Erno Ruismäki

Data-Driven Decision-Making in Fantasy Premier League: Correlations, predictions and optimization strategies

School of Technology and Innovations
Master's thesis in Industrial Management
Master of Science in Economics and Business Administration

Vaasa 2026

UNIVERSITY OF VAASA
Tekniikan ja innovaatiojohtamisen yksikkö

Author:	Erno Ruismäki
Title of the thesis:	Data-Driven Decision-Making in Fantasy Premier League: Correlations, predictions and optimization strategies
Degree:	Master of Science in Economics and Business Administration
Degree Programme:	Industrial Management
Supervisor:	Emmanuel Ndzibah
Year:	2026 **Pages:** 52

ABSTRACT:

The increased amount of available information has, over the past decades, spread through every industry, and the world of sports has been no exception. Today, various statistics are available almost in real time and for nearly every event. This is naturally also the case with the world's most popular football league, the English Premier League. This available information is utilized not only within teams themselves, but also widely for other purposes such as betting and fantasy games. Fantasy Premier League (FPL), which focuses on the English Premier League, and the statistics related to it, are at the center of this thesis. FPL provides an ideal environment for applying the principles of data-driven management within the constraints of limited resources. The research question of this thesis is: How can data-driven decision-making enhance performance in Fantasy Premier League (FPL) through correlation analysis, predictive modeling, and optimization strategies?

The theoretical section of the thesis focuses on explaining the principles of data-driven management, predictive analytics, and optimization, particularly from the perspective of fantasy sports games. The methodology section describes the data analysis techniques used, presents the underlying mathematical formulas, and explains how the data was collected. The data used in this thesis is numerical, and the research method is quantitative.

The empirical section of the thesis consists of a test created using the Python programming language. The first objective was to identify the factors that correlate most strongly with FPL points and recorded match events. After this, the mean absolute error and the root mean squared error were determined to evaluate the effectiveness of the predictive regression models. Finally, a simulation utilizing optimization was carried out, in which the created optimization model selected an FPL team and competed against a team representing a human minded, heuristic model.

The results of the thesis indicate that data-driven management can also be applied in the world of fantasy sports. The created model did not outperform the human-like team in every individual round, but overall, it accumulated more points. In addition, it utilized the available budget considerably more efficiently. It should be remembered that FPL is a complex game, and this thesis alone cannot account for all variables, meaning that some simplifications had to be made. The observation period was also too short to draw broader conclusions. Nevertheless, it can be concluded that data-driven management and various forms of data analytics have become a permanent part of the lives of FPL managers going forward.

KEYWORDS: data-driven decision-making, predictive analytics, optimization, fantasy sports analytics, fantasy premier league

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen yksikkö
Tekijä: Erno Ruismäki
Tutkielman nimi: Data-Driven Decision-Making in Fantasy Premier League: Correlations, predictions and optimization strategies
Tutkinto: Kauppatieteiden maisteri
Koulutusohjelma: Tuotantotalouden maisteriohjelma
Ohjaaja: Emmanuel Ndzibah
Vuosi: 2026 **Sivumäärä:** 52

TIIVISTELMÄ:
Lisääntynyt määrä saatavilla olevaa tietoa on kulkenut viime vuosikymmeninä läpi jokaisen toimialan eikä urheilumaailma ole ollut poikkeus. Nykyään erilaisia tilastoja on saatavilla lähes reaaliaikaisesti ja lähes kaikista tapahtumista. Näin on tietysti myös maailman suosituimman jalkapallosarjan Englannin valioliigan kanssa. Tätä saatavilla olevaa tietoa hyödynnetään paitsi joukkueiden sisällä myös laajalti muihin tarkoituksiin kuten esimerkiksi vedonlyöntiin tai fantasy peleihin. Englannin valioliigaan keskittyä Fantasy Premier League (FPL) ja siihen liittyvät tilastot ovatkin tämän työn keskiössä. FPL tarjoaa otollisen ympäristön hyödyntää tiedolla johtamisen periaatteita vallitsevien rajallisten resurssien ympäristössä. Työn tavoitteena on selvittää, pystytäänkö näitä jo muilta aloilta tuttuja keinoja hyödyntämään myös mahdollisimman hyvän FPL joukkueen rakentamiseen.

Työn kirjallinen osuus keskittyy avaamaan tiedolla johtamisen, ennustavan analytiikan ja optimoinnin periaatteita erityisesti urheilun fantasiapelien näkökulmasta. Metodologia osiossa avataan käytettävät data-analyysitekniikat, esitellään taustalla vaikuttavat matemaattiset kaavat sekä käydään läpi, miten aineisto on kerätty. Tässä työssä käytetty aineisto on numeerista ja työ on tutkimusmenetelmältään kvantitatiivinen.

Työn empiirinen osuus koostuu ohjelmointikieli Pythonilla luodusta testistä, missä tavoitteena oli ensiksi löytää ne asiat, mitkä korreloivat voimakkaimmin FPL pisteiden ja tilastoitujen pelitapahtumien välillä. Tämän jälkeen selvitettiin keskimääräinen absoluuttivirhe ja keskineliövirheen neliöjuuri, jotta voitiin havainnoida ennustavien regressiomallien tehokkuutta. Viimeiseksi toteutettiin optimointia hyödyntäen simulointi, missä luotu optimointimalli valitsi FPL joukkueen ja pelasi ihmisen valitsemaa joukkuetta edustavaa joukkuetta vastaan.

Työn tulokset osoittavat, että tiedolla johtamista on mahdollista hyödyntää myös fantasia urheilun maailmassa. Luotu malli ei voittanut ihmistä jäljittelevää joukkuetta jokaisella yksittäisellä kierroksella, mutta keräsi kokonaisuudessaan enemmän pisteitä. Lisäksi se hyödynsi käytössä olevaa budjettia huomattavasti tehokkaammin. On syytä muistaa, että FPL on monimutkainen peli eikä tämä työ pysty yksinään huomioimaan kaikkia muuttujia vaan joitain yksinkertaistuksia on jouduttu tekemään. Myös tarkastelujakso on liian lyhyt suurempien johtopäätösten tekemiseen. Kaikesta huolimatta voidaan todeta, että jatkossa tiedolla johtaminen ja erilainen data-analytiikka on tullut pysyväksi osaksi myös FPL managerien elämää.

KEYWORDS: tiedolla johtaminen, ennustava analytiikka, optimointi, fantasypelien analytiikka, Fantasy Premier League

DISCLAIMER AND DECLARATION OF INDEPENDENT AUTHORSHIP

I hereby declare that this Master of Science (MSc) thesis has been written independently and represents my own original work carried out in accordance with the academic rules, ethical standards, and research integrity principles of University of Vaasa.

I confirm that, to the best of my knowledge and intention, I have not engaged in any form of academic misconduct, including but not limited to plagiarism, falsification, ghostwriting, unauthorized collaboration, or the misuse of Artificial Intelligence (AI), Large Language Models (LLMs), or other automated content-generation tools in a manner that violates university regulations or academic integrity standards.

Furthermore, I declare that this thesis has not been subcontracted, outsourced, purchased, or produced in whole or in part by any paid service provider, third party, commercial writing agency, or external individual. All analysis, interpretation, writing, and presentation contained in this thesis are the result of my own independent academic effort unless explicitly and properly referenced.

Where AI-assisted tools or digital technologies may have been used for limited support purposes permitted under university guidelines (such as language refinement, grammar checking, formatting assistance, or idea organization), such usage has been conducted responsibly, transparently, and without compromising the originality, intellectual ownership, or academic integrity of the work.

I fully acknowledge and accept responsibility for the contents of this thesis. I understand that if evidence of plagiarism, unauthorized AI misuse, academic dishonesty, or third-party authorship is discovered at any stage before or after submission, the University reserves the right to take appropriate disciplinary and academic actions in accordance with its regulations and policies. Such actions may include rejection of the thesis, annulment of the degree, disciplinary sanctions, or any other measures deemed necessary by the University.

By signing this declaration, I affirm my commitment to honesty, transparency, and ethical academic conduct.

Student Name: Erno Ruismäki

Student Number: 2302959

Programme: Industrial Management

Title of Thesis: Data-Driven Decision-Making in Fantasy Premier League: Correlations, predictions and optimization strategies

Contents

1 Introduction 8

1.1 Background of the Study 9

1.2 Research Gap, Question and Objectives 10

1.3 Definitions and Scope of the Study 12

1.4 Structure of the Study 13

2 Literature Review 15

2.1 Data-Driven Decision-Making (DDDM) in Sports and Management 15

2.2 Predictive Analytics in Fantasy Sports 17

2.2.1 Linear and Regularized Regression Models 17

2.2.2 Time Series and Rolling Window Approaches 18

2.3 Optimization Strategies in Fantasy Football 18

2.3.1 Linear Programming 19

2.3.2 Binary Integer Programming 19

2.4 Correlation Analysis of Advanced Performance Metrics 20

2.5 Summary of Theoretical Review 21

3 Methodology 23

3.1 Data Collection 24

3.2 Data Analysis Techniques 25

3.2.1 Correlation Analysis 25

3.2.2 Predictive Modeling 25

3.2.3 Optimization Modeling 27

3.3 Data Validity and Reliability 30

3.3.1 Construct and Internal Validity 31

3.3.2 Reliability and Deterministic Verification 31

3.3.3 External Validity and Scope of Simulation 32

4 Empirical Results 34

4.1 Correlation Analysis Results 34

4.2 Predictive Model Performance 36

4.3 Optimization Results and Team Comparisons 39
4.4 Summary of Empirical Findings 42
5 Discussion 43
5.1 Interpretation of Results 43
5.2 Implications for Fantasy Sports Strategy 44
5.3 Reflection on Data-Driven vs. Heuristic Decision-Making 45
6 Conclusion 47
6.1 Summary of the Study 47
6.2 Contributions to Theory and Practice 48
6.3 Limitations and Future Research 49
References 51

Figures

Figure 1 Heatmap 35

Figure 2 Scatter Plot 38

Figure 3 Cumulative performance 41

Tables

Table 1 Comparative Performance (Gameweeks 25-29) 39

1 Introduction

In recent years, the availability of data has increased throughout every area of life. That together with advances in analytical methods have reshaped decision-making across a wide range of industries. Today organizations lean on data-driven decision-making (DDDM) to improve efficiency, reduce uncertainty and gain competitive advantage (Brynjolfsson & McElheran, 2016). This same development can be seen also in the sports industry. Performance numbers stream in without pause and shape strategy on and off the pitch. This continuous data generating is now used when taking operational decisions. (Müller et al., 2017).

