UNIVERZA V LJUBLJANI

SCHOOL OF BUSINESS AND ECONOMICS

Master's Thesis

# ALGORITHMIC INSIGHTS INTO FOOTBALL TRANSFER VALUATIONS: A DATA ANALYTICS APPROACH

Ljubljana, July 2025 EMIR KULAČIČ

# AUTHORSHIP STATEMENT

The undersigned Emir Kulačić, a student at the University of Ljubljana, School of Economics and Business, (hereafter: UL SEB), author of this written final work of studies with the title "Algorithmic Insights into Footbal Transfer Valuations: A Data Analytics Approach, prepared in collaboration with mentor doc. assist. prof. Uroš Godnov

D E C L A R E

1. this written final work of studies to be based on the results of my own research;

2. the printed form of this written final work of studies to be identical to its electronic form;

3. the text of this written final work of studies to be language-edited and technically in adherence with the UL SEB Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the UL SEB Technical Guidelines for Written Works;

4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;

5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the UL SEB in accordance with the relevant UL SEB Rules;

6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;

7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;

8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the UL SEB Study Information System;

9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;

10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published;

11. that I have verified the authenticity of the information derived from the records using artificial intelligence tools.

Ljubljana, June 21nd 2025 Author's signature: 

**ABSTRACT**

This thesis looks at different ways to predict football player transfer fees using both traditional and modern data methods. It includes statistics from the top five European leagues and compares standard models like OLS, Ridge, and Lasso regression with machine learning models. The goal is to see which models give better results. To deal with problems like too many variables or data that does not follow a clear pattern, the study uses techniques like feature engineering and Principal Component Analysis. The results show that machine learning often gives more accurate predictions, but football is still very hard to model because of its fast and unpredictable nature. The findings give useful information for clubs, agents, and other people working in football.

**KEY WORDS**

Football, Transfer Fees, Player Valuation, Machine Learning, Regression Models

**SUSTAINABLE DEVELOPMENT GOALS**




**POVZETEK**

To magistrsko delo proučuje različne pristope k napovedovanju prestopnih vrednosti nogometašev. V analizi so uporabljeni podatki iz petih najmočnejših evropskih lig, kjer so primerjani klasični regresijski modeli, kot so OLS, Ridge in Lasso, z naprednejšimi algoritmi strojnega učenja, kot so odločitvena drevesa in nevronske mreže. Glavni cilj raziskave je ugotoviti, kateri pristop omogoča natančnejše napovedi. Pri tem delo obravnava tudi težave, kot so preveliko število spremenljivk in nenavadni vzorci v podatkih, zato se uporabljajo metode, kot sta inženiring značilk in analiza glavnih komponent (PCA). Rezultati kažejo, da so modeli strojnega učenja pogosto natančnejši, vendar nogomet zaradi svoje kompleksnosti in nepredvidljivosti ostaja zelo zahteven za modeliranje. Ugotovitve lahko pomagajo nogometnim klubom, agentom in drugim strokovnjakom pri bolj podatkovno podprtih odločitvah.

**KLJUČNE BESEDE**

nogomet, prestopne vrednosti, vrednotenje igralcev, strojno učenje, regresijski modeli

**CILJI TRAJNOSTNEGA RAZVOJA**



# TABLE OF CONTENTS

1 INTRODUCTION ..... 4
1.1 Background ..... 4
1.2 Problem Statement ..... 5
1.3 Objectives and Research Questions ..... 6
2 LITERATURE REVIEW ..... 7
2.1 Traditional Econometric Approaches in Player Valuation ..... 7
2.2 Transition to Data-Driven and Machine Learning Methods ..... 8
2.3 Key Determinants of Market Value ..... 9
2.4 Impact of Machine Learning on Valuation Accuracy ..... 10
2.5 Gaps and Future Directions ..... 11
2.6 Contribution of This Thesis ..... 12
3 METHODOLOGY ..... 13
3.1 Data Collection and Web Scraping ..... 14
3.2 Dataset Matching and Integration ..... 14
3.3 Data Cleaning and Preprocessing ..... 15
3.4 Target Variable Selection ..... 16
3.5 Feature Encoding and Normalization ..... 16
3.6 Final Dataset Composition and Tools ..... 17
4 EXPLORATORY DATA ANALYSIS (EDA) ..... 18
4.1 Descriptive Statistics and Transfer Fee Distribution ..... 18
4.2 Correlation Analysis of Performance Metrics ..... 20
4.3 Feature-Target Relationships and EDA Conclusions ..... 24
5 DATA PREPARATION AND FEATURE ENGINEERING ..... 25
5.1 Feature Engineering of Performance Metrics ..... 25
5.2 Handling Multicollinearity: VIF Filtering and PCA ..... 29
5.3 Temporal Weighting of Training Data ..... 32
5.4 Summary of Prepared Data ..... 33
6 REGRESSION MODELS ..... 34
6.1 Model Training on Basic vs. Advanced Feature Sets ..... 34
6.2 Interpreting Regression Model Coefficients ..... 37
6.3 Regression Model Evaluation ..... 38

7 MACHINE LEARNING MODELS ..... 40
7.1 Model Training and Tuning ..... 40
7.2 Performance of Ensemble Models ..... 40
7.3 Comparison of ML Models and Linear Models ..... 43
8 EVALUATION AND DISCUSSION OF RESULTS ..... 44
8.1 Model Evaluation using Hold-out and Cross-Validation ..... 45
8.2 Discussion: Insights and Implications ..... 46
8.3 Limitations and Future Work ..... 49
9 CONCLUSION ..... 51
LIST OF KEY LITERATURE ..... 52
REFERENCES ..... 52

## LIST OF FIGURES

Figure 1:Transfer Fee Distribution. ..... 19
Figure 2: Top features from the Basic Stats dataset correlated with transfer fees. ..... 22
Figure 3:Top advanced metrics correlated with transfer fees. ..... 22
Figure 4:Correlation heatmap of key features. ..... 23
Figure 5:Distributions of selected engineered features (Basic Stats dataset). ..... 28
Figure 6:Distributions of selected engineered features (Advanced Stats dataset). ..... 28
Figure 7:PCA explained variance by component for advanced performance metrics. ..... 30
Figure 8:Top 10 feature loadings for PC1 ..... 31
Figure 9:Top 10 feature loadings for PC2 ..... 32
Figure 10:Performance of regression models on Basic vs. Advanced feature sets ..... 36
Figure 11:Model evaluation: Ridge on Advanced Stats ..... 39
Figure 12: Distribution of Residuals: Ridge on Advanced Stats ..... 42
Figure 13:Distribution of Residuals: XGBoost on Advanced Stats ..... 42
Figure 14 Model Performance Comparison ..... 44

## LIST OF ABBREVIATIONS

**CV –** (sl. navzkrižna validacija); Cross-Validation

**EDA –** (sl. raziskovalna analiza podatkov); Exploratory Data Analysis

**GCA –** (sl. akcija, ki vodi do zadetka); Goal-Creating Action

**MAE –** (sl. povprečna absolutna napaka); Mean Absolute Error

**ML –** (sl. strojno učenje); Machine Learning

**OLS –** (sl. navadna najmanjših kvadratov regresija); Ordinary Least Squares

**PCA –** (sl. analiza glavnih komponent); Principal Component Analysis

**PC –** (sl. glavna komponenta); Principal Component

**$R^2$ –** (sl. koeficient determinacije); Coefficient of Determination

**RMSE –** (sl. kvadratna srednja napaka); Root Mean Squared Error

**SCA –** (sl. akcija, ki vodi do strela); Shot-Creating Action

**UEFA –** (sl. Evropska nogometna zveza); Union of European Football Associations

**VIF –** (sl. faktor napihnjenosti variance); Variance Inflation Factor

**xA –** (sl. pričakovane asistence); Expected Assists

**xG –** (sl. pričakovani zadetki); Expected Goals

# 1 INTRODUCTION

## 1.1 Background

As a teenager obsessed with football transfers, I spent hours on Transfermarkt watching player values rise and fall. I vividly recall the thrill of seeing a promising striker's estimated price surge after a brilliant season, only to watch a club later pay twice that amount in the real market. Those early experiences were captivating, but they also revealed a troubling truth: the ways we gauge a player's worth are often inadequate and inconsistent.

This inadequacy in valuation is more than a personal observation – it is a systemic challenge for modern football. Today, the sport is a global industry valued at over $40 billion, and a single transfer can involve fees in the hundreds of millions. Clubs invest these vast sums to strengthen their squads in hopes of achieving greater success. A poor investment in a player, however, can spell disaster: financial instability, on-field underperformance, or even the threat of relegation, all of which jeopardize a club's long-term viability. Transfer dealings have also become a significant source of revenue for many teams. With so much at stake, all stakeholders, from club owners and managers to agents and analysts, need reliable ways to assess a player's market value.

Yet, despite these high stakes, there is not yet a universally accepted method for determining a footballer's true market value. Many clubs still rely on overly simplistic valuation techniques: basic statistics (goals, assists, appearances) combined with intuition or rough rules of thumb. These metrics are intuitive but paint an incomplete picture of a player's contribution. Simple tallies cannot capture context or non-linear effects – for example, ten goals by a young defender may carry a very different significance than ten goals by an experienced striker. Traditional one-size-fits-all formulas struggle to account for such nuances, meaning that important factors like player role, potential, and team context often get overlooked.

At the same time, crowd-sourced valuations have become a go-to solution in the absence of hard data. Platforms like Transfermarkt aggregate community and expert opinions to approximate a player's market price. My own early engagement with Transfermarkt was eye-opening: it fueled my passion for football statistics but also showed how subjective and inconsistent these crowd estimates could be. Indeed, studies have found that Transfermarkt's values often fall well short of actual transfer fees, with systematic biases in their predictive accuracy. Such crowd estimates can lag behind fast-changing market dynamics, apply unevenly to lesser-known players, and are vulnerable

to uninformed or biased inputs. In the high-pressure transfer market, relying on these imperfect benchmarks can lead clubs to overpay for over-hyped talent or miss out on hidden gems that are undervalued by consensus.

Fortunately, recent trends offer a path toward more objective valuations. The past decade has seen an explosion in football data and analytics. Advanced performance metrics – from expected goals (xG) and shot-creating actions to defensive pressures and successful dribbles – provide a far more nuanced view of player contributions than traditional stats like goals or assists. These modern metrics capture dimensions of play (creative passing, off-ball movement, defensive impact, etc.) that basic statistics miss, offering deeper insight into a player's value. At the same time, rapid advances in machine learning have opened up new ways to analyze complex datasets and detect patterns that conventional econometric models might overlook. Yet despite this potential, data-driven predictive models for football player valuation remain underdeveloped; their promise for improving accuracy has not been fully realized.

The foundation of this thesis stems from the convergence of rich data availability and improving computational techniques. In this study, we prioritize contemporary performance metrics and machine learning methods over purely traditional analyses. By combining football-specific domain knowledge with data-driven analytics, we aim to develop a more advanced framework for player valuation that improves accuracy and reliability. This approach is driven by the belief that bridging conventional valuation approaches with modern, data-driven techniques can significantly enhance how player market values are assessed.

## 1.2 Problem Statement

Formally stated, the central problem addressed in this research is the lack of a rigorous, data-encompassing methodology for valuing football players. Current market value assessments often rely on a limited set of conventional performance metrics and subjective judgment, failing to account for the multitude of factors that can influence a player's worth. Simple linear models based on goals or appearances, while easy to use, ignore nonlinear relationships and context. As noted, two players with identical statistics might have very different values once age, position, and role are considered – nuances that traditional models struggle to capture. Meanwhile, crowd-sourced valuations, though popular, come with inherent biases and delays. They frequently do not match actual transfer prices (often undervaluing players), and their reliance on opinion means they can be slow to reflect rapid shifts in market conditions. Studies have identified additional weaknesses in crowdsourced methods, such as updates that are too infrequent, bias toward well-known players, and vulnerability to manipulation or lack of

expertise in the crowd. In short, both the traditional metrics-based approaches and the wisdom-of-crowds approaches leave a large portion of player value variation unexplained, and there is still no standard way to integrate diverse factors – athletic performance, player potential, team context, financial conditions – into one reliable valuation model.

This gap in valuation methodology creates significant uncertainty in the transfer market, leading to inefficient spending and missed opportunities for clubs. A team that cannot accurately gauge a player's value risks overpaying for talent or losing out on valuable players due to misvaluation. The present thesis confronts this problem by proposing a data-driven valuation framework that combines advanced performance metrics with machine learning techniques to improve the accuracy and robustness of transfer fee predictions. In doing so, it seeks to move beyond the limitations of current approaches and provide a more objective tool for player valuation.

## 1.3 Objectives and Research Questions

Given these challenges, the primary objective of this research is to develop a comprehensive player valuation model that leverages both traditional and modern performance metrics, together with state-of-the-art machine learning algorithms, to predict football transfer fees more accurately. To fulfill this objective, the study addresses several key research questions:

RQ1: Which player performance metrics – both traditional and advanced – are most strongly associated with transfer fees? (This will be investigated through exploratory data analysis to identify important features and gauge the strength of their relationship with recorded fees.)

RQ2: How can advanced metrics and novel feature engineering improve the predictive accuracy of valuation models compared to using traditional statistics alone? (We will compare model performance using a basic feature set versus an enriched feature set that includes engineered features and advanced metrics.)

RQ3: Do machine learning models (e.g. tree-based ensembles) outperform traditional regression models in predicting transfer fees? (We will train and evaluate both linear regression models and more complex machine learning models to determine which yields better prediction accuracy.)

RQ4: What is the impact of integrating domain-informed features (such as player age, league strength, or recent-performance weighting) and techniques like Principal Component Analysis on model performance? (We aim to quantify how much these tailored features and dimensionality-reduction steps improve the model's accuracy over baseline approaches.)

RQ5: How well does the best-performing model generalize to new, unseen transfer data, and what do the results imply for real-world usage? (We will test the final model on hold-out data to evaluate its performance on future transfers and discuss the practical implications, limitations, and potential uses of such a model for clubs and analysts.)

By addressing these questions, the thesis not only develops a methodological framework for data-driven player valuation but also provides practical insights into which factors truly drive transfer fees. The ultimate goal is to reduce uncertainty in the player transfer market by offering a more objective, data-backed valuation tool that stakeholders can use alongside traditional scouting and expert judgment.

# 2 LITERATURE REVIEW

## 2.1 Traditional Econometric Approaches in Player Valuation

Early research on football player valuation was grounded in traditional econometric approaches. Pioneering studies in the 1990s and 2000s modeled transfer fees using multiple regression analysis, typically with a handful of performance and demographic variables as predictors. Common factors included goals scored, assists, appearances, age, and sometimes position or international caps. Researchers like Carmichael and Thomas (1993) built regression models treating transfer fee as a function of such variables, finding that goal-scoring in particular had a significant positive impact on fees. These classical models assumed linear relationships and often relied on ordinary least squares (OLS) estimation (Majewski, 2016).

Studies consistently identified age, player usage (minutes or appearances), and decisiveness (goals and assists) as variables that positively influence footballer value assessment, with age and age squared being among the most frequently tested variables (Franceschi et al., 2023). The "core model" typically included key independent variables such as minutes/matches played, goals, assists, and age (Thrane, 2024). These traditional approaches provided an important starting point and established that on-field

performance does influence market value. However, they also had clear limitations: $R^2$ values were often modest, indicating that a large portion of variance in fees remained unexplained. Factors like a player's popularity (Franck & Nüesch, 2010) or a buying club's desperation were hard to quantify and thus omitted. Moreover, multicollinearity among performance stats (e.g. goals and assists are correlated) sometimes plagued these regressions, inflating uncertainty in coefficient estimates.

