POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Masters's Degree Thesis

Predicting Young Football Talents' Market Value and Optimizing Team Transfers

Supervisors

Prof. Paolo GARZA

Prof. Jesús CARRETERO PÉREZ

(Supervisor at host university UC3M during mobility)

Candidate

Ali ALHAJ HASSAN

September 2025

Summary

This thesis presents a data-driven framework for improving football club transfer strategies by predicting the market value of young players and optimizing player acquisitions. The study begins by constructing two original datasets: one focused on individual player performance, and the other on overall team performance. Using these datasets, the research explores the relationship between key performance indicators and market valuation, with an emphasis on feature engineering and exploratory data analysis.

A predictive model is then developed to estimate the market value of young football talents based on historical and performance-related data. Following this, an optimization model is proposed using Pyomo and the Gurobi solver to assist clubs in selecting the most suitable players within given budgetary and squad constraints. The proposed methodology combines machine learning, statistical analysis, and operations research to support informed and strategic decision-making in professional football recruitment.

Keywords: Football analytics, young players, market value prediction, machine learning, team transfer optimization, Pyomo, Gurobi, sports data analysis, feature engineering, recruitment strategy.

Acknowledgements

I am deeply grateful to everyone who supported me throughout this work and my academic journey. To my family, thank you for your unwavering encouragement and belief in me, especially during challenging times.

I am deeply grateful to my supervisors, **Prof. Jesús Carretero Pérez (UC3M)** and **Prof. Paolo Garza (Politecnico di Torino)**, whose guidance, feedback, and constant support were vital from the very beginning of this project.

To my friends and classmates, thank you for sharing the journey, the perspective, and memorable moments.

I also wish to thank **Politecnico di Torino**, my home university, for laying the foundations of my academic and professional path, and **Universidad Carlos III de Madrid**, where this work was carried out, for providing the welcoming environment, resources, and support during my mobility that made this thesis possible.

This thesis reflects the collective contribution of all of you—thank you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Football is considered the most popular sport in the world, with an estimated 3.5 billion fans, according to World Atlas [1]. Played in more than 200 countries around the world, each with its own associations and football leagues, the sport represents a massive market, generating billions in revenues through player transfers, sponsorships, broadcast rights, and merchandise sales. In recent years, the football transfer market has grown exponentially, for instance, according to the British sports newspaper The Sun [2] Premier League teams spent around £1.3 billion on summer transfers in 2024.

Clubs are increasingly investing in young talent to secure future stars or generate significant revenue in the coming years. Some clubs specialize in talent development and sales, making significant profits from player transfers. For example, according to transfermarkt.com [3], from the 2020/2021 season to the 2024/2025 season, SL Benfica earned €303.42 million in revenue from selling young players under 23 years of age while spending €200.85 million, resulting in a net surplus of €102.57 million. Similarly, LOSC Lille generated €297.27 million from transfers while spending only €116.87 million, leading to a €180.40 million profit. Borussia Dortmund, known for its player development strategy, earned €285.85 million while spending €187.65 million, achieving a net balance of €98.20 million from selling young players.

Conversely, some top-spending clubs prioritize acquiring young talent rather than selling. According to Transfermarkt, from the 2020/2021 season to the 2024/2025 season, Chelsea FC spent €969.55 million on young players but only generated €201.44 million from player sales, resulting in a substantial deficit of €768.11 million [3], Without any significant success in winning trophies, this demonstrates an example of non-optimized spending. In contrast, Manchester City exemplifies success and effective spending optimization. The club has not only invested significantly in player acquisitions, but these investments have also been translated into financial and sporting achievements.

Financially, the club reported record revenues of £715 million for the 2023/24 season, according to their official annual report [4]. Additionally, Manchester City has generated over £260 million from player sales while spending around £300 million on young players over the past two years [3]. These investments have positioned Manchester City as one of the top teams in Europe and the world, winning the Premier League title for the last four seasons and winning the UEFA Champions League for the season 2022/2023 which is considered the most important european tournament. They remain competitive for this season as well, with the young players recruited playing a crucial role in these accomplishments.

The contrasting transfer strategies of clubs like SL Benfica, LOSC Lille, and Borussia Dortmund which profit from developing young talent for success financially and sportswise, versus clubs like Chelsea FC, which have spent heavily without immediate returns, opens the possibility of trying data-driven techniques and machine learning (ML) in football recruitment to achieve the best results. While known stars come with clear market values, the price of young unknown talents is often uncertain, making recruitment a high-risk investment. Data-driven models can bridge this gap by analyzing extensive player performance metrics, scouting reports, and historical transfer trends to estimate a player's true market value. These models help clubs and recruiters optimize their transfer strategies by aligning acquisitions with financial constraints, squad needs, and playing styles.

1.2 Objectives and Goals of the Study

The objective of this study is to develop a data-driven framework for optimizing young player acquisitions in football by predicting their market value and then developing an optimization model to select the players that are most fit for the squad. The study aims to:

- **Create new datasets** specifically gathered for this research, focusing on young players and the performance of the team. Specifically, two data sets, one for players' performance and another for teams' performance.
- **Analyze the datasets** and the attributes, identifying key features and their relationships with market value, and analyze team performance in a purely data-driven approach.
- **Develop a predictive model** capable of assessing and predicting a young player's market value based on performance metrics, feature engineering, historical transfer data, and other relevant attributes.
- **Build an optimization model** that helps teams select players based on the club's financial constraints, squad requirements, and player attributes, ensuring maximum success from transfer decisions.

1.3 Workflow of the Study

The research methodology follows a structured workflow composed of four main phases:

1. **Data Collection:**
    - Identify and select reliable data sources relevant to football performance and market value.
    - Scrape and compile data from platforms such as Transfermarkt, FBref, and SoFIFA.
    - Clean and preprocess the data to build two structured datasets: one for player-level performance and another for team-level metrics.
2. **Data Analysis and Feature Engineering:**
    - Perform exploratory data analysis (EDA) using descriptive statistics and visualizations to understand key patterns, distributions, and outliers in the data.
    - Engineer features that capture individual player performance as well as contextual factors such as team strength, league level, and nationality.
    - Calculate player contribution scores across different lines (defense, midfield, attack) using normalized performance metrics.
    - Apply unsupervised clustering techniques to group players into tactical roles based on their performance profiles and identify distinct playing styles or specialties.
    - Analyze team-level performance data to detect gaps in tactical roles, revealing the most urgent recruitment needs of each team.
    - Identify correlations and trends between engineered attributes and market value to select the most relevant predictors for modeling.
3. **Predictive Model Development:**
    - Train multiple machine learning models to predict young players' market value using the engineered features.
    - Evaluate model performence using metrics such as RMSE, MAE, and R-squared.
    - Select the best-performing model to be integrated into the final decision-support framework.

4. **Optimization Model Development:**
    - Construct a pool of candidate players whose profiles allign with the specific needs of a target team, based on performance contributions and contextual factors.
    - Identify the tactical roles most needed by the team using similarity analysis between team deficiencies and role profiles derived from clustering.
    - Develop and solve an optimization model that selects the best combination of players from the candidate pool, maximizing on-field performance while satisfying constraints such as budget limits, age restrictions, and squad composition requirements.

By combining predictive analytics with optimization techniques, this framework enables football clubs to reduce overspending, uncover undervalued talent, and build squads strategically. The result is a **data-driven decision-making tool** designed to enhance transfer strategies and long-term club performance.

1.4 Document Structure

The structure of this document is as follows:

- **Chapter 1: Introduction** presents the motivation behind the study, its objectives, methodological workflow, and an overview of the thesis structure.
- **Chapter 2: State of the Art** reviews the relevant literature on football economics, valuation techniques, and machine learning applications in sports analytics.
- **Chapter 3: Data Collection and Analyses** explains how the datasets were built, cleaned, and analyzed. It also covers feature engineering, player contribution scoring, clustering of tactical roles, and the detection of team needs.
- **Chapter 4: Market Value Prediction Model** describes the modeling process for estimating player market value, including data preparation, algorithm selection, model evaluation, and separate models for different positions.
- **Chapter 5: Optimization of Player Acquisition Strategy** introduces the optimization model for squad building, including player pool construction, tactical role gap detection, and the Pyomo-based optimization setup with real constraints and decision variables.
- **Chapter 6: Discussion** interprets the results and findings from the predictive and optimization models, as well as the case study. It also addresses limitations and practical implications.

- **Chapter 7: Conclusion and Future Work** summarizes the research contributions and suggests potential directions for future development and refinement.

As described in the accompanying GitHub repository [5], all code and data are publicly available.

Chapter 2

State of the Art

The application of data science and machine learning in football analytics has grown significantly in recent years. Several studies have explored various techniques to evaluate player performance, predict market value, and optimize team transfer strategies. This chapter provides an overview of recent research and methodologies in the field.

2.1 Foundations of Football Economics and Valuation

2.1.1 Sports Economics vs Business Economics

Neale's contributions [6] were fundamental in establishing sports economics as a distinct field, separate from traditional business economics. His work clearly outlined how the principles guiding sports economics differ from those in mainstream economics, especially in the context of valuing players. Through a set of nine propositions, Neale showed that conventional business economic theories do not always apply to sports in the same way. A well-known illustration is the Louis-Schmelling paradox, which highlights that, unlike regular industries where firms aim to eliminate competition, sports teams rely on strong opponents to succeed. This mutual dependence contradicts standard economic assumptions about competition and market behavior.

Building on Neale's foundation, later studies (El-Hodiri & Quirk (1971) [7], Fort & Quirk (1995) [8], Kesenne (2000) [9], Scully (1974) [10], Sloane (1971) [11], Vrooman(2000) [11] further reinforced that sports economics operates under different assumptions than general economics. Key distinctions include the pursuit of non-profit goals [11], cooperative financial structures [8], and a focus on winning over maximizing profits [9].

2.1.2 Valuation Typologies

Gerrard [12] categorized the existing literature on football player valuation into two main approaches: *fundamental* and *comparative*. The terms fundamental and intrinsic are used interchangeably, as are comparative and relative. In the fundamental approach, a player's value is determined by internal factors such as on-field performance, individual skill, and contribution to team achievements. In contrast, the comparative approach evaluates a player's market value by comparing them with other players of similar profiles. Gerrard advocates for a combined use of both methods to more accurately reflect a player's overall value, taking into account both their actual performance and prevailing market dynamics. The fundamental approach aligns with corporate finance theory, typically utilizing discounted cash flow (DCF) analysis to estimate the present value of a player's expected future contributions.

Hill et al. [13] introduce a comprehensive typological framework for valuing football players by adapting Damodaran's corporate finance valuation methods to the football context. The authors categorize valuation approaches into four types: intrinsic, relative, real options, and probabilistic. Each typology is analyzed in terms of applicability, limitations, and data requirements, with particular attention to the challenges of data opacity in football—especially regarding wages and transfer fees. The study argues that while intrinsic valuations using discounted cash flow (DCF) and marginal revenue product (MRP) offer theoretical rigor, they are often impractical due to unavailable or unreliable financial data. Relative valuations, based on comparing standardized metrics across players, are found to be the most pragmatic under current data conditions. The paper also explores the potential of real options and probabilistic methods (e.g., machine learning and performance-based forecasting) for capturing future value and uncertainty in player performance. Ultimately, this work unifies disparate valuation literature and offers a decision-making framework to guide both researchers and practitioners in selecting the most appropriate valuation technique based on available data.

2.2 Machine Learning Approaches

2.2.1 Transfer Fee Modeling and Market Value

Mustafa A. Al-Asadi and Sakir Tasdemir[14] introduced a machine learning-based approach utilizing data from the FIFA20 video game, originally generated by EA Sports, to estimate market values for football players. Their results demonstrated that random forest models outperformed traditional regression techniques in estimating transfer fees, achieving an R^2 value of 0.95.

In a similar study, **V. B. Jishnu et al.**[15] used FIFA 22 video game data, splitting the study and prediction models according to player positions, dividing them into GK, $Defenders$, $Midfielders$, and $Attackers$. The results showed strong performance for

both Random Forest and Gradient Boost across all positions, with Gradient Boost yielding slightly better results.

Müller et al. [16] conducted a study utilizing an extensive dataset of 4,217 players from 146 teams across the top five European leagues (British, Spanish, French, Italian, and German) spanning six playing seasons. Their research showcased the effectiveness of multilevel regression models in estimating the market values of players. By comparing their model-based findings with crowd-sourced estimates, they highlighted that a data-driven approach can overcome many of the limitations associated with crowdsourcing, all while maintaining a high degree of accuracy. Given the growing availability of player data from commercial sources and user-generated content online, the study suggests that data analytics will become increasingly important in player recruitment and transfer negotiations within the football industry.

Majewski[17] investigated the factors influencing the market value of forward players by analyzing data from 150 well-known attacking players using the Generalized Least Squares (GLS) method. The study aimed to identify key features affecting players' market value. Conducted in two stages, it first selected potential variables and verified their significance, then focused on finding the best model for describing their relationship. The analysis confirmed a linear relationship, examining various estimation methods, including OLS, GLS, and FGLS with heteroscedasticity correction. Key findings indicated that the most significant variables impacting market value were Canadian classification points (goals and assists), club value adjusted by FIFA rankings, and player goodwill or brand value.

Shen et al. [18] utilized both FIFA 19 video game data and real-world statistical reports to try and predict players' market values. The study focused on analyzing various player attributes, including physical characteristics (height, weight, age), performance metrics (dribbling, finishing, shot power), and popularity indicators derived from social network activity. The dataset used in the study comprised 53 features for 491 players from the top five European leagues. After preprocessing a refined subset of 459 players was selected for analysis. The study applied various machine learning models, including Support Vector Regression (SVR), Random Forest Regression (RFR), CatBoost (CAT), and Extreme Gradient Boosting (XGB), along with hybrid ensemble models to enhance predictive accuracy. Performance evaluation revealed that SVR-based models exhibited the weakest predictive performance, whereas XGB-based models outperformed others. The study further demonstrated that the RSCX_SC ensemble model achieved the highest predictive accuracy, with an R^2 value of 0.9931, followed by XGSC (R^2 = 0.9917) and CASC (R^2 = 0.9830). The findings underscored the effectiveness of advanced hybrid ensemble models in improving market value estimation, with optimization techniques such as SCSO outperforming HGSO in improving prediction accuracy.

Poli et al.[19] created a statistical model to estimate football players' transfer fees, using data from approximately eight thousand transfers between 2014 and 2024 domestically and globally. Their regression analysis accounts for 84.8% of the variance in

transfer prices, identifying important factors such as the duration left on a player's contract, their level of experience, market dynamics, and performance metrics. Nevertheless, the study points out that some transfers can considerably deviate from the predicted values due to external factors like financial issues within clubs, the tendency of emerging markets to overpay for players, and competition by other teams.

Parida and Thilak (2022) [20] explored ML techniques to forecast player growth using real-world statistics from WhoScored.com [21]. Their study used models such as LSTM for time series prediction and XGBoost for performance estimation, aiming to identify emerging talents and suggest replacements for aging players. The study assessed model effectiveness using Mean Squared Error (MSE), with XGBoost achieving the lowest error and KNN performing the worst.

2.2.2 Player Selection and Optimization

Galaz-Cares [22] presents a comprehensive optimization framework for football squad planning that integrates sporting performance, financial constraints, and long-term strategic goals. Drawing on workforce planning concepts traditionally used in corporate HR, the study adapts these principles to the football context, where player heterogeneity, multi-season contracts, and dynamic transfer markets introduce unique complexities. The methodology consists of five interconnected modules, beginning with the extraction of player characteristics and the estimation of performance using advanced metrics such as On-Ball Value (OBV). These performance scores, along with market value estimates, are projected into future seasons using machine learning models. This predictive data serves as input to an Integer Programming model that selects optimal squad compositions while satisfying quality thresholds, financial limits, and regulatory requirements.

Boon and Sierksma [23] propose a set of linear optimization models for optimal team formation across various domains, with a particular focus on football (soccer) in their paper *Team formation: Matching quality supply and quality demand*. The authors begin by developing a basic optimization model for constructing football teams, where the transformation of position-based and player-based scoring tables into parameter values serves as the core input for model formulation. While the approach is demonstrated in the context of football, it is generalizable to a range of team-based settings including American football, hockey, rowing, and even project or management teams.

The paper introduces the concept of the *Computer Coach*, a decision support system designed to aid scouting and recruitment by aligning individual qualities of players with the structural and functional requirements of the team. A key insight of the paper is the contrast between sports like soccer—where positions are relatively fixed—and sports such as volleyball, where the absence of static positions introduces additional complexity into the optimization process. The volleyball model presented in the second half of the paper reflects this increased difficulty.

The authors emphasize that while the Computer Coach system can offer valuable

recommendations, it is intended to supplement rather than replace expert judgment from coaching staff. Despite resistance to such systems in some parts of the sports world—particularly with regard to incorporating metrics like absenteeism or player health—there is growing interest in using decision support tools to manage human capital more efficiently. This interest is driven in part by increasing access to computational tools and the adoption of mathematical modeling techniques within professional sports.

Onwuachu and Enyindah [24] present a machine learning-based decision support system aimed at assisting football managers in selecting suitable players based on performance attributes. The study employs artificial neural networks to process a wide range of player characteristics and generate performance evaluations. The system aggregates player capabilities into four major categories to support clearer comparisons and selection decisions by team managers.

The authors validate their model by comparing system-generated results with real-world online performance data collected on December 16, 2014, for prominent players including Franck Ribéry, Cristiano Ronaldo, Andrés Iniesta, Arjen Robben, Lionel Messi, Luis Suárez, and Eden Hazard. Line graphs illustrate the alignment between predicted and actual performances, demonstrating that the neural network model achieves a close approximation with minimal error.

2.2.3 Expected Goals (xG) Models and Interpretability

Mead et al. [25] present a comprehensive study aimed at enhancing expected goals (xG) models in football by incorporating previously underexplored features such as player ability, team quality, and psychological effects. Utilizing event-level data from the top five European leagues (sourced from Wyscout), the authors apply various machine learning techniques—including XGBoost, AdaBoost, logistic regression, and neural networks—to model shot outcomes probabilistically. Their study reveals that integrating contextual variables like match importance, team Elo ratings, and player market value can significantly improve the predictive power of xG models. Among the findings, distance to goal and shot angle remain dominant predictors, but new variables such as player PlayeRank scores and team transfer spending also show substantial influence. The authors validate their models against traditional football metrics and existing xG systems, demonstrating superior performance in predicting match outcomes. This work not only provides methodological advancements in football analytics but also strengthens the practical utility of xG in decision-making processes for clubs and analysts.

Cavus and Biecek [26] propose an advanced expected goals (xG) model focused on interpretability and accuracy by leveraging explainable artificial intelligence (XAI) tools. Using over 315,000 shot events from the top-five European football leagues across seven seasons (2014–2021), the authors train various machine learning models including Random Forests, CatBoost, XGBoost, and LightGBM. Their methodology tackles class imbalance using oversampling techniques and emphasizes model evaluation using both

conventional metrics and balanced accuracy measures. A key innovation lies in the introduction of Aggregated Ceteris-Paribus (CP) profiles to interpret xG predictions at the team and player level. These profiles allow practitioners to analyze how shot characteristics like angle and distance impact scoring probabilities, and to simulate performance improvements through tactical adjustments. This work significantly contributes to football analytics by combining high-performing black-box models with transparent evaluation frameworks, aiding both decision-makers and analysts in performance assessments.

2.3 Key Machine Learning Models

Gradient Boosting is an ensemble machine learning technique introduced by **Friedma** [27] that builds a strong predictive model by sequentially combining multiple weak learners, typically decision trees. Each subsequent model is trained to correct the residual errors of the previous one by minimizing a specified loss function using gradient descent in function space. This iterative method enables the algorithm to focus on difficult-to-predict samples, leading to high performance on structured data. Gradient Boosting has become foundational in modern machine learning, with implementations such as XGBoost, LightGBM, and CatBoost widely used in practice due to their speed and accuracy.

Random Forest, proposed by **Breiman** [28], is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode (for classification) or mean prediction (for regression) of the individual trees. The model introduces randomness by using bootstrap sampling and random feature selection at each split, which reduces variance and prevents overfitting. Random Forests are known for their robustness, high accuracy, and ability to handle large feature spaces and missing data. They are widely applied across domains for classification, regression, and feature importance analysis.

Linear Regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The approach assumes a linear relationship and is widely used due to its interpretability and computational efficiency. As explained in **Draper and Smith** [29], the model estimates coefficients by minimizing the sum of squared residuals, providing a best-fit line. Despite its simplicity, linear regression serves as a foundational technique for more advanced models and remains a key tool in econometrics, forecasting, and scientific research.

2.4 Summary

This chapter has provided a comprehensive overview of the current landscape in football analytics, particularly focusing on how data science and machine learning are applied to evaluate players and predict their market values.

We began by exploring the fundamental principles of football economics and valuation. It's crucial to understand that sports economics operates differently from traditional business economics. As Neale and others have shown, unlike typical industries where companies aim to eliminate competition, sports teams often rely on strong opponents for their own success (the Louis-Schmelling paradox). This unique dynamic, along with differing goals like focusing on winning over purely maximizing profits, shapes how players are valued. We then looked at various valuation approaches, categorizing them as either fundamental (intrinsic) or comparative (relative). While intrinsic methods, such as using discounted cash flow, offer a strong theoretical basis, they can be impractical due to the difficulty in accessing reliable financial data like wages and transfer fees. Because of this, comparative approaches, which assess a player's value by comparing them to similar players, are often more practical with the data currently available. Hill et al. also presented a broader framework, including probabilistic methods, emphasizing the need for flexible valuation techniques given data limitations.

Next, the chapter reviewed Machine Learning approaches to player valuation. Early models demonstrated the effectiveness of techniques like Random Forest and Gradient Boosting, often using data from video games like FIFA. More recent and advanced studies, utilizing extensive real-world datasets, have shown that sophisticated models, including multilevel regression and hybrid ensemble approaches (like XGBoost and CatBoost), can achieve high accuracy in predicting player market values. Our review also touched on transfer fee modeling, identifying key factors such as contract length, player experience, and market trends that influence transfer prices, while acknowledging that external factors can also cause deviations. We also examined the evolution of Expected Goals (xG) models, which have become more advanced by incorporating contextual details and using Explainable AI (XAI) tools to offer clearer insights into shot outcomes and tactical analysis. Finally, the chapter provided a brief introduction to commonly used Machine Learning models like Gradient Boosting, Random Forest, and Linear Regression, noting their strengths in prediction and interpretability.

In conclusion, this review confirms that significant progress has been made in using machine learning for football analytics, especially in player valuation and performance assessment. Current research provides strong predictive models and a solid understanding of the many factors that influence player market values.

However, our analysis of the current literature also points to several key areas that need further attention and will be the focus of the upcoming chapters:

- While various data sources have been used, there's often a lack of detailed explanation on how comprehensive, real-world datasets are collected, integrated, and

processed to support both precise player valuation and strategic decision-making. Our study will explicitly address this.

- Although some research segments models by player position, there's still room for more rigorous comparison and refinement of machine learning models for each distinct playing position. This will involve using a rich, consolidated dataset to capture the unique performance aspects of different roles.
- Most importantly, while existing work excels at predicting player values and performance, it often doesn't fully bridge the gap to actionable strategic recommendations for clubs. There's a clear need for developing integrated optimization frameworks that can translate these predictions into practical player acquisition strategies. This means considering a club's specific budget, squad needs, and how new players might impact broader team performance and financial goals, such as "team scores" or overall club success.

These identified areas—comprehensive data handling, position-specific model refinement, and, critically, the development of an integrated decision-support framework for strategic player acquisition—will form the primary contributions and focus of the subsequent chapters of this research.

Chapter 3

Data Collection and Analyses

3.1 Data Sources

This study uses three primary data sources:

- Transfermarkt [3],
- FBref [30],
- and SoFIFA [31].

These platforms were selected due to their extensive databases, historical coverage, and established reliability in football analytics. Many previous studies have leveraged these sources, demonstrating their credibility in providing accurate and comprehensive football data [14] [18].

3.1.1 FBref

FBref [30] is an advanced statistical platform dedicated to football analytics. It is powered by Sports Reference [32], a well-known provider of sports statistics that covers multiple sports, including basketball, baseball, American football, and hockey. FBref specializes in football data, offering an extensive range of statistics for individual players and teams, encompassing domestic leagues, international tournaments, and club competitions around the world.

One of the primary strengths of the platform is its integration of advanced analytics, made possible through its collaboration with Opta [33], a leading provider of sports data and performance analysis. FBref includes key performance metrics such as Expected Goals (xG) and Expected Assists (xA), which provide deeper insights into shot quality and passing effectiveness. Additionally, the platform offers detailed progressive passing and dribbling statistics, which are essential for evaluating a player's influence on ball

progression and attacking movements. Other advanced metrics, such as pressures, tackles, and interceptions, provide crucial insights into defensive contributions.

FBref is widely utilized by football analysts, researchers, coaches, and scouts who depend on data-driven insights for evaluating player performance, understanding team tactics, and refining strategic decision-making. The platform's ability to break down player and team performances through statistical models allows for more informed assessments and comparisons.

With data covering competitions across more than 230 nations worldwide, FBref features historical and real-time statistics on thousands of players spanning multiple years. Its comprehensive dataset includes match-by-match breakdowns, seasonal summaries, and in-depth scouting reports. The extensive and detailed nature of FBref's statistics makes it an invaluable resource for gathering real-world and analytical data, supporting this study by providing objective, quantifiable measures of player and team performance.

3.1.2 SoFIFA

SoFIFA [31] is a platform that aggregates and presents ratings and attributes of football players derived from the **EA Sports FIFA** series [34]. EA Sports employs a network of real-life scouts, analysts, and data reviewers who assess player performance based on real-world matches.

Sports journalist Lukas Scherbaum, in his article **FIFA Ratings Explained: How Does A Player Get His OVR** [35], states that a player's in-game statistics determine their overall rating. He explains:

> ***"There are more than 30 leagues, more than 700 clubs, and more than 19,000 players each year. And every single one of these players has to be scouted in order to get fair stats. Even players who have been promoted to the third tier (or even lower, if you look at England) have to be watched in order to determine the player's pace, dribbling, and passing ability. Of course, this can only be done with an extensive scouting team."***

According to Scherbaum, the FIFA scouting team consists of approximately 30 producers and more than 400 freelancers who scout players for EA. Additionally, more than 6,000 data reviewers contribute to the team by providing assessments and insights.

Their evaluations, combined with algorithmic processing, generate detailed ratings for various player attributes. This dataset serves as a valuable resource for analyzing player performance and transfer value, as it reflects both subjective expert assessments and data-driven metrics. The dataset includes over 30 attributes per player, categorized into physical (e.g., strength, speed, height), technical (e.g., passing, shooting, defending), and mental (e.g., aggression, vision, composure). The publicly available nature of this data makes it an essential source for research and analytical studies in football performance evaluation.

3.1.3 Transfermark

Transfermarkt [3] is one of the most comprehensive football databases, primarily known for its extensive coverage of player market values, transfer history, and contract details. Established as a community-driven platform, it has grown into a widely trusted source of football-related information, frequently referenced by analysts, clubs, and media outlets. The platform covers a vast range of data, including player statistics, team performances, competition overviews, and historical transfer records.

One of the key features of Transfermarkt is its player market valuation system, which estimates the potential transfer value of footballers based on various factors such as performance, age, contract duration, positional demand, and market trends. Although these valuations are not official, they provide valuable insights into a player's perceived worth in the footballing economy.

In addition to player valuations, Transfermarkt provides detailed transfer histories, tracking completed, rumored, and record-breaking transfers across global leagues. The platform also offers statistics on contract expirations, agent affiliations, and salary information, making it a vital tool for scouts, journalists, and football enthusiasts interested in transfer market dynamics.

Transfermarkt's database spans professional leagues, national teams, and youth academies, covering thousands of players across different tiers of football. It also includes managerial records, club financial data, and squad value assessments, giving a holistic view of team-building strategies. The platform is widely utilized by analysts, researchers, and clubs seeking data-driven insights into player recruitment and squad planning.

With its extensive coverage of leagues worldwide, Transfermarkt plays a crucial role in football analytics by providing real-time updates on transfers, player performances, and squad valuations. Its robust dataset makes it an invaluable resource for this study, offering detailed insights into market trends, player movements, and financial aspects of football.

3.2 Data Collection and Extraction

This study involves the extraction of data from three primary sources: FBref [30], SoFIFA [31], and TransferMarkt [3]. The goal of this extraction process is to construct a custom dataset collected and formed specifically for this study, ensuring that the collected data aligns with the research objectives. This dataset serves as the foundation for further analyses and models development.

