Key Metrics Influencing Player Wages: An Empirical Analysis of European Football's Top Five Leagues

University *of* Exeter

Submitted by Waleed Eltigani to the University of Exeter as a dissertation for the degree of Master of Science in Applied Data Science & Statistics, August 2024

I certify that all material in this dissertation which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signature: 

Table of Contents

ABSTRACT

The aim of this study is to identify the key on-field performance metrics that influence Football player wages across different positions in Europe's top five football leagues in 2024. By focusing on performance indicators—defined as a chosen set of variables that represent specific aspects of performance and contribute to athletic success (Hughes and Bartlett, 2002)—this research narrows the scope to on-field metrics that are critical in today's football. In addition, factors like age, league, and nationality are included to provide contextual understanding, acknowledging that these elements also play a significant background role in shaping player wages and should be coupled with their performance metrics. Using machine learning models, including Random Forest, Gradient Boosting, and Bayesian Additive Regression Trees (BART), the study develops predictive models to analyze the relative importance of these metrics in wage determination. In addition to these models, the application of traditional regression techniques such as Stepwise Regression and Quantile Regression provided insights into wage patterns across different wage levels. For goalkeepers, advanced metrics like Post-Shot Expected Goals per Shot on Target (PSxG/SoT) and distribution abilities were important, reflecting their evolving role in both shot-stopping and necessity for on-ball ability in the modern game. Additionally, team results were also an important factor in determining wages. The evolution of defenders into more ball-playing roles made metrics like pass completion % and passing accuracy, significant in determining wages. For midfielders, Expected Goals metrics like xG Plus/Minus highlighted the importance of versatility in both offensive and defensive roles, aligning with the demands of the modern midfielder. For forwards, metrics such as xG ± per 90 minutes and Non-penalty Expected Goals per Shot underscored their crucial role in influencing team goal-scoring opportunities, reflecting the shift towards a more comprehensive evaluation of their contributions beyond just scoring. Playing in the Premier League emerged as a key indicator of higher wages across all positions, while Ligue 1 had the opposite effect, highlighting the financial disparities between leagues. The results display the influence of football's tactical evolutions, forcing a shift from specialised roles to multi-faceted roles, with influence on overall team influence of a player being heavily weighted in wage determination. The findings reveal that specific performance metrics are consistent in models for each of the four positions; offering valuable information for players, clubs, and agents in optimising wage structures and contract negotiations. By providing a data-driven understanding of the factors driving wage disparities in modern football, this study contributes to more equitable and informed decision-making for all parties.

1. INTRODUCTION

The world of professional football is a multi-billion pound industry, particularly in Europe, where the top five leagues—the English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, and French Ligue 1—dominate both viewership and revenue all over the globe. In such a financially intense environment, determining fair wages for professional football players is crucial. Accurate wage determination is essential not only for maintaining a balanced club budget but also for ensuring team chemistry and player satisfaction. A club's ability to sign new players, maintain harmony within the team, and effectively manage player transfers can be highly influenced by how much they are paying players on their wage bill. Decisions on salaries influence the football market by creating benchmarks and establishing new reference points for wage expectations and negotiations across the industry.

Objective of the Study

The primary aim of this dissertation is to identify the key on-field performance metrics that influence player wages for each position—goalkeepers, defenders, midfielders, and forwards—in Europe's top five football leagues; . While many past studies have focused on factors such as player popularity, media influence, and team-oriented metrics, this research narrows the focus to player performance metrics, which are the most direct contributors to a team's success on the pitch. Factors like league, nationality, and age were used to support contextualising where, and at what stage of his career, the player was producing these performance metrics. To achieve our objective, we employed statistical models to analyse how these performance metrics correlate with wages, with the aim of developing models that would determine the key predictors and predict players' wages as accurately as possible.

Implications

This study will be valuable for both players and clubs by providing a data-driven approach to understanding wage determination. For players, it offers insights into the specific performance metrics that can enhance their market value today, enabling them to focus on improving specific areas that directly impact their earnings. Additionally, it equips players with a deeper awareness of their market value, helping them make more strategic decisions in contract negotiations and career development by understanding which leagues they should target for the best opportunities. For clubs, the research helps optimise wage structures by identifying the most cost-effective metrics that correlate with on-field success, leading to more informed contract negotiations and better financial management. These findings can also serve as a crucial tool for player agents, club negotiators, and analysts in shaping strategic decisions during contract offer or renewal discussions.

Assumptions and Limitations

This study assumes that on-field performance metrics, in addition to age, nation, and league, are the primary drivers of player wages and can be applied universally across Europe's top five leagues, despite variations in playing styles and environments. The study excludes other significant factors, such as player popularity like social media presence, the financial health of clubs, and broader market dynamics to focus solely on what happens on the pitch.

While leagues often have overarching styles of play, it's crucial to recognize that individual teams within those leagues can have vastly different approaches to playing. Each manager has their own philosophy for how they want their teams to play. For example, Sheffield United, a team in the Premier League for the 2023/24 season, is known for its direct style, frequently playing long balls to their attackers to win headers and maintain pressure in the opponent's half, rather than focusing on slow, possession-based buildup. In contrast, Manchester City emphasises dominating possession and intricate passing. As a result, Sheffield United might prioritise paying higher wages to a taller, more physical striker with a strong aerial duel success rate, whereas Manchester City would place more value on a player excelling in shot creation and passing actions. The diversity in tactical priorities for teams places different levels of importance on various metrics when deciding player wages. This study extracts the top predictors from all of Europe's top 5 leagues rather than by league, or by individual teams.

Furthermore, the study assumes a level-playing field regarding the financial capabilities of clubs, which is a simplification taken to focus solely on the player impact aspect. Wage dynamics are closely related to the ambitions of the club's management and the targets they are pursuing (Bernardo, Ruberti, & Verona, 2022). Wealthier clubs often pay much higher wages due to their high ambitions and financial power, independent of player performance; meaning the same player with the same performance levels, could possibly make a far more lucrative salary at one club than with another club. This assumption was made to maintain a clear focus on the impact of performance metrics, though it is acknowledged that the exclusion of financial disparities may limit the study's ability to fully capture the complexities of wage determination in professional football.

Importance of the study

Wages of a football player are determined by a wide combination of factors; popularity, financial power of the buying club, negotiating ability of the agents and club negotiators, and more. Of course a player's ability to impact merchandise sales and sponsorships can influence their value to a club, however the core driver of a footballer's value ultimately is their contribution to winning matches. How are they impacting results *on* the pitch?

In North America, teams and organisations prioritise profit maximisation as the main objective, thus the player's salaries are made to be proportional to team revenues (Bernardo, Ruberti, & Verona, 2022). This approach creates a more level playing field, where multiple teams have a realistic chance of winning, rather than financially dominant teams overwhelming the

competition. This increased competition keeps fans engaged and grows the overall appeal of the league, leading to higher revenues from sports events (Szymanski 2003). In contrast, European teams, particularly in football, focus on athletic success, driving up wages for top players and opposing the implementation of wage caps (Fort 2000; Sloane 2006), resulting in a huge disparity in the financial capability of the top clubs versus the rest.

In European football, winning is the most important target. Success on the pitch drives popularity and profit, attracts the best talent and coaches, and thus creates a cycle of sustained success. A footballer's impact on winning is generally the most important indicator of their wage. Through the results, we can analyse what the teams in the top 5 European League, in 2024, are prioritising as the most important predictors for each position on the pitch, that ultimately create a winning side. This understanding can help teams and players alike to align their strategies with what is valued most in the current football landscape.

Methodology

The study employs advanced machine learning models, including Random Forest, Gradient Boosting (GBM), and Bayesian Additive Regression Trees (BART), to develop robust predictive models. These models are chosen for their ability to handle complex data structures and provide insights into the relative importance of various predictors. By implementing different machine learning models, I was able to compare their performance to determine the models that most effectively captured the nuances of wage determination; using that strategy for each position. This ensured the most accurate and insightful predictions were achieved in this study. Stepwise regression was also utilised to systematically identify the most relevant predictors, refining the model by focusing on the variables with the strongest influence on wage. Finally, Quantile Regression was applied to provide a more nuanced understanding across different wage levels. Unlike traditional regression models that focus on the average effect of predictors, Quantile Regression allows us to explore the "sources of heterogeneity in statistical relationships"(Koenker, 2005), or in our case, the examination of how these predictors influence different points in the wage distribution, such as the lower, median, and upper quantiles. This approach is particularly valuable for understanding variations in wage determination that may not be evident when only considering average outcomes, thereby offering a more comprehensive analysis of how performance metrics impact wages across the full spectrum of players.

Dissertation Structure

This dissertation is structured to first provide a comprehensive Literature Review in Chapter 2, discussing previous studies on the subject and analysing what questions they were exploring, their conclusions, and what data and methodologies they used. Chapter 3 entails an introduction to the data sources, what the data encompasses, how it was collected, and the initial data exploration and pre-processing steps. Following this, the Methodology section in Chapter 4 will detail the statistical

models and theoretical framework used in this study. In Chapter 5, the Results & Analyses section will present the key findings for each position, beginning with goalkeepers, followed by defenders, midfielders, and finally forwards; discussing the predictive power of the models across different player positions and the top predictors in each model. Finally, in Chapter 6, the dissertation will conclude with a summary and discussion of our findings as well as suggestions for future research.

2. LITERATURE REVIEW

Research in other team sports, such as basketball, baseball, rugby, and American football, has made significant strides in developing performance profiles for individual players, largely due to the structured nature of these sports. The ability to break down the game into distinct plays allows for easier identification and measurement of individual contributions (Lago-Peñas et. al, 2011). In contrast, football presents unique challenges for performance analysis because of the sport's continuous flow. Fewer set plays and relatively low scoring make the isolation and measurement of individual performances much more difficult to achieve. Football is very dynamic, and much harder to quantify without the continuous stoppages of other sports. For example, a forward's contribution may not only be judged by goals scored but also by their movement away from the ball, positioning, and ability to create space for teammates— actions that are more challenging to measure precisely. As opposed to American Football, where each player's role is more specialised and discrete like a wide receiver's running routes, making it easier to track and quantify their contributions within each play. As a result, many studies have been more team-oriented, looking at the effects of formations and playing styles on winning. For example, González-Rodenas et al. (2020) investigated the combined effects of tactical and contextual indicators on achieving offensive penetration and scoring opportunities in English Premier League matches. Similarly, Lago-Peñas, Lago-Ballesteros, and Rey (2011) examined these challenges by comparing performance indicators between winning and losing teams in the UEFA Champions League. Their study highlighted that despite the clear difficulties in football performance analysis, it is still possible to identify key indicators in the sport, with their study concluding that winning teams exhibited significantly higher average values in total shots, shots on goal, effectiveness, passes, successful passes, and ball possession. Conversely, losing teams had significantly higher values in more negative indicators related to discipline, such as yellow and red cards (Lago-Peñas et. al, 2011).

Past research on wage determination in professional football has also examined the influence of both individual performance metrics and broader contextual factors. Bernd Frick's study (2011) demonstrated that player salaries in the German Bundesliga are strongly linked to individual performance indicators such as career games played, goals scored, and international appearances, emphasising the role of experience and on-field contributions in salary determination. Frick applied Ordinary Least Squares (OLS) regression and Random Effects models, which are effective for estimating average effects but may oversimplify the complex relationships in player

performance data by assuming linearity and independence. To address these limitations, Frick also employed quantile regression, which allowed him to examine how salary's relationship with player characteristics varies across different points in the salary distribution, thus providing a more nuanced understanding of pay and performance relationships (Frick, 2011). However, simply because they weren't recorded at the time, the metrics involved in the study were not as advanced as present-day statistics. Football has integrated data-driven approaches and created new metrics and data-gathering techniques exponentially over the past decade.

Bryson, Rossi, and Simmons' study (2014) investigated the wage differential between migrant and domestic football players in Italy's top two divisions over a seven-year period, using OLS models, Quantile Regression, and Oaxaca-Blinder decomposition techniques. They found that migrant players, particularly those from the EU, earned significantly higher wages than their domestic counterparts, even after accounting for individual productivity metrics. The study concluded that this wage premium is indicative of a "superstar" effect, where migrant players are valued not only for their superior talent but also for their greater popularity, which contributes to higher team performance and increased crowd attendance (Bryson, Rossi, & Simmons, 2014).

Similarly, Lucifora and Simmons' study (2003) explored the "superstar" effect in the Italian league using a Mincer-type salary model, another form of OLS. This model was useful for highlighting key trends, particularly in showing that top players command disproportionately high wages due to their ability to attract spectators and generate revenue. However, like other OLS-based approaches, this model struggled to capture nonlinear interactions between variables. Lucifora and Simmons controlled for team-specific influences using team fixed effects, but acknowledged the need for further research to explore other performance measures that could provide a more comprehensive understanding (Lucifora & Simmons, 2003). Most past studies on football salary determination have focused on broader factors, such as media presence and superstar status, which can detract from what happens on the pitch.

There have, however, been a few recent studies that's main focus was on performance metric's influence on salary. Berri et. al (2023) used a Mincer-type salary model, focused on analysing goalkeeper pay in Europe's top five leagues, a scope similar to that of this study. They concluded that clubs primarily use basic defensive statistics to determine goalkeeper wages, focusing on team results and co-production rather than individual workload of the goalkeeper. Additionally, they noted that ball distribution skills positively impact salaries, highlighting the importance of goalkeepers in initiating the build-up to offensive moves. The study concluded that clubs could improve their evaluation of goalkeepers by incorporating more advanced metrics, suggesting that current approaches by clubs may undervalue individual contributions.

Yaldo and Shamir's study (2017) used data from sofifa.com, with player attributes ranked on a 0-99 scale to quantify abilities to determine wages through pattern recognition algorithms. The attributes used were particularly interesting with aggression, reactions, position, and vision all examples of the statistics incorporated in the study rather than traditional statistics like goals and

assists. This is because these attributes and their ratings were developed by football experts for both video game simulations and real-world scouting; providing a standardised measure for consistent comparisons across the dataset of 6,082 players (Yaldo, Shamir, 2017). The study found a strong Pearson correlation (~0.77) between predicted and actual salaries, demonstrating the potential of this method in wage negotiations. Unlike this study, Yaldo and Shamir included data of players from 91 different leagues, rather than focusing on a smaller subset of leagues. Overall, they concluded that while their proposed method of determining football player wages based on performance and skills offers a more objective and quantitative approach, it does not fully account for the non-performance-related factors that often inflate salaries, particularly for superstars. They observed that football superstars' salaries are not always proportional to their on-field performance, highlighting the influence of factors such as marketability and the "superstar effect" on wage determination (Bryson al., 2014).

James Liu (2023) addressed the gap in recent studies investigating on-pitch performance data for outfield players by using stepwise regression to analyse how different European leagues reward certain performance metrics. Liu's study focused only on forwards and midfielders and was more concerned with differences between leagues rather than identifying the overall most important wage predictors in Europe's top five leagues. Liu's methodology involved three rounds of regression analysis. Initially, he identified general variables like minutes played, height, and weight as strongly correlated with salaries. However, seeking more specific insights into the relationship between on-pitch performance and wages, Liu progressively refined his data by removing various factors and focusing on position-specific comparisons. This adjustment led to more significant findings, with R-squared values surpassing 0.8 across different leagues, and highlighted expected assists (xA) as a particularly strong predictor of midfielder salaries (Liu, 2023). However, a key limitation of Liu's study is the assumption that removing certain general factors would lead to more meaningful insights, which might have overlooked the combined impact of these variables in salary determination when analysed alongside position-specific metrics.

These variations in findings suggest that while team-level metrics and superstar status have been the focus of much of the existing literature, they can be highly situational and dependent on external factors like the style of play or the quality of opposition. The relationship between performance indicators and team success in football remains complex and often yields conflicting results. For instance, while Hughes and Franks (2005) found that successful teams in the 1990 World Cup were better at converting possession into shots on goal, Hughes and Churchill (2005) found no significant differences in play patterns between successful and unsuccessful teams during the 2001 Copa America. Such contradictions underscore the challenges of using team-based indicators as consistent predictors of success or salary.

This study shifts the focus to individual player performance metrics, which, despite being influenced by situational factors, offer a more consistent and objective basis for analysing wage determination. Today, it is easier to detail the impact players are making on the pitch. Advanced

metrics have played an instrumental role in making more informed decisions regarding player contracts and salary negotiations. Past studies on the relationship between performance and salary have been limited by the availability of metrics at the time. With the advancement of big data creating access to more in-depth metrics, current studies can now focus on the exact statistics that truly make a difference. For example, Declan Rice, a defensive midfielder for Arsenal in the English Premier League, averaged 0.11 "defensive actions that led to a shot attempt" per 90 mins, placing him in the 87th percentile among players in his position (FBref, 2024). He also averaged 7.79 progressive passes per 90 minutes, placing him in the 90th percentile and demonstrating his great ability to advance the ball up the field to the attackers (FBref, 2024); a critical contribution that teams highly value, as it directly influences their ability to create scoring opportunities and maintain possession (Liu, 2023). These detailed statistics did not exist in the past. This underscores the vital role that advanced metrics play in pinpointing the specific on-field contributions that influence a player's value (Liu, 2023). Teams can make more informed decisions about a player's wages by effectively analysing these important actions, ensuring that compensation truly reflects a player's impact on the pitch.

Overall, the existing literature on wage determination in professional football has provided valuable insights into the roles of individual performance, superstar status, and contextual factors such as media presence. This dissertation addresses the gap in studies that focus on performance metrics in europe's top five leagues by employing advanced machine learning and traditional regression models to analyse individual performance metrics across multiple positions in Europe's top five football leagues. By focusing on the specific, measurable actions that contribute directly to a player's value, this study provides a more nuanced and comprehensive understanding of wage determination in the modern football environment.

3. DATA

3.1 Data Sources

The data for this study were obtained from two prominent football statistics web platforms, FBref.com and Transfermarkt.com. FBref is well-regarded for its comprehensive and detailed football statistics, offering a wide array of both traditional and advanced performance metrics. It also includes per 90 (minutes) data, which is crucial for understanding a player's efficiency relative to their playing time. Additionally, FBref provides player wage information, sourced from Capology.com. The site specifically has a big five European football leagues section for the English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, and French Ligue 1. Transfermarkt is another well-renowned football statistics platform, known for its extensive coverage of player transfer values, contract details, and injury histories—offering detailed data on career injuries, which was vital in the study to assess the impact of player availability on wage

determination. Transfermarkt's data also includes club and player financials, making it a valuable resource for analysing the economic aspects of football.