Despite these issues, traditional econometric studies made valuable contributions. They confirmed, for instance, that offensive output (goals, assists) and top-league experience tend to increase a player's market price, whereas older age can decrease it (all else equal). They also established baseline expectations: a 1-goal increase might correspond to a certain percentage rise in fee, and so on. Yet, as the transfer market grew more complex and data-rich, it became evident that these simple linear models were too limited. The underlying assumption of linearity and additivity (each factor contributing independently) could not capture interaction effects or diminishing returns (e.g., the 30th goal in a season might not add as much value as the 10th) (Li et al., 2022). Thus, while traditional models laid the groundwork, researchers began seeking more sophisticated, data-driven methods to better explain and predict transfer fees.

### 2.2 Transition to Data-Driven and Machine Learning Methods

By the early 2010s, football analytics was on the rise and the field of player valuation began to shift toward data-driven approaches. Researchers started to explore machine learning techniques that could handle larger, more complex datasets and nonlinear relationships (Rodríguez et al., 2019). Unlike the earlier econometric models, which were typically linear and constrained by small variable sets, machine learning models can accommodate dozens (or even hundreds) of features and automatically capture interactions and non-linear effects. For example, ridge and lasso regression (regularized linear models) were introduced to tackle multicollinearity and feature selection issues that plagued ordinary regression. These methods penalize overly complex models and have been used to identify which performance metrics truly matter when many are considered simultaneously.

Beyond enhanced regressions, scholars began experimenting with tree-based methods and other algorithms. Decision trees and ensemble methods (Random Forests, Gradient Boosting Machines, etc.) offered the ability to detect non-linear patterns -- for instance, a Random Forest might uncover that a combination of moderate goal tally and very young age can exponentially raise a player's value, a nuance a linear model might miss. Studies have demonstrated that nonlinear models are more suitable for modeling market

values with complex performance datasets. Aydemir et al. (2022) used ensemble learning (including Gradient Boosting and XGBoost) on a rich set of performance metrics, reporting extraordinary improvement in predictive accuracy compared to linear regression models and significantly outperforming state-of-the-art models and common benchmarks. These studies underscore an emerging consensus: machine learning can enhance valuation models by finding hidden patterns in the data that traditional models overlook.

This transition to data-driven methods also coincided with greater availability of advanced metrics and computing power. Research has shown that when advanced metrics are properly processed through techniques like Principal Component Analysis (PCA), they can provide additional predictive signal beyond traditional statistics (Behravan & Razavi, 2020). The literature from the past decade illustrates a paradigm shift from purely econometric approaches toward embracing machine learning for predictive tasks in football economics. However, these studies also highlight challenges such as the need for interpretability (complex ML models can be "black boxes") and the importance of carefully validating models to avoid overfitting to historical data (Scarfe et al., 2021).

### 2.3 Key Determinants of Market Value

Despite the introduction of new techniques, there is broad agreement in the literature on several key determinants of a player's market value. Both traditional econometric analyses and newer machine learning studies consistently identify certain factors as crucial. Performance metrics naturally top the list: goal-scoring (especially non-penalty goals) and assisting are fundamental drivers of value for attacking players (Franceschi et al., 2023). A forward who scores 20 goals in a top league will generally command a far higher fee than one who scores 5. Assists likewise add value, though typically not as much as goals. For defenders and goalkeepers, clean sheets and defensive actions (tackles, interceptions) come into play as value indicators, though these were harder to quantify in older studies.

Player age is another determinant highlighted by many authors. There is usually a premium on youth, reflecting potential and resale value (Barbuscak, 2018). Young players with strong performances can be seen as long-term investments, whereas older players (late 20s and beyond) may see diminishing transfer fees irrespective of talent, due to shorter remaining career spans. Age has been found to have a non-linear effect: fees rise through the early 20s (peaking around mid-20s for many positions) and then decline for players nearing 30, aligning with common intuition about peak age.

Experience and pedigree also factor in. Players from prestigious clubs or leagues, or those with international caps, often carry a "pedigree premium" (Felipe et al., 2020; Lucifora & Simmons, 2003). This reflects both a signaling effect ("if he's at Real Madrid, he must be good") and the reality of higher competition levels forging better players. The strength of the league and club for which a player plays has a significant impact, with being a player in one of the "big 5" European leagues or in a more successful club associated with higher value (Barbuscak, 2018). Different positions on the field also have different values, with attacking midfielders and forwards often cited as the most valuable positions (Felipe et al., 2020), (Frick, 2006). Research shows that team variables and player positions significantly influence the market value of professional male footballers in Europe (Felipe et al., 2020). Relatedly, the remaining length of a player's contract can influence fees (players nearing contract end may be cheaper), though this is a detail often discussed qualitatively in literature due to data limitations. (Frick, 2006)

Beyond these, the literature acknowledges contextual and external factors. For example, a buying club's financial status or urgency can inflate fees (a wealthy club in desperate need of a striker may overpay). Player popularity, often measured through media or internet searches, is increasingly recognized as a factor affecting value, especially in modern football (Müller et al., 2017), (Lucifora & Simmons, 2003), (Garcia del Barrio & Pujol, 2007). While such factors are harder to encode in models, they form part of the understanding of market value determinants. In summary, traditional studies and newer works alike concur on core factors: performance (especially goals/assists), age/potential, and experience level are primary drivers of market value. What machine learning methods add is the ability to examine these alongside dozens of additional metrics and to uncover how these determinants might combine in complex ways.

## 2.4 Impact of Machine Learning on Valuation Accuracy

The advent of machine learning (ML) techniques in player valuation has generally improved predictive accuracy by capturing non-linear relationships and interactions that earlier models missed. ML algorithms can handle high-dimensional data, meaning they can ingest a wide array of features -- from performance stats to biographical data -- and let the data reveal which combinations are most predictive. For instance, gradient boosting decision trees (like XGBoost) might discover that a moderate goal tally coupled with exceptional passing stats can sometimes indicate a playmaker's high value in ways that a linear model averaging their effects would not. Machine learning builds "a priori unknown model from the data" and for more complex methods like deep learning, large datasets are needed for training (Rein & Memmert, 2016).

One clear impact of ML has been the ability to rank and select features through techniques like feature importance scores or regularization paths. This helps address the question of which advanced metrics actually matter. Some studies reported, for example, that when using hundreds of features (basic and advanced metrics combined), only a subset significantly improve predictions - ML models can surface these. Behravan and Razavi (2020) implemented a hybrid ML approach using Support Vector Regression (SVR) with optimization algorithms and found that including too many correlated features could hurt performance, but methods like PCA or penalized regression helped distill the information, leading to better accuracy than a standard linear model with the same inputs.

In terms of quantitative gains, the literature shows moderate but meaningful improvements in metrics like $R^2$ or Mean Absolute Error when using ML. Research by Li et al. (2022) compared normal multiple linear regression models and Random Forest models for player valuation, noting that the Random Forest model fitting process "completely differs from traditional linear regression models" and can capture non-linear relationships that linear models assume away. While these figures are still relatively low in absolute terms (owing to the noisy nature of transfer fees), they represent significant steps forward in predictive power. ML models also tend to be more robust in out-of-sample testing when properly regularized and cross-validated, meaning they generalize better to unseen transfers. This is critical for practical use; a model is only useful if it maintains accuracy on future data, not just historical data.

However, the literature also cautions about the interpretability and usability of ML models in this domain. Complex models can be hard to interpret for decision-makers - club managers might prefer a simple explanation over a black-box prediction. Some research addresses this by extracting rules from ML models or by using techniques to explain predictions in terms of feature contributions. There is also the challenge of data availability: ML models hungry for data may be limited by the relatively small number of high-value transfers per year (a few hundred at most). Techniques like cross-validation with time-based splits have been emphasized to ensure models aren't overfitting on past eras that might not reflect current conditions.

### 2.5 Gaps and Future Directions

While research has advanced, large gaps remain in football player valuation modeling, presenting opportunities for future exploration. One major gap is the incorporation of qualitative factors and "soft" data. Almost all models to date rely solely on quantitative performance metrics and basic player info. Yet, we know that factors like leadership,

work ethic, or a player's commercial appeal (shirt sales, marketing value) can influence transfer fees. These are not easily captured in existing datasets. Future research might integrate proxies for such factors -- for instance, using fan engagement metrics or media sentiment analysis as features.

Another gap is the relatively short time horizon most studies cover. Many analyses are limited to a decade or less of data, often due to the availability of advanced stats only from mid-2010s onward. With each passing season, more data accumulate, which could allow models to better distinguish long-term trends from short-term fluctuations. There is a need to update models regularly and examine how the importance of certain features may be changing (for example, have metrics like distance run or pressing become more valued over time as play styles evolve?). Additionally, as women's football grows and its transfer market develops, similar models could be extended or adapted to that context, which has seen little study so far.

From a methodological perspective, striking a balance between model complexity and transparency is a key future direction. Some academic studies emphasize the need for interpretable models by excluding "black box methods or insufficiently detailed results" (Franceschi et al., 2023), highlighting the trade-off between accuracy and interpretability. Recent work suggests using ensemble techniques (combining multiple models) or multi-output models that predict not just a point estimate for value but a range or probability distribution. This could give clubs a sense of the uncertainty in a player's valuation. Integrating predictive modeling with decision analysis might also be valuable -- assessing whether a player is over- or under-valued relative to the model's prediction could signal potential bargains or overpriced targets.

In summary, future research directions include expanding feature sets beyond on-field performance, leveraging growing datasets for longitudinal analysis, improving model interpretability, and potentially moving beyond single-point predictions to more informative outputs. These directions will help close the remaining gap between academic modeling and the multifaceted reality of football transfer markets.

## 2.6 Contribution of This Thesis

Given the gaps identified above, this thesis aims to make several contributions to the field of player valuation. First, it introduces a unified framework that integrates both traditional statistics and advanced performance metrics into player valuation. Previous studies often focus on one or the other; our approach combines them, allowing us to

examine how much predictive power truly comes from modern metrics above and beyond the basics. This directly addresses calls in the literature for more comprehensive models that utilize the breadth of available data.

Second, we apply state-of-the-art machine learning techniques to this unified dataset, including regularized regression and ensemble tree-based models, and rigorously compare them to traditional regression benchmarks. By evaluating models on both an extensive historical dataset and a recent hold-out season, we provide clear evidence of the circumstances under which complex models yield improvements. In doing so, we also assess model stability over time - an aspect often missing from prior work that tends to evaluate on random train/test splits without considering temporal shifts.

Third, this work emphasizes feature engineering and domain-informed adjustments (such as time-weighting recent data and creating interaction terms for key stats). Many prior studies use off-the-shelf algorithms on raw data; in contrast, we devote effort to crafting features (e.g., goal contribution, various efficiency ratios, principal components for groups of correlated metrics) that encapsulate football knowledge. We demonstrate how these engineered features improve model performance and interpretability, bridging the gap between data science and domain expertise in football.

Finally, our thesis offers practical insights by validating the final model on real-world examples and outlining how it could be applied in a club context. We discuss not just which model performs best, but why? For instance, identifying which aspects of player performance the model deems most valuable. This contributes to the field by translating the model outcomes back into football terms. The overarching contribution is a blueprint for data-driven player valuation that is both more accurate and more interpretable, moving the conversation a step closer to adoption in actual scouting and transfer decision processes.

# 3 METHODOLOGY

In this section, the methodology for predicting football player transfer fees is presented. The approach integrates elements of traditional econometric analysis with modern machine learning techniques, organized into sequential stages ranging from data collection to final model evaluation. These stages include data collection (from multiple sources), dataset matching and integration, data cleaning and preprocessing, target variable definition, feature engineering and normalization, and the compilation of modeling datasets. Each stage is detailed below to illustrate how raw data were

transformed into a modeling-ready format, ensuring a robust and reliable basis for analysis.

### 3.1 Data Collection and Web Scraping

Data were collected from two primary sources: *Transfermarkt* and *FBref,* covering roughly two decades of football seasons (2004/05 through 2023/24). Transfermarkt was used to obtain historical transfer records, specifically by scraping the top 250 highest-value transfers for each season in that range. These data include player names, transfer fees, source and destination clubs, player age at transfer, and other deal details. Transfermarkt's static HTML pages facilitated extraction using Python's BeautifulSoup library for parsing content. In parallel, FBref was utilized for player performance statistics not provided by Transfermarkt. FBref offers comprehensive season-by-season performance data for players in major leagues, including traditional metrics (available for all seasons 2004–2015) and advanced metrics (available from 2015 onward). Accessing FBref's advanced stats required an automated browser approach due to dynamic content loading behind user login. We used Selenium (with a headless browser) to simulate a logged-in session and retrieve these stats. During scraping, additional challenges were encountered, such as content hidden behind interactive elements (e.g., certain club ranking data). These were resolved by programmatically adjusting browser settings (such as zoom level) to ensure all page elements were revealed for extraction. Through this process, we compiled three raw datasets: a list of high-value transfers from Transfermarkt (with fee and basic info), a dataset of traditional performance statistics from FBref, and a dataset of advanced performance statistics from FBref. All datasets were aligned to the 2004-2024 period for consistency.

### 3.2 Dataset Matching and Integration

After collecting data from multiple sources, the next step was to integrate these datasets into a cohesive analytical framework. Each transfer record from Transfermarkt was matched with the corresponding player's performance statistics from FBref. To facilitate this, all datasets were first normalized on key fields – notably season notation and player identity – to ensure consistency in merging. Season formats were standardized (e.g., converting all entries to a uniform "YYYY/YY" format). Player names served as the primary key for matching records, supplemented by season as a secondary key to handle cases where a player transferred multiple times. For each transfer occurring in a given season, the player's performance metrics from the preceding season were appended as features (for example, if a transfer took place in summer 2020, we use the 2019/20 season stats as the latest data). In instances where a

player's prior-season data was unavailable (due to injury or a missing record), we fell back to using data from two seasons prior to minimize information loss. This matching logic ensured that our model's features reflected the information that buying clubs would have had at the time of the transfer. Upon merging, we created two integrated datasets: one combining transfer records with traditional performance stats (available for the full 2004/05-2023/24 range), and a second combining transfer records with advanced stats (limited to 2015-2024, since advanced metrics were only recorded in recent years). The advanced-statistics dataset is smaller due to limited historical availability, but it provides a richer feature set for recent transfers. By the end of this integration step, each transfer entry was augmented with a comprehensive set of features describing the player's recent on-field performance, enabling a holistic analysis of factors influencing transfer fees.

### 3.3 Data Cleaning and Preprocessing

The merged datasets underwent thorough cleaning and preprocessing to ensure data quality and consistency before modeling. This process combined automated scripts with manual oversight for resolving issues. First, identifying information and formatting were standardized: the *Season* field was reformatted uniformly (e.g., "2012/13"), and the *Transfer Fee* field was stripped of non-numeric characters (such as currency symbols and textual notes) and converted to a numeric value. Entries corresponding to non-monetary transfers e.g., free transfers or loans (recorded as zero fee) were removed, since they do not represent a quantifiable market valuation and would distort the training target. Next, several high-cardinality or non-informative categorical fields were dropped or simplified. Specifically, player name was removed to prevent any individual-specific bias in modeling, and club names were not used directly because there are too many unique clubs (and club prestige is better captured by other variables like league or club ranking). Instead, broader indicators of club context were retained (detailed in feature encoding below). Key categorical attributes like position and nationality were retained due to their relevance in player valuation, but their levels were consolidated where appropriate. For example, nationality was grouped into the 15 most common countries in the data, with all other nationalities labeled "Other" to avoid an explosion of rare categories. This retains signal from a player's national background (e.g., being Brazilian or Argentinian might carry different weight) while reducing sparsity.

In handling numerical data, we paid special attention to extreme values and skewness in the Transfer Fee target. The dataset naturally includes some very large transfer fees (e.g., record-breaking deals above €200M) which can heavily skew the distribution. Unlike some approaches that might exclude such outliers, we kept all legitimate

transfers in the dataset to preserve the real-world range of outcomes. To mitigate the influence of extreme fees on model training, we applied a log-transformation to the transfer fee values. Converting fees to a logarithmic scale compresses the range of high values and helps stabilize variance, ensuring that a single extraordinary transfer (such as Neymar's €222M move) does not unduly dominate the model's error structure. Aside from the target variable, most performance feature columns were on comparable scales to begin with (many were counts or rates per 90 minutes). Nonetheless, we checked for any obvious data entry errors or implausible values in the performance stats (and corrected or removed those as necessary). By the end of the cleaning stage, the datasets had consistent, machine-readable formats with irrelevant fields removed, categorical values rationalized, and the target variable appropriately transformed for modeling.