Each of these sources has a distinct website structure, data format, and type of information available, forcing the use of different web scraping techniques to ensure efficient and accurate data retrieval. The extraction process required careful consideration of dynamic content, HTML parsing complexities, and data consistency across sources.

By implementing appropriate scraping methodologies, we ensured that the gathered data was comprehensive, structured, and suitable for subsequent analysis.

3.2.1 Data Scraping and Data Scraping Techniques

The extraction process employed a combination of Python-based web scraping techniques, leveraging both automated browsing and direct HTTP request methods to access and structure the data.

FBref and SoFIFA Extraction

For FBref and SoFIFA, the primary scraping method involved `Selenium` [36], allowing for automated interaction with dynamic webpage elements. The workflow included:

- **Configuring Suitable Links for Leagues and Seasons:** Data extraction began by identifying and setting appropriate URLs corresponding to the leagues and seasons relevant to this study. Rather than relying on dynamically generated requests, predefined links were structured to ensure direct access to pages containing the required player statistics. This method allowed for efficient data retrieval while maintaining focus on relevant leagues and timeframes.
- **Locating and Identifying Relevant Data Sections:** Once the appropriate links were configured, web pages were examined to determine the exact locations of tables, features, and statistics. A detailed analysis of the webpage structure helped in identifying relevant data elements, ensuring that only necessary information aligned with the study's objectives was considered.
- **Extracting and Filtering Data:** After pinpointing the relevant sections, the data was extracted with necessary filters applied to retain only records within the scope of the study. Players older than 23 years were excluded for the prediction model data , and leagues or seasons outside the predefined range were omitted. Additional preprocessing steps addressed inconsistencies, including missing values, formatting variations, and duplicate entries, to enhance data quality.
- **Structuring the Extracted Data into a Tabular Format:** The extracted data was then organized into a structured tabular format to facilitate further analysis. The finalized dataset was stored in CSV format.

TransferMarkt Extraction

Unlike FBref and SoFIFA, TransferMarkt's website posed challenges due to its intricate HTML structure. Instead of `Selenium`, we primarily used the `requests` library [37] in combination with `lxml` to parse the HTML content. Key steps in this extraction process included:

- **Setting Links for Data Extraction:** The process began by configuring the appropriate web links for each league and season, ensuring direct access to relevant data sources.
- **Identifying XPath Selectors and Webpage Structure:** Before extraction, the webpage structure was analyzed to determine the correct XPath expressions for locating key elements such as player names, market values, and corresponding clubs.
- **Looping Through Team Links and Extracting Data:** For each league, the script iterated through team links, accessing individual team pages on TransferMarkt to extract detailed player information, ensuring full coverage of all players in a league.
- **Saving Extracted Data to a CSV File:** Once extracted, the data was processed, structured, and stored in a CSV format for further analysis, ensuring consistency and usability across different modeling and analytical tasks.

3.2.2 Overview of Collected Data

The dataset comprises player data for athletes aged 23 and younger from multiple leagues. Specifically, we focused on:

- The top five European leagues: Premier League (UK), LaLiga (Spain), Ligue 1 (France), Serie A (Italy), and Bundesliga (Germany).
- Their corresponding second divisions: Championship (UK), LaLiga 2 (Spain), Ligue 2 (France), Serie B (Italy), and Bundesliga 2 (Germany).
- Additional leagues: Belgian Pro League (Belgium), Eredivisie (Netherlands), Primeira Liga (Portugal), Brazil Série A (Brazil), and Liga Profesional Argentina (Argentina).

The dataset spans four seasons, from 2020/2021 to 2023/2024, covering a total of 60 league seasons.

In addition to individual player data, we also collected squad-level data for teams in the top five European leagues. This aggregated team data was used to calculate tactical and statistical needs, which were then matched against player profiles in the optimization model.

3.2.3 FBref Data

FBref [30] provides a comprehensive set of player performance metrics, covering general player attributes, match statistics, disciplinary records, and advanced analytics. The

availability of certain features varies depending on the league's importance, with major leagues offering more detailed performance insights.

The fundamental attributes available for all players across leagues include:

- **General Information**: Player, Nation, Position, Squad, Age, Year of Birth.
- **Match Participation**: Matches Played, Starts, Minutes, 90s Played.
- **Basic Performance Metrics**: Goals, Assists, Goals + Assists, Non-Penalty Goals, Penalty Kicks Made, Penalty Kicks Attempted.
- **Disciplinary Data**: Yellow Cards, Red Cards.

Beyond the general attributes, outfield players [1] have access to additional performance-based features, depending on the level of coverage available for their league:

- **Per 90-Minute Metrics**: Goals/90, Assists/90, Goals + Assists/90, Non-Penalty Goals/90, Non-Penalty Goals + Assists/90.
- **Advanced Analytics (Major Leagues Only)**:
  - **Expected Goals (xG)**: A measure of shot quality and goal probability.
  - **Non-Penalty Expected Goals (npxG)**: xG excluding penalty kicks.
  - **Expected Assisted Goals (xAG)**: The likelihood of an assist leading to a goal.
  - **xG + xAG**: A combined metric of goal-scoring and assisting potential.
  - **Progressive Metrics**: Progressive Carries, Progressive Passes, Progressive Passes Received.

For defensive players, additional defensive metrics are recorded, which vary in availability depending on league coverage:

- **Defensive Actions**: Tackles, Tackles Won, Tackles in Defensive/Midfield/Attacking Thirds.
- **Dribbler Engagements**: Dribblers Tackled, Dribbles Challenged, % of Dribblers Tackled, Challenges Lost.
- **Blocking and Interceptions**: Blocks, Shots Blocked, Passes Blocked, Interceptions, Tackles + Interceptions (Tkl+Int).
- **Clearances and Errors**: Clearances, Errors Leading to Goals or Shots.

[1] An outfield player refers to any football (soccer) player who is not a goalkeeper. This includes defenders, midfielders, and forwards.

For goalkeepers, FBref provides specialized statistics tailored to their role:

- **General Goalkeeping Data**: Matches Played, Starts, Minutes, 90s Played.
- **Shot-Stopping Ability**: Goals Against, Goals Against/90, Shots on Target Against, Saves, Save Percentage.
- **Match Outcomes**: Wins, Draws, Losses, Clean Sheets, Clean Sheet Percentage.
- **Penalty Performance**: Penalty Kicks Attempted, Penalty Kicks Allowed, Penalty Kicks Saved, Penalty Kicks Missed, Save Percentage for Penalty Kicks.

It is important to note that not all leagues provide the same level of data coverage.

- **Major European Leagues (Premier League, La Liga, Bundesliga, Serie A, Ligue 1)**: These leagues have the most extensive data coverage, including advanced analytics such as xG, xAG, progressive passing metrics, and detailed defensive breakdowns.
- **Lower Divisions and Other Leagues**: Secondary leagues, including La Liga 2, Serie B, and leagues outside Europe's top five, have more limited data availability. These leagues primarily provide basic player information, match participation stats, and general performance metrics, but lack detailed advanced analytics.

This variability in feature availability reflects the depth of statistical analysis provided for each league, with higher-profile competitions receiving more comprehensive data tracking.

A set of examples of the data extracted :

TABLE 3.1. FBREF PERFORMANCE STATISTICS FOR A GOALKEEPER: ILLAN MESLIER (2023/2024)

Attribute	**Value**	**Attribute**	**Value**
Player	Illan Meslier	Matches Played	34
Nation	FRA	Starts	34
Position	GK	Minutes	3,060
Squad	Leeds United	90s Played	34.0
Age	22	Goals Against	67
Year of Birth	2000	Goals Against/90	1.97
Shots on Target Against	158	Saves	92
Save Percentage	59.5%	Clean Sheets	5
Clean Sheet %	14.7%	Wins	7
Draws	9	Losses	18
Penalty Kicks Attempted	3	Penalty Kicks Allowed	3
Penalty Kicks Saved	0	Penalty Kicks Missed	0
Save% (Penalty Kicks)	0.0%		

TABLE 3.2. FBREF PERFORMANCE STATISTICS FOR A DEFENDER: JOŁKO GVARDIOL (2023/2024)

Attribute	**Value**	**Attribute**	**Value**
Player	Joško Gvardiol	Dribblers Tackled	21
Nation	CRO	Dribbles Challenged	38
Position	DF	% Dribblers Tackled	55.3%
Squad	Manchester City	Challenges Lost	17
Age	21	Blocks	36
Year of Birth	2002	Shots Blocked	12
90s Played	25.9	Passes Blocked	24
Tackles	55	Interceptions	27
Tackles Won	33	Tackles + Interceptions	82
Tackles (Def 3rd)	21	Clearances	37
Tackles (Mid 3rd)	24	Errors	3
Tackles (Att 3rd)	10		

TABLE 3.3. FBREF PERFORMANCE STATISTICS FOR A MIDFIELDER:JUDE BELLINGHAM (2023/2024)

Attribute	**Value**	**Attribute**	**Value**
Player	Jude Bellingham	Penalty Kicks Attempted	1
Nation	ENG	Yellow Cards	5
Position	MF	Red Cards	1
Squad	Real Madrid	xG (Expected Goals)	11.1
Age	20	npxG (Non-Penalty xG)	10.3
Year of Birth	2003	xAG (Expected Assisted Goals)	5.3
Matches Played	28	npxG + xAG	15.6
Starts	27	Progressive Carries	85
Minutes	2,315	Progressive Passes	196
90s Played	25.7	Progressive Passes Rec	177
Goals	19	Goals/90	0.74
Assists	6	Assists/90	0.23
Goals + Assists	25	Goals + Assists/90	0.97
Non-Penalty Goals	18	Non-Penalty Goals/90	0.70
Penalty Kicks Made	1	Non-Penalty Goals + Assists/90	0.93
		xG/90	0.43
		xAG/90	0.21
		xG + xAG/90	0.64
		npxG/90	0.40
		npxG + xAG/90	0.61

TABLE 3.4. FBREF PERFORMANCE STATISTICS FOR A FORWARD: NICO WILLIAMS (2023/2024)

Attribute	**Value**	**Attribute**	**Value**
Player	Nico Williams	Penalty Kicks Attempted	0
Nation	ESP	Yellow Cards	6
Position	FW	Red Cards	1
Squad	Athletic Club	xG (Expected Goals)	6.0
Age	21	npxG (Non-Penalty xG)	6.0
Year of Birth	2002	xAG (Expected Assisted Goals)	5.7
Matches Played	31	npxG + xAG	11.7
Starts	29	Progressive Carries	143
Minutes	2,263	Progressive Passes	66
90s Played	25.1	Progressive Passes Rec	302
Goals	5	Goals/90	0.20
Assists	11	Assists/90	0.44
Goals + Assists	16	Goals + Assists/90	0.64
Non-Penalty Goals	5	Non-Penalty Goals/90	0.20
Penalty Kicks Made	0	Non-Penalty Goals + Assists/90	0.64
		xG/90	0.24
		xAG/90	0.22
		xG + xAG/90	0.46
		npxG/90	0.24
		npxG + xAG/90	0.46

In addition to player-level data, the same features were also computed at the squad level, enabling a direct comparison between team needs and individual player contributions. These team-level values were extracted using aggregated statistics, allowing for a more accurate assessment of how well a player's profile fits the tactical and statistical gaps of a target team.

The player pool used for optimization includes footballers from all age groups. However, the model allows for flexible filtering, such as restricting candidates by age (e.g., under 23), to suit different transfer strategies or club philosophies (e.g., youth development vs. experienced reinforcements).

TABLE 3.5. TEAM-LEVEL SUMMARY STATISTICS FOR REAL MADRID (2023/2024)

Attribute	**Value**	**Attribute**	**Value**
Season	2023–2024	League	La Liga
Squad	Real Madrid	League Rank	1
Matches Played	38	Wins / Draws / Losses	29 / 8 / 1
Goals For (GF)	87	Goals Against (GA)	26
Goal Difference (GD)	+61	Points	95
Possession %	59.2	Avg Attendance	72,061
Shooting		**Passing**	
Goals Scored	85	Passes Attempted	24,691
Shots (Sh)	593	Pass Completion %	88.3
Shots on Target (SoT)	242	Progressive Pass Distance	112,682
xG	68.8	xA (Expected Assists)	52.5
npxG (Non-Penalty xG)	65.2	Progressive Passes	1935
Defending		**Goalkeeping**	
Tackles (Tkl)	557	Goals Against (GA)	26
Interceptions (Int)	297	Saves	93
Blocks	401	Save %	78.8
Clearances (Clr)	853	Clean Sheets (CS)	21
Defensive Errors	14	CS %	55.3

3.2.4 SoFIFA Data

The extracted data from SoFIFA [31] includes a large set of player attributes and performance metrics that are consistently available across all leagues. These features represent player characteristics.The fundamental characteristics that define each player include:

- **Age**
- **Overall Rating**: A numerical representation of the player's overall skill level.
- **Potential**: Estimated peak performance capability.
- **Height & Weight**
- **Preferred Foot**
- **Best Overall Rating**: The highest possible rating based on position.
- **Best Position**: The most suitable playing position.
- **Growth**: Expected improvement in the player's abilities over time.

Beyond general attributes, SoFIFA provides a detailed breakdown of player abilities, categorized into different skill domains based on EA Sports FIFA [34]. These categories include:

- **Attacking**: Total Attacking, Crossing, Finishing, Heading Accuracy, Short Passing, Volleys.
- **Skill**: Total Skill, Dribbling, Curve, Free Kick Accuracy, Long Passing, Ball Control.
- **Movement**: Total Movement, Acceleration, Sprint Speed, Agility, Reactions, Balance.
- **Power**: Total Power, Shot Power, Jumping, Stamina, Strength, Long Shots.
- **Mentality**: Total Mentality, Aggression, Interceptions, Attacking Positioning, Vision, Penalties.
- **Defending**: Total Defending, Defensive Awareness, Standing Tackle, Sliding Tackle.
- **Goalkeeping**: Total Goalkeeping, GK Diving, GK Handling, GK Kicking, GK Positioning, GK Reflexes.
- **Additional Attributes**: Weak Foot, Attacking Work Rate, Defensive Work Rate, International Reputation.

All of these attributes are consistently available for every player across all leagues in the SoFIFA dataset. The completeness and accuracy of these statistics are maintained through SoFIFA's data collection process, which is based on EA Sports' scouting network and player evaluations. This uniform availability ensures that all players can be analyzed comprehensively, regardless of the league in which they compete.

An example of the data extracted :

TABLE 3.6. SOFIFA STATS FOR PEDRI (2023/2024)

Pedri			
Name	Pedri	Agility	88
Age	20	Reactions	87
Overall Rating	86	Balance	90
Potential	92	Total Power	369
Height	174 cm	Shot Power	68
Weight	60 kg	Jumping	73
Foot	Right	Stamina	87
Best Overall	88	Strength	73
Best Position	MF	Long Shots	68
Growth	6	Total Mentality	355
Value[2]	€105M	Aggression	62
Wage[2]	€165K	Interceptions	73
Total Attacking	334	Att. Position	79
Crossing	68	Vision	88
Finishing	72	Penalties	53
Heading Accuracy	50	Total Defending	210
Short Passing	88	Def. Awareness	68
Volleys	56	Standing Tackle	77
Total Skill	402	Sliding Tackle	65
Dribbling	87	Total Goalkeeping	46
Curve	79	GK Diving	12
FK Accuracy	62	GK Handling	7
Long Passing	86	GK Kicking	11
Ball Control	88	GK Positioning	8
Total Movement	422	GK Reflexes	8
Acceleration	81	Weak Foot	4
Sprint Speed	76	Attacking Work Rate	High
International Reputation	3	Defensive Work Rate	High

3.2.5 TransferMarkt Data

The dataset extracted from TransferMarkt [3] includes essential market value attributes such as:

- Player name and club.
- Estimated market value in euros.

This kind of data was collected for all the players in all the leagues under our scope for the past four seasons. TransferMarkt's valuations are influenced by multiple factors, including player performance, contract status, and positional demand. The inclusion of these valuations provides critical insights into market trends and player worth over time.

A set of examples of the data extracted :

TABLE 3.7. TRANSFERMARKT DATA EXTRACTED FOR THE EXAMPLE PLAYERS (2023/2024)

Club	**Player**	**Market Value (€)**
Manchester City	Joško Gvardiol	€75,000,000
FC Barcelona	Pedri	€80,000,000
Real Madrid	Jude Bellingham	€180,000,000
Athletic Bilbao	Nico Williams	€60,000,000
Leeds United	Illan Meslier	€18,000,000

3.3 Dataset Preparation and Market Value Adjustments

The data that was collected from multiple sources as shown before is merged into several datasets. This process was crucial for performing detailed analysis, uncovering patterns, and finding important connections between player characteristics, performance data, and market values. This helped meet the research goals.

3.3.1 Market Value Inflation Adjustment

The dataset was initially organized by position and season to facilitate the study and analysis of patterns across different time periods. This preliminary structuring enabled an in-depth investigation of variations and trends specific to each position throughout the seasons.

Subsequently, the data from all seasons were consolidated by position to provide a holistic view. However, a significant challenge emerged due to economic inflation

affecting the football market. This inflation results in a continuous increase in player market values year after year, complicating direct comparisons across seasons.

To mitigate the impact of inflation and ensure meaningful comparisons, market values were adjusted using specific inflation rates recommended by the CIES Football Observatory Monthly Report No. 82 (February 2023) [38]. According to this report, the annual inflation rates for different player positions are:

TABLE 3.8. ANNUAL INFLATION RATES BY PLAYER POSITION
[38]

Position	**Inflation Rate (% per year)**
Goalkeeper (GK)	5.2%
Defender (DF)	12.5%
Midfielder (MF)	8.5%
Forward (FW)	8.2%

Adjusted market values were calculated using the exponential growth formula to accurately account for annual compounding inflation.

$$\text{Current Value} = \text{Original Value} \times (1 + r)^n \tag{3.1}$$

where:

- Current Value is the inflation-adjusted market value.
- Original Value is the original market value from previous seasons.
- r is the annual inflation rate (expressed in decimal form).
- n is the number of years that have elapsed since the original valuation.

After performing the calculations, adjusted market values were rounded to ensure clarity and readability, making the data easier to interpret and apply practically.

For example, consider the market value evolution of Rafael Leão, a Forward (FW), adjusted to the base season of 2023/2024, using the provided exponential growth formula:

$$\text{Adjusted Value} = \text{Original Value} \times (1 + r)^n$$

With an annual inflation rate for forwards (FW) at 8.2% (Table 3.8), the adjusted market values based on original market valuations are calculated as follows:

- **2020/2021 season (21 years old)**: Original Value = €25,000,000, multiplier = 3

$$25{,}000{,}000 \times (1 + 0.082)^3 \approx 32{,}000{,}000$$

- **2021/2022 season (22 years old)**: Original Value = €70,000,000, multiplier = 2

$$70{,}000{,}000 \times (1 + 0.082)^2 \approx 82{,}000{,}000$$

- **2022/2023 season (23 years old)**: Original Value = €90,000,000, multiplier = 1

$$90{,}000{,}000 \times (1 + 0.082)^1 \approx 97{,}000{,}000$$

The rounded adjusted market values are summarized clearly in the following table:

TABLE 3.9. INFLATION-ADJUSTED MARKET VALUES FOR RAFAEL LEÃO

Season	Age	Original Value (€)	Adjusted Value (€)
2020/2021	21	25,000,000	32,000,000
2021/2022	22	70,000,000	82,000,000
2022/2023	23	90,000,000	97,000,000

This method ensures practical interpretation by clearly presenting inflation-adjusted market values across seasons.

3.4 Dataset Description and Analysis

After scaling the players' market values according to their respective positions and seasons using defined inflation rates (Table 3.8), datasets were systematically generated for each player position. Each **position-specific dataset** compiles all players from the corresponding position across all seasons, resulting in four distinct datasets: **Goalkeepers (GK)**, **Defenders (DF)**, **Midfielders (MF)**, and **Forwards (FW)** each including the features from all data sources.

Additionally, a **unified comprehensive dataset** combining all positions across all seasons was created. This dataset shows the scaled market values and key features that are in common and of a high and equal importance for all positions and significant for comparative analysis.

The final datasets prepared for analysis are:

- **Position-specific datasets**: GK, DF, MF, and FW, each containing adjusted market values for all seasons.
- **Unified comprehensive dataset**: All players across all positions and seasons with essential attributes.
- **Teams dataset**: a data set formed of all the teams from the top-5 European Leagues over the last 4 seasons including a set of performance features and statistic.

These structured datasets enable detailed, robust, and clear analyses across positions and time periods, enhancing the interpretability of research findings.

3.4.1 Goalkeeper (GK) Dataset

The Goalkeeper (GK) dataset contains scaled market values and relevant performance attributes for goalkeepers across multiple seasons (around **400** players). Given the distinct role of goalkeepers compared to outfield[3]. Players, specific features such as **save percentage, clean sheets, reactions**, and other features play a crucial role in evaluating their market value. This dataset allows for a position-specific analysis of factors influencing a young goalkeeper's market value.

Key Features: To ensure a comprehensive analysis of goalkeeper market value, the dataset includes the following key attributes:

- **General Information:** Nation, Position, League, Squad, Age, Year of Birth, Market Value, and Season.
- **Performance Metrics:** Matches Played, Starts, Minutes, 90s Played.
- **Goalkeeping-Specific Metrics:** Goals Against, Goals Against/90, Shots on Target Against, Saves, Save Percentage, Wins, Draws, Losses, Clean Sheets, Clean Sheet Percentage, Penalty Kicks Attempted, Penalty Kicks Allowed, Penalty Kicks Saved, Penalty Kicks Missed, Save
- **Physical and Technical Attributes (FIFA EA Sports [34]):** Height, Weight, Foot, Best Overall, Best Position, Reflexes, Diving, Handling, Positioning, Reactions.
- **Valuation ((FIFA EA Sports [34])):** Value [4], Wage [4], Growth, Source League, International Reputation.
- **Additional Performance Indicators(FIFA EA Sports [34]) :** Total Goalkeeping (GK Diving, GK Handling, GK Kicking, GK Positioning, GK Reflexes), Total Movement (Acceleration, Sprint Speed, Agility, Balance), and Total Power (Shot Power, Jumping, Stamina, Strength).

Visual Analysis: To provide deeper insight into the valuation of the goalkeeper market, we include several visualizations.

[3]an outfield player refers to any football (soccer) player who is not a goalkeeper. This includes defenders, midfielders, and forwards

[4]This value and wage are the values used in the EASports FIFA video game, not the real life values

- **Age Distribution:** A pie chart illustrating the distribution of goalkeeper ages.

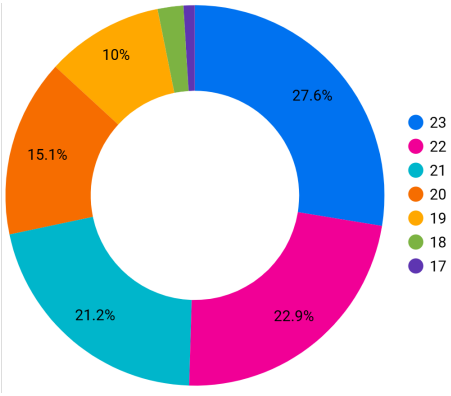


Fig. 3.1. Pie chart showing the age distribution of goalkeepers.

- **Market Value by Age:** A box plot comparing market values across different Ages.

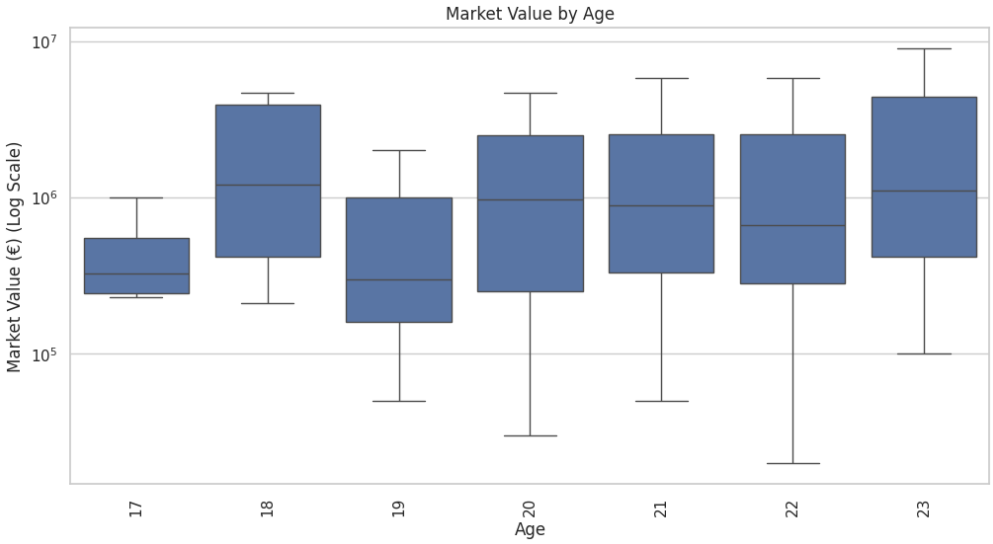


Fig. 3.2. Box plot of market value distribution across leagues.

The distribution of goalkeepers by age, as shown in Figure 3.1, reveals that the majority of goalkeepers in the dataset are aged **21 to 23**, with **23-year-olds** making up the largest share (27.6%). This indicates that the dataset primarily focuses on players approaching their peak developmental stage. Figure 3.2 illustrates the relationship between market value and age, showing a general trend of increasing market value with age. The median market value progressively rises from **17 to 23 years old**, suggesting

that as goalkeepers gain experience, their valuation improves. However, the interquartile range expands at older ages, particularly from **19 to 23**, indicating greater variability in valuation. Notably, **18-year-olds** show a temporary spike in median market value, possibly due to the early recognition of top talents. The overall trend highlights that while age positively correlates with market value, individual player growth and league exposure significantly impact valuation.

- **Feature Correlation:**

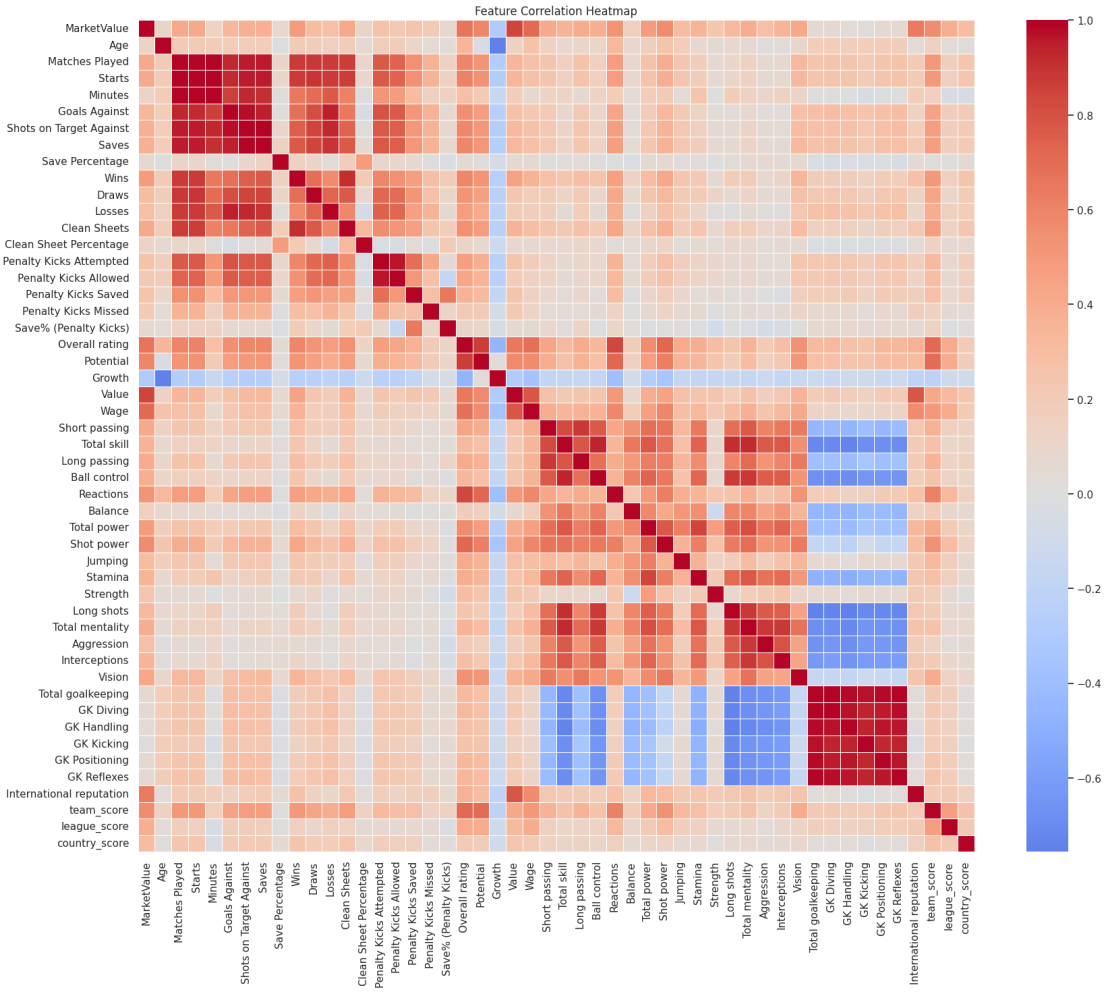


Fig. 3.3. Feature correlation heatmap showing relationships between key attributes in the dataset.

