applied sciences

MDPI

Article

Predicting Soccer Player Salaries with Both Traditional and Automated Machine Learning Approaches

Davronbek Malikov, Pilsu Jung and Jaeho Kim *

Department of AI Convergence Engineering, Gyeongsang National University (GNU), Jinjudaero 501, Jinjusi 52828, Republic of Korea; davronbekmalikov96@gmail.com (D.M.); psjung@gnu.ac.kr (P.J.)
* Correspondence: jaeho.kim@gnu.ac.kr

Abstract

Soccer's global popularity as the world's favorite sport is driven by many factors, with high player salaries being one of the key reasons behind its appeal. These salaries not only reflect on-field performance, but also capture a broader evaluation of player value. Despite the increasing use of performance data in sports analytics, a critical gap remains in establishing fair compensation models that comprehensively account for both quantifiable and intangible contributions. To address these challenges, this study adopts machine learning (ML) techniques that model player salaries based on a combination of performance metrics and contextual features. This research focuses on reducing bias and improving transparency in salary decisions through a systematic, data-driven approach. Utilizing a dataset spanning the 2016–2022 seasons, we apply both traditional and automated ML frameworks to uncover the most influential factors in salary determination. The results indicate a nearly 17% improvement in R^2 and about a 30% reduction in MAE after incorporating the newly constructed features and methods, demonstrating a significant enhancement in model performance. Gradient Boosting demonstrates superior effectiveness, revealing a group of significantly underestimated and overestimated players, and showcasing the model's proficiency in detecting valuation discrepancies.

Keywords: soccer player salaries; salary prediction; performance metrics; predictive modeling; actual goals; expected goals; machine learning; European soccer leagues

Academic Editor: Firstname Lastname

Received: 10 June 2025
Revised: 17 July 2025
Accepted: 17 July 2025
Published:

Citation: Malikov, D.; Jung, P.; Kim, J. Predicting Soccer Player Salaries with Both Traditional and Automated Machine Learning Approaches. *Appl. Sci.* **2025**, *1*, 0. https://doi.org/

1. Introduction

Soccer is the most popular sport in the world, with nearly 5 billion fans, making it the most-watched and played sport globally [1,2]. The sport's high level of competition and its ability to bring together people from different cultures have helped it remain dominant on the world stage. As a result, soccer has become a global phenomenon that crosses national borders. Beyond its cultural significance, the high salaries offered to professional soccer players play a key role in attracting and retaining top talent globally, giving the sport a competitive edge over others. Table 1 presents information on the highest-paid athletes worldwide in 2024. Soccer dominates the 2024 global earnings list, with six of the top ten highest-paid athletes coming from the sport. Cristiano Ronaldo exemplifies this, earning an estimated USD 280 million last year, including USD 220 million from his contract with Al Nassr, making him the highest-paid athlete for the fourth time [3]. This highlights soccer's global appeal and financial clout.

Table 1. Earnings of the 10 highest-paid athletes in the world in 2024 [3].

Player	Sport	On-the-Field (USD Million)	Off-the-Field (USD Million)
Cristiano Ronaldo	Soccer	220	60
Jon Rahm	Golf	198	20
Lionel Messi	Soccer	65	70
LeBron James	Basketball	48.7	80
Neymar	Soccer	80	30
Stephen Curry	Basketball	55.8	50
Karim Benzema	Soccer	100	4
Giannis Antetokounmpo	Basketball	48.8	45
Kylian Mbappé	Soccer	70	20
Lamar Jackson	Football	53	32.6

Player salaries are influenced by factors such as club wealth, league exposure, endorsements, and global fanbase. On-field performance metrics, such as assists, expected goals (xG), and actual goals (aG), also play key roles in salary evaluation [4]. Additionally, demographic and contextual features, such as age, nationality, and career experience, impact marketability and salary within league and club dynamics [5]. However, prior research often neglects positional roles and league-specific factors [6,7], focusing mainly on transfer valuations rather than salaries [8,9]. Salary decisions in professional soccer are frequently influenced by incomplete data and subjective judgments, leading to the potential undervaluation of players who contribute in less visible but essential roles, as well as the overvaluation of more popular or high-scoring players. This imbalance can affect team dynamics, player motivation, and financial planning within clubs. Establishing a transparent, data-driven framework for salary evaluation not only fosters fairness and equity across different playing positions but also provides clubs with more reliable tools for contract negotiations and budgeting. Moreover, as the use of advanced analytics becomes increasingly integral to sports management, developing interpretable and unbiased machine learning models is crucial for gaining trust among stakeholders and facilitating informed decision-making in a competitive and financially complex environment.

Despite advances in machine learning, accurately predicting soccer salaries remains complex due to many interacting factors [6,10,11]. Furthermore, automated machine learning (AutoML) techniques, which automate key tasks such as model selection, tuning, and feature engineering, are increasingly applied in various fields [12]. However, their use in soccer salary prediction remains limited, with most studies relying on traditional models. Yet, even with advanced modeling approaches like AutoML, determining fair and accurate salaries involves a complex interplay of measurable and intangible factors. Qualitative elements such as leadership, team influence, and positional roles often difficult to quantify continue to play a crucial role in salary decisions. ML offers a promising approach but also brings its own challenges, such as difficulty in capturing intangible factors and ensuring balanced, unbiased predictions across different player positions.

This study addresses these issues through a systematic analysis using both traditional ML models and an automated ML framework to uncover the most influential factors in salary determination. We analyze a wide range of characteristics that contribute to players' salaries to develop a comprehensive and interpretable evaluation. Moreover, this approach provides a balanced evaluation across all player positions, ensuring that no group is unfairly represented. By bridging ML methodologies with domain-specific considerations, this study not only identifies the key drivers of player salaries but also provides actionable insights into the importance of balanced feature selection, advancing the field of sports analytics. In our comparative evaluation, we assess model performance between the baseline configuration and our proposed approach. Moreover, we analyze

players from five major European leagues—the English Premier League, Bundesliga, La Liga, Serie A, and Ligue 1—and classify players as either underestimated or overestimated based on their salaries.

The main contributions of this study are as follows:

- Identifies key factors influencing professional soccer player salaries, enhancing understanding of compensation drivers.
- Classifies players as underestimated or overestimated based on salary performance gaps across leagues.
- Introduces a fairness-aware method that balances positional influence, preventing bias towards attacking players.
- Shows that traditional machine learning models outperform AutoML in both accuracy and interpretability.
- Provides actionable insights for clubs, managers, and analysts to support data-driven decision-making.

The remainder of the paper reviews related work (Section 2), outlines data collection and feature engineering (Section 3), describes the ML algorithms (Section 4), comparative evaluation and evaluation results (Sections 5 and 6), discusses limitations (Section 7), and concludes our study (Section 8).

2. Literature Review and Study Overview

Research on soccer player salary prediction has advanced from traditional econometric methods to modern machine learning approaches. Table 2 summarizes key studies, followed by a discussion of major developments in the field.

Table 2. Key studies on soccer player salary prediction using traditional and machine learning approaches.

Traditional Econometric and Statistical Approaches	
Frick (’07) [13]	Salary depends on productivity, position, and career longevity.
Lucifora et al. (’09) [14]	International experience and seniority strongly impact wages.
Késenne (’07) [15]	League prestige and financial strength shape salary levels.
Bryson et al. (’14) [16]	Assesses returns to education and experience using a human capital model.
Garcia-del-Barrio et al. (’07) [17]	Considers brand value, off-field visibility, and player reputation.
Müller et al. (’17) [18]	Market dynamics and negotiation power influence pay beyond performance.
Machine Learning and Data-Driven Approaches	
Malikov et al. (’24) [4]	Uses xG and aG metrics to model salary fairness and performance impact.
Rong et al. (’24) [21]	Employs player stats like goals, assists, and appearances for salary prediction.
Bhilawa et al. (’22) [22]	Applies ML to identify key predictors using detailed match data.
Majewski et al. (’16) [20]	Focuses on prediction using performance ratings and time on field.
Herm et al. (’14) [28]	Investigates salary using in-game activity and player involvement.
Margareta et al. (’22) [23]	Combines market value, salary, age, and performance for ML modeling.
Elahi et al. (’23) [5]	Proposes feature engineering to improve salary prediction accuracy.
Lee et al. (’22) [25]	Uses ensemble models across positions and leagues.
AlAsadi et al. (’22) [9]	Applies boosting methods for player value estimation.
Huang et al. (’23) [10]	Integrates FIFA and real performance metrics for hybrid prediction.
Yaldo et al. (’17) [24]	ML-based salary prediction using FIFA video game data.
Kaggle (’23) [29]	Uses FIFA 20 attributes (e.g., potential, reputation) for salary modeling.
Li et al. (’22) [6]	Focuses on on-field metrics but lacks contextual features.
Stafylidis et al. (’24) [7]	Highlights the need for position-aware salary models.
Shen et al. (’25) [8]	Predicts transfer values as proxy indicators of wage.
Huang et al. (’23) [10]	Emphasizes explainable ML but not directly tied to salary.
Bayesian Sports Analytics (’22) [11]	Applies Bayesian learning for general soccer predictions.
Pieper and Rehm (’23) [26]	Explore goalkeeper-specific performance metrics in salary modeling.
Kim and Vukoja (’24) [27]	Investigate behavioral and institutional factors influencing salaries.