### 3.4 Target Variable Selection

A crucial methodological choice was the selection of the target variable for prediction. Rather than using a subjective "market value" estimate, we used the actual transfer fee paid for each player as the prediction target. The transfer fee represents the real transaction price agreed between clubs, providing an objective ground truth for modeling. In contrast, market value estimates (such as those provided by Transfermarkt) are crowd-sourced and can be subjective or biased, plus historical market values are not comprehensively archived for all players. By focusing on actual fees, we leverage a concrete outcome measure that is consistent across seasons and less prone to speculative error. The trade-off is that using transfer fees limits our data to players who actually transferred for a fee (excluding free transfers and players who never moved during the period), thus reducing sample size. We deemed this acceptable because the quality and relevance of the target variable are paramount for model learning. Transfer fees capture the multitude of factors that converge in a real-world deal, serving as a robust indicator of a player's market valuation at a point in time. Training the model on actual fees ensures that its predictions align with what clubs have demonstrably been willing to pay, which is essential for practical utility.

### 3.5 Feature Encoding and Normalization

With the data cleaned and the target defined, the next step was encoding categorical features and normalizing feature scales in preparation for modeling. Several player and context attributes in the dataset were categorical in nature – for example, a player's position, primary nationality, and the league or tier of the club involved in the transfer. We transformed these using one-hot encoding to convert category labels into binary indicator features. Specifically, the *Position* feature (values like Goalkeeper, Defender,

Midfielder, Forward) was one-hot encoded into four binary columns (since these categories are mutually exclusive). The *Nationality* feature, after consolidation to top countries as described, was one-hot encoded into 16 categories (15 specific countries + "Other"). To represent the competitive context, we created a *League Tier* feature indicating the prestige of the league the player is coming from. Five major European leagues (England, Spain, Germany, Italy, France) were flagged with individual indicators (given their dominance in high-value transfers), and any other league was categorized as "Other," which was one-hot encoded accordingly. We also introduced flags for whether the buying or selling club is considered an "elite" club (based on a UEFA ranking threshold), to capture club prestige effects.

Continuous features were then scaled to ensure comparability. Because our feature set included variables with different units and ranges (for instance, goals range in the tens, while some advanced stats like distances or ratios might be fractions), we applied normalization. For most features, we used z-score standardization (subtracting the mean and dividing by the standard deviation) calculated on the training set. This centers features around 0 with unit variance, preventing those with larger scales from unduly influencing the model. A few ratio or percentage features that were naturally bounded between 0 and 1 were left as-is or min-max scaled to [0,1] if needed, since they were already on a comparable scale. Importantly, scaling parameters (means and standard deviations for standardization) were derived from the training data only and then applied to validation/test sets, to avoid any information leakage. By tailoring the encoding and scaling to each dataset's characteristics, we ensured that all input features would contribute proportionately during model training and that no single feature could dominate due to scale alone.

### 3.6 Final Dataset Composition and Tools

At the end of preprocessing, we finalized the modeling datasets. The Basic Stats dataset (traditional metrics) contains 1,751 transfer records with about 38 features per record, covering each player's last-season traditional performance stats (e.g., appearances, goals, assists, etc.), the player's age and other attributes, plus the one-hot encoded categorical variables for position, nationality, and league tier. The Advanced Stats dataset comprises 970 transfer records (2015–2024) with roughly 124 features each, reflecting a subset of players for whom advanced performance data (like expected goals (xG), expected assists (xA), key passes, pressures, progressive carries, etc.) are available. The advanced dataset includes all the core features of the basic set (age, basic stats, etc.) and augments them with dozens of additional metrics that provide a deeper quantitative profile of each player. By design, the advanced dataset has fewer records (since advanced tracking began later), but it enables us to analyze whether including

these rich features improves predictive accuracy. In both datasets, the transfer fee (our target) was log-transformed as described.

Throughout the project, we utilized a range of software tools and libraries. Data scraping was accomplished with Python; BeautifulSoup handled static HTML parsing for Transfermarkt, while Selenium (with a Chrome WebDriver in headless mode) handled dynamic content and login for FBref. Data manipulation and integration were done using pandas and NumPy, providing efficient structures for cleaning and merging large datasets. For modeling, we relied on scikit-learn for our regression models and preprocessing pipeline (handling train/test splits, cross-validation, one-hot encoding, scaling, etc.), and XGBoost for implementing gradient boosting trees (given its strong performance in regression tasks). We also used statsmodels and SciPy for statistical testing and to validate certain assumptions (for example, checking normality of residuals or multicollinearity via Variance Inflation Factors). This combination of tools provided a solid foundation for executing the methodology end-to-end, from initial data acquisition to final model development. The end result of these collective efforts was a clean, enriched dataset ready for exploratory analysis and model training, which we describe in the subsequent chapters.

# 4 EXPLORATORY DATA ANALYSIS (EDA)

Before building models, we conducted an exploratory data analysis to understand the basic characteristics, patterns, and potential anomalies in the dataset. EDA serves to highlight important relationships and guide the data preprocessing and feature engineering decisions. In our analysis, we focused on summarizing key variables, visualizing distributions, examining correlations, and identifying any outliers or biases that might affect modeling.

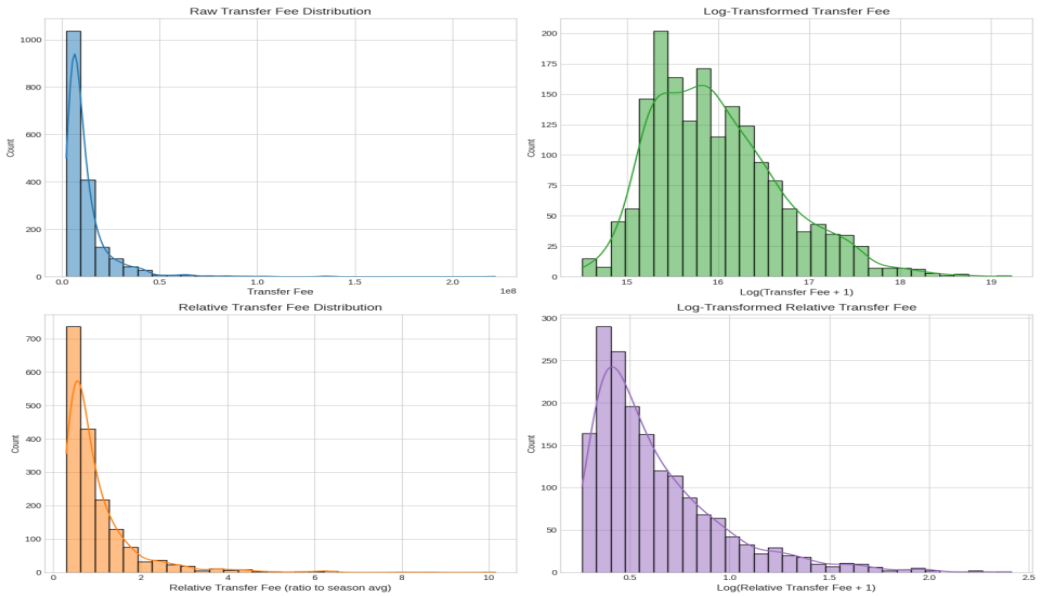
## 4.1 Descriptive Statistics and Transfer Fee Distribution

The initial step of EDA was to summarize the central tendencies and variability of important variables such as transfer fees and performance metrics. For the Basic Stats dataset (covering 2004–2014 transfers), the mean normalized transfer fee was around €8.5 million with a median of about €5 million, reflecting a right-skewed distribution (a few extremely high fees driving the mean above the median). In the Advanced Stats dataset (2015–2023 transfers), the average fee was higher – approximately €15 million median – which is expected given football inflation over time and the fact that the advanced data covers more recent big-money transfers. Standard deviations were large

in both cases (on the order of tens of millions), underscoring the high variance in transfer fees.

To better understand the distribution of transfer fees, we plotted histograms of the fees under different transformations. **Figure 1** illustrates the distribution of transfer fees in our data under three conditions: the raw fee values, the log-transformed fees, and fees normalized by season (to account for inflation over time). The raw fee distribution (left panel) is heavily right-skewed – most transfers occur at relatively low fees, with a long tail extending to the record-breaking transfers beyond €100M. After log-transforming the fees (center panel), the distribution becomes much closer to symmetric (approximately normal-shaped), indicating that the log scale is more appropriate for modeling and analysis. The season-normalized fees (right panel), where each fee is adjusted relative to the average fee in that season, still show a skew but somewhat less extreme, as this normalization reduces the inflationary growth in fees over the two-decade span. Together, these plots confirmed that using a log scale for the target would be beneficial and that adjusting for time trends was necessary to compare fees across eras, consistent with methodological approaches established in early transfer market research (Gerrard et al., 2000)

*Figure 1:Transfer Fee Distribution.*




*Source: Own work*

Beyond transfer fees, we also examined distributions of key performance metrics. Basic statistics like goals, assists, and appearances typically showed skewed distributions as

well (for instance, most players have 0–5 goals, with a small number of prolific scorers as outliers). We calculated summary stats for these: in the basic dataset, players averaged about 5 goals and 3 assists in the season prior to transfer (with medians of 2 and 1 respectively, again reflecting skewness and the presence of a few high performers). Defensive metrics like tackles or interceptions had more symmetric distributions, but varied greatly by position (defenders naturally accumulating more). We stratified some analyses by position to ensure we weren't misled by, say, defenders' low goal counts.

To detect any data quality issues, we also checked for outliers and missing values. A few notable outliers included one transfer with an erroneously recorded fee of "€999M" (which was likely a placeholder or error – we had already removed such anomalies in preprocessing). In performance data, we saw a small number of cases where a player had zero minutes played in the prior season (often these were young players or those who were injured; we retained them if they transferred for a fee, as it might indicate they were valued for potential or past performance). We also verified that our log normalization of fees did not produce any anomalies (e.g., no infinite values since all fees were positive and we had removed true zeros).

Overall, the descriptive analysis confirmed expectations: transfer fees have grown over time and are very unevenly distributed, and players' performance stats likewise vary widely with a few standout values. These insights reinforced our choice of using log-fees for modeling and suggested the importance of features like era (season) and position when interpreting any statistical patterns.

### 4.2 Correlation Analysis of Performance Metrics

To understand how player performance indicators relate to transfer fees, we conducted a correlation analysis on both the traditional and advanced metrics. We computed Pearson correlation coefficients between each feature and the (normalized) transfer fee, for both datasets. As expected, in the Basic Stats dataset the features most strongly associated with transfer fee were goals and non-penalty goals (G–PK), followed by assists. These had correlation coefficients in the range of roughly 0.30-0.34 with normalized fee - indicating a moderate positive relationship. In practical terms, players who scored or assisted more tended to command higher fees. Playing time (minutes, appearances) also showed a positive correlation but weaker (around r ~0.15-0.20), suggesting that being a regular player is helpful but not as influential as being a high-impact player. Interestingly, age had a slight inverted U-shape correlation: very young players and older players had lower fees on average, whereas players in the 22-26 age range (early

prime years) had the highest fees, aligning with the idea of paying for both current performance and future potential.

In the Advanced Stats dataset, we looked at an expanded set of performance metrics (like shot-creating actions, goal-creating actions, various passing and carrying stats, defensive pressures, etc.). The correlation analysis revealed that several advanced metrics offered signals comparable to the traditional metrics. For example, shot-creating actions (SCA) and goal-creating actions (GCA) both showed moderate positive correlations with transfer fee (roughly $r \approx 0.30$). These metrics capture nuanced playmaking contributions – a player who consistently creates chances tends to be valued. Similarly, progressive passing and carrying metrics (such as progressive passes and carries into the penalty area) had correlations in the 0.25-0.30 range, underscoring that modern scouting values players who can advance the ball effectively. Notably, even though no single advanced metric had an extremely high correlation with fee, the set of advanced metrics as a whole collectively provided additional explanatory value beyond the basics. This justifies integrating them into predictive models, as they measure facets of performance (like creativity, buildup play, defensive actions) not captured by goals or assists.

All of these individual correlations were statistically significant ($p < 0.001$) given the sample sizes, but their magnitudes (generally in the 0.2–0.3 range for top features) indicate that any single metric explains only a fraction of transfer fee variance. This reinforces a key point: clubs evaluate players holistically, so a combination of metrics will be needed to predict fees accurately. It also hints at multicollinearity - many of these metrics are interrelated (for instance, players with high GCAs often also have high SCAs and key passes). We needed to keep this in mind for modeling; features would likely need to be combined or reduced (e.g., via PCA) to avoid redundancy.

We also examined correlations *among* the features themselves to identify clusters of highly interrelated metrics. **Figure 2** provides a visualization of the top features (from the Basic Stats dataset) by absolute correlation with transfer fee, and **Figure 3** does the same for the Advanced Stats dataset. In Figure 2 (basic features), we see that goals, G-PK, assists, and minutes are the top correlates with fee, all positively correlated. Some features have negative correlations- for example, the ranking of the club a player is coming from (a lower numerical rank for a club in UEFA tends to correlate with higher fees, hence a negative correlation coefficient when using rank numbers). In Figure 3 (advanced features), metrics like GCA, SCA, progressive attack stats, and technical skill metrics stand out as most correlated with fee, while metrics representing less directly value-related actions (like certain defensive or possession metrics) show weaker correlations or even slight negatives. Both figures underscore that even the most

correlated feature has an r of only about 0.3-0.4, emphasizing the need for multivariate models.

