy sports and sports wagers in general were exempted from the ban, which contributed to a significant shift toward sports gambling in the country through fantasy sports.

Football (soccer)

With a current player pool of 11.4 million, Fantasy Premier League (FPL) stands as the world's largest Fantasy Football game and is the focus of this report. A higher percentage of UK residents participate in FPL compared to the number of Indians playing fantasy cricket or Americans playing Fantasy American Football. FPL experiences consistent growth, with an average annual increase of over a million participants and has become a permanent hobby for football fans, who often form leagues with their friends and family members. FPL in the UK does not usually involve any monetary wagers.

How Does Human Bias Impact FPL?

Like any decisions made by humans, the process of selecting players for your Fantasy Premier League (FPL) team is susceptible to various inherent biases.

One of the most prevalent biases is **team loyalty**, wherein players often exhibit a preference for players from their favourite Premier League club. This loyalty, whether consciously or subconsciously, leads to a certain level of favouritism towards players from their team of choice, as well as a reluctance to include players from rival teams. As an Arsenal fan, for instance, it is often the case for my team to consist of multiple Arsenal players, even if they may not be the most optimal choices, while Tottenham players (Arsenal's biggest rival) are rarely considered. This trend seems to emerge every year, albeit unintentionally.

Recency bias also plays a role in influencing team selections. If one has recently witnessed a team or player perform exceptionally well or poorly, there is a tendency to overemphasize these recent performances when making decisions for the upcoming weeks selections.

Another powerful form of bias, which can be both helpful and misleading, is the players **selected percentage** metric. This metric, displayed alongside each player during the FPL team selection process, indicates the percentage of FPL players who have included that player in their teams. While high selected percentages often indicate players who have been performing well or offer good value, it is not always a reliable indicator. Nonetheless, many FPL players tend to select certain players simply because others have chosen them, without considering other relevant factors.

The objective of this project is to mitigate these biases in all their forms. By developing algorithms and selection processes solely based on historical data, my goal is to eliminate human bias and stick to purely data-based decision-making in FPL team selection.

FIGURE 1

Project Steps

Collecting data

Core objective

Our decision-making process relies solely on data-driven models, abandoning all personal football knowledge and other players' selections. Consequently, our choices are determined exclusively by the collection of indicative data points that effectively predict team and player performance.

Key steps

- Historical team data
- Historical player data
- Elo data
- XG data

Predicting goals

Core objective

The collected data from Step 1 will be used to develop a linear regression model with the purpose of predicting goals scored and conceded by teams participating in future matches.

Key steps

- Assessing the optimal form period
- Comparing XG and actual metric accuracy

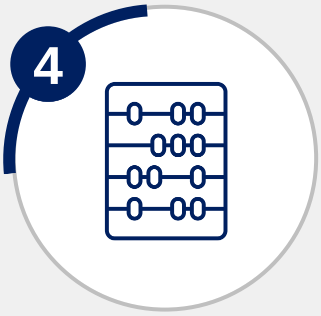
Predicting points

Core objective

To convert predicted team goals into individual player points, we will combine the projected team results with the individual player's current form. By incorporating both aspects, we can estimate the potential contribution of each player towards the team's overall performance.

Key steps

- Create induvial player form
- Distinguishing positional points
- Integrate specific FPL point metrics



Selecting squads

Core objective

Using predicted player points, we can construct weekly optimal squads that adhere to the criteria set by the Fantasy Premier League (FPL) game.

Key steps

- Combine top performers with value picks
- Formations and Substitutes

1

Collecting Data

To create, train, and test models, it is necessary to collect relevant historical data that encompasses both team performances and individual player performances. This data should include both actual data and expected data, enabling a comparison between their predictive capabilities.

Elo ratings

Elo equation

One method to measure relative strength levels is the Elo rating system, initially created by physics professor Arpad Elo for improving chess player ratings. This widely adopted system can be applied to football teams, assigning each team a single rating. The difference in ratings between two teams serves as a predictor of the match outcome. The general Elo equation used is as follows:

$$E = 1 / (10^{(-dr/400)} + 1)$$

Where dr is the Elo point difference of the two teams.

Points exchange

After a match between two teams, the winning team will take points from the losing team based on their relative strengths prior to the match. If a team with a considerably higher Elo rating defeats a team with a significantly lower Elo rating, the higher-rated team will only gain a few points from their lower-rated opponent. However, if a team with a much lower rating beats a considerably higher-rated team, the exchange of rating points will be more substantial. The number of points exchanged must be determined so that a certain win rate between two clubs makes the Elo difference between both clubs converge towards the Elo difference that corresponds to this win rate. The following equation satisfies this constraint:

$$\Delta Elo\_1X2 = (R - E) * k$$

Where R is the results (1 for win, 0.5 for draw, 0 for loss) and k is the weighted index, a higher k will have the ratings converge quicker to their true values but will suffer from more variation. The Elo's we will use a weight index of k = 20.