Figure 3.3 presents a heat-map displaying the correlation between different goalkeeper attributes. Strong **positive correlations (dark red)** indicate attributes that tend to increase together, while **negative correlations (blue)** reveal inverse relationships. The **market value** exhibits high correlations with attributes such as **overall rating, potential, value,**

and international reputation, suggesting that higher-rated goalkeepers are typically valued higher. Performance metrics like **matches played, minutes, and wins** also show moderate correlations with market value, reinforcing the idea that consistent playing time impacts valuation. Notably, **goalkeeping-specific attributes (e.g., save percentage, GK positioning, GK reflexes)** show weak or negative correlations, indicating that a goalkeeper's value may depend more on general reputation and playing experience rather than isolated shot-stopping ability.

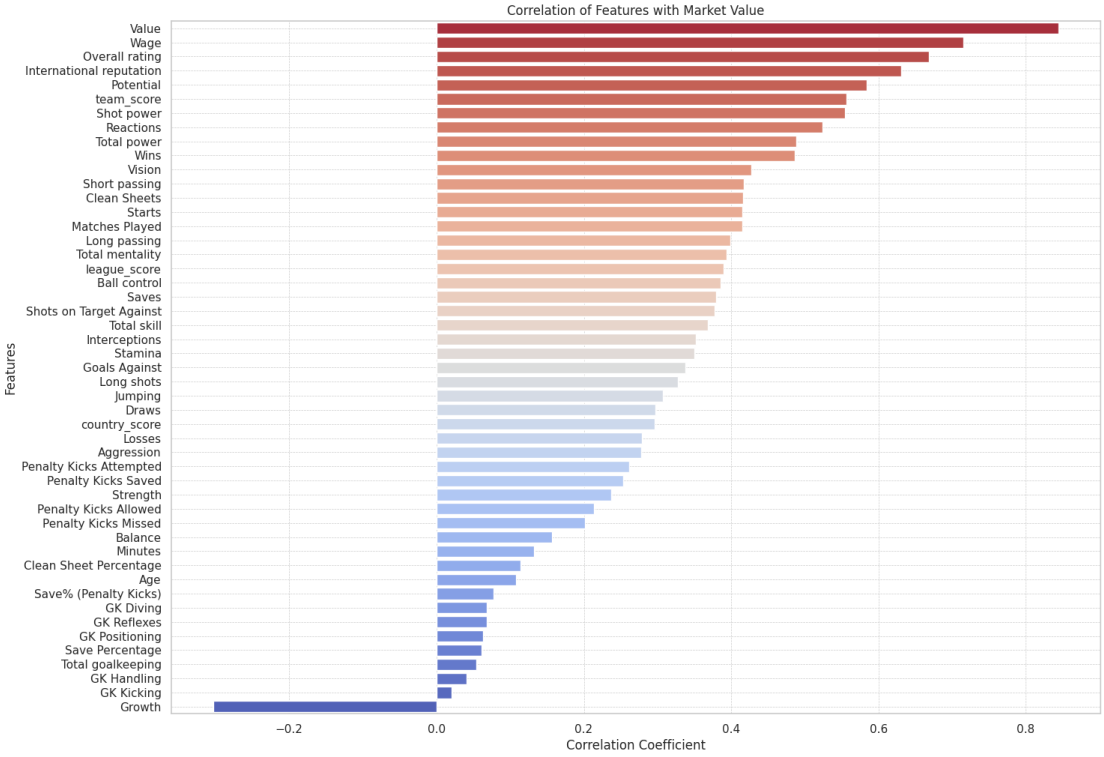


Fig. 3.4. Correlation of various features with market value.

Figure 3.4 ranks the attributes most correlated with market value. Positive correlations indicate factors that contribute to higher valuations. The strongest correlations appear with **value, wage, overall rating, and international reputation**, highlighting that financial indicators and subjective ratings play a crucial role in determining a young goalkeeper's worth. Additionally, **potential and team performance (team score)** exhibit strong correlations, suggesting that goalkeepers from stronger teams receive higher valuations. Surprisingly, **Growth, penalty performances, and goalkeeping-specific attributes (GK diving, GK reflexes, GK handling)** show weak or even negative correlations with market value. This implies that a goalkeeper's technical ability alone does not dictate their market price; rather, club reputation, international recognition, and

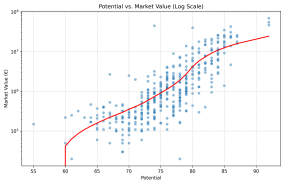
financial valuation metrics play a more dominant role.

The correlation analysis suggests that **market value in young goalkeepers is heavily influenced by reputation, financial valuation, and overall performance rather than isolated shot-stopping abilities**. Attributes such as **team success, player reputation, and wage expectations** strongly impact how goalkeepers are valued in the market. This insight highlights the importance of exposure, playing time, and team performance in determining a goalkeeper's worth rather than just their raw technical goalkeeping skills.

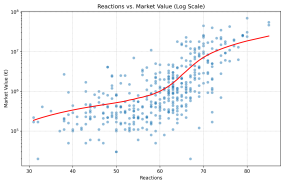
Scatter Plots

In this section, we present several scatter plots. The selection of features for these plots was based on the correlations shown in Fig. 3.4.

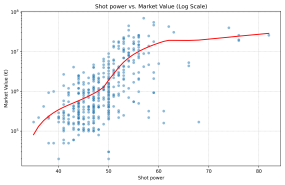
Scatter Plots of Key Features vs. Market Value :



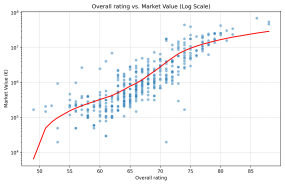
(a) Potential vs. Market Value



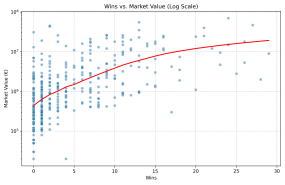
(b) Reactions vs. Market Value



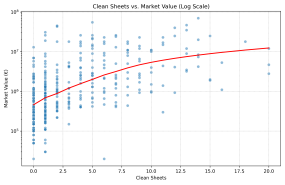
(c) Shot Power vs. Market Value



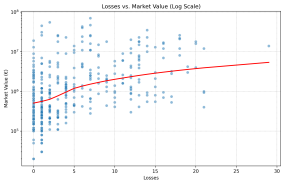
(d) Overall Rating vs. Market Value



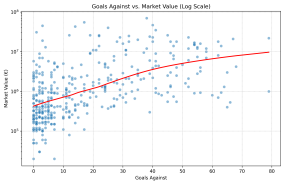
(e) Wins vs. Market Value



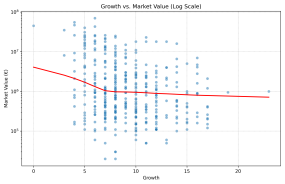
(f) Clean Sheets vs. Market Value



(g) Losses vs. Market Value



(h) Goals Against vs. Market Value



(i) Growth vs. Market Value

Fig. 3.5. Scatter plots displaying the relationship between key features and market value.

Each scatter plot includes a red trend line representing the **LOWESS** (Locally Weighted Scatterplot Smoothing). The **Y-axis is logarithmic**, emphasizing exponential growth trends in market value.

Analysis and Insights :

Scatter plots provide key information on how different attributes of players influence the market value.

Strong Positive Correlations:

- **Potential, Overall Rating, and Reactions** exhibit a **strong positive correlation** with market value, meaning highly skilled or highly rated players tend to be worth significantly more. This trend follows a nearly **exponential curve**, indicating that **elite players are valued disproportionately higher**.
- **Shot Power and Wins** also displays a strong positive correlation, suggesting that **offensive ability and team success contribute to a player's worth**.

Moderate to Weak Correlations:

- **Clean Sheets** [5] has a **moderate correlation** with market value, but the relationship is more scattered. This suggests that other factors, such as **team quality and player reputation**, might be more significant in determining goalkeeper valuations.

Negative Correlations:

- **Losses and Goals Against** show a **negative correlation**, meaning that players on **losing teams or those conceding more goals** tend to have **lower market values**. This is particularly relevant for defenders and goalkeepers, as their valuation depends heavily on defensive stability.
- **Growth vs. Market Value** has a **weak negative correlation**, implying that **players who have already reached their peak tend to be valued lower than younger players with room for improvement**.

3.4.2 Defenders (DF) Dataset

The Defender (DF) dataset consists of approximately **3,500** player samples, each annotated with scaled market values and a wide range of attributes reflecting both performance and physical characteristics. Given the unique responsibilities of defenders compared to other outfield players, specific metrics such as **interceptions, tackles, clearances**, and **defensive duels** are particularly important in evaluating their market value. This

[5] Clean Sheets is the number of matches played without conceding any goal

dataset enables a detailed, position-specific analysis of the key factors that influence the valuation of defenders in the football market.

Key Features: To ensure a comprehensive analysis of defender market value, the dataset includes the following categories of attributes:

- **General Information:** Nation, Position, Squad, Age, Year of Birth, Name, Season, Market Value.
- **Playing Time:** Matches Played, Starts, Minutes, 90s Played.
- **Offensive Metrics:** Goals, Assists, Goals + Assists, Non-Penalty Goals, Penalty Kicks Made and Attempted, Offensive Metrics per 90 Minutes (Goals/90, Assists/90, Goals + Assists/90, etc.).
- **Expected Goals Statistics:** Expected Goals (xG), Non-Penalty xG (npxG), Expected Assisted Goals (xAG), Combined Metrics (npxG + xAG), and their respective per 90 metrics.
- **Progression Indicators:** Progressive Carries, Progressive Passes, Progressive Passes Received.
- **Disciplinary Stats:** Yellow Cards, Red Cards.
- **Physical and Technical Attributes (FIFA EA Sports [34]):** Height, Weight, Preferred Foot, Best Overall Rating, Best Position, Growth, Overall Rating, Potential.
- **Valuation (FIFA EA Sports [34]):** Estimated Market Value[6], Wage[6], International Reputation, Source League.
- **Performance Skills (FIFA EA Sports):**
  - **Attacking:** Total Attacking, Crossing, Finishing, Heading Accuracy, Short Passing, Volleys.
  - **Skill:** Dribbling, Curve, FK Accuracy, Long Passing, Ball Control.
  - **Movement:** Acceleration, Sprint Speed, Agility, Reactions, Balance.
  - **Power:** Shot Power, Jumping, Stamina, Strength, Long Shots.
  - **Mentality:** Aggression, Interceptions, Attacking Position, Vision, Penalties.
  - **Defending:** Total Defending, Defensive Awareness, Standing Tackle, Sliding Tackle.

[6]As provided in EA Sports FIFA video game, not real-world values

  - **Goalkeeping (for completeness):** GK Diving, GK Handling, GK Kicking, GK Positioning, GK Reflexes.

- **Defensive Actions:** Tackles, Tackles Won, Tackles by Zone (Defensive Third, Middle Third, Attacking Third), Dribblers Tackled, Dribbles Challenged, Percentage of Dribblers Tackled, Challenges Lost.
- **Defensive Contributions:** Blocks, Shots Blocked, Passes Blocked, Interceptions, Tackles + Interceptions (Tkl+Int), Clearances, Errors.
- **Additional Metadata:** Attacking and Defensive Work Rate, Team Score, League Score, Country Score.

Visual Analysis: To gain deeper insight into the valuation of defenders, the dataset is complemented with several visualizations

- **Age Distribution:** A pie chart showing the distribution of defenders by age.

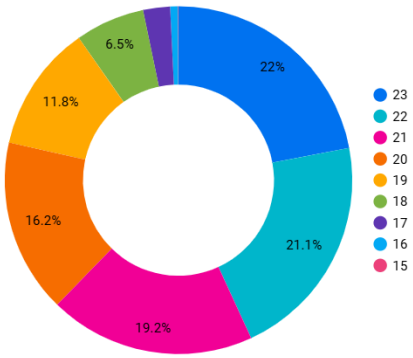


Fig. 3.6. Pie chart displaying the age distribution of defenders.

Figure 3.6 shows that the defender dataset is dominated by players aged **21 to 23**, with **23-year-olds** making up the largest segment (**22%**), followed closely by 22-year-olds (**21.1%**) and 21-year-olds (**19.2%**). This highlights that the sample is mainly composed of players in the later stages of youth development, nearing peak performance age.

- **Market Value by Age:** A box plot visualizing the distribution of market values across age groups for defenders.

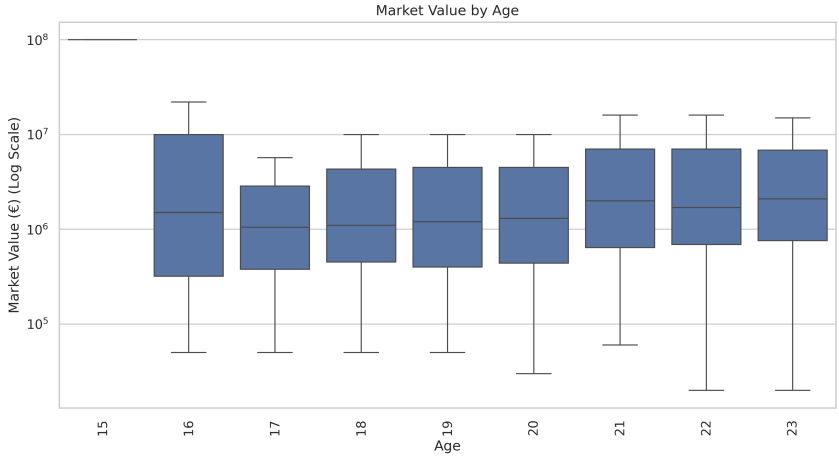


Fig. 3.7. Box plot of defender market value across age groups (log scale).

Figure 3.7 depicts the relationship between defender market value and age. While median market value generally increases with age, the highest variability is seen in players aged **16 and 23**, indicating a wide disparity in individual valuation at these ages. Notably, defenders aged **18 and 21** also show strong value distribution, with some outliers reaching valuations above €10 million. The box plot confirms that although age contributes to higher valuations, performance consistency, reputation, and club context are critical for top-tier market value.

- **Feature Correlation Analysis:**

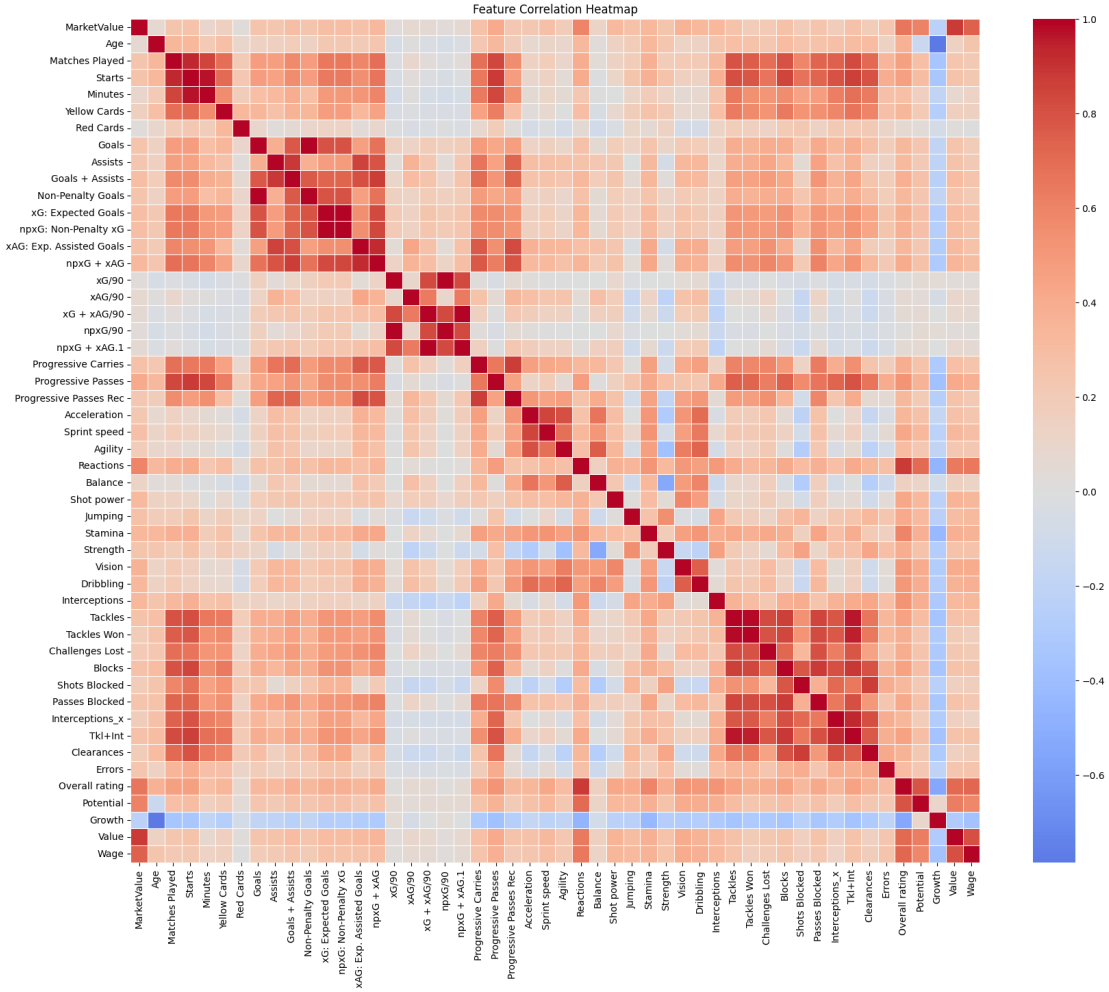


Fig. 3.8. Feature correlation heatmap showing pairwise relationships among all numerical features.

Figure 3.8 presents a correlation heatmap that illustrates the linear relationships between all numeric features in the defender dataset. As expected, attributes such as **overall rating**, **potential**, **value**, and **wage** form a highly correlated cluster, confirming that FIFA video game metrics are strongly interconnected and predictive of market value. Defensive performance metrics such as **tackles**, **interceptions**, **blocks**, and **clearances** also show positive correlations among themselves but form a less distinct relationship with market value.

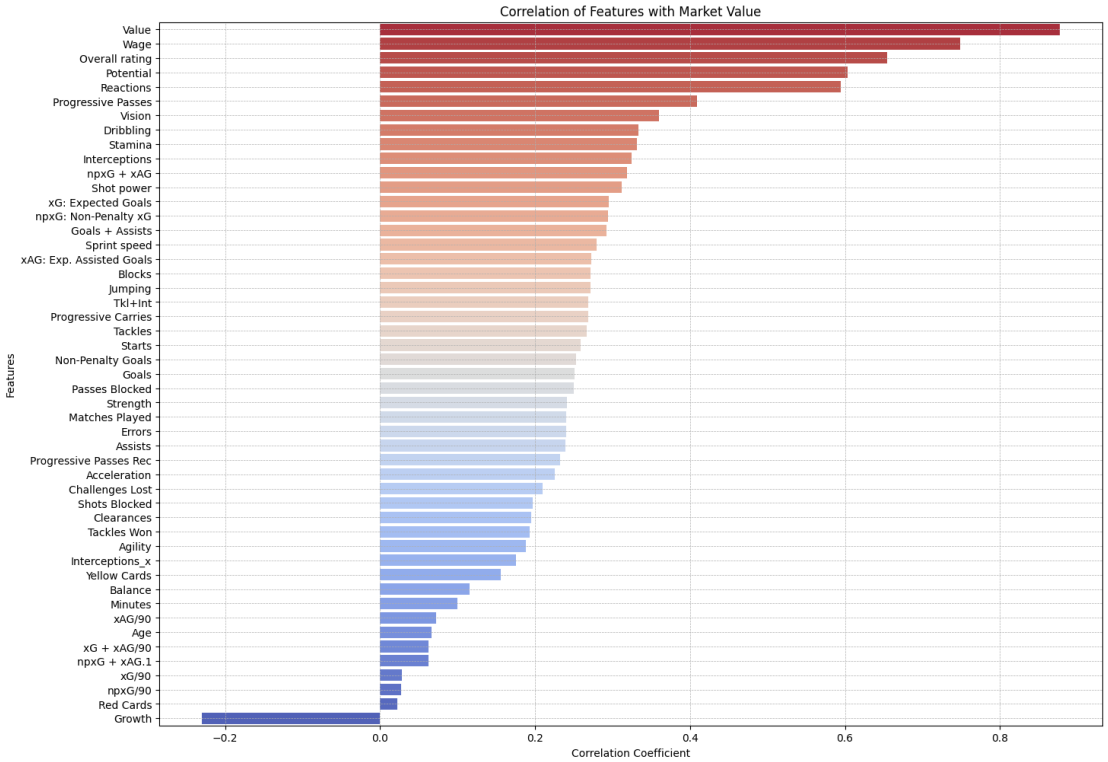


Fig. 3.9. Correlation of individual features with market value.

In Figure 3.9, we can directly observe which features are most correlated with market value. The top positively correlated features include:

- **Value** and **Wage**: unsurprisingly show the highest correlation, as these are FIFA-derived estimations tied closely to market perception.
- **Overall Rating** and **Potential**: strong indicators of general ability and future promise.
- **Reactions, Progressive Passes, Vision, Dribbling**: suggest that offensive contribution and passing abilities are more highly valued than basic defensive actions in determining a defender's worth.

Interestingly, classic defensive stats like **tackles**, **interceptions**, and **blocks** are only moderately correlated with market value, which may reflect how modern defenders are increasingly expected to contribute to build-up play and possess technical abilities.

On the other hand, negatively correlated features include:

- **Red Cards** and **Errors**: expected, as disciplinary issues and critical mistakes lower a player's market perception.

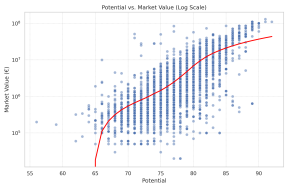
– **Growth**: has a slightly negative correlation, possibly due to players with high initial potential already reaching their ceiling.

– **xG/90** and **npxG/90**: weakly negatively correlated, reflecting their limited relevance to defenders' valuation.

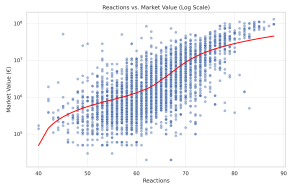
Overall, the correlation analysis confirms that market value is influenced more by reputation, overall skill metrics, and offensive or ball-progressive capabilities than purely defensive metrics. These insights support the growing view of defenders as multi-functional players within modern tactical systems.

Scatter Plots of Key Features vs. Market Value:

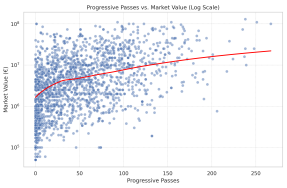
(a) Overall Rating vs. Market Value



(b) Potential vs. Market Value

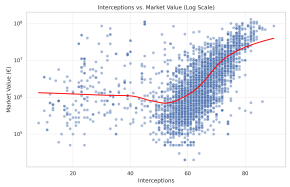


(c) Reactions vs. Market Value

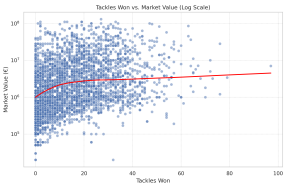


(d) Progressive Passes vs. Market Value

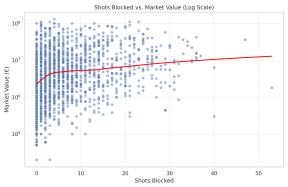
(e) Growth vs. Market Value



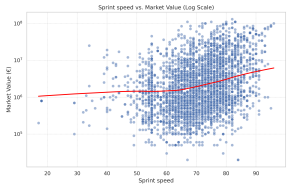
(f) Interceptions vs. Market Value



(g) Tackles Won vs. Market Value



(h) Shots Blocked vs. Market Value



(i) Sprint Speed vs. Market Value

Fig. 3.10. Scatter plots illustrating the relationships between various key player features and market value. Each plot includes a LOWESS (Locally Weighted Scatterplot Smoothing) regression line (in red) to highlight trends and non-linear relationships.

Analysis and Insights: Strong Positive Correlations

- **Overall Rating, Potential, and Reactions**: These features show strong positive correlations with market value. The **red LOWESS trend line** rises sharply, particularly after certain thresholds, indicating elite players with high ratings or reactions are valued disproportionately higher.
- **Progressive Passes**: A moderate positive correlation is evident. Players capable of progressive passing generally have higher market valuations, although the trend line indicates diminishing returns beyond a certain level of passes.

Moderate and Non-linear Relationships

- **Interceptions**: Exhibits a distinct non-linear trend. Low interception counts initially have minimal impact, but after a threshold (approximately 60 interceptions), player value rises significantly. This suggests a premium valuation for defenders who demonstrate exceptional interceptive abilities.
- **Sprint Speed**: Displays a weak to moderate positive trend. Although faster defenders generally hold slightly higher market values, the significant scatter indicates speed alone is not strongly decisive for valuation.

Weak or Negligible Correlations

- **Tackles Won and Shots Blocked**: These defensive metrics show weak positive correlations, with relatively flat red trend lines. This implies these specific defensive actions have limited direct influence on market valuations individually.
- **Growth**: Shows a weak negative correlation with market value. Players with higher growth potential (typically younger or less established) are slightly less valued currently, indicating market valuation favors established performance over mere future potential.

The **red LOWESS trend line** effectively captures the nuanced relationships between player attributes and market value. For attributes with strong positive correlations such as *Overall Rating*, *Potential*, and *Reactions*, the red line sharply ascends, particularly after certain threshold values, indicating that elite performers experience exponential valuation increases. Moderate correlations, like those seen with *Progressive Passes* and *Interceptions*, show the LOWESS line initially rising before flattening or steepening after specific points, highlighting diminishing returns or threshold effects. Conversely, attributes with weak correlations (*Tackles Won*, *Shots Blocked*, *Growth*) are represented by relatively flat or gently sloping LOWESS lines, signifying minimal direct influence on market valuation.

3.4.3 Midfielders (MF) Dataset

The Midfielder (MF) dataset consists of approximately **5,300** player samples, each annotated with scaled market values and the set of attributes covering performance metrics, technical skills, physical characteristics, and valuation indicators. Midfielders play a pivotal role in football teams, often bridging the defensive and offensive lines. Consequently, their evaluation necessitates a balanced consideration of both offensive contributions (such as goals, assists, passes, and progressive plays etc.) and defensive actions (including interceptions, tackles, blocks, and clearances etc.).

Having detailed offensive statistics allows the assessment of midfielders' abilities to create scoring opportunities, maintain possession, and facilitate attacks. Simultaneously, defensive metrics provide insights into their effectiveness in disrupting opponent play, regaining possession, and offering crucial defensive support. The combination of these offensive and defensive statistics ensures a holistic evaluation, accurately capturing the diverse responsibilities midfielders fulfill on the pitch.The dataset uses the same detailed features as those listed previously for the defenders.

Visual Analysis:

- **Age Distribution:** A pie chart showing the distribution of midfielders by age.

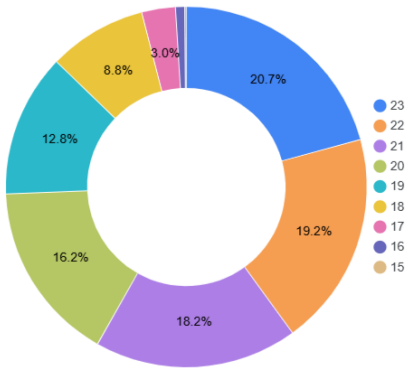


Fig. 3.11. Pie chart displaying the age distribution of midfielders.

Figure 3.11 indicates that midfielders aged **23** comprise the largest portion of the dataset at **20.7%**, followed closely by 22-year-olds at **19.2%**. Midfielders aged **21 and 20** also represent significant proportions, at **18.2% and 16.2%**, respectively. Conversely, the youngest age group (**15-year-olds**) accounts for less than **1%**, highlighting a potentially limited representativeness and cautioning against strong conclusions from this segment.

- **Market Value by Age:** A box plot visualizing the distribution of market values across age groups for midfielders.