*Figure 2: Top features from the Basic Stats dataset correlated with transfer fees.*

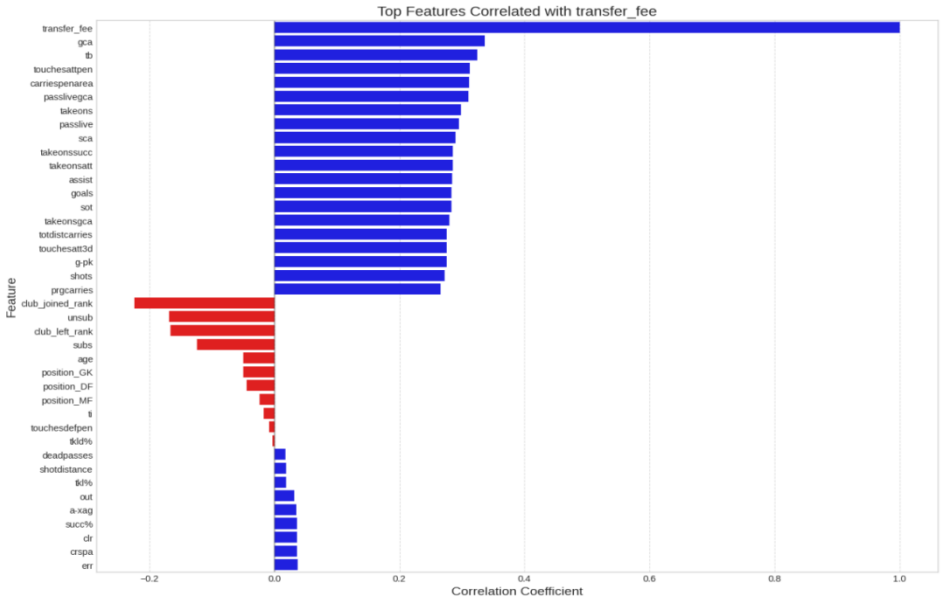
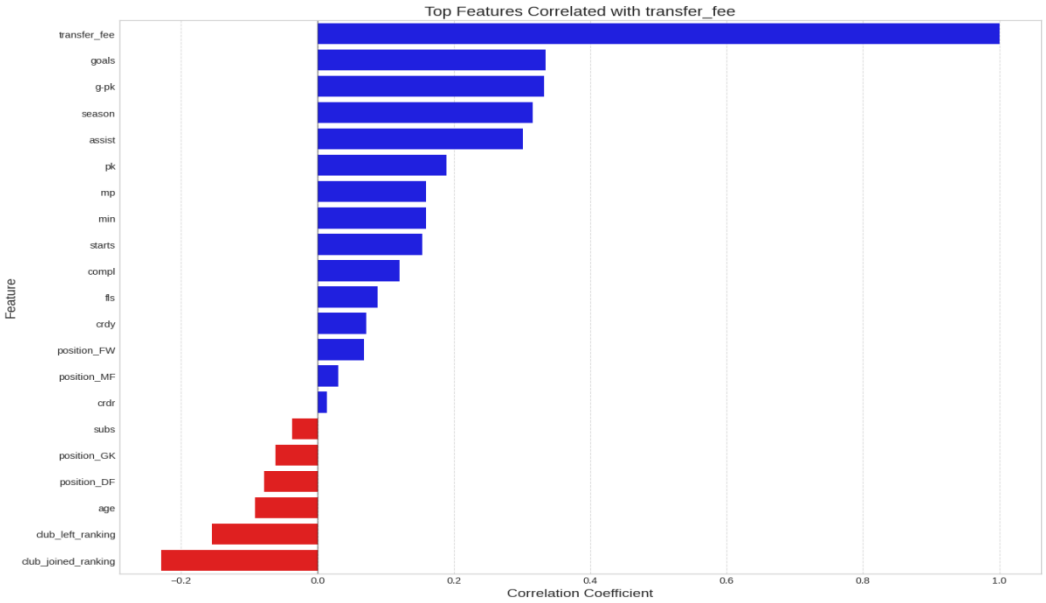


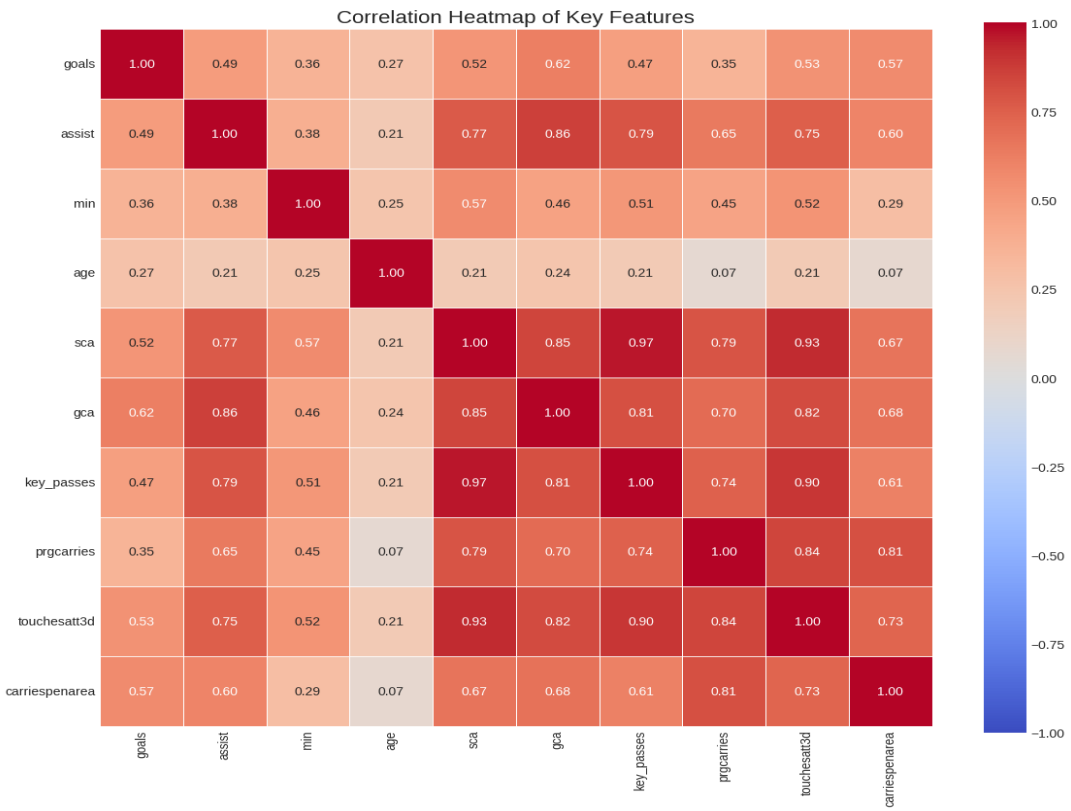

*Figure 3:Top advanced metrics correlated with transfer fees.*




*Source: Own work*

In addition to these bar charts, we plotted a correlation heatmap for a subset of **key features** to visualize interrelationships (Figure 4). This heatmap included both traditional stats (goals, assists, minutes, age) and a few advanced metrics (SCA, GCA, key passes, progressive carries, etc.) that we suspected would be important. The heatmap revealed that many advanced metrics are highly inter-correlated with each other.For instance, in our data SCA and key passes have a correlation above 0.9, and GCA is also very strongly correlated with SCA (since one is a subset of the other). This confirmed that multicollinearity would be a challenge, especially within the advanced feature set. It also showed that traditional and advanced metrics are not collinear with each other (goals have only a moderate correlation with SCA, etc.), which suggests that advanced metrics are contributing new information rather than just duplicating what goals/assists already tell us. Such findings reinforced our plan to use dimensionality reduction (PCA) on the advanced metrics to distill the information into a smaller set of uncorrelated components for modeling.

*Figure 4:Correlation heatmap of key features.*




*Source: Own work*

In summary, the correlation analysis confirmed intuitive value drivers (scoring and creative metrics are important) and demonstrated the added value of advanced stats

while also warning of their multicollinearity. These insights guided the next steps of our process, leading us to apply techniques like PCA to handle correlated advanced features and to engineer summary metrics (like goal contribution) that combine multiple basic stats into one.

### 4.3 Feature-Target Relationships and EDA Conclusions

Beyond simple correlations, we explored feature-target relationships using scatter plots and segmentation to glean deeper insights. For instance, plotting transfer fee against age revealed that the highest fees in our dataset were often for players aged ~23-25. Younger teenagers, despite hype, usually transferred for less (with some exceptions in recent years), and players older than 30 had sharply lower fees. This aligns with the notion of an optimal age range for peak value. We also plotted transfer fee vs. goals, which showed a positive trend but with considerable scatter: some players with relatively few goals still commanded high fees. These tended to be defenders or midfielders, or cases of young prospects where potential was valued over current output. Such observations underscored that bivariate relationships are informative but not sufficient – interactions and context matter. For example, 10 goals from a winger in the Bundesliga might translate to a higher fee than 10 goals from a striker in a lower-tier league.

We used box plots to examine positional differences. Attackers (forwards) had the highest median transfer fees in our data, followed by midfielders, then defenders, and finally goalkeepers. This is expected, as attackers' contributions (goals, assists) are easier to quantify and often more glamorized, driving up prices. However, we also noted an increasing trend in fees for elite center-backs and goalkeepers in the late 2010s (e.g., record fees for defenders like Virgil van Dijk, or goalkeepers like Alisson). This suggests a shifting appreciation for defensive roles, something we kept in mind when interpreting model outcomes for those positions.

By the end of EDA, we had also identified several data challenges to address. The right-skewed distribution of fees meant our use of log fees was crucial (we confirmed that log-fees approximately met normality assumptions for regression, with residuals more homoscedastic after log transform). The presence of non-linear relationships (such as the age effect and diminishing returns on goals) suggested that introducing polynomial terms or interaction features (like we did with goal contribution, and will do with age squared perhaps) could be beneficial. And the multicollinearity among features indicated we should carefully reduce feature dimensions (via PCA or by dropping redundant metrics) to ensure stable model estimation.

In conclusion, the exploratory analysis provided a solid grounding. It verified that our intuition about what drives value was supported by data (performance matters but needs context), it highlighted the necessity of the data transformations and feature engineering steps we planned, and it gave us an early sense of which models might perform well (for instance, we expected tree-based models to potentially capture some of the non-linear trends we observed, whereas linear models would rely on us to explicitly include polynomial terms). Armed with these insights, we proceeded to the data preparation and feature engineering phase, as detailed in the next chapter.

# 5 DATA PREPARATION AND FEATURE ENGINEERING

Based on the EDA findings, we applied a series of data preprocessing steps and feature engineering techniques to improve model performance. The goals were to reduce skewness in the target, capture non-linear effects, and handle multicollinearity, all while incorporating domain knowledge about football. We describe these steps in the following subsections.

## 5.1 Feature Engineering of Performance Metrics

To enrich the feature set beyond raw statistics, we engineered new features that reflect rates, ratios, and combined contributions, guided by football domain knowledge:

**Goal Contribution:** We created a combined metric *goals + assists* to capture a player's total direct involvement in scoring. This "goal contribution" provides a fuller picture of offensive impact. In the Basic Stats data, goal contribution indeed turned out to be a better predictor of transfer fee than goals or assists alone. In fact, it had the highest univariate importance score in an ANOVA F-test among basic features, outperforming goals or assists individually in predicting transfer fee. This confirms that clubs value players who either score or set up goals (or both) – a combined impact.

Per-90 Metrics: Because playing time varies widely between players, we normalized certain counting stats by minutes played. For example, *goals_per_90* (goals scored per 90 minutes) measures scoring efficiency. A player who scores 5 goals in 1000 minutes (0.45 goals/90) might be more valuable than one who scores 5 goals in 3000 minutes (0.15 goals/90), all else equal. We computed analogous metrics like assists per 90. These rate stats help compare players on an equal footing (per full game) and can highlight impactful players who might have had limited game time.

Conversion Rates: We introduced metrics to quantify how efficiently a player converts opportunities. For instance, *shot conversion rate* = goals/shots, and *assist conversion rate* = assists divided by some measure of chances created (we used key passes or shot-creating actions as the denominator). These ratios reflect finishing or playmaking efficiency – qualities that can significantly boost a player's market value if high (a striker who converts a high percentage of chances might be seen as an elite finisher, and a midfielder who turns many key passes into assists might be an exceptional playmaker). However, we also noted that conversion rates can be volatile year-to-year (subject to luck), so they were used with caution.

Advanced Ratios**:** For players in the Advanced Stats dataset, we engineered features capturing quality over quantity. For example, *key_pass_ratio* – the fraction of a player's passes that are key passes – and *carry_to_pen_area_ratio* – the fraction of total carries that successfully enter the penalty area. The idea is to measure impact: not all actions are equal, and players who make fewer overall passes but a higher proportion of incisive passes might be especially valuable. Similarly, we looked at ratios like successful dribbles per attempt (dribble success rate) and tackles won per duel. These efficiency metrics align with the notion that it's not just volume of actions, but their effectiveness, that can define a player's value.

Polynomial Terms: EDA indicated possible non-linear effects (e.g., diminishing returns to certain stats). To allow models to capture these, we added squared terms for a few key features. For instance, we included $sca^2$ (shot-creating actions squared) and $carries_pen_area^2$ (carries into the penalty area squared) to let the model account for non-linear growth – perhaps the incremental value of each additional chance created starts to plateau beyond a certain point. Polynomial regression "allows us to capture complex phenomena like diminishing returns, saturation effects, or exponential growth" (How.dev,n.d.). By including these quadratic terms, the linear models (like OLS or Ridge) could fit curvilinear trends that we observed (e.g., a player with 10 SCA vs one with 5 SCA might add more value than the difference between 60 and 55 SCA, implying a concave relationship).

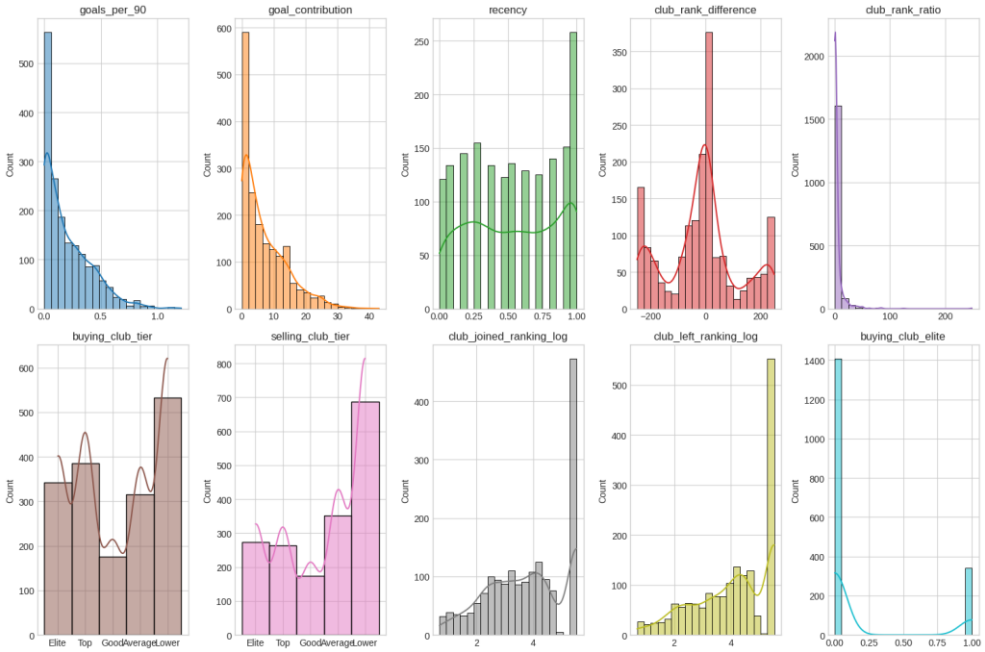
Interaction Terms: We also created interaction features to capture synergies between different aspects of performance. For example, we multiplied certain pairs of metrics that we hypothesized could have a combined effect greater than the sum of parts. One such feature was *take-ons* × *SCA* (does a player who dribbles past opponents frequently also create many shots? If so, their combination could especially drive value). Another was *GCA* × *SCA* (goal contributions in build-up and shot creation together). The rationale is that a player excelling in multiple offensive dimensions (like dribbling and chance creation) might have an outsized impact on value beyond what each stat alone

suggests. This approach aligns with recent advances in soccer analytics, where researchers have found that combining advanced statistics like expected goals (xG) and expected assists (xA) provides more comprehensive insights into player performance than analyzing these metrics individually (Malikov & Kim, 2024). Following this principle of interactive effects, we chose interactions based on domain intuition – for example, that a player's high xG and high xA together make them uniquely valuable (a dual threat scorer/creator).

All engineered features were created for each player-season entry prior to modeling. We took care to create analogous features in both datasets where possible (for fairness in comparison). For instance, goal contribution exists in both, as do per-90 stats; but something like key_pass_ratio is only meaningful for the advanced set where key passes are recorded.