Goal difference and home field advantage

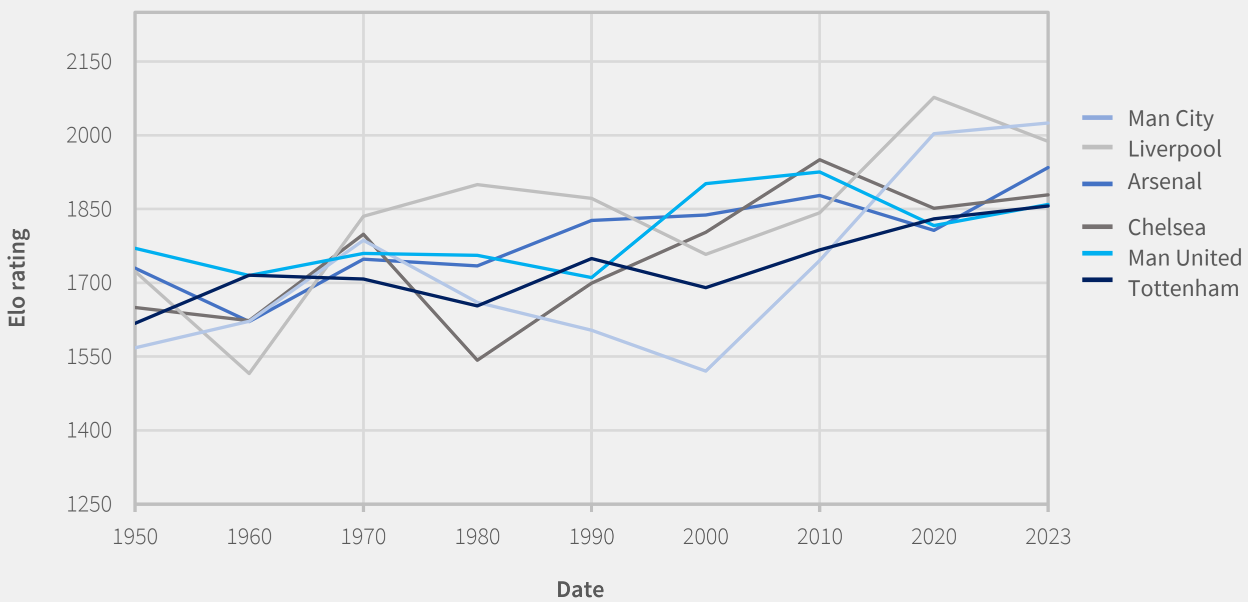
The Elo rating system we will use has been altered slightly to include certain metrics that are specific to football. These two metrics are:

1. Adjusting points exchange passed on the severity of the win / loss (goal difference), in other words if a team wins by a bigger margin, they will gain more points compared to a narrow victory and vice versa.

2. To account for home team advantage. Teams perform better when they are playing at their home stadium compared to when playing away from home. As a result, points exchange is also adjusted based on if the teams are home or away.

FIGURE 2

Elo ratings of the 'big six' teams in the English Premier League between 1950 and 2023



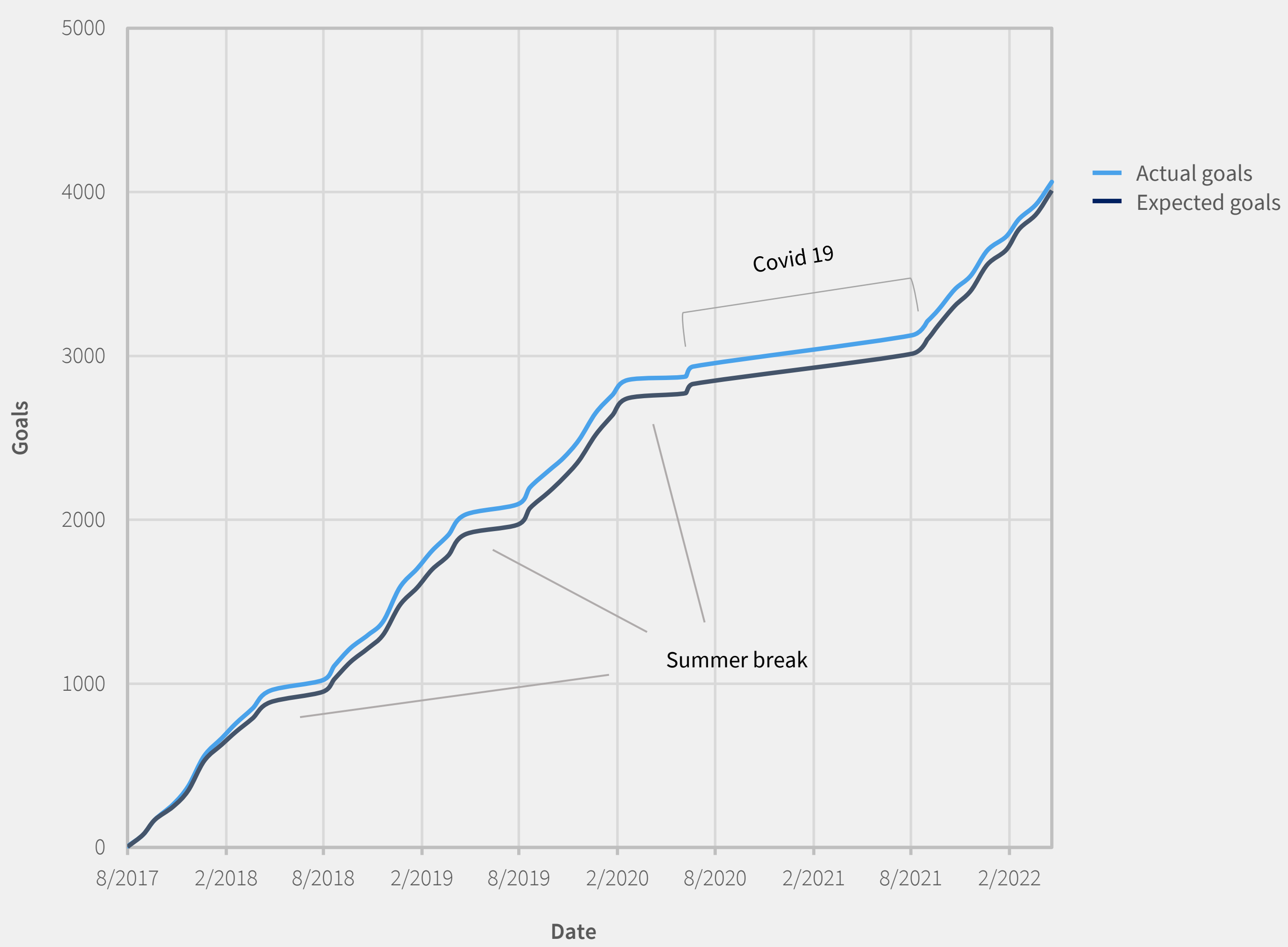
Source
www.clubelo.com

Expected Goals (XG)