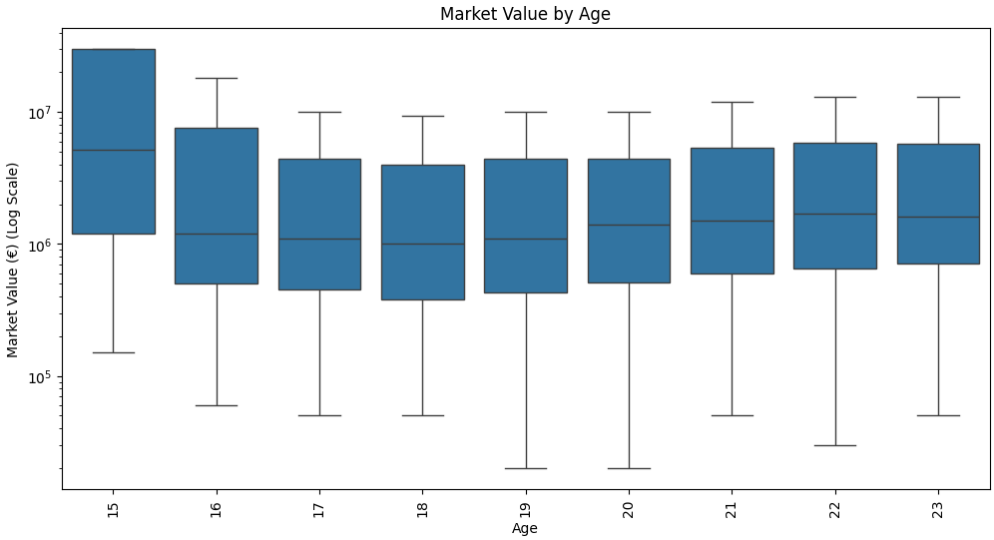


Fig. 3.12. Box plot of midfielder market value across age groups (log scale).

Figure 3.12 illustrates the relationship between midfielder market value and age. A notable observation is the wide variability among midfielders aged **15**, whose market value range is disproportionately large despite representing less than **1%** of the dataset. This exaggerated variability arises primarily due to the very small sample size, which can create misleading impressions of extreme value differences within this age group. Therefore, conclusions drawn from this segment should be interpreted cautiously.

From age **17 onward**, a clearer trend emerges, with median market values steadily rising, reflecting greater market valuation consistency with increasing age. This supports the inference that while age generally correlates positively with market value, reliable market valuation assessments require sufficient sample sizes to avoid distorted interpretations.

- **Feature Correlation Analysis:**

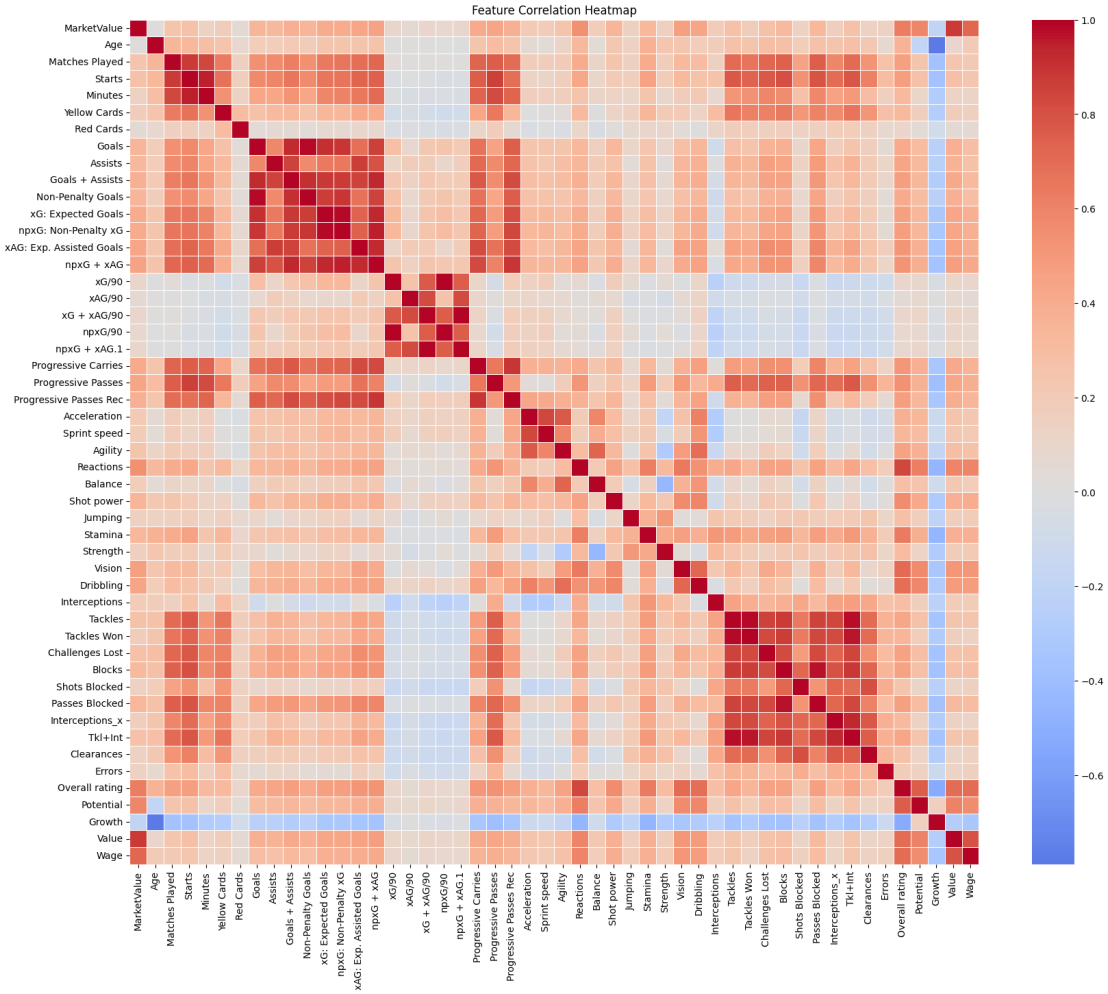


Fig. 3.13. Feature correlation heatmap showing pairwise relationships among all numerical features for midfielders.

Figure 3.13 shows clear positive correlations among attacking and progressive metrics such as goals, assists, expected goals (xG), expected assisted goals (xAG), progressive passes, and carries. This suggests these metrics collectively contribute to effective offensive midfield performance. Defensive actions, including interceptions and tackles, exhibit weaker correlations with market value, showing a potential undervaluation of defensive contributions or a stronger market preference for offensive skills in midfielders. The negative correlation between age and growth underlines the diminishing market valuation of potential in older players.

- **Detailed Feature Correlation Heatmap:**

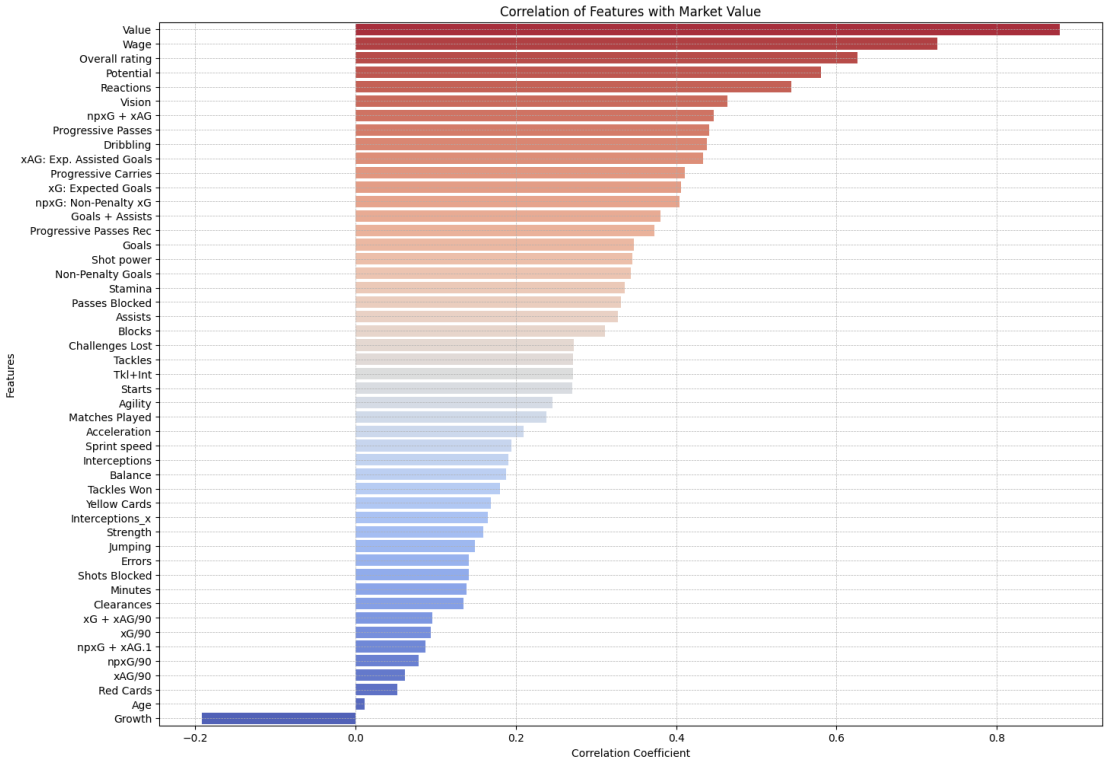


Fig. 3.14. Detailed heatmap illustrating correlations between all numerical features.

The analysis of the correlation coefficients in 3.14 for midfield players reveals several key patterns. Features such as **Overall Rating**, **Potential**, and **Reactions** exhibit the strongest positive correlations with market value, underscoring the importance of general skill level, development potential, and decision-making responsiveness in evaluating midfield talent. Additionally, creative and playmaking attributes including **Vision**, **Progressive Passes**, **Dribbling**, **Expected Assisted Goals (xAG)**, and **Progressive Carries** all show high correlations (above 0.35), emphasizing that technical creativity and ball progression are highly valued traits in midfielders.

Offensive metrics such as **Expected Goals (xG)**, **Non-Penalty xG (npxG)**, **Goals + Assists**, and **Shot Power** also correlate positively, though to a slightly lesser degree (approximately between 0.25–0.35), indicating that midfielders contributing to goal creation and scoring significantly boost their market appeal. In contrast, defensive metrics such as **Passes Blocked**, **Tackles**, **Interceptions**, and **Blocks** show moderate to low correlation, reflecting a comparatively weaker influence on valuation for players in these roles.

Physical attributes like **Stamina**, **Agility**, **Acceleration**, and **Sprint Speed** have moderate correlations (~0.15–0.25), suggesting that athleticism plays a supporting rather than a leading role. Metrics such as **Jumping**, **Strength**, and defensive volume stats (**Clearances**, **Shots Blocked**) are weakly correlated with market value. Interestingly, normalized performance metrics like **xG/90**, **xAG/90**, and **npxG/90** display negative correlations, likely due to lower reliability across varying minutes or inconsistent match presence.

Finally, **Age** and **Growth** have the most negative correlations, with Age being a particularly strong negative driver, highlighting the preference for younger talents with higher resale potential. These insights collectively indicate that market value in midfielders is driven more by technical skill, creativity, and offensive contribution than by defensive workload or physical prowess.

For deeper visual exploration, we present a selection of scatter plots that focus on the most influential features identified in the correlation analysis. These plots offer clarity on potential non-linear relationships, outliers, or thresholds, and reinforce which traits tend to drive higher market valuation among midfielders.

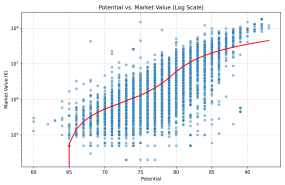
Scatter Plots of Key Features vs. Market Value:

3.4.4 Attackers/Forward (FW) Dataset

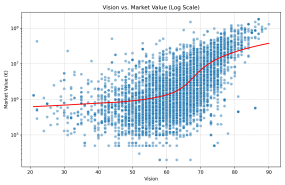
The dataset for attackers or forwards (FW) comprises approximately **3,500** player samples, each annotated with a scaled market value, as described in Section 3.3.1. The features collected include a mix of match statistics, expected goal metrics, physical and technical attributes, as well as contextual and scouting-based ratings.

While many features are shared with the midfielder and defender datasets, the attackers' dataset places greater emphasis on end-product indicators such as goals, expected goals (xG), non-penalty goals, and shooting accuracy. These metrics are critical for assessing a forward's efficiency and impact in the final third. In addition, dynamic metrics like `Progressive Passes Received`, `xG/90`, and `xAG/90` allow for a normalized evaluation of offensive involvement and creativity on a per-90-minute basis, offering a fair comparison regardless of total minutes played.

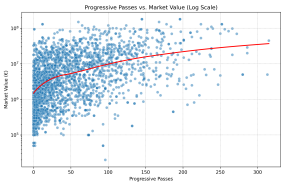
The dataset also incorporates FIFA-style attributes such as finishing, dribbling, positioning, and sprint speed, which are known to correlate with attacking effectiveness. These features provide additional nuance when modeling player value and contribution, particularly in roles that rely heavily on individual skill and movement without the ball.



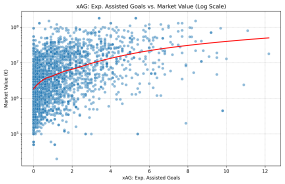
(a) Potential vs. Market Value



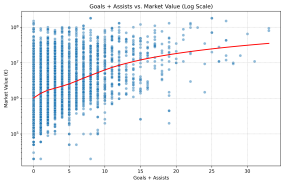
(b) Vision vs. Market Value



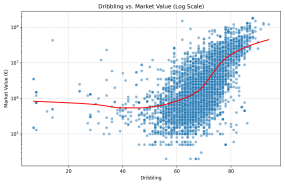
(c) Progressive Passes vs. Market Value



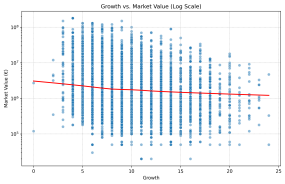
(d) xAG vs. Market Value



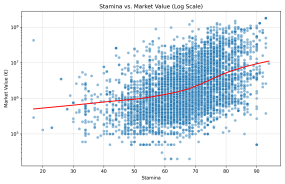
(e) Goals + Assists vs. Market Value



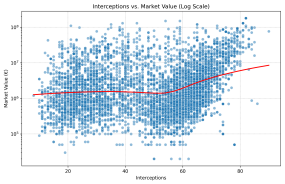
(f) Dribbling vs. Market Value



(g) Growth vs. Market Value



(h) Stamina vs. Market Value



(i) Interceptions vs. Market Value

Fig. 3.15. Scatter plots illustrating the relationships between selected key features and market value for midfield players. Each plot includes a LOWESS (Locally Weighted Scatterplot Smoothing) regression line to emphasize underlying trends and capture any non-linear associations.

Visual Analysis:

- **Age Distribution:** A pie chart showing the distribution of forwards by age.

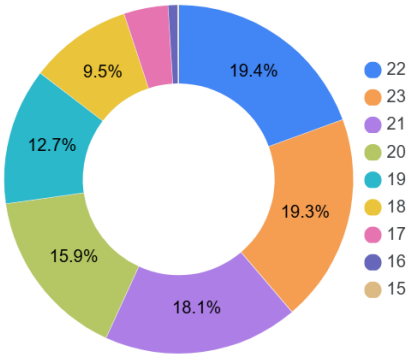


Fig. 3.16. Pie chart displaying the age distribution of forwards.

- **Market Value by Age:** A box plot visualizing the distribution of market values across age groups for forwards.

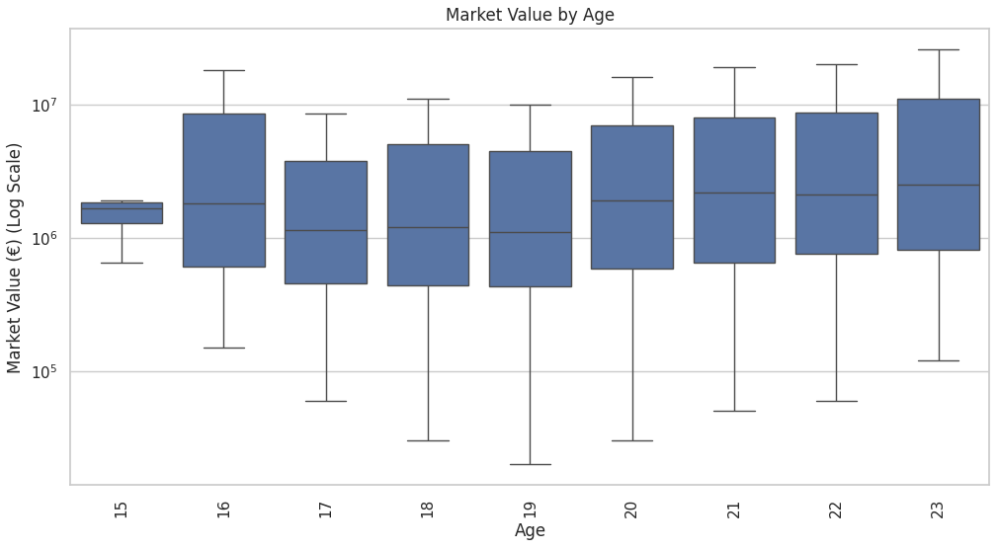


Fig. 3.17. Box plot of fowrards market value across age groups (log scale).

The distribution of market value across age groups for forwards Fig.3.17 reveals a non-linear but clearly upward trend in both value magnitude and variability as players mature. From the first boxplot, it is evident that the median market value

increases from age 15 to 23, with a notably wider spread beginning around age 17. While 15-year-old players show tightly packed valuations—reflecting early-stage scouting and relatively low financial commitment—the spread significantly expands from age 16 onward, indicating more diverse valuation based on performance, potential, and exposure. The use of a log scale highlights this disparity: although median values grow moderately, the upper whiskers for ages 21–23 extend into the multi-million euro range, suggesting a segment of highly valued elite prospects.

The age distribution in the donut chart 3.16 complements this by showing that the dataset is skewed toward older youth players, with the majority of forwards concentrated between ages 20 and 23. Notably, age 22 accounts for the largest share (19.4%), closely followed by age 23 (19.3%) and age 21 (18.1%). The younger age groups (15–17) are underrepresented, making up only a small portion of the sample. This imbalance reinforces the idea that forwards are most heavily evaluated and priced during the transition to senior football, particularly between ages 20 and 23, which also coincides with the most financially valuable years in their early professional careers.

- **Feature Correlation Analysis:**

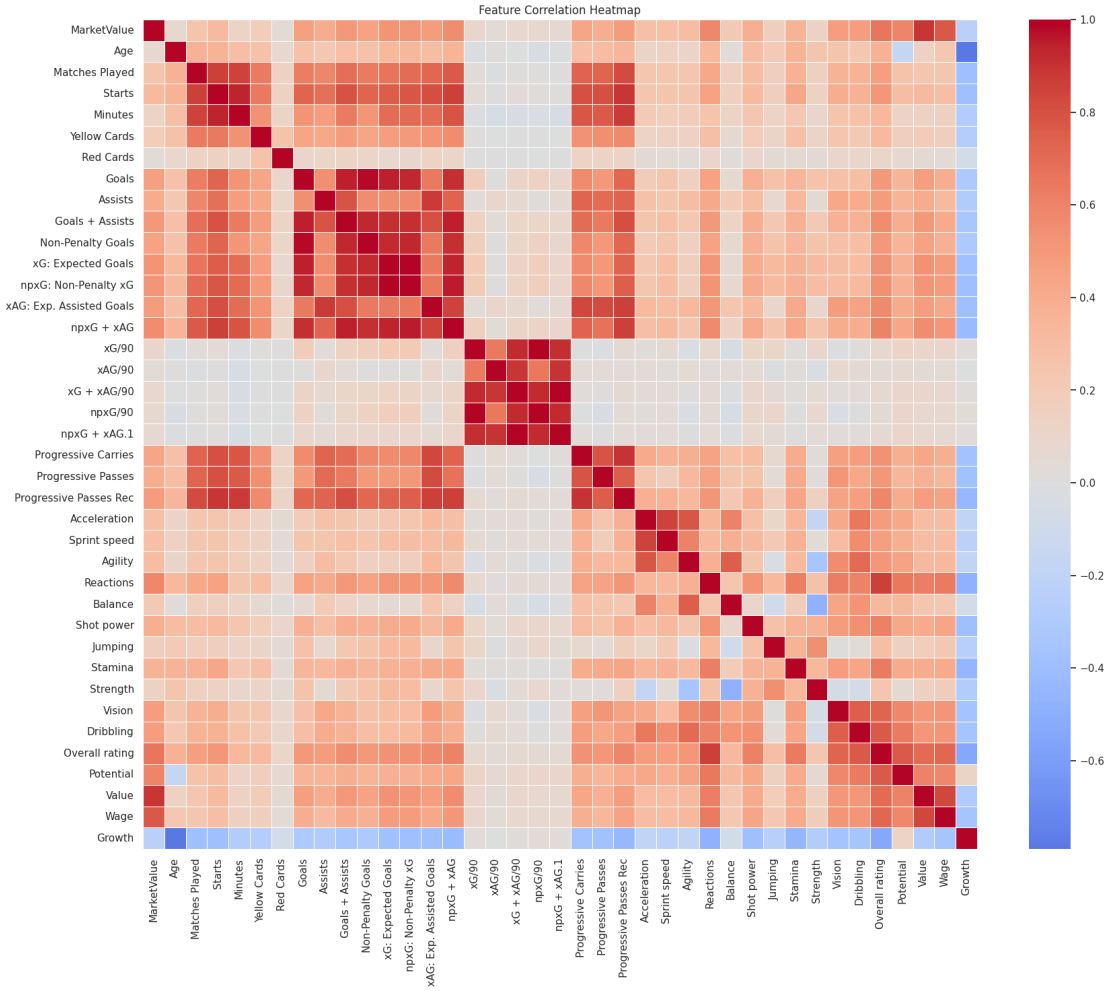


Fig. 3.18. Feature correlation heatmap showing pairwise relationships among all numerical features for forwards.

As expected, the `MarketValue` feature shows strong positive correlations with attributes such as `Overall Rating`, `Potential`, `Value`, and `Goals + Assists`, all of which are consistent indicators of performance and perceived ability. Notably, `Progressive Passes`, `Progressive Carries`, and `xAG (Expected Assisted Goals)` also exhibit moderate correlations with market value, highlighting the importance of forward involvement in build-up and chance creation beyond just scoring goals.

Several goal-related statistics—such as `xG`, `npxG`, and other per-90 expected metrics—show near-zero or weak correlation with market value. However, this is not

indicative of their irrelevance; rather, it stems from the lack of availability of these advanced metrics across all players in the dataset. As explained in Section 3.2.3, these features are often only reported for players in leagues with advanced tracking data, leading to substantial missing values. This absence reduces the reliability of the computed correlations, and as such, these metrics should not be dismissed based solely on their low statistical association here.

Finally, the heatmap also reveals groups of tightly correlated variables, such as among `Goals`, `Assists`, `Goals + Assists`, and `xG`-related metrics, which reflect underlying shared contributions to attacking output. Physical traits like `Sprint Speed`, `Agility`, and `Acceleration` show mild to moderate associations with both technical and performance attributes, confirming their supportive but not dominant role in determining a player's market valuation.

Further insights into the specific contributions of each feature to market value can be observed in the detailed bar chart of correlation coefficients shown in Figure 3.19. This visualization highlights which individual attributes are most linearly associated with market value for forwards. As reported in the graph, attributes such as `Overall Rating`, `Potential`, and `Reactions` show the strongest positive correlations, underscoring their importance in evaluating attacking talent. Meanwhile, physical traits and per-90 metrics such as `xG/90`, `npxG/90`, and `xAG/90` display minimal or even negative correlation. As previously mentioned in Section 3.2.3, this is largely due to missing data coverage for these advanced metrics across many players, rather than their actual irrelevance. The full ranked list in the chart allows for a clearer prioritization of relevant features for modeling and evaluation tasks.

Scatter Plot Analysis of Key Features. To gain a more granular understanding of the relationship between specific player attributes and market value for forwards, we present a series of scatter plots in Figure 3.20. These plots were selected based on both statistical correlation and domain relevance. They include attributes with strong positive associations (e.g., `Overall Rating`, `Potential`, `Reactions`), direct performance indicators (e.g., `Goals + Assists`, `xG + xAG`, `Dribbling`), and features with more complex or indirect influence (e.g., `Progressive Passes Received`, `Age`). Additionally, we include `xG/90`, which exhibits low correlation due to partial data coverage (see Section 3.2.3), to visually examine the extent of sparsity or non-linearity in this feature. These visualizations not only support the interpretation of the correlation results but also help detect patterns, outliers, and non-linear effects that may influence predictive modeling.

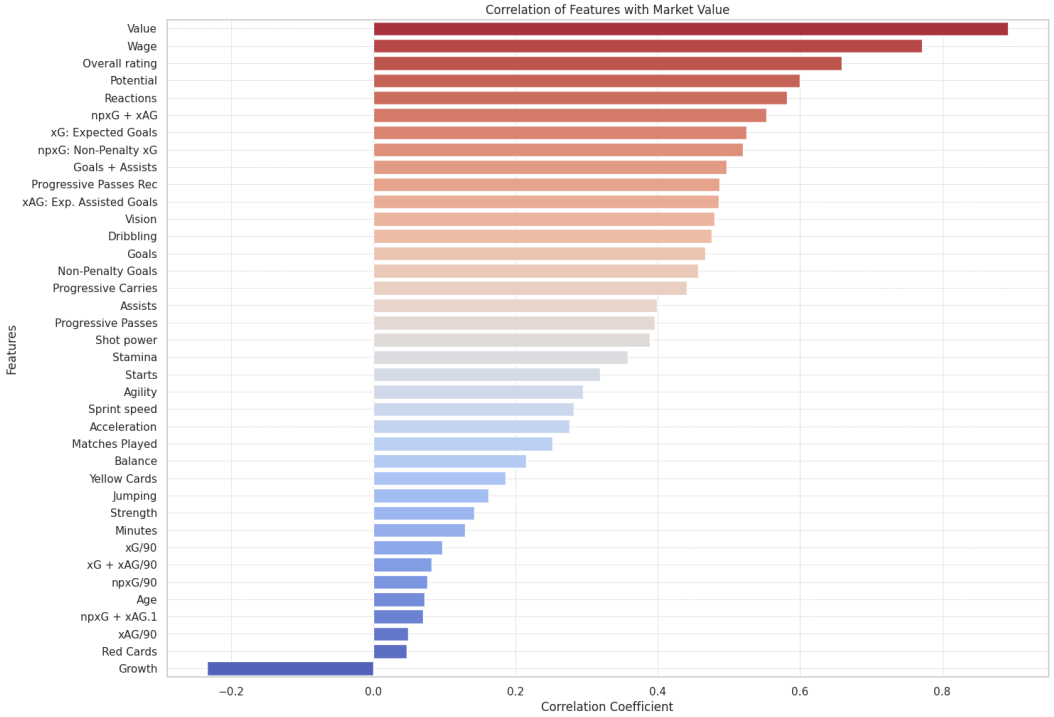
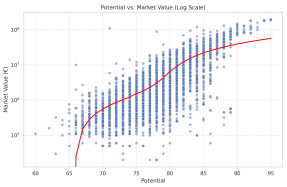
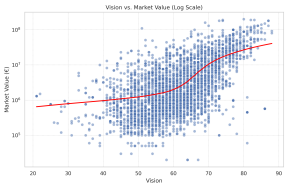


Fig. 3.19. Bar chart of Pearson correlation coefficients between each feature and the market value for forwards. Higher bars indicate stronger positive association; negative values imply inverse relationships.