2.1. Literature Review

Traditional econometric and statistical methods remain foundational in analyzing and predicting soccer player salaries. These approaches typically use regression models, correlation analysis, and human capital theory to explain how individual and institutional factors influence wages. Frick (2007) employed multiple regression analysis on German football data, identifying player productivity and career longevity as key salary determinants [13]. Similarly, Lucifora and Simmons (2009) studied the Italian Serie A and showed that experience, past season performance (e.g., goals scored), and international appearances significantly explain wage variation, reinforcing the role of ability and seniority in compensation [14]. Késenne (2007) expanded the analysis by incorporating club-level economic factors, such as financial resources and league prestige, framing salary structures within the broader market and institutional environment [15]. Meanwhile, Bryson et al. (2014) applied human capital theory to show the influence of formal education and on-the-job learning in wage outcomes [16]. Other works, such as that of Garcia-del-Barrio and Pujol (2007), emphasized brand value and player visibility, while Müller et al. (2017) highlighted the role of negotiation power and market dynamics alongside performance [17,18].

The application of machine learning (ML), a subfield of artificial intelligence (AI), offers advanced capabilities for predicting soccer players' salaries and market values [9,19]. While many studies have explored player performance through on-field metrics such as expected goals (xG), actual goals (aG), assists, and playing time, they often focus more on transfer valuation than direct salary prediction [8,9]. Despite growing interest, salary modeling remains underexplored due to the complex interplay between performance, market dynamics, and contextual variables [10,11]. Early ML-based research typically relied on core statistics, such as goals, assists, and minutes played [20]. More recent work has expanded feature sets to include defensive metrics (e.g., interceptions, clearances), career dynamics (e.g., appearances, starting frequency), and contextual variables (e.g., age, nationality, and market value) [21,22]. Several studies have integrated financial data like transfer fees and existing salaries, combining them with performance metrics to enhance predictive accuracy [4,23]. These multi-dimensional models underscore the importance of modeling both on-field output (e.g., xG, aG) and broader influences on compensation. Advanced ML models, such as Random Forest, XGBoost, and Support Vector Regression, show strong predictive performance. For instance, Yaldo et al. [24] used pattern recognition to achieve a Pearson correlation of 0.77, while Lee et al. [25] optimized MLS salary distributions using ensemble learning. Hybrid approaches combining game statistics, FIFA data, and metaheuristic optimization techniques further demonstrate the versatility of ML in soccer salary prediction [10]. Moreover, building upon earlier work, more recent studies have further advanced this line of analysis. For instance, Pieper and Rehm (2023) focused on goalkeeper-specific wage determinants, highlighting clean sheets and save efficiency as key factors [26]. Moreover, Kim and Vukoja (2024) introduced behavioral insights into salary structures, emphasizing perceived fairness and institutional inequality [27]. Together, these studies enriched the traditional econometric approach by incorporating position-specific performance metrics, advanced regression techniques, and interdisciplinary perspectives, thereby providing a more nuanced understanding of soccer salary determinants.

2.2. Research Gaps and Study Overview

Despite the growing use of machine learning in sports analytics, salary prediction studies often rely on general performance metrics (e.g., goals, assists) without addressing mismatches between player contributions and compensation [13,14,28]. This study fills this gap by classifying players as underestimated or overestimated based on salary–performance alignment, highlighting compensation imbalances across leagues.

Additionally, previous research tended to overlook positional effects, favoring attacking players due to their visibility. We introduce a positional ratio feature to balance salary modeling across roles, ensuring the fair evaluation of defenders and midfielders and enhancing model robustness. The study also compares traditional ML models with AutoML frameworks. While AutoML automates tuning and selection [30–32], traditional models outperform AutoML in accuracy and interpretability, providing clearer insights into salary–performance relationships vital for club's financial decisions. Finally, key salary predictors include performance metrics (e.g., xGg, aGg), market features (league and position ratios), and engineered variables, offering both academic and practical values for fair contract negotiations.

3. Overview of Data and Preprocessing

In this section, we present the dataset description and tools used in this study. We also cover data preprocessing, feature encoding, general feature engineering, positional normalization, contribution metrics, the introduction of new features, and feature specification with explanations.

3.1. Dataset Description and Analytical Tools

In this study, we utilize two distinct datasets to investigate the relationship between soccer player performance and salary dynamics. The first, sourced from Understat [33], offers advanced performance metrics, such as expected goals (xG), assists, shots, and other key indicators essential for evaluating on-field contributions. The second dataset, obtained from Capology [34], provides detailed salary data, including weekly wages, contract values, and financial commitments across professional soccer. Both datasets are publicly curated and made accessible by an independent analyst, serving as a valuable resource for research in sports analytics [35]. The combined dataset spans the top five European leagues, covering six seasons from 2016–2017 to 2021–2022. It includes approximately 45,000 rows representing 5238 unique players across various positions, such as defenders, midfielders, forwards, goalkeepers, and specialized roles like attacking or defensive midfielders. This extensive scope enables a holistic analysis of performance and compensation trends across roles and leagues.

For data processing and analysis, we employ Python (3.11.7) alongside libraries such as Pandas (2.1.4), NumPy (1.26.4), Matplotlib (3.7.5), Seaborn (0.12.2), Klib (1.3.2), and Scikit-learn (1.4.2). PyCaret (3.3.2) is used for the AutoML component of our modeling workflow. These tools facilitate data cleaning, the handling of missing values, feature engineering, and the construction of robust predictive models. Their integration ensures analytical rigor and enhances reproducibility throughout the study.

3.2. Data Preprocessing

Before analysis, the dataset undergoes preprocessing to ensure integrity, consistency, and readiness for modeling. Preprocessing is a vital step in data-driven research, involving the cleaning, transformation, and organization of raw data into a structured format suitable for exploration and prediction. This section outlines the preprocessing techniques applied, including handling missing values, data encoding, feature engineering, and scaling. As detailed in Section 3.1, the study combines two datasets from different seasons and leagues. The raw data are first collected and then merged to create a unified dataset for analysis. The klib library is employed to improve data quality by removing duplicates, standardizing column names, and ensuring overall consistency. To address missing values, we adopt a combination of mode and mean imputation techniques. Specifically, for categorical variables, mode imputation is applied by replacing missing values with the most frequently

occurring category, thereby preserving the distributional consistency. For numerical variables, mean imputation is utilized, filling missing entries with the average of available observations to maintain the overall statistical integrity of the dataset while minimizing potential bias.

3.2.1. Feature Encoding

To transform categorical variables into numerical formats suitable for machine learning, we use Label Encoding [36] for low-cardinality features and Hashing Encoding [37] for high-cardinality features. Label Encoding assigns a unique integer to each category:

$$LE(c_i) = i \quad \forall c_i \in C \tag{1}$$

where $LE(c_i)$ is the encoded value of category c_i and C is the set of unique categories. While efficient, this may introduce ordinal relationships not present in the data. Hashing Encoding applies a hash function to map categories to a fixed number of buckets, mitigating issues of order and high dimensionality:

$$HE(c_i) = \text{hash}(c_i) \mod k \quad \forall c_i \in C \tag{2}$$

To select an encoder, we use the following rule based on cardinality:

$$E(F) = \begin{cases} HE(F) & \text{if } |C(F)| > 10 \\ LE(F) & \text{if } |C(F)| \leq 10 \end{cases} \tag{3}$$

This strategy balances computational efficiency and model performance by adapting to the nature of the categorical features.

3.2.2. Feature Engineering

General Feature Engineering

Feature engineering is a critical component of data preprocessing, enabling the transformation of raw data into meaningful features that enhance model performance. In this study, we generated several new features to better capture player performance dynamics and improve predictive accuracy. Table 3 summarizes the key engineered features, their descriptions, and their calculations.