Football, one of the world's biggest and most lucrative games, has now become highly data-driven sport environment. The data figures reach far beyond professional clubs and analysts. They have become available to anyone with an internet connection. Fantasy Premier League, the official fantasy game of the English Premier League, draws millions who must keep making choices under tight budgets, constant uncertainty and sharp competition (O'Brien et al., 2021). These decisions closely resemble resource allocation and managerial decision-making problems studied in industrial management and operations research (Wright, 2009).

Even though the data has become increasingly available to everyone there is still a remaining gap between professional and amateur analytics. Elite sides have integrated data analytics into daily operations to squeeze out every inefficiency (Morgulev et al., 2018). Individuals playing the FPL seem to strongly believe intuition rather than using any systematic ways which creates a clear gap compared to professional organizations.

This is a usual warning sign highlighted in behavioral economics. In environments where there are a lot of factors varying and no supporting tools available for analytical decision making, people are falling back to intuition-based decisions and simple heuristics. Kahneman (2011) argues that when making decisions in environment like this and acting quickly a cognitive bias is happening. This leads to a value loss. Clear path forward from

this is bringing into action data-driven decision-making tools and finding out could individual managers gain advantage by using those.

1.1 Background of the Study

Data-driven decision-making has become a regular way of working in this era of large data volumes and available computer power. DDDM in its core is trusting numbers more than a human mind when making both operational and strategic decisions. There is a wide understanding in the area that companies who follow this simple rule and avoid taking intuitional decisions are making better overall results when it comes to performance and productivity (Brynjolfsson & McElheran, 2016).

Nowadays football and other sport analytics have become one of the areas that this change from trusting your intuition to trusting the data provided is the most visible. For example, in football matches these days new technology allows to track player actions and match events in real time and at a very precise level. This would not have been possible only a few decades ago. (Müller et al., 2017). Especially in football the use of advanced metrics such as Expected Goals (xG) have become a standard way of analyzing performance and it is widely being replacing traditional statistics with less information so things like just shots or shots on target (Rathke, 2017). Naturally there are also commercial sides to all this. Companies like Opta are doing the work to gather and standardize the data and providing it to consumers which so far has meant that there is data available for anyone who is interested.

Fantasy Premier League is excellent place to figure out how all this comes together and affects the decisions. Predictive analytics is essential part of the weekly squad choosing for all managers whether they use analytical tools for it or not. Everyone is trying to predict who will score the most points and choose the players based on that which is exactly how predictive analytics is known in information system research area (Shmueli & Koppius, 2011). This all must be done within the budget limits, which means the squad building in this sense can be seen as a Knapsack Problem.

Fantasy Premier League rules are changing every now and then depending on the evolution of the game. Regarding this study the season 2024-2025 data is being used as a source and rules are being adopted accordingly. The main difference between the 2024-2025 season and 2025-2026 season is the additional points coming from defensive contributions in the new season (*Fantasy Premier League, Official Fantasy Football Game of the Premier League*, n.d.) In this study the points are therefore coming from simpler ways; goals, assist, clean sheet and are only involving defensive actions via the bonus points system which has been in use for a long time already. This rule change raises the value of using data-driven decision-making tools when building a squad because it complicates the game even more. Therefore, the role of DDDM can be seen to increase also in this sector in the future.

As mentioned, the availability of the data has not been an issue during recent years if it was the case at some point, however, there has not been much research that would have attacked into this topic of utilizing DDDM into fantasy football world. O'Brien et al., (2021) stated there is an advantage of using analytics over time. Deep look into how xG for example works compared to older statistics such as shots or how much advantage a manager could gain for using systematic tools and optimization remains unclear. This study is targeting to utilize correlation analysis, predictive modeling and optimization methods directly into the FPL environment and provide therefore more precise and direct information on the new area of advantage performance metrics and DDDM methods in fantasy sports.

1.2 Research Gap, Question and Objectives

There exists a remarkable amount of studies these days that have been gathering evidence on how the data-driven decision-making is improving results in organizations across the industries (Brynjolfsson & McElheran, 2016). At the same time researchers have noticed that with the modern performance metrics and predictive analytics improving results can be gained in sports as well (Müller et al., 2017). Combining these two and tracking the using of DDDM systematically in fantasy sports world lacks research.

So far, the studies which have used DDDM methods have been studies that have either focused on betting or studies that have focused on professional teams' strategies. These both have a wide set of research already. (Wright, 2009). There is a clear lack of studies that are putting into action correlation analytics, forecasting and optimization into the field of fantasy sports. The game of Fantasy Premier League contains several factors with budget, position rules and start-bench player issue so that there is a clear need for studies regarding how advantage could be gained with DDDM. O'Brien et al. (2021) confirmed there has not been a lot of attention to any human vs algorithms comparisons although studies do hint there is a skill included in the game between the managers.

This study targets to fulfill this existing gap. Correlation analysis, predictive modeling and optimization are being used in this study in the Fantasy Premier League world directly. Data-driven decision-making methods are being applied so there so the ideas and theories themselves are not anything new but applying these in the fantasy premier league world is yet an understudied field. Secondly, there exists an evidence gap this study is also partly fulfilling as even though there are many studies regarding data-driven approaches versus human the coverage is still quite narrow when it comes to specifically fantasy sports.

The central research question for this study is: **How can data-driven decision-making enhance performance in Fantasy Premier League (FPL) through correlation analysis, predictive modeling, and optimization strategies?**

To further support the answer to the research question these objectives are being used:

1. Relationship between advantage performance metrics and Fantasy Premier League points are being studied to find out the correlations that can then support decision making.
2. Historical rolling averages are being used to build predictive regression model to forecast player performance against heuristic model

3. Optimization strategy for FPL team selection is being built and tested whether it can gain advantage over traditional human based selection within budget limits.

1.3 Definitions and Scope of the Study

Before moving forward, the key terms of this work are being defined. For the purporses of this research the following operational definitions are adopted.

- **Data-Driven Decision-Making (DDDM):** The systematic practice of basing operational and strategic decisions on verifiable data analysis rather than intuition or heuristic judgment. In this study, DDDM specifically mean weaving together correlation analysis, predictive modeling, and mathematical optimization to steer squad selection inside Fantasy Premier League.
- **Predictive Analytics:** The use of statistical techniques on historical data to forecast what is likely to happen next. This study relies on interpretable regression models, Ridge Regression in particular, to produce a predicted points figure for every player. That figure then feeds straight into the optimization engine.
- **Optimization:** A mathematical approach for picking the best available solution from a range of options while respecting stated constraints. Here the problem is cast as a weighted Integer Linear Programming task. The goal is to push the projected points of the starting XI as high as possible, with a lower weight attached to bench players so that the model reflects the real trade-offs managers face between immediate payoff and keeping depth.
- **Advanced Performance Metrics:** Quantitative Numbers meant to capture the quality of performance beyond the usual statistics. This includes things like Expected Goals (xG) and Expected Assists (xA) and the derived Expected Points (xP). In the analysis these are smoothed through historical rolling averages to avoid short-term noise and improve stability.

This study is staying on strictly limited scope using performance metrics from Premier League 2024-2025 season. Data is player based but it is being used for building a 15-

player squad using optimization methods. The FPL game has some extra parts in the game, so called chips. These chips (Triple Captain, Free Hit, Wildcard) are not being involved in this study. This decision is solely based on the complication the additional chips would create. The aim of this study is to focus on the key performance units and not to cover the special cases that the chips create for the game. Also, the data set for the 2024-2025 season does not include the new defensive contributions and therefore those are not included in the study. However, it is to be noted that unlike the chips the defensive contribution points work the same way as the regular scoring and therefore it can be stated that the model can be adapted to 2025-2026 in that sense that it would just be one new column in the data set without further changes needed.

Methodologically the emphasis lies on keeping the models readable as well as accurate. Breiman (2001) framed the classic tension as Occam's dilemma: complex algorithms such as neutral networks can hit high accuracy numbers, yet they remain black boxes whose inner workings stay hidden, exactly the part needed if we want to pit them against human heuristics. To sidestep that trap this study turns to ridge regression. As Shmueli & Koppius (2011) stress, genuine predictive strength has to be tested out of sample: Ridge regression delivers that test while leaving the coefficients transparent enough for direct comparison with the decisions real managers make.

1.4 Structure of the Study

This study follows a clear structure typical of thesis work. In chapter 1 the background of the study is being presented. A Research gap, a research question and objectives are being introduced. The definitions and overall explanation of what will be covered and what not is also being stated here.

Chapter 2 is looking into a literature review of this research. Main themes of the study are being presented: Data-Driven Decision-Making, Predictive analytics, Optimization and Correlation analysis. Fantasy sports' point of view is also being included in this section.