3.2 The Data

The player performance metrics in the data encompassed a very broad range of variables, including traditional statistics like goals, assists, clean sheets, and passes completed, alongside advanced metrics such as expected goals (xG) and expected assisted goals (xAG). These metrics, sourced from FBref, are among the most crucial in modern football analytics as they offer a deeper understanding of a player's contribution beyond basic statistics.

Expected Goals (xG) is defined by FBref as the probability that a shot will result in a goal based on various factors such as the location of the shooter, the body part used, the type of pass, and the type of attack (FBref, 2024). This metric is invaluable in assessing the quality of scoring opportunities a player generates and what kind of positions he's able to put himself in to generate goal-scoring chances. Expected Assisted Goals (xAG) — as defined by FBref— specifically isolates the xG on passes that directly lead to a shot, reflecting a player's ability to set up scoring chances without relying on the outcome of the shot itself (FBref, 2024). xG and xAG are crucial for accurately capturing a player's goal contributions, offering a clearer reflection of their true impact on the game.

In addition to these advanced metrics, per 90 minutes statistics were also included. This metric normalises player performance statistics based on the time they spend on the pitch, allowing for fair comparisons between players who may have played different amounts of time. This approach ensures that the analysis reflects a player's efficiency and contributions relative to their playing time.

To provide further context in the analysis, we also included categorical variables such as age, nationality, and league. Age is an important factor as it can influence a player's experience and physical capabilities, which in turn can affect their performance and market value. Nationality was included to explore potential differences in player valuation across different nationalities, as cultural and marketability factors can influence wages. The league in which a player competes is also crucial, as different leagues may place varying levels of emphasis on certain attributes, or have more financial power, thus impacting player performance metrics and wages.

Wage data detailing annual salaries for players was integral to the analysis of the relationship between player performance and compensation, helping to establish connections between on-field contributions and financial rewards.

3.3 Data Collection

The data collection process involved several stages to ensure the comprehensive collection of all the relevant metrics. Initially, raw performance data from the 2023/24 season were extracted in CSV format across various categories using the 'worldfootballR' package in R, specifically through the

"fb_big5_advanced_season_stats" function (Zivkovic, 2024). This function is designed to retrieve detailed statistics for individual players or teams across the Big 5 European leagues. The "stat_type" argument in the function was used to specify the required performance categories. For outfield players, data were gathered across several categories, including 'standard', 'shooting', 'passing', 'passing_types', 'gca', 'defence', 'possession', 'playing_time', and 'misc'. For goalkeepers, more specialised categories, 'keepers' and 'keepers_adv', were employed to capture both basic and advanced goalkeeping statistics. Once the data frames were returned for each of these "stat_types", I then compiled the dataframes into one large dataset comprising all of the statistics. This approach ensured comprehensive data collection, covering all relevant metrics for subsequent analysis. Additionally, I employed web scraping techniques using the 'pandas' package in Python to gather all the wage data for players in our dataset, enabling the automation of data extraction directly from FBref. In addition to player performance metrics and wages, injury data was collected from Transfermarkt using 'worldfootballR' again. This allowed me to gather detailed records on the duration of total career injuries for each player prior to the 2023/24 season, as injuries during the latest season would not impact contracts that were already in place. These datasets were subsequently merged to create a unified database that integrated performance metrics with wage information. During the merging process, particular attention was paid to ensuring the alignment of data across player names, teams, and positions, which was crucial for maintaining consistency and accuracy. The result was a cleaned, comprehensive dataset ready for robust analyses to investigate the relationship between player performance and wages. Finally, I split the goalkeepers from the dataset, creating their own dataset as they had their own features that were relevant only to them.

3.4 Initial Data Summaries and Preprocessing

The initial phase of data analysis involved generating descriptive statistics to summarise the dataset and understand the distribution and range of key variables. Measures such as means, medians, standard deviations, and quartiles were calculated for each performance metric, providing insights into the central tendencies and variability within the data. This statistical summary was essential for identifying any potential outliers or anomalies that could influence the analysis.

Before applying the chosen methodology, the data underwent a series of preprocessing steps to ensure that it was ready for analysis. The wages data had distinct currencies for players in different leagues, so they were all converted to British pound sterling. During the pre-processing, I found the wages to be skewed. To normalise the data, I applied a logarithmic transformation to the annual wages variable. Variables that were not relevant to wages such as Season Year, Player, and Born (as we already had an Age variable) were excluded from all analysis and modelling. The categorical variables like Competition—containing the top five league the player played in—and Nationality were one hot-encoded allowing me to effectively incorporate them into the statistical models.

4. METHODOLOGY

4.1 Machine Learning Algorithms

I chose to use the Random Forest machine learning approach via the 'randomForest' package in R for this study due to its ability to handle large datasets that have complex interactions among the variables (Liaw & Wiener, 2002). As an ensemble learning method, the technique builds multiple decision trees and merges them to get a more accurate prediction. This is effective for managing non-linear relationships and capturing interactions between predictors, which are common in our final dataset which includes various types of statistics. Furthermore, Random Forest is also known for its resilience to overfitting, especially when dealing with noisy data by averaging multiple trees to smooth out the predictions. This is important for analysing football data, where a player's performance metrics can vary widely due to factors like fluctuations in form, injuries, or match-specific conditions, which can result in noise and outliers into the dataset.

In addition, I used the Gradient Boosting (GBM) machine learning technique for this study, through the 'gbm' package in R, due to its ability to optimise predictive accuracy through iterative learning (Ridgeway & Developers, 2024). GBM constructs models sequentially, with each iteration refining the errors of the previous one making it is well-suited to handle the relationships within football performance metrics, where nuanced improvements can make the difference. Additionally, GBM manages both linear and non-linear relationships effectively, which assists with the wide range of variables in our dataset that range from basic metrics like goals and assists to advanced statistics such as expected goals (xG) and expected assists (xA). This versatility ensures the model can adapt to the nature of football performance data, where the influence of different metrics is often not straightforward or linear.

Similarly, the Bayesian Additive Regression Trees (BART) method was used in this study—through the 'bartMachine' package in R(Kapelner & Bleich, 2016)—due to its ability to effectively capture complex, non-linear relationships in our data. BART functions by summing multiple regression trees, allowing it to model the intricate patterns that we often find in football performance metrics. This method is particularly useful in high-dimensional settings, where interactions between variables can significantly impact the results (Kapelner & Bleich, 2016). The package introduces several advanced features for our data analysis, such as variable selection, interaction detection, and model diagnostic plots, which were instrumental in this study for identifying the key predictors in wage determination. BART's underlying Bayesian probability model distinguishes it from other ensemble methods because it provides a framework that manages uncertainty and variability in the data well, enhancing predictive power (Kapelner & Bleich, 2016). Furthermore, the regularisation process prevents any single tree from dominating the model, ensuring balanced and reliable predictions (Kapelner & Bleich, 2016). This was a crucial aspect when analysing the complex relationships between performance metrics and wage valuation.

4.2 Regression Approaches

In addition to the machine learning techniques, traditional regression methods were employed to provide a complementary approach. Stepwise Regression, via the 'MASS' package in R, was chosen for its ability to identify the most significant predictors among a large set of variables, which is particularly useful in a dataset as complex as ours (Venables & Ripley, 2002). The method works by iteratively adding or removing predictors based on their statistical significance in explaining the variance in our dependent variable, player wages. This ability to simplify the model by retaining only the most impactful variables reduces multicollinearity and improves interpretability. This approach was especially valuable in narrowing down which specific performance metrics were most strongly associated with higher wages, allowing for a more focused analysis. It also provided a baseline model to compare with more complex machine learning approaches, helping to validate the results by highlighting consistent predictors across different methodologies.

Quantile Regression was included in the study through the 'quantreg' package in R to address the limitations of models that only estimate the average effects of predictors on the dependent variable (Koenker, 2023). Unlike traditional regression methods, Quantile Regression allows for examining the effects of predictors across different points in the wage distribution, such as the lower, median, and upper quantiles. This is particularly important in football, where the value of certain wage predictors can vary significantly between quantiles. P-values were instrumental in assessing which metrics had statistically significant effects at different quantiles, further enhancing the model's ability to provide a nuanced understanding of wage determination.

4.3 Application of Methodology

The machine learning models were trained on the dataset, using a carefully planned approach to split the data into training and test sets (80% and 20% respectively). Before training, the data underwent preprocessing, including dropping irrelevant columns, encoding categorical variables and applying log transformations to the annual wages to normalise the distribution and reduce skewness, ensuring that the models could accurately capture the relationships between the predictors and wages. Particular attention was given to the selection of hyperparameters, such as the number of trees in the Random Forest model and the burn-in periods in the BART model, to enhance their performance. I employed grid search techniques to explore combinations of hyperparameters to ensure the models were finely tuned for the data. By partitioning the data into multiple folds and iteratively training and validating the models, overfitting was minimised, and consistent performance across different subsets of the data was ensured.

The outputs were carefully interpreted using variable importance scores providing insights into which predictors had the most significant impact on wages. Diagnostic evaluations, including comparisons between predicted and observed wage values and assessments of model error patterns, were also used to gauge model performance. By visually comparing predicted wages with the actual observed values, I was able to highlight the models' accuracy, while the examination of error

distributions helped interpret whether the models accurately captured the underlying data patterns without systematic biases.

To assess the sensitivity and robustness of the models, analyses such as cross-validation were conducted to examine how variations in input variables affected the predictions. Finally, the results from all of the different models were integrated and compared. The alignment of key findings across the machine learning models and the traditional regression methods, such as the consistent identification of significant predictors, validated by statistical significance in various quartiles, reinforced the credibility of the conclusions drawn from the study. Any discrepancies were carefully analysed to understand the underlying reasons, contributing to a more nuanced interpretation of the wage determination process in professional football.

5. RESULTS & ANALYSES

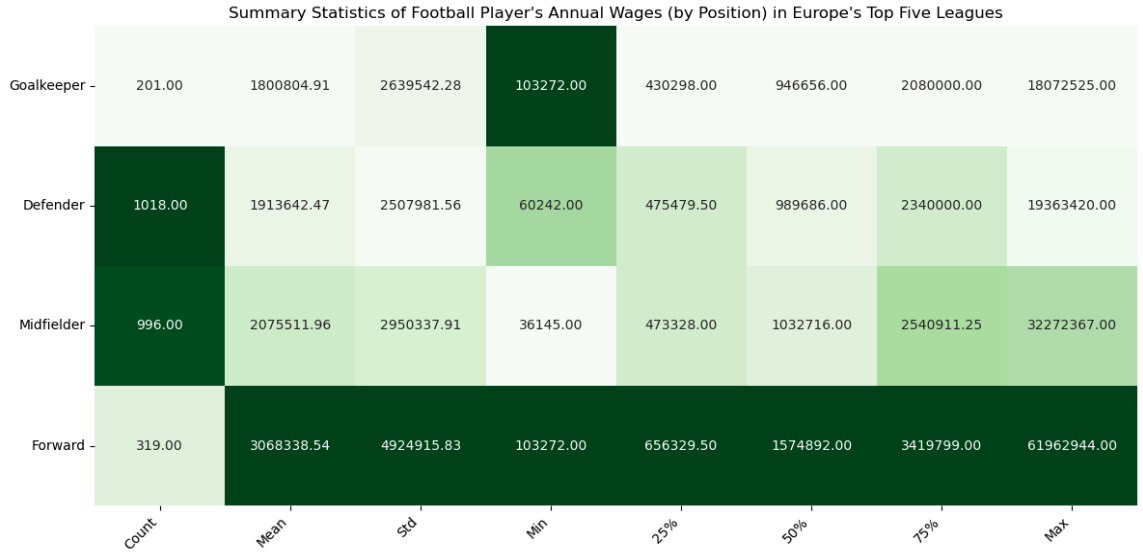

Table 1: Summary statistics of the annual wages of football players in our dataset, categorized by position. The statistics include the count of players, mean, standard deviation, minimum, various percentiles (25th, 50th, 75th), and the maximum annual wages; with darker shade representing higher values.

5.1 Statistical Overview of Dataset

Table 1 presents a summary of the annual wages of men's football players in Europe's top five leagues categorised by position, offering a preliminary glimpse into the distribution of wages across different roles on the pitch. As expected, the dataset includes a higher number of defenders and midfielders as teams have more players in these positions than forwards and goalkeepers due to fewer positions available within a team for those roles.

Notably, forwards lead in almost every wage-related category, including mean, median, and maximum annual wages. Fourteen of the top 25 wage-earners in the dataset were forwards. This trend is consistent with the pivotal role that forwards play in scoring goals, capturing headlines, and

attracting the most popularity and media attention. Their high visibility and goal-scoring contributions often translate into higher compensation, as clubs are willing to pay a premium for players who can deliver match-winning performances. Goals are without a doubt, the highlight of the sport. The most well-known and internationally recognized players, who are overwhelmingly goal-scoring forwards, significantly enhance the team's brand value, leading to increased profits from merchandise sales and broadcasting rights, while also attracting new sponsors and opening up additional investment opportunities (Bernardo, Ruberti, & Verona, 2022).

Conversely, goalkeepers, despite their critical role in preventing goals, tend to have lower wages on average, likely reflecting the lesser emphasis placed on defensive roles in the overall wage structure of football clubs.

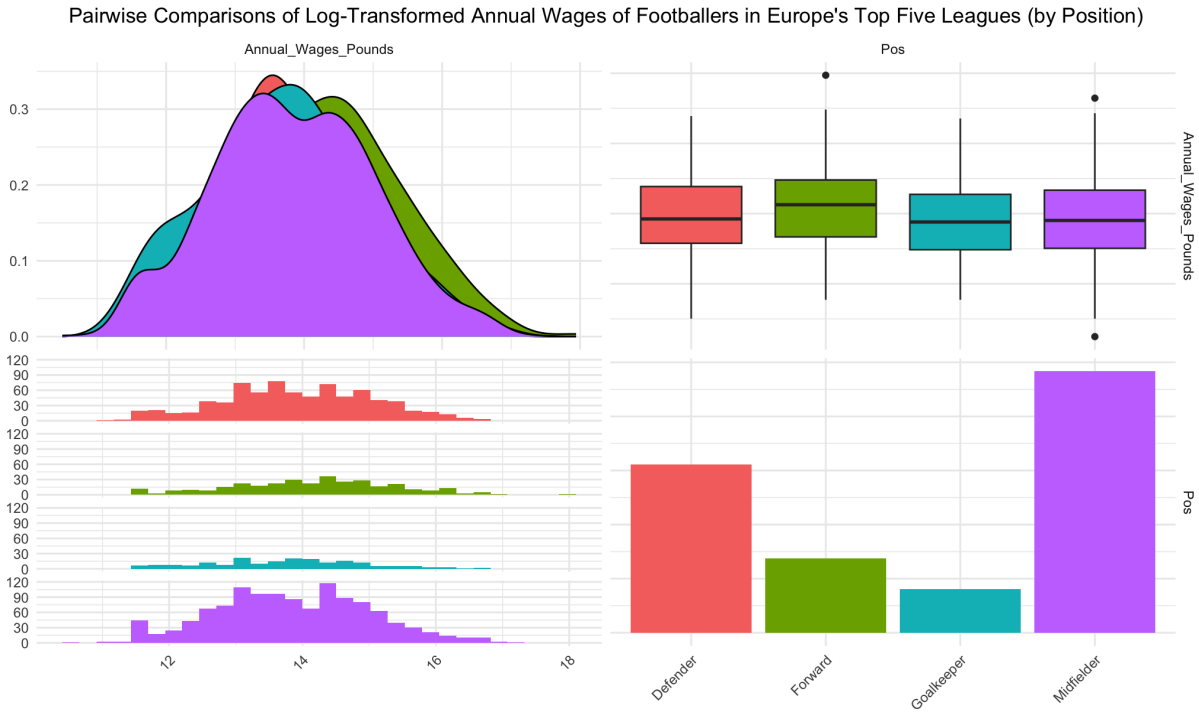

Figure 1: Pairwise comparisons of annual wages, log-transformed, of footballers in Europe's top five leagues by position. The density plots on the left depict the distribution of wages, highlighting the spread and skewness for each position, while the boxplots on the right compare the median and variability of wages across positions. The histograms below the density plots show the frequency distribution of these wages, and the bar plot further illustrates the count of players per position.

Figure 1 provides a detailed pairwise comparison of the log-transformed annual wages of footballers across Europe's top five leagues, categorised by position. The density plots on the left reveal how these wages are distributed, with midfielders showing a broad spread, indicating a wide range of salaries within this group. Forwards, on the other hand, display even higher variability in wages, as suggested by their higher standard deviation in the summary table. This is further confirmed by the boxplots on the right, where forwards, represented by green, have the highest median wages, followed by midfielders, represented by the purple. The boxplots also highlight the greater variability among forwards, reflected in their larger interquartile range. The histogram at the bottom of the density plots highlights the frequency distribution of these wages, reinforcing the observation that midfielders are the most represented group in the dataset, followed by defenders,

forwards, and goalkeepers. The wages of Kylian Mbappe, a forward for Real Madrid, can be spotted on the histogram as the outlier, earning 92% more annually than the second-highest earner in Europe, Frenkie De Jong. Finally, the bar plots provide a visual comparison of the size of each position group in our dataset.

The decision to log-transform the annual wages before generating this pairwise plot was essential for several reasons. Firstly, wage data in professional football tend to be right-skewed, as ours was, and log transformation helps to normalise this distribution, making the data more symmetric and thus more suitable for analysis. Secondly, the transformation compresses the scale of wages, making it easier to identify patterns for each position. Lastly, interpreting wages in log-transformed terms allows for a more meaningful analysis, giving us the ability to consider differences in relative (percentage) terms rather than just absolute values, which can be more insightful.