Table 3. Engineered features with descriptions and calculations.

Feature	Description	Calculation
aGg	Actual goals scored per game	$aGg = \frac{goals}{games}$
gpm	Goals scored per minute	$gpm = \frac{goals}{time}$
apg	Assists per game	$apg = \frac{assists}{games}$
apm	Assists per minute	$apm = \frac{assists}{time}$
shpg	Shots taken per game	$shpg = \frac{shots}{games}$
shpm	Shots taken per minute	$shpm = \frac{shots}{time}$
kppg	Key passes per game	$kppg = \frac{key\_passes}{games}$
kppm	Key passes per minute	$kppm = \frac{key\_passes}{time}$

Table 3. *Cont.*

Feature	Description	Calculation
ypg	Yellow cards per game	$ypg = \frac{yellow\_cards}{games}$
ypm	Yellow cards per minute	$ypm = \frac{yellow\_cards}{time}$
rpg	Red cards per game	$rpg = \frac{red\_cards}{games}$
rpm	Red cards per minute	$rpm = \frac{red\_cards}{time}$
xGdiff	Difference between actual and expected goals	$xGdiff = aGg - xGg$
xGg	Expected goals per game	$xGg = \frac{xG}{games}$

Features such as *aGg* and *apg* quantify a player's direct offensive contributions. Normalized metrics like *gpm* and *apm* adjust for playing time, offering a clearer view of efficiency. Attacking involvement is captured by *shpg* and *shpm*, while playmaking ability is represented by *kppg* and *kppm*. Disciplinary behaviors are measured through *ypg* and *rpg*, which track yellow and red cards, respectively. Additionally, *xGdiff*, the difference between actual and expected goals, provides insight into a player's finishing ability relative to chance quality. The player's position is also considered, recognizing that forwards naturally have higher goal-scoring opportunities, which can influence salary outcomes. Together, these engineered features not only enrich the dataset but also offer deeper insights into performance, behavior, and value assessment.

Table 4. Player position balancing and normalization.

Feature Name	Description	Calculation Method
goalsPerPosAvg	Goals relative to average in position	$\frac{goals}{\text{mean } goals \text{ in position}}$
xGPerPosAvg	Expected goals normalized by position	$\frac{xG}{\text{mean } xG \text{ in position}}$
assistsPerPosAvg	Assists normalized by position	$\frac{assists}{\text{mean } assists \text{ in position}}$
xGgPerPosAvg	xG per game normalized by position	$\frac{xGg}{\text{mean } xGg \text{ in position}}$
aGgPerPosAvg	Actual goals per game normalized by position	$\frac{aGg}{\text{mean } aGg \text{ in position}}$
shotsPerPosAvg	Shots normalized by position	$\frac{shots}{\text{mean } shots \text{ in position}}$
keyPassesPerPosAvg	Key passes normalized by position	$\frac{key\_passes}{\text{mean } key\_passes \text{ in position}}$
goalsZScore	Goals Z-score	$\frac{goals - \mu goals}{\sigma goals}$
xGZScore	Expected goals Z-score	$\frac{xG - \mu xG}{\sigma xG}$
assistsZScore	Assists Z-score	$\frac{assists - \mu assists}{\sigma assists}$
xGgZScore	xG per game Z-score	$\frac{xGg - \mu xGg}{\sigma xGg}$
aGgZScore	aG per game Z-score	$\frac{aGg - \mu aGg}{\sigma aGg}$
shotsZScore	Shots Z-score	$\frac{shots - \mu shots}{\sigma shots}$
keyPasses	Key passes Z-score	$\frac{key\_passes - \mu key\_passes}{\sigma key\_passes}$
xGRatio	xG to aGg ratio	$\frac{xG}{aGg}$
xGgRatio	xGg to aGg ratio	$\frac{xGg}{aGg}$
aGgRatio	aGg to xGg ratio	$\frac{aGg}{xGg}$

Positional Normalization and Contribution Metrics

Position significantly influences a player's market value alongside team performance and skill [38,39]. We engineered features to normalize player performance by position, playing time, and expectations, preventing attackers from being overvalued simply owing to more scoring chances. Table 4 summarizes these features, which balance performance

across positions. Metrics like *goalsPerPosAvg*, *xGPerPosAvg*, and *assistsPerPosAvg* compare a player's stats to their positional averages, identifying over- or underperformance relative to peers. Similar normalization applies to *xGgPerPosAvg*, *aGgPerPosAvg*, *shotsPerPosAvg*, and *keyPassesPerPosAvg*, capturing differences in scoring and playmaking.

We also use Z-score standardization (e.g., *goalsZScore*, *xGZScore*, *assistsZScore*) to measure deviations from the league mean, enabling fair comparisons across leagues and teams. Ratio-based features like *xGRatio*, *xGgRatio*, and *aGgRatio* assess efficiency in converting chances, enhancing both model accuracy and interpretability in salary valuation.

3.2.3. Introduction of New Features

In addition to engineered features, we introduce the *league_weight* feature to improve model accuracy by capturing league competitiveness. This feature combines two key factors:

1. **Average Annual Salary Ranking** (S): The average player salary in million GBP, reflecting a league's financial strength and ability to attract top talent. Higher salaries often correlate with better resources and player performance [34].
2. **UEFA Coefficient** (U): Measures league strength based on club performance in European competitions over the past five seasons. Higher coefficients indicate more competitive leagues with better financial and institutional backing [40].

We compute the *league_weight* using a weighted normalization formula:

$$\textit{league\_weight} = \alpha \cdot \frac{S}{\max(S)} + \beta \cdot \frac{U}{\max(U)} \tag{4}$$

where $\max(S)$ and $\max(U)$ are the maximum salary ranking and UEFA coefficient values across leagues, respectively. The weights $\alpha = 0.6$ and $\beta = 0.4$ reflect the relative importance of salary and UEFA coefficient, with more emphasis on salary due to cases like France, where high wages exist despite lower UEFA rankings. We select a range of appropriate weight values based on our domain knowledge and test combinations of candidate weight values, determining the optimal set based on experimental results. Moreover, Table 5 presents the final results of the league weight calculation, derived from the average annual salary and UEFA coefficients for the top five leagues.

Table 5. *league_weight* calculation based on average annual salary and UEFA coefficients for top five leagues.

League	Average Annual Salary (GBP)	UEFA Coefficient	League Weight
EPL (England)	GBP 3,433,450	106.624	1.0000
La Liga (Spain)	GBP 1,950,083	87.739	0.6695
Bundesliga (Germany)	GBP 1,568,559	82.581	0.5845
Serie A (Italy)	GBP 1,477,455	92.668	0.6060
Ligue 1 (France)	GBP 1,161,575	69.093	0.4620

3.3. Feature Specification and Explanation

This subsection provides a concise overview of the features used in this study, categorized by type and relevance. Table 6 summarizes features from the raw dataset, engineered attributes, and newly introduced variables.

- **Performance Metrics**: Include key statistics such as goals, *xG*, assists, key passes, *xGChain*, and *xGBuildup*, reflecting direct and indirect offensive contributions.
- **Disciplinary Actions**: Captured through yellow/red cards, and normalized measures like per game or per minute rates, indicating player discipline and its impact.

- **Playing Time & Efficiency**: Combines games, minutes played, and rate-based metrics such as *gpm* and *apm* to evaluate contribution relative to time on the field.
- **Positional & League Influence**: Includes categorical variables for position, position encoding, and the *league_weight* to model contextual competitiveness.
- **Team & Nationality Encoding**: Encoded club and country data preserve privacy while allowing contextual analysis.

This classification improves interpretability and ensures balanced integration of diverse factors in evaluating player performance and modeling outcomes.

Table 6. Specification of features used for training and prediction.