Chapter 3 covers the methodology going through the work step-by-step, from data collection all the way to data validation and reliability. Rolling average, ridge regression training and mathematical side of Integer Linear Programming are being investigated here.

Chapter 4 then presents the empirical findings. It opens with the correlations between advanced metrics (xG, xA) and actual point returns, moves on to the predictive accuracy of the regression model, and closes with the optimization outcomes tested against the heuristic strategies across the 2024-2025 season. Clear summary of all parts is being then presented at the end of the chapter.

Chapter 5 is results. Full overview of Data-Driven vs. human model is being analyzed and comprehensive Fantasy Sports strategy discussion is involved.

Final chapter of the work, Chapter 6, is going to present a summary of the study. Firstly, it covers the main areas of the study from both theoretical and practical side and then repeats the results. At the end of the study a future opportunities and limitations of this study are being discussed.

2 Literature Review

Theoretical literature regarding data-driven decision-making will be covered in this chapter. The theory is being adapted to Fantasy Premier League. The main target for this chapter is to provide further information on how predictive analytics, optimization and correlation analysis can be used in this environment. After these topics have been introduced the following empirical chapters will test them in real life.

The chapter is organized around the core thematics that run through the central research question. It begins by reviewing the literature on data-driven decision-making (DDDM) to establish a workable theoretical setting for analytics-based decision support. This leads to an examination of predictive analytics and forecasting methods, with particular focus on their established uses in performance evaluation and sports analytics. Final part of this section is diving into mathematical optimization techniques, which is also an important part of fantasy sports analytics. When working with high uncertainty and a clear budget limit they must be factored in already when looking things theoretically. The aim of the last paragraphs of this chapter is to provide the reader with a wide understanding of how these theories can be utilized in the fantasy premier league environment which will be then followed in the next chapter.

2.1 Data-Driven Decision-Making (DDDM) in Sports and Management

Data-Driven Decision-Making (DDDM) refers to the practice of making decision based on trust to the numerical data provided rather than anything else such as intuition. In the management literature Provost & Fawcett (2013) describe it as a data-analytic mindset a thinking habit that treats the data itself as a core strategic asset for improving decision quality. Studies show that companies and different organizations that follows this approach are constantly being delivering better results than the ones still using other decision making methods. (Brynjolfsson & McElheran, 2016). By putting statistical evidence ahead of subjective calls, the whole process improves consistency, transparency and reproducibility.

Same traps are being on the way of human decision making constantly so whenever there is rising uncertainty humans keep running into those same mistakes. Tversky & Kahneman (1974) traced most of them to heuristics, those mental shortcuts that simplify messy probabilities but regularly tilt the scales towards bias. In the world of sports the issue is always present since a human mind naturally remembers the highlights from last game and overweight it in the decision making forgetting to stop and analyze longer horizon of performance. Gilovich et al. (1985) captured this neatly in their work on the “hot hand” in basketball where observers kept imposing patterns on what was random noise and ended up over valuing streaks that carried no predictive weight. Algorithmic routes cut through that noise cleanly. Numbers do not have feelings and therefore they are always staying in a consistent way of doing things.

Sport environments in general have made DDDM more visible as tracking technologies have sharpened. Football analytics now works with event-level detail that lets analysts price a player’s contribution without the usual halo of public perception (Müller et al., 2017). Fantasy sports is a wonderful environment to investigate all this since it has all it takes to analyze the decisions with identical budgets, rules and repeating system of doing same things weekly throughout the season (Morgulev et al., 2018).

The scoring methods can of course change between the seasons. Bonus systems, heavier defensive weighting, which only swells the data stream. That growing complexity makes plain why purely intuitive selection starts to become heavy and weak. More numbers come in with the data every year so the human mind eventually will not be able to keep track of all. This means humans will fall into intuitional choices. To avoid those a more systematic way of working is a necessary way to survive (Davenport & Harris, 2007).

Even as artificial intelligence and machine learning components slide into the toolkit, DDDM remains the larger frame that decides how they are deployed. In this study it

supplies the theoretical warrant for combining predictive analytics, correlation analysis, and optimization inside Fantasy Premier League.

2.2 Predictive Analytics in Fantasy Sports

In DDDM the predictive analytics are in major role. In the process of predicting the historical data is being captured and turned into a prediction of what is most likely to happen in the future. In this context the best possible outcome would be the highest possible score. Machine learning rules from other industries are true here also. (Bunker & Thabtah, 2019). Football being the high variance game it is, no one expects certainty. The realistic aim is simply to shift the probabilities in favor of the better assets over a long enough sequence of decisions.

2.2.1 Linear and Regularized Regression Models

Sport analytics require visible information and therefore linear regression models are favored. The results directly provide numerical information on how much coefficients there is with each factor. (James et al., 2023).

Yet ordinary least squares quickly run into trouble once the predictors start talking to one another so when there is a multicollinearity situation. Shots on target and actual goals for instance are structurally entangled. Tiny shifts in the data set can send the estimates swinging wildly. That multicollinearity inflates variance and damages predictive reliability of the matches (Hastie et al., 2009).

With a penalty term (L_2 regularization) the ridge regression is a matching tool for this issue. Mathematically, while OLS minimizes the sum of squared residuals, ridge regression minimizes:

$$\hat{\beta}_{ridge} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \quad (1)$$

Here $\lambda \geq 0$ is a tuning parameter that controls the strength of the penalty. The extra term shrinks coefficients towards zero cutting variance at the modest cost of a little extra bias. In fantasy-sports work that trade-off is usually welcome since the priority is robust out-of-sample prediction, not unbiased recovery of historical relationships (James et al., 2023).

2.2.2 Time Series and Rolling Window Approaches

Performance in football is always changing. Team strength drifts, player form and fitness come and go, so long run career averages can actually be misleading (Hvattum & Arntzen, 2010). At the same time, looking only at the last match is also significantly raising more random noise. Rolling windows strike the practical compromise. By averaging the performance for fixed period of time recently, short term luck gets faded away while recent shifts in ability still get registered. The technique, often called temporal feature engineering, turns raw sequences into lagged predictors that carry a sense of momentum. It is crucial that the lagged feature is strict. Only information available before the deadline of the decision can be used. Anything else is a look-ahead bias and that would mean the whole validation exercise to collapse (Hyndman & Athanasopoulos, 2018). This discipline is non-negotiable if the forecasts are meant to mirror the real-world information set facing a manager on deadline day.

2.3 Optimization Strategies in Fantasy Football

Optimization strategies are required when making decisions under budget constraints. In Fantasy Premier League the task is always the same: squeeze the highest possible expected points out of a squad while staying inside rigid budget limits, positional

requirements and three player per club ceiling. Problems of exactly this shape, resource allocation, have occupied operations research for generations. (Wright, 2009).

On paper the squad selection exercise sits neatly inside the Knapsack Problem because several resources press simultaneously. Since money, formation slots and club quotas are all involved, it is more accurately described as a Multidimensional Knapsack Problem (MKP) with cardinality constraints. In this case the goal still remains unchanged. The target is to pick the subset of players that returns the largest total projected value (points) without breaching any capacity dimension (Martello & Toth, 1990).

2.3.1 Linear Programming

Linear programming (LP) is the basic engine for optimization. It works whenever the objective and every constraint are in linear relationships. Inside FPL this already lets the model scan combinatorial spaces that would leave any human exhausted. Following the standard general formulation laid out by Winston (2004), the model maximizes a linear objective subject to linear constraints. In matrix notation the canonical form is expressed as:

$$\begin{gathered} \text{Maximize } Z = c^T x \\ \text{Subject to: } Ax \ \leq b \\ \text{and } x \ \geq 0, \end{gathered} \tag{2}$$

Where *x* is the vector of decision variables, *c* holds the objective coefficients (projected points) and *A* and *b* encode the constraint coefficients and limits (e.g. budget.)

2.3.2 Binary Integer Programming

Standard Linear Programming provides a good baseline; however, it can return fractional results. In this case it would mean choosing a 0.5 of a player which of course is not possible and would therefore be meaningless here. Every decision is binary: a player is either selected or not. The formulation therefore tightens into a Binary Integer Programming

(BIP) model. Its core structure maps directly to the classic 0/1 Knapsack Problem, defined as:

$$\begin{aligned} &\text{Maximize } z = \sum_{i=1}^{n} v_i x_i \\ &\text{Subject to: } \sum_{i=1}^{n} w_i x_i \leq W \\ &x_i \in \{0,1\}, i = 1, \dots, n \end{aligned} \tag{3}$$

In this formulation, v_i represents the value (projected points) of item i, w_i represents its weight (cost), and *W* denotes the total capacity (budget) (Martello & Toth, 1990).

Research in decision analytics keeps showing that optimization engines of this kind outperform pure heuristic strategies (Davenport & Harris, 2007) They can satisfy every constraint at once so never more than three from the same real-life club, exact formation rules and so on and never collapses under the cognitive load that forces human managers into mental shortcuts.

2.4 Correlation Analysis of Advanced Performance Metrics

Correlation analysis remains one of the quickest, most revealing checks a researcher can run on sports data. It is used repeatedly to find out which performance metrics actually sit closest to the outcome that matters. By placing a precise figure on the linear link between the predictors and the target, it turns feature selection from guesswork into something closer to evidence.

The usual metric used is the Pearson product-moment correlation coefficient. It measures both the strength and the direction of the linear relationship between two continuous variables. In the standard statistical learning formulation, the coefficient is calculated as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (4)$$

Where x_i and y_i represent the individual sample points and $\bar{x}$ and $\bar{y}$ denote the sample means.