5.2 Goalkeepers

Overview

The analysis of goalkeeper wages in this study yielded a variety of results that highlight the evolving role of goalkeepers in modern football. The performance of the Random Forest, Gradient Boosting Machine (GBM), and Bayesian Additive Regression Trees (BART) models, alongside the stepwise and quantile regression models, provided a comprehensive view of the key factors that influence goalkeeper salaries. This section explores the performance of these models, the common features that consistently emerged as important predictors, and the broader implications for how goalkeepers are valued in today's game compared to traditional standards. Additionally, we will compare our findings with those of Berri et al.'s wage study (2023) on goalkeeper wages, and discuss the practical significance of the actual vs. predicted results and residuals observed in our models.

Model Performance and Interpretation

Each of the models used in this study provides insights into the determinants of goalkeeper wages, reflecting the complexity of the position in modern football. As seen in Table 2, the Random Forest model achieved an R^2 of 0.48, indicating a moderate level of accuracy in predicting wages based on the features included. The GBM and BART models both performed similarly, with R^2 values of 0.46 each, suggesting that while they captured some important aspects of goalkeeper performance, there were still unexplained variances.

The stepwise regression model outperformed the other models, achieving an R^2 of 0.56. This superior performance could be attributed to the model's ability to eliminate less predictive variables, allowing it to focus on the most relevant features. This selective feature inclusion process likely enabled noise reduction and enhanced predictive power. In contrast, the Random Forest and GBM models, which include a broader set of variables, may have been slightly hindered by the

inclusion of features that, while potentially important, did not contribute as much to the overall predictive accuracy. The quantile regression models offered valuable insights by highlighting how different features influenced wages across the distribution of salaries. As can be seen in Table 2, the R^2 values for these models were 0.23 at Tau = 0.25, 0.52 at Tau = 0.5, and 0.28 at Tau = 0.75. These values suggest that the model's features better explain wage variability at the median quantile (Tau = 0.5) compared to the lower and upper quantiles. This finding indicates that while certain predictors are particularly relevant for goalkeepers earning median wages, other factors might be more influential at the relative extremes of the salary spectrum, where wage determination can be more complex and influenced by unobserved variables. The varying model fit across quantiles underscores the importance of considering different segments of the wage distribution when analysing salary determinants in professional football.

Goalkeepers: Key Predictors and Performance Insights Across Multiple Models

Feature	Random Forest	BART	Gradient Boosting	Stepwise Regression	Quantile Regression (Tau = 0.25)	Quantile Regression (Tau = 0.5)	Quantile Regression (Tau = 0.75)
R^2	0.48	0.46	0.46	0.56	0.23	0.52	0.28
Top Feature 1	Wins	Wins	Wins	Premier League	Ligue 1	Premier League	Post-Shot Expected Goals per Shot on Target (PSxG/SoT)
Top Feature 2	Passes Attempted (GK)	Age	Age	Post-Shot Expected Goals per Shot on Target (PSxG/SoT)	Premier League	Post-Shot Expected Goals per Shot on Target (PSxG/SoT)	Premier League
Top Feature 3	Age	Premier League	Launch % (Goal Kicks)	Bundesliga	Post-Shot Expected Goals Minus Goals Allowed (PSxG-GA)	Post-Shot Expected Goals Minus Goals Allowed (PSxG-GA)	Ligue 1
Top Feature 4	Throws Attempted	Avg. Length of Goal Kicks	Premier League	Ligue 1	Post-Shot Expected Goals (PSxG)	Bundesliga	Defensive Actions Outside Penalty Area per 90 (OPA/90)
Top Feature 5	Premier League	Clean Sheet %	Pass Completion % (Launched)	Wins	Post-Shot Expected Goals per Shot on Target (PSxG/SoT)	Ligue 1	Free Kick Goals Against
Top Feature 6	Pass Completion % (Launched)	Launch % (Goal Kicks)	Passes Attempted (GK)	Age	Wins	Post-Shot Expected Goals (PSxG)	Clean Sheets
Top Feature 7	Minutes per 90	Passes Attempted (GK)	Shots on Target Against	Launch % (Goal Kicks)	Shots on Target Against	Shots on Target Against	Post-Shot Expected Goals Minus Goals Allowed (PSxG-GA)
Top Feature 8	Launch % (Goal Kicks)	Shots on Target Against	Defensive Actions Outside Penalty Area	Save %	Saves	Saves	Losses
Top Feature 9	Avg. Length of Goal Kicks	Goals Against	Clean Sheets	Passes Completed (Launched)	Draws	Free Kick Goals Against	Draws
Top Feature 10	Avg. Length of Passes	Clean Sheets	Clean Sheet %	Launch % (Passes)	Bundesliga	Minutes per 90	Age

Table 2: Summarises model performance, via R-Squared, and top predictors of goalkeeper wages across different models and quantile regression analyses. Colour coding is used to visually group predictors in other models.

Top Predictors

In Table 2, several features consistently emerged across multiple models as key predictors of goalkeeper wages, including "Post-Shot Expected Goals per Shot on Target (PSxG/SoT)," "Premier League," "Ligue 1," and "Clean Sheets", and ball distribution related statistics. The significance of these predictors varied across the different quantiles, highlighting their nuanced impact on goalkeeper wages.

PSxG/SoT (Post-Shot Expected Goals per Shot on Target) measures the difficulty of the shots a goalkeeper faces, with higher values indicating that the shots on target are harder to stop and

more likely to result in a goal (FBref, 2024). PSxG/SoT (Post-Shot Expected Goals per Shot on Target) consistently appeared across the models, particularly in the median quantile (Tau = 0.5), with a negative coefficient, although it was not statistically significant as the p-value in Table A1 indicates. This negative association suggests that goalkeepers facing more difficult shots (higher PSxG/SoT) might be earning lower wages, potentially reflecting the increased challenge of their role. By facing more difficult shots, this means the team is conceding dangerous chances frequently, indicating poor defensive performance. It can be insinuated the team the goalkeeper plays for may not be a top club. Additionally, "Post-Shot Expected Goals - Goals Allowed" (PSxG-GA) emerged as an important indicator across the three quantiles of the quantile regression models. This metric, which represents the difference between the quality of shots a goalkeeper faces and the actual goals allowed, provides insight into a goalkeeper's shot-stopping ability (FBref, 2024). A higher PSxG-GA value suggests that the goalkeeper is allowing fewer goals than expected based on the difficulty of the shots they face, highlighting their effectiveness. Overall, the frequent inclusion of these metrics across quantiles underscores their growing importance as indicators of a goalkeeper's effectiveness in the modern game.

Team statistics like Clean Sheets, Losses, Draws, and Wins emerged as frequent predictors throughout all of the models, with Clean Sheets being a statistically significant predictor in the upper quantile (Tau = 0.75). This indicates overall that these traditional measures of goalkeeping success still play a crucial role in determining higher wages, especially for top-earning goalkeepers. This finding is consistent with Berri's study, which highlighted that goalkeepers are often rewarded based on team outcomes and traditional defensive statistics, rather than solely on individual performance metrics (Berri et. al, 2023). This alignment underscores the ongoing relevance of these established indicators in wage determination for goalkeepers, particularly those at the higher end of the wage spectrum.

Several ball-playing features, such as Passes Attempted (GK), Passes Completed (GK), Launch % and Pass Completion % (Launched), emerged as key predictors in some of the models, reflecting the growing importance of a goalkeeper's distribution abilities. Launch % measures the percentage of passes longer than 40 yards (excluding goal kicks), while Pass Completion Percentage (Launched) indicates the accuracy of these long passes (FBref, 2024). In today's game, proper distribution is crucial, as goalkeepers are increasingly expected to contribute to the team's build-up play rather than merely clearing the ball upfield, a more traditional approach. These metrics were particularly significant in specific quantiles, such as Pass Completion % (Launched), which was statistically significant at (Tau = 0.5), highlighting the value placed on precise long-ball distribution in leagues where playing out from the back is a key strategy. The significance of these features underscores the evolving role of goalkeepers, who are now required not only to stop shots but also to maintain possession and accurately initiate attacks.

The Premier League was found as a top predictor in every model and positively associated with wages across multiple quantiles and was statistically significant, especially in the median (Tau

= 0.5) and upper quantiles (Tau = 0.75), as seen in Table A1. This finding indicates that playing in the Premier League significantly boosts a goalkeeper's market value, likely due to the higher level of competition and visibility in this league. This aligns with the data as Premier League goalkeepers make up ten of the top 20 highest goalkeeper wages in Europe. Conversely, Ligue 1 was negatively associated with wages across the quantiles, with statistical significance at Tau = 0.25, suggesting that goalkeepers in Ligue 1 may command lower wages compared to those in other top leagues, likely due to lower overall revenue and media exposure.

Age remained a crucial factor in the upper quantile and in our Random Forest, GBM, and BART models; showing a positive association with wages, reflecting its traditional role in assessing a player's experience and potential longevity. The inclusion of age alongside modern metrics like PSxG/SoT and league-specific impacts indicates a broader and more sophisticated understanding of what makes a goalkeeper valuable in today's football landscape.

The varying R^2 values across quantiles further suggest that while certain predictors are particularly influential for goalkeepers with median wages, the dynamics at the lower and upper ends of the wage spectrum may be influenced strongly by additional, unobserved factors. This highlights the importance of considering both traditional and modern performance metrics when evaluating goalkeeper wages.

The Impact of the Back-Pass Law on Goalkeeping

The transformation of the goalkeeper's role in modern football can be traced back to a pivotal moment in 1992, when the back-pass law was introduced. This rule change, which prevented goalkeepers from handling the ball after a deliberate pass from a teammate, fundamentally altered the position and its demands. As Cox (2018) notes, this change forced goalkeepers, and subsequently other positions like defenders, to broaden their skill sets, evolving from mere shot-stoppers into all-around players who are integral to their team's passing moves. This evolution is clearly reflected in our models, where distribution metrics like "Passes Attempted (GK)," "Launch %," and "Pass Completion % (Launched)" emerged as significant predictors of wages. Specifically, Pass Completion % (Launched) was statistically significant in the median quantile (Tau = 0.5), emphasising the importance of accurate long-ball distribution in today's game.

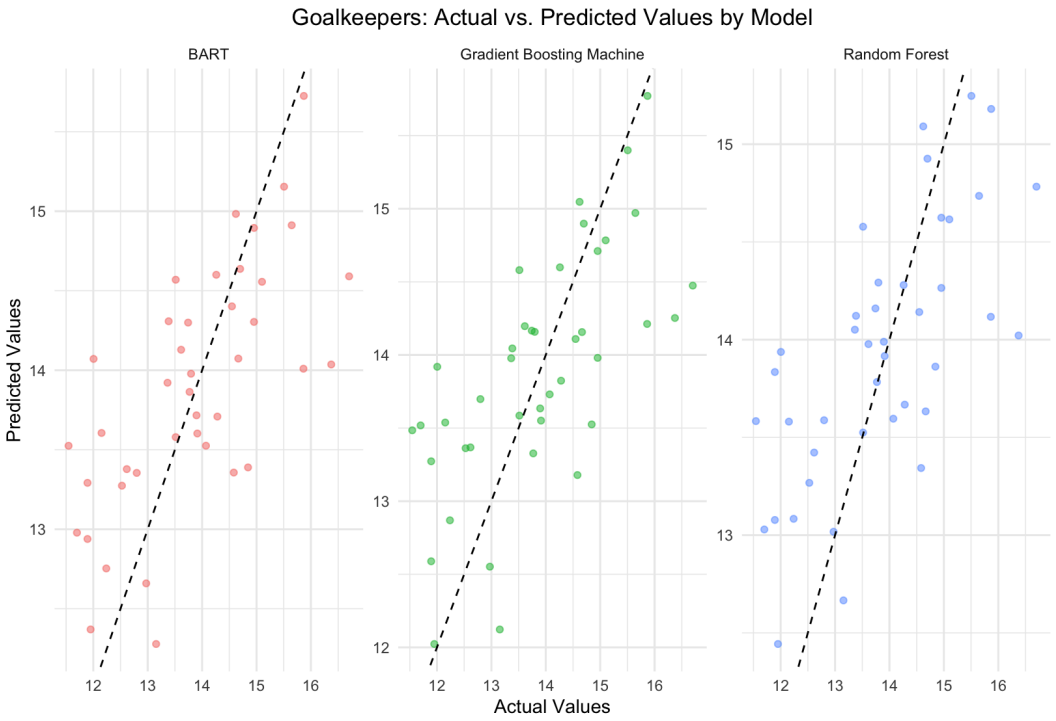

Figure 2: This plot compares actual and predicted goalkeeper wages across three models: BART, GBM, and Random Forest. The red dashed line represents perfect predictions. Points closer to the line indicate better model accuracy, while deviations show prediction errors.

Actual vs. Predicted

The actual vs. predicted value plots for our models in Figure 2 reveal varying degrees of accuracy across the different models. The GBM model demonstrated the most consistent performance, with predictions closely clustered around the diagonal line, indicating that it effectively captured the general trend in goalkeeper wages. However, some dispersion around the line suggests it struggled with more nuanced aspects of the data. The Random Forest model, while generally aligning with the trend, showed clear scatter. This suggests that Random Forest, although capturing the overall pattern, may miss key details to wage prediction. The BART model showed considerable variability, especially at lower actual wage values, reflecting the challenges of working with complex data and the evolving role of goalkeepers. These patterns underscore that while the models are useful for identifying broad trends, they may not fully account for the complexities and nuances that influence goalkeeper wages, highlighting the limitations of relying solely on performance metrics.

Residuals

In Figure A1, The BART model exhibits some clustering of residuals around the zero line, but with noticeable spread, particularly at the higher fitted values. This suggests that while BART generally captures the trend, it struggles with extreme predictions, leading to some bias in it's results. The GBM model, on the other hand, shows a relatively even spread of residuals, with higher variability particularly in the middle range of fitted values. The Random Forest model displays the most

significant variability, with a wide scatter of residuals, and less clustering around the zero line relative to the other two models.

Conclusion

The expanded role of goalkeepers in modern football has made it increasingly challenging to assess their value using traditional metrics alone. Goalkeepers are now integral to building play from the back and initiating attacks, meaning their contributions extend far beyond just making saves and team results like keeping clean sheets. This shift in responsibilities may explain why some of our models struggled to predict wages accurately, particularly for goalkeepers who excel in these modern aspects of the game but may not necessarily stand out in traditional metrics. The significance of ball-playing features, such as Pass Completion % (Launched) and Launch %, in our analysis underscores the need to consider a wider range of performance indicators when evaluating the true worth of a modern goalkeeper.

Our findings align with and expand upon Berri et al.'s (2023) study, which highlighted that elite European clubs often rely on primitive defensive statistics to determine goalkeeper pay, potentially undervaluing the multifaceted role goalkeepers now play and overvaluing team results. While our approach incorporated a broader set of advanced metrics that reflect these evolving responsibilities, the unexplained variance in wages in our models suggests that clubs may still not fully capture the true value of these contributions. This echoes Berri et al.'s (2023) conclusion that clubs could improve their valuation processes by integrating a more comprehensive set of performance data. In this evolving landscape, traditional metrics alone may no longer suffice, and clubs could benefit from incorporating a broader set of advanced statistics into their valuation processes. Enhancing the accuracy of wage predictions by considering both traditional and modern performance metrics will better reflect the complex, evolving role of the modern goalkeeper and ensure their value is appropriately recognized in today's game.

5.3 Defenders

Overview

In the analysis of defender wages, we applied the same modelling approaches used for goalkeepers, including Random Forest, Gradient Boosting Machine (GBM), Bayesian Additive Regression Trees (BART), Stepwise Regression, and Quantile Regression models. The outputs, depicted in the table and graphical representations, reveal insights into which factors were found to be the most influential in determining the wages of defenders in professional football.

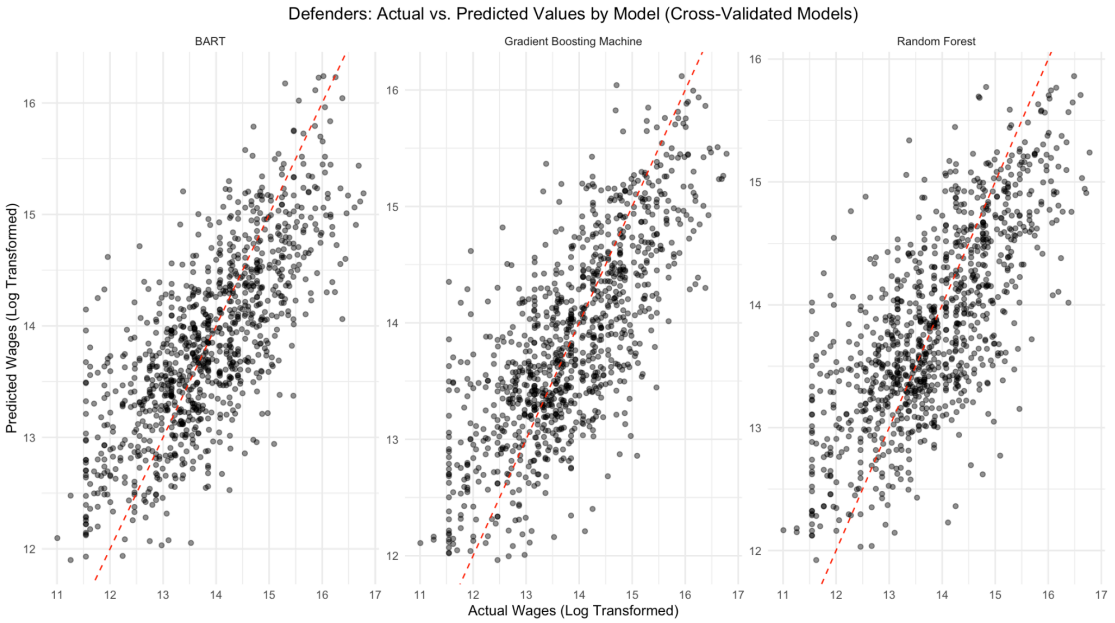

Figure 3: This plot compares actual and predicted defender wages across three models: BART, GBM, and Random Forest. The red dashed line represents perfect predictions. Points closer to the line indicate better model accuracy, while deviations show prediction errors.

