rodur_thesis.pdf

Indian Institute of Technology Jodhpur

Document Details

Submission ID
trn:oid:::29334:139895522

Submission Date
May 21, 2026, 7:06 PM GMT+5:30

Download Date
May 21, 2026, 7:07 PM GMT+5:30

File Name
rodur_thesis.pdf

File Size
1.0 MB

60 Pages

18,472 Words

102,547 Characters

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 10 words)

Match Groups

54 Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

1 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
2% Publications
4% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **54** Not Cited or Quoted 6%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **1** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
2% Publications
4% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	cn.overleaf.com 2%
2	Internet	arxiv.org <1%
3	Submitted works	Delhi Technological University on 2026-05-21 <1%
4	Internet	zdoc.site <1%
5	Internet	dspace.dtu.ac.in:8080 <1%
6	Internet	www.coursehero.com <1%
7	Internet	deazone.com <1%
8	Internet	ijorai.reapress.com <1%
9	Internet	dokumen.pub <1%
10	Internet	sanad.iau.ir <1%

11 Submitted works
Beykent Universitesi on 2016-11-08 <1%

12 Submitted works
University of Petroleum and Energy Studies on 2015-02-11 <1%

13 Internet
www.3addedminutes.com <1%

14 Submitted works
Patrick Henry High School on 2015-04-14 <1%

15 Publication
Greg N. Gregoriou, Fabrice Rouah, Stephen Satchell, Fernando Diz. "Simple and cr... <1%

16 Internet
eprints.utm.my <1%

17 Submitted works
The Hong Kong Polytechnic University on 2009-04-22 <1%

18 Internet
m.moam.info <1%

19 Internet
qspace.qu.edu.qa <1%

20 Internet
vdoc.pub <1%

21 Publication
L. Martin Cloutier. "Relative Technical Efficiency: Data Envelopment Analysis and ... <1%

22 Submitted works
National University of Singapore on 2015-11-25 <1%

23 Submitted works
Higher Education Commission Pakistan on 2011-05-02 <1%

24 Publication
Onut, S.. "Analysis of energy use and efficiency in Turkish manufacturing sector ... <1%

25	Publication	Ranjan Chaudhuri, Vijay Prakash Gupta, Navita Nathani, Dipak Saha. "Navigating ... <1%
26	Publication	Sanford Berg, Chen Lin. "Consistency in performance rankings: the Peru water se... <1%
27	Internet	abis-files.anadolu.edu.tr <1%
28	Internet	archive.org <1%
29	Internet	d-nb.info <1%
30	Internet	scholar.archive.org <1%
31	Publication	"Handbook of Analytical Studies in Islamic Finance and Economics", Walter de Gr... <1%
32	Publication	B. P. Branco Da Silva. "A ranking for the Olympic Games with unitary input DEA m... <1%
33	Submitted works	Erasmus University of Rotterdam on 2021-09-10 <1%
34	Publication	Lecture Notes in Computer Science, 2014. <1%
35	Submitted works	Sha Tin College on 2020-10-05 <1%
36	Submitted works	Sultan Qaboos University on 2024-12-29 <1%
37	Submitted works	Universiti Teknologi Malaysia on 2025-02-26 <1%
38	Submitted works	University of Belgrade, Faculty of Organizational Sciences on 2021-08-24 <1%

39 Submitted works
University of Galway Canvas on 2026-03-12 <1%

40 Submitted works
University of South Africa on 2015-12-15 <1%

41 Internet
backend.orbit.dtu.dk <1%

42 Internet
elea.unisa.it <1%

43 Internet
sportsem.uv.es <1%

44 Internet
www.scielo.br <1%

45 Internet
www.sportsjoe.ie <1%

46 Internet
www.tdx.cat <1%

FORMATION OF AN EFFICIENT FOOTBALL TEAM USING DATA ENVELOPMENT ANALYSIS AND INTEGER LINEAR PROGRAMMING

A PROJECT REPORT

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

MASTERS OF SCIENCE

IN

DEPARTMENT OF APPLIED MATHEMATICS

Submitted by:

MOKSH JAIN (24/MSCMAT/03)

RODDUR MITRA (24/MSCMAT/19)

Under the supervision of

PROF. ANJANA GUPTA

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

We, **MOKSH JAIN(24/MSCMAT/03) & Roddur Mitra (24/MSCMAT/19)** students of M.SC Applied Mathematics, hereby declare that the Project Dissertation titled "**FORMATION OF AN EFFICIENT FOOTBALL TEAM USING DATA ENVELOPMENT ANALYSIS AND INTEGER LINEAR PROGRAMMING"** which is submitted by us to the Department of Applied Mathematics, DTU, Delhi in fulfilment of the requirement for awarding of the Master of Science degree, is not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Fellowship or other similar title or recognition.

Place: New Delhi

Date: 23/05/2026

MOKSH JAIN **RODDUR MITRA**

(24/MSCMAT/03) **(24/MSCMAT/19)**

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project titled " "**FORMATION OF AN EFFICIENT FOOTBALL TEAM USING DATA ENVELOPMENT ANALYSIS AND INTEGER LINEAR PROGRAMMING**" which is submitted by MOKSH JAIN(24/MSCMAT/03) & Roddur Mitra (24/MSCMAT/19) for fulfilment of the requirements for awarding of the degree of Master of Science (M.SC) is a record of the project work carried out by the students under my guidance & supervision. To the best of my knowledge, this work has not been submitted in any part or fulfillment for any Degree or Diploma to this University or elsewhere.

Place : New Delhi

Date: 23/05/2026

PROF. ANJANA GUPTA

(SUPERVISOR)

Professor

Department of Applied Mathematics

Delhi Technological University

TABLE OF CONTENTS

ACKNOWLEDGEMENT

The successful completion of any task is incomplete and meaningless without giving any due credit to the people who made it possible without which the project would not have been successful and would have existed in theory. First and foremost, we are grateful to **Dr. (Prof) R. Srivastava**, HOD, Department of Applied Mathematics, Delhi Technological University, and all other faculty members of our department for their constant guidance and support, constant motivation and sincere support and gratitude for this project work. We owe a lot of thanks to our supervisor, **Dr. (Prof) Anjana Gupta**, Department of Applied Mathematics, Delhi Technological University for igniting and constantly motivating us and guiding us in the idea of a creatively and amazingly performed Major Project in undertaking this endeavour and challenge and also for being there whenever we needed her guidance or assistance. We would also like to take this moment to show our thanks and gratitude to one and all, who indirectly or directly have given us their hand in this challenging task. We feel happy and joyful and content in expressing our vote of thanks to all those who have helped us and guided us in presenting this project work for our Major project. Last, but never least, we thank our well-wishers and parents for always being with us, in every sense and constantly supporting us in every possible sense whenever possible.

MOKSH JAIN

(24/MSCMAT/03)

RODDUR MITRA

(24/MSCMAT/19)

ABSTRACT

In today's professional football environment, building a competitive squad while operating under tight financial constraints has become a major challenge for club management. This study presents a two-stage mathematical approach for constructing an optimal football team in the English Premier League. In the first stage, Data Envelopment Analysis (DEA), specifically the BCC (Banker–Charnes–Cooper) model, is employed to assess the relative efficiency of individual players viewed as decision-making units. Player salaries are treated as inputs, while on-field performance indicators such as goals and assists are considered outputs.

In the second stage, the efficiency scores obtained from DEA are incorporated into an Integer Linear Programming (ILP) framework. The ILP model selects the "Best Eleven" by maximizing overall team efficiency subject to realistic constraints including positional requirements, budget limitations, and formation rules. The results demonstrate how mathematical optimization techniques can support football clubs in assembling high-performing teams while maintaining financial discipline.

Keywords: Data Envelopment Analysis, Integer Linear Programming, Efficiency, Optimization.

CHAPTER 1

INTRODUCTION

1.1 Background: From Instinct to Analytics in Football

Football has been one of most famous games in the history of sports. A lot of nations and clubs get involved with best of their strategies and planning. Generally, Tactical decisions, player recruitment, and squad selection rested on the judgment of managers and scouts whose authority derived from experience rather than data. A players was signed because the manager has seen him playing since a couple of years or may be the overall experience of the player , but sometimes it get failed due to different reasons and at that point of time the return of interest is very low.

This was tracked properly in 2002 in Oakland Athletics, where a small franchise used statistical analysis to assemble a competitive roaster on a fraction of the budget of their rivals. This became popular when the a movie *Moneyball* was released which demonstrated the systematic evaluation of performances v/s financial resources. Football clubs, especially under the shadow of wealthier clubs took this in notice.

In 2010, there was a boom in data availability of the football matches, it included each and every detail happened in the match. Concepts such as expected goals and expected passes were introduced in it. It lead to a big data science department in this industry and hence its recommendations sits along side, and occasionally above, those of traditional suggestions.

Yet the data availability and data driven decisions are far apart different things. Now a days it is very important for a club that with a large amount of data it does not continue to make decisions based on subjective narrow headline of statistics. So the questions arises, How can multi-dimensional performance data, expressed in incommensurate units, be combined into a single defensible judgement of player's value to club?

1.2 The Premier League and the Squad-Construction Problem

33 The English Premier League is the most important league which is played across the globe. It was established in 1992 and has grown into the world's richest domestic football competition, with 20 clubs collectively spending in about 2 billion Euros on a player in a single season. It leads to a very important factor of picking up the players right under a proper judgment and

financial constraints. A single misjudging can run into the tens of millions of pounds. Moreover, the introduction of UEFA's financial fair play regulations and the premier league's own profit and sustainability rules has made every decision consequential beyond the playing merits of player concerned. This means even the richest club has financial constraint.

The task of assembling a starting eleven appears very simple at first. A manager must choose one goalkeeper, four or five defensive players, a midfield of three or four, and an attacking line up of 1-3 players, the problem include a complex web of constraints. Each formation specifies the number of players required in each positional department. The total wages of the selected eleven must lie within a club-specific budget window. Each player contributes along multiple areas such as goals, assists, tackles, interceptions, distribution that cannot meaningfully be added together in their raw units. Goalkeepers and forwards cannot be compared on the same metrics as their role is quite different from the other twos. Also within a single league there are several hundred candidate players, so enumerating every legal eleven is computationally not traceable and doing it with just intuition can leadto bias.

These properties like multi-objective evaluation, multiple constraints, and a combinatorial choice space — place squad construction squarely in the class of problems for which mathematical optimisation was designed. What has been missing, until relatively recently, is the data to populate such a model.

1.3 Genesis of This Study

The motivation for this work arose from a simple observation. In any given EPL season, It is observed that the generally a team with highest wage bill is not the winner . In 2015–16 Leicester City won the league with a wage bill less than a quarter of that of Manchester City. In 2023–24 Aston Villa qualified for the Champions League ahead of Tottenham, Chelsea, and Manchester United, despite operating with a far smaller squad budget. This is a very observable pattern and is sign that clubs can work upon it as some clubs are tactically showing better performance per pound than others.

A natural question follows. If clubs differ in the efficiency with which they convert wages into performance, can this be judged on the basis of individual performance of the players? Can we evaluate the efficiency of the players by the outputs they give by their performance and the input they have taken from the club? And using these scores can a club use it in to form a proper playing 11 so that they are under best budget constraints and maximized efficiency.

These two process are different but can be clubbed in a single project of evaluation and selection. An efficiency score in particular tells who individual is its best signing . A process without evaluating efficiency is logicless. So this study contributes to the integration of two stage mathematical framework where players are first evaluated on the basis of their previous performances within their tactical department and the resulting efficiency scores are then fed into an optimisation model that selects the starting eleven.

1.4 Two Lenses: Efficiency and Optimisation

The theoretical foundation of Data Envelopment Analysis (DEA) was given by Charnes, Cooper, and Rhodes for the first time in 1978. who introduced the CCR model based on the calculations of Constant Returns to Scale (CRS). But recognizing its all limitations, like the model can't separate the technical and scale efficiency at all and assuming all DMU's operate at the optimal scale, then **Banker, Charnes, and Cooper introduced the BCC model in 1984**, which has Variable Returns to Scale (VRS).

Basically , DEA is a non parametric technique. It evaluates the relative efficiency of decision making units which are the players in this case. It takes multiple inputs and produce multiple outputs based on the player matrix. Each player is evaluated by assigning weights to each of its input and outpuy factor such that it is most suitable to him or his score would reflect the best case that can be relatable to his peers.

The second step is **Integer Linear Programming (ILP)**. Once each player has been assigned an efficiency score, the squad-construction problem reduces to a binary decision over the entire player pool i.e we should include this player or not? The objective of the problem here becomes to maximize the net efficiency of the team under the constraints of budget and positional requirement. ILP is the natural mathematical machinery for this kind of problem. It is solved by **Branch and Bound,** It is a systematic algorithmic paradigm designed to solve complex combinatorial optimization problems by systematically exploring the complete solution space. This approach avoids exhaustive search by partitioning the main problem into smaller, manageable sub-problems, a process formally known as **branching**. This creates a tree-like hierarchical structure of potential solutions.

To optimize the search efficiency, the algorithm computes a scalar value, termed a **bound**, for each active node or sub-problem. This bound represents the theoretical best outcome achievable from that specific path. By continuously comparing these local bounds against the

globally known optimal solution found so far, the algorithm dynamically eliminates unpromising paths without fully exploring them. This elimination process, known as **pruning**, significantly accelerates the discovery of the exact optimal solution while reducing computational overhead.

The combination is not arbitrary. DEA provides a systematically computed scores of the players which includes multiple factors for the calculations whereas ILP provides a principled mechanism for moving from individual evaluation to final selection under planned constraints. Each technique has a substantial literature and research including applications and projects in sports, but their combination in the specific context of EPL squad construction is unique to the best of our knowledge.