Feature	Description	Category	Feature	Description	Category
goals	Total goals scored	Raw Data	*yellow_cards*	Number of yellow cards	Raw Data
xG	Expected goals	Raw Data	*red_cards*	Number of red cards	Raw Data
assists	Total assists	Raw Data	*ypg*	Yellow cards per game	Engineered
xA	Expected assists	Raw Data	*ypm*	Yellow cards per minute	Engineered
shots	Total shots taken	Raw Data	*rpg*	Red cards per game	Engineered
key_passes	Total key passes	Raw Data	*rpm*	Red cards per minute	Engineered
npg	Non-penalty goals	Raw Data	*games*	Total games played	Raw Data
npxG	Non-penalty expected Ggoals	Raw Data	*time*	Total minutes played	Raw Data
xGChain	xG contribution in possession chains	Raw Data	*gpm*	Goals per minute played	Engineered
xGBuildup	xG excluding shots and assists	Raw Data	*apg*	Assists per game	Engineered
xGdiff	Difference between xG and goals	Engineered	*shpg*	Shots per game	Engineered
xGg	Expected goals per game	Engineered	*shpm*	Shots per minute played	Engineered
aGg	Actual goals per game	Engineered	*kppg*	Key passes per game	Engineered
position_weight	Encoded positional category	Engineered	*kppm*	Key passes per minute	Engineered
league_weight	League-specific weight factor	New Feature	*goalsPerPosAvg*	Goals normalized to position average	Engineered
xGPerPosAvg	xG normalized to position average	Engineered	*assistsPerPosAvg*	Assists normalized to position average	Engineered
xGgPerPosAvg	xGg normalized to position average	Engineered	*aGgPerPosAvg*	aGg normalized to position average	Engineered
shotsPerPosAvg	Shots normalized to position average	Engineered	*goalsZScore*	Z-score of goals	Engineered
xGZScore	Z-score of xG	Engineered	*assistsZScore*	Z-score of assists	Engineered
xGgZScore	Z-score of xGg	Engineered	*aGgZScore*	Z-score of aGg	Engineered
shotsZScore	Z-score of shots	Engineered	*keyPasses*	Z-score of key passes	Engineered
xGRatio	xG to actual goals ratio	Engineered	*xGgRatio*	xGg to aGg ratio	Engineered
aGgRatio	aGg to xGg ratio	Engineered	*teamEncoded*	Team-specific encoding	Engineered
positionEncoded	Encoded-position category	Engineered	*age*	Player's age	Raw Data

4. Workflow and Machine Learning Algorithms

The overall workflow of the study is illustrated in Figure 1, outlining the step-by-step pipeline from data preprocessing to model evaluation. This structured workflow is important as it ensures transparency, reproducibility, and a clear understanding of the modeling process. Moreover, this study employs both traditional ML and AutoML to predict player salaries. AutoML, implemented via PyCaret [12,41], serves as a benchmark to evaluate whether automated model selection and optimization can rival or surpass manually-tuned models. PyCaret's efficient structure enables rapid, consistent experimentation across algorithms, ensuring a fair and comprehensive comparison.

Figure 1. Study workflow from data collection to prediction.

4.1. Traditional Machine Learning Models

This study applies supervised regression models to predict the continuous target variable, *salaryAdjusted*, using player performance features. Both linear and non-linear models are employed to capture diverse data relationships. Linear Regression (LR) is introduced as a baseline owing to its simplicity and interpretability [42]. To address non-linearity and interactions, we include tree-based and boosting methods.

The models include:

- **Linear Regression (LR)**—Establishes a linear relationship between input features and target values [42].
- **Decision Tree (DT)**—Splits data into hierarchical branches based on feature values [43].
- **Gradient Boosting (GB)**—Iteratively improves performance by correcting previous errors [44].
- **XGBoost (XGB)**—An efficient and scalable gradient boosting method optimized for structured data [45].

Model performance is evaluated using multiple metrics: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2), providing a comprehensive overview of accuracy and model fit.

4.2. Automated Machine Learning (AutoML)

To complement traditional models, we employ an AutoML framework to automate model selection, tuning, and evaluation [12]. AutoML reduces manual effort by streamlining preprocessing, feature engineering, and model training. We use PyCaret (v3.3.2) [41], a low-code AutoML framework that enables rapid prototyping and benchmarks models based on predefined metrics. This supports a fair comparison with traditional approaches. Integrating AutoML allows us to evaluate its effectiveness in predicting salaries and determine whether automated pipelines can match or exceed the performance of manually-tuned models.

5. Comparative Evaluation

This section compares four Regression ML models predicting player salaries. Models are evaluated using R², MAE, MSE, and RMSE to assess predictive accuracy. The results offer insight into model performance and guide future enhancements through hyperparameter tuning.

5.1. Traditional Machine Learning Approach

Four ML models are employed to predict professional soccer players' salaries. As shown in Table 7, notable performance improvements are observed after hyperparameter tuning within the traditional machine learning approach. Gradient Boosting (GB) exhibits the most substantial enhancement, with a 62.5% increase in performance. This is followed by XGBoost (XGB), with a 27.7% increase, and Decision Tree (DT), with a modest 4.3% improvement. In contrast, the baseline LR model shows a 28.6% decrease in performance. These results confirm that traditional models with hyperparameter tuning applied using RandomizedSearchCV [46] consistently improve predictive accuracy, with GB demonstrating the highest responsiveness to tuning and emerging as the most effective model for salary prediction in this study.

The tuned parameters in Table 8 enhance model performance by improving generalization and reducing overfitting. After tuning, all three models show improvement. GB achieves the highest R^2 with reduced errors. XGBoost improves in R^2 but shows slightly higher error values, suggesting a bias–variance trade-off. DT shows minor improvement, but error increases, indicating instability. LR underperforms in all metrics, confirming its limitations in modeling non-linear relationships. Overall, GB provides the best balance of accuracy and stability, making it the most reliable model.

Table 7. Performance comparison of traditional ML and AutoML models (pre vs. post tuning).

Model	Approach	Tuning	R^2	MAE (GBP)	MSE (GBP)	RMSE (GBP)
XGBoost (XGB)	Traditional	Pre	0.65	9.96×10^5	3.49×10^{12}	1.87×10^6
	Traditional	Post	0.83	9.95×10^5	3.66×10^{12}	1.91×10^6
	AutoML	Pre	0.69	9.55×10^5	2.94×10^{12}	1.72×10^6
	AutoML	Post	0.60	9.84×10^5	3.18×10^{12}	1.78×10^6
Gradient Boosting (GB)	Traditional	Pre	0.56	1.17×10^6	4.41×10^{12}	2.10×10^6
	Traditional	Post	0.91	1.28×10^4	6.50×10^8	2.55×10^4
	AutoML	Pre	0.46	1.15×10^6	4.21×10^{12}	2.05×10^6
	AutoML	Post	0.46	1.16×10^6	4.27×10^{12}	2.07×10^6
Decision Tree (DT)	Traditional	Pre	0.47	1.01×10^6	5.33×10^{12}	2.31×10^6
	Traditional	Post	0.49	1.13×10^6	5.10×10^{12}	2.26×10^6
	AutoML	Pre	0.35	9.90×10^5	5.13×10^{12}	2.25×10^6
	AutoML	Post	0.31	1.03×10^6	5.42×10^{12}	2.33×10^6
Linear Regression (LR)	Traditional	Pre	0.28	1.37×10^6	7.24×10^{12}	2.69×10^6
	AutoML	Pre	0.20	1.39×10^6	6.32×10^{12}	2.50×10^6

Table 8. Hyperparameter optimization results for different models.

Model	Hyperparameter	Optimized Value
Decision Tree	max_depth	15
	min_samples_leaf	6
	min_samples_split	15
XGBoost	subsample	1.0
	n_estimators	100
	max_depth	9
	learning_rate	0.05
	colsample_bytree	0.6
Gradient Boosting	n_estimators	200
	min_samples_split	15
	min_samples_leaf	2
	max_depth	9
	learning_rate	0.1

5.2. AutoML Framework

To ensure consistency and enable a fair comparison, the same four models, XGB, GB, DT, and LR, are applied in both the traditional ML and AutoML frameworks. This enables for an objective evaluation of AutoML's automated preprocessing, feature selection, and hyperparameter tuning, relative to the manually implemented traditional approach. To optimize performance, RandomizedSearchCV is applied across models for systematic hyperparameter tuning. Table 7 compares PyCaret's AutoML model performance before and after tuning. Among AutoML models, XGB experiences the largest decline in R^2 by 13.0%, while DT shows a smaller reduction of 11.4%, and GB exhibits no improvement. LR, serving as a baseline, is only evaluated before tuning and demonstrates a relatively low R^2. These results suggest that, unlike traditional models, hyperparameter tuning in the AutoML framework does not consistently improve predictive accuracy for this dataset.