Expected goals (XG) is a relatively recent metric in football, introduced in the Premier League in 2017. Although it has only been used for a few seasons, XG has significantly transformed the way we analyse performance statistics in the game. XG is a numerical measure, typically ranging from 0 to 1, indicating the likelihood that a given shot will result in a goal. This measure is determined by considering various factors and historical data, enabling us to assess the number of goals a player or team should have scored based on the quality of their scoring opportunities. Expected goals provide us with a means to account for factors such as clinical ability, current form, and luck, allowing us to determine the "true" number of goals a team or player should have scored based on the outcomes of similar chances in previous games. Alongside expected goals, metrics like expected assists and expected saves have also emerged as vital data points for analysing player and team performance. To evaluate the effectiveness of expected goals in predicting future outcomes, we will compare them to the actual data points (goals vs expected goals ect) equivalents to determine whether expected metrics outperform, underperform, or perform similarly to actual metrics in terms of their predictive accuracy. Figure 3 illustrates the relationship between expected goals and actual goals scored in the Premier League from 2017 to 2022, highlighting the consistency of XG as an indicator of actual performance.

FIGURE 3

Running English Premier League goals and expected goals from August 2017 to May 2022



Note: The global covid 19 pandemic between 2020 and 2021 resulted in a complete halt to spectator sports in the UK including the Premier league for several months, and so no goals were recorded during this period.

Source
www.fbref.com

FPL data

While historical data is crucial for creating models and predicting player points, we still rely on Fantasy Premier League (FPL) data to assist us in squad selection and as a benchmark for comparison. FPL player values, a metric exclusive to the game, play a significant role in our decision-making process. These values are assigned to players based on their form, popularity, and other factors. Considering the budget constraint of 100 million, a player's cost is the most important factor when assessing their value in a team.

Two additional FPL metrics we collect are player positions and weekly points. To calculate a player's position and points, we use data from the football data website FBREF. In most cases, this aligns with the actual FPL position and point metrics. However, there are a few exceptions. FPL players are initially assigned a single position (GK, DEF, MID, or FWD) for the entire season, even if they sometimes play in different positions. For example, a midfielder may occasionally play as a centre-back in a formation with 3 centre-backs. Consequently, their position in FBREF data is determined game by game and so can fluctuate. The accuracy of player positions is essential as player points are determined by their position. For instance, a midfielder will only receive 1 point if their team keeps a clean sheet, whereas a defender earns 4 points (see chapter 2 for detailed points scoring system). Aligning the actual FPL positions with the FBREF positions allows us to ensure consistency. Another metric that may sometimes differ between our data and the official FPL data is the awarding of bonus points. Bonus points are assigned based on over 30 different in game performance data points, such as creating a big chance or making an error leading to a goal. The top three players in each game, based on total bonus points, receive 1, 2, or 3 extra points added to their weekly score. While FBREF provides most of these data points, they use a different data supplier than FPL (FPL uses OPTA data). As a result, the criteria for defining a "big chance" or "error" may slightly differ, leading to minor discrepancies in final bonus points.

TABLE 1

Data collected that effect player points

Data Set	Source	Metrics
Team Data Overall team data	— http://clubelo.com — https://fbref.com	— Historical fixtures and results — Future fixtures — Elo data
Player Data Individual player and match data	— https://fbref.com	— Minutes played — Position played — Goals / Expected goals — Assists / Expected Assists — Yellow and red cards — Goals conceded / Expected goals conceded — Post shot expected goals (GK) — Own goals
Fantasy Data Player FPL data	— https://github.com/vaastav	— Player points / values / positions — Double & blank game weeks

2

Predicting goals

The concept of fantasy football is straightforward. The objective is to assemble a squad of 15 players, including 11 starters and 4 substitutes, within a budget of £100 million. Each player is assigned a value based on their historical performance in terms of scoring points. They are designated one of four positions: Goalkeeper (GK), Defender (DEF), Midfielder (MID), or Forward (FWD). Points are awarded to players based on their real-life performances, with most points being attributed to goals scored or conceded.

Goalkeepers and defenders primarily earn points through clean sheets, which occur when their team avoids conceding any goals during a match. On the other hand, midfielders, and forwards gain points mostly through goals scored or assists provided. Consequently, accurately predicting the number of goals a team will score or concede in each game becomes the most important metric for anticipating the points that both the team and its players will score on a weekly basis.

.