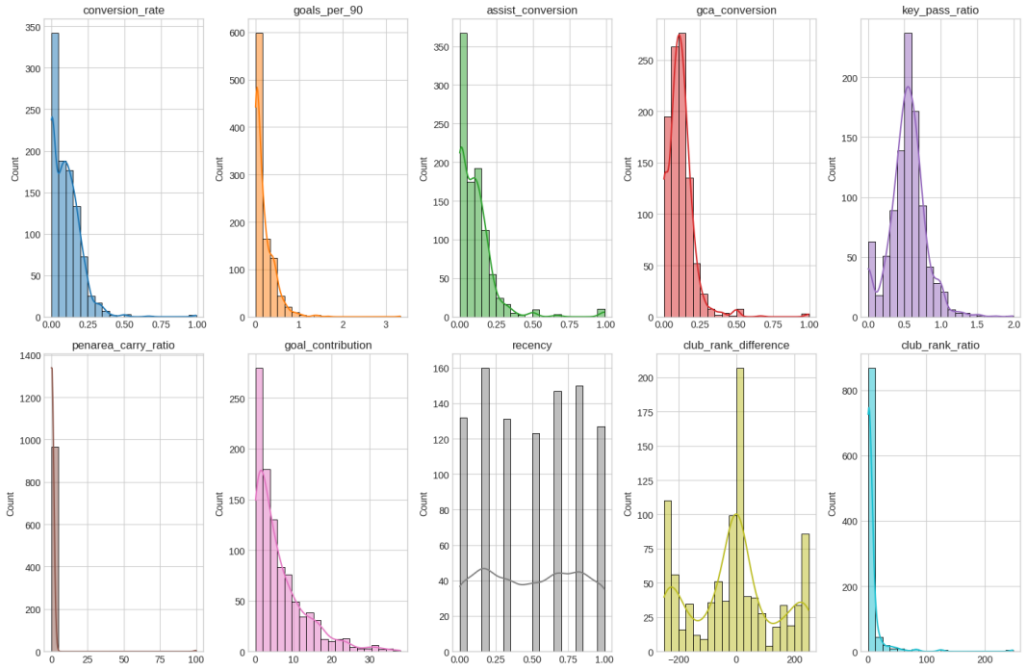
**Figures 5 and 6** illustrate the distribution of several engineered features in the Basic and Advanced Stats datasets, respectively, confirming that many of these features are highly skewed and may need transformations as well. In Figure 4 (Basic Stats features), we see distributions for examples like *goals_per_90*, *goal_contribution*, *recency* (a time-weighting factor we added, discussed later), and club strength indicators. Most distributions are right-skewed (e.g., goals_per_90 has many players around 0–0.3 and a long tail of prolific scorers approaching 1.0; goal_contribution has a bulk at low values and a tail of star players above 20). The club-related features like *club_rank_difference* (difference in ranking between selling and buying club) show a different pattern – a somewhat symmetric distribution centered near zero, since many transfers occur between relatively similar-level clubs, but with notable spikes (some players moving from far smaller clubs to top clubs, indicated by large positive rank differences). In Figure 6 (Advanced Stats features), we plot distributions for advanced efficiency metrics like *conversion_rate*, *assist_conversion*, *gca_conversion* (goal-creating action conversion), and others like *key_pass_ratio*. These too are heavily skewed towards 0 for most players (e.g., conversion rates cluster at low values for most players, with only a few having very high conversion percentages). The takeaway is that many engineered features, especially those defined as ratios or rates, ended up skewed and bounded. We applied logit transforms to certain ratio features (like conversion rates) when appropriate to normalize them, and we were prepared to use non-linear models that can inherently handle these distributions if needed.

*Figure 5:Distributions of selected engineered features (Basic Stats dataset).*




*Source: Own work*

*Figure 6:Distributions of selected engineered features (Advanced Stats dataset).*




*Source: Own work*

By engineering these features, we aimed to present the models with more informative inputs that encapsulate key concepts (overall contribution, efficiency, non-linear effects) without making the models figure them out from raw data alone. This approach, grounded in domain knowledge, was intended to improve model learning and also help interpret the results (coefficients or importances on these features have clearer meanings).

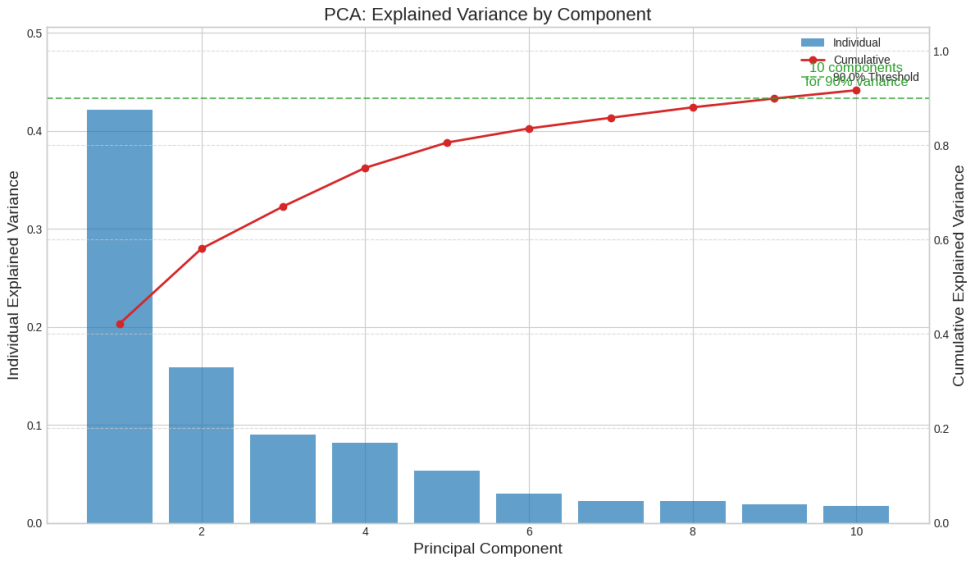
### 5.2 Handling Multicollinearity: VIF Filtering and PCA

With dozens of features (especially after our feature engineering), multicollinearity became a concern. Highly correlated features can destabilize linear model coefficients and make it harder to discern each feature's unique contribution. We addressed this in two ways:

Variance Inflation Factor (VIF) Filtering: We calculated VIF scores for the features in the Basic Stats dataset to identify extreme multicollinearity. As expected, features like goals and assists (which compose goal contribution) exhibited high VIF when all were included, as did minutes and appearances (which are almost redundant). We applied a threshold (VIF > 10) to flag problematic cases and then made principled reductions: for instance, we retained goal contribution but not raw goals and assists in the same linear model, and we dropped either minutes or appearances since they convey overlapping information. Similarly, among club strength indicators, we kept only one of rank difference vs. rank ratio to avoid duplication. This filtering process removed or consolidated features that were essentially linear combinations of others.

Principal Component Analysis (PCA): For the Advanced Stats dataset, which included a large set of correlated performance metrics, we applied PCA to create a smaller number of uncorrelated components. PCA allowed us to capture most of the variance in the advanced metrics with fewer composite features, thus alleviating multicollinearity and the "curse of dimensionality." We decided to retain enough principal components to explain ~90% of the variance in the advanced metric subset. As shown in **Figure 7**, the first few principal components account for the majority of variance (the green dashed line marks the 90% cumulative variance threshold). Specifically, the first principal component (PC1) explains about 42% of variance, the second (PC2) around 16%, and by the time we include up to PC10, we reach roughly 90% of total variance captured. We thus kept 10 principal components (PC1–PC10) as new features representing the advanced stat space. Using 10 PCs (out of 100+ original advanced features) preserved most information (~90% of variance) while eliminating multicollinearity (by construction, PCs are orthogonal).

*Figure 7:PCA explained variance by component for advanced performance metrics.*




*Source: Own work*

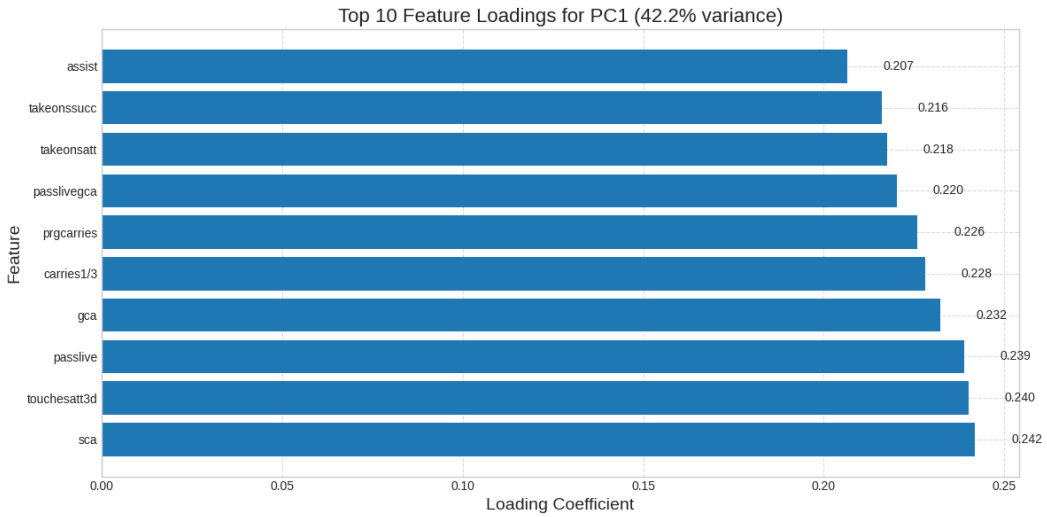
Importantly, while PCA helped reduce the feature count, we still wanted to interpret what these principal components represent in football terms. We examined the component loadings – the correlation of each original feature with the principal components – to understand their meaning. PC1, which explains 42% of variance, had high positive loadings on virtually all the advanced offensive metrics. Essentially, PC1 turned out to be an "overall offensive involvement" factor – players who were high on PC1 were those who had large numbers of touches, passes, carries, and shot or goal contributions. It was like a general activity/usage metric (one might even call it a "volume of play" component). **Figure 8** shows the top 10 feature loadings for PC1. We see that metrics such as shot-creating actions (SCA), goal-creating actions (GCA), touches in the attacking third, progressive passes, progressive carries, key passes, and even assists all load heavily (positively) on PC1. This indicates that PC1 is essentially measuring a single underlying trait: how much a player is involved in creating offensive plays. A player with a high PC1 score is heavily involved in attack – they frequently touch the ball in attacking areas and generate chances – which understandably correlates with higher value.

PC2 (explaining ~16% of variance) showed a contrasting pattern between finishing and playmaking. It loaded negatively on scoring metrics (goals, shots on target) and positively on passing metrics (especially completed passes, short pass attempts, and

possession metrics). **Figure 9** illustrates the top loadings for PC2, where we see "Completed Passes" and "Short Pass Attempts" on the positive side, and "Goals" and "Shots on Target" on the negative side (among others). We interpret PC2 as a "playmaker vs. scorer" axis. A player with a high PC2 score is one who contributes through build-up and distribution (lots of passes, perhaps a midfielder dictating play) but doesn't score much, whereas a player with a low (negative) PC2 score is more of a pure finisher (many goals and shots, fewer passes – think of a poacher striker). This component is valuable because it differentiates players by style: two players might both be high on PC1 (very involved in offense), but one could be a high-PC2 midfield orchestrator and the other a low-PC2 scoring forward. Interestingly, such distinctions can influence value depending on what a buying club needs.

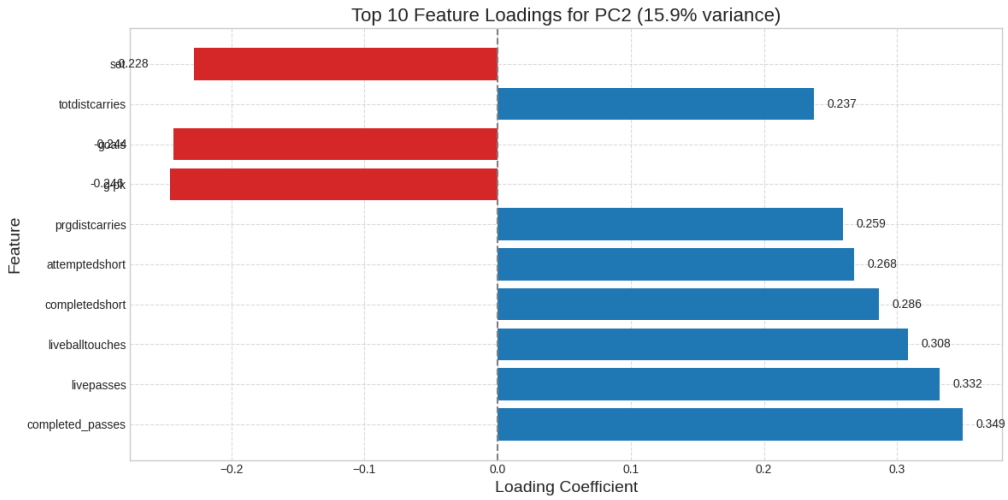
We repeated this interpretative exercise for further PCs. PC3 (not visualized here for brevity) had loadings suggesting it captured defensive actions vs. attacking actions (defenders vs. attackers separation), and so on. By examining these loadings, we ensured that the PCA process, while abstracting the features, still produced components with plausible football interpretations. These PCs were then used as input features for modeling (alongside other non-PCA features like age, position dummies, and a few aggregate stats that we kept). We acknowledge that one limitation of PCA is that the resulting components are not as directly interpretable as the original metrics (Jolliffe & Cadima, 2016), but by inspecting loadings we can at least qualitatively describe what they represent. The benefits of reduced multicollinearity and noise, as well as dimensionality reduction, outweighed this cost and indeed, we will see improved model stability and performance as a result.

*Figure 8:Top 10 feature loadings for PC1*




*Source: Own work*

*Figure 9:Top 10 feature loadings for PC2*




*Source: Own work*

By reducing our feature set via VIF filtering and PCA, we arrived at a streamlined list of inputs for each model: the Basic Stats models used a curated set of traditional and engineered features with minimal redundancy, and the Advanced Stats models used the 10 PCA components (representing advanced metrics) along with the traditional features and engineered features (like age, position, goal contribution, etc.). This approach not only mitigated multicollinearity but also likely improved generalization, as it removed noise and less relevant information from the modeling process.

## 5.3 Temporal Weighting of Training Data

An innovative step in our preprocessing was to apply time-based weighting to the training data. Football transfer markets evolve over time,because a €20M transfer in 2005 might correspond to a €50M transfer in 2020 due to inflation in fees and changing market dynamics. To ensure our model remained relevant to the latest market conditions, we implemented a scheme of temporal weighting: more recent transfers in the training set were given higher weight during model training, while older transfers were down-weighted. Concretely, we assigned weights to each training example that decay backward in time. For the Basic Stats dataset (2004-2014 used for training, with 2015 as test), we gave transfers from 2014 a higher weight than those from 2005, on a

linear scale. For the Advanced Stats dataset (2017-2022 training, 2023 test), we did similarly (2022 transfers weighted most, 2017 least).

This approach has the effect of biasing the model to fit recent data more closely, under the assumption that recent patterns are more indicative of the near future (which we care about for prediction) than decade-old patterns. We chose a moderate weighting scheme (the most recent season was weighted about 1.5x the earliest season in training) to avoid completely ignoring older data while still reflecting inflation and trends. The weights were also used in cross-validation folds to maintain consistency.

Additionally, we incorporated a "recency" feature as mentioned earlier, which captures how old the performance data is relative to the transfer. A player's performance metrics from one year ago might be more predictive than those from three years ago. By including recency both as a feature and by weighting, we addressed the issue of time from two angles.

### 5.4 Summary of Prepared Data

After all these preprocessing and feature engineering steps, we had two finalized datasets ready for modeling:

Basic Dataset (2004-2015 transfers): ~38 features (post one-hot encoding and feature engineering) per player-season, including goal contribution, per-90 stats, conversion rates, polynomial terms, interaction terms, and context features (age, position dummies, nationality dummies, league tier, club elite flags). The transfer fee target is log-transformed and the data is weighted by recency for model fitting. There are 1,751 training examples (with 2015 held out).

Advanced Dataset (2015-2024 transfers): ~40 features including the 10 PCA components for advanced metrics, plus all the features above that are applicable (age, position, etc., and a few basic performance features like goal contribution to complement the PCA components). The target is log-fee as well. We have 970 training examples (2017–2022) and we will test on 2023.

Both datasets were normalized and cleaned as described. We also set aside a portion of data for validation during model development – ensuring that our feature engineering

did not inadvertently leak information (for example, PCA was fit on the training data only within each CV fold to avoid lookahead bias).

With data preparation complete, we moved to the modeling phase, confident that our input data were as informative and non-redundant as possible. The next chapters detail the modeling approaches (regression and machine learning models), their performance, and the insights gained.