Model Performance and Interpretation

The Random Forest model yielded the highest R^2 value at 0.81, suggesting it had the strongest fit among the models in capturing the variability in defender wages. This strong performance might be attributed to Random Forest's robustness in handling large sets of data with many variables, effectively capturing the non-linear interactions between predictors. However, the Actual vs. Predicted plot in Figure 3 indicates clear scatter, particularly at the extremes of the wage spectrum, suggesting that while Random Forest captures the overall trend, it may struggle with precision underestimating at higher and overestimating at lower wage levels. The GBM model followed with an R^2 of 0.71. Despite this slightly lower R^2, GBM's iterative improvement process makes it a strong performer, particularly in the mid-range of wages. Figure 3 shows GBM's predictions are more consistently aligned with the diagonal line, indicating better accuracy across the board, though there is still some variability at higher wage levels. The BART model, with an R^2 of 0.65, offers a different perspective. BART's flexibility and Bayesian approach allow for a nuanced interpretation, but Figure 3 reveals that the model began to underestimate wages as they got higher. This may be due to BART's regularisation, which prevents overfitting but can lead to underestimation of higher earners.

The Stepwise Regression model produced an R^2 of 0.57, which, although lower than the other models, is valuable for its interpretability. The model's selection of significant predictors highlights the most impactful variables without the complexity of non-linear interactions making it useful in understanding the fundamental drivers of defender's wages. The quantile regression models, with R^2 values of 0.17, 0.33, and 0.17 for the lower, median, and upper quantiles

respectively, indicates that these models struggled to capture the variability in wages across different wage levels.

Defenders: Key Predictors and Performance Insights Across Multiple Models

Feature	Random Forest	BART	Gradient Boosting	Stepwise Regression	Quantile Regression (Tau = 0.25)	Quantile Regression (Tau = 0.5)	Quantile Regression (Tau = 0.75)
R^2	0.81	0.65	0.71	0.57	0.17	0.33	0.17
Top Feature 1	Plus/Minus Contribution to Team Success	Age	Age	Age	Age	Competition: Premier League	Plus/Minus Contribution to Team Success
Top Feature 2	Age	Total Career Games Missed	Plus/Minus Contribution to Team Success	Competition: Premier League	Competition: Premier League	Age	Age
Top Feature 3	Competition: Premier League	Plus/Minus Contribution to Team Success	Competition: Premier League	Competition: Ligue 1	Plus/Minus Contribution to Team Success	Plus/Minus Contribution to Team Success	Competition: Premier League
Top Feature 4	Plus/Minus Contribution per 90 Minutes	Plus/Minus Contribution per 90 Minutes	Total Career Games Missed	Plus/Minus Contribution to Team Success	Competition: Ligue 1	Competition: Ligue 1	Pass Completion %
Top Feature 5	Total Career Games Missed	Pass Completion %	Plus/Minus Contribution per 90 Minutes	Nationality: Germany	Goals	Plus/Minus Contribution per 90 Minutes	Total Career Games Missed
Top Feature 6	Total Career Injuries Duration	Total Career Injuries Duration	Pass Completion %	Pass Completion %	Total Career Games Missed	Total Career Games Missed	Competition: Ligue 1
Top Feature 7	Pass Completion %	Competition: Premier League	Total Career Injuries Duration	Total Career Games Missed	Total Career Injuries Duration	Pass Completion %	Total Career Injuries Duration
Top Feature 8	Long Pass Completion %	Competition: Ligue 1	Total Passes Attempted	Total Passes Attempted	Competition: La Liga	Total Career Injuries Duration	Progressive Passes
Top Feature 9	Total Passes Attempted	Nationality: Brazil	Long Pass Completion %	Nationality: Spain	Aerial Duel Win %	Progressive Passes	Long Pass Completion %
Top Feature 10	Aerial Duel Win %	Expected Goals + Expected Assisted Goals Per 90 Minutes	Competition: Ligue 1	Nationality: Colombia	Pass Completion %	Aerial Duel Win %	Plus/Minus Contribution per 90 Minutes

Table 3: Summarises model performance & top predictors of Defender wages across different models and quantile regression analyses. Colour coding is used to visually group predictors in other models.

Top Predictors

The consistency of specific features appearing across all models in Table 3 signals their importance in wage determination. Age is a standout predictor, showing up as a top factor in all models, and statistically significant in all quantiles of our quantile regression model, as seen in Table A2. Age is often correlated with experience, physical prime, and career longevity, all of which are to be taken into account in the valuation of a defender. Younger defenders might be seen as prospects with potential growth, whereas those in their late twenties to early thirties are likely to be hitting their peak, commanding higher wages due to their reliability. In contrast, older defenders might see a decline in their wages as they approach the latter stages of their careers with their reduced physical capabilities.

The “Plus/Minus (+-) Contribution to Team Success” metric also features prominently, appearing no lower than fourth in all of the model’s top features. This statistic reflects a defender’s effectiveness in contributing to the overall success of the team, which is directly linked to the player’s value. Defenders who consistently help their teams maintain positive goal differences are more likely to be seen as assets that are worth investing in, justifying higher wages. This metric is particularly relevant in modern football, where defenders are not only tasked with preventing goals but are also integral to the team’s build-up play and overall tactical setup.

"Pass Completion %" and "Long Pass Completion %" are additional key predictors that align with modern football's tactical evolution. In the early 1990s, players had clearly defined roles—defenders focused solely on defence, and attackers on offence. However, over time, these roles have expanded, with defenders now initiating attacks and attackers contributing to defensive efforts (Cox, 2018). With the increasing emphasis on playing out from the back and ball retention, defenders are expected to be proficient with the ball, making passing ability a crucial aspect of their worth. The presence of these passing metrics across multiple models highlights how the game has evolved, with technical proficiency now as valued as the traditional defensive skills. A defender with a high pass completion rate is likely seen as reliable in maintaining possession and initiating attacks, which are highly valued attributes in modern football tactics that emphasise building from the back.

The league the player plays in is clearly a crucial factor. The Premier League proved statistically significant in all 3 quantiles in Table A2. As the most financially lucrative league and one of the most competitive globally, wages are naturally inflated for the players in it. Broadcasting rights alone soared from £51 million per season between 1992 and 1997 to an astounding £2.75 billion per season by 2016—a fifty-fold increase (Cox, 2018). Defenders performing at this level are often compensated more due to the higher stakes, visibility, and the financial muscle of Premier League clubs. In contrast, Ligue 1, the premier French league, has a negative coefficient in our quantile regression models and is statistically significant (Table A2). This suggests the league is not financially powerful relative to the other leagues and is unable to provide high wages to players. Interestingly, Paris Saint Germain, the top club in Ligue 1 for the last 10 years, has the highest wage bill in all of Europe, with being state-sponsored by Qatar. Four of their defenders are in the top twelve highest paid defenders in world football. The club however has a monopoly over the league, and no other club comes close to their financial backing. Overall, top leagues attract more viewership, sponsorship deals, and thus, higher revenues, enabling clubs to offer substantial wages to secure top talents.

"Total Career Games Missed" and "Total Career Injuries Duration" are both important predictors in all positions and prove to be here in the models for defenders. Career Injuries Duration reflects either frequency in nagging injuries or massive injuries that could change the physical capabilities of a player. For example, a study in 2016 concluded that only 65% of professional football players with ACL reconstruction for a total rupture maintained the ability to play at the same level 3 years after returning in professional football (Waldén, Hägglund, Magnusson, Ekstrand, 2016). On the other hand, "Total Career Games Missed" directly quantifies the impact of injuries on a player's availability, specifically highlighting how those absences affected a player's participation in matches, providing a clearer picture of their reliability and consistency. These factors are particularly important for defenders, as the physical demands of the position mean that clubs are likely to be wary of investing heavily in players with a history of frequent or long-term injuries. A defender with a clean injury record is more likely to be available for selection

throughout the season, making them a more valuable asset and justifying higher wages. Conversely, a history of missed games due to injuries could lower a player's market value due to the potential risk of re-injury.

"Aerial Dual %" appears in three of our models, as the 9th highest predictor in one, and 10th in the other two. In previous years, this metric would likely have been even more prominent, as aerial dominance was a crucial aspect of defensive play. However, as modern football has evolved, passing ability has become increasingly important for the defenders and even preferred in some systems.

Quantile Regression Insights

As shown in Table A2, the top predictors of our quantile regression have varying influence across different wage levels. The impact of age is significant across all quantiles, yet its influence varies depending on wage levels. In lower quantiles, where wages are generally lower, younger defenders might be viewed as rising talents with future potential, which could justify moderate wages. In contrast, in higher quantiles, age is strongly correlated with higher wages, reflecting the premium placed on experience and peak physical condition. This differentiation highlights that while age is a universal factor, its impact on wages is more pronounced at the higher end of the wage spectrum, where seasoned defenders are being rewarded. The Plus/Minus Contribution to Team Success metric is statistically significant across all quantiles, which aligns with what would be expected. The metric highlights the importance of a player's consistent impact on team performance at every wage level. The competitive league factor also shows varying effects across quantiles. While playing in the Premier League is a significant positive predictor across all quantiles, its impact is especially strong at higher wage levels. Premier League defenders make up 42.93% of the top 25% of wage-earners among defenders in Europe's top 5 leagues. Conversely, Ligue 1 shows a negative coefficient, particularly in the lower and middle quantiles, indicating that while it is a top European league, it does not offer the same wage premiums as leagues like the Premier League— only making up 9.09% of the top 25% of wage-earners among defenders in Europe's top 5 leagues. This difference reflects the financial disparities between these leagues and the varying market values attributed to players within them. Pass Completion % and Long Pass Completion % are consistently relevant across quantiles, proving their importance in modern football tactics. These metrics are not only valued just at the top end of the wage scale but also at lower and middle levels, where players who can maintain possession and contribute to build-up play are valuable assets in a team's overall strategy.

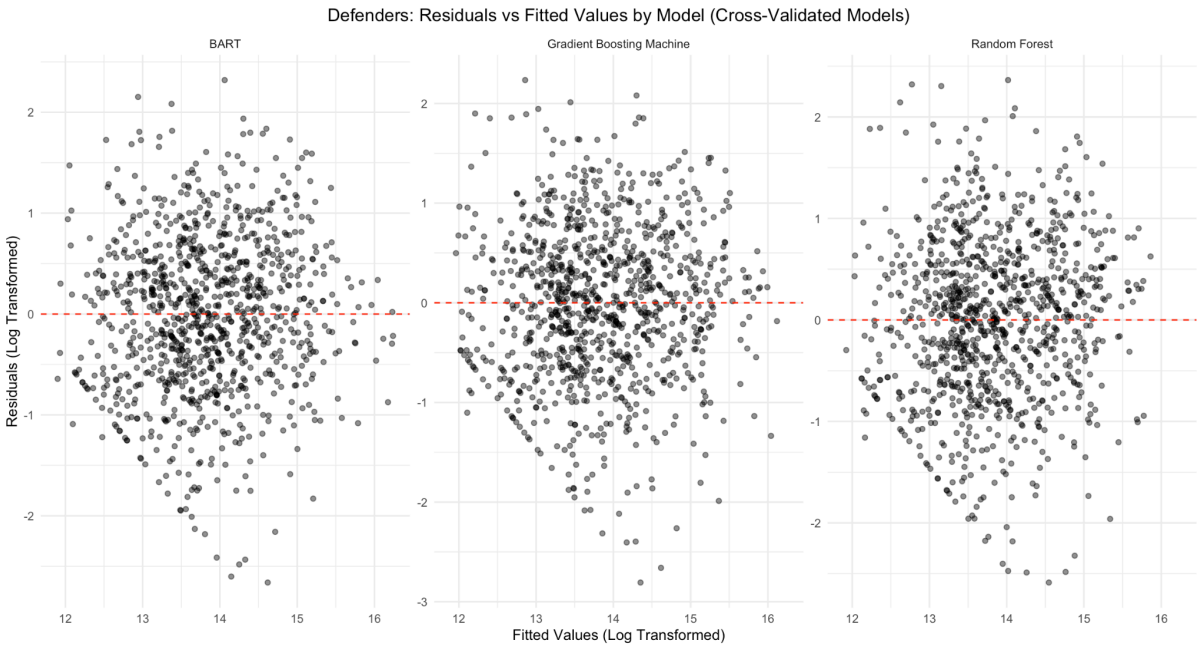

Figure 4: Residuals versus fitted values plotted for defenders for our machine learning models: BART, Gradient Boosting Machine, and Random Forest. These plots illustrate the distribution of residuals (log-transformed) against the predicted values (log-transformed), highlighting the degree of variance and any potential patterns or biases in the model predictions. The red dashed line represents the zero residual line, which indicates perfect prediction. The spread and symmetry of the residuals around this line suggest the accuracy and consistency of our model predictions.

Residual Analysis

The residuals vs. fitted values plots in Figure 4 provide further insights into model performance. The Random Forest model shows residuals fairly evenly spread across the range of fitted values, though there is a noticeable angled pattern at the lower end, suggesting some potential model misspecification. GBM also displays a consistent spread of residuals, with a similar angled pattern at the bottom, indicating that while it captures the general trend well, it may still face challenges with predictions. BART's residuals plot reveals more clustering around zero, and it also shows this angled pattern which might be because of its regularisation effect.

Conclusion

In conclusion, the analysis reveals that team contribution metrics, passing abilities, age, and the league the defenders are playing in, are consistently influential in determining defender wages across various models. Each model provided unique insights. The residuals analysis suggests that while these models are indeed effective, they each have limitations. The inclusion of quantile regression added another layer of understanding for defenders, showing how the impact of these predictors varies across different wage levels.

5.4 Midfielders

Overview

In the analysis of Midfielder wages, we employed the same model approaches. Our results, depicted in the table and figures, reveal insights into the model performances and which factors are most influential in determining the wages of midfielders in professional football.

Figure 5: Actual versus predicted wages for midfielders using Gradient Boosting, Random Forest, and BART models. The plots display how well each model's predictions align with the actual log-transformed annual wages of midfielders.

Model Performance and Interpretation

The performance of the models for midfielders was varied, with the GBM and BART models both achieving an R^2 of 0.50, while the Random Forest model slightly outperformed them with an R^2 of 0.51. These R^2 values suggest that the models had moderate predictive power, capturing approximately half of the variance in wages among midfielders. This level of performance is expected in a highly complex and diverse position group like midfielders. The actual vs. predicted plots in Figure 5 further illustrate this, showing that while the models generally captured the trend, there were instances of over- and under-prediction. The Random Forest model seems to have underestimated the lower wages whereas the GBM model fits the general trend better. However, the GBM model displays more variability than the Random Forest especially at the higher wage levels. The BART model shows clustering around the mid-level predictions, possibly reflecting its tendency to make more conservative predictions due to its Bayesian framework and regularisation

process, both of which contribute to a cautious approach when dealing with uncertain values. The moderate R^2 values and the observed prediction errors suggest that while the models are reasonably effective at predicting average wages, they are less capable of capturing the extremes, likely due to unmeasured variables and the diverse roles midfielders play.

Midfielders: Key Predictors and Performance Insights Across Multiple Models

Feature	Random Forest	BART	Gradient Boosting	Stepwise Regression	Quantile Regression (Tau = 0.25)	Quantile Regression (Tau = 0.5)	Quantile Regression (Tau = 0.75)
R^2 / Pseudo R^2	0.51	0.50	0.50	0.63	0.24	0.38	0.24
Top Feature 1	**Competition: Premier League**	**Age**	**xG Plus/Minus (Team xG - Team xGA)**	**Age**	**Age**	**Competition: Premier League**	**Competition: Premier League**
Top Feature 2	**Age**	**Competition: Premier League**	**Age**	**Competition: Premier League**	**Competition: Premier League**	**Age**	**Age**
Top Feature 3	**xG Plus/Minus (Team xG - Team xGA)**	**Plus/Minus Contribution per 90 Minutes**	**Competition: Premier League**	**Minutes per Match Played**	**Competition: Ligue 1**	**Plus/Minus Contribution per 90 Minutes**	**xG On-Off (Net xG On Pitch vs Off)**
Top Feature 4	**Total Career Games Missed**	**Minutes per Match Played**	**Plus/Minus Contribution per 90 Minutes**	**Competition: Ligue 1**	**Plus/Minus Contribution per 90 Minutes**	**Competition: Ligue 1**	**xG Plus/Minus (Team xG - Team xGA)**
Top Feature 5	**Plus/Minus Contribution per 90 Minutes**	**xG Plus/Minus (Team xG - Team xGA)**	**Total Career Games Missed**	**Plus/Minus Contribution per 90 Minutes**	**Minutes per Match Played**	**Minutes per Match Played**	**Competition: Ligue 1**
Top Feature 6	**Minutes per Match Played**	**Total Career Games Missed**	**Total Passes Attempted**	Passes Received	**Unused Substitutes**	**Unused Substitutes**	**Plus/Minus Contribution per 90 Minutes**
Top Feature 7	**Total Passes Attempted**	Nationality: Spain	**xG Plus/Minus per 90 Minutes (Team xG - Team xGA)**	**xG Plus/Minus (Team xG - Team xGA)**	**Fouls Drawn Leading to Goals**	**xG Plus/Minus (Team xG - Team xGA)**	**Unused Substitutes**
Top Feature 8	Expected Assisted Goals (xAG)	**xG On-Off (Net xG On Pitch vs Off)**	**xG On-Off (Net xG On Pitch vs Off)**	**Unused Substitutes**	**xG Plus/Minus (Team xG - Team xGA)**	**Total Career Games Missed**	**Minutes per Match Played**
Top Feature 9	**Unused Substitutes**	**Pass Completion %**	**Pass Completion %**	**Dispossessed Carries**	Nationality: Germany	**xG On-Off (Net xG On Pitch vs Off)**	**Fouls Drawn Leading to Goals**
Top Feature 10	**Competition: Ligue 1**	**Dispossessed Carries**	Expected Non-Penalty Goals and Assists	**Total Passes Attempted**	Successful Take-Ons Leading to Shot	Aerial Duels Lost	**Pass Completion %**

Table 4: This table summarises the model performance including R-Squared values for each model & top predictors of Midfielder wages across the different models. Colour coding is used to visually group predictors that are featured as significant in other models.