5.3. Traditional ML vs. AutoML: Performance Comparison

Table 7 presents a comparison between traditional ML and AutoML approaches, before and after hyperparameter tuning, highlighting the relative increases and decreases in

predictive performance metrics. Traditional models show substantially greater gains in R^2, with GB showing a 97.8% relative increase, followed by DT (58.1%) and XGB (38.3%). In contrast, the LR model shows a 28.6% lower R^2 under the AutoML approach compared to the traditional approach, as no hyperparameter tuning is applied, reflecting baseline differences rather than tuning effects and some AutoML models exhibit minimal gains or declines in R^2 after tuning. GB also outperforms AutoML's XGB in MAE and RMSE, indicating greater accuracy and robustness in limiting large prediction errors. This demonstrates the advantage of expert-guided hyperparameter tuning and customization in traditional ML, which better captures complex, domain-specific relationships in soccer salary data. Although AutoML aims to automate model selection and tuning, our results show that traditional ML models outperform AutoML in this context. Several factors may explain this outcome. First, the AutoML search space may have been constrained or misaligned with the data characteristics, limiting its ability to explore optimal model configurations. Second, the dataset used in this study contains complex patterns and potential distributional shifts (e.g., between positions or leagues), which AutoML might not fully adapt to without domain-specific guidance. Third, although we apply consistent preprocessing across all models, AutoML's internal pipeline may introduce its own steps, leading to potential inconsistencies or redundancies. This may partly explain its underperformance compared to manually tuned traditional models.

5.4. Comparative Analysis of Gradient Boosting with and Without Feature Enhancements

In this section, we examine the effectiveness of our proposed methods and newly engineered features in improving model performance. As shown in Sections 5.1 and 5.2, the GB model outperforms other traditional ML and AutoML approaches, making it the most effective model for salary prediction. We conduct a comparative evaluation of the GB's performance with and without the inclusion of new features and methods.

Table 9 summarizes the results after hyperparameter tuning using RandomizedSearchCV in both cases to ensure consistency. The results show that incorporating new features and methods leads to a substantial improvement in model performance. Specifically, we enhance our model by introducing a new feature, *league_weight*. In addition, we apply engineered features derived from performance data and include position-normalized features to better account for role-specific contributions. R^2 increases by 17%, indicating a better fit and higher explanatory power. Moreover, error metrics such as MAE, MSE, and RMSE decrease, reflecting enhanced predictive accuracy. The improvements confirm that the enhanced feature set significantly boosts GB's accuracy in salary prediction.

Table 9. Performance comparison of the GB model without and with new features and methods.

Approach	R^2	MAE (GBP)	MSE (GBP)	RMSE (GBP)
Without new features & methods	0.78	1.84×10^4	8.69×10^8	2.95×10^4
With new features & methods	0.91	1.28×10^4	6.50×10^8	2.55×10^4
Improvement (%)	16.67%	30.43%	25.20%	13.56%

6. Evaluation Results

6.1. SHAP-Based Feature Importance Analysis

To improve interpretability and performance, (SHAP) analysis was employed for feature selection. SHAP assigns importance values to features based on their contribution to model predictions, ensuring consistent and locally accurate attributions [47,48]. We applied SHAP to four models: LR, DT, GB, and XGB, identifying the top 25 features influencing salary prediction. Moreover, for all SHAP analyses, the feature importance

values are computed using the test dataset to ensure unbiased and reliable interpretation of the models' predictive behavior on unseen data.

Figure 2a–d show these features ranked by SHAP values. LR serves as a baseline, highlighting only three meaningful features: *goals*, *xG*, and *xGdiff*. This reflects its limitation to direct linear relationships, ignoring collinear or weaker predictors.

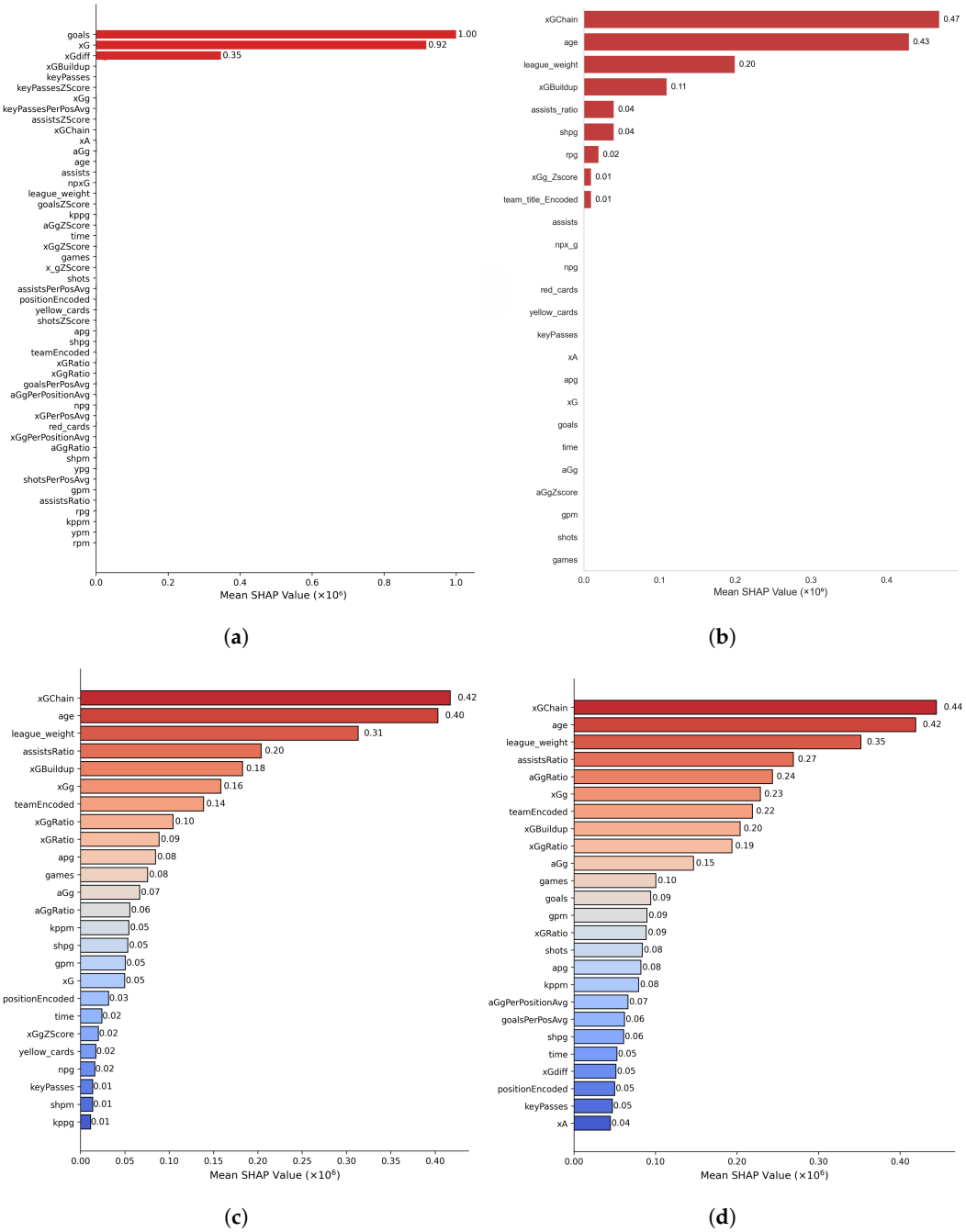


Figure 2. Comparison of feature importance using SHAP values across ML models. (**a**) Linear Regression (LR); (**b**) Decision Tree (DT); (**c**) Gradient Boosting (GB); (**d**) XGBoost Regressor (XGB).

DT captures non-linear interactions, identifying *xGChain*, *age*, and *league_weight* as top features. It emphasizes players' involvement in buildup play, career stage, and league strength, with *xGBuildup* and *assistsRatio* also being influential. Traditional metrics like *goals*, *xG*, and *key_passes* show little impact, indicating that DT values broader offensive and career factors over isolated scoring stats.

GB similarly ranks *xGChain*, *age*, and *league_weight* highest, reinforcing the importance of dynamic offensive roles and league quality. Features like *assistsRatio*, *xGBuildup*, and *xGg* support the significance of playmaking and scoring potential. Moderate influence appears for shooting and passing metrics (*kppm*, *shpg*, *gpm*), while defensive and discipline metrics have less effect. Lower-ranked features include *key_passes* and *npg*, highlighting a focus on overall attacking contribution.

XGB, an optimized GB variant, also prioritizes *xGChain*, *age*, and *league_ratio*, emphasizing passing sequences, career stage, and league competitiveness. Additional important features include *assistsRatio*, *aGgRatio*, *xGg*, and encoded categorical variables such as *teamEncoded*, and *positionen*, reflecting team strength and positional roles, respectively.