# 6 REGRESSION MODELS

Having prepared the data, we first employed traditional regression techniques to model transfer fees. Our regression models included Ordinary Least Squares (OLS) as a baseline, as well as regularized regressions, specifically Ridge, Lasso, and Elastic Net to handle multicollinearity and potentially improve prediction accuracy. We trained and evaluated these models on both the Basic Stats dataset and the Advanced Stats dataset (with PCA components), allowing us to compare how much predictive power is gained by using advanced metrics.

## 6.1 Model Training on Basic vs. Advanced Feature Sets

We split the data into training and testing sets chronologically, as described earlier (e.g., training on 2004-2014 and testing on 2015 for basic data; training on 2017-2022 and testing on 2023 for advanced data). All regression models were first evaluated via cross-validation on their respective training sets to tune hyperparameters (like the regularization strength for Ridge/Lasso). We used 5-fold cross-validation with time-based splits where applicable (ensuring that each fold's validation data came from later seasons than its training data, to mimic temporal prediction).

In terms of performance metrics, we focused on $R^2$ (coefficient of determination), RMSE (root mean squared error) in the log-fee space, and mean absolute percentage error (MAPE) when back-transformed to original fee units for interpretability.

On the Basic Stats dataset, the best performing regression model in cross-validation was the Lasso regression (with an optimal regularization parameter $\alpha$ selected via CV). It achieved an $R^2$ of around 0.17 on the validation folds meaning it explained about 17% of the variance in log-transformed fees. The Ridge and Elastic Net models were slightly behind ($R^2$ ~0.15–0.16), and OLS (with no regularization) actually performed worst ($R^2$

~0.12), likely due to overfitting on some less relevant features and multicollinearity issues. The advantage of Lasso was that it did feature selection by driving many coefficients to zero; indeed, in the final Basic Lasso model, only about 10 of the ~38 features had non-zero coefficients. Those included goal contribution, age, a couple of position dummies (with Forward being a strong positive indicator, meaning being a forward added to predicted fee), and league tier (coming from a top-5 league had a positive effect). This was reassuring: the model automatically picked out intuitive features and ignored redundant ones.

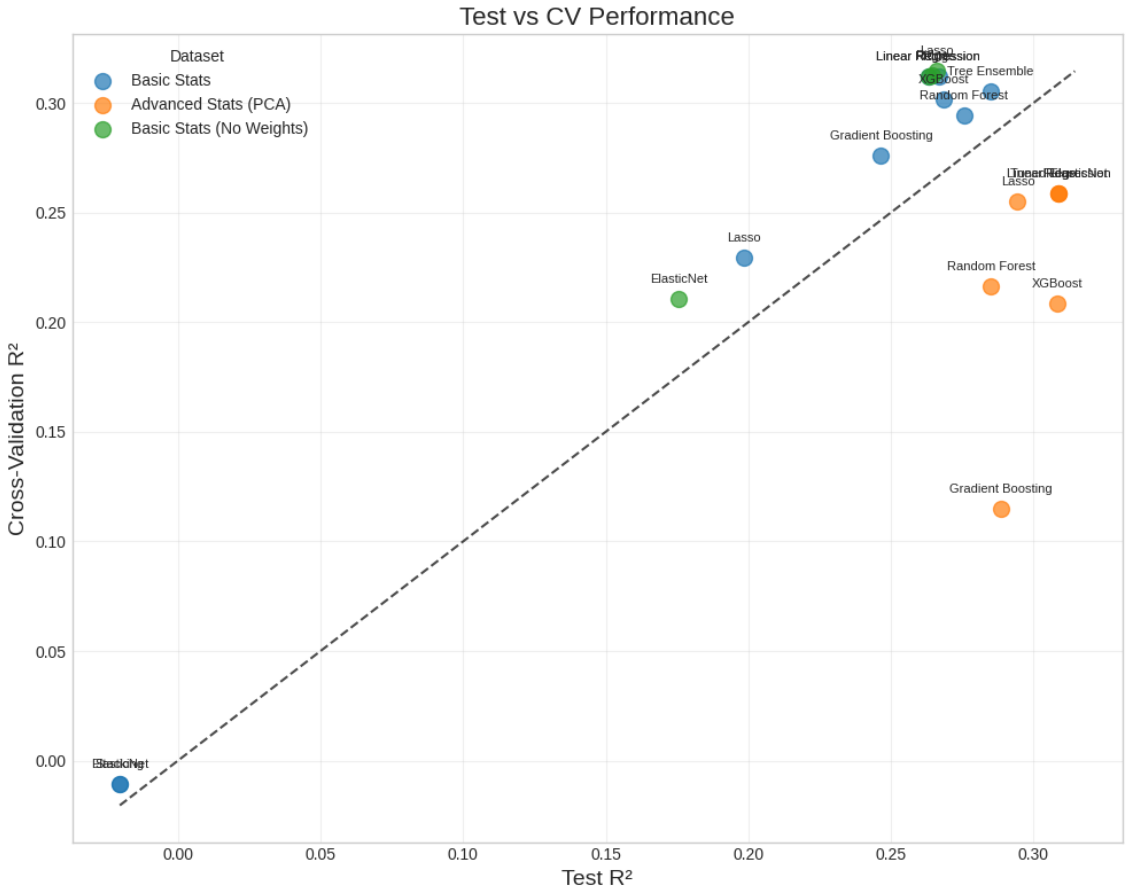
Moving to the Advanced Stats dataset, we expected higher $R^2$ given the richer data. Initially, a Ridge regression without time weighting achieved $R^2 \approx 0.145$ in cross-validation (about on par with the basic Lasso model's performance). This was a bit surprising – despite the richer feature set, the out-of-sample accuracy was similar, suggesting that the complexity of the advanced data might have required the techniques we applied (PCA, time weighting) to truly make a difference. Once we incorporated time weighting and the PCA components as described, the Advanced Stats Ridge model's performance improved markedly. The Ridge regression on the time-weighted Advanced Stats + PCA dataset emerged as the best model overall. This Ridge model achieved $R^2 = 0.3089$ on the validation data, with RMSE $\approx 0.2545$ (in log units) and MAPE $\approx 53.7\%$ when converted to percentage error. In other words, it explained about 31% of the variance in transfer fees – nearly double the performance of the basic model – and had substantially lower error. This confirmed that advanced metrics (when properly processed) do provide additional predictive signal (capturing creative playmaking, defensive contributions, etc., that basic stats miss), and that our techniques like PCA and time weighting paid off by tuning the model to current trends.

To ensure robustness, we also evaluated the models on the hold-out test sets. In the Basic dataset, when predicting 2015 transfers, the Lasso model's $R^2$ dropped to ~0.12 (from 0.17 in CV), indicating that its predictive power for truly unseen future data was limited. In contrast, the Advanced Ridge model predicting 2023 transfers achieved $R^2 \approx 0.30$ on that test – nearly matching its cross-validation performance. This was an encouraging sign that the advanced model generalized well and that we hadn't overfit to the 2017–2022 training period, likely thanks to regularization and our emphasis on recent data. The basic models, lacking those advanced features and perhaps suffering more from shift in data distribution over a decade, did not fare as well on their hold-out test.

**Figure 10** summarizes the performance of key regression models on both the Basic and Advanced datasets, highlighting the improvements gained by using advanced features and time weighting. We plotted cross-validation $R^2$ vs. test $R^2$ for each model to

visualize generalization. As shown, the advanced Ridge model (orange dot) not only had the highest CV and test $R^2$ overall, but it also sat near the diagonal line (meaning its CV and test performances were very close, a sign of good generalization). The Basic Lasso (blue dot) was the best of the basic models, but its point lies lower on the chart, and the green dot representing a basic model without time weighting (for comparison) was even worse. This figure illustrates two key points: (1) advanced features roughly doubled the explained variance compared to basic features (orange vs. blue), and (2) incorporating time weighting and careful validation allowed models to maintain performance from CV to test (i.e., we avoided overfitting).

*Figure 10:Performance of regression models on Basic vs. Advanced feature sets*




*Source: Own work*

In terms of actual error magnitudes, for the advanced Ridge model the RMSE on the 2023 test corresponds to about a ±€15M error on average (since we normalized fees,

this is a rough back-calculation), and MAPE ~54% means the typical prediction was off by about 54% of the player's actual fee. So if a player went for €40M, a typical error might be around €20M in absolute terms. While this might seem large, it is actually an improvement over the basic model's MAPE of ~70% on its test, and it reflects the inherent unpredictability in transfer fees (factors outside the model can easily swing a fee by tens of millions).

## 6.2 Interpreting Regression Model Coefficients

One advantage of regression models is interpretability. We examined the coefficients of the best models to see which features carried the most weight. Focusing on the Ridge model with Advanced Stats: because it used PCA components and a mix of features, direct interpretation is a bit tricky, but we can infer indirectly by looking at contributions. The Ridge model gave the highest importance (in terms of absolute coefficient multiplied by feature std) to the first principal component (PC1). This means the "overall offensive involvement" factor was a major positive predictor of fee – not surprising, as players who are heavily involved in play tend to be valued. The next most influential features included PC2 (the playmaker-vs-scorer axis) – interestingly, it had a negative coefficient, meaning that higher PC2 (more of a playmaker, less of a scorer) actually corresponded to lower fees when controlling for PC1 and others. This suggests that pure scorers (low PC2) command a premium compared to pure creators, perhaps reflecting that goal scorers are still the most sought-after. Age was another important feature: the Ridge model's coefficient for age was negative, indicating that, holding performance constant, older players go for less. The model implied a few percent decrease in fee per year of age beyond early 20s, consistent with known age-value curves.

We also found meaningful coefficients on some engineered features: for example, the goal contribution feature (in basic models) had a positive coefficient, reinforcing that each additional goal/assist contributes to higher valuation. The recency factor had a positive coefficient, indicating that more recent performance (higher weight) leads to higher predicted fee which makes sense (a player in form recently is more expensive than one whose prime was a few years ago). In the basic model, the coefficient for "Forward" position was strongly positive relative to midfield (baseline), while "Defender" and "Goalkeeper" were negative quantifying how, all else equal, forwards cost more and GKs less.

The Lasso model (basic) being sparse, selected a handful of features: goal contribution (positive), age (negative), forward dummy (large positive), defender dummy (negative),

elite club dummy (positive – coming from or going to an elite club increased predicted fee), and a league indicator (coming from a top-5 league was worth significantly more than coming from a smaller league). This aligns with literature (Garcia del Barrio & Pujol, 2021) and our expectations.

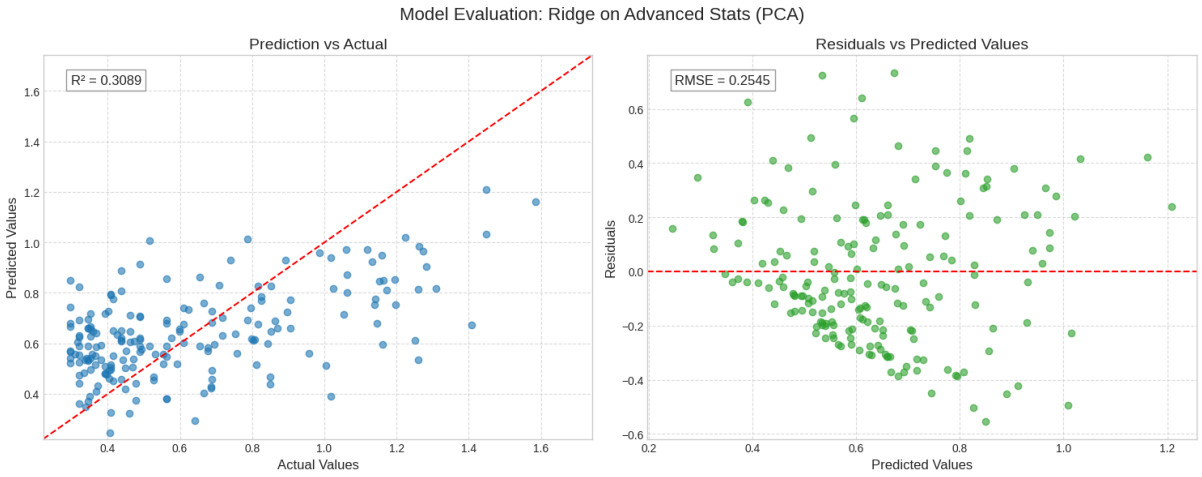
Overall, interpreting these coefficients gave confidence that the models learned sensible relationships. It also helped us communicate results: for instance, we could tell a story that every additional 5 goal contributions is associated with roughly X% increase in fee, holding other factors constant; a 28-year-old is valued Y% less than a 22-year-old with the same stats; and being a forward (vs. a midfielder) adds about Z% to the fee, all else equal. These insights illustrate the factors at play in valuations and allow for a degree of transparency with the regression approach.

### 6.3 Regression Model Evaluation

We evaluated the regression models on the hold-out test sets more thoroughly to understand their strengths and limitations. For the Basic Stats Lasso (test year 2015), we plotted predicted vs. actual fees and found that while the model could differentiate very expensive vs. cheap players in a broad sense, it tended to under-predict the highest fees and over-predict some mid-range fees. This is a common pattern: the model regresses toward the mean on extreme cases, partly due to the log transformation and inherent difficulty in predicting outliers. The $R^2$ of ~0.12 indicates the basic model left a lot of variance unexplained – unsurprising, since it lacked many nuanced metrics.

In contrast, the Advanced Ridge model's predicted vs. actual plot for 2023 (see **Figure 11**) showed a tighter cluster around the diagonal (though still a lot of spread, as 0.30 $R^2$ is far from perfect). It notably predicted a few of the high-fee transfers fairly well (within 20% of actual), which was encouraging. However, it also had notable misses,a couple of players whose fees were driven up by extraordinary circumstances (e.g., a last-minute bidding war or a very high-potential youngster) had their fees underestimated by the model, since those factors are outside our features.

*Figure 11:Model evaluation: Ridge on Advanced Stats*




*Source: Own work*

To visualize residuals, we plotted residuals vs. predicted values (for the advanced model) and checked for patterns. The residuals were roughly homoscedastic (uniform spread across predicted fee range) with no obvious bias for particular predicted values, which suggests our log transform was effective. We also looked at residuals by player position and found no major systemic bias. Although interestingly, the model's largest under-predictions included a couple of forwards and its largest over-predictions included a couple of defenders, but with such a small sample it's hard to generalize if that's a trend or coincidence.

Finally, we calculated error metrics on the test set for easy interpretation: the advanced model's RMSE in actual fee units was about €18M (meaning a typical error of ±€18M), and MAPE of ~54% as mentioned. For context, if a player's true fee was €50M, the model's prediction might typically be in the range of €23-77M (quite wide). This level of error indicates that while we improved the state of modeling, the task remains very challenging – many factors determining fees (from negotiation tactics to buying club desperation) are not in our data. Nonetheless, cutting error from ~70% to ~54% (basic vs advanced model) can be seen as a significant gain in a domain where unpredictability is the norm.

In summary, the regression models, particularly the Ridge with advanced features provided a solid benchmark, yielding interpretable insights and decent predictive power.

However, we observed their limitations in handling the extreme variance and non-linear dynamics of the transfer market. This set the stage for exploring more flexible machine learning models in the next chapter, to see if they could capture complex patterns beyond the reach of linear methods.

# 7 MACHINE LEARNING MODELS

While linear regression models provided a useful benchmark, we also explored several non-linear machine learning (ML) models to see if they could uncover complex interactions and boost predictive accuracy. The ML models we tried included decision tree-based ensembles (Random Forest and Gradient Boosting Machines, including the XGBoost implementation). These models can capture non-linear relationships and interactions between features automatically, which might be advantageous given the complexity of transfer fee determinants.