In this study the focus in mainly on the Opta derived advanced metrics so especially expected goals and expected assist and how tightly they track total FPL points. Unlike the regularly used statistics xG is a probability estimation which estimates the likelihood of a goal coming from a shot. It is calculated by using historical data of things like distance from the goal and angle of the shot. A penalty kick for example usually carries an xG value above 0.75 (Rathke, 2017). In our context so looking from an analytical perspective this means the xG is more of measurement of a process rather than an outcome alone. It strips away the luck that can distort any single match (Brechot & Flepp, 2020).

Running these correlations does two jobs at once. First, it supplies the empirical backing for feeding advanced metrics into the predictive model showing they carry a stable signal for point accumulation. Second, it flags the multicollinearity that is bound to appear so goals and xG are structurally entangled by design. That overlap is precisely why regularized regression so in our case Rigde regression becomes necessary if the model is to hold together (James et al., 2023).

2.5 Summary of Theoretical Review

This chapter has drawn together the main theoretical and empirical threads that are relevant to data-driven decision-making in fantasy sports DDDM has a major role in providing better results by trusting the provided information more than by trusting a human mind that seems to always, sooner or later, stumble to same recency bias or other similar traps. Operating under strict rules in FPL the environment fits perfectly to test DDDM concepts.

Predictive analytics in the literature are bringing in regularized ridge regression and rolling window-features. Both are well suited in this environment where there are constant new things happening and therefore more noise and more data affecting.

Predictive modeling generates estimates and the next logical step is to use those estimates for optimization. Knapsack formulation is chosen to be the model here because it fits perfectly to the squad selection task. Therefore, Binary Integer Programming provides a robust result while human managers tend to take shortcuts when aiming for easy wins.

Finally, the correlation works around advanced metrics made clear why underlying process indicators such as expected goals deserve priority over raw outcome counts. Earlier literature mainly treats these separately, so this study fills in the gap of integrating all this into Fantasy Premier League environment. Having all this covered it is possible to continue with the methodology chapter that follows, where the data collection, feature engineering and modeling will be applied to 2024-2025 Fantasy Premier League season.

3 Methodology

This chapter sets out the methodological approach used to test whether data-drive decision-making can lift performance in Fantasy Premier League. However, a comprehensive explanation of the methodology will be presented first moving layer by layer towards the data collection and analysis. Everything begins with research philosophy. In this study the numerical data is being tested in correlations and regression, and the target is to receive objective results. Therefore, the research philosophy of the study is positivism. Study moves from theory to test the data so the research approach used is deductive. Used data is coming from fantasy premier league as secondary data which is then being statistically modeled and optimized. This is a must since the decision-making process requires prediction and optimization. There are no interviews or qualitative work included but purely numerical analysis. All this leads to the fact that this study is a quantitative study. Time horizon of the study is longitudinal because it uses gameweeks over time and has a walk-forward validation step included. This was chosen because it reflects a real-world situation and as explained in the study, avoids a look-ahead bias. The robustness of the study does not come from mixing an actual method but through a multiple techniques being used with correlation, regression and optimization and a comparison against the heuristic baseline. (Saunders, 2007)

The purpose is to focus tightly on the research question and the three objectives by running correlation analysis, predictive modeling and optimization techniques together inside one quantitative process. Historical player performance data provides the empirical core: they are used first to map the links between performance metrics and realized FPL points, then to train and validate forecasting models and finally to evaluate optimization routines that respect the game's hard budget and formation limits.

The chapter is structured as follows. Section 3.1 describes the data sources, the variables retained and the full sequence of preprocessing steps. Section 3.2 presents the analytical techniques applied, including correlation analysis, predictive modeling with Rigde regression and optimization modeling with Integer Programming. Section 3.3 addresses

issues of data validity and reliability, spelling out the safeguards built in at every stage to keep results grounded.

3.1 Data Collection

There are three main sources of data in this empirical section of the study: the official Fantasy Premier League API, Opta performance metrics and club Elo ratings. All were assembled through the open-source FPL historical archive compiled by Vastaav Anand (Anand, 2024), which standardizes the raw data into a single, workable relational format. The study covers player-level performance data for the 2024–2025 English Premier League season.

In the beginning the raw data was converted to a gameweek level. This was to avoid the classic FPL challenge regarding the double and blank gameweeks where the same team can have two or zero games instead of regular one game. This way the total points was always going to be aligned because the possibility of 0 or 2 games was eliminated which improved the validity.

The finalized data had basic background information such as price, team or position. On top of the total points variable and standard counts like goals and assists also a advanced metrics were in use. As Rathke (2017) showed, xG already filters out much of the finishing variance that makes raw goal tallies unreliable: the same probabilistic logic is applied here to xA when judging creative reliability.

Feature engineering followed the rolling-window logic set out by Baboota & Kaur (2019). A fixed four-gameweek window was chosen to strike a workable balance. It is recent enough to reflect current tactical and physical state yet long enough to weaker the random noise that is natural part of any single match. These rolling averages were then lagged by one full gameweek$(t-1)$. In other words, the predictors used for Gameweek *N* are built strictly from the mean performance across Gameweeks $N-4$ through $N-1$. The strict lag is non-negotiable as it enforces that the model is only using the

information that is available prior to kick off. This ensures the fundamental time-series forecast principles are being followed by preventing the data leakage and look ahead bias. (Hyndman & Athanasopoulos, 2018) Without it the entire validation exercise would collapse.

3.2 Data Analysis Techniques

This study applies to a multi-stage process which consists of correlation analysis, predictive modeling, and optimization modeling. Each of the methods mentioned has a unique role in the data-driven decision-making process. Combining all three is essential for this work to be able to cover the methodological gap of the study and utilize these in FPL context.

3.2.1 Correlation Analysis

Correlation analysis is used to quantify the strength and direction of linear relationships between player performance metrics and FPL points. Pearson's correlation coefficient r is employed due to its suitability for continuous variables and its widespread use in sports performance research (O'Donoghue, 2010)

The analysis focuses on identifying which advanced metrics (xG, xA) correlate most strongly with the target variable (total points). This step was mainly performed to validate that these metrics are reasonable to be used and not just to trust the historical outcomes.

3.2.2 Predictive Modeling

Predictive modeling provides an expectation on how many points the player will provide on the gameweek based on the historical data available. Although simple linear regression is frequently used, football datasets routinely exhibit multicollinearity, for example the high correlation between goals and expected goals.

To address this issue the study uses Ridge Regression (L_2 regularization.) Ridge regression modifies the standard least squares objective function by adding a penalty term proportional to the square of the coefficient magnitudes. This regularization technique shrinks the coefficients of correlated predictors, effectively trading a small increase in bias for a significant reduction in variance, which improves generalization on unseen data (James et al., 2023).

The model estimates the future points $\hat{y}_{i,t}$ for player i in gameweek t as a linear combination of their lagged rolling average metrics, especially for minutes played, goals, and expected metrics. The regression equation is defined as:

$$\hat{y}_{i,t} = \beta_0 + \beta_1\overline{Mins}_{i,t-1} + \beta_2\overline{G}_{i,t-1} + \beta_3\overline{A}_{i,t-1} + \beta_4\overline{CS}_{i,t-1} + \beta_5\overline{xG}_{i,t-1} + \beta_6\overline{xA}_{i,t-1} + \beta_7\overline{ICT}_{i,t-1}\ \epsilon_{i,t} \quad (5)$$

While Ordinary Least Squares (OLS) regression estimates coefficients by minimizing the residual sum of squares, the highly collinear explanatory variables in football performance data can lead to unstable coefficient estimates and therefore reduce generalization performance. To solve this issue the model is using Ridge Regression with L_2 regularization penalty. The main thing that is then happening is that it shrinks coefficient magnitudes and stabilizes the estimation process.

The coefficients (β) are estimated by minimizing the following objective function:

$$\min_{\beta}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\sum_{j=1}^{p}\beta_j^2\right) \quad (6)$$

Where the first term represents the residual sum of squares, and the second term acts as the penalty, shrinking the magnitude of the coefficients to reduce model variance.

All continuous feature variables were standardized to a mean of zero and a standard deviation of one as of z-score normalization. This was done before the model training started. It is a critical step for Ridge Regression because the penalty L_2 is so sensitive to the scale of input variables.

It was essential to ensure that there are no leaks in walk-forward validation. This was done by only using parameters from training to the test. Model was therefore kept clean from any other information on future gameweeks that could have ruined the whole process.

In this model there were no static penalty involved. Instead, the study generated a range of regularization strengths $(\lambda \in \{0.1, 1.0, 10.0, 100.0\})$ for every gameweek. The one that minimized the validation error was then chosen. This way the different changes that always come up during the long season were being adapted to.

Simulation was not just done using train-test split but by retraining repeatedly. This way more accurate environment of a fantasy manager was reached. A walk forward validation was adapted.

For every gameweek N in the validation set, the model is trained on all historical data available prior to N (Gameweeks $1\ to\ N-1$. Performance is assessed using Mean Absolute Error (MAE), averaging the deviation between predicted and actual points across multiple gameweeks to moderate the variance of single-week outcomes (Hyndman & Athanasopoulos, 2018).

3.2.3 Optimization Modeling

The team selection problem is formulated as a Binary Integer Linear Programming (BILP) model. While the standard Knapsack Problem optimizes for a single set of items, FPL requires a more complex structure involving distinct roles (Starters vs. Bench) and conditional logic (Captaincy).

Captain picking in this study is based on the highest predicted score, so the model assigns that player as a captain who has the most predicted points. The main target is to maximize the expected points of the starting XI. However, to have some optimization regarding the squad depth, which is in reality an important factor of the game due to all kind of changing elements in real life, the model uses a reduced weight ($\lambda = 0.1$) to the bench. This confirms it is being noted but no further sacrifices are being made for immediate returns.