Top Predictors

Across the models in Table 4, several key predictors consistently emerged as significant for midfielders. Age once again appeared as a top predictor in all models while the Premier League was also a statistically significant (Table A3) repeated predictor, mirroring the same level of importance we found for goalkeepers and defenders. Nearly 40% of the top 25% of midfielder wage earners are in the Premier League. Also similar to goalkeepers and defenders results, Ligure 1 was found to be a statistically significant top predictor with an inverse relationship with wages.

"xG Plus/Minus (Team xG - Team xGA)" and "Plus/Minus Contribution per 90 Minutes" also appeared as high predictors in every model we tested, reflecting the importance of a midfielder's contribution to team success. These metrics, which account for the expected goals scored and conceded while a player is on the pitch, are crucial indicators of a midfielder's effectiveness in both attack and defence. Midfielders who positively influence their team's expected

goals while minimising expected goals are more valuable and, therefore, command higher wages. It is clear demands are less focused on being specialised in one area of the game, and instead geared towards overall influence on the team's play. Modern day midfielder's are being asked to perform multiple roles. In a 2004 interview published in The Times, Pep Guardiola said, "Football is played at a higher pace and it's a lot more physical. The tactics are different now, you have to be a ball-winner, a tackler, like Patrick Vieira or Edgar Davids. If you can pass too, well, that's a bonus. But the emphasis, as far as central midfielders are concerned, is all on defensive work"(Cox, 2018). Football has changed drastically in the 20 years since this quote, ushering in different demands of midfielders throughout the last twenty years. Midfielder's have to be able to do it all, rather than specialise in passing or tackling, making it more difficult to pinpoint the most significant features for the entire position group.

Passing statistics such as Total Passes Attempted, Pass Completion %, and Progressive Passes emerged as key predictors in the models as passing has historically been associated with midfielders due to their central role in team play. Midfielders who are heavily involved in matches tend to be pivotal players, which explains why these metrics significantly influence wage determination. Over the past decade, the English Premier League has seen substantial changes, with a notable increase in high-intensity running and sprinting by 30-50%, along with a 40% rise in the number of passes made during games (Barnes, Archer, Hogg, Bush, & Bradley, 2016). This evolution is not by chance, it is by influence. Barcelona and Spain's greatest contribution to modern football was not just their remarkable success but their ability to convince other European teams to adopt possession-based football (Cox, 2018). This style, exemplified by their star midfielders, Iniesta and Xavi, who embodied the principle of "Receive, pass, offer," as described by Iniesta, and "I get the ball, I pass the ball," as Xavi said, became known as 'tiki-taka'. This approach, characterised by short, quick passes and patient build-up play, became a dominant force in football (Cox, 2018). The influence of this playing style is evident in its widespread implementation in top leagues. As a result, passing metrics have become crucial for assessing a midfielder's contribution to the team, reflecting the broader shift towards a more analytical focus on passing figures, as clubs increasingly depend on data to evaluate player performance (Cox, 2018).

Availability is clearly a crucial trait for a midfielder to have. "Unused Substitute", a recurring top feature in most of the models, is when a player does not get substituted into a match at any point and remains on the bench for the entire game. This could mean the player is injured and working back from an injury, is going through a bad patch of playing form, being outplayed by another player in his position at the time, among other possible reasons. There is something the player is doing that is not good enough to warrant the manager playing him on the pitch. This aligns with the variable having a negative coefficient and being statistically significant in all three quartiles in Table A3. This variable coincides with the "Minutes per Match Played" variable that features in all but one model. It is often said in sports that availability is the best ability. "Total Career Games Missed" being another significant predictor, particularly in our machine learning

models, highlights the importance of availability once again. It is an especially crucial trait for midfielders to have as their role often demands high levels of exertion and being a consistent presence on the pitch. Players who miss fewer games are likely seen as more reliable and durable, therefore, more valuable to teams, justifying higher wages.

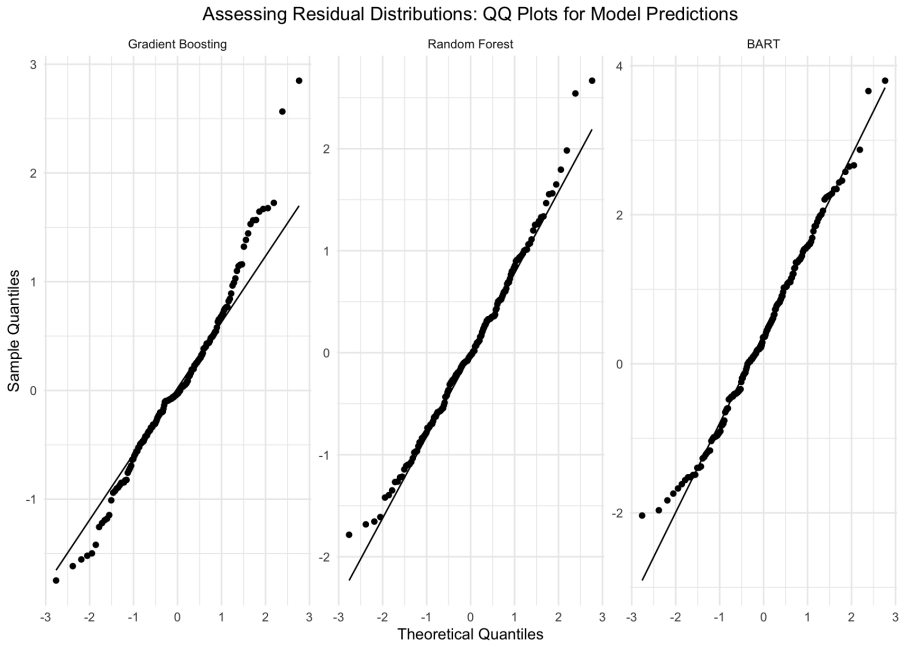

Figure 6: The QQ plots here compare the residual distributions of the Gradient Boosting, Random Forest, and BART models for Midfielders against a theoretical normal distribution. Deviations from the diagonal line, particularly at the tails, indicate that the models struggle to fit extreme values, with BART and Random Forest showing the most significant departures.

QQ Plot Analysis

The QQ plots for the residuals in Figure 6 indicate that the models struggled with extreme predictions, as can be seen by the deviations from the line, particularly at the tails. This suggests that the models have difficulty accurately capturing wage determinants for midfielders with higher or lower wages. The Gradient Boosting model, in particular, showed significant deviations, highlighting larger errors for values on both spectrums. While key predictors like Total Passes Attempted, Pass Completion %, xG Plus/Minus, and categorical factors like age and league, were consistent across models, the challenges in handling extreme cases suggest that other unmeasured factors in the study play a crucial role in determining midfielder wages, warranting further research.

Residual Analysis

The residual vs. fitted plots in Figure A2, further highlight the strengths and limitations of each model. The GBM model shows a wider spread of residuals, which indicates a broader range of prediction errors. However, as fitted values increase, the residuals of the model tend to cluster closer to the zero line, suggesting improved accuracy in predicting higher wage levels. The Random Forest model has a more concentrated distribution of residuals compared to the other models. This suggests fewer extreme errors, but similar to GBM, the residuals get closer to the zero line as the fitted values increase, indicating better performance at higher wage levels. The BART model shows

a tighter clustering of residuals around the mid-level fitted values and a more consistent spread unlike the other two models.

Quantile Regression Insights

The quantile regression analysis reveals that certain predictors consistently influence midfielder wages across different wage levels, while others vary in their significance. Age and Premier League participation are significant positive predictors across all the quantiles, indicating their impact on wage determination for midfielders across the board. However, variables like xG On-Off, the net team expected goals while the midfielder is on the pitch, and "Fouls Drawn Leading to Goals" show more pronounced effects at higher wage quantiles, indicating that these metrics may be more critical for the top-earning midfielders in Europe's top five leagues. In contrast, the negative impact of Unused Substitutes is consistent but most influential at the upper quantiles, highlighting the importance of consistent playing time for securing higher wages. This analysis suggests that while some factors universally affect midfielder wages, others are more important at different points in the wage distribution.

5.5 Forwards

Overview

In the analysis of Forward's wages, we implemented the same model approaches as the other positions. The outputs, depicted in the following table and figures, reveal insights into which factors are most influential in determining the wages of forwards in professional football.

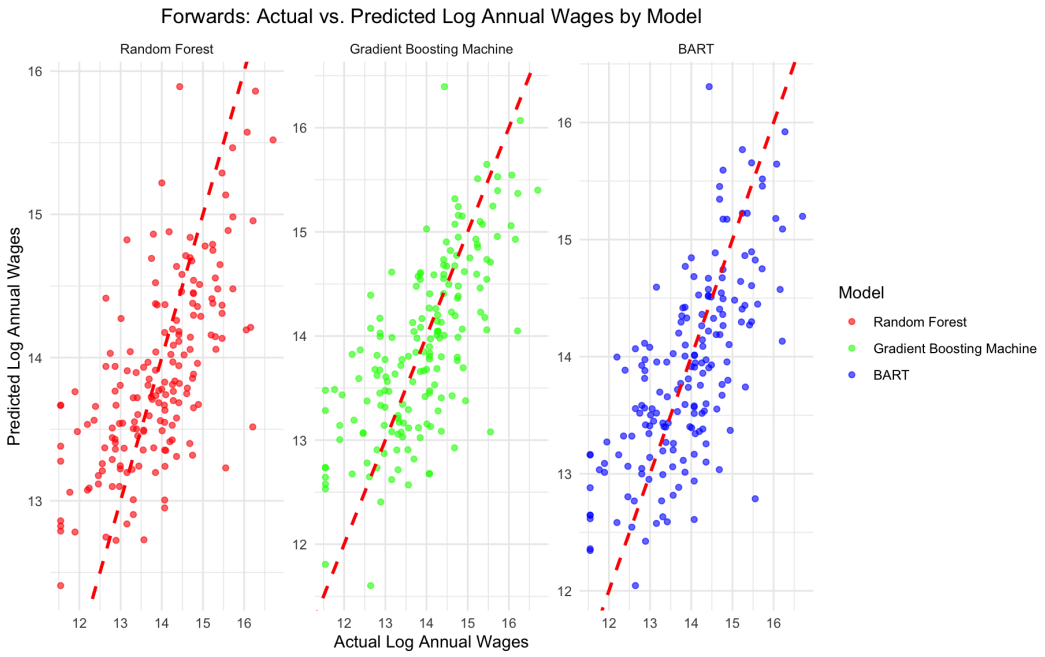

Figure 7: The actual vs. predicted wage plots for Forwards compare model predictions to actual log-transformed wages. The Random Forest model shows the most scatter, indicating more prediction variability. GBM and BART are more consistent but still struggle with extreme wage values, especially at the higher end. These plots highlight the models' general accuracy but also their limitations.

Model Performance

In analysing the predictors of the log-transformed annual wages for forwards, each model provided varying levels of insight into the factors that influence wages. Notably, the Stepwise Regression model achieved the highest R^2 value of 0.68, indicating its superior explanatory power in this context. This was followed by the Random Forest, BART, and Gradient Boosting models, each with R^2 values of 0.49, and the Quantile Regression models, which showed R^2 values ranging from 0.44 to 0.61 depending on the quantile. Figure 7 highlights distinct patterns that reveal the strengths and limitations of each machine learning approach. The Random Forest model shows a strong overall fit, with predictions clustering close to the diagonal line, yet there is noticeable dispersion, particularly at the mid-level. At the highest level of wages for all three models there are outliers, with the most dispersion there by the Random Forest. All of this suggests potential overfitting, where the model is capturing noise rather than the underlying wage-determining patterns. The GBM model, while displaying an even spread of predictions along the diagonal, also struggles with mid-level values, reflecting its difficulty in managing the variability. Meanwhile, the BART model demonstrates less variability with tighter clustering than the other models around the diagonal. Collectively, these observations indicate that while all three models capture the general trend in wage prediction, they differ in their handling of extreme wage values, with Random Forest and GBM showing more variability, and BART leaning towards more conservative estimates.

Forwards: Key Predictors and Performance Insights Across Multiple Models

Feature	Random Forest	BART	Gradient Boosting	Stepwise Regression	Quantile Regression (Tau = 0.25)	Quantile Regression (Tau = 0.5)	Quantile Regression (Tau = 0.75)
R^2	0.49	0.49	0.49	0.68	0.45	0.61	0.44
Top Feature 1	**Points per Match (while on pitch)**	**Age**	**Points per Match (while on pitch)**	Non-penalty Goals per 90 minutes	**Competition: Premier League**	**Non-penalty Expected Goals (xG) per Shot**	**Non-penalty Expected Goals (xG) per Shot**
Top Feature 2	**xG +- (Expected Goals - Expected Goals Against) while on pitch**	**Points per Match (while on pitch)**	**Age**	Goals per 90 minutes	**Non-penalty Expected Goals (xG) per Shot**	**Competition: Premier League**	**Competition: Premier League**
Top Feature 3	**+- (Goals scored - Goals allowed) while on pitch**	**Competition: Premier League**	**Competition: Premier League**	Non-penalty Goals + Assists per 90 minutes	**xG +- (Expected Goals - Expected Goals Against) per 90 minutes**	**xG +- (Expected Goals - Expected Goals Against) per 90 minutes**	**xG +- (Expected Goals - Expected Goals Against) per 90 minutes**
Top Feature 4	**Age**	**Career Games Missed**	**Team Goals Scored (while on pitch)**	Goals + Assists per 90 minutes	**+- (Goals scored - Goals allowed) while on pitch**	**+- (Goals scored - Goals allowed) while on pitch**	**+- (Goals scored - Goals allowed) while on pitch**
Top Feature 5	**Competition: Premier League**	**Competition: Ligue 1**	**Career Games Missed**	**xG +- (Expected Goals - Expected Goals Against) per 90 minutes**	**Expected Assisted Goals (xAG) per 90 Minutes**	**Competition: Ligue 1**	**Nationality: Brazil**
Top Feature 6	**Team Goals Scored (while on pitch)**	**+- (Goals scored - Goals allowed) while on pitch**	**xG +- (Expected Goals - Expected Goals Against) while on pitch**	**+- (Goals scored - Goals allowed) while on pitch**	**Nationality: Brazil**	**Expected Assisted Goals (xAG) per 90 Minutes**	**Competition: Ligue 1**
Top Feature 7	Team Expected Goals (xG) while on pitch	Passes Received	**+- (Goals scored - Goals allowed) while on pitch**	Pass Completion %	**Competition: Ligue 1**	**Points per Match (while on pitch)**	**Discipline: 2 Yellow Cards**
Top Feature 8	**xG +- (Expected Goals - Expected Goals Against) per 90 minutes**	Unsubstituted Games	Expected Goals (xG) + Expected Assisted Goals (xAG) per 90 minutes	**Expected Assisted Goals (xAG) per 90 Minutes**	**Discipline: 2 Yellow Cards**	**Discipline: 2 Yellow Cards**	**Points per Match (while on pitch)**
Top Feature 9	+- (Goals scored - goals allowed) per 90 Minutes while on pitch	**Minutes per Match Played**	**Competition: Ligue 1**	**Points per Match (while on pitch)**	**Team Goals Scored per 90 Minutes (while on pitch)**	**Nationality: Brazil**	**Expected Assisted Goals (xAG) per 90 Minutes**
Top Feature 10	Non-penalty Expected Goals (xG) + Expected Assisted Goals (xAG)	**xG +- (Expected Goals - Expected Goals Against) per 90 minutes**	**Minutes per Match Played**	**Competition: Premier League**	**Points per Match (while on pitch)**	**Team Goals Scored per 90 Minutes (while on pitch)**	**Team Goals Scored per 90 Minutes (while on pitch)**

Table 5: Displaying the key predictors & R-squared's for our models (Random Forest, BART, Gradient Boosting, Stepwise Regression, and Quantile Regression at various quantiles (0.25, 0.5, 0.75) for predicting wages of Forwards. Repeated predictors are colour-coded.

Top Predictors

In Table 5, the models revealed a clear pattern in the key factors influencing forward wages. Across various models, certain predictors consistently stood out, including Age, competing in the Premier League, and advanced metrics related to a player's impact on overall team performance. These findings underscore the significant role these factors play in shaping the earnings of forwards, aligning with trends observed across other positions.

For Forwards, Age is only found as a top predictor in the Random Forest, BART, and GBM models and not in the regression models. This difference likely stems from the non-linear relationship between age and wages for forwards. The machine learning models are better suited to capture these non-linear patterns within the data, making them more likely to identify age as a significant predictor. The regression models, which assume linear relationships, may struggle to capture the nuanced impact of age on forward wages.

The Premier League was a highly significant (Table A4) predictor across all models and quantiles for forwards. Forwards have the highest exposure and popularity in football, due to the fact that they are usually the players responsible for providing the rare and climactic moments in the sport, goals. Coupling the popularity of the position with the most profitable football league in the world, forwards in the Premier League are at a big advantage relative to their counterparts in the other top 5 leagues. All the positions in the study have found the English Premier League to be a significant predictor in wage determination with players benefitting from high visibility. Similar to other positions as well, Ligue 1 is a top predictor but with an inverse relationship with wages as can be seen with the negative coefficient in Table A4.

Table 5 shows how important 'Expected' statistics have become in the football data world. Rather than simply looking at goals and assists like past studies have had to settle with, expected statistics provide far more context about a player's influence. "xG ± (Expected Goals - Expected Goals Against) per 90 minutes" and "+/- (Goals Scored - Goals Allowed) per 90 minutes" emerged as significant predictors in all but one model, and third and fourth highest importance respectively in each quantile of the quantile regression model. These metrics reflect a player's overall impact on the pitch, both offensively and defensively. In the forwards case, it is about how they are influencing goals for their team. A high xG ± per 90 minutes indicates that the player's presence significantly improves the team's expected goal difference, a critical factor in modern football where forwards are evaluated not only on their scoring abilities but also on their contribution to the team's ability to consistently generate dangerous situations to score goals. "Non-penalty Expected Goals (xG) per Shot" was significant in the 0.50 quantile (Table A4), indicating that forwards who consistently create or capitalise on high-quality chances are rewarded with higher wages. This metric is especially important for evaluating a forward's efficiency and effectiveness in goal-scoring opportunities, determining how often they are getting in these positions relative to their total shots on goal. "Expected Assisted Goals (xAG) per 90 Minutes" also appeared as a significant predictor but in the lower and middle quantiles. This reflects the importance of a forward's playmaking

ability and their contribution to creating goal-scoring opportunities for their teammates, which is increasingly valued in the modern game.