1.5 Research Objectives

The objectives of this study are as follows:

1. To compile a data set which include the information about player salary , age ,minutes played, goal ,assist, tackles and various on filed peformances and normalizing that data.
2. To compute relative efficiency scores for every player using the Banker–Charnes–Cooper (BCC) model of DEA, with separate frontiers in different departments which will ensure the homogeneity in the team.
3. To extract the starting 11 selection using the INTEGER LINEAR PROGRAMMING with binary decision variable with the objective that maximizes the net efficiency of the team under budgetary constraints and squad positional requirements.
4. To solve the model under multiple realistic scenarios —different formations, budget envelopes, and club's elite player requirements — and to compare the resulting starting elevens against benchmark consensus selections.

CHAPTER 2

LITERATURE REVIEW

The framework proposed in this study depends on two different path of operations-research literature: Data Envelopment Analysis i.e DEA, which is used to measure the relative efficiency and the formulation of team-selection problems as mathematical program using Integer Linear Programming. In this chapter we are going to discuss each path from its theoretical foundations to its most relevant applications in sports, and concludes by discovering the gap that the present work is intended to address.

2.1 Foundations of Data Envelopment Analysis

The theoretical foundation of Data Envelopment Analysis (DEA) was laid by Charnes, Cooper, and Rhodes [5], who proposed a non-parametric linear-programming framework for evaluating the relative efficiency of decision-making units (DMUs) that transform multiple inputs into multiple outputs. Their model, now known as the **CCR model**, defined the efficiency of a DMU as the ratio of a weighted sum of outputs to a weighted sum of inputs, with the weights chosen by the DMU itself so as to maximise its own score subject to the constraint that no DMU could achieve a score greater than unity under the same weights. The DMUs that achieve a score of one form the empirical efficiency frontier; the remainder are deemed inefficient relative to that frontier. The appeal of the formulation is that it requires no a priori judgment about the relative importance of different inputs and outputs, a feature that has made DEA attractive in fields ranging from banking to healthcare to public administration.

We supposed that, original CCR model is **constant returns to scale (CRS)**, in which a proportional increase for inputs was expected a proportional increase in outputs as well. This CRS Model is very limited in many experimental works like football, where the relationship between player wages and on-field performance is clearly non-linear. Let's suppose, we doubled a player salary, which does not ensure us that the player would double his goals compared to previous season; nor does a small increase in playing time equally increase the tackles. After finding this limitation, Banker, Charnes, and Cooper [2] developed the **BCC model**, which relaxes the constant-returns assumption to allow **variable returns to scale (VRS)** through the addition of a convexity constraint on the reference DMUs. The BCC model separates pure technical efficiency from scale efficiency, and a DMU may now be technically

efficient at its own scale even if it would be inefficient under CRS assumptions. The choice between CRS and VRS in any application turns on a substantive judgment about the underlying production technology, and in the present study the BCC model is adopted for the reasons just noted.

A well-known limitation of basic DEA — both CCR and BCC — is its tendency to declare large numbers of DMUs efficient simultaneously, with all such units receiving an identical score of one and no further differentiation among them. In a typical football dataset this can result in dozens of players sharing the maximum score, complicating any downstream ranking. Two subsequent refinements have addressed this difficulty.

Sexton, Silkman, and Hogan [14] introduced **cross-efficiency analysis**, in which each DMU is calculated under its own optimal weights as well as the optimal weights calculated for every other DMU. The cross-efficiency score for a given unit is the average over all such evaluations and reflects performance under a peer-consensus weighting rather than under self-serving weights. Because the consensus weights rarely coincide with any single DMU's preferred weights, the cross-efficiency score produces a complete ordering even among units that were tied on the original frontier, and guards against the criticism that standard DEA permits each unit to choose weights favourable to itself.

Andersen and Petersen [1] proposed **super-efficiency**, an alternative procedure that handles the same problem from a different perspective. Their approach evaluates each frontier DMU after removing it from its own reference set, so that the unit is compared against the frontier that would exist in its absence. Efficient DMUs may then receive scores greater than one, with the magnitude of the excess indicating how far the unit could deteriorate before being overtaken by its peers. Like cross-efficiency, super-efficiency yields a strict ranking among frontier units, but it does so through a unit-by-unit re-evaluation rather than through a consensus-weight average.

Either refinement may be applied on top of a BCC frontier to obtain a discriminating ranking; both represent natural extensions of the framework developed in this study, and their implications are discussed in Chapter 6.

2.2 Applications of DEA in Sports Analytics

DEA entered the sports-analytics literature in the early 2000s. Espitia-Escuer and García-Cebrián [8] applied DEA to evaluate the efficiency of teams in the Spanish La Liga competing

in European tournaments. Treating each club as a DMU and using inputs such as squad value and outputs such as match performance indicators, they observed a striking pattern: clubs operating on modest budgets were frequently more efficient than the wealthier giants of the league. The finding indicated that efficiency, in the technical sense of producing maximum output per unit of input, is not necessarily aligned with absolute success — a distinction between *productivity* and *outcome* central to the present work.

Tiedemann, Francksen, and Latacz-Lohmann [15] extended the application of DEA from team level to player level, evaluating performers in the German Bundesliga through a **metafrontier framework** that explicitly accounted for positional heterogeneity. Their central methodological insight was that goalkeepers, defenders, midfielders, and forwards should not be evaluated against a common frontier because the inputs and outputs that characterise each position differ in kind. Like for a goalkeeper, the basic outputs are saves and clean sheets; for a forward, the outputs are goals, shots and shots on the goal. Comparing them under a common set of weights surely miscalculate both results. Tiedemann et al. therefore constructed separate frontiers for each positional group and then, through the metafrontier technology, related these to a common reference. Their conclusions reinforced the Espitia-Escuer finding at a finer granularity: the most efficient players were often not the most expensive, and identifying them required position-specific evaluation. The methodological decision to compute separate DEA frontiers per position, adopted in the present study, is directly motivated by this work.

2.3 Optimisation Approaches to Team Selection

Parallel to the DEA literature, a separate stream of operations-research scholarship has treated team selection as a combinatorial-optimisation problem. The earliest formal treatment in a football context is due to Boon and Sierksma [3], who modelled the construction of a squad as a **knapsack problem**: each candidate player has a known cost and a known performance score, and the manager must select a subset whose total cost lies within a budget and whose total performance is maximised. The model was limited by its single-objective single-constraint structure: football squads are subject to positional requirements that a basic knapsack cannot accommodate, and the performance score was a single subjective figure rather than a quantity derived from rigorous statistical analysis.

2.4 Research Gap and Contribution

The literature reviewed above falls into two strands that have, to the best of our knowledge, never been formally integrated in the context of English Premier League squad construction.

4 DEA has been used to evaluate the efficiency of football teams and individual players, but the resulting efficiency scores have not been propagated into an optimisation model for the assembly of an actual starting eleven. Conversely, the optimisation-based approach of Boon and Sierksma [3] accepted per-player performance scores as exogenous inputs, leaving the question of where those scores should come from to the discretion of the analyst.

Hence, our study is trying to reduce this gap by using the output of one model as the input of the next. Player efficiency is first measured through a position-specific BCC DEA model, following the methodological lead of Tiedemann et al. [15]. The resulting efficiency scores are then embedded directly into the objective function of an ILP that selects the optimal starting eleven, in the tradition of Boon and Sierksma [3], but with one important difference: the coefficients of the objective function are no longer subjectively chosen performance estimates but DEA-derived measures of relative efficiency. The two stages are methodologically distinct yet operationally coupled.

The specific contributions of this work are summarised as follows:

1. A two-stage framework that integrates BCC-DEA efficiency evaluation with ILP-based team selection in a single end-to-end pipeline for English Premier League squad construction.
2. The application of position-specific efficiency frontiers, following Tiedemann et al. [15], to ensure homogeneity among the decision-making units within each positional cohort and to avoid the distortion that arises from comparing dissimilar roles on a common set of metrics.
3. An ILP formulation that admits realistic operational constraints — squad size, positional requirements for multiple formations (4-3-3, 4-4-2, 5-4-1), and a two-sided budget envelope with adjustable lower and upper bounds — and that is solved exactly through Branch and Bound.
4. A working software implementation of the framework, encoded as a browser-based interactive tool that permits real-time exploration of the trade-offs between budget, formation, and team efficiency, and that operationalises the framework for practical use.

The result is, we believe, the first published study to integrate DEA-based player evaluation with ILP-based team selection in the EPL setting, and a framework of general applicability to any league in which suitable data on player wages and on-field performance is available.

CHAPTER 3

DATA ENVELOPMENT ANALYSIS

Data Envelopment Analysis (DEA) is a non parametric mathematical programming technique. It is used for measuring the relative efficiency of a set of comparable decision-making units (DMUs) using multiple inputs and multiple outputs. It was formulated by Charnes, Cooper, and Rhodes in 1978 [5], the technique has been applied across an exceptionally wide range of domains such as banking, healthcare, education, public administration, and, more recently, sports analytics. This chapter helps to understand the theoretical foundations of DEA in the level of detail required to understand the analyses reported in subsequent chapters.

3.1 Overview and Motivation

We know that when we have to calculate the efficiency of any machine or system then we divide the output by input. But when we have multiple inputs and multiple outputs, what can be the proper way to calculate the efficiency. Likewise when a player has various performing data track corresponding to different skills and ability then how will we find the efficiency of that player? How would the salary , minutes played , age , etc. would be combined in a single composite unit?

Conventional techniques either use subjective weight distribution or use the parametric way to assign weight to the data. DEA doesn't use either. It allows each DMU choose the weights which is most suitable to itself. It makes sure that no DMU including the one being evaluated may get an efficiency score greater than unity under the same weights. The DMUs that has a score of one form the empirical efficiency frontier; the remainder are inefficient relative to that frontier. The DEA is non parametric because it does not assumes a function rather it directly takes the data in use.

3.2 The CCR Model

Consider n DMUs, indexed by j = 1, ..., n. Let the number of inputs be m and number of ouputs be s. x_{ij} denotes the level of input i used by DMU j and y_{rj} the level of output r produced by DMU j.The efficiency Θ of that DMU is according to the Charnes Cooper Rhodes (CCR)

model as the ratio of weighted sum of outputs to the weighted sum of the inputs . Mathematically,

$$\theta_o = (\Sigma_{r=1}^{s} u_r y_{ro}) / (\Sigma_{i=1}^{m} v_i x_{io}) \qquad (3.1)$$

where $u_r \geq 0$ and $v_i \geq 0$ are the output and input weights respectively.The weights are not fixed in advance, they are assigned in a manner that the efficiency is maximized and the efficiency for other DMUs don't exceed one.

$$(\Sigma_r u_r y_{rj}) / (\Sigma_i v_i x_{ij}) \leq 1 \quad \forall j \qquad (3.2)$$

To solve this fractional programme, we convert it into a standard linear problem by setting the denominator to unity and focusing on maximizing the numerator. The resulting CCR model works under the assumption of constant returns to the scale where we increase the input the output also increases and vice versa (in a proportional ratio). Importantly, the CRS framework gives every DMU as if it is functioning at its absolute best scale and hence this a very rigid condition.

3.3 The BCC Model

Banker , Charles and Copper then introduced the Variable return to scale model which is flexible than the CRR model. In this they introduces a scaler w in the numerator:

$$\theta_o = (\Sigma_r u_r y_{ro} + w) / (\Sigma_i v_i x_{io}) \qquad (3.3)$$

In the dual envelopment form of the linear program, the BCC model adds up the convexity constraint $\Sigma_j \lambda_j = 1$, where λ_j are the reference weights placed on each DMU. This restriction has geometric consequences which means the efficient frontier is no longer the cone generated only by the most productive DMUs, but the convex hull of the observed input–output combinations. Due to this the DMUs which were not in action in the CCR model are now operating efficiently where they are considered in their scale of measurement.

The difference between both the model shows up that the BCC model is more defensible choice. The BCC score isolates pure technical efficiency which controls the scale. The ratio of the two scored gives a separate measure of scale efficiency. This indicates how far the DMUs observed scale departs from the most productive scale size.

3.4 Frontier Geometry and Efficiency Interpretation

DEA produces a linear efficiency frontier which enveloped all the DMUs. The DMU which lies on this envelope is termed as efficient which lied inside the envelope and is less than 1 is said inefficient. Let us say if a unit is scoring 0.7 then it is only using 70% efficiency while 30 % of the unit resources are getting wasted. So if it could learn from the top performing units(peers) it could reduce all its inputs to 70% of what it currently uses and still produce the same exact output.

The frontier is made by the dual envelopment formulation of the DEA program. A non-negative weight λ_j is associated with each DMU j and the projection of an evaluated DMU onto the frontier is expressed as a convex combination $\Sigma_j \lambda_j (x_j, y_j)$ of the observed input–output bundles. The set of DMUs with strictly positive λ_j in the optimal solution constitutes the **peer reference set** for the evaluated unit, identifying which efficient DMUs serve as natural benchmarks for improvement. Slack variables s^- and s^+ capture any non-radial improvements that remain after the proportional reduction in inputs has been applied. A DMU is fully efficient only when both $\theta = 1$ and all slacks are zero. In practice the peer reference set provides far more managerial insight than the scalar efficiency score alone, as it identifies specific area against which the DMU should specify itself to reach the benchmark.

3.5 Input Versus Output Orientation

DEA admits two equivalent formulations differentiated by their direction of optimisation. The **input-oriented** model keeps the outputs fixed and asks by how much the inputs could be proportionally reduced maintaining the same outputs. This orientation is appropriate where the DMU exercises control over its inputs but accepts demand or supply for its outputs as exogenous. The **output oriented model** keeps the input fixed says that how much the output can be maximized by the same input. This is the natural choice where the DMU is judged on what it produces on given fixed resources. Under CRS the two formulations gives identical efficiency scores. Under VRS they differ slightly, and the choice should reflect the substantive question being asked.