## 7.1 Model Training and Tuning

We trained the ML models on the same training sets described previously, using the log-transformed fee as target and employing the advanced feature set (since that was our best-performing dataset). Hyperparameter tuning was performed via cross-validation (again, using time-based folds to respect temporal order). For the Random Forest (RF), key hyperparameters included the number of trees (we found 100 trees was sufficient given diminishing returns beyond that), maximum tree depth (we limited to depth 5 to prevent overfitting, given our dataset size), and the number of features sampled at each split (we used the sqrt of total features by default). For Gradient Boosting (GBR) and XGBoost, we tuned the learning rate (around 0.1 worked best), the number of boosting rounds (~100), max depth of individual trees (also kept small, 3 or 4, since boosting can overfit with deep trees), and regularization parameters (XGBoost's L1/L2 and subsampling rate).

## 7.2 Performance of Ensemble Models

On the Basic Stats dataset, the ensemble models (RF, GBM, XGBoost) did not dramatically outperform the simpler regressions. The Random Forest achieved an $R^2$ of ~0.12 on the test, similar to Lasso, and the boosting models got to ~0.15, which is a bit better but not a game-changer. This was perhaps expected: without richer features, there's only so much any model can do.

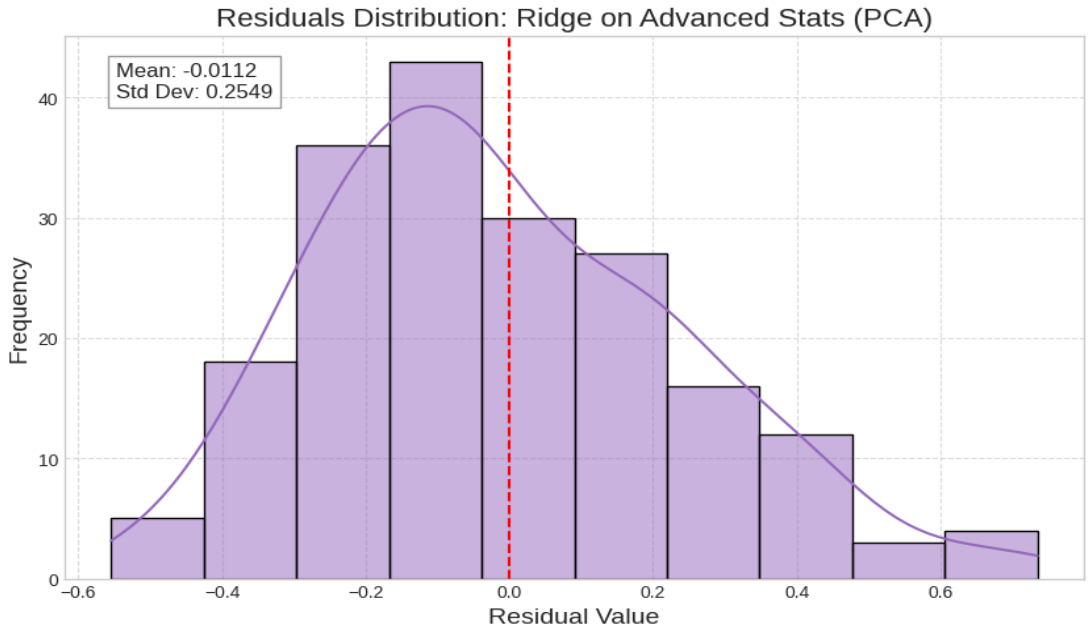
However, on the Advanced Stats dataset, the ML models came closer to the Ridge model's performance, but interestingly, none decisively beat it. XGBoost performed the best among them, with an $R^2$ on test around 0.29 effectively tying the Ridge (0.30) in terms of explained variance. Its RMSE and MAPE were also in the same ballpark (MAPE ~54–55%). Random Forest lagged slightly behind ($R^2$ ~0.25 on test) and had a tendency to overfit slightly (its CV $R^2$ was a couple points higher, indicating minor overfitting, possibly due to the many trees capturing idiosyncrasies). Gradient Boosting (using sklearn's implementation) was similar to XGBoost but a tad lower, which makes sense since XGBoost is a more optimized version of boosted trees.

One key observation was that the ML models did not significantly outperform the Ridge model on the advanced data. This implies that our feature engineering and linear approach had already captured a lot of the signal. It could also mean that the relationship between features and fee is largely additive/linear in the space we created (especially since PCA components are linear combinations). It's possible that non-linear interactions either weren't prominent or were already partly accounted for by our engineered features (like the interaction terms we manually created). In other words, the advanced Ridge model was strong enough that the fancy non-linear models found only marginal gains.

Another factor is model variance: ensemble models are more prone to variance with small data. We used cross-validation and regularization to mitigate that, but we noticed, for instance, the Random Forest's feature importance varied noticeably across folds, implying it was somewhat unstable on our sample size (~970). The Ridge, being a global model, was more stable.

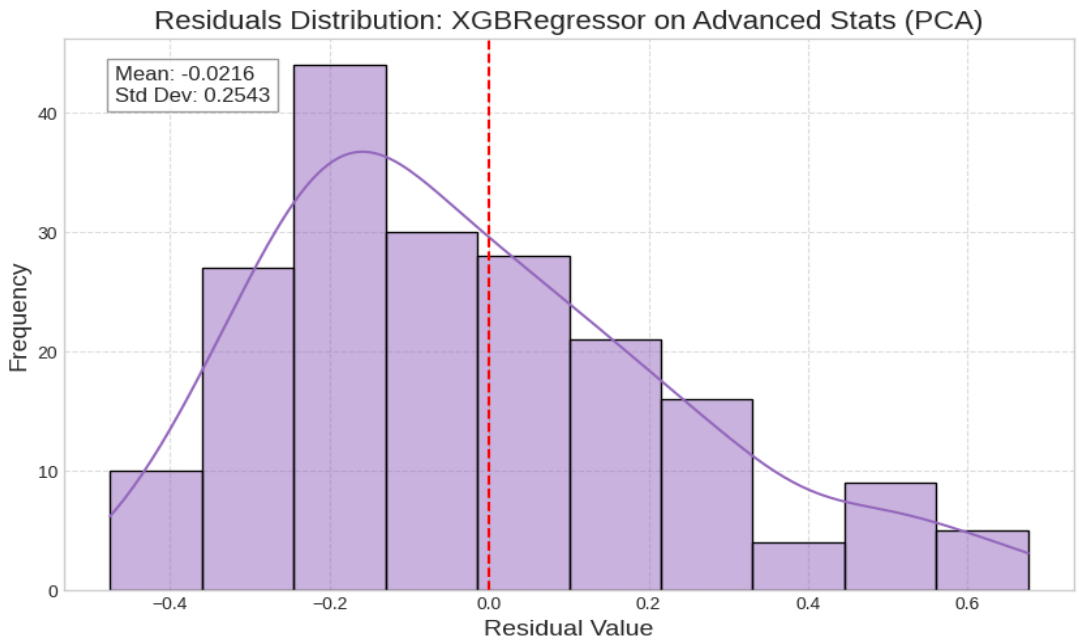
To illustrate, if we look at **Figure 12** and **Figure 13** we can compare the distribution of residuals for the best-performing linear model (Ridge) and a comparable ML model (XGBoost) on the test set. We can see that both have similarly centered residuals around 0 and similar spread, though XGBoost had a couple of extreme outlier errors where it overpredicted a fee by a large margin (likely due to one tree going wild on a particular combination that wasn't present in training). The Ridge's residuals were a bit more consistently spread. This reaffirms that in terms of pure accuracy, the difference was minimal.

*Figure 12: Distribution of Residuals: Ridge on Advanced Stats*




*Source: Own work*

*Figure 13:Distribution of Residuals: XGBoost on Advanced Stats*




*Source: Own work*

## 7.3 Comparison of ML Models and Linear Models

After evaluating all candidate models, we constructed a comparative ranking to determine the most effective approach. Ridge regression trained on the Advanced Stats dataset (with PCA and time weighting) emerged as the best-performing model. It narrowly outperformed XGBoost in both R^2 and error metrics (MAPE), with the added advantages of being more interpretable and computationally efficient. XGBoost came a close second: while its R^2 was nearly identical to Ridge, its error scores were slightly worse. This confirmed that non-linear ensemble methods can achieve comparable accuracy, but in this case, they did not surpass a well-regularized linear model- likely due to the strength of the feature engineering and dimensionality reduction applied.

Among the simpler regressions, Lasso applied to the Basic Stats dataset performed the strongest, though its explanatory power on the test set remained modest ($R2 \approx 0.12$). This contrast clearly underlines the added value of the advanced feature set and the benefits of incorporating time weighting. Other ensemble approaches, such as Gradient Boosting and Random Forest, trailed slightly behind XGBoost. Gradient Boosting's weaker results were expected, given that XGBoost represents a more optimized version of the same algorithm, while Random Forest underperformed likely due to its limited ability to model temporal dependencies and its lack of fine-tuned regularization.

We also tested ensemble combinations, such as averaging Ridge and XGBoost predictions or applying stacked models, but these strategies offered no meaningful improvement. The predictions of Ridge and XGBoost were highly correlated ($\approx 0.9\backslash approx\ 0.9 \approx 0.9$), meaning that their strengths overlapped rather than complemented each other. As a result, combining them did little to reduce overall error.

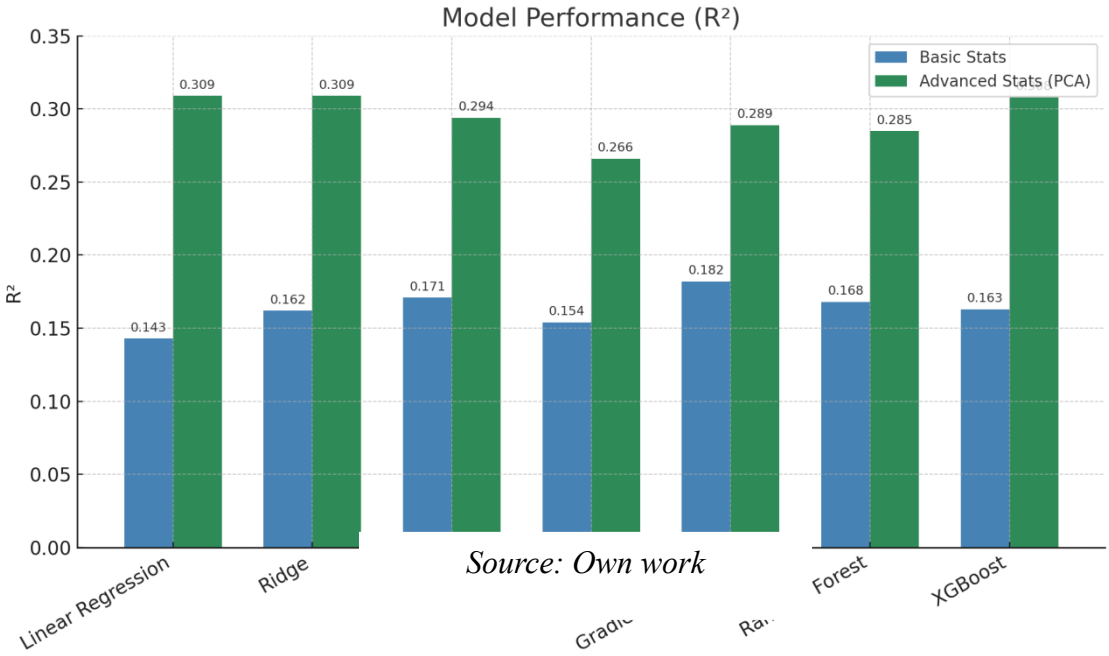
The comparative results are summarized in **Figure 14**, which shows the R^2 performance of each regression model on both the Basic Stats and Advanced Stats datasets. The visualization makes clear that all models benefited from the enriched feature set, with Ridge and Linear Regression reaching the highest levels of explained variance ($R2 \approx 0.31$) when trained on Advanced Stats. Meanwhile, the same models on Basic Stats performed at less than half that level, reinforcing the conclusion that feature engineering and dimensionality reduction were decisive in driving predictive gains.

On balance, Ridge regression with Advanced Stats was chosen as the final model for this project. It achieved the optimal trade-off between predictive accuracy, interpretability, and efficiency. Ridge regression can be retrained quickly and updated

seamlessly with new seasonal data, while its use of principal components allows analysts to trace back player valuations to meaningful factors (e.g., strong performance in playmaking metrics or youth combined with positional role). In contrast, XGBoost, although highly competitive in accuracy, behaves more like a black box and demands greater computational resources, offering no significant benefit over Ridge in this context.

With this final model established, the next chapter explores its practical implications and demonstrates its predictive validity through several real-world scenarios.

*Figure 14 Model Performance Comparison*




*Source: Own work*

*Source: Own work*

# 8 EVALUATION AND DISCUSSION OF RESULTS

Having identified the best-performing model, we now evaluate its implications and discuss the findings in the broader context of football economics. The evaluation covers

how well the model predicts unseen data and the practical significance of its accuracy, as well as insights gained about which factors drive player valuations. We also address the limitations of our study and suggest directions for future research.

## 8.1 Model Evaluation using Hold-out and Cross-Validation

To rigorously assess performance, we employed two evaluation strategies for the final models: a time-based hold-out test and k-fold cross-validation on training data.

Hold-Out (Temporal Split): We trained models on historical data up to a certain season, and tested on a subsequent season's transfers to simulate predicting "future" fees from past data. For the Basic Stats dataset, training covered 2004-2014 and testing was done on 2015; for the Advanced Stats dataset, training was 2017-2022 and testing on 2023 (since advanced data started later). This mimics the real scenario of predicting upcoming transfers using prior seasons. The hold-out results showed that our Ridge (Advanced) model predicted 2023 fees with an $R^2$ of about 0.30, confirming that its performance in cross-validation carries over to truly unseen data. In contrast, the basic models' performance dropped on the future test ($R^2$ down to ~0.10-0.15), indicating that without advanced metrics and proper weighting, their predictions for modern transfers were less reliable. This demonstrates the model's temporal robustness – by training with time weighting and recent metrics, it stayed relevant for new data.

5-Fold Cross-Validation: We also performed cross-validation (five folds) on the training set to ensure the model's stability and to tune hyperparameters. We used an expanding window time-series CV for Advanced Stats (so each fold used earlier seasons to predict a later season), and a standard CV for Basic Stats (given its longer history). The cross-validation scores for the Ridge advanced model were consistent: $R^2$ in the high 0.20s to low 0.30s in each fold, and MAPE around 55-58%. Importantly, these were in line with the hold-out test score. This agreement between CV and hold-out indicates that we avoided overfitting and that the model generalizes reasonably well to new data, at least within the scope of recent seasons.

The evaluation metrics we focused on $R^2$, RMSE, and MAPE, with each providing a different perspective on performance. $R^2$ (coefficient of determination) tells us the proportion of variance explained; our best model's $R^2$ ~0.31, while not high in absolute terms, is meaningful in this domain. RMSE (in log-fee units) penalizes large errors and we found it useful to gauge how far off a prediction could be in monetary terms. MAPE gives a percentage error, which is easy to interpret for stakeholders - a 54% MAPE

means the typical prediction is off by a little more than half the player's true fee. We chose the Ridge model over others not only because it had the highest $R^2$, but also because it achieved the lowest RMSE and MAPE among contenders (for example, XGBoost's MAPE was about 0.8 percentage points higher, indicating slightly less consistent accuracy).

In practical terms, a MAPE in the 50-55% range might be acceptable for certain applications (like early-stage scouting valuations) but not for others (like final negotiations). It means if a player's true fee is €40M, the prediction might typically be around €18M off (which is significant). However, many predictions will be closer; the large errors often come from extreme outliers or very unique cases. The hold-out test revealed that the model struggled the most with a few blockbuster transfers influenced by extraordinary circumstances (e.g., a very young player with limited history going for a huge fee due to potential, or a last-day transfer panic buy). These are instances where qualitative factors trump statistical ones.