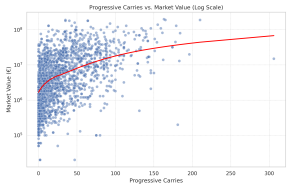
Scatter Plots of Key Features vs. Market Value:



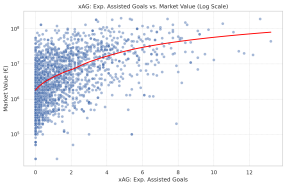
(a) Potential vs. Market Value



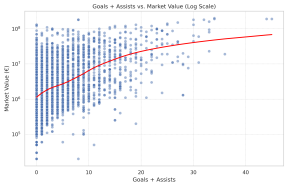
(b) Vision vs. Market Value



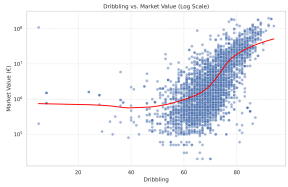
(c) Progressive Carries vs. Market Value



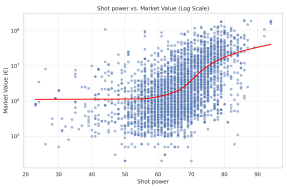
(d) xAG vs. Market Value



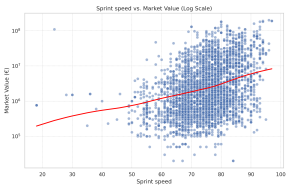
(e) Goals + Assists vs. Market Value



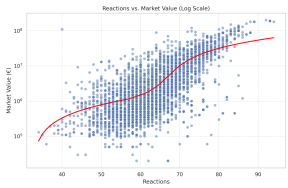
(f) Dribbling vs. Market Value



(g) Shot Power vs. Market Value



(h) Sprint Speed vs. Market Value



(i) Reactions vs. Market Value

Fig. 3.20. Scatter plots showing the relationship between selected forward-specific attributes and market value. Each plot includes a LOWESS smoothing line to visualize potential non-linear trends.

3.4.5 Unified Dataset

The complete dataset consists of approximately **12,500 football players** and is constructed by merging all the position-specific datasets previously analyzed. This unified collection brings together players from diverse league systems, seasons, and nationalities, covering the full spectrum of on-field roles and tactical profiles. Each player is described using a rich set of biographical, physical, performance, and contextual features, enabling a comprehensive, multidimensional analysis of football talent and market valuation.

In this section, we present key analytical visualizations exploring the factors associated with market value across various dimensions, including player's source league, playing position, and international reputation.

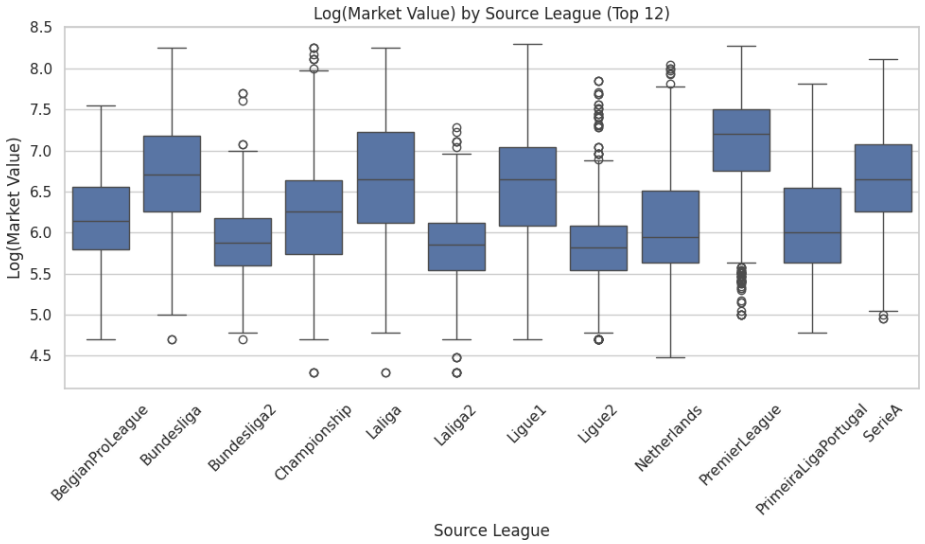


Fig. 3.21. Box plot of log(Market Value) across the top 12 source leagues.

From Figure 3.21, we observe that players from top-tier leagues (e.g., Premier League, Bundesliga, and La Liga) possess generally higher market valuations compared to second-tier leagues (e.g., Championship, Ligue 2, Bundesliga 2). The broader range of valuations within top leagues reflects both the presence of elite, highly valued players and younger, promising talents.

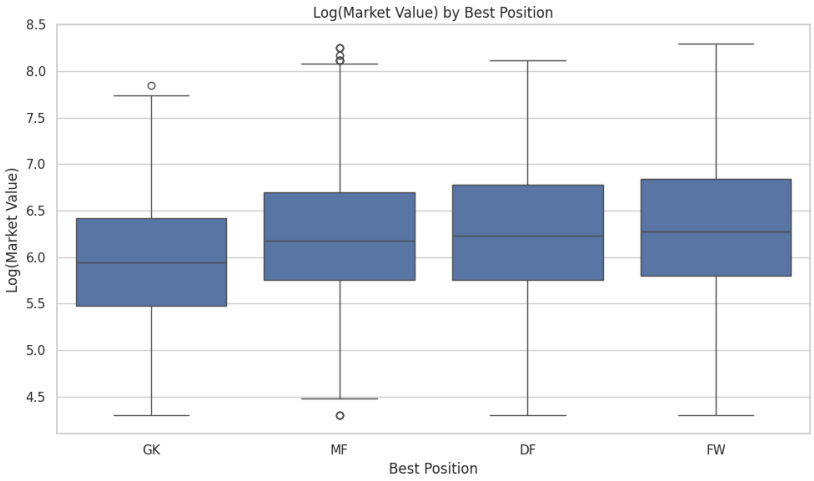


Fig. 3.22. Box plot of log(Market Value) segmented by player positions.

As illustrated in Figure 3.22, forwards (FW) and midfielders (MF) command slightly higher market valuations on average. This aligns with the market's preference for offensive and creative talents, likely due to their direct impact on scoring and match outcomes. In contrast, defensive roles (DF, GK) exhibit relatively lower median values and less variation, indicative of more stable but generally lower perceived financial value.

Fig. 3.23. Bar plot depicting average Market Value across different levels of International Reputation on a logarithmic scale.

Figure 3.23 shows a strong and clear positive correlation between international reputation and average market value. Players with the highest international reputation scores

(4–5) command significantly higher valuations compared to players with lower reputation scores (1–3). This exponential growth trend strongly suggests that international recognition substantially enhances a player's financial value and perceived marketability.

3.4.6 Team-Level Dataset

The team-level dataset consists of approximately 400 team-season entries, derived from around 100 professional football clubs tracked across four consecutive seasons (2020-2021 to 2023-2024). Each entry corresponds to a single team's performance within a given season and is annotated with an extensive set of aggregated performance metrics.

These metrics encompass offensive indicators (such as goals scored, expected goals (xG), and shot efficiency), defensive records (including goals conceded, tackles, interceptions, blocks, and errors), and possession or build-up attributes (such as pass completion rates, progressive passes, and average possession percentage). This multidimensional representation enables a holistic understanding of a team's strategic strengths and weaknesses.

The inclusion of team-level data serves multiple roles within this study. It provides crucial contextual signals for evaluating player environments (e.g., team strength or playing style) and enables the construction of derived performance indexes (defensive, midfield, attacking) used exclusively in the optimization model. These indices, which summarize how well a team performed in each line of play, will be formally introduced and explained in the following section.

3.5 Feature Engineering

This section describes the set of features that were engineered to enhance both the market value prediction model and the player selection optimization model. Feature engineering was central to the project, enabling us to capture both player-specific qualities and broader contextual and tactical factors relevant to valuation and recruitment decisions.

We distinguish between two groups of engineered features:

- Features created to improve the accuracy of **player market value prediction**
- Features developed exclusively for the **optimization model**

3.5.1 Features for Market Value Prediction

To account for the external context in which a player performs, three new features were engineered: *team_score*, *league_score*, and *country_score*. These features reflect the competitive environment and reputation of a player's surroundings and were developed to improve prediction accuracy.

The process for constructing these features involved the following steps:

- *Weighted mean calculation*: For each team, league, and country, a weighted average market value $\bar{M}_g$ was computed using the number of matches played as weights:

$$\bar{M}_g = \frac{\sum_{i \in g} M_i \times w_i}{\sum_{i \in g} w_i}$$

  where:

  - M_i is the market value of player i,
  - w_i is the number of matches played by player i,
  - The summations are taken over all players i in group g.

  Groups with zero total matches were assigned missing values and handled accordingly.

- *Score assignment based on percentiles*: The computed weighted means were transformed into discrete scores ranging from 1 to 5, according to their position in the empirical distribution:

$$\text{score}_g = \begin{cases} 5 & \text{if } \bar{M}_g \geq Q_{80} \\ 4 & \text{if } Q_{60} \leq \bar{M}_g < Q_{80} \\ 3 & \text{if } Q_{30} \leq \bar{M}_g < Q_{60} \\ 2 & \text{if } Q_{10} \leq \bar{M}_g < Q_{30} \\ 1 & \text{if } \bar{M}_g < Q_{10} \end{cases}$$

  Here, Q_p denotes the p-th percentile of the distribution across all groups. Groups with insufficient data were assigned a default score of 3.

This feature engineering step enriched the prediction dataset with contextual indicators reflecting the strength of a player's environment. Although discretization introduced a minor reduction in information granularity, it enhanced interpretability and robustness, important considerations for practical applications.

3.5.2 Features for Optimization Modeling

Player Contribution Dataset for Optimization

To support the optimization model, a dedicated subset of the player-level dataset was derived from the original data. This subset retained only real-world, match-based

performance statistics — excluding subjective ratings or video game attributes — in order to preserve the objectivity and interpretability of the resulting metrics.

The goal of this process was to compute how much each player contributed to their team's overall performance in each statistical category. For a given feature, the player's contribution was calculated as the ratio between their individual value and the corresponding team-level total:

$$\text{Contribution}_{i,f} = \frac{X_{i,f}}{X_{\text{team}(i),f}}$$

where:

- $X_{i,f}$ is the value of feature f for player i,
- $X_{\text{team}(i),f}$ is the total value of feature f aggregated over all players in player i's team.

This transformation resulted in a set of relative contribution scores that quantify each player's importance within their tactical unit for specific performance metrics (e.g., tackles, interceptions, progressive passes).

After calculating raw contribution ratios, all features were rescaled using a Min-Max normalization approach:

$$X'_{i,f} = 20 + \frac{X_{i,f} - \min(X_f)}{\max(X_f) - \min(X_f)} \times 80$$

This normalization step made sure that all features were scaled while avoiding zero values which could lead to interpretability issues or disproportionately penalize players with low (but non-zero) contributions. Scaling was used to:

- Preserve relative differences in contribution while eliminating scale-related bias.
- Integrate heterogeneous performance metrics (e.g., blocks, assists, xG) into unified modeling pipelines.
- Avoid assigning zero to any feature, which could create logical inconsistencies in downstream modeling.
- Simplify interpretation — ensuring that all players receive a baseline contribution score, avoiding complications from zeros.

Role Detection via Clustering

To move beyond traditional positional labels and gain a deeper understanding of tactical player roles, we employed an unsupervised clustering approach across the three major positional lines: defenders (DF), midfielders (MF), and forwards (FW). The aim was to identify and group players into role-based archetypes according to their on-field contributions, rather than predefined labels.

For each positional line, a subset of the full player dataset was created by filtering players who were classified — either primarily or secondarily — under that line.

A set of features was selected for each line, based on domain knowledge and statistical importance. Importantly, some features commonly associated with other lines (e.g., progressive passes or key passes in defenders) were still included when relevant. This reflects the fluid nature of modern football, where players are not confined strictly to their traditional positional responsibilities. For example, ball-playing defenders may contribute significantly to progression metrics, while defensive forwards may impact pressing and recoveries.

To reduce dimensionality and enhance cluster interpretability, Kernel Principal Component Analysis (Kernel PCA) with a radial basis function (RBF) kernel was applied prior to clustering.

We used **Gaussian Mixture Models (GMM)** to identify roles within each line. The number of clusters was empirically selected per line based on distribution and interpretability criteria and on the **Silhouette Score**. Each resulting cluster represents a tactical role with a distinct statistical fingerprint.

For each line, the average profile (centroid) of every cluster was computed and interpreted as a role vector. These profiles were visualized using radar plots, which are presented in the figures below. The plots help reveal which features dominate each role and provide a visual foundation for manually assigning descriptive labels to the detected roles.

Finally, after defining the role clusters based on high-performing players in each line, the remaining players were assigned to the most appropriate roles using cosine similarity. This approach was adopted after observing that traditional classification models (e.g., decision trees, support vector machines) yielded unsatisfactory results, likely due to the limited size and imbalanced structure of the clustered data. Instead of relying on poorly performing classifiers, cosine similarity offered a more direct and reliable method for role assignment. By comparing each player's normalized contribution profile to the previously learned role vectors, players were matched to the tactical role they were most similar to in terms of feature distribution.

As a result of this process, a new role feature was added for each player, indicating the

tactical role they most closely resemble based on their contribution profile. This role label enriched the player dataset with interpretable role-level context, enabling more informed matching between individual player characteristics and team-level needs in the optimization model.

Visualizing and Interpreting Tactical Roles

To illustrate the differences between detected roles, the following list summarizes the tactical role profiles per positional line. Each role label reflects the dominant patterns observed across key contribution features.

- **Defenders (DF)**

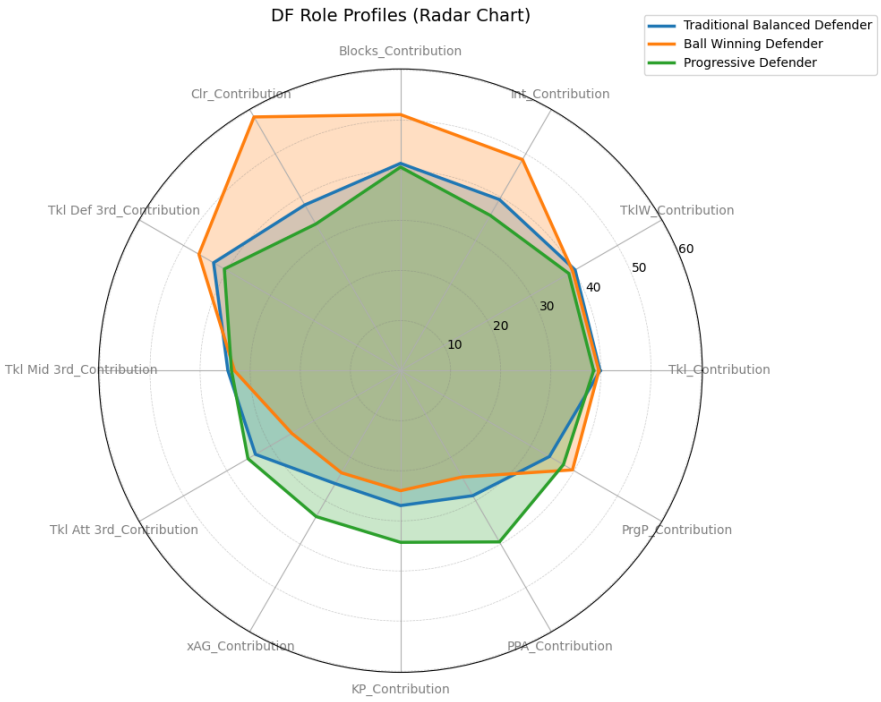


Fig. 3.24. Defensive Role Profiles (Radar Chart)

  - **Traditional Balanced Defender**: Well-rounded contribution across core defensive actions (tackles, interceptions, clearances); typical of center-backs with conservative progression.
  - **Ball Winning Defender**: Strong peaks in defensive metrics (e.g., tackles won, interceptions) indicating an aggressive, duel-oriented profile focused on regaining possession.
  - **Progressive Defender**: Higher progressive passing and xAG values, reflecting a ball-playing defender involved in transitions and build-up.

- **Midfielders (MF)**

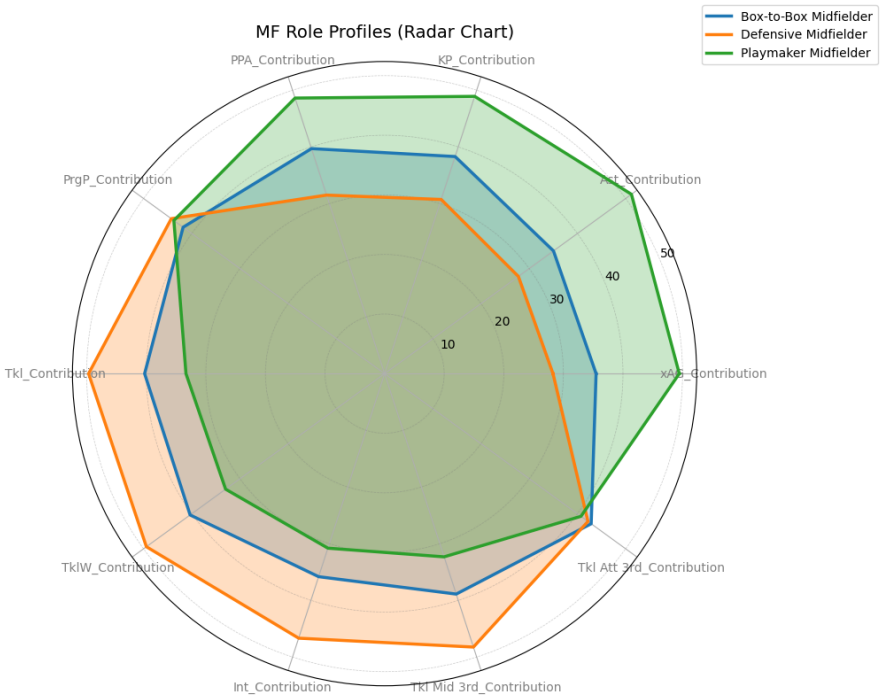


Fig. 3.25. Midfield Role Profiles (Radar Chart)

  - **Box-to-Box Midfielder**: Balanced profile across offensive and defensive metrics; suited for dynamic contributors to all phases.
  - **Defensive Midfielder**: Peaks in interceptions, tackles, and clearances; a midfield anchor shielding the back line.
  - **Playmaker Midfielder**: Dominated by key passes, xAG, and assist contributions; central to attacking build-up and chance creation.

- **Forwards (FW)**

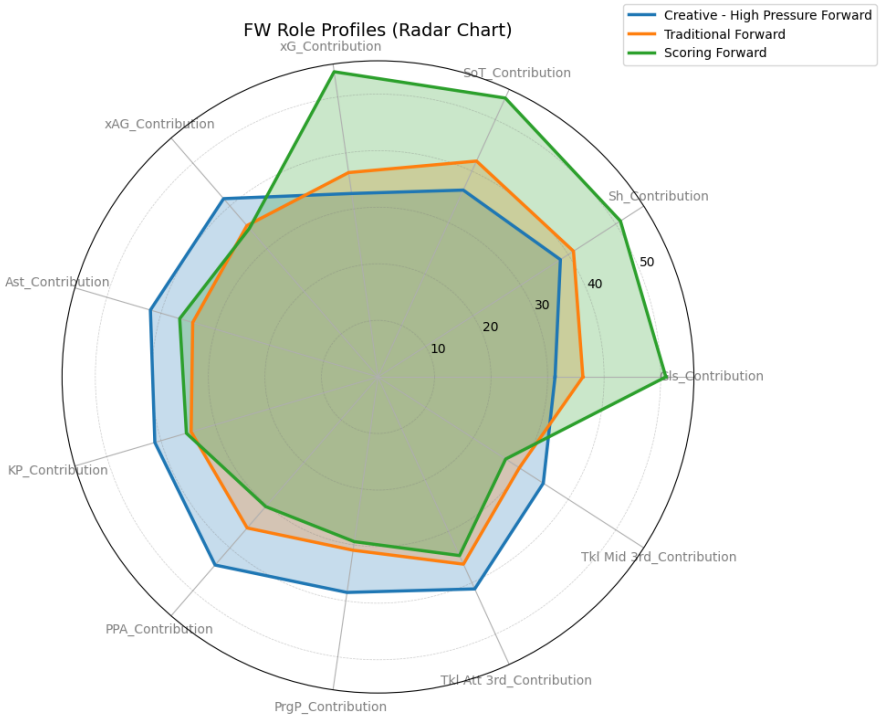


Fig. 3.26. Forward Role Profiles (Radar Chart)

  - **Creative – High-Pressure Forward**: Balanced goal threat with strong chance-creation signals (xAG, assists); often a wide forward or support striker.
  - **Traditional Forward**: Average across dimensions, indicating a general striker role involved in both finishing and link-up play.
  - **Scoring Forward**: Strong peaks in finishing metrics (goals, xG, shots); a poacher-style forward primarily focused on scoring.

Team-Level Line Performance Scores

Beyond individual player profiling, we evaluated team-level strength across the three tactical lines — defense, midfield, and attack — by constructing composite performance scores for each dimension. These scores offer a high-level summary of how well each team performed in the respective phase of play, based solely on objective, match-based statistics.

To ensure fairness and comparability, all features were normalized using Min-Max scaling over the full dataset spanning **all teams and all seasons** from 2020-2021 to

2023-2024. This allowed us to place every team on a common scale, thus making the scores season-independent and robust to year-specific outliers or fluctuations.

Defensive Score: For the defensive phase, we used a simple but interpretable inverse metric based on goals conceded:

$$\text{Team\_Def\_Score} = \frac{1}{\text{Goals Conceded (GA)}}$$

This formulation rewards teams that concede fewer goals and penalizes those with leaky defenses. After computing this inverse value, it was scaled to a $[20, 100]$ range across all teams to obtain the final normalized defense score for the same reasons explained earlier.

Midfield Score: The midfield score was designed to capture aspects of playmaking, progression, and creativity. It aggregates several normalized components:

$$\text{Team\_Mid\_Score} = \text{PrgP} + \text{KP} + \text{xAG} + \text{PPA}$$

This formulation emphasizes:

- **Progressive Passes (PrgP)** — breaking lines and advancing play
- **Key Passes (KP)** — leading to shot attempts
- **Expected Assists (xAG)** — creative threat
- **Passes into the Penalty Area (PPA)** — final third penetration

The sum was again normalized across all teams and seasons to ensure consistency.

Attacking Score: To measure attacking effectiveness, we combined two key normalized metrics:

$$\text{Team\_Att\_Score} = \text{Goals} + \text{xG}$$

This approach accounts for both actual finishing output (**Goals**) and chance quality (**Expected Goals**). It balances realized scoring efficiency with underlying attacking threat.

Top 5 Teams by Line (2023–2024)

TABLE 3.10. TOP 5 TEAMS DEFENSIVELY. 2023-2024

Team	Defense Score
Inter	91.00
Leverkusen	83.49
Real Madrid	77.14
Arsenal	69.26
Nice	69.26

TABLE 3.11. TOP 5 TEAMS IN MIDFIELD. 2023-2024

Team	Midfield Score
Liverpool	96.69
Arsenal	90.53
Manchester City	90.45
Bayern Munich	87.87
Tottenham	83.74

TABLE 3.12. TOP 5 TEAMS ATTACKING. 2023-2024

Team	Attack Score
Bayern Munich	96.01
Manchester City	93.36
Liverpool	90.78
Inter	88.03
Arsenal	86.07

3.6 Summary

This chapter presented a detailed overview of the data collection and analysis framework underlying this study. Data were sourced from three major football statistics platforms: FBref, SoFIFA, and Transfermarkt that were used in most of the prevoius work in this domain, selected for their reliability and complementary data types.

From these sources, multiple structured datasets were constructed, including position-specific datasets for goalkeepers, defenders, midfielders, and forwards, a unified player dataset all with inflation-adjusted market values, and a team-level dataset with aggregated

performance indicators. Additional engineered features—such as contextual environment scores (*team_score*, *league_score*, *country_score*) and player contribution ratios as well as performance scores per team —were introduced to enrich the dataset and enhance modeling precision.

Exploratory data analysis, supported by correlation matrices and scatter visualizations, was conducted to understand how individual attributes relate to player valuation. These insights directly inform the feature selection process for subsequent prediction models by identifying the most informative and relevant variables for each player position.

Contribution-based metrics were used to quantify individual player impact relative to their team's totals. These normalized values were critical in uncovering latent tactical roles through clustering, moving beyond traditional positional categories. The resulting role profiles—derived via Gaussian Mixture Models and visualized through radar charts—offered a more nuanced, data-driven view of how players operate on the pitch that will be of core importance in the optimization model.

On the team level, raw statistics were aggregated into interpretable composite scores (*Team_Def_Score*, *Team_Mid_Score*, *Team_Att_Score*) that capture each squad's effectiveness across different phases of play. These ratings offer a meaningful abstraction layer, making it easier to assess and compare team performance in strategic terms, and are of major importance for the optimization model.

Together, these datasets, engineered features, and analytical insights form the foundation for the next chapters.

The per-position feature analyses and correlations with market value will be the core of the predictive models for estimating players' market value in the next chapter.

Chapter 4

Market Value Prediction Model

4.1 Data Preparation

The dataset used in this study consisted of rich player-level information, combining performance statistics, physical characteristics, and contextual variables such as team and league affiliation. The primary objective was to develop accurate models capable of predicting each player's market value using this diverse set of inputs.

The data preparation pipeline consisted of two main stages: **feature selection** and **outlier detection and removal**.

4.1.1 Feature Selection

Feature selection was guided by both domain knowledge and statistical relevance. Variables were selected based on their theoretical connection to market valuation and supported by exploratory correlation analysis.

The final set included attributes related to:

- Technical quality (e.g., Finishing, Vision)
- Physical attributes (e.g., Stamina, Sprint Speed)
- Playmaking impact (e.g., Progressive Passes, Expected Assisted Goals)
- Defensive actions (e.g., Interceptions, Blocks)
- Contextual environment (e.g., *team_score*, *league_score*, *country_score*)

To reflect the unique skill profiles of different positions, feature selection was performed separately for the four major positional groups: *Goalkeepers (GK)*, *Defenders (DF)*, *Midfielders (MF)*, and *Forwards (FW)*. Each group had its own tailored set of predictive

features. Accordingly, separate models were trained for each position, ensuring that valuation estimates aligned with role-specific requirements.

4.1.2 Outlier Detection and Removal

Following feature selection, a two-pronged outlier removal procedure was applied:

- *Z-score filtering*: Observations with a z-score greater than 3 (in absolute value) relative to the distribution of market values were considered statistical outliers and removed.
- *Cross-validation with Sofifa value*: Players whose reported market value exceeded 1.5 times their in-game Sofifa Value were excluded, as such deviations suggested inconsistencies between performance and valuation data.

This dual strategy ensured that both statistical anomalies and logically inconsistent records were eliminated, resulting in a cleaner, more reliable modeling dataset.

4.1.3 Target Variable Definition

The prediction target was the player's *Market Value* in Euros (€). Due to the highly skewed nature of market values, a logarithmic transformation was applied to stabilize variance and improve learning performance:

$$\log(\text{MarketValue}_i) = \ln(\text{MarketValue}_i)$$

where MarketValue_i is the raw market value of player i, and $\ln(\cdot)$ denotes the natural logarithm.

This transformation offered several benefits:

- Reduced sensitivity to extreme values
- Improved residual normality
- Proportional error scaling, which aligns better with economic significance

To ensure robustness, models were evaluated under both targets:

- Raw *Market Value (€)*
- Transformed log(MarketValue), with predictions back-transformed via:

$$\text{MarketValue}_{\text{predicted}} = e^{\widehat{\log(\text{MarketValue})}}$$

This dual evaluation allowed for a more comprehensive interpretation of model accuracy across the entire valuation spectrum.

4.2 Model Building Strategy

The model development strategy was carefully structured to ensure robust, interpretable, and position-specific market value predictions.

4.2.1 Separate Models per Position

Recognizing the distinct roles and skill sets required for different player positions as explained earlier, separate models were built and trained for each of the four primary groups:

- **Goalkeepers (GK)**
- **Defenders (DF)**
- **Midfielders (MF)**
- **Forwards (FW)**

Each group had its own feature selection, model development, and evaluation pipeline, allowing models to specialize in the valuation dynamics specific to each role.

4.2.2 Choice of Modelling Algorithms

The selection of machine learning models was driven by both theoretical suitability and empirical evidence gathered from prior research, as discussed in the State of the Art section. Previous studies in the field of football analytics and market value prediction consistently emphasized the importance of flexible, ensemble-based models capable of capturing complex, non-linear interactions among features.

Based on these insights and their proven effectiveness in structured prediction tasks, the following modelling approaches were selected:

- **Random Forest Regressor (RF)**: An ensemble-based method that aggregates predictions from multiple decision trees. It handles non-linear feature relationships and interactions well, while also being robust to noise and overfitting. Its popularity in previous sports analytics literature further justifies its inclusion.
- **Extreme Gradient Boosting (XGBoost)**: A boosting technique that builds additive models in a sequential manner. XGBoost is known for its predictive power, scalability, and regularization capabilities, and has been used with notable success in previous market value estimation research as shown in the SoA section.

- **Stacked Ensemble Models**: A meta-learning approach that combines multiple base models—specifically RF and XGBoost—into a unified predictor using Ridge regression as the final estimator. This method aims to capture complementary patterns learned by each base model, as suggested by ensemble learning studies in the literature.