"Points per Match" and "Team Goals Scored" proving importance in Table 5 aligns with the theme of team success, while on the pitch, being crucial to the wage determination for a forward. This metric projects the idea that forwards who are consistently part of winning teams and influencing a higher amount of goals being scored are more highly valued, and thus, demand higher wages.

Interestingly, the Stepwise Regression model highlighted more traditional statistics such as "Goals per 90 Minutes", "Non-penalty Goals per 90 Minutes", and other goal contribution statistics per 90 minutes as the leading predictors, with the advanced 'Expected' metrics also playing a big role. Rather than team statistics, the model focused on individuality and actual production rather than the more contextual 'Expected' metrics.

Contextual Analysis

Evaluating forwards has become increasingly complex due to the evolving nature of their role in modern football. Traditionally, strikers were often strong players who remained in the penalty box to capitalise on crosses and final passes. However, in recent years, this role has shifted significantly. Many teams in recent history have employed a "false nine," a forward who drops deeper to link up with teammates rather than staying near the opposition's box, contrasting sharply with the classic striker role (Cox, 2018). Some tactical systems even forgo the concept of a traditional striker altogether, and distribute the goal-scoring responsibilities among midfielders and wide players. This evolution complicates the use of traditional metrics to assess forward performance and value.

The repeated emphasis on metrics related to a forward's overall impact on team performance, such as Plus/Minus and xG ± (Expected Goals - Expected Goals Against) per 90 minutes, reflects the evolving expectations placed on modern forwards. As Cox (2018) noted, the traditional striker's role has expanded; modern forwards are now expected to contribute across various phases of play, including pressing, creating chances, and linking up with midfielders. This broader role is mirrored in our results, where metrics capturing a player's entire contribution to team success—beyond just goals and assists—are increasingly valuable in wage negotiations.

The inclusion of advanced metrics like Non-Penalty Expected Goals (npxG) across multiple models further illustrates the growing reliance on data-driven insights in football. Unlike traditional goal metrics, npxG provides a more nuanced gauge of a forward's effectiveness by considering the quality of chances they're generating or receiving outside of penalties; which don't necessarily represent a player's performance. This shift towards advanced analytics aligns with the broader trend in the sport, where deeper insights are being used to inform player valuations and wage determinations.

Residual Analysis

Residual plots in Figure A3 provide further insight into the performance of each model. The residuals of the Random Forest model show more spread than the other two models indicating it is struggling with precision despite capturing the general trend. Both the BART and GBM models display slightly more consistency, yet they still show deviations from the ideal random scatter, particularly at the extremes where we can see clear outliers. This displays the challenges in predicting wages for forwards with more unusual wage characteristics.

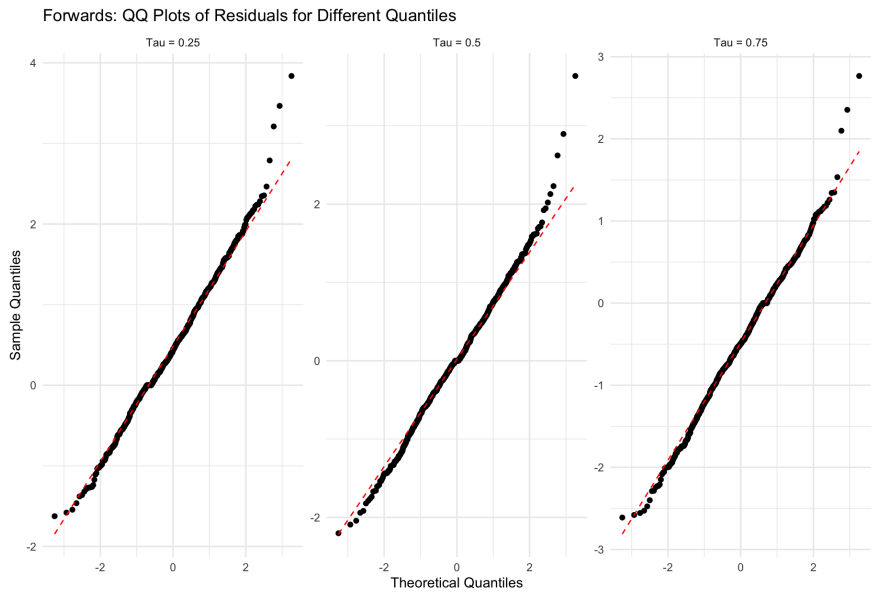

Figure 8: QQ plots displaying the residuals from the Quantile Regression models at three different quantiles (τ = 0.25, 0.5, 0.75) for Forwards, comparing the sample quantiles of the residuals against theoretical quantiles of a normal distribution.

QQ Plots Interpretation

The QQ plots for the residuals at different quantiles (Tau = 0.25, 0.5, and 0.75) in Figure 8 show that the residuals generally follow the theoretical quantiles closely, however, there are clear deviations at the extremes, particularly at the upper end. This indicates potential issues with the model's accuracy in predicting the highest wages, where the model tends to underestimate. The deviations at the tails suggest that the models may struggle to capture the full variability in wages for forwards.

Conclusion

The analysis highlights the multifaceted factors that influence forward wages, with advanced metrics related to a player's plus/minus and 'Expected' impact to team success, and playing in the Premier League standing out as key predictors. The machine learning models were particularly effective in capturing the non-linear influence of age while the consistent significance of the Premier League reaffirms its financial monopoly in world football. The negative association with

Ligue 1 wages reflects the opposite end of that monopoly. Metrics like xG ± per 90 minutes and Non-penalty Expected Goals (xG) per Shot emphasise the importance of a forward's contribution to goal-scoring opportunities and overall attacking play for his team.

6. SUMMARY & DISCUSSION

This study aimed to explore the factors that determine the wages of professional football players across different positions—goalkeepers, defenders, midfielders, and forwards—by using a combination of advanced machine learning models and traditional regression techniques. By analysing a comprehensive dataset that included annual wages and various advanced and traditional performance metrics coupled with complimentary categorical factors from Europe's top five leagues, I sought to identify the key predictors that influence player compensation. The approach included Random Forest, GBM, BART, Stepwise Regression, and Quantile Regression models, each providing unique insights into wage determination across the various positions on the pitch.

The evolution of football, particularly since the introduction of the back-pass rule in 1992, has significantly impacted how players are analysed and valued. This rule change forced goalkeepers and defenders to adapt from being primarily defensive players to becoming integral parts of a team's passing game. This shift marked the beginning of an era where every player on the pitch needed to possess a high level of technical skill, particularly in ball distribution. As Cox (2018) notes, the back-pass rule redefined the goalkeeper's role, making them a key component in the build-up play rather than just shot-stoppers. It is my estimate that this has made it harder to predict wages based purely on on-field metrics exclusively—as can be seen with our R-squared results that hovered around 0.47 for the machine learning models for goalkeepers. Goalkeepers have to be able to do it all in the modern game; especially in the top leagues, however different teams value different attributes. For example, Barcelona, known for possession play, will value a goalkeeper's Pass Completion % more than a defensive team like Atletico Madrid. This variety in what is important across the top leagues, makes it more difficult to pinpoint the overarching best performance metrics than it would have been when a goalkeeper's job was more defined years ago—especially, as Berri's study (2023) found, when the teams are using more team-focused statistics to determine wages for goalkeepers like Clean Sheets and team results.

In the years following this rule change, particularly with the tiki-taka era of Spain and Barcelona teams from 2008 to 2012, the emphasis on possession and counter-pressing further heightened the technical demands that were placed on players across all positions. The modern game, characterised by faster ball movement and higher intensity, has created a need for players to excel in multiple facets of play. This is evident in the increased number of passes performed during matches, with research by Barnes et al. (2014) highlighting a 40% increase in passes over recent seasons, coupled with a rise in passing tempo and success rates. This evolution underscores the

necessity for players, regardless of position, to be proficient in passing and to contribute to the team's overall strategy, reflecting the broader tactical shift in football towards possession-based play.

While these tactical evolutions affect every position on the pitch, the most important for a team's style of play is the midfielder. As the central players on the pitch, they dictate the tempo and orchestrate how the team plays. Thus, their role has particularly evolved with the tactical shifts, requiring them to embody the team's philosophy and perform both offensive and defensive duties. Modern-day midfielders are no longer confined to specific tasks like passing or tackling but are expected to contribute across all phases of the game. This broadening of responsibilities makes it challenging to isolate the most significant performance metrics for midfielders, as they must excel in multiple aspects, including both creating and preventing goals. Although "Total Passes Attempted" and "Pass Completion %", more traditional markers of midfield play, appeared in our top feature results, they were not as influential as the team-related statistics. The importance of metrics like "xG Plus/Minus (Team xG - Team xGA)" and "Plus/Minus Contribution per 90 Minutes" in our models underscores the multifaceted demands placed on midfielders, who have to influence their team's success on both ends of the pitch.

Our models showed notably higher performance for defenders, with R-squared values ranging from 0.65 to 0.81, while the R-squared values for midfielders and forwards were generally lower, hovering around 0.49 to 0.51. This higher predictive accuracy for the defenders could be attributed to the fact that despite also evolving in their roles, defenders still have a more defined set of responsibilities compared to midfielders and forwards. The results of the defender's models showed team metrics like "Plus/Minus (+-) Contribution to Team Success" are indeed crucial in determining defender's wages among the traditional statistics. The relative clarity in the expectations for defenders—focused on defensive stability, aerial duels, and passing accuracy metrics like "Aerial Dual %" and "Pass Completion %" —may contribute to the models' better performance in predicting their wages. Defenders' roles, while also expanding, remain more specialised and easier to quantify using the available metrics, leading to stronger model fits.

Tactical evolution has also impacted the results of our modelling for forward's wages. The introduction of the 'false nine' role significantly altered the expectations placed on forwards. Instead of focusing solely on goal-scoring, forwards are now evaluated on their ability to enhance overall attacking play. Our results highlighted the importance of advanced metrics like "xG ± (Expected Goals - Expected Goals Against) per 90 minutes" and "+/- (Goals Scored - Goals Allowed) per 90 minutes" in assessing a forward's broader impact on team success. Although measuring the player's impact on the entire team's attacking play rather than focusing on individual metrics might be a more accurate assessment of player's influence, it makes it harder to analyse what the player is actually doing on the field to increase the team's expected play. Further research into what the on-field metrics are that result in higher team Expected Goals statistics would be beneficial.

The emergence and prominence of 'Expected' statistics in our model results underline a critical shift in how player contributions are evaluated. Unlike traditional statistics, which might not capture a player's full impact on the game, expected statistics provide a more nuanced analysis. For instance, a forward who sets up a high-quality scoring opportunity for a teammate—only for the chance to be missed—would have gone unrecognised in the past. Today, this contribution is acknowledged through expected assists (xA) or expected goals (xG) metrics, which measure the quality of scoring chances created or taken by players. This shift is not limited to creating goal-scoring opportunities though. For goalkeepers, expected statistics like Post-Shot Expected Goals per Shot on Target (PSxG/SoT) have become crucial indicators of performance. This metric, which appeared as a top predictor in our analysis, reflects the difficulty of the shots a goalkeeper faces and provides a more comprehensive assessment of their shot-stopping ability. By accounting for the quality of the chances they face, expected statistics allow for a more accurate and fair evaluation of a goalkeeper's performance, moving beyond traditional metrics like clean sheets or save percentage. In conclusion, this study has displayed the multifaceted factors that influence football wages, demonstrating the importance of advanced metrics in capturing the full scope of a player's contributions. As football continues to evolve, so must the tools and methods used to evaluate player performance and value. The shift towards expected statistics and the varying performance of our models across different positions highlight the complexity of modern football and the need for continued progress and refining in analytical approaches.

Although the primary aim of this study was to identify the key on-field performance metrics that influence player wages, it is essential to note that playing in the Premier League emerged as a highly significant factor across all positions. Incorporating the league where players compete is important to the study to analyse contextual effects, as two players with similar performance metrics can earn vastly different wages, particularly if one is in the Premier League and the other is in Ligue 1. Premier League clubs have a considerable financial advantage, as demonstrated by Deloitte's 2024 report, which noted that total wage costs in the league surpassed £4 billion for the first time in the 2022/23 season, with wages growing by £377 million. This substantial revenue allows Premier League clubs to offer higher wages, attracting top talent from around the world. In contrast, Ligue 1, apart from its state-sponsored giant Paris Saint-Germain (PSG), struggles to compete financially with other top European leagues. The negative coefficient associated with playing in Ligue 1 in our quantile regression models reflects this disparity. This stark contrast between the Premier League and Ligue 1 highlights the significant role that league-specific revenue and marketability play in determining player wages.

The overall moderate performance of our R-squared and diagnostic plot results reflects the complexity and murkiness in defining the roles of players in the modern game. These modern multifaceted roles can lead to a dilution in the predictive power of the models and create a broad spectrum of responsibilities—making it challenging to pinpoint key on-field predictors of wage determinants with high precision. Additionally, the results underscore the impact of factors off the

pitch, such as the economics of the sport and the popularity of both the player and the league they compete in, which significantly influence player wages beyond their on-field performance metrics.

Future research could expand on these findings by incorporating more of the prior seasons as lagging indicators for wage. Researching for deeper insights into player performance and what actions they do on the pitch to positively or negatively impact team xG +- statistics that we found consistently in the study would also be very interesting. Extending the study by focusing on the key performance predictors of wages in each of the top 5 European leagues would give detailed results on the distinct variables that weigh more heavily in each league. Incorporating the popularity factor in addition to the on-field variables would provide a more comprehensive study on wage determination in Europe's elite league. Exploring these avenues would enhance the understanding of wage determination in Europe's elite leagues and provide a deeper and more nuanced perspective on the key predictors of wage for footballers.

BIBLIOGRAPHY

Barnes, C., Archer, D. T., Hogg, B., Bush, M., & Bradley, P. S. (2014). The Evolution of Physical and Technical Performance Parameters in the English Premier League. *International Journal of Sports Medicine*, *35*(13), 1095–1100. https://doi.org/10.1055/s-0034-1375695

Berri, D., Butler, D., Rossi, G., Simmons, R., & Tordoff, C. (2024). Salary determination in professional football: Empirical evidence from goalkeepers. European Sport Management Quarterly, 24(3), 624-640. https://doi.org/10.1080/16184742.2023.2169319

Boulier, B. L., & Stekler, H. O. (2003). Predicting the outcomes of National Football League games. International Journal of Forecasting, 19(2), 257-270. https://doi.org/10.1016/S0169-2070(02)00048-6

Bradley, P. S., Archer, D. T., Hogg, B., Schuth, G., Bush, M., Carling, C., & Barnes, C. (2016). Tier-specific evolution of match performance characteristics in the English Premier League: It's getting tougher at the top. Journal of Sports Sciences, 34(10), 980-987. https://doi.org/10.1080/02640414.2015.1082614

Bryson, A., Rossi, G., & Simmons, R. (2014). The migrant wage premium in professional football: A superstar effect? Kyklos, 67(1), 12–28.

Cox, M. (2018). *The mixer: the story of Premier League tactics, from route one to false nines*. HarperCollins.

Csataljay, G., O'Donoghue, P., Hughes, M., & Dancs, H. (2009). Performance indicators that distinguish winning and losing teams in basketball. *International Journal of Performance Analysis in Sport, 9*(1), 60-66. doi:10.1080/24748668.2009.11868464

Deloitte. (2024). Annual review of football finance 2024. Deloitte. https://www.deloitte.com/content/dam/assets-zone2/uk/en/docs/services/financial-advisory/2024/deloitte-uk-annual-review-of-football-finance.pdf

FBref. (n.d.). FBref.com: Football Statistics and History. Retrieved June 5, 2024, from https://fbref.com

FBref. (2024). Declan Rice scouting report. Retrieved August 10, 2024, from https://fbref.com/en/players/1c7012b8/scout/365_m1/Declan-Rice-Scouting-Report

Frick, Bernd. (2011). Performance, Salaries and Contract Length: Empirical Evidence from German Soccer. International Journal of Sport Finance. 6. 87-118.

Giovanni Bernardo, Massimo Ruberti & Roberto Verona (2022) Image is everything! Professional football players' visibility and wages: evidence from the Italian Serie A, Applied Economics, 54:5, 595-614, DOI: 10.1080/00036846.2021.1967863

Gonzalez-Rodenas J, Aranda-Malaves R, Tudela-Desantes A, Nieto F, Uso´ F, Aranda R (2020) Playing tactics, contextual variables and offensive effectiveness in English Premier League soccer matches. A multilevel analysis. PLoS ONE 15(2): e0226978. https://doi.org/10.1371/journal.Pone.0226978

Hughes MD, Bartlett RM. The use of performance indicators in performance analysis. J Sport Sci, 2002. 20(10): 739-754.

Hughes MD, Churchill S. Attacking profiles of successful and unsuccessful teams in Copa America 2001. In T Reilly, J Cabri, D Araújo (eds): Science and Football V. London and New York: Routledge, 2005, pp 219-224.

Hughes MD, Franks I. Analysis of passing sequences, shots and goals in soccer. J Sport Sci, 2005. 23(5): 509- 514.

Ibáñez, S. J., García, J., Feu, S., Lorenzo, A., & Sampaio, J. (2008). Effects of consecutive basketball games on the game-related statistics that discriminate winner and losing teams. *Journal of Sports Science & Medicine, 7*(3), 458-462. Retrieved from https://www.jssm.org

Jones, N. M., James, N., & Mellalieu, S. D. (2004). Possession as a performance indicator in rugby union. *International Journal of Performance Analysis in Sport, 4*(1), 61-71. doi:10.1080/24748668.2004.11868289

Kapelner, A., & Bleich, J. (2016). bartMachine : Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software*, *70*(4), 1–40. https://doi.org/10.18637/jss.v070.i04

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R. (2023). *quantreg: Quantile Regression*. R package version 5.97. Retrieved from https://CRAN.R-project.org/package=quantreg.