TABLE 2

How players earn points in Fantasy Premier League

Action	Goalkeeper	Defender	Midfielder	Forward
Playing up to 60 minutes	+1	+1	+1	+1
Playing 60 minutes or more	+2	+2	+2	+2
For each goal scored	+6	+6	+5	+4
For each assist for a goal	+3	+3	+3	+3
For a clean sheet	+4	+4	+1	-
For every 3 shots saved	+1	-	-	-
For each penalty save	+5	-	-	-
For each penalty miss	-	-2	-2	-2
Bonus points for the best players in a match*	+1 - 3	+1 - 3	+1 - 3	+1 - 3
For every 2 goals conceded	-1	-1	-	-
For each yellow card	-1	-1	-1	-1
For each red card	-3	-3	-3	-3
For each own goal	-2	-2	-2	-2

*Bonus points are calculated by the sum of several performance statistics in a game such as passing completion percentage and key tackles known as bps. The 3 players in each match that score the most bps receive 1,2, 3 bonus points respectively.

Source
www.premierleague.com

Model

To predict the number of goals a team will score and concede in each match; we will employ a linear regression model. A linear regression model describes the relationship between a dependent variable, y and one or more independent variables x. In our case the goals scored by a team is the y variable and our features are the x variables. Since football is a zero-sum game, the number of goals one team scores will be equal to the number of goals the opposing team concedes. Therefore, once we can predict the goals scored by each team, we can also predict the goals conceded.

Features

After conducting testing and experimentation with different combinations of features, the combination of the following features yielded the best results:

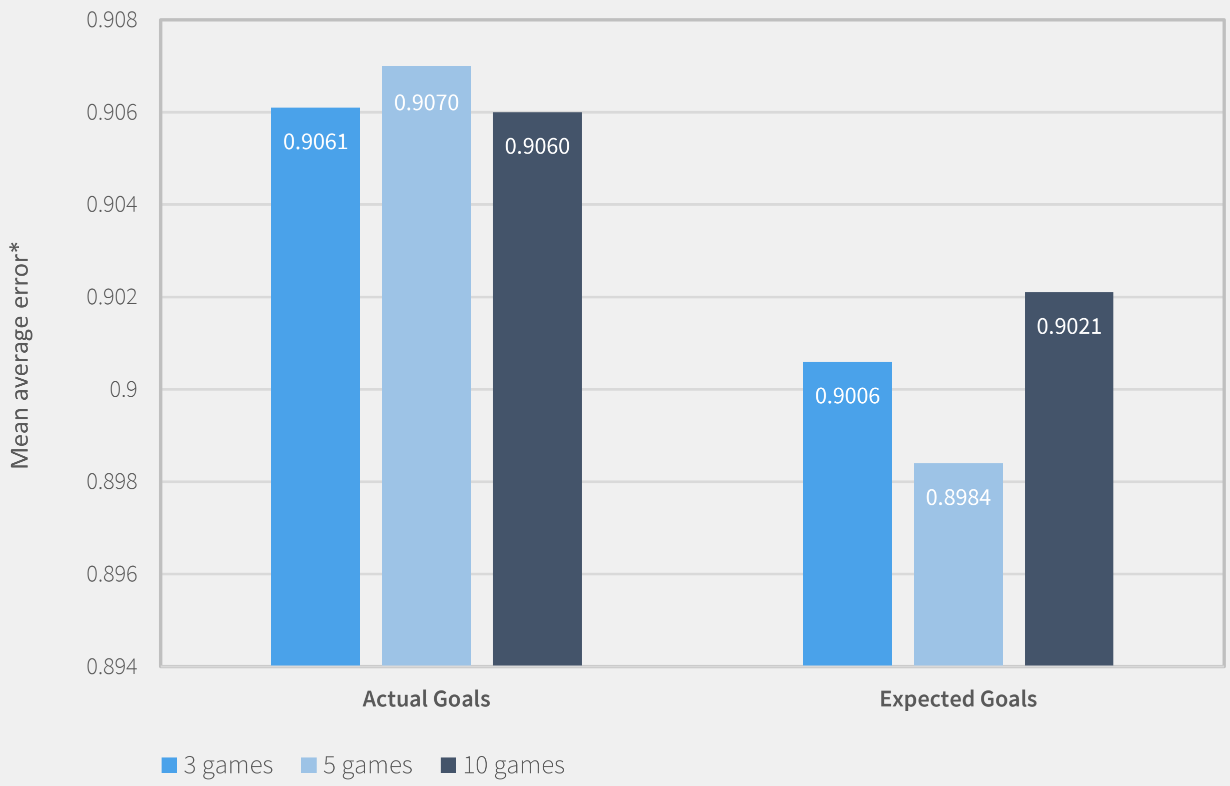
Elo: The difference in Elo ratings between two teams is the most significant feature when predicting goals scored based on historical data. Both the team's Elo rating and the opponent's Elo rating are used as separate features in the model.

Home or away: The venue of the match, whether the team is playing at home or away, has a substantial impact on their performance. This feature, when included in the training data, significantly improves the accuracy of the model even with the additional adjustment in Elo ratings based on home field advantage.

Goals scored and conceded form: This feature considers the number of goals scored by the team and the number of goals conceded by their opponent in recent past games. Two factors were considered: whether to use actual goals or expected goals, and the number of games to include in the metric. After comparing 3, 5 and 10 game periods with actual and expected metrics, using a 5-game expected goal form yielded the best results. It strikes a balance between capturing recent form and avoiding overemphasis on short-term fluctuations.

It is important to note that the goal of this model is not to predict exact scores with precision, but rather to identify which teams and players are more likely to perform better than others in each game week. The model's predictions are intended to provide insights into relative performance levels rather than precise outcomes.

FIGURE 4 **Accuracy of different goal scoring form metrics for seasons 2017-18 and 2018-19 (lower is better)**



* Mean average error (MAE) is the average difference between the number of goals scored the model predicted and the actual goals scored. If the model predicted a team to score 2 goals and the team scored 3 goals then the MAE would be 1.