Each modelling technique was applied and evaluated under two different target configurations: the raw market value in Euros, and the logarithmically transformed market value. This dual setup ensured that model performance could be assessed across both economic and statistical dimensions and that model selection was grounded in theory and practice.

4.3 Model Evaluation and Results

4.3.1 Hyperparameter Optimization

To ensure accurate and interpretable predictions, all models were evaluated using both the raw market value target and its logarithmic transformation, log(MarketValue). For models trained on the log-transformed target, predictions were back-transformed into Euros using the exponential function to allow direct economic interpretation.

To maximize model performance, a systematic hyperparameter tuning procedure was employed. **Grid Search** with **cross-validation** was used to identify the optimal combinations of parameters for each model. Specifically:

- **Random Forests:** parameters such as maximum depth, number of estimators, minimum samples per leaf, and maximum features were optimized.
- **XGBoost:** parameters including learning rate, maximum depth, number of estimators, subsample ratio, and column sampling were tuned.
- **Stacked Models:** the tuning followed a rigorous strategy in which the base models (Random Forest and XGBoost) were first optimized independently. Then, multiple stacking configurations were tested, and the best-performing combination—based on validation metrics—was selected. This approach ensured that the final ensemble was constructed from the strongest base learners and stacking logic.

Given computational constraints, the hyperparameter grids were carefully designed to balance search comprehensiveness with practical feasibility. Instead of performing exhaustive searches, grids were limited to a targeted range of around 1000–2000 combinations, ensuring a meaningful yet efficient search space.

For models trained on log(MarketValue), predictions were back-transformed into Euros using the exponential function to allow direct economic interpretation.

4.3.2 Evaluation Metrics

Model performance was assessed using multiple complementary metrics to capture both absolute and relative prediction errors. The following evaluation metrics were used:

Mean Absolute Error (MAE) The MAE measures the average absolute deviation between the predicted values and the true values:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

where:

- n is the number of observations,
- $\hat{y}_i$ is the predicted value for observation i,
- y_i is the true value for observation i.

Interpretation: MAE provides a direct measure of the average magnitude of errors in Euros without considering their direction (under- or over-estimation). Lower MAE indicates better model accuracy.

Root Mean Squared Error (RMSE) The RMSE emphasizes larger errors by squaring the deviations before averaging:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

Interpretation: RMSE penalizes large errors more heavily than MAE and is thus sensitive to outliers. It provides a measure of the typical magnitude of prediction errors in the original unit (Euros).

Coefficient of Determination (R-squared, R^2) R^2 measures the proportion of variance in the true values explained by the predictions:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the true values.

Interpretation: R^2 ranges from 0 to 1. Higher values indicate better predictive performance. An R^2 close to 1 means that the model explains a large proportion of the variance in the target variable.

Root Mean Squared Percentage Error (RMSPE) RMSPE measures the relative prediction error as a percentage:

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \times 100$$

Interpretation: RMSPE expresses the average relative error in percentage terms. It is particularly useful when dealing with targets of varying scales, as it normalizes errors relative to the size of the true value. Lower RMSPE indicates better relative performance.

Separate evaluations were performed on both the raw Market Value (€) and the log-transformed target (log(MarketValue)), enabling a complete understanding of model behaviour across scales. For models trained on log(MarketValue), predictions were back-transformed into Euros using the exponential function to allow direct economic interpretation.

4.3.3 Goalkeeper (GK) Model Evaluation and Selection

1. Best Parameters Found

Random Forest Regressor – Raw Target

```
RandomForestRegressor(
    n_estimators=100,
    max_depth=15,
    max_features=0.5,
    min_samples_split=5,
    min_samples_leaf=1,
    random_state=42
)
```

Random Forest Regressor – Log Target

```
RandomForestRegressor(
    n_estimators=300,
    max_depth=15,
    max_features='auto',
    min_samples_split=2,
    min_samples_leaf=1,
    bootstrap=True,
    random_state=42
)
```

XGBoost Regressor – Raw Target

```
XGBRegressor(
    n_estimators=100,
    learning_rate=0.05,
    max_depth=3,
    subsample=0.8,
    colsample_bytree=0.8,
    objective='reg:squarederror',
    random_state=42)
```

XGBoost Regressor – Log Target

```
XGBRegressor(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=3,
    subsample=0.8,
    colsample_bytree=0.8,
    objective='reg:squarederror',
    random_state=42)
```

Stacked Regressor – Raw Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
                  n_estimators=100,
                  max_depth=None,
                  random_state=42)),
        ('xgb', XGBRegressor(
                  n_estimators=150,
                  learning_rate=0.1,
                  max_depth=3,
                  subsample=0.8,
                  colsample_bytree=0.8,
                  objective='reg:squarederror',
                  random_state=42))
                  ],
        final_estimator=make_pipeline(
                  StandardScaler(),
                  Ridge(alpha=10.0)),
                  passthrough=True)
```

Stacked Regressor – Log Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
                  n_estimators=100,
                  max_depth=10,
                  random_state=42)),
        ('xgb', XGBRegressor(
                  n_estimators=150,
                  learning_rate=0.05,
                  max_depth=5,
                  subsample=0.8,
                  colsample_bytree=0.8,
                  objective='reg:squarederror',
                  random_state=42))
                  ],
        final_estimator=make_pipeline(
                  StandardScaler(),
                  Ridge(alpha=10.0)),
                  passthrough=True)
```

2. Model Results Tables

A. Results on Log Target

Model	MAE (log)	RMSE (log)	R^2 (log)	RMSPE (%)
Random Forest	0.334	0.459	0.907	3.434
XGBoost	0.372	0.475	0.900	3.477
Stacked	0.368	0.493	0.892	3.63

TABLE 4.1. PERFORMANCE OF GK MODELS ON LOG-TRANSFORMED TARGET

B. Results on Raw Target

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest	1,173,965	2,166,532	0.942	127.09
XGBoost	1,357,409	2,859,486	0.899	166.15
Stacked	2393350	4674017	0.730	606.52

TABLE 4.2. PERFORMANCE OF GK MODELS ON RAW (UNTRANSFORMED) TARGET

C. Results on Back-Transformed Predictions from Log Models

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest (Log → €)	1,171,542	2,933,230	0.893	77.29
XGBoost (Log → €)	1,532,561	4,223,970	0.779	62.28
Stacked (Log → €)	2,647,370	11,892,404	-0.751	84.58

TABLE 4.3. PERFORMANCE AFTER INVERSE LOG-TRANSFORMING PREDICTIONS- GK

3. Result Analysis and Model Selection

Although XGBoost on the log-transformed target achieved the lowest RMSPE (62.28%) after back-transformation, the *Random Forest (Raw Target)* model exhibited the highest R^2 score of 0.942, reflecting stronger overall explanatory power. However, the relatively high RMSPE values across all models—especially in the raw and back-transformed settings—suggest that the results for goalkeepers are not fully reliable. This is likely due to the small sample size for this position, as goalkeepers are significantly fewer in number compared to outfield players. The stacked model, in particular, showed overfitting in log space and very poor generalization when transformed back.

Conclusion: Despite limitations in data quantity, the *Random Forest on Raw Target* remains the most balanced option based on R^2 and acceptable trade-offs in RMSPE.

4.3.4 Defender (DF) Model Evaluation and Selection

1. Best Parameters Found

Random Forest – Raw Target

```
RandomForestRegressor(
    n_estimators=100,
    max_depth=10,
    max_features=None,
    min_samples_split=2,
    min_samples_leaf=5,
    random_state=42
)
```

Random Forest – Log Target

```
RandomForestRegressor(
    n_estimators=300,
    max_depth=10,
    max_features=None,
    min_samples_split=2,
    min_samples_leaf=5,
    random_state=42
)
```

XGBoost – Raw/Log Target

```
XGBRegressor(
    n_estimators=200,
    learning_rate=0.05,
    max_depth=3,
    subsample=1.0,
    colsample_bytree=0.7,
    objective='reg:squarederror',
    random_state=42)
```

Stacked Regressor – Raw Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
            random_state=42,
            max_depth=10,
            max_features='sqrt',
            min_samples_leaf=1,
            min_samples_split=2,
            n_estimators=200
        )),
        ('xgb', XGBRegressor(
            objective='reg:squarederror',
            random_state=42,
            colsample_bytree=0.9,
            learning_rate=0.05,
            max_depth=3,
            n_estimators=200,
            subsample=1
        ))
    ],
    final_estimator=make_pipeline(StandardScaler())
)
```

Stacked Regressor – Log Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
            random_state=42,
            max_depth=10,
            max_features=None,
            min_samples_leaf=5,
            min_samples_split=2,
            n_estimators=300
        )),
        ('xgb', XGBRegressor(
            objective='reg:squarederror',
            random_state=42,
            colsample_bytree=0.7,
            learning_rate=0.05,
            max_depth=3,
            n_estimators=200,
            subsample=1.0
        ))
    ],
    final_estimator=make_pipeline(StandardScaler())
)
```

2. Model Results Tables

A. Results on Log Target

Model	MAE (log)	RMSE (log)	R^2 (log)	RMSPE (%)
Random Forest	0.193	0.234	0.959	1.609
XGBoost	0.183	0.225	0.962	1.553
Stacked	0.184	0.226	0.962	1.560

TABLE 4.4. PERFORMANCE OF DF MODELS ON LOG-TRANSFORMED TARGET

B. Results on Raw Target

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest	945,412	1,755,158	0.956	26.368
XGBoost	850,694	1,639,581	0.962	25.988
Stacked	898,630	1,668,106	0.960	24.480

TABLE 4.5. PERFORMANCE OF DF MODELS ON RAW (UNTRANSFORMED) TARGET

C. Results on Back-Transformed Predictions from Log Models

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest (Log → €)	948,821	1,796,933	0.954	24.885
XGBoost (Log → €)	847,769	1,531,520	0.967	24.012
Stacked (Log → €)	854,658	1,583,783	0.964	23.950

TABLE 4.6. PERFORMANCE AFTER INVERSE LOG-TRANSFORMING PREDICTIONS-DF

3. Result Analysis and Model Selection

Table 4.4 shows that all models perform similarly in log space, with XGBoost slightly outperforming others in both RMSE and RMSPE. However, log-scale metrics alone are not sufficient to judge real-world predictive accuracy.

Looking at Table 4.5 (Raw Target), XGBoost again delivers the best performance in terms of both RMSE (€1,639,581) and MAE (€850,694), with a strong R^2 of 0.962. Stacked models also perform competitively, suggesting that combining learners may help slightly in this context.

When evaluating back-transformed results in Table 4.6, the XGBoost model trained on log-transformed data and inverse-transformed during inference yields a RMSPE (24.012%) and lowest RMSE (€1,531,520). It also retains a high R^2 of 0.967, confirming its strong explanatory power post-transformation.

Conclusion:

The best-performing model overall is *XGBoost* trained on log-transformed targets, as it achieves the best balance between low prediction error (RMSE, MAE) and high goodness-of-fit (R^2) on both raw and back-transformed scales (see Figure 4.1).

Figure 4.2 illustrates the feature importances derived from the XGBoost model trained on the logarithmic transformation of the target variable (log(MarketValue)). These importances reflect how much each feature contributed to reducing the model's loss function (i.e., improving predictive accuracy).

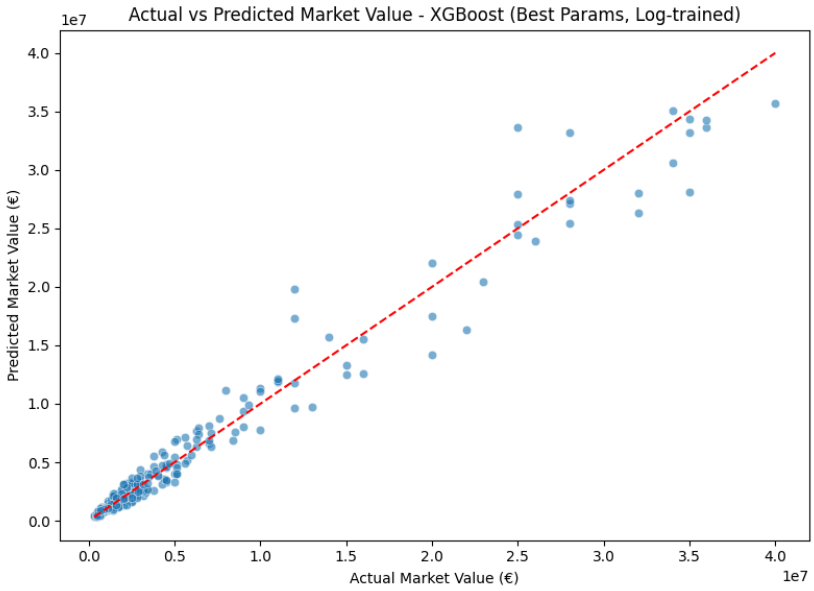


Fig. 4.1. DF Actual vs Predicted Log Market Value - XGBoost (log Target)

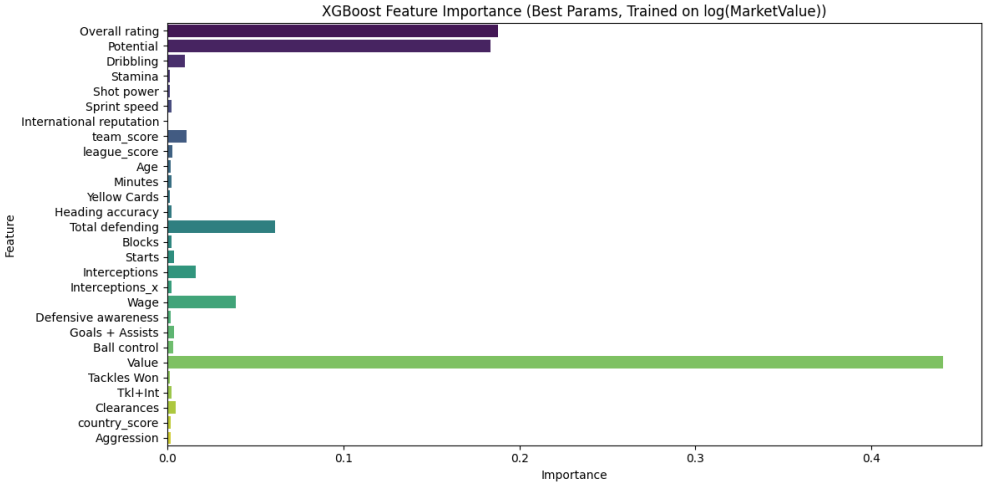


Fig. 4.2. Top DF Feature Importances - XGBoost

As seen in the figure 4.2 , the most influential feature is **Value**, which likely captures a proxy for the market perception already embedded in the FIFA-style attribute. This is followed by *Overall rating* and *Potential*, both of which summarize a player's current performance level and future development, respectively.

Other notable features include:

- **Total defending** and **Heading accuracy** — key technical attributes especially relevant for defensive players.

- **Wage** — another economically linked feature likely correlated with market value.

- **team_score**, **Interceptions**, and **country_score** — reflecting contextual or situational value influenced by league/team/nation quality.

Interestingly, some classical defensive metrics such as *Tackles Won* or *Blocks* were found to be less impactful than broader indicators like *Total defending*, *Interceptions*, and economic proxies.

4.Example Predictions Using RMSPE-Based Prediction Ranges

Table 4.7 illustrates example predictions produced by the final model trained on defenders. The table includes five players sampled across different market value ranges. For each player, we report the actual market value, the predicted market value (in euros), and a prediction range derived using the model's RMSPE.

The prediction bounds are not statistical confidence intervals but are calculated using the model's Root Mean Squared Percentage Error (RMSPE), which quantifies the typical relative prediction error. Specifically, we compute the lower and upper bounds of the range as:

$$\text{Lower Bound} = \hat{y} \cdot (1 - \text{RMSPE}), \quad \text{Upper Bound} = \hat{y} \cdot (1 + \text{RMSPE})$$

where $\hat{y}$ is the predicted value and RMSPE is expressed as a decimal (e.g., 24.01% becomes 0.2401). This yields an intuitive range around each prediction, showing the expected magnitude of error based on the model's overall performance.

Player	Actual (€)	Predicted (€)	Lower Bound (€)	Upper Bound (€)
K. Coronel	450,000	515,886	392,022	639,750
O. Kabak	11,000,000	9,279,018	7,051,125	11,506,910
M. Dembélé	28,000,000	27,142,246	20,625,392	33,659,100
M. Tambedou	1,100,000	1,089,971	828,269	1,351,673
J. Storey	2,100,000	2,199,082	1,671,082	2,727,082

TABLE 4.7. PREDICTED MARKET VALUES WITH RMSPE-BASED PREDICTION RANGES (RMSPE = 24.01%) - DF

4.3.5 Midfielder (MF) Model Evaluation and Selection

1. Best Parameters Found

Random Forest – Raw Target

```
RandomForestRegressor(
    n_estimators=200,
    max_depth=10,
    max_features='sqrt',
    min_samples_split=5,
    min_samples_leaf=1,
    random_state=42
)
```

Random Forest – Log Target

```
RandomForestRegressor(
    n_estimators=300,
    max_depth=15,
    max_features=None,
    min_samples_split=2,
    min_samples_leaf=2,
    random_state=42
)
```

XGBoost – Raw Target

```
XGBRegressor(
    n_estimators=100,
    learning_rate=0.05,
    max_depth=3,
    subsample=0.9,
    colsample_bytree=0.9,
    objective='reg:squarederror',
    random_state=42
)
```

XGBoost – Log Target

```
XGBRegressor(
    n_estimators=200,
    learning_rate=0.05,
    max_depth=5,
    subsample=0.7,
    colsample_bytree=0.7,
    objective='reg:squarederror',
    random_state=42
)
```

Stacked Regressor – Raw Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
            random_state=42,
            max_depth=10,
            max_features='sqrt',
            min_samples_leaf=1,
            min_samples_split=2,
            n_estimators=200
        )),
        ('xgb', XGBRegressor(
            objective='reg:squarederror',
            random_state=42,
            colsample_bytree=0.7,
            learning_rate=0.05,
            max_depth=3,
            n_estimators=200,
            subsample=0.9
        ))
    ],
    final_estimator=make_pipeline(StandardScaler())
)
```

Stacked Regressor – Log Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
            random_state=42,
            max_depth=10,
            max_features='sqrt',
            min_samples_leaf=1,
            min_samples_split=2,
            n_estimators=200
        )),
        ('xgb', XGBRegressor(
            objective='reg:squarederror',
            random_state=42,
            colsample_bytree=0.9,
            learning_rate=0.05,
            max_depth=3,
            n_estimators=100,
            subsample=0.9
        ))
    ],
    final_estimator=make_pipeline(StandardScaler())
)
```

2. Model Results Tables

A. Results on Log Target

Model	MAE (log)	RMSE (log)	R^2 (log)	RMSPE (%)
Random Forest	0.190	0.229	0.963	1.573
XGBoost	0.180	0.217	0.967	1.484
Stacked	0.186	0.220	0.966	1.510

TABLE 4.8. PERFORMANCE OF MF MODELS ON LOG-TRANSFORMED TARGET

B. Results on Raw Target

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest	970,032	1,906,675	0.949	30.179
XGBoost	1,014,038	1,995,469	0.944	44.778
Stacked	954,867	1,859,923	0.951	27.280

TABLE 4.9. PERFORMANCE OF MF MODELS ON RAW (UNTRANSFORMED) TARGET

C. Results on Back-Transformed Predictions from Log Models

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest (Log → €)	1,006,397	2,013,129	0.943	23.428
XGBoost (Log → €)	986,301	1,972,812	0.945	22.229
Stacked (Log → €)	976,763	1,892,288	0.950	22.500

TABLE 4.10. PERFORMANCE AFTER INVERSE LOG-TRANSFORMING PREDICTIONS-MF

3. Result Analysis and Model Selection

Table 4.8 shows that all three models perform comparably when trained on the log-transformed target. The Stacked model achieves the one of the lowest RMSE (0.220) and RMSPE (1.510%), followed closely by Random Forest and XGBoost. All models achieve high R^2 scores above 0.96, indicating strong explanatory power in log space.

However, as Table 4.9 shows that when trained and evaluated on raw target values, the stacked model achieves the lowest MAE (€954,867) and RMSE (€1,859,923), with a high R^2 of 0.951. The other models also performs competitively with similar results and slightly haigher MAE and RMSE but the Satcked model is showing really better results regarding the RMSPE.

To fairly compare models trained in log scale, we also evaluate their inverse-transformed predictions in Table 4.10. Here, the *Stacked model (Log → €)* outperforms others with the lowest MAE (€976,763), lowest RMSE (€1,892,288), and a great RMSPE (22.50%), while maintaining the strongest R^2 of 0.950 between the Back-Transformed models.

Conclusion: The best-performing model for midfielders is the *Stacked model trained on the log-transformed target*, as it achieves superior performance across all major metrics after inverse transformation,demonstrating both accurate prediction and generalization.

Figures 4.5 and 4.4 illustrate the feature importance rankings for the XGBoost and Random Forest models, respectively, which were used as base learners in the

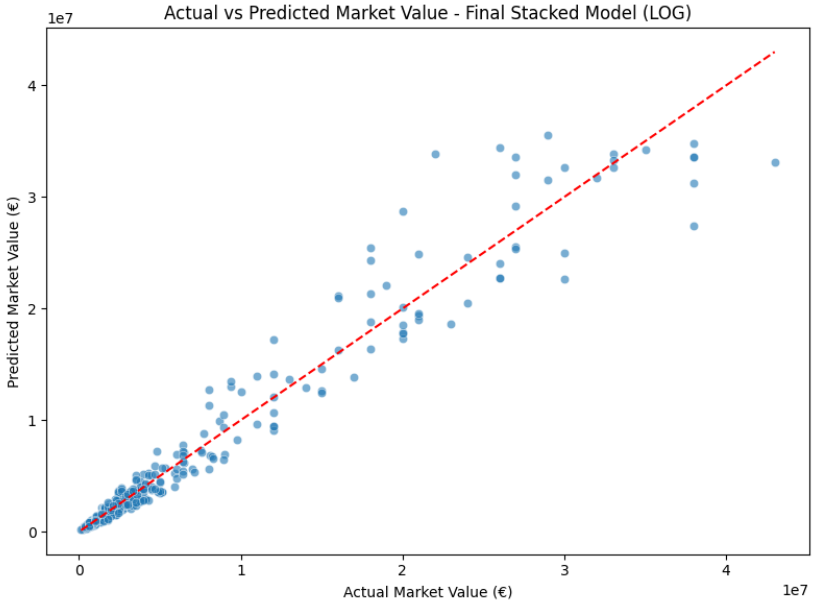


Fig. 4.3. MF Actual vs Predicted Log Market Value - XGBoost (log Target)

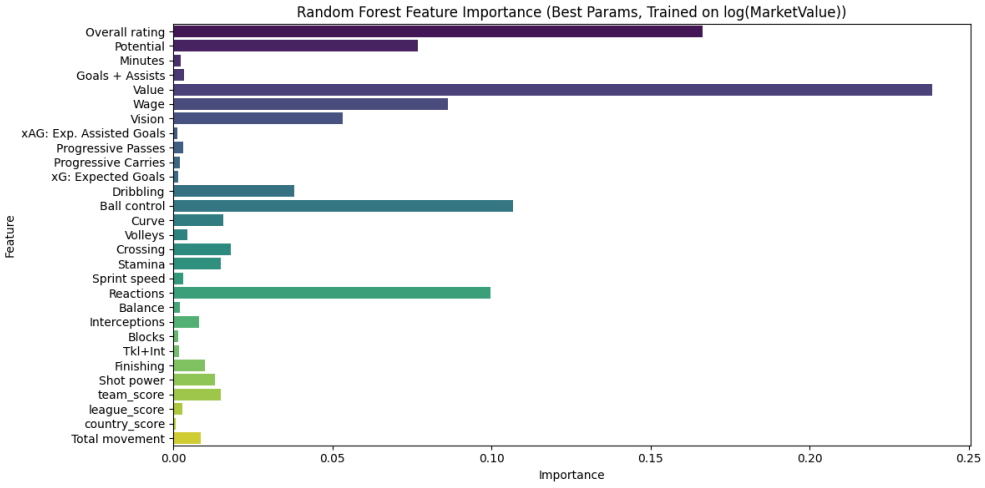


Fig. 4.4. Top MF Feature Importances - RF

stacked regression model for midfielders. These importances are extracted after training each model using the best hyperparameters on the log-transformed market value (`log(MarketValue)`) as the target variable.

In both models, the *Value* attribute provided by SoFIFA emerges as the most influential predictor. This is expected since it represents an internal valuation that already captures performance and reputation metrics. Additionally, attributes such as *Overall rating*, *Ball*

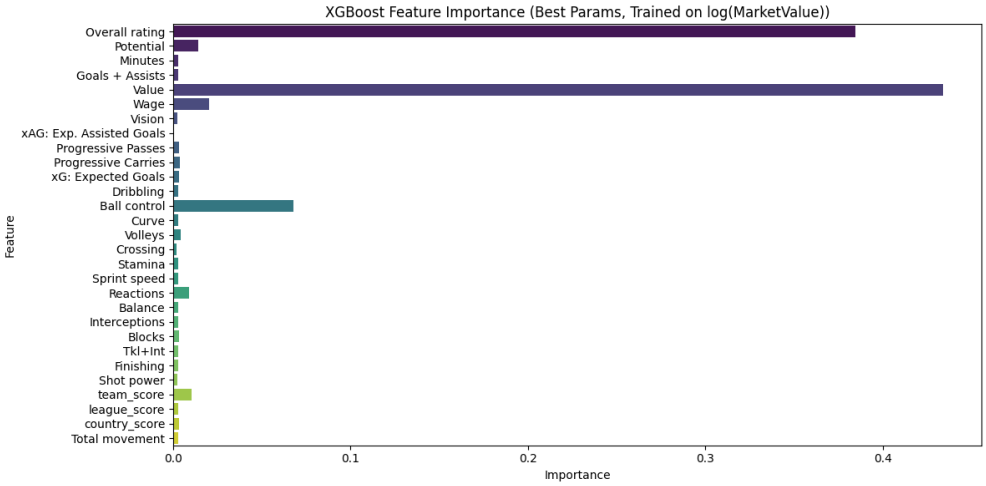


Fig. 4.5. Top MF Feature Importances - xgb

control, and *Reactions* appear prominently across both models, supporting the intuition that core technical and physical abilities are key in determining a midfielder's value. Features like *team_score*, *league_score*, and *country_score*, introduced in the feature engineering phase to reflect environmental context, also demonstrate moderate influence.

These plots not only provide model transparency but also help validate the relevance of the engineered and selected features, reinforcing both the methodological and empirical soundness of the modeling pipeline.

4.Example Predictions Using RMSPE-Based Prediction Ranges

Table 4.11 presents an example of predicted market values for five midfield players, along with their actual values and RMSPE-Based Prediction Ranges usinf the RMSPE of 22.5% obtained from the best-performing stacked model trained on log-transformed market values.