Overall, SHAP analysis consistently identifies league competitiveness (*league_weight*), goal-related contributions, and positional context as key drivers in predicting player salaries. The growing importance of *league_weight* across more complex models highlights their enhanced capacity to capture contextual performance factors. Notably, *league_weight* ranks among the top three contributors in all models analyzed, with SHAP values of approximately 0.20, 0.31, and 0.35 for DT, GB, and XGB, respectively. This demonstrates that our newly engineered league weighting feature significantly improves the models' explanatory power regarding salary discrepancies, affirming its relevance in capturing league-specific salary effects.

Individual goal contributions, specifically *goals* and *assists*, consistently rank as key predictors across all models due to their fundamental role in evaluating player performance and value. Goals directly impact match results and are often the primary factor fans, clubs, and sponsors associate with a player's effectiveness. Assists highlight a player's ability to create scoring opportunities for teammates, reflecting creativity and vision—qualities highly prized in professional soccer. These two statistics are widely recognized as the most concrete and quantifiable measures of offensive contribution, which strongly influence contract negotiations and salary levels. Their consistent importance across different models underscores the centrality of direct involvement in goal-scoring actions as a primary driver of player salary.

Moreover, two advanced performance metrics, *xGChain* and *xGBuildup*, emerge as key contributors in the SHAP analysis. *xGChain* quantifies the total expected goals (xG) value of all attacking actions a player participates in during a possession sequence culminating in a shot. This metric effectively captures the player's cumulative influence throughout buildup and passing sequences, highlighting their integral role in offensive plays beyond merely taking shots. Conversely, *xGBuildup* isolates the expected goals generated during the buildup phase by excluding key passes and shots. This emphasizes a player's contribution in advancing the ball and facilitating scoring opportunities prior to decisive attacking actions. Collectively, these metrics enrich the evaluation of a player's offensive impact by encompassing both direct and indirect contributions to goal-scoring chances, underscoring the multifaceted nature of attacking involvement. While we acknowledge the presence of highly correlated features such as *xG* and *xGg* or *aG* and *aGg*, these engineered per game metrics provide important standardization for fair comparisons across players with differing playtime. Given the robustness of tree-based models used in our study to multicollinearity, we retain these features to capture complementary information.

6.2. Salary Discrepancy Analysis Using GB

As shown in Section 5, the GB model is the most effective for predicting player salaries. Using this model, we analyze salary discrepancies by comparing actual versus predicted earnings. Players are classified as underestimated (predicted salary higher than actual) or overestimated (actual salary higher than predicted). To improve prediction

reliability, we use the top 25 influential features identified via SHAP analysis. This approach highlights potential market inefficiencies and league-specific salary patterns. We also examine factors such as age, league, and position to better understand the drivers behind these discrepancies.

Table 10 lists the top five underestimated and overestimated players. Each has an actual salary of GBP 2.1 M, yet their predicted salaries reveal large gaps. L. Messi shows the largest shortfall with a predicted GBP 26.3 M, a GBP 24.2 M difference. Neymar (GBP 21.4 M) and A. Griezmann (GBP 11.9 M) also have significant underestimations. These gaps suggest valuation misalignments influenced by contracts, wage policies, salary structures, and aging effects. Among overestimated players, actual salaries exceed predicted values, indicating potential overvaluation. Marcelo earns GBP 21.95 M compared to a predicted GBP 1.99 M, a GBP 19.96 M gap. E. Hazard (GBP 24.9 M), L. Suárez (GBP 24.7 M), David de Gea (GBP 16.82 M), and G. Bale (GBP 20.7 M) follow with similar discrepancies. Furthermore, Table 11 shows that most players (3449) are underestimated, while 1789 are overestimated, indicating that the model generally predicts lower than actual salaries.

Table 10. Top 5 underestimated and overestimated players based on salary difference (in GBP M).

Player	Adj. Salary (GBP M)	Pred. Salary (GBP M)	Diff. (GBP M)
Underestimated players			
L. Messi	2.1	26.3	24.2
Neymar	2.1	23.5	21.4
A. Griezmann	2.1	14.0	11.9
R. Lewandowski	2.1	13.8	11.7
K. De Bruyne	2.1	13.1	11.0
Overestimated players			
Marcelo	21.95	1.99	−19.96
E. Hazard	28.1	3.3	−24.9
L. Suárez	27.0	2.3	−24.7
David de Gea	19.50	2.68	−16.82
G. Bale	28.1	7.4	−20.7

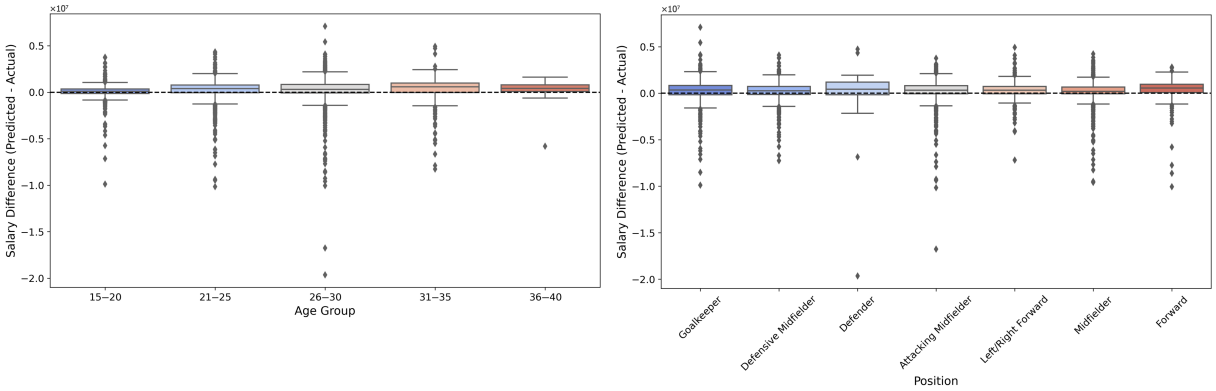
Table 11. Distribution of salary categories for unique players.

Salary Category	Number of Unique Players
Overestimated	1789
Underestimated	3449

We analyzed how salary predictions vary across different age groups and player positions to gain deeper insights into the internal factors influencing salary estimations. Both features consistently emerge as significant predictors in our ML models, underscoring their strong influence on salary outcomes. Moreover, age and position are well-established determinants of player compensation in professional football, significantly affecting a player's market value, performance expectations, and career trajectory. These factors are critical for understanding salary disparities.

Figure 3a presents a detailed comparison of predicted and actual salary differences across career stages. The results show that the model effectively captures salary trends by age, particularly for mid-career players (26–30 years), where predictions closely match actual salaries. This indicates that the model incorporates key performance factors influencing peak-year salaries. For younger players (15–25 years), some underestimations likely reflect real-world contracts, where emerging talents earn lower base wages before renegotiations.

Thus, these predictions may realistically represent early-stage salary trajectories rather than systematic bias. For older players (31+ years), occasional overestimations may stem from legacy contracts or wage structures extending beyond peak performance, indicating that the model accounts for salary stability due to long-term agreements. Figure 3b analyzes the impact of player position on salary predictions, revealing close alignment between predicted and actual salaries across positions. Goalkeepers and defensive midfielders show consistent salary differences, reflecting more predictable salary structures, likely driven by standardized contracts. In contrast, forwards and midfielders exhibit greater variability, with some earning significantly less than predicted, indicating more fluctuation in salary expectations for these roles.



(**a**) (**b**)

Figure 3. Comparison of salary discrepancies by demographic groups. (**a**) Analysis by age groups; (**b**) analysis by position groups.

In summary, the distribution of salary discrepancies reveals a predominance of underestimated players, suggesting the model tends to predict lower salaries overall. Predictions align more closely with actual salaries for mid-career players, while younger and older players experience some under and overestimations, respectively. Positional analysis further highlights that goalkeepers and defensive midfielders have more stable salary patterns, whereas forwards and midfielders show greater variability. Moreover the group-level patterns illustrated in Figure 3a,b provide important context for the individual-level discrepancies reported in Table 10. The consistent underestimation of prominent forwards and attacking midfielders in their mid-career stage (e.g., Messi, Neymar) appears to stem from the model's reliance on measurable performance indicators, which may not fully capture broader determinants of salary such as commercial value or contractual nuances. Conversely, the overestimation of older defenders and goalkeepers (e.g., Marcelo, David de Gea) suggests that the model does not always reflect the influence of long-term or legacy contracts that may not align with current on-field contributions. This alignment between group-level trends and individual-level deviations enhances the interpretability of the results and offers insight into the structural factors influencing salary prediction accuracy.