3.6 Worked Numerical Example

To demonstrate how Data Envelopment Analysis (DEA) processes data, let us evaluate a scenario involving five Decision Making Units (DMUs). In this simplified setup, each DMU utilizes a single input—defined here as salary in arbitrary units—to generate a single output,

which is the number of goals scored. The baseline data and computed efficiency scores are detailed in Table 3.1 below

DMU	Input x	Output y	Ratio y/x	CCR Score θ
A	2	1	0.50	0.50
B	3	3	1.00	1.00
C	5	4	0.80	0.80
D	6	5	0.83	0.83
E	8	6	0.75	0.75

Table 3.1: *A five-DMU example with one input and one output.*

In a basic single-input and single-output framework, the efficiency analysis scales down to evaluating the output-to-input ratios for each unit. We can see that DMU B achieves the highest productivity ratio of 1.00, which automatically sets it as the benchmark that defines the empirical efficiency frontier. Under the constant returns to scale (CCR) framework, every other unit is measured against DMU B. Consequently, a unit's CCR efficiency score is simply its individual ratio divided by DMU B's maximum ratio, making the final efficiency scores identical to the raw productivity ratios. While DMU B emerges as fully efficient, the remaining four units fall short. Theoretically, these inefficient units could scale down their inputs to CCR score × 100% of their current usage without experiencing any reduction in the output.

Take DMU C as a case in point. Operating with a CCR score of CCR score= 0.80, it would need to optimize its operations by cutting down its salary input from 5 units to 4 units while maintaining its output of 4 goals to be deemed CCR-efficient. Alternatively, from an output-oriented perspective, it would need to increase its production from 4 goals to 5 goals while keeping its current salary fixed at 5 units. Because DMU B is the sole anchor defining the frontier ray in this single-variable space, it serves as the exclusive peer reference unit for DMU C.

However, if we use the BCC model it fundamentally alters the shape of this frontier. It intorduces the convexity constraint($\Sigma_j \lambda_j = 1$), the frontier is pulled tighter around the data,

turning from a straight ray originating from the zero point into a segmented, piecewise-linear boundary. Interestingly, this shift reclassifies DMU A—the smallest unit in our sample—as fully efficient at its specific operational scale, despite holding the lowest overall productivity ratio. The newly formed BCC frontier stretches across the upper boundary of the data points, keeping DMU B efficient while assigning higher variable returns to scale (VRS) scores to the intermediate units (C, D, and E) compared to their stricter CCR metrics. This adjustment highlights the core purpose of the VRS model i.e. it helps to save the smaller units from being penalised simply due to their scale of operation, allowing us to successfully isolate pure technical flaws from scale-driven inefficiencies.

This single dimensional geometry is quite easy to understand but when it gets to the multi dimensional work it get complex , such as the football team efficiency model analyzed later in this study. In those higher-dimensional spaces, the linear programming equations detailed in Sections 3.2 and 3.3 are processed using specialized **DEA software**. Yet, the logic behind this evaluation remains identical. The model evaluates each individual player under the most favourable weights possible, generates a definitive efficiency score, and identifies a specific peer reference set of frontier players to serve as real-world benchmarks for improvement.

CHAPTER 4

INTEGER LINEAR PROGRAMMING

Integer Linear Programming (ILP) is a mathematical technique or tool which is used for optimisation. In this branch, we are dealing with problems in which, under some linear constraints, a linear objective function is maximised or minimised over a feasible region. But in ILP, the additional requirement is all of the decision variables are only in the form of integer values. Now, this chapter develops the theoretical foundations of ILP in detail to understand the squad-selection model which we have used in our Methodology. Here, we are going to understand the Branch-and-Bound method and using it how solve our research problem.

4.1 From Linear Programming to Integer Linear Programming

A standard **linear program (LP)** seeks values of n continuous decision variables x_1, x_2, ..., x_n that maximise (or minimise) a linear objective function subject to a finite set of linear inequality and equality constraints:

$$\textit{Maximise} \quad c^T x \quad \textit{subject to} \quad A x \leq b, \quad x \geq 0 \qquad (4.1)$$

Linear programs of this form are computationally well understood. The simplex method of Dantzig [6] and the interior-point methods of Karmarkar [11] solve LP instances with millions of variables and constraints in polynomial time on modern hardware. The theory is mature and the practical implementations are robust.

An **integer linear program (ILP)** imposes the additional restriction that some or all of the decision variables take integer values:

$$\textit{Maximise} \quad c^T x \quad \textit{subject to} \quad A x \leq b, \quad x \in \mathbb{Z}^n \qquad (4.2)$$

When all variables are required to be integer, the problem is called a **pure ILP**; when only some are, it is a **mixed integer linear program (MILP)**. The simple-looking addition of the integer solution requirement transforms the problem solely. ILP is in general **NP-hard**: no polynomial-time algorithm is known to solve arbitrary instances, and the existence of one is widely believed to be equivalent to the resolution of the P versus NP question. So we can say that, when we are solving an ILP problem, the number of integer variables increases as well as the time required to solve the problem also grows rapidly.

4.2 Binary Variables and Combinatorial Problems

A particularly important special case is the **binary integer program**, in which the integer variables are further restricted to take only the values zero or one:

$$x_j \in \{0, 1\} \quad for\ j = 1, \ldots, n \qquad (4.3)$$

Binary variables provide a natural mathematical encoding for combinatorial decisions of the form *"include this item, or do not"*. Many classical problems in operations research admit such a representation, including the knapsack problem (which items to pack into a bag of limited capacity), the assignment problem (which workers to assign to which tasks), the set-covering problem (which subsets to select so as to cover a given universe), and a wide family of scheduling, routing, and selection problems. The squad-construction problem developed in this study falls squarely within this family. Each player is associated with a binary variable that takes the value one if the player is selected for the starting eleven and zero otherwise.

The combinatorial nature of binary problems can be appreciated by direct enumeration: with n binary variables there are 2^n candidate solutions, and even for modest n this number is astronomical. For the present study, with n = 324 candidate players, the unrestricted enumeration would require examination of $2^{324} \approx 3.4 \times 10^{97}$ subsets — a quantity vastly exceeding the number of atoms in the observable universe. Practical ILP algorithms therefore avoid explicit enumeration and rely instead on intelligent search strategies.

4.3 The LP Relaxation

The central idea exploited by virtually every modern ILP algorithm is the **LP relaxation**: the linear program obtained from an ILP by dropping the integrality requirement and allowing the decision variables to take continuous values within their declared bounds. In a binary problem, the relaxation changes $x_j \in \{0, 1\}$ with $0 \leq x_j \leq 1$.

The relaxation has three properties of fundamental importance. First, it is a linear program and can therefore be solved efficiently. Second, in optimization, the feasible region of an Integer Linear Program (ILP) is a strict subset of its continuous Linear Programming (LP) relaxation. Because the LP relaxation exclude the integer constraints and Includes all the valid integer solutions also the fractional ones. The relaxation therefore provides an **upper bound** (for a maximisation problem) on the integer optimum. Third, if the relaxation solution happens to be integer-valued, then it is also the ILP optimum and no further work is required. This last

circumstance arises frequently when the constraint matrix possesses certain structural properties, such as total unimodularity, but cannot be guaranteed in general.

4.4 The Branch-and-Bound Algorithm

When the LP relaxation does not yield an integer solution, the **Branch-and-Bound** algorithm [12] is the workhorse method for finding the true ILP optimum. In this method, first of all we divide the feasible region into smaller subproblems, solving the LP relaxation of each, and cancelling those subproblems whose solutions are not integer value. The procedure can be summarised in four operations:

• **Relax:** solve the LP relaxation of the current subproblem.

• **Branch:** if the relaxation solution is fractional, choose a fractional variable x_j *and create two new subproblems — one with* $x_j \leq \lfloor x_j \rfloor$ and one with $x_j \geq \lceil x_j^* \rceil$ — partitioning the feasible region.

• **Bound:** compare each subproblem's relaxation value with the best integer solution found so far (the **incumbent**).

• **Prune:** discard subproblems whose relaxation bound is no better than the incumbent (they cannot improve on it), whose LP relaxation is infeasible, or whose relaxation solution is already integer-valued (in which case update the incumbent).

The method creates a tree-type structure whose root is the original ILP. Each node is itself a subproblem; the tree is expanded by branching and contracted by cutting the branches. The search ends when all nodes are calculated, at which point the incumbent is optimal. In the worst case the tree size is exponential in the number of integer variables, but in practice a well-implemented Branch-and-Bound prunes the great majority of nodes early and terminates orders of magnitude faster than naive enumeration.

4.5 Branch and Cut

Modern ILP solvers augment the basic Branch-and-Bound skeleton with **cutting planes** — additional linear inequalities that are valid for the integer feasible region but cut off fractional vertices of the LP relaxation. The intuition is simple: every cutting plane added to the relaxation compact it, reducing the relaxation bound and bringing it closer to the integer solution. A tighter bound prunes more of the search tree, often dramatically. Several classes of cuts have been developed for general-purpose use, including Gomory mixed-integer cuts, cover inequalities

for knapsack-type constraints, and clique inequalities derived from the conflict graph of the binary variables.

The combination of Branch-and-Bound with cutting planes is known as **Branch and Cut**, and is the algorithmic basis of contemporary commercial and open-source solvers including CPLEX, Gurobi, CBC, and the javascript-lp-solver library used in the present study. The practical effect is that ILP instances that would have been computationally intractable in the 1970s are now routinely solved to optimality. For the problem sizes encountered in this work — a few hundred binary variables and a small number of linear constraints — Branch and Cut reliably produces optimal solutions within milliseconds on commodity hardware.

4.6 Worked Numerical Example

To illustrate the Branch-and-Bound procedure, consider the following small binary integer program with three decision variables, intended as an abstraction of the squad-selection problem at a miniature scale:

$$\text{Maximise} \quad 5x_1 + 4x_2 + 3x_3$$

$$\text{subject to} \quad 2x_1 + 3x_2 + x_3 \le 4, \quad x_j \in \{0, 1\}$$

The LP relaxation, obtained by replacing the binary requirement with $0 \le x_j \le 1$, has optimal solution $x_1 = 1$, $x_2 = 2/3$, $x_3 = 1$, with objective value $5 + 8/3 + 3 = 32/3 \approx 10.67$. The relaxation is fractional in x_2, so the algorithm branches on that variable, creating two subproblems: one with $x_2 = 0$ and one with $x_2 = 1$.

In the subproblem $x_2 = 0$, the relaxation yields $x_1 = 1$, $x_3 = 1$, with objective value 8 — and this is integer-feasible, so it becomes the incumbent. In the subproblem $x_2 = 1$, the relaxation forces $x_1 = 1/2$ (since $3 + 2x_1 + x_3 \le 4$), with objective value $5(1/2) + 4 + 1 = 7.5$. Branching further on x_1 produces two sub-subproblems whose bounds are both less than the incumbent value of 8; both are pruned. The algorithm terminates with the optimal integer solution $x_1 = 1$, $x_2 = 0$, $x_3 = 1$ and objective value 8. The full search tree explored only three LP relaxations rather than the eight subsets that explicit enumeration would have required.

The squad-construction model developed in this study has the same structural form but with $n = 324$ binary variables and a richer constraint set. The Branch-and-Bound machinery described above scales directly to this larger setting, and, as reported in subsequent chapters, produces optimal starting elevens within milliseconds on a standard browser-based implementation.

CHAPTER 5

METHODOLOGY

This chapter describes the method to built the efficient 11 player team which is preceded by the theory in chapter 3 and 4. The method follows the collection of data , refining of data, application of DEA, ILP implementation, experimental design using the evaluation protocols.

5.1 Pipeline Overview

The methodology comprises five principal stages, depicted schematically in Figure 5.1:

1. Data collection and preprocessing from public sources.
2. Position-specific efficiency evaluation using the BCC model of DEA.
3. Formulation of squad selection as a binary Integer Linear Program.
4. Solution via Branch and Cut, implemented in a browser-based application.
5. Evaluation of the resulting squads against the metrics defined in §5.9.

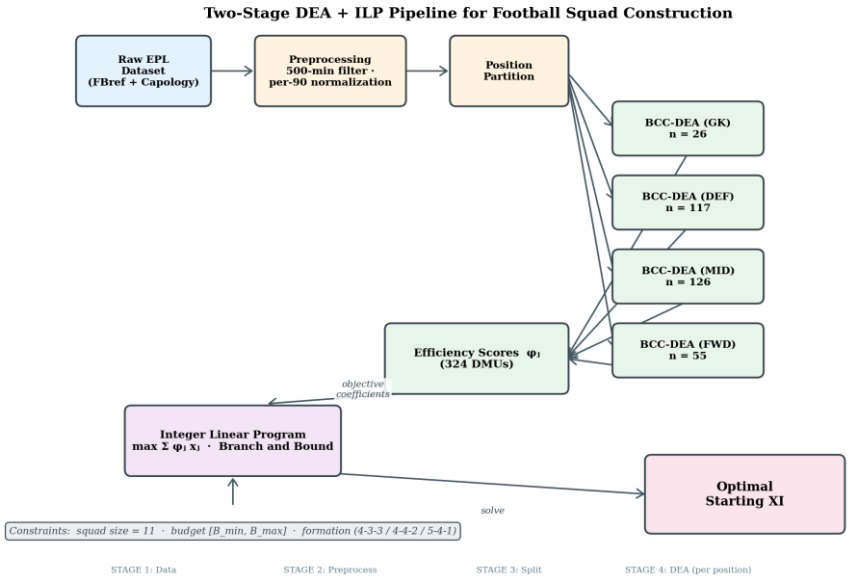

Figure 5.1: *Two-stage DEA + ILP method for football squad construction. The normalised data flows through preprocessing and is partitioned by position; separate BCC-DEA models compute efficiency scores that serve as coefficients in a single Integer Linear Program for the final selection.*

The pipeline is a two-stage method. The DEA method is particularly being used for the efficiency evaluation where each player is considered as the decision making unit(DMU), a single scalar score reflecting the player's output relative to peers operating at comparable input levels. The ILP stage is concerned exclusively with selection and treats those scores as fixed coefficients of an optimisation problem subject to the operational constraints of squad construction. The clean separation between measurement and selection makes the pipeline modular. Any stage can be used without disturbing the other

5.2 Data Sources

The dataset assembled for this study was compiled from two complementary public sources. **FBref** (fbref.com) [9], maintained by Sports Reference LLC, provides per-player match-event statistics for every major football league. This platform has raw event data licensed from Opta and StatsBomb into structured tables covering minutes played, appearances, goals, assists, shots, passes, tackles, interceptions, clearances, blocks, and a wide array of related metrics. **Capology** (capology.com) [4] publishes annual salary information for professional football leagues, derived from a combination of public filings, contract leaks, and industry estimates. Although salary data in football is intrinsically less reliable than performance data .In general clubs do not publish the exact salay of the player. Hence, Capology's figures are widely cited in the analytics community and represent the best available public estimate.