Player	**Actual (€)**	**Predicted (€)**	**Lower Bound (€)**	**Upper Bound (€)**
N. Chadli	600,000	635,444	492,469	778,419
J. Hove	2,500,000	2,883,263	2,234,529	3,531,998
M. Camara	33,000,000	31,722,060	24,584,596	38,859,523
Carlos Mendes Gomes	1,200,000	1,426,094	1,105,223	1,746,966
S. Esposito	5,000,000	4,411,287	3,418,747	5,403,826

TABLE 4.11. PREDICTED MARKET VALUES WITH
RMSPE-BASED PREDICTION RANGES (RMSPE = 22.5%) -MF

4.3.6 Forward (FW) Model Evaluation and Selection

1. Best Parameters Found

Random Forest – Raw Target

```
RandomForestRegressor(
    n_estimators=300,
    max_depth=None,
    max_features=None,
    min_samples_split=2,
    min_samples_leaf=5,
    random_state=42
)
```

Random Forest – Log Target

```
RandomForestRegressor(
    n_estimators=300,
    max_depth=15,
    max_features=None,
    min_samples_split=2,
    min_samples_leaf=5,
    random_state=42
)
```

XGBoost – Raw Target

```
XGBRegressor(
    n_estimators=100,
    learning_rate=0.05,
    max_depth=3,
    subsample=0.7,
    colsample_bytree=1.0,
    objective='reg:squarederror',
    random_state=42
)
```

XGBoost – Log Target

```
XGBRegressor(
    n_estimators=100,
    learning_rate=0.05,
    max_depth=3,
    subsample=0.7,
    colsample_bytree=0.5,
    objective='reg:squarederror',
    random_state=42
)
```

Stacked Regressor – Raw Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
            random_state=42,
            max_depth=10,
            max_features='sqrt',
            min_samples_leaf=1,
            min_samples_split=2,
            n_estimators=200
        )),
        ('xgb', XGBRegressor(
            objective='reg:squarederror',
            random_state=42,
            colsample_bytree=0.9,
            learning_rate=0.05,
            max_depth=3,
            n_estimators=100,
            subsample=0.9
        ))
    ],
    final_estimator=make_pipeline(StandardScaler())
)
```

Stacked Regressor – Log Target