Sets and Indices:

- $i \in I$ Set of all available players $\{1, \dots, n\}$
- $j \in J$: Set of Premier League teams $\{1, \dots, 20\}$
- $k \in K$: Set of positions$\{\text{GK, DEF, MID, FWD}\}$

Decision Variables:

$$start_i = \begin{cases} 1, \text{if player i is in Starting XI} \\ 0, \text{otherwise} \end{cases}$$

$$bench_i = \begin{cases} 1, \text{if player i is in Starting XI} \\ 0, \text{otherwise} \end{cases}$$

$$capt_i = \begin{cases} 1, \text{if player i is in Starting XI} \\ 0, \text{otherwise} \end{cases}$$

Objective Function:

The objective is to maximize the total projected score Z. The function includes a term for the captain ($capt_i$) to strictly ensure the solver assigns the armband to the player with the highest expected points (p_i)

$$\text{Maximize } Z = \sum_{i \in I} p_i(start_i + capt_i + 0.1 \cdot bench_i) \qquad (7)$$

(Where p_i is the predicted points $\hat{y}_{i,t}$ calculated in section 3.2.2)

While the objective function above includes the captaincy multiplier to drive optimal selection, the empirical results presented in Chapter 4 are calculated based on the aggregate base points of the squad (excluding the double-points multiplier). This distinction is made to isolate the predictive quality of the squad selection algorithm from the high variance inherent in the captaincy mechanic.

Constraints:

1. Squad Structure and Roles:

The squad must consist of exactly 11 starters, 4 substitutes, and 1 captain.

$$\sum_{i \in I} start_i = 11, \quad \sum_{i \in I} bench_i = 4, \sum_{i \in I} capt_i = 1 \tag{8}$$

2. Exclusivity Logic:

A player cannot be both a starter and, on the bench, and a player can only be captain if they are already selected as a starter.

$$\begin{aligned} start_i + bench_i &\leq 1, \quad \forall i \in I \\ capt_i &\leq start_i, \quad \forall i \in I. \end{aligned} \tag{9}$$

3. Budget Constraint:

The total cost of all selected players (starters and bench) must not exceed the £100m budget.

$$\sum_{i \in I} c_i(start_i + bench_i) \leq 100 \tag{10}$$

4. Team Constraints:

No more than 3 players may be selected from any single Premier League club j.

$$\sum_{i \in I:\mathrm{Team}(i)=j} (start_i + bench_i) \leq 3 \quad \forall j \in J \tag{11}$$

5. Positional Constraints (Squad Ownership):

The total 15-man squad must adhere to strict FPL quotas. The constant $Quota_k$ represents the required number of players for position k defined as $\{\mathrm{GK}: 2, \mathrm{DEF}: 5, \mathrm{MID}: 5, \mathrm{FWD}: 3\}$

$$\sum_{i \in I:\ \mathrm{Pos}(i)=k} (start_i + bench_i) = Quota_k \quad \forall k \in K \tag{12}$$

6. Valid Formation Constraints (Starting XI):

To ensure the Starting XI adheres to FPL regulations, the following bounds are enforced on the starters. These constraints ensure the model selects exactly one goalkeeper and a valid combination of outfield players (e.g., 3-4-3, 4-5-1, 5-3-2):

$$\sum_{i \in GK} start_i = 1$$
$$3 \leq \sum_{i \in DEF} start_i \leq 5$$
$$2 \leq \sum_{i \in MID} start_i \leq 5$$
$$1 \leq \sum_{i \in FWD} start_i \leq 3 \tag{13-16}$$

3.3 Data Validity and Reliability

Data validity and reliability are key elements of each study and to ensure both are being fulfilled comprehensive tracking should be performed, this chapter will be all about it from both point of views a scientific side and a practical side.

Data inputs will be investigated and confirmed that they are being used correctly as well as other tools related to handling the data itself. Therefore, it can be confirmed that the study and its results can be seen trusted and reducible.

3.3.1 Construct and Internal Validity

For all studies the starting point is that the data used must be valid. Otherwise, the empirical results cannot be seen to provide correct information on the topic of the study. In this study the validity of the data is confirmed by using the official FPL points and standardized performance metrics. These together with transparent rating system for Premier League performance provide objective data for usage. (McHale et al., 2012).

On the other hand, equally careful attention must be given to internal validity. In this case, all predictive features are lagged by one gameweek $t-1$. to eliminate any risk of data leakage. Future information leaking to a model in this case would cause a chaos with unstable and invalid results. Used walk-forward validation will ensure there is a clear and complete separation between the validation sets and the training data. It is only with historical data that is used with scaling and model fitting. This way it is to simulate as realistic as possible the scenario of every FPL manager.

3.3.2 Reliability and Deterministic Verification

Data preprocessing, predictive modeling, and squad optimization were implemented entirely in Pyhon 3.10. Parts of the coding process have been done with the support of artificial intelligence Gemini (Google, 2026). Feature scaling and the hyperparameter-tuned Ridge Regression were handled through Scikit-learn (Pedregosa et al., 2011) while the Binary Integer Linear Programming (BILP) formulation for squad selection was solved using the COIN-OR Branch and Cut (CBC) solver (Forrest & Lougee-Heimer, 2005) via the PuLP library (Mitchell et al., 2011).

Reliability rests on consistency and exact reproducibility. The study eliminates human intervention by means of fully automated end-to-end route in Python. Both Ridge Regression and Binary Integer Linear Programming (BILP) are deterministic, given the identical raw dataset and historical window, the algorithms will always return the same predicted values and the same optimal squad.

The optimization output is not trusted to heuristic judgement; it is verified mathematically by the BILP solver itself. Every squad generated is forced to respect the official £100m FPL budget and the positional formation rules, for example requirement of having at least three defenders etc., guaranteeing that each solution is not only optimal but immediately viable in the actual game.

3.3.3 External Validity and Scope of Simulation

The analysis is confined to the 2024-2025 English Premier League season. Consequently, the findings remain tied to the specific tactical environment and variance profile that campaign, whether injuries, team form or other season-specific factors and may not generalize without adjustment to seasons exhibiting markedly different scoring distributions.

To protect internal validity in the comparative evaluation, the study deliberately abstracts away longitudinal transfer constraints. In the real FPL game, the managers will get penalties, point reductions, if they make too many substitutions in between the gameweeks. So, a manager cannot change the whole squad every week. However, in our case this penalty is ignored. Reason for this is the necessary simplification. This way pure predictive signal of the model from the confounding effect of multi-week resource management is being isolated. This enables a direct skill vs. skill comparison against the heuristic baseline under identical conditions. In FPL context this is a “Free Hit” scenario as that chip allows manager to recreate the whole squad without being penalized.

As has already been noted, FPL scoring parameters did change in the 2025-2026 season. There were no major changes in the rules, but additional defensive contribution points

were added. However, the methodology developed here remains highly adaptable. Because the predictive-modeling and optimization components are modular, the framework can be reapplied to future seasons by simply recalibrating the objective function to reflect the new scoring weights thereby preserving the overall external validity of the approach. (*Fantasy Premier League, Official Fantasy Football Game of the Premier League*, n.d.)

4 Empirical Results

This chapter presents the empirical findings of the study, arranged so that each section speaks directly to one of the three research objectives set out in Chapter 1. As established earlier, the analysis follows the same three phases as the methodology. So, section 4.1 examines first the correlation analysis and thereby sets statistical grounding for the selected performance metrics. This is related to objective 1.

Section 4.2 evaluates the predictive performance of the used Ridge Regression model placing its error metrics alongside the heuristic benchmark. This is directly related to objective 2.

Section 4.3 reports the outcomes of the full optimization simulation. There cumulative points are being compared, squad differentiation is being analyzed and resource efficiency between the data-driven strategy and the heuristic version is being looked at. This provides information to fulfill the set objective 3.

During the chapter few different visualizations are used to support the findings. Each figure is being explained in text form also to provide comprehensive understanding of the results. At the end of the chapter there will also be a summary section combining all three parts together even though a separate chapter analyzing the total findings will follow afterwards.

4.1 Correlation Analysis Results

To address the first research objective correlation analysis was carried out to find out which indicators carry the most weight for FPL performance. The results are being presented in a heatmap which will provide a visual sight of how strong the correlation between the selected variables is. The darker colour the stronger correlation.