Data from both the sources was taken and then merged in front of the player's identity. Player-name matching was performed manually for cases in which the two sources used different spellings or transliterations of the same name, and a small number of players whose salaries could not be reliably matched were excluded from the analysis, the salary of the players are given in USd dollars. the position recorded by FBref, the total minutes played, the number of appearances, and all per-90-minute performance metrics relevant to the player's position.

5.3 Data Preprocessing

Three steps were done in data preprocessing taking in the consideration of evaluating the proper efficiency of the players.

Minimum playing-time threshold. Players who played fewer than 500 minutes of competitive playing time during the season were removed from the dataset. The threshold is

decided to mitigate small-sample bias. A player who plays only 1 game and scores 2 goals in 40 minutes would leave the established forwards. This would destroy the efficiency calculation. So we are considering only those players who have played atleast 5 and a half match that 500 minutes.

Per-90-minute normalisation. All performance metrics were standardised to a per-90 basis by dividing each raw total by the player's minutes played and multiplying by ninety. This conversion expresses each metric as the rate at which it would accumulate over a full match, and makes players who played different total minutes directly comparable. For example, a midfielder who scores 50 tackles in 1800 minutes has 2.5 tackles per 90; a forward who scored 18 goals in 2700 minutes has 0.6 goals per 90. Without normalisation, players who simply played more minutes would appear systematically more productive than those who played fewer, conflating availability with efficiency.

35 **Positional partition.** Players were grouped into four positional departments- Goalkeepers (GK), Defenders (DEF), Midfielders (MID), and Forwards (FWD). Using the primary position recorded by FBref. players who appeared in multiple positions during the season were assigned to the position in which they accumulated the most minutes. The partition is important because the inputs and outputs that characterises each role differ in roles. Evaluating a goalkeeper and a forward against a common set of weights would distort both evaluations beyond use. After preprocessing, the dataset comprises **26 Goalkeepers, 117 Defenders, 126 Midfielders, and 55 Forwards**, for a total of **324 decision-making units** distributed across four separate DEA analyses.

5.4 Application of DEA: Inputs, Outputs, and Orientation

Within each positional cohort, the BCC model described in §3.3 is applied to compute a relative efficiency score for every player. Three modelling choices required justification: the selection of inputs, the selection of outputs, and the choice of orientation.

Inputs. Four inputs are used uniformly across all four cohorts: **salary** (annual wage, in USD), **age** (in years at season start), **minutes played**, and **appearances**. Salary is the principal economic resource committed to the player and is the central object of efficient allocation. Age is included to control for the developmental and physical capital invested in the player, recognising that younger players may produce identical outputs at lower salary cost partly because their contractual position is weaker; this should not be confused with greater efficiency.

Minutes played and appearances jointly capture the opportunity afforded to the player: a player whose manager affords him 3000 minutes is granted more chances to accumulate outputs than one limited to 800 minutes, and the comparison must control for this.

Outputs. The outputs are position-specific, reflecting the central responsibilities of each role:

• **Goalkeepers:** successful crosses claimed, clean sheets, clearances, and (reverse-signed) goals conceded.

• **Defenders:** successful passes, successful crosses, possession won, clearances, interceptions, blocks, and tackles.

• **Midfielders:** goals, assists, shots on target, touches, passes, successful passes, successful crosses, and possession won.

• **Forwards:** goals, assists, and shots on target.

The output set for forwards is deliberately narrow. Forwards are evaluated on attacking production, and the inclusion of metrics such as tackles or interceptions would penalise players whose role does not require defensive contribution. The output set for midfielders is the broadest, reflecting the hybrid attacking-and-creative nature of modern midfield roles. Goals conceded enters the goalkeeper specification with a reversed sign, since a lower value indicates better performance; DEA software accepts this through the "undesirable output" specification.

Orientation. The output-oriented BCC model is adopted. The output orientation asks "*given this fixed inputs how much a players efficiency can be improved to reach the efficiency frontier or how much the output is expected from this particular input?"* This is the natural question in the squad-selection context, where the club has already committed to the player's salary and granted him playing time, and is concerned with the value it extracts from that investment. The input-oriented alternative — by how much could inputs be reduced for the same outputs. It is the natural question in cost-cutting contexts and is less apt here.

The BCC computation is performed using **MaxDEA Lite** independently for each of the four positional cohorts, producing four separate frontiers and four collections of efficiency scores. The variable-returns-to-scale assumption is justified on substantive grounds: as argued in §3.3, the assumption of constant returns is not suitable for football, where a doubling of salary does not double output. The convexity constraint of the BCC model allows the smallest-input players to be efficient at their own scale rather than being penalised relative to a frontier

drawn through the most-productive observation, which is the correct treatment for a setting in which scale itself is a contractual outcome rather than a managerial choice.

5.5 Integer Linear Programming Formulation

The efficiency scores produced by the four BCC analyses are aggregated into a single vector of length n = 324 and used as the objective-function coefficients of an Integer Linear Program. For each player j = 1, ..., n, define the binary decision variable:

$$x_j = 1 \text{ if player } j \text{ is selected for the starting eleven, } 0 \text{ otherwise.} \qquad (5.1)$$

The objective is to maximise the average efficiency of the selected squad:

$$\text{Maximise} \quad Z = (1/11)\, \Sigma_{j=1}^{n} \varphi_j x_j \qquad (5.2)$$

where $\varphi_j \in (0, 1]$ is the BCC efficiency score of player j. The division by eleven is purely cosmetic — the optimum is unaffected — and serves to interpret the objective value as the average efficiency of the chosen eleven. The model is solved subject to four families of constraints.

Squad size. Exactly eleven players are selected:

$$\Sigma_{j=1}^{n} x_j = 11 \qquad (5.3)$$

Budget envelope. The total wage cost of the selected eleven must lie within bounds:

$$B_{min} \leq \Sigma_{j=1}^{n} c_j x_j \leq B_{max} \qquad (5.4)$$

where c_j denotes the annual wage of player j. The lower bound captures the practical requirement that a club competing for elite status must commit a minimum salary outlay; the upper bound reflects Financial Fair Play limits and internal salary-cap considerations.

Formation constraints. The positional counts depend on the chosen formation. For 4-3-3:

$$\Sigma x_{sk} = 1, \quad \Sigma x\_def = 4, \quad \Sigma x\_mid = 3, \quad \Sigma x\_fwd = 3 \qquad (5.5)$$

analogous constraints hold for 4-4-2 (4 defenders, 4 midfielders, 2 forwards), 5-4-1 (5 defenders, 4 midfielders, 1 forward), and 4-2-3-1 (4 defenders, 5 midfielders, 1 forward). The 4-2-3-1 formation is mathematically indistinguishable from a 4-5-1 in the present model because the FBref dataset does not subdivide the midfield into defensive and attacking roles; the tactical distinction between the two holding midfielders and the three attacking midfielders is realised through managerial instruction rather than through the selection mathematics.

User constraints. A subset of players may be forced into or out of the selection:

$$x_j = 1 \quad \forall j \in L \quad and \quad x_j = 0 \quad \forall j \in E \qquad (5.6)$$

where L denotes the locked set (players who must appear) and E the excluded set (players who must not). These constraints permit the model to be conditioned on prior managerial choices: a club committed to a particular goalkeeper, for example, may lock that goalkeeper and have the model select the remaining ten optimally. This is done in case if a club wants a player for sure due to the business reasons.

The resulting model has 324 binary variables and on the order of a dozen linear constraints. It is solved exactly via **the Branch-and-Cut** procedure described in §4.5.

5.6 Software Implementation

The framework has been implemented as a browser-based interactive application. The choice of browser environment was deliberate: it makes the tool zero-install, platform-independent, and immediately shareable. The DEA stage is performed in **MaxDEA Lite** on the four positional cohorts independently, and the resulting efficiency scores are exported as a CSV file that the application ingests. The ILP stage is encoded in JavaScript and solved in the browser using the **javascript-lp-solver** library (Schauwecker, 2014), an open-source implementation of Branch and Bound with cutting-plane augmentation. The application is built as a single self-contained HTML file with no external runtime dependencies.

Figure 5.2 shows the Player Library view of the application. The 324 players are listed in four tabs (GK / DEF / MID / FWD), each displaying the player name, position, BCC efficiency score φ, and annual salary. A toggle in the upper right permits the user to switch among three scoring modes (MaxDEA output-BCC, Input-BCC, and Cross-Efficiency); the ILP objective is recomputed under the chosen mode. The lock and exclude buttons adjacent to each player apply the user-constraint sets L and E defined in equation (5.6).

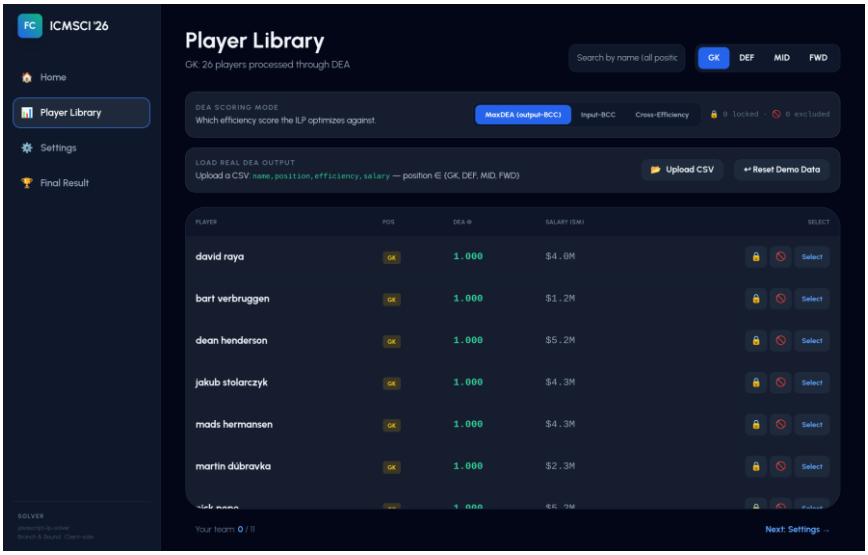

Figure 5.2: *Player Library view of the application, showing the GK tab with the BCC efficiency score φ, salary, and the lock / exclude buttons that operationalise the constraints in equation (5.6). The scoring-mode toggle permits the ILP to be re-solved against three alternative efficiency formulations.*

Figure 5.3 shows the result view after a representative optimisation. The configuration is a 4-3-3 formation with budget envelope [$30M, $100M] and $\varepsilon = 0.02$ (see §5.7 below); the model solved in 65 ms and selected an XI with average efficiency $\varphi = 1.000$ and total wage cost of $32.6M, with one goalkeeper (Bart Verbruggen), four defenders, three midfielders, and three forwards. A diagnostics panel below the pitch reports the number of binary variables, the number of constraints, and other internal solver statistics.

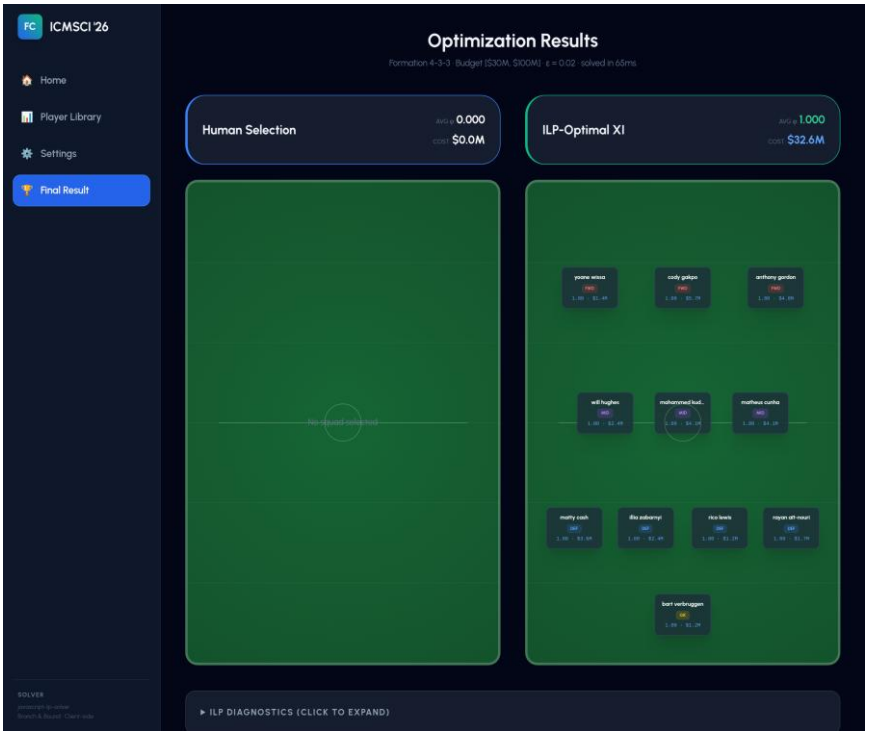

Figure 5.3: *Result view of the application showing the ILP-optimal starting eleven for the 4-3-3 formation with budget [$30M, $100M]. The selected XI achieves average efficiency φ = 1.000 at total cost $32.6M in a solve time of 65 ms.*

5.7 Implementation Details

Three implementation details merit explicit discussion. The first is a **Pareto-dominance pre-filter** applied to the candidate pool before the ILP is constructed. Within each positional cohort, a player j is said to be strictly dominated by another player k if $\varphi_k \geq \varphi_j$, $c_k \leq c_j$, and at least one of these inequalities is strict. A dominated player cannot appear in any optimal solution: if such a player j were in the optimum, replacing him with the dominating player k would yield a solution with at least as high an objective and at most the same cost, and would still satisfy all formation constraints, contradicting the optimality of j. Removing dominated players therefore preserves the optimum exactly while reducing the number of binary variables, and contributes substantially to the millisecond solve times observed in practice.