6.3. Comparison with State-of-the-Art

Several prior studies investigated player salary prediction using traditional ML technologies, often relying on basic features like goals, assists, and appearances. As shown in Table 12, these models generally lack advanced performance metrics or contextual feature engineering, which limits their predictive power.

Table 12. Comparison of existing studies with the proposed model. R^2 is a unitless metric representing the proportion of variance explained, while all monetary error metrics (MAE, MSE, RMSE) are reported in pounds (GBP).

Related Work	Key Features	Used Methods	Results
Smith et al. (2021) [49]	Goals, assists, appearances	Linear Regression, Random Forest	R^2: 0.71, MAE: 8.40×10^4, RMSE: N/A
Huang C, Zhang Sh (2023) [10]	Streamlined features	Gradient Boosting Decision Tree	R^2: 0.90, MAE: N/A, RMSE: 3.22×10^6
Rong et al. (2024) [21]	Goals, assists, minutes	Tree-based models	Not reported
Li et al. (2022) [6]	Current age, position, achievements	ML with GridSearchCV	R^2: 0.60, MAE: N/A, RMSE: N/A
Frick et al. (2007) [13]	Player performance, goals, national team membership	Not specified	Not reported
Müller & Simons (2017) [50]	Age, height, minutes played	Statistical modeling	R^2: N/A, MAE: N/A, RMSE: 1.80×10^7
Proposed Work	xGg, aGg, league/position weights	Gradient Boosting, Linear Regression	R^2: 0.91, MAE: 1.28×10^4, RMSE: 1.71×10^6

For example, Smith et al. [49] reported a moderate R^2 of 0.71 using standard performance indicators, though their error metrics such as RMSE, were not provided. Huang and Zhang [10] included an error metric, reporting an R^2 of 0.90, of approximately 3.22×10^6 , but did not incorporate predicted performance metrics such as expected goals. In contrast, the proposed model integrates refined predicted metrics like *xGg* and *aGg*, along with league and position adjustments, which enhances its predictive capability. It achieves a substantially higher R^2 of 0.91, representing a significant improvement in explained variance over previous works, and reduces prediction errors with an MAE of approximately 9.28×10^5. Additionally, our dataset and models differ from others, outperforming all existing metrics. This comparison highlights the benefits of integrating domain-specific features with advanced ML techniques, leading to more precise and dependable salary predictions in soccer.

7. Limitations and Discussion

While this study provides useful insights into soccer salary prediction, several limitations should be noted. First, the model uses historical data and does not account for real-time factors like form, injuries, or transfers, which can affect salary assessments. Second, the lack of physical and biometric data, whcih are often unavailable or confidential, limits the model's ability to reflect fitness and injury risks. Third, differences in salary structures across leagues may reduce generalizability. Additionally, the salary data, sourced from Capology via the Edd Webster repository, includes both verified and estimated values without clear distinction, introducing some uncertainty. While this may affect individual accuracy, the dataset still allows for a meaningful analysis of overall salary patterns.

A further challenge arises from discrepancies in reported salary values for some high-profile players, with certain adjusted figures appearing implausibly low, likely due to imputed or placeholder entries where exact amounts were unavailable or unverified. These inconsistencies can lead to notable gaps between reported and predicted salaries. While such anomalies reflect limitations in publicly available salary data, our model is designed to capture broader salary trends rather than validate individual cases. Importantly, the predicted salaries align reasonably well with expected market values, supporting the robustness of the results despite imperfections in the input data.

The validity of this analysis may be influenced by the reliability of the salary data employed in the study. Due to the lack of verified or standardized salary records and

the absence of quantitative measures of uncertainty, precisely assessing the accuracy of the salary information remains challenging. While our dataset includes players from all positions, the engineered features are primarily based on attacking metrics due to their wider availability and consistency. This may introduce some positional bias, particularly for defenders and goalkeepers, whose contributions are better reflected through defensive statistics such as tackles, interceptions, and saves. However, aligning such data with our salary dataset proved challenging due to inconsistent identifiers and formatting issues. To mitigate this, we applied position-based normalization techniques (see Table 4) to promote representation. While position normalization helps adjust for role-specific differences and may assist the model's performance, we acknowledge that it has limitations, particularly for non-attacking players. Metrics such as normalized *xG* are less meaningful for defenders, whose value often lies in actions not captured by goal-related statistics. Therefore, our model may be less effective in predicting salaries for non-attacking players, as the dataset lacks sufficient representation of defensive contributions.

Fourth, external economic factors like salary inflation and free agency dynamics are not considered, despite their strong influence on contracts. Additionally, non-performance factors such as commercial value, endorsements, age, injury history, and contract structures are excluded due to limited data, though they significantly affect player compensation.

8. Conclusions

This study predicts soccer players' salaries in top European leagues using enhanced feature engineering that captures league competitiveness, player roles, and contextual performance. Introducing a position ratio addresses scoring bias across roles, enabling fairer salary comparisons. We compare traditional ML models and AutoML with hyperparameter tuning, identifying GB Regressor as the best performer. A key contribution is classifying players as overestimated or underestimated based on salary prediction errors, highlighting discrepancies and factors beyond goals and assists such as playmaking, league quality, and roles. Moreover, to better understand the key drivers behind salary predictions, we employ SHAP-based feature importance analysis, which reveals that league competitiveness (*league_weight*), offensive contributions including goals, assists, and passing sequences, as well as player age and positional context, are the most influential factors. This comprehensive analysis underscores the multifaceted nature of salary determination, demonstrating that contextual and role-specific features significantly contribute alongside traditional performance metrics.

Author Contributions: Conceptualization, D.M. and J.K.; methodology, D.M. and J.K.; data curation, D.M.; validation D.M.; writing—original draft preparation, D.M.; formal analysis D.M.; writing—review and editing, P.J.; visualization, P.J.; supervision, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant RS-2023-00209720