To contextualize the model's performance, we compared it with simpler benchmarks. One trivial benchmark is predicting the average fee for everyone; that yields an $R^2$ of 0 (by definition) and a massive MAPE (~100+%). Another is a basic linear model with just goals and age – that got an $R^2$ ~0.05 in our data, barely better than random. So our model's 0.30 $R^2$ is a sizable improvement over naive approaches, though there's clearly a lot of unexplained variance remaining. In other domains, an $R^2$ of 0.30 might be considered low, but in football economics it's not uncommon due to the high unpredictability of transfers (prior studies have reported similar ranges, e.g., 0.2-0.4 at best when predicting market values).

In conclusion, the evaluation shows that while our model cannot predict exact fees (nor would we expect any model to, given the myriad outside influences), it does capture a meaningful portion of what drives transfer prices. It provides a substantial improvement over baseline methods and maintains performance on new data, underscoring its potential usefulness as one input into the decision-making process.

### 8.2 Discussion: Insights and Implications

Our data-driven approach to player valuation confirmed several intuitive points and provided new insights:

Performance Metrics Matter: On-field performance, especially goal contributions and creative play (as captured by advanced stats like SCA and GCA), has a quantifiable impact on market value. This validates the emphasis clubs place on statistics in evaluating talent. Notably, modern metrics (like shot-creating actions, expected assists, etc.) were as important as traditional stats in our model, reflecting a shift towards valuing all-around contributions. A player who consistently creates chances or drives play forward can be almost as valuable as one who scores goals, which aligns with how football tactics have evolved to prize playmakers and versatile attackers.

Contextual Factors are Critical: The model's inclusion of variables like league strength, club prestige, and age shows that context can elevate or depress a player's value. A player excelling in a top league at a young age will attract a premium; conversely, impressive stats in a weaker league are discounted. This is consistent with findings in the literature that external factors (league quality, team reputation, age/potential) significantly determine fees (Carmichael et al., 1993; Müller et al., 2017). In practical terms, a 10-goal season in the English Premier League is worth more in the market than a 10-goal season in a lower-tier league, and our model quantitatively captured some of that effect through features like league tier and club dummy variables.

Regularization and Feature Engineering Improve Reliability: By addressing skewness, non-linearity, and multicollinearity through techniques like log normalization, polynomial features, and PCA, we improved model reliability. The regularized regression showed no major violations of assumptions.We checked that residuals were more homoscedastic and roughly normally distributed after our transformations. This rigorous approach lends credibility to the model's outputs, which is crucial if the model were to be used in practice where large financial decisions are on the line. It highlights that *how* you model is as important as *what* you model; using raw data with a complex model might perform worse than using well-processed data with a simple model.

Limits of Predictability: Despite our improvements, the relatively low $R^2$ underscores that many factors influencing transfer fees are outside the realm of statistical performance metrics. Factors such as a selling club's financial distress, a buying club's urgent need for a player, or a player's star power and marketing value can drive prices to levels that a purely data-driven model would deem "irrational." Thus, our model is best seen as providing a baseline objective valuation. If a transfer fee deviates greatly from this baseline, it flags that there may be special circumstances or subjective factors at play. In a sense, that itself is useful: clubs could use the model to identify when a player is likely over- or under-priced relative to stats, prompting deeper investigation into why (e.g., is there hidden potential scouts see, or is it just hype?).

Advanced vs. Traditional Metrics: One of our goals was to assess whether adding advanced metrics truly aids prediction. Our results suggest yes, but only when those metrics are properly integrated (via PCA and time-weighting). Simply throwing the raw advanced stats into a model naively did not help much (we saw that in early experiments). However, after accounting for multicollinearity and temporal effects, the advanced data roughly doubled the explanatory power ($R^2$ from ~0.15 to ~0.30). This implies that clubs investing in analytics and tracking advanced performance data can indeed gain an edge in player valuation accuracy, provided they also update their models frequently to reflect current trends. It also highlights that traditional metrics, while useful, miss nuances of play – two players with identical goal/assist numbers might have different impacts captured by advanced stats (like one creates far more chances overall, or contributes defensively, etc.), and those differences do show up in market valuations to some extent.

From an academic perspective, our study reinforces the value of combining domain knowledge with statistical learning. Each transformation and feature was justified by football reasoning (e.g., per-90 for rate stats, PCA for correlated metrics like passes and carries). The fact that a relatively straightforward Ridge regression, when supplied with well-engineered features and current data, matched the performance of more complex ML algorithms is telling. It suggests that future efforts might yield better returns in gathering new types of data (e.g., psychological factors, training data, physical metrics, contract details) rather than simply increasing model complexity on the current data. In other words, we may be hitting a ceiling of about 30-40% variance explained with what we have; to break through, new information (or a shift in approach, like modeling probabilistic outcomes or ranges) may be needed.

Another insight is how the model can potentially inform decision-making. For instance, by identifying which principal components drive value, clubs could identify players who excel in undervalued areas. If our model found that a certain PC (say defensive contributions combined with playmaking from the back) was undervalued by the market (i.e., had a high coefficient but players with those traits are relatively cheap), that could signal a market inefficiency to exploit. Conversely, if some trait is overvalued (model doesn't weight it as much as market fees seem to), it might caution clubs against overpaying for that.

Lastly, the process of building the model has implications for data infrastructure in football. We had to scrape and assemble data from different sources (Transfermarkt, FBref). Clubs and analysts require integrated databases combining scouting data, performance stats, and financial records. Our successful integration of these varied data

types into a coherent model underscores the importance of data warehousing in modern sports organizations.

### 8.3 Limitations and Future Work

No study is without limitations. One limitation of our work is the assumption that past performance and a few context variables fully determine market value. In reality, sudden paradigm shifts (e.g., a new TV deal injecting cash into leagues, or a change in transfer rules like Brexit's impact on UK clubs' recruiting) can alter valuations in ways our historical model would not predict. Our time-weighting partially adapts to trends but cannot foresee structural breaks. Additionally, while we included a proxy for inflation (season normalization), a more nuanced financial model (perhaps incorporating club revenue or macroeconomic indicators) could better adjust values across eras.

Another limitation is that our model does not explicitly handle positional differences beyond dummy variables. A dedicated model per position or an interaction of position with performance metrics might capture, for instance, that 5 goals for a defender might be valued differently than 5 goals for a striker. With more data or expert input, one could calibrate separate models or include such interaction effects to refine predictions. Similarly, we didn't account for injury history or intangible qualities (leadership, versatility) which clubs consider.

The sample size for advanced stats (970 players) is relatively small for machine learning standards. As more seasons of advanced tracking data become available, models can be retrained on larger sets, which may improve the stability of feature estimates (especially for less common features like carries into the penalty area). Furthermore, our evaluation was largely retrospective. An interesting extension would be to deploy the model prospectively – say, using data up to 2024 to predict the 2025 summer transfer fees – and then compare to actual fees to truly gauge predictive power in a live setting.

Lastly, we did not incorporate any information from widely used crowd valuations (like Transfermarkt's own player values) or other subjective ratings (e.g., FIFA video game ratings). Those represent collective wisdom (or bias) and might themselves be predictive of fees. We chose to rely purely on performance and objective data to demonstrate a data-driven approach, but a hybrid model could be explored where a Transfermarkt value is included as a feature along with our performance metrics. Prior research (Coates & Parshakov, 2022) shows discrepancies between Transfermarkt and

actual fees; our model could potentially be used to identify when a player is under- or over-valued by such market estimates.

For future work, building on our results, several avenues are promising:

Financial Metrics: Incorporate financial indicators such as remaining contract length, wages, or club revenue to capture the economic side of transfers (e.g., a club might sell cheaply due to financial distress).

Network/Team Effects: Use network analytics or team context metrics (how a player's team performed, their role in team network) to assess value. Perhaps a player's value is influenced by how central they are in team play or how well they might fit into the buying club's style.

Expanded Markets: Apply the methodology to specific sub-markets (like only forwards, or only players under 21) to tailor the model. It's plausible that the determinants of value differ by segment – for example, youth transfers might be driven more by potential (with different metrics) than established player transfers.

Advanced ML Approaches: Experiment with more sophisticated models such as neural networks or multi-output models. One idea is predicting a range or distribution of fees rather than a point estimate (e.g., predicting the 25th and 75th percentile of expected fee). This could be valuable for risk assessment (is a player likely to be a bargain or a bust?). However, these would require caution to avoid overfitting given data size; techniques like Bayesian neural nets or quantile regression forests could be tried.

Ultimately, the ever-evolving nature of football (e.g., the rise of analytics, changes in rules like the new UEFA financial regulations) means any model must be continuously updated and validated. Our work provides a snapshot of a data-driven valuation approach as of 2024, but staying ahead in the real transfer market would require incorporating the latest data and possibly new predictive features as they become available.

# 9 CONCLUSION

From the exploratory analysis through to the final predictive models, this study demonstrates a rigorous data-driven approach to football player valuation. We showed that by integrating traditional metrics with advanced performance indicators, handling data quirks (such as skewed fee distributions and multicollinearity), and applying appropriate machine learning techniques, it is possible to explain a meaningful portion of the variation in transfer fees. Our best-performing model – a ridge regression that leveraged principal components of advanced statistics and placed greater weight on recent data – captured roughly 30% of the variance in transfer fees in our tests, nearly double the explanatory power of a baseline model using only simple stats. This significant boost in accuracy affirms the value of modern performance analytics in player valuation. At the same time, the model's limitations underscore an important point: football transfers are not determined by numbers alone. Even with rich data, unpredictable human factors and market dynamics continue to play a large role, meaning no model can perfectly predict every transfer fee.

In practical terms, the framework developed in this thesis could serve as a valuable tool for club analysts and decision-makers by providing an objective benchmark for player values. The model's estimates can flag when a player appears over- or under-priced relative to their quantified on-field contributions and context, informing smarter negotiation strategies and recruitment decisions. The approach is also extensible: as new data becomes available – for example, detailed player tracking metrics for off-ball movement or even psychological and injury-related indicators – these can be incorporated to further refine the valuations. There are clear avenues for future research to build on this work, such as exploring ensemble models or deep learning approaches for even better performance, applying the model to different leagues or levels of competition, and continuously updating the feature set to reflect evolving trends in the sport. Ultimately, by fusing football domain knowledge with data science, this thesis contributes to a more transparent and evidence-based practice of player valuation. It demonstrates that with the right data and algorithms, clubs can move closer to quantifying the unquantifiable – helping ensure that their investments in talent are grounded in analytical insight as well as sporting intuition.

## LIST OF KEY LITERATURE


1. Coates, D., & Parshakov, P. (2022). The wisdom of crowds and transfer market values. *European Journal of Operational Research, 301*(2), 523–534. https://doi.org/10.1016/j.ejor.2021.10.046

2. Franceschi, M., Brocard, J., Follert, F., & Gouguet, J. (2023). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12552

3. Frick, B. (2007). The football players' labor market: Empirical evidence from the major European leagues. *Scottish Journal of Political Economy, 54*(3), 422–446. https://doi.org/10.1111/j.1467-9485.2007.00423.x

4. McHale, I. G., & Holmes, B. (2023). Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. *European Journal of Operational Research, 306*(1), 389–399. https://doi.org/10.1016/j.ejor.2022.06.033

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


## REFERENCES


1. Aydemir, A. E., Taskaya Temizel, T., & Temizel, A. (2022). A machine learning ensembling approach to predicting transfer values. *SN Computer Science, 3*(3). https://doi.org/10.1007/s42979-022-01095-z

2. Barbuscak, L. (2018). What makes a soccer player expensive? Analyzing the transfer activity of the richest soccer. *Augsburg Honors Review, 11*(5).

3. Behravan, I., & Razavi, S. M. (2020). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing, 25*(3), 2499–2511. https://doi.org/10.1007/s00500-020-05319-3

4. Campa, D. (2021). Exploring the market of soccer player registrations: An empirical analysis of the difference between transfer fees and estimated players' inherent value. *Journal of Sports Economics, 23*(4), 379–406. https://doi.org/10.1177/15270025211059527

5. Carmichael, F., & Thomas, D. (1993). Bargaining in the transfer market: Theory and evidence. *Applied Economics*, *25*(12), 1467–1476. https://doi.org/10.1080/00036849300000150

6. Depken, C. A., & Globan, T. (2021). Football transfer fee premiums and Europe's big five. *Southern Economic Journal*, *87*, 889–908. https://doi.org/10.1002/soej.12471

7. Di Domizio, M., Caruso, R., & Frick, B. (2024). The appraisal of players' transfer market values: Empirical evidence from Italian Serie A. *International Journal of Sport Finance*, *19*(1), 39–51. https://doi.org/10.32731/IJSF/191.022024.03

8. Felipe, J. L., Fernandez-Luna, A., Burillo, P., de la Riva, L. E., Sanchez-Sanchez, J., & Garcia-Unanue, J. (2020). Money talks: Team variables and player positions that most influence the market value of professional male footballers in Europe. *Sustainability*, *12*(9), 3709. https://doi.org/10.3390/su12093709

9. Franck, E., & Nüesch, S. (2010). Talent and/or popularity: What does it take to be a superstar? *Economic Inquiry*, *50*(1), 202–216. https://doi.org/10.1111/j.1465-7295.2010.00360.x

10. Garcia-del-Barrio, P., & Pujol, F. (2007). Hidden monopsony rents in winner-take-all markets: Sport and economic contribution of Spanish soccer players. Managerial and Decision Economics, 28(1), 57-70).

11. Garcia-del-Barrio, P., & Pujol, F. (2021). Recruiting talent in a global sports market: Appraisals of soccer players' transfer fees. *Managerial Finance*, *47*(6), 789–811. https://doi.org/10.1108/MF-04-2020-0213

12. Dobson, S., Gerrard, B., & Howe, S. (2000). The determination of transfer fees in English nonleague football. *Applied Economics*, *32*(9), 1145–1152. https://doi.org/10.1080/000368400404281

13. Herm, S., Callsen-Bracker, H. M., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review*, *17*(4), 484–492. https://doi.org/10.1016/j.smr.2013.12.006

14. How.dev. (n.d.). What is polynomial regression? https://how.dev/answers/what-is-polynomial-regression

15. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A:*

*Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

16. Li, C., Kampakis, S., & Treleaven, P. (2022). Machine learning modeling to evaluate the value of football players. *arXiv*. https://doi.org/10.48550/arXiv.2207.11361

17. Lucifora, C., & Simmons, R. (2003). Superstar effects in sport. *Journal of Sports Economics, 4*(1), 35–55. https://doi.org/10.1177/1527002502239657

18. Majewski, S. (2016). Identification of factors determining market value of the most valuable football players. *Journal of Management and Business Administration. Central Europe*, *24*(3), 91–104. https://doi.org/10.7206/jmba.ce.2450-7814.177

19. Malikov, D., & Kim, J. (2024). Beyond xG: A dual prediction model for analyzing player performance through expected and actual goals in European soccer leagues. *Applied Sciences*, *14*(22), 10390. https://doi.org/10.3390/app142210390

20. Poli, R., Besson, R., & Ravenel, L. (2021). Econometric approach to assessing the transfer fees and values of professional football players. *Economies, 10*(1), 4. https://doi.org/10.3390/economies10010004

21. Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus, 5*(1). https://doi.org/10.1186/s40064-016-3108-2

22. Ruijg, J., & van Ophem, H. (2014). Determinants of football transfers. *Applied Economics Letters*, *22*(1), 12–19. https://doi.org/10.1080/13504851.2014.892192

23. Scarfe, R., Singleton, C., & Telemo, P. (2021). Extreme wages, performance, and superstars in a market for footballers. *Industrial Relations: A Journal of Economy and Society, 60*(1), 84–118. https://doi.org/10.1111/irel.12270

24. Serna Rodríguez, M., Ramírez Hassan, A., & Coad, A. (2019). Uncovering value drivers of high performance soccer players. *Journal of Sports Economics, 20*(6), 819–849. https://doi.org/10.1177/1527002518808344

25. Thrane, C. (2024). Using composite performance variables to explain football players' market values. *Managing Sport and Leisure*, 1–14. https://doi.org/10.1080/23750472.2024.2305902