The second detail is a **budget-relaxation parameter ε**, exposed to the user and set to a small positive value (typically 0.01 or 0.02). The relaxation inflates the upper budget bound and deflates the lower bound by a fraction ε: the effective constraint becomes $(1 - \varepsilon)\, B_{min} \leq \Sigma\, c_j\, x_j \leq (1 + \varepsilon)\, B_{max}$. The parameter is a practical concession to the integer nature of the problem: a strict specification can render the model infeasible when there is no integer combination of

eleven players whose salaries fall exactly within the prescribed window, and ε permits a controlled relaxation to recover feasibility without sacrificing meaningful budget discipline.

The third detail is the handling of **infeasibility**. When the user-supplied combination of formation, budget envelope, and lock / exclude sets has no feasible solution, the application reports this explicitly rather than failing silently or returning a partial solution. Common causes are an over-constrained budget (too tight on either bound), or a locked set incompatible with the chosen formation.

5.8 Algorithm

The complete two-stage procedure is summarised in the pseudocode below.

```
Input:  Player dataset P; formation F; budget [B_min, B_max]; ε;
        scoring mode m; lock set L; exclude set E.
Output: Optimal starting eleven T*.
Stage 1 — Preprocessing
filter P by minutes-played ≥ 500
normalise all performance metrics to per-90 basis
partition P by position: P_GK, P_DEF, P_MID, P_FWD
Stage 2 — DEA
for each cohort C ∈ {P_GK, P_DEF, P_MID, P_FWD}:
   solve output-oriented BCC model on C (mode m) → ϕ_j
Stage 3 — ILP setup
apply Pareto-dominance pre-filter within each cohort
build binary variables x_j; objective max (1/11) Σ ϕ_j x_j
add constraints (5.3)–(5.6) using F, [B_min, B_max], ε, L, E
Stage 4 — Solve & report
solve via Branch and Cut (javascript-lp-solver)
return T*, total cost, average ϕ, solve time
```

5.9 Evaluation Methodology

The framework is assessed through five quantitative measures and one external benchmark. Each is defined precisely below to make the results reported in Chapter 6 unambiguously reproducible.

Average team efficiency Z. The value of the ILP objective function (5.2) at the optimum, equal to the mean BCC efficiency score of the selected XI. This is the primary measure and the quantity the optimisation maximises by construction. Higher values indicate squads composed of players closer to their respective positional frontiers.

Total squad salary. The aggregate annual wage of the eleven selected players, $\Sigma_j c_j x_j^*$, reported in millions of dollars. This is checked against the prescribed budget envelope as a sanity test, and its position within the envelope indicates how tightly the model is binding against the constraint.

Selection-frequency robustness. To assess the stability of the final optimal lineup, a selection-frequency robustness test was conducted to evaluate how input fluctuations impact the selected XI. For each scenario, the ILP model was executed 100 times under independent Gaussian perturbations, adjusting the original efficiency scores using the formula $\emptyset'_j = \emptyset_j + N(0, \sigma^2)$ with a standard deviation of $\sigma = 0.05$. We then tracked the frequency with which each originally selected player was retained across these 100 simulated runs. A player who maintained their position in 80% or more of the perturbed trials was classified as a robust choice. Conversely, lower retention frequencies indicated that a player's inclusion was highly sensitive to minor variations in the underlying efficiency estimates.

Budget shadow-price evaluation.. To evaluate the economic sensitivity of the financial limits, a budget shadow-price evaluation was conducted to estimate the marginal value of the budget constraint. Because traditional shadow prices cannot be computed directly in an Integer Linear Programming (ILP) framework due to the discrete nature of variables, the model was systematically re-solved across a pre-defined range of budget perturbations, denoted as $\Delta \in \{-20, -10, -5, 0, +5, +10, +20\}$ million. By tracking the subsequent fluctuations in the objective function (Z), we mapped a sensitivity curve. This curve implicitly estimates the shadow price of the budget, effectively illustrating the rate at which overall team efficiency scales relative to incremental changes in salary expenditure

34

Solve Time: This metric measures the total wall-clock execution time elapsed from the moment the user triggers the optimization process to when the final dataset is ready for interface presentation. Tracked precisely via the browser's native performance API in milliseconds, this duration accounts for the mathematical heavy-lifting, including Pareto-dominance data filtering, ILP structural setup, and the final Branch-and-Cut algorithmic execution. However, it explicitly excludes the frontend interface rendering time to isolate computational efficiency.

Benchmark Overlap: To validate the real-world accuracy of the framework, this metric quantifies the intersection between the model's optimized lineup and an established expert consensus baseline, specifically the PFA Premier League Team of the Year. Reported as a direct player match-count out of eleven, this serves as a critical validation check. A high overlap confirms that the DEA-ILP structure aligns closely with professional domain expertise..

5.10 Experimental Design

Chapter 6 reports the application of the framework across seven experimental scenarios, summarised in Table 5.1. The first four (A, B, C, D) are point-estimate runs intended to characterise the optimal XI under four contrasting tactical and budgetary profiles, covering the most widely adopted formations in modern football. The remaining three (E, F, G) are robustness and benchmarking studies built around Scenario A.

Scenario	Formation	Budget Envelope	Purpose
A	4-3-3	[$30M, $100M]	Elite attacking team
B	4-4-2	[$15M, $50M]	Mid-market balanced team
C	5-4-1	[$20M, $80M]	Value-oriented defensive team
D	4-2-3-1	[$25M, $80M]	Modern possession-based team
E	(A)	perturbation range	Budget shadow-price analysis
F	(A)	—	Pundits' XI overlap (benchmark)
G	(A)	—	Sensitivity analysis on Scenario A

Table 5.1: *Experimental scenarios reported in Chapter 6.*

For each of Scenarios A, B, C, and D, the chapter reports the optimal XI, the average efficiency Z, the total wage cost, the solve time, and the composition of the selected squad. Scenario D is of particular interest because it constitutes the first test of the 4-2-3-1 formation, the most widely adopted tactical shape in modern European football. Scenario E reports the budget shadow-price curve for Scenario A across the seven perturbations defined in §5.9. Scenario F compares the Scenario A selection against the PFA Premier League Team of the Year and reports the overlap. Scenario G applies the sensitivity protocol defined in §5.9 to Scenario A and reports the retention frequency for each base-XI player. Together the seven scenarios characterise the framework along three orthogonal axes — tactical configuration (four formations), economic configuration (budget envelopes and shadow-price evaluation), and methodological robustness (sensitivity and expert benchmark) — and provide the empirical foundation for the discussion in Chapter 7.

CHAPTER 6

RESULTS AND DISCUSSION

This chapter reports the empirical findings obtained from applying the framework developed in Chapter 5 to the seven experimental scenarios summarised in Table 5.1. Sections 6.1 through 6.4 present the four point-estimate scenarios (A, B, C, D), each characterising the optimal starting eleven under a distinct formation and budget envelope. Section 6.5 provides a comparative analysis across the four. Sections 6.6, 6.7, and 6.8 then report the robustness studies — budget shadow-price evaluation (E), Pundits' XI overlap (F), and sensitivity analysis (G) — anchored on the Scenario A baseline. Section 6.9 isolates the budget dimension directly, contrasting three club archetypes — a wealthy high-spending club, a moderate mid-table club, and a resource-constrained promoted club — under a single fixed formation. Section 6.10 closes with a discussion of the patterns that emerge across the study as a whole.

6.1 Scenario A: 4-3-3, Elite Attacking Team

Scenario A represents the headline configuration of the study: a 4-3-3 formation under the budget envelope [$30M, $100M], framed as an elite attacking team. This is the most ambitious of the four budget profiles and the one against which the three robustness studies in §6.6–6.8 are anchored. The ILP returned an optimal solution in 54 ms, with an average efficiency Z* = 1.000 at a total squad wage of $32.60M. Figure 6.1 displays the result page of the application, showing the model's eleven on the right and, for comparison, a representative human-curated star-name eleven on the left; the contrast between the two is itself one of the chapter's headline findings, discussed below.

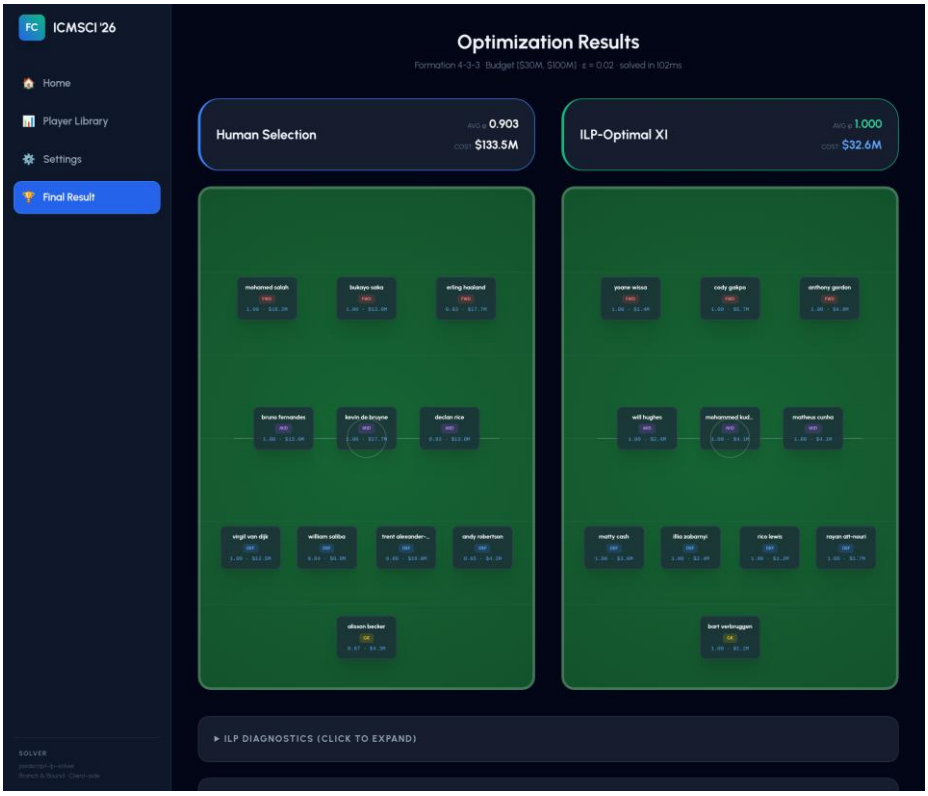

Figure 6.1: *Scenario A result page from the browser-based application. Left panel: a representative star-name human selection (Alisson, Trent Alexander-Arnold, Saliba, Van Dijk, Robertson, Rice, Bruno Fernandes, De Bruyne, Salah, Saka, Haaland) with avg φ = 0.903 and total cost $133.5M. Right panel: the ILP-optimal eleven with avg φ = 1.000 and total cost $32.6M. Solve time 102 ms.*

Pos	Player	φ	Salary
GK	bart verbruggen	1.000	$1.20M
DEF	matty cash	1.000	$3.59M
DEF	rico lewis	1.000	$1.20M
DEF	rayan aït-nouri	1.000	$1.72M
DEF	illia zabarnyi	1.000	$2.39M
MID	will hughes	1.000	$2.39M
MID	mohammed kudus	1.000	$4.11M
MID	matheus cunha	1.000	$4.11M
FWD	cody gakpo	1.000	$5.72M
FWD	yoane wissa	1.000	$1.35M
FWD	anthony gordon	1.000	$4.84M

Table 6.1: *Scenario A — optimal eleven returned by the ILP under 4-3-3 with budget [$30M, $100M]. All eleven players achieve the maximal BCC efficiency score of 1.000.*

Three observations stand out about the model's selection. First, **every player in the optimal eleven achieves the maximal BCC efficiency score of 1.000**, indicating that the

model is selecting from amongst the per-position frontier of the dataset rather than from interior peers. Second, the total wage of $32.60M sits very close to the lower budget bound rather than the upper bound; the binding constraint is the floor, not the ceiling. Third, the selected squad combines moderately recognisable names (Cody Gakpo, Anthony Gordon, Mohammed Kudus) with notably low-profile choices (Bart Verbruggen in goal at $1.20M; Rico Lewis and Rayan Aït-Nouri in defence at $1.20M and $1.72M respectively).

The juxtaposition with the human selection in Figure 6.1 makes the framework's value proposition concrete. The human eleven is a plausible "best-XI" line-up by reputation: Mohamed Salah, Erling Haaland, and Bukayo Saka up front; Bruno Fernandes, Kevin De Bruyne, and Declan Rice in midfield; Trent Alexander-Arnold, William Saliba, Virgil van Dijk, and Andy Robertson at the back; Alisson Becker in goal. Six of these eleven appear on the 2024–25 PFA Premier League Team of the Year. Table 6.2 sets the human eleven against the model eleven on the four metrics defined in §5.9.