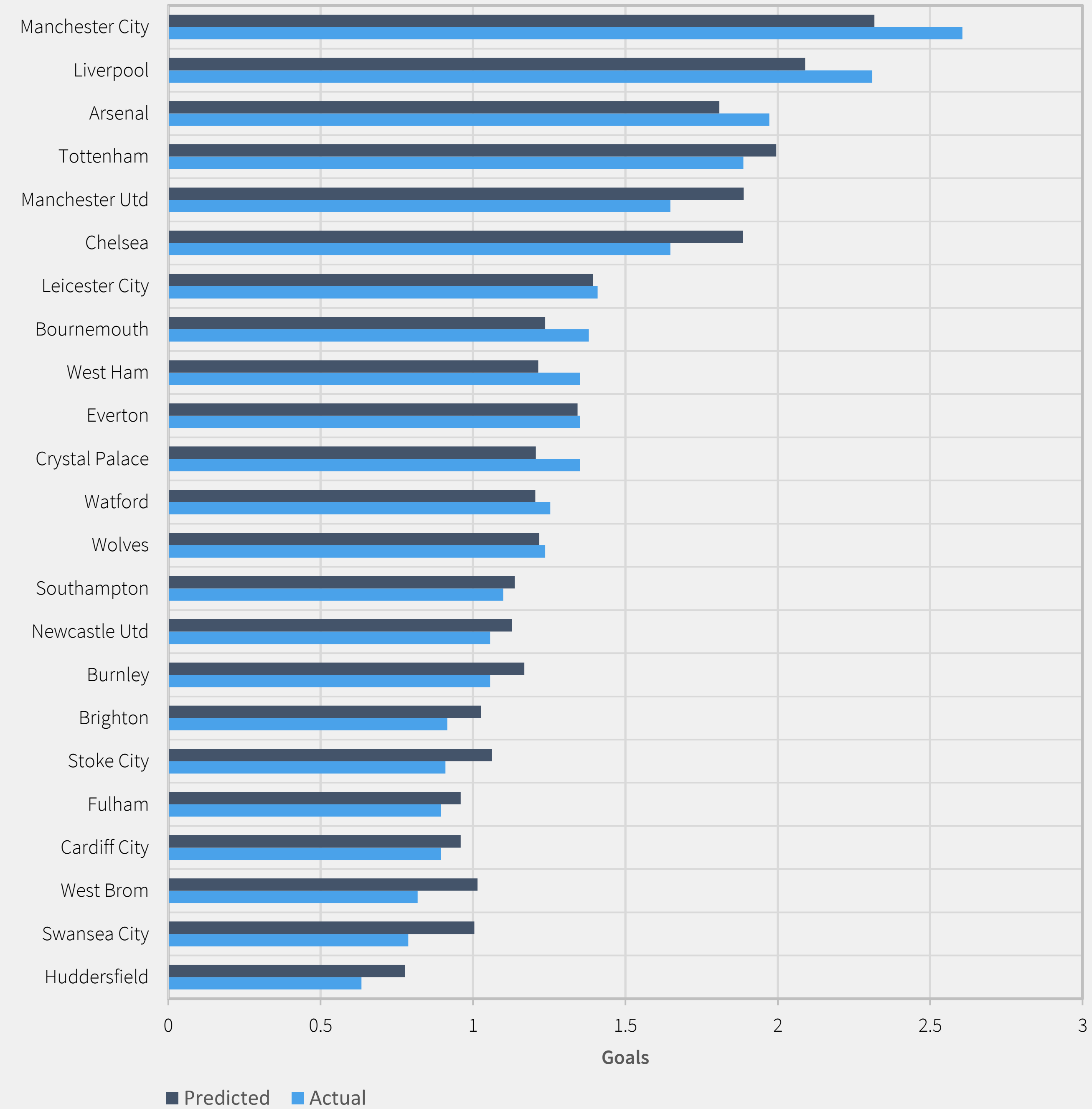
A Bayesian Ridge linear regression model is used as it was the most effective among the tested linear regression models. Its ability to incorporate probability distributions instead of point estimators makes it well-suited for our purposes, especially considering the relatively small data sample we have available.

To train the model, we use Premier League game data from the 2017-18 season to the 2018-19 season. Using the Python package Scikit-Learn, we split this data into training and testing sets, with a 75%/25% split. The model will be trained on the training set and then applied to the test set to evaluate its accuracy. The trained model will then be applied to the data relevant to the matches we want to predict goals for, excluding the first five games (due to the lack of goal-scoring form data).

This methodology relies solely on Elo ratings, expected goals form, and home/away advantage, without any potential human bias or prior knowledge influencing the predictions. While this approach may not always outperform a human-based approach, its value lies in its ability to generate unbiased results, aligning with the objective of this project. This model will also be applied to future seasons' data to predict goals based on the same methodology.

FIGURE 5

Average goals scored per game predicted by final model compared to actual goals per game, 2017-18 and 2018-19 seasons



3

Predicting points

To predict player points, we combine the predicted goals and conceded metrics generated by our Bayesian model with player form, specifically using player 5-game expected form (as its shown to be the most accurate indicator). Most points are determined by:

Clean Sheets: Based on our predicted goals conceded data, we calculate a clean sheet probability metric for each team.

Assists/Goals: Like clean sheet points, assist and goal points are determined by combining the team's expected goals scored.

To predict the total points of an outfield player who is expected to play over 60 minutes, we use the following equation:

Total Points = ((2 + (CSP*PCS) + (3*(AF*PGS)) + (PG*(GF*PGS)) + (BSP)) – ((PM*2) + (YC*1) + (RC*3) + (OG*2) + ((DG*PGC)//2))

Where:

CSP is the expected probability of the player's team keeping a clean sheet, ranging from 0 to 1.