```
StackingRegressor(
    estimators=[
        ('rf', RandomForestRegressor(
            random_state=42,
            max_depth=10,
            max_features='sqrt',
            min_samples_leaf=1,
            min_samples_split=2,
            n_estimators=200
        )),
        ('xgb', XGBRegressor(
            objective='reg:squarederror',
            random_state=42,
            colsample_bytree=0.9,
            learning_rate=0.05,
            max_depth=3,
            n_estimators=100,
            subsample=0.9
        ))
    ],
    final_estimator=make_pipeline(StandardScaler())
)
```

2. Model Results Tables

A. Results on Log Target

Model	MAE (log)	RMSE (log)	R^2 (log)	RMSPE (%)
Random Forest	0.187	0.219	0.972	1.489
XGBoost	0.193	0.223	0.971	1.519
Stacked	0.190	0.221	0.971	1.510

TABLE 4.12. PERFORMANCE OF FW MODELS ON LOG-TRANSFORMED TARGET

B. Results on Raw Target

C. Results on Back-Transformed Predictions from Log Models

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest	1,251,841	2,402,192	0.960	23,209
XGBoost	1,256,706	2,393,789	0.960	51.673
Stacked	1,230,041	2,433,257	0.959	27.840

TABLE 4.13. PERFORMANCE OF FW MODELS ON RAW (UNTRANSFORMED) TARGET

Model	MAE (€)	RMSE (€)	R^2	RMSPE (%)
Random Forest (Log → €)	1,255,483	2,484,884	0.957	22.226
XGBoost (Log → €)	1,330,162	2,645,485	0.951	22.898
Stacked (Log → €)	1,296,331	2,578,169	0.953	22.70

TABLE 4.14. PERFORMANCE AFTER INVERSE LOG-TRANSFORMING PREDICTIONS-FW

3. Result Analysis and Model Selection

Table 4.12 shows that all three models Random Forest, XGBoost, and the Stacked perform comparably in log space, with Random Forest slightly outperforming the others in both RMSE (0.219) and RMSPE (1.489%). This suggests a high-quality fit in logarithmic terms across all models.

In Table 4.13, when evaluating on the raw (untransformed) target, the Random Forest model again performs best in terms of MAE (€1,251,841), RMSE (€2,402,192), and R^2 (0.960). Notably, XGBoost shows a much higher RMSPE (51.673%), indicating instability in the raw scale despite similar error magnitudes.

Table 4.14 presents the evaluation results after inverse-transforming the log-predicted values. Here, the Random Forest model still achieves the best MAE (€1,255,483), RMSE (€2,484,884), and the highest R^2 (0.957). The stacked model remains competitive with slightly higher error but a lower RMSPE than XGBoost (22.70% vs. 22.898%).

Conclusion: The *Random Forest model trained on log-transformed target* yields the best overall performance for forwards. It consistently outperforms or matches other models across both scales and demonstrates the strongest explanatory power (highest R^2) with the lowest overall prediction errors.

Figure 4.7 illustrates the feature importances from the best-performing Random Forest model trained on the log-transformed market value for forwards. Remarkably, only two features—*Value* and to a much lesser extent *Overall rating*—contribute meaningfully to the model's predictions, with "Value" overwhelmingly dominating the importance distribution.

This behavior contrasts sharply with the feature importance distributions observed in the defender and midfielder models and even the other models used on the same forwards

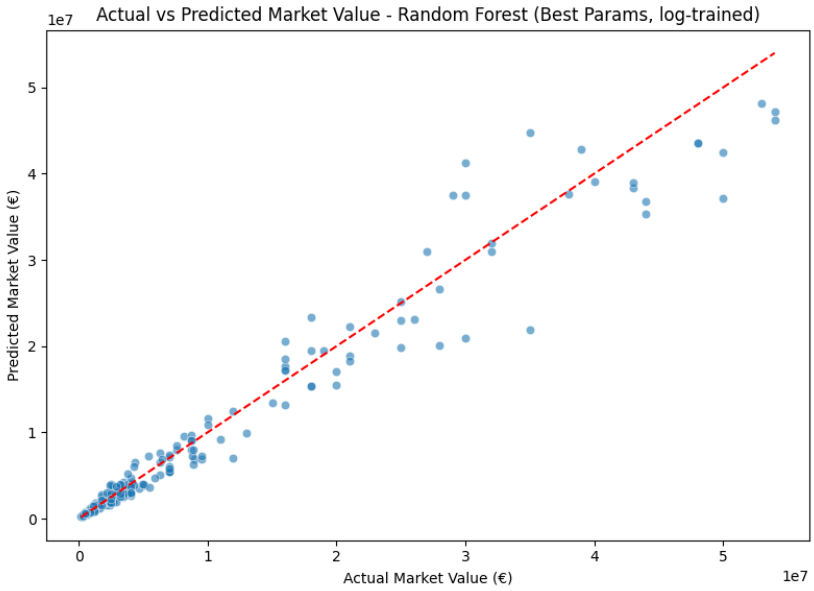


Fig. 4.6. FW Actual vs Predicted Log Market Value - RF (log Target)

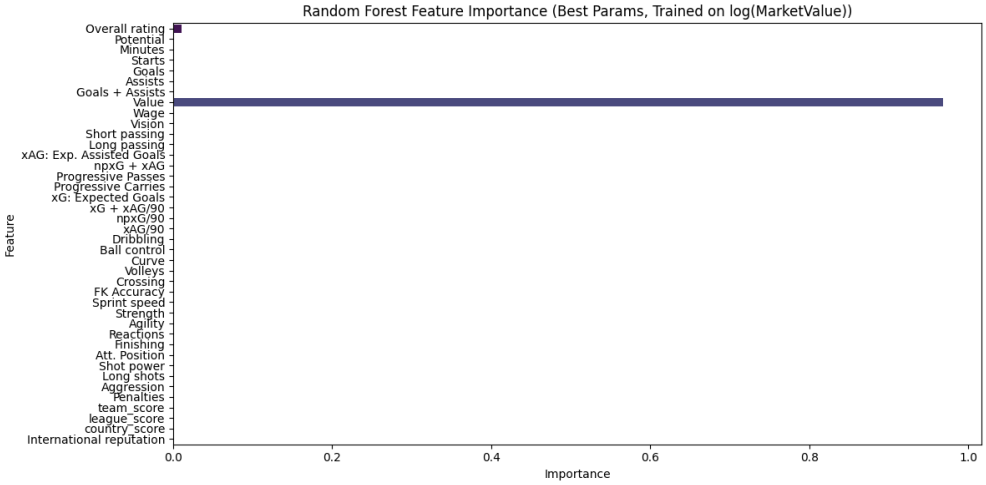


Fig. 4.7. Top FW Feature Importances - RF(log Target)

data set, where a more diverse set of technical, physical, and contextual attributes were leveraged (see previous subsections). Despite this narrow reliance on only two predictors, the model achieved the best performance across all evaluation metrics for forwards, as reported in Tables 4.12 - 4.14.

4.Example Predictions Using RMSPE-Based Prediction Ranges

Table 4.15 presents an example of predicted market values for five forward players, along with their actual values and prediction intervals computed using the RMSPE of 22.226% obtained from the best-performing Random Forest model trained on log-transformed market values.

Player	Actual (€)	Predicted (€)	Lower Bound (€)	Upper Bound (€)
M. El Idrissy	940,000	1,028,646	800,019	1,257,274
D. Camara	1,300,000	1,309,140	1,018,170	1,600,110
M. Ødegaard	53,000,000	49,210,615	38,273,063	60,148,166
L. Zeefuik	1,000,000	1,178,338	916,440	1,440,235
P. Nebel	3,500,000	3,614,951	2,811,492	4,418,410

TABLE 4.15. PREDICTED MARKET VALUES WITH
RMSPE-BASED PREDICTION RANGES (RMSPE = 22.226%) – FW

4.4 Summary

This study evaluated the performance of three machine learning models—*Random Forest*, *XGBoost*, and a *Stacked Regressor*—for predicting football players' market values across four positional groups: goalkeepers (GK), defenders (DF), midfielders (MF), and forwards (FW). Each model was tested under both raw and log-transformed target settings, with evaluations conducted using a comprehensive suite of metrics: MAE, RMSE, R^2, and RMSPE. Additionally, back-transformed predictions from log-space were assessed to approximate real-world performance.

Across positions, log transformation generally improved the stability of models and reduced percentage-based error (RMSPE), particularly for outfield players. However, in some cases, notably for goalkeepers, raw target modeling yielded better R^2 and lower absolute errors, suggesting limited value of log-transformation in low-data regimes.

For **Goalkeepers (GK)**, the *Random Forest on the raw target* achieved the highest R^2 (0.942) and lowest RMSE, making it the most balanced choice despite a relatively high RMSPE. The stacked model, although complex, showed poor generalization post back-transformation, likely due to overfitting and small sample size.

In contrast, for **Defenders (DF)**, the *XGBoost model trained on the log-transformed target* and evaluated after inverse transformation produced the best results: lowest RMSE (€1.53M), lowest RMSPE (24.01%), and highest R^2 (0.967).

Similarly, for **Midfielders (MF)**, the *Stacked Regressor trained on log-transformed targets* achieved the strongest overall performance. It yielded the lowest RMSE (€1.89M) and best R^2 (0.950) among back-transformed predictions, while offering a favorable

RMSPE (22.50%). The use of combined learners allowed this model to benefit from both Random Forest's robustness and XGBoost's precision.

For **Forwards (FW)**, the *Random Forest model on the log-transformed target* was the most reliable. It consistently outperformed or matched other models in RMSE (€2.48M), R^2 (0.957), and RMSPE (22.23%) and very close results for the same model applied on raw data. Interestingly, this model relied heavily on only two features—*Value* and *Overall rating*—highlighting the extent to which forward valuations may be concentrated around reputation and general ability.

To further illustrate the models' predictive capabilities, several example predictions were presented for players in each position. These included the actual market value, predicted value, and an estimated prediction range calculated using the model's RMSPE. The lower and upper bounds provided an intuitive interval estimate of the expected prediction error based on each model's typical relative accuracy.

While optimal model configurations vary by player role, models trained on log-transformed targets and evaluated in the back-transformed space generally provide the best trade-off between accuracy and interpretability. For defenders and midfielders, log-transformed modeling significantly improved generalization. For goalkeepers and forwards, simpler Random Forest models proved more effective, especially when raw target values were retained. The results validate the importance of positional segmentation and transformation choice in football player valuation, and they underscore the necessity of combining contextual and technical features for robust performance.

Chapter 5

Optimization of Player Acquisition Strategy

5.1 Team Needs Vector Construction

A core element of the optimization framework is the **team needs vector**, which quantifies how urgently a team requires improvement across various performance metrics. This vector serves as the reference point in multiple components of our analysis, including player selection, role gap identification, and optimization modeling.

5.1.1 Source: Team-Level Performance Metrics

The raw input for generating the needs vector is the *team-level performance dataset*, which includes standardized performance scores across multiple footballing dimensions:

- Defensive metrics (e.g., tackles, interceptions, blocks),
- Midfield metrics (e.g., progressive passes, passes into the penalty area),
- Attacking metrics (e.g., key passes, expected goals, goals).

These features are initially normalized using Min-Max scaling to ensure comparability across teams.

5.1.2 Inverse Normalization and Rescaling

To convert performance into a measure of need, we apply an **inverse normalization** followed by a **rescaling to the range [20, 100]**. The transformation is defined as:

$$\text{Need}_{i,j} = 20 + \left(1 - \frac{x_{i,j} - \min_j}{\max_j - \min_j}\right) \cdot 80$$

where:

- $x_{i,j}$ is the value of team i for feature j,
- $\min_j$ and $\max_j$ are the minimum and maximum values of feature j across all teams,
- $\text{Need}_{i,j}$ is the final score indicating how much team i lacks in feature j, scaled between 20 (lowest need) and 100 (highest need).

This transformation ensures that:

- Teams performing poorly on a metric will have high need scores (closer to 100),
- Teams performing well will have low need scores (closer to 20).

This approach assigns a minimum value of 20 instead of 0 to reflect that even the best-performing teams may still have room for improvement, and thus should not have a need score of zero.

5.1.3 Output: The Team Needs Dataset

The result of this transformation is the **team needs dataset**, a matrix where rows represent teams and columns represent feature-specific needs. This dataset is used to:

1. Identify external players whose feature contributions closely match a team's deficiencies (Section 5.2).
2. Detect roles that are underrepresented based on tactical patterns and unmet feature needs (Section 5.3).
3. Inform the optimization model with a quantitative profile of team requirements, indicating how much a team lacks in each specific feature.
4. Derive team-level performance scores per tactical line (DEF, MID, ATT) from the needs dataset, providing an overview of how poorly each line is performing.

This formulation aligns with the central goal of the optimization pipeline: to improve a team's weakest areas by selecting players whose contributions fill the most significant gaps.

5.2 Player Pool Construction Based on Team Needs

The first step in the optimization framework involves constructing a pool of suitable transfer targets for a given team. This is accomplished by identifying players whose feature contributions closely align with the team's most pressing needs, as defined in Section 5.1.

5.2.1 Line-Based Feature Grouping

To ensure position-specific relevance, player contributions are divided into three tactical lines—defense (DEF), midfield (MID), and attack (ATT). Each line is associated with a curated set of performance features that reflect core responsibilities on the pitch:

- **DEF:** `Tkl_Contribution`, `TklW_Contribution`, `Int_Contribution`, `Blocks_Contribution`, `Clr_Contribution`, `Tkl Def 3rd_Contribution`, `Tkl Mid 3rd_Contribution`, `Tkl Att 3rd_Contribution`
- **MID:** `PPA_Contribution`, `PrgP_Contribution`, `xAG_Contribution`, `Ast_Contribution`, `KP_Contribution`
- **ATT:** `Sh_Contribution`, `SoT_Contribution`, `Gls_Contribution`, `xG_Contribution`

These feature groups allow the model to focus on players whose contributions are most relevant for improving a given tactical line.

5.2.2 Similarity-Based Matching Procedure

For each line, the following steps are applied:

1. **Extract Team Profile:** Filter the current squad's players who belong to the targeted line and are under a specified age threshold (e.g., ≤ 2). Compute the average contribution vector across the relevant features to represent the team's profile for that line.
2. **Generate Candidate Pool:** Identify players from other squads who play in the same line and also meet the age constraint. These form the set of potential transfer targets.
3. **Compute Cosine Similarity:** For each candidate, calculate the cosine similarity between their individual feature vector and the team's profile vector:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{||A||\,||B||}$$

where A is the team's average feature vector, and B is the candidate player's vector.

4. **Rank and Select:** Rank all candidates based on their similarity score and retain the top N most similar players for each tactical line.

5.2.3 Output Structure and Practical Utility

The output is a structured player pool categorized by tactical line, containing only those players whose statistical profiles align most closely with the team's current needs. This serves two essential purposes:

- It **narrows the scouting scope**, focusing the search on players who are statistically the best fit for reinforcing weak points.
- It **recommends players likely to add immediate value**, as their strengths directly target the areas where the team is underperforming.

This filtered player pool is then passed to the optimization model described in Section 5.4, where further constraints (e.g., budget, role balance) are applied to finalize transfer decisions.

5.2.4 Illustrative Examples: Strong vs Weak Teams

To demonstrate the effectiveness of the similarity-based player pool construction approach, we present two contrasting case studies: **Real Madrid**, representing a high-performing team with relatively balanced needs, and **Almería**, a lower-ranked team with more pronounced deficiencies across all lines. In both examples, an age limit of 30 years was applied to the candidate pool in order to provide a broader perspective on available talent. If a team prefers to prioritize youth development, a stricter age filter can be applied to focus specifically on younger prospects.

Example 1: Real Madrid (Top Team) Despite being a strong team, Real Madrid still exhibits some areas that could be bolstered. The following tables display the top 5 highest market value players recommended for each line, alongside the top 3 team needs extracted from the team's performance data.

TABLE 5.1. TOP 5 MARKET VALUE DEF PLAYERS FOR REAL MADRID

Player	Age	Squad	Market Value	Similarity
Alphonso Davies	22	Bayern Munich	€50M	0.997
Levi Colwill	20	Chelsea	€50M	0.998
Lisandro Martínez	25	Man Utd	€45M	0.999
Jakub Kiwior	23	Arsenal	€30M	0.999
Andrew Robertson	29	Liverpool	€30M	0.998

TABLE 5.2. TOP 3 DEF NEEDS FOR REAL MADRID

Feature	**Need Value**
Clr	87.01
tkl Def 3rd	78.27
tkl Mid 3rd	77.55

TABLE 5.3. TOP 5 MARKET VALUE MID PLAYERS FOR REAL MADRID

Player	**Age**	**Squad**	**Market Value**	**Similarity**
K. Thuram	22	Nice	€35M	1.000
João Gomes	22	Wolves	€35M	1.000
J. Ramsey	22	Aston Villa	€35M	1.000
Enzo Millot	21	Stuttgart	€30M	1.000
John McGinn	28	Aston Villa	€30M	1.000

TABLE 5.4. TOP 3 MID NEEDS FOR REAL MADRID

Feature	**Need Value**
PPA	51.72
KP	46.81
xAG	44.64

TABLE 5.5. TOP 5 MARKET VALUE ATT PLAYERS FOR REAL MADRID

Player	**Age**	**Squad**	**Market Value**	**Similarity**
Ademola Lookman	25	Atalanta	€40M	0.999
Serge Gnabry	28	Bayern Munich	€40M	0.999
A. Mitrovic	28	Fulham	€28M	0.999
J. Lindström	23	Frankfurt	€22M	0.999
Romain Faivre	25	Bournemouth	€15M	0.999

TABLE 5.6. TOP 3 ATT NEEDS FOR REAL MADRID

Feature	**Need Value**
Sh	53.05
xG	46.43
SoT	31.14

Example 2: Almería (Weaker Team) In contrast, Almería's top needs are far more severe across all lines. The system recommends more critical reinforcements, and the overall team scores per line are significantly higher, indicating lower current performance and thus higher need.

TABLE 5.7. TOP 5 MARKET VALUE DEF PLAYERS FOR ALMERÍA

Player	**Age**	**Squad**	**Market Value**	**Similarity**
Éder Militão	25	Real Madrid	€60M	0.998
Lisandro Martínez	25	Man Utd	€45M	0.999
M. Akanji	28	Man City	€45M	0.998
Luke Shaw	28	Man Utd	€32M	0.997
Jakub Kiwior	23	Arsenal	€30M	0.999

TABLE 5.8. TOP 3 DEF NEEDS FOR ALMERÍA

Feature	**Need Value**
tkl Att 3rd	87.16
tkl Mid 3rd	72.95
TklW	58.94

TABLE 5.9. TOP 5 MARKET VALUE MID PLAYERS FOR ALMERÍA

Player	**Age**	**Squad**	**Market Value**	**Similarity**
N. Barella	26	Inter	€80M	1.000
F. de Jong	26	Barcelona	€70M	1.000
W. Zaïre-Emery	17	Paris SG	€60M	1.000
Curtis Jones	22	Liverpool	€35M	1.000
A. Rabiot	28	Juventus	€35M	1.000

TABLE 5.10. TOP 3 MID NEEDS FOR ALMERÍA

Feature	**Need Value**
PPA	87.88
PrgP	86.10
Ast	82.37

TABLE 5.11. TOP 5 MARKET VALUE ATT PLAYERS FOR ALMERÍA

Player	**Age**	**Squad**	**Market Value**	**Similarity**
A. Mitrović	28	Fulham	€28M	1.000
R. Faivre	25	Bournemouth	€15M	1.000
Ben Doak	17	Liverpool	€10M	1.000
B. Traoré	27	Aston Villa	€9M	1.000
J. Anthony	23	Bournemouth	€6M	1.000

TABLE 5.12. TOP 3 ATT NEEDS FOR ALMERÍA

Feature	**Need Value**
Gls	80.29
xG	74.40
Sh	73.27

5.3 Team Role Gap Identification

To move from feature-level needs to actionable recruitment strategy, we identify which **tactical roles** are most urgently required by each team. This is achieved by comparing the team's feature-based need vector with predefined role vectors derived from unsupervised player clustering (see Section 3.5.2).

Analytical Framework

Let $T \in \mathbb{R}^d$ denote the team's normalized need vector across d features, and let $\{R_1, \ldots, R_k\} \subset \mathbb{R}^d$ be the set of role centroids obtained from clustering top players per line. For each role vector R_i, we compute a cosine similarity score:

$$\text{sim}(T, R_i) = \frac{T \cdot R_i}{||T|| \cdot ||R_i||}$$

This measures how well a role's contribution profile aligns with the team's current deficiencies. The result is then scaled by the team's overall need in that line (DEF, MID, ATT) via a weight w_p:

$$\text{Weighted Similarity}(R_i) = \text{sim}(T, R_i) \cdot w_p$$

Examples:

Figures 5.1 and 5.2 present the top 10 most needed tactical roles for Real Madrid and Almería, respectively, based on the above similarity scores.

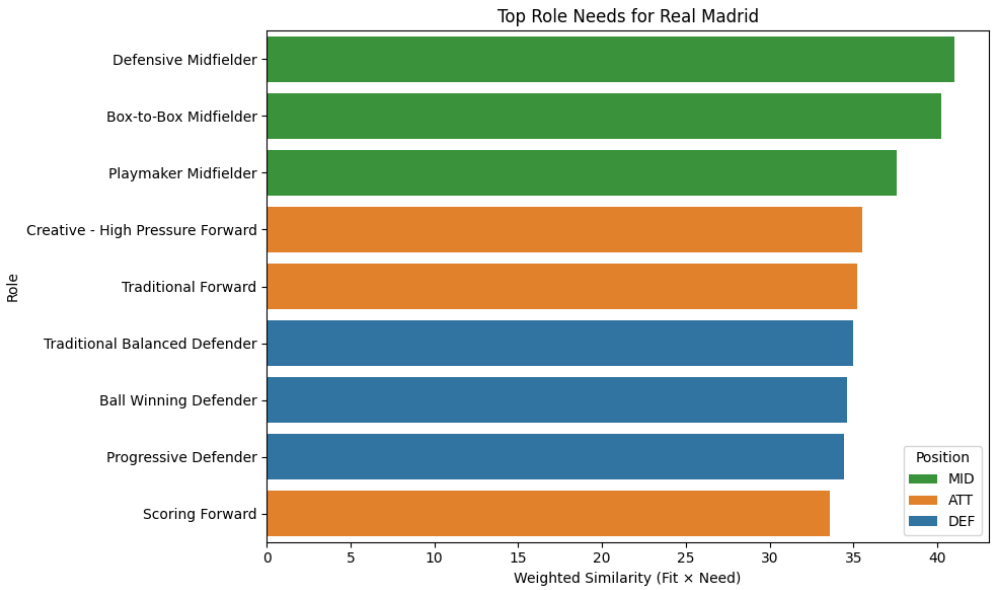


Fig. 5.1. Top Tactical Role Needs for Real Madrid (Fixed Color Mapping)

Real Madrid Despite being a high-performing team, Real Madrid still exhibits targeted tactical gaps. The top three needed roles are all midfield-based:

- **Defensive Midfielder**
- **Box-to-Box Midfielder**
- **Playmaker Midfielder**

This indicates that while their general team balance is strong, there is a relative underperformance in midfield contributions — particularly in defensive coverage, ball progression, and chance creation. Forward roles like *Creative Forward* and *Traditional Forward* also rank high, suggesting secondary deficiencies in support and finishing. Interestingly, defensive roles are ranked lower, which aligns with the club's overall strong defensive statistics.

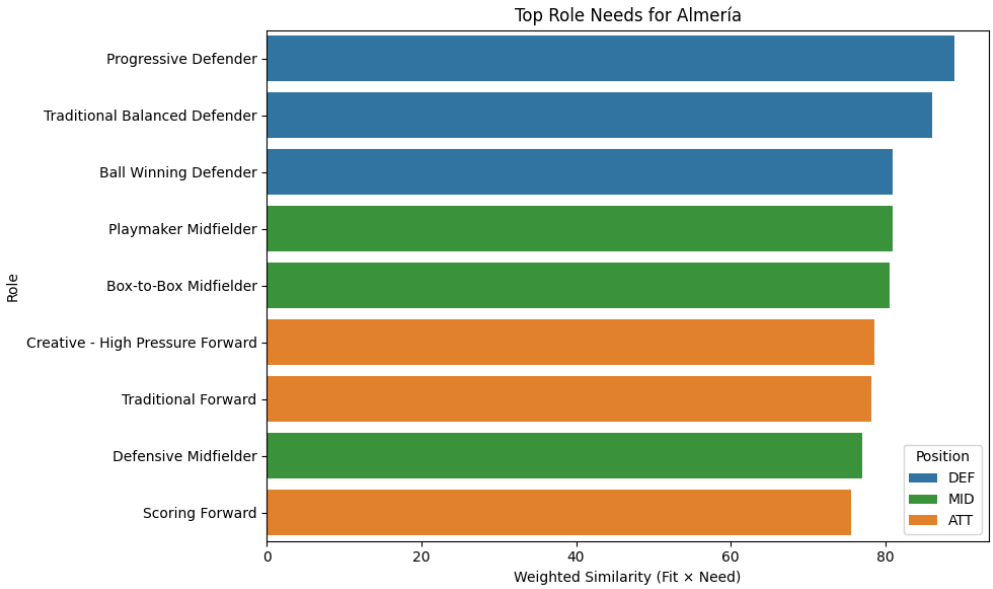


Fig. 5.2. Top Tactical Role Needs for Almería (Fixed Color Mapping)

Almería In contrast, Almería — a struggling team — shows severe tactical deficiencies across all lines. The highest-ranked needs include:

- **Progressive Defender**
- **Traditional Balanced Defender**
- **Ball Winning Defender**

This suggests acute structural issues in defense, where the team lacks both possession-breaking and ball-progressing profiles. Additionally, midfield roles like *Playmaker* and *Box-to-Box Midfielder* are ranked highly, pointing to weak progression and linking play. Finally, forward roles like *Creative Forward*, *Scoring Forward*, and *Traditional Forward* are also needed — indicating low attacking efficiency.

Strategic Implication

The role ranking analysis provides a more nuanced interpretation of team needs than line-based scores. For instance:

- Real Madrid does not need "midfielders" in general — it needs specific profiles such as *defensive anchors* and *advanced creators*.
- Almería needs both quantity and variety — defenders who can recover possession, midfielders who can progress the ball, and forwards who can convert chances.

These insights directly feed into the optimization model (Section 5.4), which uses these ranked role needs to prioritize players that match both the tactical function and statistical contribution required by each squad.

5.4 Optimization Model for Player Recruitment

To convert team-specific tactical needs into optimal transfer decisions, we formulate a constrained optimization model. The goal is to select a set of players who best fill the team's functional gaps while respecting practical constraints such as budget, positional roles, and tactical role diversity.

5.4.1 Modeling Framework

The optimization model is implemented using the **Pyomo** modeling language in Python and solved using the **Gurobi** optimizer. The problem is formulated as a **Binary Integer Program (BIP)**, where each decision variable represents the selection status of a player. That is, each variable $x_i \in \{0,1\}$ denotes whether player i is included in the selected squad.

The **player pool** used in the model is constructed through the similarity-based filtering procedure described in Section 5.2. It consists of players from other clubs who statistically match the target team's needs based on their contribution profile.
The **minimum and maximum number of players per specific tactical role** are automatically determined as follows:

- The team's ranked tactical role needs (from similarity-based role ranking) are used to proportionally allocate player quotas across roles.
- For each role with sufficient need and available players, a non-zero minimum is assigned.
- Quotas are capped based on the total number of players to be selected and adjusted to match player availability.

This ensures that the optimization respects both the relative urgency of different tactical roles and practical feasibility based on the current transfer market.

5.4.2 User-Defined Inputs

Before building the model, the user specifies:

- **Target team** (`target_squad`) – the club that seeks to reinforce its squad.
- **Number of players to sign** (k).
- **Budget constraint** (in euros).
- **Age limit** for eligible players.

5.4.3 Decision Variables

Let $x_i \in \{0,1\}$ be a binary variable defined as:

$$x_i = \begin{cases} 1 & \text{if player } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

5.4.4 Parameters and Inputs

The model relies on the following key inputs:

- f_{ij}: Contribution of player i to feature j.
- n_j: Need of the target team for feature j.
- s_i: Strength of player i's current team in the relevant line (DEF, MID, ATT).
- N_l: Need score of the target team for line $l \in \{\text{DEF}, \text{MID}, \text{ATT}\}$.
- c_i: Market value (cost) of player i.
- r_i: Specific tactical role assigned to player i.

- R: Set of all specific tactical roles.
- P: Set of all candidate players.
- F: Set of all contribution features.

5.4.5 Objective Function

The model aims to **maximize the total weighted contribution** of selected players. The score for each player combines:

1. **How well they contribute to features that the team needs** ($f_{ij} \cdot n_j$),
2. **The quality of the player's source team** (s_i),
3. **The need of the target team in the player's broad line** (N_l).

The mathematical form of the objective is:

$$\max \sum_{i \in P} x_i \cdot s_i \cdot N_{l_i} \cdot \left(\frac{1}{|F_{l_i}|} \sum_{j \in F_{l_i}} f_{ij} \cdot n_j \right)$$

Where:

- l_i is the broad role of player i,
- F_{l_i} is the set of features associated with role l_i.

This formulation ensures that players are rewarded not simply for absolute ability, but for how well their contributions align with the team's most pressing tactical and structural needs. To avoid bias toward lines with a greater number of features, the contribution term is normalized by the number of features in each broad role, ensuring that each line (DEF, MID, ATT) contributes comparably to the objective regardless of dimensionality.

5.4.6 Constraints

The model includes the following constraints:

1. **Squad Size Constraint**:
$$\sum_{i \in P} x_i = k$$
2. **Budget Constraint**:
$$\sum_{i \in P} x_i \cdot c_i \leq \text{Budget}$$

3. **Minimum Role Coverage**:

$$\sum_{i \in P : r_i = r} x_i \geq \text{Min}_r \quad \forall r \in R$$

4. **Maximum Role Quota**:

$$\sum_{i \in P : r_i = r} x_i \leq \text{Max}_r \quad \forall r \in R$$

These constraints ensure that the selected squad respects not only financial and numeric limitations but also the tactical needs diagnosed earlier in the pipeline.

5.4.7 Solution Output

Once solved, the model returns:

- A list of selected players with their tactical role, position, age, cost, and source club.
- The total market value and remaining budget.
- Distribution of selected players by tactical line (DEF, MID, ATT) and specific roles.
- The final objective score achieved by the selected squad.

5.4.8 Solver

The optimization problem is solved using the `Gurobi` solver, which offers high performance for binary and mixed-integer programming. Its ability to handle large-scale decision spaces makes it suitable for real-world football recruitment problems where hundreds of candidates must be evaluated under multi-dimensional constraints.

5.4.9 Illustrative Examples: Real Madrid vs. Almería

To demonstrate the versatility of our optimization framework, we present two contrasting case studies: **Real Madrid**, a top-tier club with a well-balanced squad and moderate reinforcement needs, and **Almería**, a lower-ranked team with significant tactical deficiencies and tighter resource constraints. The player pool, team needs, and tactical role rankings used in these examples were generated using the procedure outlined in Sections 5.2 and 5.3.

The following table summarizes the user-defined inputs for both cases:

TABLE 5.13. USER-DEFINED MODEL INPUTS FOR REAL MADRID AND ALMERÍA

Parameter	**Real Madrid**	**Almería**
Target Team	Real Madrid	Almería
Max Age	27	30
Max Players (k)	4	5
Budget	€120,000,000	€40,000,000

For each team, the optimization model selected a set of players maximizing the performance and whose contributions match the club's most urgent tactical needs, while satisfying all imposed constraints. The results are summarized below.

Selected Players for Real Madrid

TABLE 5.14. SELECTED PLAYERS FOR REAL MADRID

Player	**Role**	**Pos**	**Age**	**Market Value (€)**	**Squad**
João Gomes	Defensive Mf	MF	22.0	35,000,000	Wolves
Habib Diarra	Box-to-Box Mf	MF,FW	19.0	18,000,000	Strasbourg
Angel Gomes	Playmaker Mf	MF	22.0	25,000,000	Lille
Adu Ares	Creative – High Pressure Fw	FW	21.0	2,000,000	Athletic Club

Total Market Value: €80,000,000.00

This solution reflects a strategy aimed at reinforcing midfield creativity and attacking depth, while respecting the club's focus on youth and long-term value. The optimizer not only maximized overall performance but also saved money by identifying high-impact players who are not necessarily the most expensive — effectively uncovering valuable, under-the-radar talents that fit Real Madrid's tactical role needs (see Table 5.1).

Selected Players for Almería

TABLE 5.15. SELECTED PLAYERS FOR ALMERÍA

Player	Role	Pos	Age	Market Value (€)	Squad
David Raum	Progressive Df	DF	25.0	20,000,000	RB Leipzig
Jorge Cuenca	Ball Winning Df	DF	23.0	6,000,000	Villarreal
Ricardo Rodríguez	Traditional Balanced Df	DF	30.0	3,500,000	Torino
Julien Ponceau	Box-to-Box Mf	MF	22.0	3,500,000	Lorient
Kevin Stöger	Playmaker Mf	MF	29.0	5,000,000	Bochum

Total Market Value: €38,000,000.00

Almería's strategy focuses on addressing severe gaps in defense and midfield through cost-efficient transfers, including older, experienced players. The model adapts by selecting mature profiles that offer tactical fit and immediate impact, even within a limited budget. Notably, the optimizer successfully maximized performance while staying €2 million under budget, uncovering valuable players who deliver strong contributions without commanding high transfer fees — demonstrating the model's ability to generate smart, resource-aware recruitment solutions aligned with the high priority needs shown in Figure 5.2.

Together, these two case studies validate the model's flexibility in adapting to both elite and resource-constrained clubs by tailoring player selections to context-specific objectives and constraints.

5.5 Summary

This chapter presented a complete optimization framework for data-driven football player recruitment. The methodology integrates team performance analysis, player evaluation, tactical role assessment, and mathematical optimization to support strategic transfer decisions. The process is structured into the following key components:

- **Team Needs Construction:** Team-level performance data is processed through normalization and inversion to quantify feature-specific needs, forming a vector that reflects the areas requiring reinforcement.
- **Player Pool Generation:** Candidate players are selected by computing cosine similarity between team needs and individual player contributions. This ensures that only context-relevant players are considered for recruitment.

- **Tactical Role Gap Detection:** Using unsupervised clustering, tactical roles are identified and matched against team deficiencies to determine which roles are most lacking, guiding the model to prioritize certain player profiles.
- **Optimization Model:** A constrained optimization problem is formulated and solved using Pyomo and Gurobi. The objective is to maximize total team fit by selecting players whose contributions match the team's most urgent needs, subject to role, budget, and squad size constraints.

The framework yields optimized player recommendations that are both tactically aligned and financially feasible. It enables clubs to make objective, role-aware recruitment decisions, improving performance while controlling costs.

Chapter 6

Discussion

6.1 Empirical Outcomes: Market Value Prediction and Optimized Squad Selection

This thesis delivers a comprehensive, two-phase pipeline encompassing position-specific market value prediction and role-aware player selection through optimization. The empirical results from both stages strongly validate the proposed approach.

For the market value prediction models, the top-performing algorithms—Random Forest, XGBoost, and Stacked Regressors—achieved high predictive accuracy across all player lines as we can see in the results summary 4.4.

For the optimization phase, the model followed a structured decision-making pipeline: (1) constructing a pool of suitable players based on cosine similarity between player contributions and team need profiles, (2) identifying tactical role gaps using cluster centroids derived from unsupervised learning, and (3) applying a Pyomo-based optimization model that maximizes contribution while respecting budget and role constraints.

Two clubs were selected as case studies to demonstrate the system's versatility:

Real Madrid: The optimizer selected high-potential, tactically fitting players who aligned with the club's most pressing role deficiencies in midfield and attack. Selected players included João Gomes (Defensive Midfielder), Habib Diarra (Box-to-Box Midfielder), Angel Gomes (Playmaker Midfielder), and Adu Ares (Creative High-Pressure Forward). Despite varying market values—ranging from €2M to €35M—these players offered excellent value-for-money while filling specific tactical gaps. The optimizer prioritized attacking depth and midfield creativity, all while staying under budget.

Almería: With a limited budget, the model suggested experienced, undervalued players such as David Raum (Progressive Defender), Jorge Cuenca (Ball Winning Defender), Ricardo Rodríguez (Traditional Balanced Defender), Julien Ponceau (Box-to-Box Midfielder), and Kevin Stöger (Playmaker Midfielder). These selections directly addressed the most urgent deficiencies in defense and midfield. Notably, the optimizer avoided overpaying for reputation, instead favoring affordable profiles that matched team-specific tactical gaps with precision.

In both cases, the optimizer balanced quality, cost, and role-specific tactical need—producing selections that are both practical and innovative. By jointly considering role gaps and player suitability during pool formation, the system demonstrates strong capability in generating value-driven recruitment strategies tailored to club identity and constraints while saving money.

6.2 Superiority of Position-Oriented Market Value Modeling

Previous research on market value prediction, such as Al-Asadi and Tasdemir [14], Jishnu et al. [15], and Shen et al. [18], demonstrates strong predictive power, with reported R^2 values exceeding 0.95 in some cases.

This thesis improves upon those approaches in three key ways:

- **Position specificity:** Rather than one-size-fits-all modeling, the prediction phase here is segmented by position (GK, DF, MF, FW), resulting in improved relevance and performance.

- **Feature integration:** The model blends real match data (from FBref) with FIFA-style attributes (from SoFIFA), creating a richer and more realistic feature set than video-game-based inputs.

- **Contextual applicability:** While prior models often focus on performance, ours is tied directly to transfer decision-making, providing valuation that feeds into a broader recruitment pipeline.

Hence, while previous models achieve strong metrics, they are often abstracted from practical use. The models presented here are not only accurate but also built for operational decision support.

6.3 Optimization Model Advantages over Existing Frameworks

In the optimization literature, Galaz-Cares [22], Boon and Sierksma [23], and Onwuachu and Enyindah [24] each present valuable contributions to decision support in football. However, the model introduced in this thesis addresses several key limitations found in prior work:

1. **Team-specific player pool generation:** Unlike models that optimize from fixed lists, this work uses cosine similarity to dynamically generate a player pool per club based on its needs.

2. **Tactical role analysis:** Player roles are discovered via clustering, and team role gaps are identified using similarity-based matching. This enables precise targeting of not just positions, but tactical functions (e.g., ball-winning midfielder vs. deep-lying playmaker).

3. **Multiple realistic constraints:** The Pyomo model incorporates constraints on total budget, number of players, and minimum coverage of the most urgent roles—producing solutions tailored to real club situations.

4. **Performance-aware scoring:** The objective function accounts for team need intensity, player performance contribution, and club context (e.g., source team strength), offering better recruitment impact.

Together, these elements form an optimization system that is more tactical, adaptive, and cost-sensitive than previously proposed models. Most importantly, it translates abstract modeling into concrete recruitment suggestions for both elite and budget-restricted clubs.

6.4 End-to-End Practicality and Strategic Depth

Most prior studies operate in silos—either predicting value (e.g., [18, 17, 15]) or optimizing selection (e.g., [23, 22]), but rarely both. This thesis integrates both stages, building a pipeline that goes from valuation to decision-making in a manner that reflects real-world club workflows.

Furthermore, the focus on young talents adds strategic depth. Young players, as discussed in Chapter 3, are typically undervalued or hard to assess due to limited exposure. By concentrating predictive and optimization modeling on this segment, the thesis contributes to fairer and smarter talent identification—an area under-represented in football analytics research.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis proposes a comprehensive, data-driven framework for football club decision-making—one that spans from predicting player market value to the tactical, budget-constrained selection of transfer targets. The methodology focuses specifically on the emerging talents segment, enabling clubs to identify, evaluate, and acquire young players with both performance value and long-term strategic fit.

Unlike prior studies that treat prediction and optimization as separate processes, this work introduces an integrated pipeline where predicted market values inform a role-aware optimization model. The system builds contextual team need profiles, clusters player roles, infers tactical gaps, and applies constrained mathematical programming to produce feasible and impactful recruitment suggestions.

Key outcomes include:

- Position-specific market value models with strong performance (e.g., $R^2 > 0.9$ for forwards and midfielders).
- Role-based clustering of player types and gap detection per team.
- A Pyomo-based optimization model that selects cost-effective, tactically aligned players under budget and positional constraints.
- Case studies (e.g., Real Madrid and Almería) demonstrating the flexibility of the framework across different budget levels and team strategies.

Overall, the thesis contributes a realistic and actionable tool for clubs aiming to enhance their scouting, recruitment, and financial efficiency, particularly in the competitive and uncertain market for young footballers.

7.2 Future Work

While the presented framework already demonstrates strong predictive and prescriptive power, several promising directions exist for its extension and refinement:

- **Time-aware forecasting models:** Incorporating dynamic data (e.g., player form, injury history, recent transfers) using time series models or recurrent neural networks could improve the reliability of both performance forecasting and market value prediction over multiple windows.
- **Multi-objective optimization:** The current model optimizes tactical fit and budget usage. Future work could extend this to multi-objective formulations that also account for age diversity, potential resale value, injury risk, or contract duration. Pareto-efficient frontiers could help clubs navigate trade-offs.
- **Transfer success modeling:** Instead of predicting market value alone, future pipelines could assess historical transfer outcomes and learn to predict the likelihood of success (e.g., minutes played, performance improvement, resale gain), offering an additional validation layer for recommendations.
- **Cross-market generalization:** Applying the fr, mework to new segments—such as women's football, second-tier leagues, or emerging football markets—would help test its generalizability and promote broader equity in player evaluation and recruitment practices.
- **Interactive decision interfaces:** Developing a user-facing dashboard where club analysts and scouts can input team preferences, constraints, and adjust optimization weights could bridge the gap between backend analytics and day-to-day club operations.
- **Incorporating network effects:** Future models could consider team chemistry, tactical cohesion, and existing squad dynamics by incorporating graph-based representations of passing networks or match co-participation data.

7.3 Final Remarks

By integrating advanced machine learning, clustering, and optimization techniques, this thesis delivers a robust and scalable recruitment tool grounded in both data science and football expertise. It addresses real-world constraints, produces interpretable recommendations, and adapts to different club profiles and tactical systems. As the football industry continues to evolve toward data-informed strategies, frameworks like this one will be crucial in ensuring smarter, fairer, and more sustainable transfer decisions.

Bibliography

[1] *worldatlas.com*. URL: https://www.worldatlas.com (cit. on p. 1).

[2] *The Sun*. URL: https://www.thesun.co.uk (cit. on p. 1).

[3] *transfermarkt.com*. URL: https://www.transfermarkt.com (cit. on pp. 1, 2, 14, 16, 27).

[4] *Manchester City Annual Report*. URL: https://www.mancity.com/annualreport2024/ (cit. on p. 2).

[5] Ali Alhaj Hassan. *Predicting Young Football Talents Market Value and Optimizing Team Transfers*. https://github.com/AliHajHasan/Predicting-Young-Football-Talents-Market-Value-and-Optimizing-Team-Transfers. Accessed: 2025-06-01. 2025 (cit. on p. 5).

[6] Walter C. Neale. «The Peculiar Economics of Professional Sports*». In: *The Quarterly Journal of Economics* 78.1 (Feb. 1964), pp. 1–14. ISSN: 0033-5533. DOI: 10.2307/1880543. eprint: https://academic.oup.com/qje/article-pdf/78/1/1/5246371/78-1-1.pdf. URL: https://doi.org/10.2307/1880543 (cit. on p. 6).

[7] Mohamed El-Hodiri and James Quirk. «An Economic Model of a Professional Sports League». In: *Journal of Political Economy* 79.6 (1971), pp. 1302–1319. DOI: 10.1086/259837. URL: https://doi.org/10.1086/259837 (cit. on p. 6).

[8] Rodney Fort and James Quirk. «Cross-Subsidization, Incentives, and Outcomes in Professional Team Sports Leagues». In: *Journal of Economic Literature* 33.3 (1995), pp. 1265–1299. URL: https://www.jstor.org/stable/2729122?seq=1#metadata_info_tab_contents (cit. on p. 6).

[9] Stefan Kesenne. «Revenue Sharing and Competitive Balance in Professional Team Sports». In: *Journal of Sports Economics* 1.1 (2000), pp. 56–65. DOI: 10.1177/152700250000100105. URL: https://doi.org/10.1177/152700250000100105 (cit. on p. 6).

[10] Gerald W. Scully. «Pay and Performance in Major League Baseball». In: *American Economic Review* 64.6 (1974), pp. 915–930 (cit. on p. 6).

[11] Peter J. Sloane. «The Economics of Professional Football: The Football Club as a Utility Maximiser». In: *Scottish Journal of Political Economy* 18.2 (1971), pp. 121–146. DOI: 10.1111/j.1467-9485.1971.tb00979.x. URL: https://doi.org/10.1111/j.1467-9485.1971.tb00979.x (cit. on p. 6).

[12] Bill Gerrard and. «Rigour and relevance in sport management: reconciling the competing demands of disciplinary research and user-value». In: *European Sport Management Quarterly* 15.5 (2015), pp. 505–515. DOI: 10.1080/16184742.2015.1085714. eprint: https://doi.org/10.1080/16184742.2015.1085714. URL: https://doi.org/10.1080/16184742.2015.1085714 (cit. on p. 7).

[13] Danny F. Hill, James Skinner, and Anna Grosman and. «A review of football player metrics and valuation methods: a typological framework of football player valuations». In: *Managing Sport and Leisure* 0.0 (2025), pp. 1–24. DOI: 10.1080/23750472.2025.2459727. eprint: https://doi.org/10.1080/23750472.2025.2459727. URL: https://doi.org/10.1080/23750472.2025.2459727 (cit. on p. 7).

[14] Mustafa A. Al-Asadi and Sakir Tasdemır. «Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques». In: *IEEE Access* (2022) (cit. on pp. 7, 14, 113).

[15] V B Jishnu, P V Hari Narayanan, Surya Aanand, and Preetha Theresa Joy. «Football Player Transfer Value Prediction Using Advanced Statistics and FIFA 22 Data». In: *2022 IEEE 19th India Council International Conference (INDICON)*. 2022, pp. 1–6. DOI: 10.1109/INDICON56171.2022.10040117 (cit. on pp. 7, 113, 114).

[16] «Beyond crowd judgments: Data-driven estimation of market value in association football». In: *European Journal of Operational Research* 263.2 (2017), pp. 611–624. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2017.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0377221717304332 (cit. on p. 8).

[17] Sebastian Majewski. «Identification of Factors Determining Market Value of the Most Valuable Football Players». In: *Journal of Management and Business Administration. Central Europe* 24 (Sept. 2016), pp. 91–104. DOI: 10.7206/jmba.ce.2450-7814.177 (cit. on pp. 8, 114).

[18] Qijie Shen. «Predicting the value of football players: machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets». In: *Applied Intelligence* 55 (Jan. 2025). DOI: 10.1007/s10489-024-06189-0 (cit. on pp. 8, 14, 113, 114).

[19] Raffaele Poli, Roger Besson, and Loïc Ravenel. «Statistical Modeling of Football Players' Transfer Fees Worldwide». In: *International Journal of Financial Studies* 12 (Sept. 2024), p. 93. DOI: 10.3390/ijfs12030093 (cit. on p. 8).

[20] Subham Parida, K. Deepa Thilak, and Ransher Singh. «Enhancing the Prediction of Growth of Footballers using Real-Life Statistics and Machine Learning». In: *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. 2022, pp. 471–475. DOI: 10.1109/ICAAIC53929.2022.9792997 (cit. on p. 9).

[21] *WhoScored.com*. URL: https://www.whoscored.com (cit. on p. 9).

[22] Pablo Galaz-Cares. «Optimizing Football Squad Planning: Balancing Competitive Success and Financial Sustainability». In: *European Journal of Operational Research* (2023) (cit. on pp. 9, 114).

[23] Bart H. Boon and Gerard Sierksma. «Team formation: Matching quality supply and quality demand». In: *European Journal of Operational Research* 148.2 (2003). Sport and Computers, pp. 277–292. ISSN: 0377-2217. DOI: https://doi.org/10.1016/S0377-2217(02)00684-7. URL: https://www.sciencedirect.com/science/article/pii/S0377221702006847 (cit. on pp. 9, 114).

[24] Uzochukwu C. Onwuachu and P. Enyindah. «A Machine Learning Application for Football Players' Selection». In: *International Journal of Engineering Research and Technology (IJERT)* 3.10 (2014). Paper ID: IJERTV4IS100323, pp. 323–328. URL: https://www.ijert.org/a-machine-learning-application-for-football-players-selection (cit. on pp. 10, 114).

[25] James Mead, Anthony O'Hare, and Paul McMenemy. «Expected goals in football: Improving model performance and demonstrating value». In: *PLOS ONE* 18.4 (Apr. 2023). DOI: 10.1371/journal.pone.0282295. URL: https://doi.org/10.1371/journal.pone.0282295 (cit. on p. 10).

[26] Mustafa Cavus and Przemyslaw Biecek. *Explainable expected goal models for performance analysis in football analytics*. June 2022. DOI: 10.48550/arXiv.2206.07212 (cit. on p. 10).

[27] Jerome H. Friedman. «Greedy Function Approximation: A Gradient Boosting Machine». In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation--A-gradient-boosting-machine/10.1214/aos/1013203451.full (cit. on p. 11).

[28] Leo Breiman. «Random Forests». In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324. URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf (cit. on p. 11).

[29] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. 3rd. Wiley-Interscience, 1998 (cit. on p. 11).

[30] URL: https://fbref.com/en/ (cit. on pp. 14, 16, 18).

[31] *soFIFA*. URL: https://sofifa.com/ (cit. on pp. 14–16, 25).

[32] *Sports Reference*. URL: https://www.sports-reference.com/ (cit. on p. 14).

[33] *opta*. URL: https://www.statsperform.com/opta/ (cit. on p. 14).

[34] *EAsports FIFA*. URL: https://www.ea.com/es-es/games/fifa (cit. on pp. 15, 25, 30, 36).

[35] Lukas Scherbaum. «FIFA Ratings Explained». In: *earlygame.com* (2024). URL: https://earlygame.com/fifa/fifa-ratings-explained-overall-rating-1 (cit. on p. 15).

[36] *Selenium*. URL: https://www.selenium.dev/documentation/ (cit. on p. 17).

[37] *Requests*. URL: https://docs.python-requests.org/en/latest/index.html (cit. on p. 17).

[38] *CIES Football observatory*. URL: https://football-observatory.com/Inflation-in-the-football-players-transfer-market (cit. on p. 28).

Appendix. Project Plan

Workplan

- **Subject Familiarization & Brainstorming — 3 days** (19 hours) Initial domain research and setup of tools, datasets, and project structure.
- **Literature Review and Scope Definition — 10 days** (63 hours) Review of key studies on player valuation and optimization. Defined research contribution and scope.
- **Data Sourcing and Integration — 13 days** (75 hours) Collection of player and team datasets. Merging, cleaning, and preparing data for analysis.
- **Feature Engineering and EDA — 13 days** (85 hours) Computation of player contribution metrics, contextual scores, and data visualization to support modeling.
- **Market Value Prediction (Machine Learning Models) — 24 days** (150 hours) Development of Random Forest, XGBoost, and Stacking models for market value estimation. Included tuning, evaluation, and position-wise segmentation.
- **Model Evaluation and Interpretation — 8 days** (50 hours) Metrics calculation (MAE, RMSE, R^2), visual analysis of predictions, and feature interpretation.
- **Role Detection and Clustering — 3 days** (19 hours) Clustering of player performance profiles using PCA and KMeans. Assignment of tactical roles.
- **Optimization Modeling and Simulations — 20 days** (126 hours) Pyomo-based implementation of an optimization model with constraints on role fit, budget, and team needs. Simulated scenarios for different clubs.
- **Writing and Documentation — 23 days** (145 hours) Structured writing of all chapters, creation of plots and tables, integration of methodology, and preparation of reproducible code.
- **Final Review and Submission — 3 days** (20 hours) Final edits, GitHub release, and submission checks.

In total, the thesis work required 120 days and 752 hours, averaging approximately 6 to 7 hours of work per day.

Gantt Chart

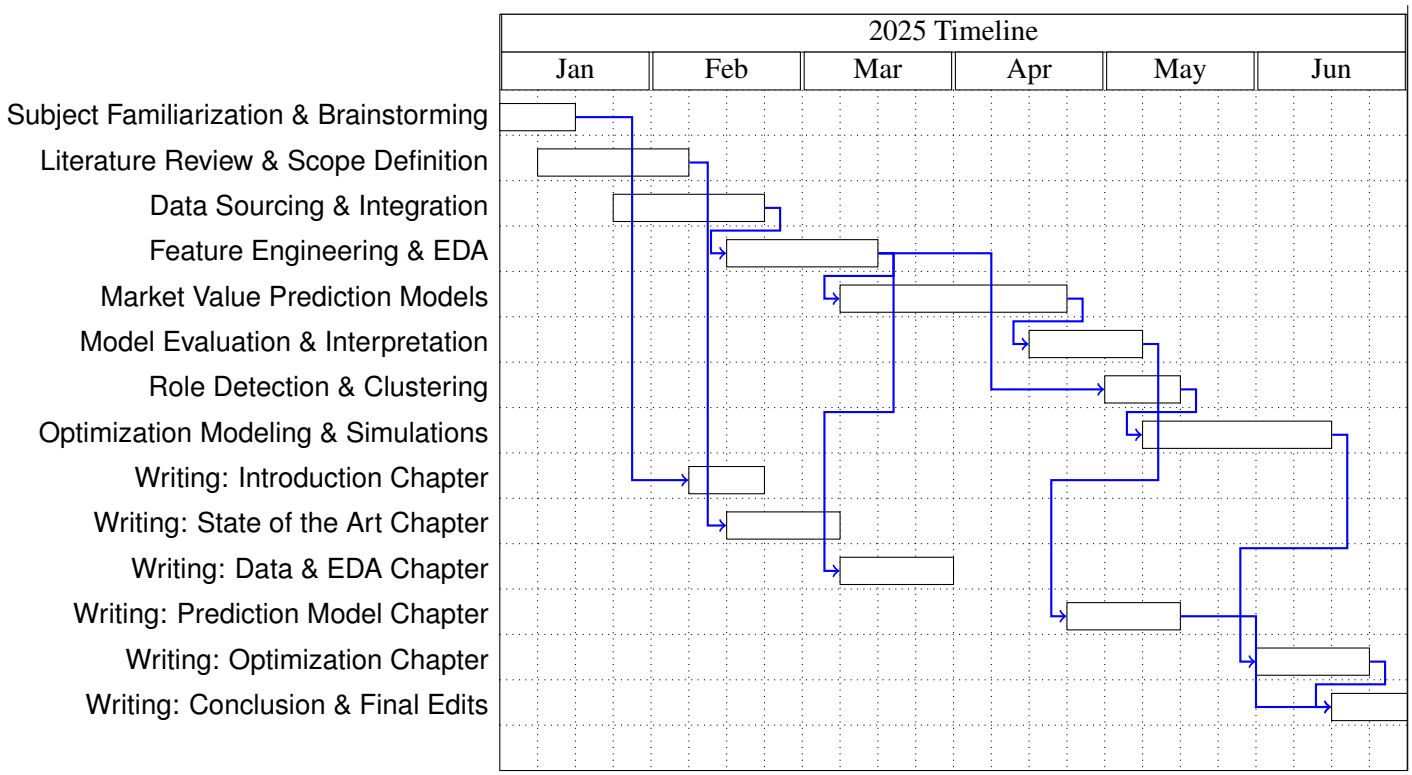


Fig. 7.1. Project Gantt Chart – Thesis Progression (Jan–Jun 2025)

Budget

The following table outlines the actual cost associated with this thesis. It includes tool subscriptions, hardware usage, utilities, and the time invested over 120 working days (756 hours total). All estimates reflect real costs incurred during the project period (January–June 2025).

Item	**Cost (EUR)**
Google Colab Pro subscription (6 months × €12)	72
Grammarly Premium subscription (6 months × €12)	72
Electricity and internet for PC usage (6 months)	50
Laptop amortization (6 months of a €1200 laptop, over 5 years)	120
Software/tools used (LaTeX, Python, GitHub – open source)	0
Material and service subtotal	**314**
Student research time (752 hours × €13.5/hour)	**10,152**
Total Estimated Cost	**€10,466**

TABLE 7.1. ACTUAL BUDGET BREAKDOWN FOR THE THESIS PROJECT

The majority of the cost is attributed to the time invested in analysis, modeling, and writing. Paid services like Colab Pro and Grammarly supported performance and writing quality. Hardware costs were proportionally amortized, and open-source tools ensured zero licensing fees.