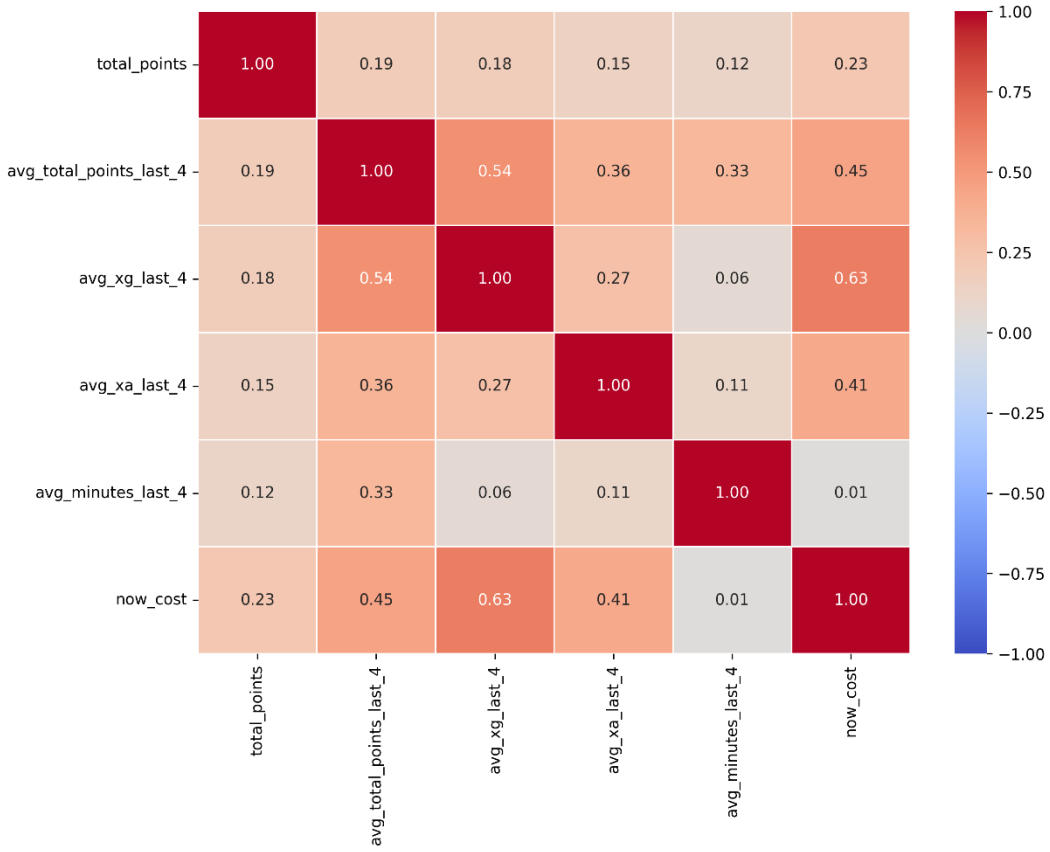

Figure 1 Heatmap

Figure 1 displays the Pearson correlation coefficients r between the selected player metrics and the target variable total points.

This correlation analysis shows that current price of the player (now_cost) stands out as the single strongest correlation at $r \approx 0.23$. These results suggest that the game's pricing mechanism functions are functioning very efficiently as the price of the player indicated the expected points, so the more player costs the more points he is expected to provide. Above all this seems to confirm that the game's underlying factors are built functionally. The price of the player in the game is set at the beginning of the season to a certain level, but it varies throughout the season depending on how many managers buy, price goes up, or sells, price goes down, the player each gameweek. Therefore, at this stage of the season the price – points correlation can be seen to be as expected.

Among the performance metrics, the four-week rolling form (avg_total_points_last_4) recorded the highest correlation number at $r \approx$ 0.19. This was narrowly ahead of the advanced statistics. Expected Goals followed at $r \approx 0.17$ and Expected Assists $r \approx 0.15$. Minutes played, by contrast, showed the weakest positive link at $r \approx 0.11$. The lower coefficient is telling; while playing time is clearly a necessary condition for accumulating points, it is not a linear driver of high returns in the same way as attacking threat. A player can earn the baseline points simply by appearing, yet the decisive difference comes from the quality actions.

What can be seen from here is that there are a clear signal of price and recent form that provides the strongest information together with the advanced metrics of xA and xG being also significant. All mentioned four parameters gave a positive correlation r > 0.1. The main purpose of the heatmap was to confirm that there is a positive correlation. Now, it can be stated that both deeper metrics and price and form can be taken into the model because they provide predictive value. This step showed a clear relationship and provided information about multicollinearity which strengthens the choice of using ridge regression.

4.2 Predictive Model Performance

The second objective set for this study was to find out the effectiveness of regression model. As mentioned, a walk-forward validation step was in use and therefore it was confirmed that the prediction was tested with unseen data. Performance was quantified with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) calculated across all active players so the one with $minutes$ > 0.

- Mean Absolute Error (MAE): 2.16
- Root Mean Squared Error (RMSE): 2.94

An MAE of 2.16 means that, on average, the model's forecast for a player's score missed the actual outcome by roughly two points. In a sport where single-match variance is

notoriously high, this level of error still offers a solid platform for finding consistent performers. The somewhat larger RMSE of 2.94 reflects the model's sensitivity to occasional haul weeks, hattricks or other outlier performances that linear methods inevitably struggle to capture.

This comparison also gave more insight into the model. The Ridge Regression model recorded a lower MAE of 2.16 than the heuristic baseline of 2.32 corresponding to a 6.9% relative improvement in accuracy. Ridge Regression clearly does, by regularization, shrink the coefficients of highly correlated features. This prevents short-term noise of the model for overfitting. Whereas the heuristic approach simply reacts to the most recent "lucky" spikes in points, the AI model achieves greater stability by keeping on using the balance between recent form against the signals coming from xG and xA. A better future gameweeks can be generalized by these forecasts.

It is important to note that even though the error values seem quite small by first sight the effect in practice is significant in FPL context. It is of course not possible to predict individual matches; however, an average of two-point prediction error is large enough to say the prediction is informative in this context. As noted above, the RMSE value does indicate there are occasionally large deviations. With linear statistical methods it is not realistic to predict the unexpected high impact events that do naturally happen in football every now and then. These random events impact highly to the squared error metric. In practice it can be stated that the model is mainly a tool for long-term squad selection not necessarily a magic way to get all high point players included in a single gameweek squad because even the best statistical models couldn't automatically eliminate the prediction error totally.

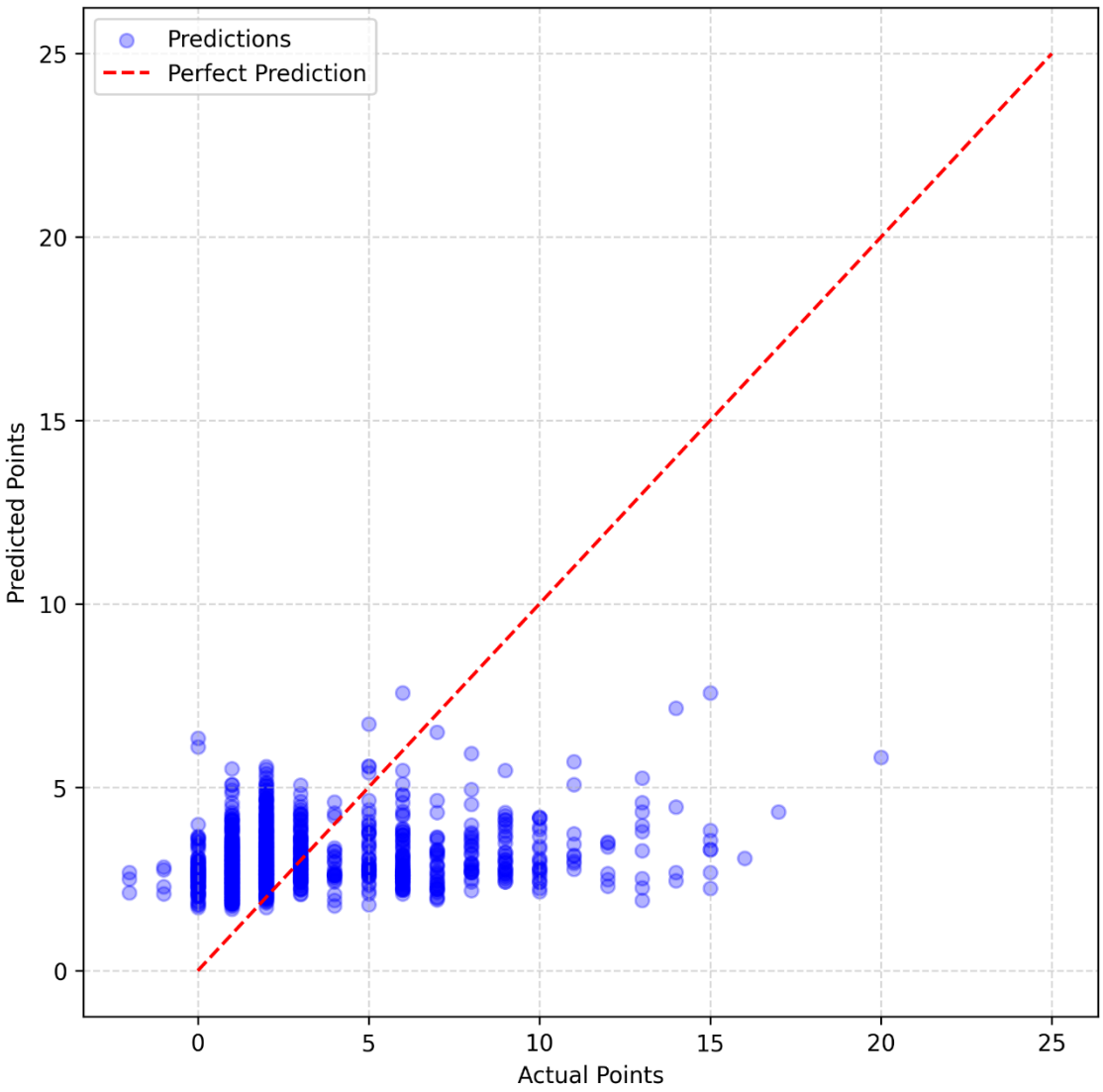

Figure 2 Scatter Plot

Figure 2 visualizes this performance, showing the relationship between predicted and actual points. The red diagonal line would be the perfect prediction so predicted 5 points would give you 5 actual points. As pointed out, that would be impossible scenario in this kind of environment. The blue dots in the figure represent individual player observations so each dot is an individual player during the evaluation window, a single gameweek in this model.

What we can see clearly is that the scatter distribution is quite narrow. Most of the dots are in between two and five points. In FPL scoring this is the expected outcome and confirms that major part of the players earns modest amounts of points in single gameweek. Therefore, the regression model naturally aims to produce quite conservative prediction that is around the common numbers.