PCS is the number of points awarded to the player's position for achieving a clean sheet.

AF is the player's expected assists based on their performance over the past 5 games.

PGS is the number of goals the player's team is predicted to score.

PG is the number of points awarded to the player's position for scoring a goal.

GF is the player's expected goals based on their performance over the past 5 games.

BSP is the expected number of bonus points awarded to the player.

PM is the number of penalties missed by the player.

YC/RC is the expected number of yellow/red cards received by the player.

OG is the expected number of own goals scored by the player.

DG is 1 if the player's position is a goalkeeper or defender, and 0 otherwise.

PGC is the number of expected goals conceded by the player's team.

FIGURE 6

Correlation between total points and predicted points using either actual or expected data points in seasons 2017-18 and 2018-19 (higher is better)

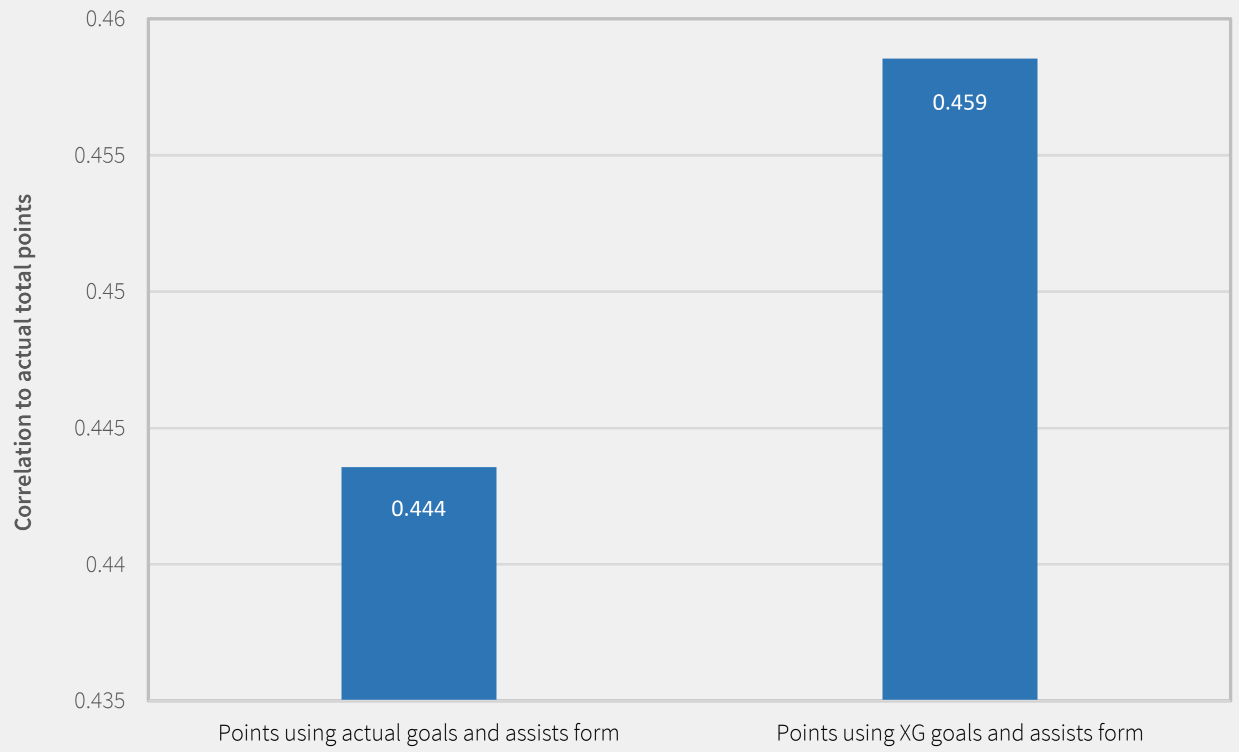


FIGURE 7

Top 50 scoring players, 2018-19 season

Note: The points scored in our shown in Figure 7 may slightly differ from the official FPL.com points due to variations in the calculation of certain performance indicators. We are obtaining our player data from fbref.com, which may have small differences in the calculation of metrics such as bonus points and assists compared to the official FPL calculation.

Source: www.fbref.com

TABLE 3

Comparing predicted points and actual points for top 50 best predicted performers in the English Premier League between 2017-18 and 2018-19 seasons