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We sincerely thank the members of our laboratory at Gyeongsang National University for their thoughtful discussions and insightful feedback, which have significantly contributed to the progress and completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Şener, İ.; Karapolatgil, A.A. Rules of the game: Strategy in football industry. *Procedia-Soc. Behav. Sci.* **2015**, *207*, 10–19. https://doi.org/10.1016/j.sbspro.2015.10.143.
2. Richter, F. The Global Game of Football. 2025. Available online: https://www.statista.com/chart/31460/world-football-day/ (accessed on 15 January 2025).
3. Srinivasan, H. These Are the Highest-Paid Athletes in the World as of 2025. 2025. Available online: https://www.investopedia.com/highest-paid-athletes-8770167 (accessed on 15 January 2025).
4. Malikov, D.; Kim, J. Beyond xG: A Dual Prediction Model for Analyzing Player Performance Through Expected and Actual Goals in European Soccer Leagues. *Appl. Sci.* **2024**, *14*, 10390. https://doi.org/10.3390/app142210390.
5. Elahi, M.; Pandey, S.; Malhi, S.S. Market Value Prediction of Football Players. In Proceedings of the KILBY 100 7th International Conference on Computing Sciences (ICCS 2023), India Kilby 2023,5th May. https://doi.org/10.2139/ssrn.4485449.
6. Li, C.; Kampakis, S.; Treleaven, P. Machine learning modeling to evaluate the value of football players. *arXiv* **2022**, arXiv:2207.11361. https://doi.org/10.48550/arXiv.2207.11361.
7. Stafylidis, A.; Mandroukas, A.; Michailidis, Y.; Vardakis, L.; Metaxas, I.; Kyranoudis, A.E.; Metaxas, T.I. Key Performance Indicators Predictive of Success in Soccer: A Comprehensive Analysis of the Greek Soccer League. *J. Funct. Morphol. Kinesiol.* **2024**, *9*, 107. https://doi.org/10.3390/jfmk9020107.
8. Shen, Q. Predicting the value of football players: Machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets. *Appl. Intell.* **2025**, *55*, 265. https://doi.org/10.1007/s10489-024-06189-0.
9. Al-Asadi, M.A.; Tasdemir, S. Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. *IEEE Access* **2022**, *10*, 22631–22645. https://doi.org/10.1109/ACCESS.2022.3154767.
10. Huang, C.; Zhang, S. Explainable Artificial Intelligence Model for Identifying Market Value in Professional Soccer Players. *arXiv* **2023**, arXiv:2311.04599. https://doi.org/10.48550/arXiv.2311.04599.
11. Lee, H.; Tama, B.A.; Cha, M. Prediction of football player value using Bayesian ensemble approach. *arXiv* **2022**, arXiv:2206.13246. https://doi.org/10.48550/arXiv.2206.13246.
12. AutoML.org, F.H.T. AutoML: The Standard in Automated Machine Learning. Available online: https://www.automl.org/automl/ (accessed on 15 February 2025).
13. Frick, B. The football players' labor market: Empirical evidence from the major European leagues. *Scott. J. Political Econ.* **2007**, *54*, 422–446. https://doi.org/10.1111/j.1467-9485.2007.00423.x.
14. Ribeiro, A.S.; Lima, F. Labour Mobility Effect on Wages: The Professional Football Players' Case. 2013. Available online: https://www.academia.edu/67498148/Labour_mobility_effect_on_wages_the_professional_football_players_case (accessed on 15 January 2025).
15. Késenne, S. *The Economic Theory of Professional Team Sports: An Analytical Treatment*; Edward Elgar Publishing: Cheltenham, UK, 2007. Available online: https://archive.org/details/economictheoryof0000kese/page/n5/mode/2up (accessed on 30 January 2025).
16. McLeod, C.M.; Li, H.; Nite, C. What Enables Human Capital Investment Sharing in Elite Sport? *Sustainability* **2022**, *14*, 10628. https://doi.org/10.3390/su141710628.
17. Singla, P. Player Power: Exploring the Impact of Player Metrics on the Valuation of Football Clubs. *SSRG Int. J. Econ. Manag. Stud.* **2024**, *11*, 44–51. https://doi.org/10.14445/23939125/IJEMS-V11I8P106.
18. Müller, O.; Simons, A.; Weinmann, M. Beyond crowd judgments: Data-driven estimation of market value in association football. *Eur. J. Oper. Res.* **2017**, *263*, 611–624. https://doi.org/10.1016/j.ejor.2017.05.005.
19. Elahi, M.; Pandey, S.; Malhi, S.S. Market Value Prediction of Football Players. In Proceedings of the KILBY 100 7th International Conference on Computing Sciences, Kilby, 2023. https://doi.org/10.2139/ssrn.4485449.
20. Majewski, S. Identification of factors determining market value of the most valuable football players. *Cent. Eur. Manag. J.* **2016**, *24*, 91–104. https://doi.org/10.7206/jmba.ce.2450-7814.177.
21. Rong, Z.; Wang, L.; Xie, S. Factors that Influence Player Market Value in Different Positions: Evidence from European Leagues. *Adv. Econ. Manag. Political Sci.* **2024**, *82*, 50–63. https://doi.org/10.54254/2754-1169/82/20230718.
22. Bhilawa, L.; Fahriansyah, R. The Influence of Performance, Age, and Nationality on the Market Value of Football Players. *Assets: J. Akunt. Dan Pendidik.* **2022**, *11*, 1–9. https://doi.org/10.25273/jap.v11i1.8422.
23. Margareta, L.M.; Malinda, O. The Effect of Performance, Age, Transfer Fee and Salary to the Market Value of Professional Players: Empirical Studies in European Leagues Football Clubs. *Int. J. Glob. Oper. Res.* **2022**, *3*, 148. https://doi.org/10.47194/ijgor.v3i3.148.
24. Yaldo, L.; Shamir, L. Computational Estimation of Football Player Wages. *Int. J. Comput. Sci. Sport* **2017**, *16*, 18–38. https://doi.org/10.1515/ijcss-2017-0002.

25. Lee, H.; Tama, B.A.; Cha, M. Prediction of Football Player Value using Bayesian Ensemble Approach. *arXiv* **2022**, arXiv:2206.13246. https://doi.org/10.48550/arXiv.2206.13246.
26. Berri, D.; Butler, D.; Rossi, G.; Simmons, R.; Tordoff, C. Salary determination in professional football: Empirical evidence from goalkeepers. *Eur. Sport Manag. Q.* **2023**, *23*, 624–640. https://doi.org/10.1080/16184742.2023.2169319.
27. David, B.; Alex, F.R.S. Do sports analytics affect footballer pay? *Front. Behav. Econ.* **2024**, 3,*1490871*. https://doi.org/10.3389/frbhe.2024.1490871.
28. Smark, C. Editorial: AABFJ Volume 8, Issue 5. *Australas. Account. Bus. Financ. J.* **2015**, *8*, 1–2. https://doi.org/10.14453/aabfj.v8i5.1.
29. Ahmad, A.; Slem, O. Football Players Full Analysis and Modelling, 2023. Kaggle Project. Available online: https://www.kaggle.com/code/anasahmad25/football-players-full-analysis-and-modelling#Missing-Values (accessed on 11 July 2025).
30. He, X.; Zhao, K.; Chu, X. AutoML: A Survey of the State-of-the-Art. *Knowl.-Based Syst.* **2021**, *212*, 106622. Available online: https://www.sciencedirect.com/science/article/pii/S0950705120307516 (accessed on 15 February 2025).
31. Hutter, F.; Kotthoff, L.; Vanschoren, J. *Automated Machine Learning: Methods, Systems, Challenges*; Springer Nature: Cham, Switzerland, 2019. Available online: https://link.springer.com/book/10.1007/978-3-030-05318-5 (accessed on 15 February 2025).
32. Julia, M. Towards Explainable Automated Machine Learning. ACM. 2023. Available online: https://doi.org/10.5282/edoc.32176 (accessed on 15 February 2025).
33. Understat. Understat Professional Soccer Website. 2022. Available online: https://understat.com/ (accessed on 4 February 2025).
34. Capology. Capology: Soccer Salaries and Contracts. 2024. Available online: https://www.capology.com/ (accessed on 4 February 2025).
35. Webster, E. Soccer Analytics. 2022. Available online: https://github.com/eddwebster/football_analytics/tree/master/data/understat/raw/metadata (accessed on 4 February 2025).
36. Scikit-Learn Developers. LabelEncoder—Scikit-Learn Documentation. Scikit-Learn. 2024. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html (accessed on 5 February 2025).
37. GeeksforGeeks. HashingEncoder. Education Website. Available online: https://www.geeksforgeeks.org/dsa/encryption-encoding-hashing/ (accessed on 5 February 2025).
38. Acco, B. What Is the Most Important Position in Soccer? 2024. Available online: https://www.playermaker.com/blog/most-important-position/ (accessed on 21 February 2025).
39. Alcheva, M. Which Position in Soccer Gets Paid the Most? 2023. Available online: https://worldsoccertalk.com/news/which-position-in-soccer-gets-paid-the-most-20231116-WST-470643.html (accessed on 21 February 2025).
40. Kassiesa, B. UEFA Coefficients for Club Competitions. 2025. Available online: https://kassiesa.net/uefa/data/method5/crank2025.html (accessed on 5 February 2025).
41. Ali, M. *PyCaret: An Open-Source, Low-Code Machine Learning Library in Python*, PyCaret version 1.0.0. 2020. Available online: https://pycaret.org/ (accessed on 25 February 2025).
42. learn Developers, S. LinearRegression—Scikit-Learn. 2024. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (accessed on 25 February 2025).
43. learn Developers, S. Decision Trees. 2024. Available online: https://scikit-learn.org/stable/modules/tree.html (accessed on 10 March 2024).
44. learn Developers Boosting, S. GradientBoostingRegressor—Scikit-Learn. 2024. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (accessed on 25 February 2025).
45. Chen, T.; Guestrin, C. XGBoost: Scalable and Flexible Gradient Boosting. 2016. Available online: https://xgboost.readthedocs.io/ (accessed on 25 February 2025).
46. Scikit-Learn Developers. Sklearn.model_selection.RandomizedSearchCV. 2024. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (accessed on 11 March 2025).
47. Christoph, M. Interpretable Machine Learning 2022. Available online: https://christophm.github.io/interpretable-ml-book/shap.html#shap (accessed on 3 March 2025).
48. Lundberg, S. SHAP Documentation. 2018. Available online: https://shap.readthedocs.io/en/latest/ (accessed on 10 March 2025).
49. Yiğit, A.T.; Samak, B.; Kaya, T. Football Player Value Assessment Using Machine Learning Techniques. In *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making*; Advances in Intelligent Systems and Computing; Kahraman, C., Cebi, S., Cevik Onar, S., Oztaysi, B., Tolga, A., Sari, I., Eds.; Springer: Cham, Switzerland, 2020; Volume 1029, pp. 435–444. https://doi.org/10.1007/978-3-030-23756-1_36
50. Müller, O.; Simons, A. Beyond Crowd Judgments: Data-Driven Estimation of Market Value in Association Football. *Eur. J. Oper. Res.* **2017**, *263*, 611–624. https://doi.org/10.1016/j.ejor.2017.05.005.