When we then look further down the line, we can see that the actual points start to be in some cases further away from the red perfect prediction line. This indicates that there is a systematic underestimation of the high-scoring events so when some player scores multiply goals during the same game for example. This highlights our notice of how events with more randomness are difficult to capture by linear models.

However, limitations are not the only thing we can see from the figure. The outcome is that even with these limitations the model does catch the general scoring distribution reasonably well. Even if it is not possible to predict the outliers perfectly the typical point range prediction is still valuable on its own. Therefore, the figure verifies the reported MAE and RMSE values.

4.3 Optimization Results and Team Comparisons

To see how the optimization would work a team comparison was needed so the step was to see how the data-driven model would then compare against the heuristic methods. To evaluate the effectiveness of the proposed model, Walk-Forward Validation was run from Gameweek 25 to Gameweek 29. Here the performance of the data science model which utilizes ridge regression and BILP was compared to heuristic version where the squad was selected purely based on the average points from the last four gameweeks.

Table 1 Comparative Performance (Gameweeks 25-29)

Gameweek	Model	Heuristic	Net Difference
GW25	66	74	-8

GW26	72	66	+6
GW27	58	63	-5
GW28	61	61	0
GW29	50	35	+15
Total	307	299	+8

Table 1 above shows the total outcome of the comparison. As we can see during gameweeks 25-29 so this five gameweek period the data-driven model accumulated a total of 307 and the heuristic strategy ended up in total of 299. The net difference in final score was therefore 8 points.

When we take a closer look at the individual gameweek results we can see that the regression model did not outperform the heuristic approach in each gameweek. During gameweeks 25 and 27 the model lost by 8 and 5 points. The major change happened in gameweek 29 where the model beat the heuristic version with 15 points. These results highlighted the earlier note that unexpected random events leading to single players scoring high points can significantly influence the score on a weekly level.

Benefits of data-driven decision-making are not dominating short-term scoreline but achieving better results in long-term. FPL season lasts in total of 38 gameweeks so even modest predictive tools can accumulate into meaningful differences over time when applied repeatedly. Therefore, it should be noted that 8 points difference during 5 gameweeks is already a 2.7% overall improvement.

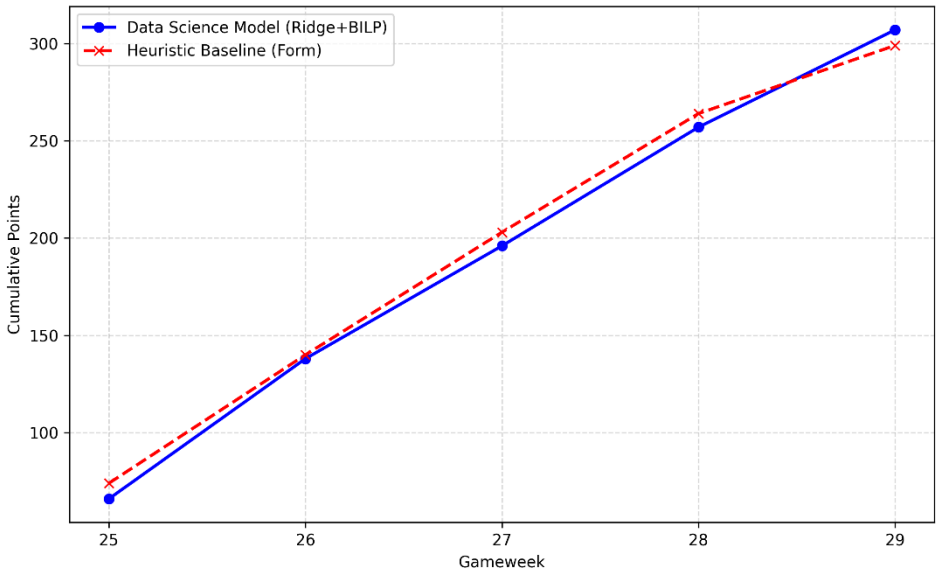

Figure 3 Cumulative performance

In Figure 3 a cumulative performance has been presented. It clearly shows the drop of the red line (heuristic) at the end, on gameweek 29, when the blue line (model) kept delivering points more steadily.

More than just total score, the model displayed huge resource efficiency. Every gameweek there is a £100m budget of which only £80.1m was used in the final gameweek by an optimized model. This is a great reminder that better players and more points are not only about the prize but looking at the underpriced options is worth it. Leaving 20% of the budget is a major move for upcoming gameweeks also and in itself a strong mark of success in any area or industry.

In total there were only 6 same players picked out of 15 in gameweek 29. Recent form was highly overweighted by heuristic model. This willingness to bet on positive regression rather than the latest hot streak allowed the model to secure the victory without even fully exhausting financial resources.

4.4 Summary of Empirical Findings

As a summary of the empirical analysis, it can be confirmed that underlying performance metrics do serve as reliable predictors of future success though they operate within a clear hierarchy. When looking at the correlation results strongest primary signals came from Player Price ($r \approx 0.23$), Past Form ($r \approx 0.19$) and Expected Goals ($r \approx 0.17$). The Ridge regression model made effective use of these signals and delivered solid predictive accuracy recording a MAE of 2.16, which was a 6.9% improvement compared to the heuristic baseline.

Finally, the optimization model was used to translate forecasts into a competitive squad selection strategy. This was done in a five-gameweek period. Data-driven approach produced a total of 307 points while the heuristic one ended up in 299 total points. The model won in the end by 8 points. During the period the heuristic baseline also did beat the model during individual gameweeks, which was due to short-term variance which is typical in football.

Another stand out point was when the squad selection was looked at from resource efficiency point of view. Model's last gameweek squad only operated with the cost of £80.1m of the available £100m budget. Taken together these findings lend empirical support to the central hypothesis: algorithmic optimization can outperform manual heuristics even if there are strict constraints in use.

5 Discussion

In this chapter the empirical findings are being put together and further discussion on the performance of the data-driven model is being set to the wider theoretical context. Heuristic decision and data-driven decision-making are being compared with resource usage, volatility and risk profiles are all being covered. This will be followed by a more practical section where some possible strategies are being presented to the FPL managers. At the end of the chapter the most important behavioral biases are being discussed such as recency bias mainly from that perspective that how it was possible to reduce the risk of it with the data model usage.

5.1 Interpretation of Results

This study was about finding out if it is possible by utilizing data-driven decision-making tools to build a fantasy premier league squad that could give clear benefits compared to just selecting the team trusting on intuition. Correlation analysis, predictive modeling and optimization were used. Finally, a comparison of DDDM team against heuristic selection squad was tested in a five-week validation period in gameweeks 25-29. The main target was to provide further insight into an existing research gap on can DDDM give advantage in fantasy sports world also as well as already confirmed in other industries. During the comparison both selections did win individual gameweeks, however, it was clear that risk management and sharper value identification ensured that the DDDM model collected more total points in aggregate.

Total points of the DDDM model were 307 while the heuristic model ended up at 299 points. This means a difference of 2,7% in the five-week period. A clear difference in volatility can be seen when looking behind the total scores. The heuristic model scored in between 35 – 74 points so the gap was 39 points when the DDDM scored more consistently by its gap between peak and low was 22 points effectively compressing the standard deviation of returns. What this points to is a distinctly lower-risk profile for the

algorithmic path, one that shields the manager from the sharp ups and downs that extreme form-chasing strategies.

The difference was particularly clear in gameweek 29. The data model scored 50 points while the heuristic model only scored 35 points. This highlights a key limitation of the heuristic approach. By relying solely on past points heuristic version ended up selecting players after strong recent performances. Even though their performance is likely to regress to the mean. On the other hand, that is exactly where the data model can bypass the trap. By using leading indicators such as xG it can spot the production potential before it translates into actual points.

The results also cast performance in terms of market efficiency. The optimization model outperformed the heuristic one while deploying just £80.1m of the available £100m budget. In doing so it demonstrated that price does not always reflect true value, homing in on underpriced assets that the pricier heuristic overlooked entirely. This is notable also because it would feel risky or even stupid if a regular heuristic FPL manager would not utilize the whole budget. During some weeks there can of course be situations where you could leave something if you're intending to make some more changes during the upcoming gameweeks, however, leaving large portion unused budget is not common.

Finally, the squads themselves differed substantially by overlapping only with 40% so 6 out of 15 players. That performance margin was therefore delivered exclusively from 60% of these differential slots. There the model operated with contrarian stance prioritizing future performance potential rather than history.

5.2 Implications for Fantasy Sports Strategy

The findings point toward a quiet but potentially important shift in how FPL managers might think about building a squad. Even the test window of five weeks used in this work is really limited, the results of the optimization model doing great still suggest that

greater emphasis should be placed on process metrics. So, use of leading indicators such as xG should be prioritized over the lagging signals of raw past points.

This hints that it is valuable to try to find the players who are showing potential in advance statistics even if their current performance shows lower points scored. The DDDM model did exactly that and avoided the recency bias by trusting the numbers. At the same time the heuristic model clearly stepped into that trap.

When it comes to resources the DDDM model performed significantly better and by leaving almost 20% on the budget on the last gameweek it proved the price solely is not the perfect value indicator in FPL. What it suggests for everyday strategy is that managers often overpay for premium assets whose cost owes more to reputation than to near-term expected output. A value-investing mindset, so consciously targeting mid-priced players who carry high statistical ceilings, can therefore prove statistically superior to the instinct to spend the whole budget. However, we need to remember that player's price was solely the strongest signal of the points, so what is important to understand here is that the more points the player is scoring the more he is being picked and the more his price is rising as defined in the rules. This highlights even more that finding the right player at the right time and not when it is too late is the way to go and the place where these kinds of models can help the player.