Player	Position	Predicted Points	Predicted Rank	Actual Points	Actual Rank
Mohamed Salah	MID	247	1	249	2
Raheem Sterling	MID	242	2	239	3
Paul Pogba	MID	203	3	183	16
Sadio Mane	MID	199	4	220	5
Eden Hazard	MID	199	5	253	1
Pierre-Emerick Aubameyang	FWD	176	6	210	7
Andrew Robertson	DEF	176	7	203	8
Aymeric Laporte	DEF	175	8	222	4
Sergio Agüero	FWD	172	9	217	6
Virgil van Dijk	DEF	171	10	193	10
Bernardo Silva	MID	170	11	178	19
Christian Eriksen	MID	170	12	187	12
Ederson	GK	168	13	196	9
César Azpilicueta	DEF	168	14	179	18
Kyle Walker	DEF	167	15	187	12
David Silva	MID	165	16	154	33
Luka Milivojevic	MID	163	17	181	17
Harry Kane	FWD	160	18	153	35
Gylfi Sigurdsson	MID	160	19	187	12
Marcos Alonso	DEF	159	20	159	28
David Luiz	DEF	159	21	165	24
Matt Doherty	DEF	157	22	157	30
Matt Ritchie	DEF	157	23	136	51
Leroy Sane	MID	153	24	168	22
Toby Alderweireld	DEF	151	25	123	69
Trent Alexander-Arnold	DEF	150	26	184	15
Antonio Rüdiger	DEF	150	27	102	100
Kieran Trippier	DEF	148	28	138	48
Lewis Dunk	DEF	147	29	103	99
Alexandre Lacazette	FWD	147	30	157	30
Richarlison	MID	146	31	155	32
Son Heung-min	MID	146	32	163	26
Patrick van Aanholt	DEF	146	33	147	39
Luke Shaw	DEF	145	34	95	119
Raúl Jiménez	FWD	145	35	163	26
Shane Duffy	DEF	145	36	125	65
Roberto Firmino	FWD	144	37	148	38
Ryan Fraser	MID	144	38	184	15
Jamie Vardy	FWD	143	39	163	26
Nathan Ake	DEF	142	40	121	74
Romelu Lukaku	FWD	142	41	119	76
David de Gea	GK	142	42	126	63
Ashley Young	DEF	141	43	75	177
Michael Keane	DEF	140	44	118	79
Lucas Digne	DEF	140	45	175	20
James Maddison	MID	138	46	152	36
Nathan Redmond	MID	137	47	138	48
Marcus Rashford	FWD	136	48	140	45
Ayoze Pérez	MID	136	49	150	37
Wilfried Zaha	FWD	135	50	156	31

Goalkeepers Defenders Midfielders Forwards

4 Selecting Squads

To optimize the scoring potential of a Fantasy Premier League (FPL) team, it is important to evaluate the value of each player. This involves comparing a player's cost (as determined by FPL) with their predicted points metric. By comparing these two data points, we can assess the true value of a player. For example, a player like Mohamed Salah, who consistently performs well and ranks among the top point scorers each season, is understandably expensive and highly selected in FPL. However, by comparing his cost to his expected points, we can determine if he offers good value for his price. This approach allows us to identify players who may be undervalued and offer greater returns than what is commonly perceived.

By using pure data and avoiding selected percentage data and human bias, we aim to uncover players who may have gone unnoticed by others, providing us with an advantage in team selection. In addition to evaluating player value, there are certain requirements and criteria that must be considered when selecting an FPL squad:

Budget: The total cost of the 15 players in your squad must not exceed £100 million.

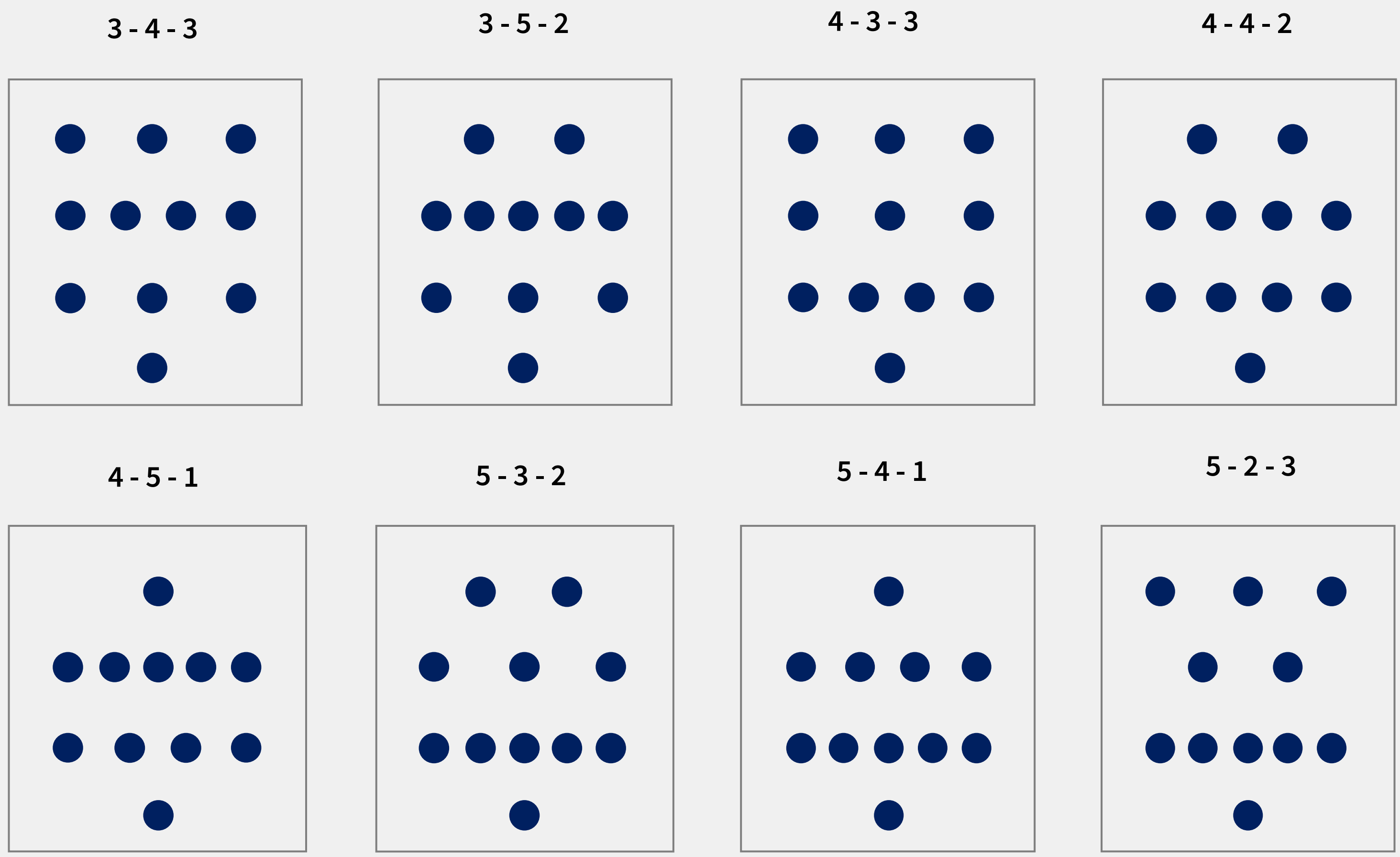
Players: Your squad must include 2 goalkeepers, 5 defenders, 5 midfielders, and 3 forwards.