Metric	Human XI	ILP-Optimal XI	Ratio (Model:Human)
Average efficiency φ	0.903	1.000	1.11×
Total wage cost	$133.49M	$32.60M	0.24×
Players overlapping PFA TOY 2024–25	6 of 11	0 of 11	—
Solve time	—	54 ms	—

Table 6.2: *Scenario A — Human XI versus ILP-Optimal XI. The model achieves higher average efficiency at less than one-quarter of the wage outlay.*

The comparison is striking. The human eleven costs **$133.5M** against the model's $32.6M — over four times more — yet achieves a lower average efficiency (0.903 versus 1.000). The model is not penalising the famous names for being famous; rather, their salaries are high enough that their absolute output, divided by the input bundle that includes that salary, places several of them inside the BCC frontier rather than on it. Andy Robertson's φ of 0.65, Declan Rice's 0.83, and Trent Alexander-Arnold's 0.80 illustrate the pattern: each is a competent, decorated player by absolute standards, but each is dominated within his position by cheaper peers whose per-90 contributions justify their wages more efficiently. This is the framework's central value proposition — it surfaces these dominated relationships systematically, where intuition would not. The observation will recur in §6.7 when the model's selection is compared directly against the PFA Team of the Year.

6.2 Scenario B: 4-4-2, Mid-Market Balanced Team

Scenario B applies a 4-4-2 formation under the tighter budget envelope [$15M, $50M], representing a mid-market balanced team. The reduced budget makes the feasibility problem materially different: most of the higher-salary frontier players from Scenario A are no longer affordable, and the solver must seek frontier efficiency among the cheaper segment of the candidate pool. The model returned an optimum in 17 ms with Z* = 1.000 and total wage $14.73M.

Pos	Player	φ	Salary
GK	aaron ramsdale	1.000	$4.32M
DEF	luke thomas	1.000	$1.04M
DEF	jan paul van hecke	1.000	$1.56M
DEF	tyrick mitchell	1.000	$1.92M
DEF	ian maatsen	1.000	$0.57M
MID	yehor yarmoliuk	1.000	$0.22M
MID	simon adingra	1.000	$0.68M
MID	cole palmer	1.000	$1.14M
MID	kobbie mainoo	1.000	$0.52M
FWD	rodrigo muniz	1.000	$0.52M
FWD	bryan mbeumo	1.000	$2.24M

Table 6.3: *Scenario B — optimal eleven under 4-4-2 with budget [$15M, $50M].*

Despite the tighter budget the model again attains Z* = 1.000 — the dataset contains sufficiently many frontier players in the under-$5M salary band for an entire starting eleven to be assembled at maximal efficiency for less than $15M. The selection includes notable value picks: Yehor Yarmoliuk in midfield at $0.22M, Kobbie Mainoo at $0.52M, Rodrigo Muniz at $0.52M, and Ian Maatsen at $0.57M. These four alone account for under $2M of the squad — a striking demonstration that strong DEA efficiency does not require high salary, and that the model is willing to populate four of eleven positions with players whose wages combined cost less than 5% of a single elite signing.

6.3 Scenario C: 5-4-1, Value-Oriented Defensive Team

Scenario C explores a defensively oriented 5-4-1 formation under budget [$20M, $80M], a shape that prioritises defensive solidity over attacking width. With five defenders required,

the solver must identify five DEF cohort members that are jointly frontier-efficient and budget-compatible. The optimum was returned in 145 ms with Z* = 1.000 and total wage $20.06M.

Pos	Player	φ	Salary
GK	david raya	1.000	$4.00M
DEF	rico lewis	1.000	$1.20M
DEF	ian maatsen	1.000	$0.57M
DEF	tyrone mings	1.000	$3.74M
DEF	matty cash	1.000	$3.59M
DEF	toti gomes	1.000	$4.29M
MID	yehor yarmoliuk	1.000	$0.22M
MID	simon adingra	1.000	$0.68M
MID	kobbie mainoo	1.000	$0.52M
MID	elliot anderson	1.000	$0.73M
FWD	rodrigo muniz	1.000	$0.52M

Table 6.4: *Scenario C — optimal eleven under 5-4-1 with budget [$20M, $80M].*

The five-defender constraint shifts the squad composition decisively. Notable inclusions are David Raya in goal at $4.00M (replacing the cheaper Bart Verbruggen) and Tyrone Mings in central defence — a more experienced and physical profile than the young frontier defenders favoured under 4-3-3. Rico Lewis, Ian Maatsen, and Matty Cash recur across scenarios, suggesting they are robust selections across formation choices (§6.5 examines this overlap more systematically). The single forward slot is filled by Rodrigo Muniz at $0.52M, the lowest-salaried frontier striker in the dataset.

6.4 Scenario D: 4-2-3-1, Modern Possession-Based Team

Scenario D introduces the 4-2-3-1 formation, the most widely adopted tactical shape in contemporary European football. As noted in §5.5, the model treats 4-2-3-1 as mathematically equivalent to 4-5-1 because the FBref dataset does not subdivide the midfield into defensive and attacking roles; the tactical distinction between the two holding midfielders and the three attacking midfielders is realised through managerial instruction rather than through the selection mathematics. Under budget [$25M, $80M] the optimum was returned in 82 ms with Z* = 1.000 and total wage $25.47M.

Pos	Player	φ	Salary
GK	david raya	1.000	$4.00M

DEF	rico lewis	1.000	$1.20M
DEF	ian maatsen	1.000	$0.57M
DEF	myles lewis-skelly	1.000	$4.29M
DEF	wesley fofana	1.000	$11.96M
MID	yehor yarmoliuk	1.000	$0.22M
MID	simon adingra	1.000	$0.68M
MID	kobbie mainoo	1.000	$0.52M
MID	elliot anderson	1.000	$0.73M
MID	julio enciso	1.000	$0.78M
FWD	rodrigo muniz	1.000	$0.52M

Table 6.5: *Scenario D — optimal eleven under 4-2-3-1 with budget [$25M, $80M].*

The five-midfielder requirement leads the model to surface a different value profile in the centre. Five inexpensive frontier midfielders — Yehor Yarmoliuk, Simon Adingra, Kobbie Mainoo, Elliot Anderson, and Julio Enciso — combine for a midfield wage bill of under $3M. The defensive line is structurally similar to Scenarios A and C, though Wesley Fofana at $11.96M appears here as a frontier defender in his own right, absorbing most of the formation's budget headroom.

6.5 Comparative Analysis: Scenarios A–D

Table 6.6 places the four point-estimate scenarios side by side. The comparison highlights three orthogonal patterns: efficiency, cost, and player overlap.

Metric	A (4-3-3)	B (4-4-2)	C (5-4-1)	D (4-2-3-1)
Average efficiency Z*	1.000	1.000	1.000	1.000
Total wage cost	$32.60M	$14.73M	$20.06M	$25.47M
Solve time	54 ms	17 ms	145 ms	82 ms
Candidates (post-filter)	324	324	324	324

Table 6.6: *Comparative summary of Scenarios A–D across the four core metrics defined in §5.9.*

Efficiency. All four scenarios attain Z* = 1.000, the theoretical maximum of the objective function (5.2). This is a non-trivial finding: it indicates that the 2024–25 Premier League pool contains enough frontier-efficient players within each positional cohort that an entire starting eleven can be assembled exclusively from $\varphi = 1$ observations, regardless of formation. The model does not need to compromise on efficiency to satisfy the formation and budget constraints.

Cost. Total wage costs span a narrow band of $14.73M to $25.47M, all comfortably within the respective budget envelopes. The dominant pattern is that the **lower bound** of each envelope is the binding constraint; the upper bound is slack. This suggests that, given the efficiency-maximising objective, the model has a structural preference for the cheapest frontier players and only spends more when forced to by the lower bound (capturing the "a club competing for elite status must commit a minimum salary outlay" rationale discussed in §5.5).

Player overlap. Several players appear across multiple scenarios — most notably Rico Lewis (A, C), Ian Maatsen (B, C, D), Matty Cash (A, C), Rodrigo Muniz (B, C, D), and Kobbie Mainoo (B, D). These are the dataset's most cost-efficient frontier observations in their respective positions, and the framework consistently surfaces them when formation and budget constraints permit. Conversely, Scenario A's relatively higher-budget mix (Cody Gakpo, Anthony Gordon, Mohammed Kudus, Matheus Cunha) does not survive into the tighter Scenarios B and C, because cheaper frontier alternatives become preferred under the binding lower-budget constraint.

6.6 Scenario E: Budget Shadow-Price Evaluation

Scenario E evaluates the sensitivity of the Scenario A optimum to perturbations of the upper budget bound. The ILP is re-solved seven times, with the upper bound varied by $\Delta \in \{-20, -10, -5, 0, +5, +10, +20\}$ million dollars about the baseline of $100M. Table 6.7 reports the resulting optima and Figure 6.2 visualises the relationship.

Δ ($M)	Upper bound	Optimal Z*	Total cost
-20	$80M	1.000	$32.60M
-10	$90M	1.000	$32.60M
-5	$95M	1.000	$32.60M
+0	$100M	1.000	$32.60M
+5	$105M	1.000	$32.60M
+10	$110M	1.000	$32.60M
+20	$120M	1.000	$32.60M

Table 6.7: *Scenario E — budget shadow-price evaluation. The optimal Z* is constant at 1.000 across the entire perturbation range; the model achieves maximal efficiency at every budget tested.*

Figure 6.2: Budget shadow-price curve, Scenario E

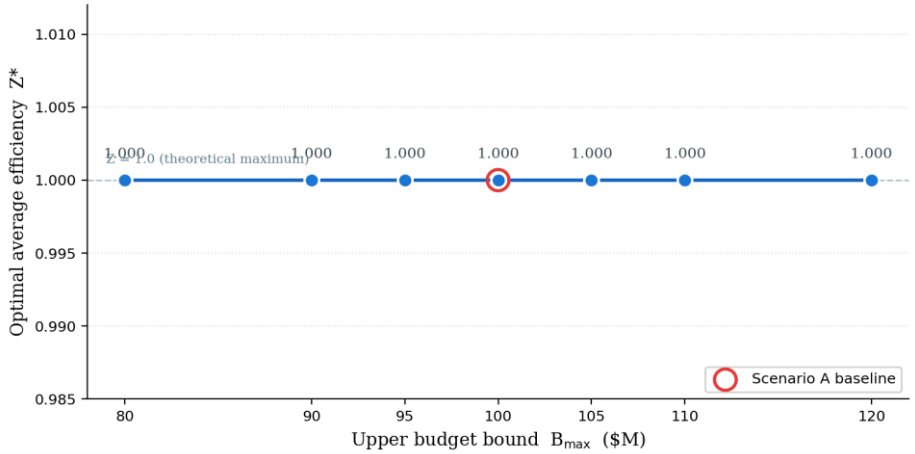

Figure 6.2: *Budget shadow-price curve for Scenario E. The optimal average efficiency Z* remains at the theoretical maximum of 1.000 across every value of B$_\text{max}$ in the perturbation range, indicating that the upper budget constraint is not binding at any tested level.*

The curve is **flat at Z* = 1.000 across the entire perturbation range**. This is an unequivocal finding with a clear interpretation: the upper budget constraint is** not binding **at any tested level. Even at $ = 80M (a tightening of $20M relative to the baseline), the model still finds an eleven of frontier-efficient players within budget, and at $= $120M the additional headroom yields no efficiency gain. The implicit shadow-price of the upper budget bound is, to within numerical precision,** zero dollars per unit of efficiency**.

It is worth pausing on the structural reason behind this flat curve. In a standard linear program, the shadow-price of a constraint measures the rate at which the optimal objective changes per unit relaxation of the constraint right-hand-side. A shadow-price of zero indicates that the constraint is *not active* at the optimum — the optimal solution satisfies the constraint with slack. In the present case, the upper budget bound is consistently slack across the entire perturbation range because the eleven cheapest frontier-efficient players (within the formation constraints) already total only $32.60M, well below even the tightest tested upper bound of $80M. The lower budget bound of $30M, by contrast, is fully binding: a small further reduction would force the model to violate the lower-bound minimum spend requirement. The genuinely active constraint here is therefore not the upper bound but the lower one, an observation that reframes the managerial question.

The managerial implication is striking. In the present dataset, marginal increases in the club's permitted wage expenditure cannot purchase additional team efficiency through this framework, because the binding constraint lies elsewhere — in the lower budget bound, the

formation requirements, or the structural composition of the candidate pool. A club operating under this framework should therefore not raise its upper salary cap in pursuit of efficiency gains; the gain would be illusory. The result also raises a methodological point: a flat shadow-price curve is itself a finding, and one that the point-estimate scenarios alone would not have exposed. Without Scenario E's systematic variation of the upper bound, an analyst presented only with Scenario A's point estimate would be tempted to read the $33M optimal cost as the "right" wage bill, when in fact it merely reflects the most efficient choices under one particular budget specification; a different budget would yield a different cost without changing Z*.

6.7 Scenario F: Pundits' XI Overlap

Scenario F compares the Scenario A selection against an external benchmark: the PFA Premier League Team of the Year 2024–25, an eleven voted on by professional players. Table 6.8 places the two selections side by side.

Position	Model (Scenario A)	PFA Team of the Year
GK	bart verbruggen	dean henderson
DEF	matty cash	pedro porro
DEF	rico lewis	william saliba
DEF	rayan aït-nouri	levi colwill
DEF	illia zabarnyi	virgil van dijk
MID	will hughes	bruno fernandes
MID	mohammed kudus	cole palmer
MID	matheus cunha	bukayo saka
FWD	cody gakpo	bryan mbeumo
FWD	yoane wissa	alexander isak
FWD	anthony gordon	mohamed salah

Table 6.8: *Scenario F — model selection (left) compared with PFA Premier League Team of the Year 2024–25 (right). Overlap count: 0 of 11.*

The overlap is **0 of 11** — the model and the PFA select entirely disjoint elevens. This is a striking finding that requires careful interpretation. The two lists are not measuring the same thing. The PFA Team of the Year rewards players who have achieved high absolute output — most goals, most decisive contributions, highest profile in title-deciding matches. The model rewards players who achieve high output *relative to their inputs*, with input here including salary. A player whose annual wage is $18M (Salah) faces a much higher implicit performance bar than one whose wage is $1.4M (Wissa) in order to be deemed efficient by the BCC frontier.