Lago-Peñas, C., Lago-Ballesteros, J., & Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA Champions League. *Journal of Human Kinetics*, *27*(2011), 135–146. https://doi.org/10.2478/v10078-011-0011-3

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22. Retrieved from https://CRAN.R-project.org/doc/Rnews/.

Liu, J. (2024). The Relation of Salary to Players' On-Pitch Performance: The Economics of Football. *Review of Business*, *44*(2), 1–38.

Lucifora, C., & Simmons, R. (2003). Superstar Effects in Sport: Evidence From Italian Soccer. Journal of Sports Economics, 4(1), 35-55. https://doi-org.uoelibrary.idm.oclc.org/10.1177/1527002502239657

Ortega, E., Villarejo, D., & Palao, J. M. (2009). Differences in game statistics between winning and losing rugby teams in the Six Nations Tournament. International Journal of Performance Analysis in Sport, 9(1), 121-129. https://doi.org/10.1080/24748668.2009.11868470

Ridgeway, G., & Developers, G. (2024). gbm: Generalized Boosted Regression Models (R package version 2.2.2). Retrieved from https://CRAN.R-project.org/package=gbm

Sampaio, J., Ibáñez, S., Lorenzo, A., & Gómez, M. A. (2010). Discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues. *European Journal of Sport Science, 10*(2), 73-78. doi:10.1080/17461390903271539

Szymanski, Stefan. 2003. "The Economic Design of Sporting Contests." *Journal of Economic Literature*, 41 (4): 1137–1187.
DOI: 10.1237/002205103771800004

Waldén, M., Hägglund, M., Magnusson, H., & Ekstrand, J. (2016). ACL injuries in men's professional football: a 15-year prospective study on time trends and return-to-play rates reveals only 65% of players still play at the top level 3 years after ACL rupture. *British Journal of Sports Medicine*, *50*(12), 744–750. https://doi.org/10.1136/bjsports-2015-095952

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer. ISBN 0-387-95457-0.

Zivkovic, J. (2024). worldfootballR: Extract and Clean World Football (Soccer) Data (R package version 0.6.5.0008). Retrieved from https://github.com/jaseziv/worldfootballr

7. APPENDIX

Figures & Tables:

Table A1.

Top 10 Features and P-values by Quantile

Quantile Regression Model Results

Feature	Coefficient	P-value	Quantile
Quantile 0.75			
Intercept	12.9283	0.0000	0.75
Post-Shot Expected Goals per Shot on Target (PSxG/SoT)	−0.7413	0.7199	0.75
Premier League	0.6654	0.0199	0.75
Ligue 1	−0.4257	0.1670	0.75
Defensive Actions Outside Penalty Area per 90 (OPA/90)	−0.1831	0.5210	0.75
Free Kick Goals Against	−0.1765	0.1824	0.75
Clean Sheets	0.1653	0.0427	0.75
Post-Shot Expected Goals Minus Goals Allowed per 90 (PSxG-GA/90)	−0.1135	0.2536	0.75
Losses	−0.1040	0.4440	0.75
Draws	−0.0572	0.5827	0.75
Quantile 0.50			
Intercept	11.3218	0.0000	0.50
Premier League	1.0692	0.0001	0.50
Post-Shot Expected Goals per Shot on Target (PSxG/SoT)	−0.8425	0.1289	0.50
Post-Shot Expected Goals Minus Goals Allowed (PSxG-GA)	0.2062	0.1124	0.50
Bundesliga	0.2020	0.1165	0.50
Ligue 1	−0.1996	0.0944	0.50
Post-Shot Expected Goals (PSxG)	−0.1976	0.1196	0.50
Shots on Target Against	0.1623	0.0923	0.50
Saves	−0.1591	0.1112	0.50
Free Kick Goals Against	−0.1164	0.1708	0.50
Quantile 0.25			
Intercept	11.9445	0.0000	0.25
Ligue 1	−0.5823	0.0133	0.25
Premier League	0.5261	0.0303	0.25
Post-Shot Expected Goals Minus Goals Allowed (PSxG-GA)	−0.2746	0.1433	0.25
Post-Shot Expected Goals (PSxG)	0.2725	0.1361	0.25
Post-Shot Expected Goals per Shot on Target (PSxG/SoT)	−0.2693	0.8363	0.25
Wins	0.1950	0.4011	0.25
Shots on Target Against	−0.1648	0.3696	0.25
Saves	0.1585	0.3442	0.25
Draws	0.1512	0.4847	0.25

Table A1: displays the top 10 features influencing Goalkeeper's wages across different quantiles, along with their coefficients and p-values from the Quantile Regression model. The quantiles represent different points in the wage distribution (0.25, 0.50, and 0.75), highlighting how the impact of specific features varies across different wage levels. Coefficients indicate the direction and magnitude of the effect, while p-values assess the statistical significance of each feature's impact within each quantile.

Figure A1.

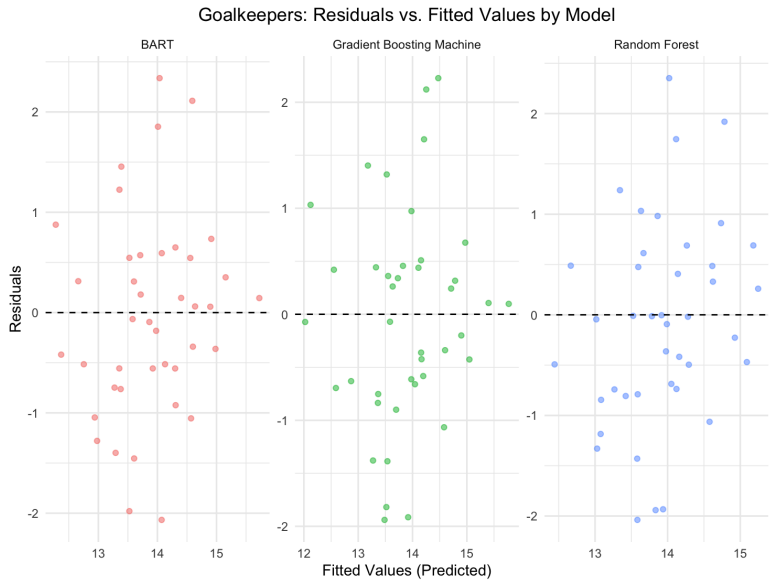

Figure A1: This plot shows the residuals (prediction errors) versus the predicted goalkeeper wages for three models: BART, GBM, and Random Forest. The horizontal dashed line at zero indicates perfect predictions. Points above or below the line show over- or under-predictions, respectively, with more spread indicating greater error variance in the model.

Table A2.

Defenders: Top 10 Features and P-values by Quantile

Quantile Regression Model Results

Feature	Coefficient	P-value	Quantile
Quantile 0.75			
Intercept	10.0908	0.0000	0.75
Competition: Premier League	0.6266	0.0000	0.75
Competition: Ligue 1	−0.3353	0.0007	0.75
Age	0.0705	0.0000	0.75
Competition: La Liga	0.0448	0.5367	0.75
Pass Completion %	0.0266	0.0000	0.75
Plus/Minus Contribution to Team Success	0.0265	0.0000	0.75
Plus/Minus Contribution per 90 Minutes	0.0169	0.3773	0.75
Total Career Games Missed	0.0070	0.0000	0.75
Progressive Passes	0.0042	0.0057	0.75
Quantile 0.50			
Intercept	10.0247	0.0000	0.50
Competition: Premier League	0.7964	0.0000	0.50
Competition: Ligue 1	−0.5006	0.0000	0.50
Age	0.0696	0.0000	0.50
Competition: La Liga	−0.0524	0.6337	0.50
Plus/Minus Contribution per 90 Minutes	0.0442	0.0000	0.50
Goals	0.0356	0.0659	0.50
Plus/Minus Contribution to Team Success	0.0216	0.0000	0.50
Pass Completion %	0.0169	0.0001	0.50
Total Career Games Missed	0.0066	0.0000	0.50
Quantile 0.25			
Intercept	9.2910	0.0000	0.25
Competition: Premier League	0.8174	0.0000	0.25
Competition: Ligue 1	−0.4556	0.0000	0.25
Competition: La Liga	−0.2683	0.0191	0.25
Age	0.0860	0.0000	0.25
Goals	0.0825	0.0003	0.25
Plus/Minus Contribution per 90 Minutes	0.0366	0.1371	0.25
Plus/Minus Contribution to Team Success	0.0162	0.0000	0.25
Pass Completion %	0.0110	0.0707	0.25
Total Career Games Missed	0.0056	0.0003	0.25

Table A2: This table summarizes the top 10 features and their corresponding p-values across different quantiles (0.25, 0.50, 0.75) for the quantile regression model applied to defenders. The features include key variables such as competition level, age, pass completion percentage, and contributions to team success. These variables significantly impact wage determination at different points in the wage distribution. The p-values indicate the statistical significance of each feature's impact on wages within each quantile, highlighting how these factors influence salaries across different segments of the wage spectrum.

Figure A2.

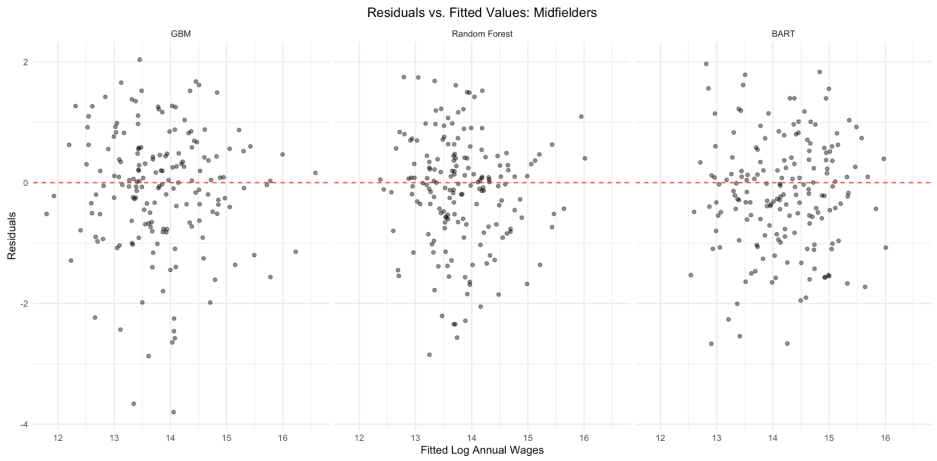

Figure A2: This plot shows the residuals (prediction errors) versus the predicted midfielder wages for three models: BART, GBM, and Random Forest. The horizontal dashed line at zero indicates perfect predictions. Points above or below the line show over- or under-predictions, respectively, with more spread indicating greater error variance in the model.

Table A3.

Midfielders: Top 10 Features and P-values by Quantile

Quantile Regression Model Results

Feature	Coefficient	t-value	P-value	Quantile
Quantile 0.75				
Intercept	10.4295	30.0396	0.0000	0.75
Competition: Premier League	0.8643	13.3808	0.0000	0.75
Age	0.0703	7.0219	0.0000	0.75
xG On-Off (Net xG On Pitch vs Off)	−0.0827	−6.5886	0.0000	0.75
xG Plus/Minus (Team xG - Team xGA)	0.0278	6.4140	0.0000	0.75
Competition: Ligue 1	−0.5781	−6.2185	0.0000	0.75
Plus/Minus Contribution per 90 Minutes	0.4320	5.8611	0.0000	0.75
Unused Substitutes	−0.0367	−4.8464	0.0000	0.75
Minutes per Match Played	0.0114	4.7143	0.0000	0.75
Fouls Drawn Leading to Goals	−0.2130	−3.2590	0.0011	0.75
Quantile 0.50				
Intercept	10.2131	28.3190	0.0000	0.50
Competition: Premier League	0.8920	12.3592	0.0000	0.50
Age	0.0907	11.4328	0.0000	0.50
Plus/Minus Contribution per 90 Minutes	0.4514	7.0352	0.0000	0.50
Competition: Ligue 1	−0.5490	−5.6124	0.0000	0.50
Minutes per Match Played	0.0121	5.5730	0.0000	0.50
Unused Substitutes	−0.0300	−5.2339	0.0000	0.50
xG Plus/Minus (Team xG - Team xGA)	0.0208	4.8038	0.0000	0.50
Total Career Games Missed	0.0019	4.2043	0.0000	0.50
xG On-Off (Net xG On Pitch vs Off)	−0.0680	−4.0757	0.0000	0.50
Quantile 0.25				
Intercept	9.3211	18.5488	0.0000	0.25
Age	0.0917	10.0277	0.0000	0.25
Competition: Premier League	1.0552	9.2116	0.0000	0.25
Competition: Ligue 1	−0.6023	−6.4553	0.0000	0.25
Plus/Minus Contribution per 90 Minutes	0.4004	5.8353	0.0000	0.25
Minutes per Match Played	0.0142	5.8350	0.0000	0.25
Unused Substitutes	−0.0270	−3.2907	0.0010	0.25
Fouls Drawn Leading to Goals	−0.2643	−3.2727	0.0011	0.25
xG Plus/Minus (Team xG - Team xGA)	0.0216	2.9983	0.0027	0.25
Nationality: Germany	−0.4039	−2.9287	0.0034	0.25

Table A3: This table summarizes the top 10 features and their corresponding p-values across different quantiles (0.25, 0.50, 0.75) for the quantile regression model applied to midfielders. The features include key variables such as competition level, age, pass completion percentage, and contributions to team success. These variables significantly impact wage determination at different points in the wage distribution. The p-values indicate the statistical significance of each feature's impact on wages within each quantile, highlighting how these factors influence salaries across different segments of the wage spectrum.

Figure A3.

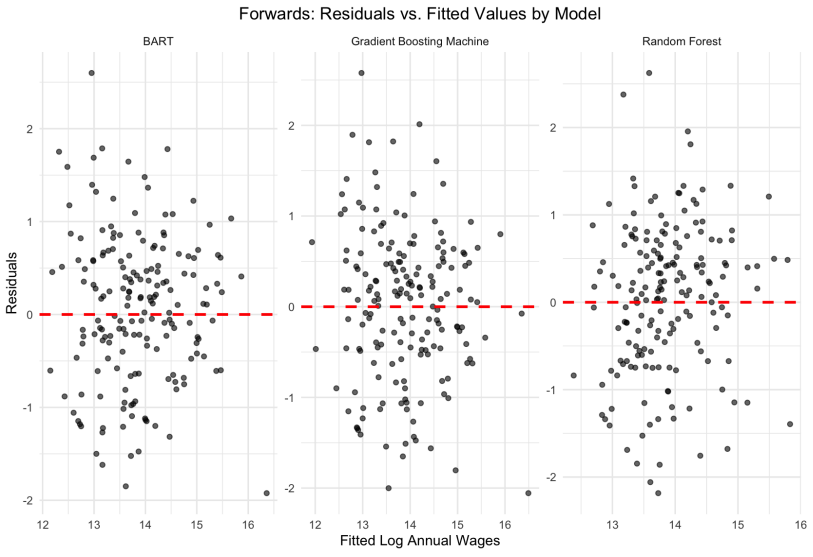

Figure A3: This plot shows the residuals (prediction errors) versus the predicted forward wages for three models: BART, GBM, and Random Forest. The horizontal dashed line at zero indicates perfect predictions. Points above or below the line show over- or under-predictions, respectively, with more spread indicating greater error variance in the model.

Table A4.

Top 10 Features and P-values by Quantile

Quantile Regression Model Results

Feature	Coefficient	P-value	Quantile
Quantile 0.75			
Intercept	11.6438	0.0000	0.75
Non-penalty Expected Goals (xG) per Shot	−0.9459	0.3395	0.75
Competition: Premier League	0.7136	0.0000	0.75
xG +- (Expected Goals - Expected Goals Against) per 90 minutes	0.5594	0.0000	0.75
+- (Goals scored - Goals allowed) while on pitch	−0.5122	0.0000	0.75
Competition: Ligue 1	−0.4977	0.0000	0.75
Nationality: Brazil	0.4227	0.0245	0.75
Points per Match (while on pitch)	0.3553	0.0000	0.75
Discipline: 2 Yellow Cards	0.3057	0.1539	0.75
Expected Assisted Goals (xAG) per 90 Minutes	0.1777	0.3106	0.75
Quantile 0.50			
Intercept	11.0973	0.0000	0.50
Non-penalty Expected Goals (xG) per Shot	−1.3902	0.0470	0.50
Competition: Premier League	0.9180	0.0000	0.50
xG +- (Expected Goals - Expected Goals Against) per 90 minutes	0.6455	0.0000	0.50
+- (Goals scored - Goals allowed) while on pitch	−0.5996	0.0000	0.50
Competition: Ligue 1	−0.3404	0.0000	0.50
Expected Assisted Goals (xAG) per 90 Minutes	0.2948	0.0001	0.50
Points per Match (while on pitch)	0.2545	0.0002	0.50
Discipline: 2 Yellow Cards	0.2137	0.1358	0.50
Nationality: Brazil	0.1802	0.3138	0.50
Quantile 0.25			
Intercept	10.5744	0.0000	0.25
Non-penalty Expected Goals (xG) per Shot	−0.9517	0.1872	0.25
Competition: Premier League	0.9259	0.0000	0.25
xG +- (Expected Goals - Expected Goals Against) per 90 minutes	0.5176	0.0000	0.25
Expected Assisted Goals (xAG) per 90 Minutes	0.4800	0.1169	0.25
+- (Goals scored - Goals allowed) while on pitch	−0.4608	0.0000	0.25
Competition: Ligue 1	−0.4185	0.0000	0.25
Nationality: Brazil	0.3360	0.0480	0.25
Discipline: 2 Yellow Cards	0.2446	0.2534	0.25
Points per Match (while on pitch)	0.2221	0.0033	0.25

Table A4: displays the top 10 features influencing Forward's wages across different quantiles, along with their coefficients and p-values from the Quantile Regression model. The quantiles represent different points in the wage distribution (0.25, 0.50, and 0.75), highlighting how the impact of specific features varies across different wage levels. Coefficients indicate the direction and magnitude of the effect, while p-values assess the statistical significance of each feature's impact within each quantile.

Code

R.