All in all, it can be said in conclusion that optimization tools are in need to support the human mind. They are not to replace the intuition, but they can serve as a powerful check on it. The major help comes when a manager is in need of finding a differential pick that has not been performing lately.

5.3 Reflection on Data-Driven vs. Heuristic Decision-Making

The focus of this comparison test was to find out how can DDDM compete against heuristic model. The heuristic model did score quite well but the main finding of its weaknesses can well be seen in the last gameweek when it scored very low. This failure falls

directly into the Recency Bias (Tversky & Kahneman, 1974). Heuristic model trusted too much on the current form of the players. The shorter the window is the more possibility of course is to avoid the total collapse of points, however, already in this five-week period it happened which proves that if one wants stable success a DDDM model way of working can provide support.

The DDDM model trusted the indicators and chose players that were perhaps not “hot” put providing good metrics with areas such as xG which was proved to have magnificent correlation. The ridge regression model proved that it is capable of this. It leads to a discussion that in the most random environments which football clearly is, analytical tools cannot just help to find the right choices but perhaps even more importantly allow the decision maker to put own, misguiding, feelings aside and trust to make the decision that the data shows are the best one.

6 Conclusion

Conclusion chapter of this thesis presents findings and mirrors them back to the original research question of how can data-driven decision-making enhance performance in Fantasy Premier League?

Chapter 6.1 is a summary of the whole study. It provides the key findings and also compares the results to the original objectives and analyzes if those were met. This follows by 6.2. which takes into account how does this study places in wider DDDM framework and practically to a fantasy premier league managers. Final chapter 6.3 opens a discussion on the possible future studies related to this topic and notes the limitations of the current study.

6.1 Summary of the Study

This thesis focused on how data-driven decision-making tools can be used to be more successful in Fantasy Premier League. The study followed a clear structure of data collection to a final team comparison where DDDM model competed against the heuristic model. The study included correlation analysis, predictive modeling and optimization. The final test was done in five gameweek validation set. Optimization was based on Integer Programming while the predictive modeling was using ridge regression. The data used in the study was from 2024-2025 season.

The results of the comparison were clear and the DDDM model scored 2,7% more points than the heuristic model. Another major difference was in predictive error where 6,9% difference was recorded in favor of the DDDM model. Other findings were spotted when looked more closely at the gameweeks. The DDDM model was clearly more stable and efficient throughout the validation period. One of the clearest examples of DDDM model advantages was also when it used only around 80% of the budget and yet scored more points than the heuristic model. Overall, all volatility numbers where lover with DDDM

model. The outcome of the study is that there exists a clear benefit of using DDDM tools in Fantasy Premier League.

6.2 Contributions to Theory and Practice

This study was another one for the long history of the decision-making process topic where data is trying to overcome human mind as in the lens of Bounded Rationality (Simon, 1955). Even just the five-week validation period proved that in noisy environments a human mind is tempted to fall into traps while numbers stay on track. In this work the heuristic model failure in gameweek 29 was clear evidence of how recency bias works when more complexity is added. The DDDM model meanwhile showed that trusting Integer Programming formulation a more systematic approach can be achieved. This is a small step in combining these two topics of DDDM and fantasy sports, but it clearly does provide shows that the latter one can benefit from the analytical tools when used properly.

Practically looking, this work was great example of how things related to DDDM have become more available to everyone since the entire work was built with Python which is an open-source tool and with data that is publicly available to everyone. Competitive performance does not necessarily need huge enterprise level environment. The work offers a fully replicable model that any independent analyst can adopt. It lowers the barrier to use of optimization models.

There is also a clear practical lesson on resource use. The optimization squad secured its 8 point total advantage while not spending the whole budget and in single gameweeks used significantly fewer resources than the heuristic. The single finding validates a value-investing mindset: sustained success appears to come less from maximizing expenditure and more from consistently spotting underpriced utility through underlying metrics rather than chasing headline reputation. For everyday managers the implication is straightforward and sometimes it is a smart move to leave money unused.

6.3 Limitations and Future Research

The study openly acknowledges several simplifications in game mechanics, each of which helps to define the precise boundaries within which empirical findings can be interpreted.

First, the simulation treated every validation gameweek as an independent optimization task built around a fixed £100m budget, effectively replicating a Free Hit scenario on weekly basis which is one of the chips that can be used in a real game where you are allowed to only make certain amount of changes in between the gameweeks or otherwise you will be penalized by point losses from extra changes made. So, in live FPL, managers operate under dynamic economic conditions in which team value fluctuates with market movements and successful play can push the budget well beyond the initial ceiling. The present work did not model capital appreciation or the impact of transfer penalties. Future research could therefore extend the work by moving to a multi-stage stochastic programming that jointly optimizes short-term point accumulation and longer-term team-value growth.

Second, certain high-variance mechanics were deliberately excluded to isolate the core predictive signal. Single-use chips were the first thing. So, on top of the free hit there are Triple Captain, Bench Boost and Wildcard. These introduce substantial non-linearities and are usually deployed in response to fixture anomalies rather than pure player performance. Next on the list is captaincy. Results were calculated on the raw aggregate output of the starting XI without applying the double-points multiplier. This was intentional because it ensured that a single binary decision did not distort the broader evaluation of the squad-selection model. Also, bench management was overlooked. The performance metric assumed the selected starting XI would feature in every match. No automatic substitutions or bench-order optimization were simulated. Consequently, the reported totals reflect the theoretical ceiling of the primary selection rather than the safety-adjusted outcome required for a Bench Boost strategy. There will of course always be some players missing some games without information being available before the

deadline, so some bench optimization and budget allocation is needed when looking at a longer term.

Third, while the data-aggregation step correctly summed statistics for historical double gameweeks, the predictive model relied exclusively on rolling averages and contained no explicit fixture-volume features. The regression therefore treated every future gameweek as equal in volume, which may have led to under-prediction for upcoming double gameweeks. Future studies could usefully incorporate an “upcoming fixture count” variable to allow the model to weight volume-driven opportunities more accurately. Double or blank gameweeks happen during every season and mainly during the latter part of the season when the other competitions than Premier League come to a picture and require changes in the original fixture list.

Finally, player availability was assessed primarily through rolling averages. This approach inevitably lacks the real-time granularity that manager encounters with for example injury news. Future iterations could address this limitation by integrating an external API to impose hard constraints on flagged players, thereby preventing the selection of assets who sustained injuries immediately before the deadline.

References

Anand, V. (2024). *Fantasy-Premier-League/data/2023-24/players at master · vaastav/Fantasy-Premier-League*. GitHub. https://github.com/vaastav/Fantasy-Premier-League/tree/master/data/2024-25

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Brechot, M., & Flepp, R. (2020). Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. *Journal of Sports Economics*, *21*(4), 335–362. https://doi.org/10.1177/1527002519897962

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3). https://doi.org/10.1214/ss/1009213726

Brynjolfsson, E., & McElheran, K. (2016). The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review*, *106*(5), 133–139. https://doi.org/10.1257/aer.p20161016

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business Review Press.

Fantasy Premier League, Official Fantasy Football Game of the Premier League. (n.d.). Retrieved January 24, 2026, from https://fantasy.premierleague.com/help/rules

Forrest, J. J., & Lougee-Heimer, R. (2005). CBC User Guide. *Emerging Theory, Methods, and Applications*, *2005*(ORMS Today, 32(1)).

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6

Google. (2026). *Gemini* . https://gemini.google.com/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470. https://doi.org/10.1016/j.ijforecast.2009.10.002

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer Nature.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Martello, S., & Toth, P. (1990). *Knapsack problems: Algorithms and computer implementations*. Wiley & Sons.

McHale, I. G., Scarf, P. A., & Folker, D. E. (2012). On the Development of a Soccer Player Performance Rating System for the English Premier League. *Interfaces*, *42*(4), 339–351. https://doi.org/10.1287/inte.1110.0589

Mitchell, S., O'Sullivan, M., & Dunning, I. (2011). *PuLP: A Linear Programming Toolkit for Python*.

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, *5*(4), 213–222. https://doi.org/10.1007/s41060-017-0093-7

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, *263*(2), 611–624. https://doi.org/10.1016/j.ejor.2017.05.005

O'Brien, J. D., Gleeson, J. P., & O'Sullivan, D. J. P. (2021). Identification of skill in an online game: The case of Fantasy Premier League. *PLOS ONE*, *16*(3), e0246698. https://doi.org/10.1371/journal.pone.0246698

O'Donoghue, P. (2010). *Research methods for sports performance analysis*. Routledge.

Pedregosa, F., Pedregosa, F., Varoquaux, G., Varoquaux, G., Org, N., Gramfort, A., Gramfort, A., Michel, V., Michel, V., Fr, L., Thirion, B., Thirion, B., Grisel, O., Grisel, O., Blondel, M., Prettenhofer, P., Prettenhofer, P., Weiss, R., Dubourg, V., …

Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Sports Analytics*, *3*(1), 21–29. https://doi.org/10.3233/JSA-170149

Saunders, M. (2007). *Research Methods for Business Students*. Pearson Education UK. http://ebookcentral.proquest.com/lib/tritonia-ebooks/detail.action?docID=5139642

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572. https://doi.org/10.2307/41409977

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99. https://doi.org/10.2307/1884852

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Winston, W. L. (2004). *Operations research: Applications and algorithms* (4th ed.). Brooks/Cole.

Wright, M. B. (2009). 50 years of OR in sport. *Journal of the Operational Research Society*, *60*(1), S161–S168. https://doi.org/10.1057/jors.2008.178