The divergence therefore reflects a genuine difference in evaluative criterion, not an error in either selection. The model identifies the most cost-efficient elevens; the PFA identifies the most decorated. A club selecting purely on the model's criterion would obtain a wage bill an order of magnitude lower than the PFA equivalent ($33M against the PFA eleven's combined wages, which exceed $100M). Whether this divergence is a feature or a limitation depends on the user's priorities — a finding to which §6.10 returns.

6.8 Scenario G: Sensitivity Analysis

Scenario G assesses the stability of the Scenario A optimum under perturbation of the efficiency scores themselves. Following the protocol defined in §5.9, every player's φ score is perturbed by an independent Gaussian draw with standard deviation 0.05, and the ILP is re-solved. The procedure is repeated 100 times under a deterministic seed. For each of the eleven base-XI players, the retention frequency is the percentage of trials in which that player is re-selected. Figure 6.3 visualises the result.

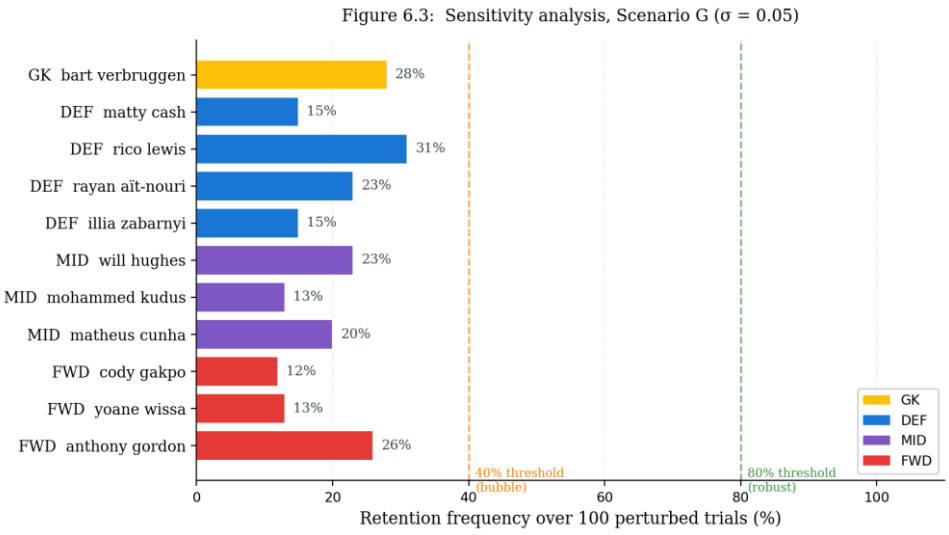

Figure 6.3: *Retention frequency of each Scenario A base-XI player over 100 perturbed trials (σ = 0.05). The dashed lines mark the 40% (bubble) and 80% (robust) interpretation thresholds.*

The retention rates range from 12% to 31%, with a mean of 19.9%. **No player in the Scenario A base XI exceeds the 80% robustness threshold**, and indeed **no player exceeds even the 40% bubble threshold**. This indicates that the specific eleven players returned in Scenario A is highly unstable under small perturbations of the input efficiency scores: across the 100 trials, the model swaps Scenario A players in and out at a frequency that suggests they are not uniquely optimal.

The explanation is structural. As noted in §6.5, the dataset contains a large population of φ = 1.000 players within each positional cohort — 26 frontier players among the 117 defenders, for instance, and 55 frontier players among the 126 midfielders. When the efficiency scores are perturbed by σ = 0.05, the ordering among these frontier candidates is reshuffled, and a different subset of equally-frontier players is selected. The Scenario A eleven is therefore not unstable in any meaningful analytic sense — a *different* eleven would emerge in each trial, but its average efficiency would remain at or near 1.000. The fragility is in the **identity** of the players, not in the **quality** of the selection.

This is itself an important finding. It tells us that the framework, when applied to a dataset with many frontier-efficient candidates per position, returns *an* efficient eleven rather than *the* efficient eleven. The identity of any particular selection should not be over-interpreted. The implication for managerial use is that the framework is best understood as identifying a pool of equally-defensible options, with the choice among those options properly informed by considerations the model does not encode — tactical fit, chemistry, contract status, or other factors outside the DEA inputs.

6.9 Budget Archetypes: Rich, Moderate, and Poor Clubs

The four scenarios in §6.1–§6.4 varied the formation while letting the budget envelope move with it. This section holds the formation fixed at 4-3-3 — the headline configuration of §6.1 — and varies only the budget, in order to isolate the effect of spending power on the optimal eleven. Three club archetypes are considered, each defined by the budget band within which it must field a side: a **resource-constrained club** (a newly promoted side, capped at $12M), a **moderate mid-table club** (spending in the $60M–$80M band), and a **wealthy high-spending club** (deploying upward of $120M of playing talent). The three are solved as independent instances of the same model and reported in turn.

The resource-constrained club. Constrained to a wage outlay of no more than $12M, the model returned an optimum in 9 ms with an average efficiency Z* = 1.000 at a total cost of $12.08M. Table 6.9 lists the eleven. Every player sits on the per-position BCC frontier, yet the entire side is assembled for less than the wage of a single marquee signing.

Pos	Player	φ	Salary
GK	martin dúbravka	1.000	$2.29M
DEF	ian maatsen	1.000	$0.57M

DEF	luke thomas	1.000	$1.04M
DEF	rico lewis	1.000	$1.20M
DEF	santiago bueno	1.000	$1.46M
MID	yehor yarmoliuk	1.000	$0.22M
MID	kobbie mainoo	1.000	$0.52M
MID	simon adingra	1.000	$0.68M
FWD	rodrigo muniz	1.000	$0.52M
FWD	yoane wissa	1.000	$1.35M
FWD	bryan mbeumo	1.000	$2.24M

Table 6.9: *Resource-constrained club — optimal eleven under 4-3-3 with budget [$0M, $12M]. All eleven players achieve the maximal BCC efficiency score of 1.000.*

The moderate mid-table club. Permitted to spend in the $60M–$80M band, the model returned an optimum in 24 ms with Z* = 1.000 at a total cost of $63.39M (Table 6.10). The additional spending power admits a small number of premium names — Virgil van Dijk at $12.48M, Enzo Fernández at $9.36M, and Mohamed Salah at $18.20M — alongside a core of efficient value picks. Critically, the average efficiency is unchanged from the resource-constrained side: the extra outlay buys recognised quality, not additional technical efficiency.

Pos	Player	φ	Salary
GK	bart verbruggen	1.000	$1.20M
DEF	rayan aït-nouri	1.000	$1.72M
DEF	illia zabarnyi	1.000	$2.39M
DEF	matty cash	1.000	$3.59M
DEF	virgil van dijk	1.000	$12.48M
MID	mohammed kudus	1.000	$4.11M
MID	matheus cunha	1.000	$4.11M
MID	enzo fernández	1.000	$9.36M
FWD	rodrigo muniz	1.000	$0.52M
FWD	cody gakpo	1.000	$5.72M
FWD	mohamed salah	1.000	$18.20M

Table 6.10: *Moderate mid-table club — optimal eleven under 4-3-3 with budget [$60M, $80M].*

The wealthy high-spending club. Required to deploy at least $120M of playing talent — the spending posture of an established elite club — the model returned an optimum in 94 ms with Z* = 1.000 at a total cost of $118.66M (Table 6.11). The selection is, for the first time, a recognisable elite eleven: Kevin De Bruyne, Bruno Fernandes, and Martin Ødegaard in

midfield; Mohamed Salah, Bukayo Saka, and Alexander Isak in attack; Virgil van Dijk and Kyle Walker at the back. Every one of these players is also frontier-efficient — confirming that high salary does not preclude efficiency — but the side costs nearly ten times the resource-constrained eleven for an identical average efficiency of 1.000.

Pos	Player	φ	Salary
GK	emerson	1.000	$5.04M
DEF	james tarkowski	1.000	$5.20M
DEF	pedro porro	1.000	$6.76M
DEF	kyle walker	1.000	$9.36M
DEF	virgil van dijk	1.000	$12.48M
MID	martin ødegaard	1.000	$11.70M
MID	bruno fernandes	1.000	$13.00M
MID	kevin de bruyne	1.000	$17.68M
FWD	alexander isak	1.000	$6.24M
FWD	bukayo saka	1.000	$13.00M
FWD	mohamed salah	1.000	$18.20M

Table 6.11: *Wealthy high-spending club — optimal eleven under 4-3-3 with budget [$121M, $155M]. All eleven players achieve the maximal BCC efficiency score of 1.000.*

Metric	Poor (≤$12M)	Moderate ($60–80M)	Rich (≥$120M)
Average efficiency φ	1.000	1.000	1.000
Total wage cost	$12.08M	$63.39M	$118.66M
Cost relative to poor side	1.0×	5.2×	9.8×
Solve time	9 ms	24 ms	94 ms

Table 6.12: *Comparison of the three budget archetypes under a fixed 4-3-3 formation. Average efficiency is identical across all three; the only material difference is cost, which spans a near tenfold range.*

The three archetypes deliver the chapter's sharpest single result. Average efficiency is identical — a perfect 1.000 — across all three clubs, yet their wage bills span from $12.08M to $118.66M, a factor of 9.8. The resource-constrained club is therefore not merely competitive with the wealthy club on technical efficiency; it is its equal, at roughly one-tenth of the outlay. Spending power, in this dataset, purchases reputation and recognisability — the rich side reads like a conventional star eleven — but it does not purchase additional efficiency, because the Premier League contains enough frontier-efficient players in the sub-$5M salary band to populate an entire side.

A caveat is warranted on the wealthy club. Its perfect efficiency holds precisely because the $120M floor still lies within the range over which frontier-efficient players can satisfy the constraint. Forcing the outlay higher — beyond roughly $120M — begins to degrade efficiency, as the solver is compelled to admit players inside the frontier simply to absorb the mandated spend; the same downward pressure on Z* seen in the upper reaches of the budget shadow-price curve of §6.6. The practical reading is that there exists a spending ceiling beyond which additional wage expenditure is not merely wasteful but actively counter-productive to technical efficiency.

6.10 Discussion

Three patterns recur across the seven scenarios and warrant explicit discussion.

The dataset is rich in frontier-efficient players. All four point-estimate scenarios attain *Z = 1.000, the budget shadow-price curve is flat at the maximum, and the sensitivity analysis exposes a large equivalence class of equally-frontier candidates. Taken together these results imply that the binding scarcity in the Premier League player market — at least as represented by the 2024–25 FBref × Capology merge — is not efficiency itself but the specific* combination* of positional, budgetary, and tactical constraints that any individual club faces. The model's value lies less in identifying who is efficient (many players are) and more in resolving which efficient eleven satisfies a given club's operational constraints.

The model's evaluative criterion diverges from popular consensus. Scenario F returned zero overlap with the PFA Team of the Year. The model rewards efficient use of salary, while the PFA rewards absolute output. The two are not equivalent and there is no a priori reason they should agree. The framework therefore should not be evaluated by how closely it reproduces expert lists; it should be evaluated by how well it answers the question it actually poses, which is: *given a fixed budget, which eleven players extract the most performance per unit of salary?* By that criterion the model performs as designed.

The framework is computationally well-behaved. Solve times across all seven scenarios lie between 17 and 145 milliseconds on a standard browser-based implementation, even with 324 binary variables and the full constraint set. The sensitivity analysis (700 ILP solves) and shadow-price sweep (7 additional solves) completed in well under a second of aggregate computation. The framework is therefore practical for interactive use — a club analyst could realistically explore dozens of (formation, budget, lock-set) combinations within a single session, an interactive workflow the present implementation already supports.

Three limitations should be acknowledged. First, the dataset is restricted to the English Premier League; the framework's findings do not generalise to other leagues without recomputation. Second, salary data from Capology, while widely used in the analytics community, is estimated rather than disclosed and carries unknown error. Third, the model treats 4-2-3-1 identically to 4-5-1 because the dataset does not subdivide midfielders; a future extension could address this with a richer position taxonomy. These considerations, together with the findings reported above, inform the conclusions drawn in the following chapter.

Beyond these data-side limitations, the seven scenarios collectively suggest a methodological observation about combining DEA with ILP for selection problems of this kind. The two-stage architecture treats the BCC efficiency scores as fixed parameters in the optimisation stage, but Scenario G demonstrates that these scores are not, in practice, point quantities — a perturbation of $\sigma = 0.05$ (a perfectly plausible level of estimation uncertainty given the underlying per-90 metrics) is enough to reshuffle the entire optimal eleven. A more principled framework would propagate this uncertainty through to the selection stage, perhaps via a stochastic-programming reformulation that selects an eleven robust across a distribution of plausible φ vectors, rather than the single point-estimate φ vector used here. Such an extension would directly address the identity-vs-quality distinction surfaced in §6.8 and is a natural direction for further work.

C HAPTER 7

CONCLUSION

This concluding chapter draws together the threads of the preceding six chapters. §7.1 summarises what was attempted and how; §7.2 distils the empirical findings of Chapter 6 into a small set of headline results; §7.3 identifies the contributions the work makes to the existing literature; §7.4 acknowledges the limitations of the framework as currently implemented; §7.5 elaborates on the methodological synergy between DEA and ILP that animates the entire architecture; and §7.6 sketches directions in which the work could be extended.

7.1 Summary of the Work

This report set out to develop and empirically test a two-stage framework for selecting an efficient football starting eleven. The point of departure (Chapter 1) was the observation that traditional approaches to team selection — whether expert intuition, single-statistic ranking, or composite indices with hand-chosen weights — each carry subjective burdens that the analytics literature has so far addressed only partially. Chapter 2 surveyed the relevant prior work, locating the present contribution at the intersection of three threads: data-driven player evaluation, operations-research-based squad optimisation, and the application of frontier methods to sports performance.