Teams: You can have a maximum of 3 players from a single team in your squad.

Formation: Your starting 11 players can be arranged in any combination if it includes 1 goalkeeper, 3-5 defenders, 2-5 midfielders, and 1-3 forwards.

Captain: You must assign a team captain each game week. This player will receive double points every match they play. You also select a vice captain which will receive double points if your captain doesn't play.

FIGURE 8 **All possible formations for starting 11 players in FPL.**



Bonus tokens and Substitutions

In FPL, the substitution system and bonus tokens play a vital role in optimizing your overall performance. Here's how they work:

Blank / Double Game Weeks

Due to various reasons such as fixture clashes with other tournaments such as Champions League and postponements due to traffic or weather, games are often rescheduled to take place during the latter stages of a premier league season. As a result, during the second half of an FPL season, there will be several game weeks where teams will play more than one match or no matches for them to catch up to the rest of the league.

Substitutions

- You receive 1 free transfer per game week, with a maximum of 2 free transfers available. If you don't use the free transfers in a game week, after you have 2 available, they won't accumulate.
- Once you've used all your free transfers, any additional transfers will cost you 4 points each.
- Substitutions are typically used to replace injured players, benched players, out-of-form players, or players with difficult upcoming matches.
- We will use a substitute predicting program that compares the predicted points of our current squad to available players and free transfers. This will help us determine the potential substitutions that can be made in any given game week.

Bonus tokens

Each player is given 4 different bonus tokens they can apply to their squad over the course of a season. The timing and execution of these tokens often splits the best FPL players from the average joe. This is how each bonus token works and how we will integrate them into our team selection process.

Wildcard

The wildcard token is a crucial bonus token in FPL, allowing players to make unlimited free transfers in a specific game week. Players are given two wildcards to use over the season, one before November 12th and one after November 12th (for the 2022 season, the specific date may change season to season). Here's how we plan to utilise the wildcard tokens:

First Wildcard: Most professional FPL players tend to play their first wildcard between weeks 6 and 12. Our algorithm will assess the predicted points for our game week 6-12 squads and determine the game week within this range that has the lowest predicted points. We will play our first wildcard on that specific game week.

Second Wildcard: The second wildcard will be used any time after week 20. We will consider various factors such as blank game weeks and double game weeks when determining the optimal timing for playing the second wildcard.

Free Hit

The Free Hit token allows players to make unlimited transfers for a single game week, but their squad returns to the original squad before the token was taken after said game week.

We will save the Free Hit token for game weeks where multiple teams have double fixtures or blank fixtures. These game weeks provide an opportunity to maximize the potential extra points that come with the Free Hit. Will be played after week 20, when most double and blank game weeks are played. Based on the points predictions for our quad, we will identify the game week where playing the Free Hit token will be most effective.

Tipple Captain

The Triple Captain bonus token in FPL triples the points earned by your chosen captain for a single game week, instead of the usual double points. We will activate the Triple Captain token when our best performing player has a double game week.

Bench Boost

The Bench Boost bonus token in FPL adds the total points earned by your bench players to your starting team for a single game week. We will use the bench boost token either to negate injuries or suspensions. To capitalise on double game weeks.

Results

Simulating past FPL seasons accurately poses several challenges, primarily due to the unavailability of injury data. In FPL, when a player is injured and unlikely to participate in future games, they are flagged with a red or yellow header, indicating the severity of the injury and a potential return date. This information is crucial for making substitutions and utilising bonus cards effectively, as it helps in replacing injured or suspended players in the team. However, the absence of detailed game to game injury data makes it difficult to accurately simulate past seasons. Relying solely on the "minute played" data point doesn't provide insights into the reason behind a player's absence from a match. It could be a tactical decision made by the manager, or it could be due to an injury or suspension. This lack of contextual information adds to the complexity of simulating seasons.

Despite these challenges, I have simulated three past Premier League seasons using the methodology outlined in this project (also replacing team and players names with random id's to further prevent bias). The results have surpassed my actual point totals, indicating the effectiveness of the approach. I plan to apply the same methodology to the upcoming FPL season (2023/24) and will share the results at the conclusion of the season.

TABLE 4

My actual vs simulated final FPL points

Season	Number of FPL players	Actual final points	Actual final rank	Siumulated final points	Simulated final rank*
2018 – 2019	6,300,000	1,891	3,186,846	2,032	-
2019 – 2020	7,600,000	1,919	2,846,081	2, 243	-
2020 - 2021	8,200,200	1,700	5,287,548	2,098	-

*Other than your own seasonal final score, final point tallies are not publicly available and so it is not possible to determine a final rank based on the simulated scores.

Sources: Fantasy.premierleague.com, mobsports.com