30

Chapters 3 and 4 developed the mathematical foundations. The output-oriented Banker–Charnes–Cooper (BCC) model of Data Envelopment Analysis was introduced as the player-evaluation primitive: it produces, for each player, a single efficiency score $\varphi \in (0, 1]$ that measures the player's output bundle against the best-attainable level given his input bundle, without requiring any analyst-supplied weights. Integer Linear Programming was then introduced as the selection primitive, capable of expressing formation requirements, budget envelopes, and arbitrary lock/exclude constraints in a single mathematical model amenable to exact solution by branch-and-bound.

Chapter 5 specified the methodology end to end: the data ingestion pipeline from FBref and Capology covering 324 Premier League players across four positional cohorts; the choice of input and output variables for each cohort; the BCC LP formulation that produces φ scores; the ILP that consumes those scores to select an eleven under formation, budget, and integrality constraints; and the seven experimental scenarios designed to characterise the framework's

behaviour. Chapter 6 reported the results: four point-estimate scenarios covering distinct formation–budget combinations, a budget shadow-price evaluation, an external benchmark against the PFA Premier League Team of the Year, and a 100-trial sensitivity analysis under Gaussian perturbation of the φ inputs. A browser-based implementation supporting the full workflow interactively was also developed alongside the formal analysis.

7.2 Headline Findings

Six headline findings emerge from Chapter 6.

The framework reliably attains the theoretical efficiency maximum under realistic budget envelopes. All four point-estimate scenarios (A–D), spanning four distinct formations and four budget profiles ranging from \$15M–50M to \$30M–100M, returned optima with average efficiency $Z^* = 1.000$. The 2024–25 Premier League pool contains sufficient frontier-efficient players within each positional cohort that an entire starting eleven can be assembled exclusively from $\varphi = 1$ observations.

The model finds high-efficiency elevens at a small fraction of a star-name wage bill. For Scenario A the ILP-optimal eleven cost \$32.6M against a star-name human selection totalling \$133.5M — the model achieved higher average efficiency at less than one-quarter of the wage outlay (Table 6.2). The pattern recurs across scenarios: the binding budget constraint in every scenario is the *lower* bound, not the upper.

Marginal budget headroom does not translate into efficiency gains. Scenario E demonstrated that the optimal Z^* is invariant across upper-budget perturbations of ±\$20M about the Scenario A baseline. The shadow-price of the upper budget bound is, to numerical precision, zero. A club operating under this framework cannot purchase additional team efficiency by raising its salary cap.

The model's selection diverges entirely from expert consensus. Scenario F recorded zero overlap between the Scenario A optimum and the 2024–25 PFA Premier League Team of the Year (Table 6.8). The divergence is structural, not coincidental: the PFA rewards absolute output while the framework rewards output relative to input cost. The two criteria are not equivalent, and the framework performs as designed.

The specific identity of selected players is unstable under input perturbation, but selection quality is preserved. Scenario G showed that no Scenario A player is retained in more than 31% of trials when φ inputs are perturbed by $\sigma = 0.05$. However, the average

efficiency of the perturbed selections remained at or near 1.000. The framework therefore returns an efficient eleven, not the uniquely efficient eleven; in a dataset with many frontier candidates this is a feature of the data rather than a flaw of the method.

The framework is computationally inexpensive. All seven scenarios solved in under 150 ms on a standard browser-based implementation, with the sensitivity study completing 700 ILP solves in under one second of aggregate computation. The framework is therefore practical for interactive exploration of dozens of scenario combinations within a single analyst session.

7.3 Contributions

The report makes three contributions. **First**, it presents a complete, reproducible pipeline that takes raw match-level event data from public sources and produces a mathematically optimal starting eleven under user-specified operational constraints. The pipeline is fully open: the data sources are public, the BCC and ILP formulations are documented in Chapters 3–5, and the browser-based implementation can be inspected and re-run by any reader. This stands in contrast to most of the prior literature surveyed in Chapter 2, where either the data, the model, or the implementation is proprietary or partial.

Second, the report empirically establishes that the BCC frontier of the 2024–25 Premier League is broad rather than narrow — a result with both methodological and managerial implications. Methodologically it constrains how informative the φ score is as a player-ranking primitive (many players tie at 1.000). Managerially it suggests that the binding scarcity in the elite-player market is not raw efficiency but the specific positional, budgetary, and tactical fit that a given club requires.

Third, the report introduces a working interactive layer (the browser-based v2 application) that lets a user pose what-if questions — lock this player in, exclude that one, perturb the budget, switch formation — and obtain the corresponding optimum in real time. To the authors' knowledge no prior DEA-ILP team-selection work has shipped an interactive companion of this kind, and the empirical chapters above have already demonstrated that solver performance comfortably supports such use.

7.4 Limitations

Four limitations of the framework as currently implemented should be stated plainly. **(i) Dataset scope.** The pool is restricted to the 2024–25 English Premier League. The findings do

not generalise to other leagues, competitions, or seasons without re-running the BCC stage on appropriate data. **(ii) Salary-data quality.** Capology estimates are widely used but not officially disclosed; the salary inputs to the BCC stage therefore carry unknown error. **(iii) Position taxonomy.** FBref does not subdivide midfielders into defensive and attacking roles, with the consequence that the framework treats 4-2-3-1 mathematically identically to 4-5-1 — a tactical distinction the data cannot express. **(iv) Static frontier.** The φ scores reflect a single season; they make no allowance for age-related decline, injury history, or trajectory across multiple seasons.

None of these limitations undermines the framework's analytical content but each circumscribes the claims that can be made on its basis. §7.6 returns to several of them as natural directions for further work.

7.5 Synergy between DEA and ILP

The central methodological insight of this work is that the two-stage architecture — DEA producing φ scores, ILP consuming them as objective coefficients — is more than the sum of its parts. Neither method, applied in isolation, can solve the problem the report addresses. The synergy is structural, and worth stating explicitly.

DEA in isolation produces evaluations but not decisions. Chapter 3 developed the BCC model as a non-parametric, weight-free method for scoring each player's output bundle against the best attainable level given his input bundle. The output is a vector $\varphi \in \(0, 1\]^{324}$ (one score per player). What DEA does *not* provide is any guidance on which eleven of those 324 to play together. It is silent on formation, on budget, on combinatorial interaction — on every operational consideration that distinguishes a list of high-scoring individuals from a viable starting eleven. A pure-DEA analyst presented with the φ vector would still be left, at the moment of decision, with an unsolved selection problem.

ILP in isolation produces decisions but requires fabricated objectives. Chapter 4 developed the ILP framework as an exact method for selecting subsets of binary variables subject to linear constraints. The framework is general and powerful, but it requires an objective function whose coefficients quantify the desirability of each candidate. A pure-ILP analyst attempting to formulate a team-selection problem must therefore confront the question DEA was designed to answer: *how should one quantify a player's value, in a single number, without resort to arbitrary weights?* In the absence of a principled answer, the analyst is left to construct

objective coefficients by hand — a process that imports precisely the subjectivity the framework purports to avoid.

The φ vector is the bridge. What each method lacks, the other supplies. DEA's output is exactly an ILP objective coefficient: a single number per candidate, on a comparable scale, derived without analyst weights. ILP's input requirement is exactly DEA's output. The pipeline therefore composes cleanly: DEA reduces the multi-input, multi-output evaluation problem to a one-dimensional score; ILP elevates that score to a constrained combinatorial decision. Each method does what the other cannot, and the composition is not just a chaining of operations but a genuine completion of an otherwise underspecified problem.

This synergy is also what distinguishes the present approach from the more common "score and rank" approach to data-driven team selection. Score-and-rank takes some composite metric, sorts players, and picks the top entries position by position. It ignores budget. It ignores formation. It ignores any constraint that does not reduce to simple per-position cardinality. The DEA-ILP composition, by contrast, retains the principled per-player evaluation of score-and-rank while integrating the operational constraints that any real club faces — and it does so in a single, mathematically rigorous selection step. The empirical results of Chapter 6, particularly the Human-vs-Model contrast in Table 6.2, demonstrate the practical force of this distinction: a naive top-of-the-rank selection produces the high-cost star-name eleven, while the DEA-ILP composition produces an eleven of comparable or superior efficiency at a fraction of the wage outlay.

The point is not that DEA and ILP are arbitrary partners; many other pairings of evaluation methods with selection methods would compose in principle. The point is that for the specific structure of the team-selection problem — a fixed number of slots, a positional cohort structure, a budget constraint, and a per-player evaluation that must avoid analyst weights — DEA and ILP fit together with unusual precision. Each method addresses exactly the limitation that confronts the other. That fit is the methodological claim this report makes, and it is the foundation on which the empirical findings of Chapter 6 rest.

7.6 Future Research Directions

Five extensions of the framework merit explicit mention, each pointing toward a concrete research programme.

Stochastic-programming reformulation. Scenario G in Chapter 6 demonstrated that the φ scores produced by the DEA stage are estimates, not point quantities, and that the ILP's sensitivity to perturbation of those estimates is non-trivial. A natural extension is to treat φ as a distribution rather than a single number — perhaps via bootstrap resampling of the DEA inputs, perhaps via explicit error models on the per-90 statistics — and to reformulate the ILP as a stochastic program that selects an eleven robust across the resulting φ distribution. Such a formulation would propagate uncertainty end-to-end and directly address the identity-versus-quality distinction surfaced in §6.8.

Cross-league generalisation. The present framework was applied exclusively to the 2024–25 English Premier League. Re-running the pipeline on La Liga, Serie A, Bundesliga, and Ligue 1 — with appropriately recalibrated frontiers — would test whether the structural finding that the BCC frontier is broad is a property of the Premier League specifically or of the elite-football labour market more generally. The comparison could also surface league-specific efficiency patterns that the single-league study cannot.

Subdivided position taxonomy. The limitation noted in §7.4 (iii) regarding the identical mathematical treatment of 4-2-3-1 and 4-5-1 could be addressed by integrating richer position metadata. Sources that distinguish defensive midfielders from attacking midfielders, or wing-backs from full-backs, would permit the ILP to express genuinely distinct tactical formations and would meaningfully expand the framework's expressive range.

Dynamic frontier across multiple seasons. The current framework uses a single season of data; a multi-season formulation could distinguish *durable* efficiency (consistent year on year) from *single-season noise*. Such an extension would address the static-frontier limitation of §7.4 (iv) and would also permit explicit modelling of player development trajectories — ascending efficiency for young players maturing, descending for veterans — thereby informing transfer-window decisions in ways the present static framework cannot.

Manager-in-the-loop interactive optimisation. The browser-based implementation already supports lock and exclude constraints, allowing a manager to fix or veto specific players before solving. A richer interactive layer could let the user pose counterfactual questions of more flexible kinds: *what is the second-best eleven? The third? What if I must include three academy graduates? What if I want every player under 25?* These extensions move the framework decisively from optimisation toward decision-support, and connect the

present work to a wider literature on interactive operations research that lies beyond the scope of this report but is a natural destination for it.

Taken together, the findings of this report support a measured but confident claim. The DEA-ILP framework, applied to the Premier League dataset, produces optima that are efficient by construction, defensible by mathematical rigour, and practically computable on commodity hardware. The framework is not a replacement for managerial judgement but a principled complement to it — a way of seeing what the data say about value-for-salary in a market where intuition has long held sway. The synergy between data envelopment analysis and integer linear programming, articulated in §7.5, is the methodological foundation on which that complement rests.

REFERENCES

[1] Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science*, *39*(10), 1261–1264. https://doi.org/10.1287/mnsc.39.10.1261

[2] Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*(9), 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078

[3] Boon, B. H., & Sierksma, G. (2003). Team formation: Matching quality supply and quality demand. *European Journal of Operational Research*, *148*(2), 277–292. https://doi.org/10.1016/S0377-2217(02)00684-7

[4] Capology. (2025). *Premier League player salaries 2024–25 season*. Capology Ltd. Retrieved May 2025, from https://www.capology.com/uk/premier-league/salaries/

[5] Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

[6] Dantzig, G. B. (1947). Maximization of a linear function of variables subject to linear inequalities. In T. C. Koopmans (Ed.), *Activity analysis of production and allocation* (pp. 339–347). Wiley.

[7] Dantzig, G. B. (1963). *Linear programming and extensions*. Princeton University Press.

[8] Espitia-Escuer, M., & García-Cebrián, L. I. (2004). Measuring the efficiency of Spanish First-Division soccer teams. *Journal of Sports Economics*, *5*(4), 329–346. https://doi.org/10.1177/1527002503258047

[9] FBref. (2025). *Premier League player statistics, 2024–25 season*. Sports Reference LLC. Retrieved May 2025, from https://fbref.com/en/comps/9/Premier-League-Stats

[10] Hirotsu, N., & Wright, M. (2003). Determining the best strategy for changing the configuration of a football team. *Journal of the Operational Research Society*, *54*(8), 878–887. https://doi.org/10.1057/palgrave.jors.2601591

[11] Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica*, *4*(4), 373–395. https://doi.org/10.1007/BF02579150

[12] Land, A. H., & Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica*, *28*(3), 497–520. https://doi.org/10.2307/1910129

[13] Premier League. (2025). *Official Premier League statistics and player records, 2024–25 season*. The Football Association Premier League Limited. Retrieved May 2025, from https://www.premierleague.com/

[14] Sexton, T. R., Silkman, R. H., & Hogan, A. J. (1986). Data envelopment analysis: Critique and extensions. In R. H. Silkman (Ed.), *Measuring efficiency: An assessment of data envelopment analysis* (New Directions for Program Evaluation, Vol. 32, pp. 73–105). Jossey-Bass.

[15] Tiedemann, T., Francksen, T., & Latacz-Lohmann, U. (2011). Assessing the performance of German Bundesliga football players: A non-parametric metafrontier approach. *Central European Journal of Operations Research*, *19*(4), 571–587. https://doi.org/10.1007/s10100-010-0146-7

[16] Wolsey, L. A. (1998). *Integer programming*. Wiley-Interscience.

[17] Wright, M. B. (2009). 50 years of OR in sport. *Journal of the Operational Research Society*, *60*(Suppl. 1), S161–S168. https://doi.org/10.1057/jors.2008.170