Pulling variety of performance stats for each player in top 5 leagues and creating combined stats dataset:

```
library(worldfootballR)
library(dplyr)

# Install and load worldfootballR package
install.packages("worldfootballR")
library(worldfootballR)

standrd <- fb_big5_advanced_season_stats(season_end_year=2024,stat_type="standard",team_or_player="player")
View(standrd)
str(standrd)
summary(standrd)
unique(standrd$Squad)

# Save the dataset as a CSV file
write.csv(standrd, "big5_2023_2024_standard_stats.csv", row.names = FALSE)

# Repeat steps for all stat_types

# Define the list of statistic types
stat_types <- c("shooting", "passing", "passing_types", "gca",
                "defense", "possession", "playing_time", "misc", "keepers", "keepers_adv")

# Define the function to fetch and save data
fetch_and_save_stats <- function(stat_type, season_end_year = 2024) {
  data <- fb_big5_advanced_season_stats(season_end_year = season_end_year, stat_type = stat_type, team_or_player =
"player")
  filename <- paste0("big5_2023_2024_", stat_type, "_stats.csv")
  write.csv(data, filename, row.names = FALSE)
  cat("Saved:", filename, "\n")
}

# Function to read CSV and standardize column types
read_and_standardize <- function(file_path) {
  df <- read.csv(file_path)
  df %>%
    mutate(across(where(is.factor), as.character)) %>%  # Convert factors to characters
    mutate(across(where(is.character), as.character)) %>%  # Convert characters to characters (ensures consistency)
    mutate(across(where(is.integer), as.numeric))  # Convert integers to numeric
}

# Read CSV files into separate data frames
standard_stats <- read_and_standardize("big5_2023_2024_standard_stats.csv")
shooting_stats <- read_and_standardize("big5_2023_2024_shooting_stats.csv")
passing_stats <- read_and_standardize("big5_2023_2024_passing_stats.csv")
passing_types_stats <- read_and_standardize("big5_2023_2024_passing_types_stats.csv")
```

```
gca_stats <- read_and_standardize("big5_2023_2024_gca_stats.csv")
defense_stats <- read_and_standardize("big5_2023_2024_defense_stats.csv")
possession_stats <- read_and_standardize("big5_2023_2024_possession_stats.csv")
playing_time_stats <- read_and_standardize("big5_2023_2024_playing_time_stats.csv")
misc_stats <- read_and_standardize("big5_2023_2024_misc_stats.csv")
keepers_stats <- read_and_standardize("big5_2023_2024_keepers_stats.csv")
keepers_adv_stats <- read_and_standardize("big5_2023_2024_keepers_adv_stats.csv")

combined_stats <- standard_stats %>%
  + left_join(keepers_adv_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born"))
%>%
  + left_join(passing_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born")) %>%
  + left_join(possession_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born"))
%>%
  + left_join(playing_time_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born"))
%>%
  + left_join(keepers_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born")) %>%
  + left_join(shooting_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born"))
%>%
  + left_join(gca_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born")) %>%
  + left_join(misc_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born")) %>%
  + left_join(passing_types_stats, by = c("Season_End_Year", "Squad", "Comp", "Player", "Nation", "Pos", "Age", "Born"))

# List of column names
column_names <- names(combined_stats)

# Identify columns with .x or .y suffixes
columns_to_remove <- column_names[grepl("\\.x$|\\.y$", column_names)]

# Remove the columns with .x and .y suffixes
cleaned_stats <- combined_stats %>%
  select(-all_of(columns_to_remove))

# Check the cleaned dataset
print(names(cleaned_stats))

# Write the cleaned dataset to a new CSV file
write.csv(cleaned_stats, "combined_stats.csv", row.names = FALSE)
```

Scraping injury data until 2022/23 for each player in European top 5 leagues:

```
# Load necessary libraries
library(rvest)
library(dplyr)
library(stringr)
library(worldfootballR)

# Define the Premier League team URLs
prem_team_urls <- c(
  "https://www.transfermarkt.com/manchester-city/startseite/verein/281/saison_id/2023",
  "https://www.transfermarkt.com/fc-arsenal/startseite/verein/11/saison_id/2023",
  "https://www.transfermarkt.com/fc-chelsea/startseite/verein/631/saison_id/2023",
  "https://www.transfermarkt.com/fc-liverpool/startseite/verein/31/saison_id/2023",
  "https://www.transfermarkt.com/tottenham-hotspur/startseite/verein/148/saison_id/2023",
  "https://www.transfermarkt.com/manchester-united/startseite/verein/985/saison_id/2023",
```

```
  "https://www.transfermarkt.com/aston-villa/startseite/verein/405/saison_id/2023",
  "https://www.transfermarkt.com/newcastle-united/startseite/verein/762/saison_id/2023",
  "https://www.transfermarkt.com/brighton-amp-hove-albion/startseite/verein/1237/saison_id/2023",
  "https://www.transfermarkt.com/crystal-palace/startseite/verein/873/saison_id/2023",
  "https://www.transfermarkt.com/west-ham-united/startseite/verein/379/saison_id/2023",
  "https://www.transfermarkt.com/fc-brentford/startseite/verein/1148/saison_id/2023",
  "https://www.transfermarkt.com/nottingham-forest/startseite/verein/703/saison_id/2023",
  "https://www.transfermarkt.com/afc-bournemouth/startseite/verein/989/saison_id/2023",
  "https://www.transfermarkt.com/wolverhampton-wanderers/startseite/verein/543/saison_id/2023",
  "https://www.transfermarkt.com/fc-everton/startseite/verein/29/saison_id/2023",
  "https://www.transfermarkt.com/fc-fulham/startseite/verein/931/saison_id/2023",
  "https://www.transfermarkt.com/fc-burnley/startseite/verein/1132/saison_id/2023",
  "https://www.transfermarkt.com/sheffield-united/startseite/verein/350/saison_id/2023",
  "https://www.transfermarkt.com/luton-town/startseite/verein/1031/saison_id/2023"
)

# Function to get player URLs for a given team URL
get_team_player_urls <- function(team_url) {
  page <- read_html(team_url)
  player_urls <- page %>%
    html_nodes(xpath = "//a[contains(@href, '/spieler/')]") %>%
    html_attr("href") %>%
    unique() %>%
    paste0("https://www.transfermarkt.com", .)
  player_urls <- player_urls[grepl("/profil/spieler/", player_urls)]
  return(player_urls)
}

# Function to get injury history for a given player URL
get_player_injuries <- function(player_url) {
  tryCatch({
    injury_history <- tm_player_injury_history(player_url = player_url)
    injury_history$player_url <- player_url
    return(injury_history)
  }, error = function(e) {
    message(paste("Error fetching data for player URL:", player_url))
    return(NULL)
  })
}

# Function to extract player name from URL
extract_player_name_from_url <- function(url) {
  name_part <- str_extract(url, "(?<=/spieler/)[^/]+")
  name_part <- gsub("-", " ", name_part)
  name_part <- str_to_title(name_part)
  return(name_part)
}

# Loop through all Premier League team URLs and get player injury data
all_player_injuries <- list()

for (team_url in prem_team_urls) {
  # Extract player URLs for the team
  player_urls <- get_team_player_urls(team_url)

  # Get injury history for all players in the team
  team_injuries <- lapply(player_urls, get_player_injuries)
```

```
  # Combine all players' injury data into one dataframe for the team
  team_injuries_df <- do.call(rbind, team_injuries)

  # Filter out injuries from the 23/24 season
  team_injuries_df <- team_injuries_df %>% filter(!str_detect(season_injured, "23/24"))

  # Check if player_name is a number and replace it with name from URL if necessary
  team_injuries_df <- team_injuries_df %>%
    mutate(player_name = if_else(grepl("^\\d+$", player_name), extract_player_name_from_url(player_url), player_name))

  # Append the team's injury data to the list
  all_player_injuries <- append(all_player_injuries, list(team_injuries_df))
}

# Combine all teams' injury data into one dataframe
all_injury_data_df <- do.call(rbind, all_player_injuries)

# Function to clean player names
clean_player_name <- function(name) {
  clean_name <- str_replace(name, "^#\\d+\\s*", "")
  return(clean_name)
}

# Apply the function to the player_name column
all_injury_data_df <- all_injury_data_df %>%
  mutate(player_name = clean_player_name(player_name))
View(all_injury_data_df)

# Summarize the total injury duration for each player
summarized_injuries <- all_injury_data_df %>%
  mutate(duration_days = as.numeric(str_extract(duration, "\\d+"))) %>%
  group_by(player_name) %>%
  summarize(Career_Injuries_Duration = sum(duration_days, na.rm = TRUE))

# Print summarized injury data
print(summarized_injuries)
View(summarized_injuries)

premierleague_injury_data <- all_injury_data_df
```

(Code above was repeated for the other 4 leagues) Now combining injury data:

```
# Combine dataframes of big 5 league injuries
combined_injury_data <- bind_rows(
  premierleague_injury_data,
  laliga_injury_data,
  seriea_injury_data,
  bundesliga_injury_data,
  ligue1_injury_data
)

# Print the combined injury data
print(combined_injury_data)
View(combined_injury_data)

#Save the combined injury data to a CSV file
```

```
write.csv(combined_injury_data, "/Users/willz/Desktop/Exeter/Dissertation/Football/combined_injury_data.csv", row.names
= FALSE)

library(stringr)


# Function to clean player names by extracting only the letters
clean_player_name <- function(name) {
  clean_name <- str_replace_all(name, "[^A-Za-z ]", "")
  clean_name <- str_trim(clean_name)
  return(clean_name)
}

# Summarize the total injury duration and games missed for each player
summarized_injuries <- combined_injury_data %>%
  mutate(duration_days = as.numeric(str_extract(duration, "\\d+")),
         games_missed = as.numeric(str_extract(games_missed, "\\d+"))) %>%
  group_by(player_name) %>%
  summarize(Career_Injuries_Duration = sum(duration_days, na.rm = TRUE),
            Career_Games_Missed = sum(games_missed, na.rm = TRUE))

# Apply the function to clean the player names in the summarized data
summarized_injuries <- summarized_injuries %>%
  mutate(player_name = clean_player_name(player_name))

# Save the cleaned summarized injury data to a CSV file
write.csv(summarized_injuries, "/Users/willz/Desktop/Exeter/Dissertation/Football/summarized_injury_data.csv",
row.names = FALSE)
```

Python.

Scraping to collect wages for all players in European top 5 leagues:

```
# Load pandas library
import pandas as pd

# Scrape wages of each league from fbref.com using read_html
premdf = pd.read_html('https://fbref.com/en/comps/9/wages/Premier-League-Wages',
                      attrs={"id":"player_wages"})[0]
ligadf = pd.read_html('https://fbref.com/en/comps/12/wages/La-Liga-Wages',
                      attrs={"id":"player_wages"})[0]
liguedf = pd.read_html('https://fbref.com/en/comps/13/wages/Ligue-1-Wages',
                      attrs={"id":"player_wages"})[0]
bundadf = pd.read_html('https://fbref.com/en/comps/20/wages/Bundesliga-Wages',
                      attrs={"id":"player_wages"})[0]
seriedf = pd.read_html('https://fbref.com/en/comps/11/wages/Serie-A-Wages',
                      attrs={"id":"player_wages"})[0]

# Combine the DataFrames
combined_wages = pd.concat([premdf, ligadf, liguedf, bundadf, seriedf], ignore_index=True)
combined_wages.describe()

# Save the combined DataFrame to a CSV file
combined_wages_path = '/Users/willz/Desktop/Exeter/Dissertation/Football/combined_wages.csv'
combined_wages.to_csv(combined_wages_path, index=False)

# Combine all wage dataframes
wage_dfs = [premdf, ligadf, liguedf, bundadf, seriedf]
```

```
combined_wages = pd.concat(wage_dfs, ignore_index=True)

# Save the combined wages dataframe to a CSV file
combined_wages.to_csv('combined_wages.csv', index=False)

# Load the cleaned stats dataframe
cleaned_stats = pd.read_csv('/Users/willz/Desktop/Exeter/Dissertation/Football/combined_stats.csv')
cleaned_stats.describe()

# Perform the merge on common columns
merged_df = pd.merge(cleaned_stats, combined_wages, on=["Player", "Squad", "Pos", "Age"], how="left")

# Drop extra columns
merged_stats = merged_df
merged_stats = merged_stats.drop('Nation_y', axis=1).rename(columns=
                                        {'Nation_x': 'Nation',
                                        'Weekly Wages':'Weekly_Wages',
                                        'Annual Wages': 'Annual_Wages'
                                        })

# Load library
import re

allstats = merged_stats

# Function to extract pound values from the wage strings
def extract_pound_wages(wage_str):
    if pd.isna(wage_str):
        return None
    pound_match = re.search(r'£ ([\d,]+)', wage_str)
    pound_value = float(pound_match.group(1).replace(',', '')) if pound_match else None
    return pound_value

# Apply the extraction function to the wage columns
allstats['Weekly_Wages_Pounds'] = allstats['Weekly_Wages'].apply(extract_pound_wages)
allstats['Annual_Wages_Pounds'] = allstats['Annual_Wages'].apply(extract_pound_wages)

allstats = allstats.drop(columns=['Weekly_Wages','Annual_Wages','Notes']).rename(columns={'Rk':'Rank'})

# Load injury data to add to combined stats
injurydata = pd.read_csv('/Users/willz/Desktop/Exeter/Dissertation/Football/summarized_injury_data.csv')
injurydata.describe()

# Merge the dataframes on the player_name column from injury data and Player column from finalstats
finalstats = pd.merge(allstats, injurydata, how='left', left_on='Player', right_on='player_name')
finalstats.head()

# Rename 'Player_x' to 'Player'
merged_data = merged_data.rename(columns={'Player_x': 'Player'})

# Remove the wage columns from merged_data
merged_data = merged_data.drop(columns=['Weekly_Wages_Pounds', 'Annual_Wages_Pounds'], errors='ignore')

# List of player names from combined_wages
wages_names = combined_wages['Player'].tolist()

# Apply fuzzy matching
cutoff = 100
```

```
merged_data['matched_name'] = merged_data.apply(
    lambda row: match_names(row, 'Player', wages_names, fuzz.ratio, cutoff), axis=1)

# Merge the dataframes using the matched names
merged_data = pd.merge(merged_data, combined_wages, left_on='matched_name', right_on='Player', how='left')

# Clean up column names and remove duplicate columns if any
merged_data = merged_data.drop(['Player_y', 'Nation_y'], axis=1, errors='ignore').rename(columns={
    'Player_x': 'Player',
    'Nation_x': 'Nation',
    'Weekly Wages': 'Weekly_Wages',
    'Annual Wages': 'Annual_Wages'
})

# Extract and convert wage values from strings to numeric values
def extract_pound_wages(wage_str):
    if pd.isna(wage_str):
        return None
    pound_match = re.search(r'£ ([\d,]+)', wage_str)
    pound_value = float(pound_match.group(1).replace(',', '')) if pound_match else None
    return pound_value

merged_data['Weekly_Wages_Pounds'] = merged_data['Weekly_Wages'].apply(extract_pound_wages)
merged_data['Annual_Wages_Pounds'] = merged_data['Annual_Wages'].apply(extract_pound_wages)

# Drop the original wage string columns and other unnecessary columns
merged_data = merged_data.drop(columns=['Weekly_Wages', 'Annual_Wages', 'Notes', 'matched_name'],
errors='ignore').rename(columns={'Rk':'Rank'})

# Create a copy of the merged_data dataframe
dataforanalysis = merged_data.copy()

# Separate dataframes for GK and outfield players

# Define the columns specific to goalkeepers
gk_columns = [
    'GA_Goals', 'PKA_Goals', 'FK_Goals', 'CK_Goals', 'OG_Goals', 'PSxG_Expected',
    'PSxG_per_SoT_Expected', 'PSxG._per__minus__Expected', 'X_per_90_Expected',
    'Cmp_Launched', 'Att_Launched', 'Cmp_percent_Launched', 'Att..GK._Passes',
    'Thr_Passes', 'Launch_percent_Passes', 'AvgLen_Passes', 'Att_Goal',
    'Launch_percent_Goal', 'AvgLen_Goal', 'Opp_Crosses', 'Stp_Crosses',
    'Stp_percent_Crosses', 'X.OPA_Sweeper', 'X.OPA_per_90_Sweeper', 'AvgDist_Sweeper',
    'GA', 'GA90', 'SoTA', 'Saves', 'Save_percent', 'W', 'D', 'L', 'CS', 'CS_percent',
    'PKatt_Penalty', 'PKA_Penalty', 'PKsv_Penalty', 'PKm_Penalty', 'Save_percent_Penalty'
]

# Identify the columns for wages and injuries
wages_injuries_columns = ['Weekly_Wages_Pounds', 'Annual_Wages_Pounds', 'Career_Injuries_Duration',
'Career_Games_Missed']

# Identify the basic player information columns
basic_info_columns = ['Player', 'Squad', 'Age', 'Pos', 'Nation', 'Mins_Per_90_Playing']

# Combine all the necessary columns for goalkeepers
gk_all_columns = basic_info_columns + gk_columns + wages_injuries_columns

# Create a boolean mask for goalkeepers
gk_mask = dataforanalysis['Pos'] == 'GK'
```

```
# Create the GK dataframe with all relevant columns
gk_data = dataforanalysis[gk_mask][gk_all_columns]

# Create the non-GK dataframe with all other columns
outfield_data = dataforanalysis[~gk_mask].drop(columns=gk_columns)

# Save the new dataframes as CSV files
gk_data.to_csv('gk_data.csv', index=False)
outfield_data.to_csv('outfield_data.csv', index=False)

# Define the columns specific to goalkeepers
gk_columns = [
    'GA_Goals', 'PKA_Goals', 'FK_Goals', 'CK_Goals', 'OG_Goals', 'PSxG_Expected',
    'PSxG_per_SoT_Expected', 'PSxG._per__minus__Expected', 'X_per_90_Expected',
    'Cmp_Launched', 'Att_Launched', 'Cmp_percent_Launched', 'Att..GK._Passes',
    'Thr_Passes', 'Launch_percent_Passes', 'AvgLen_Passes', 'Att_Goal',
    'Launch_percent_Goal', 'AvgLen_Goal', 'Opp_Crosses', 'Stp_Crosses',
    'Stp_percent_Crosses', 'X.OPA_Sweeper', 'X.OPA_per_90_Sweeper', 'AvgDist_Sweeper',
    'GA', 'GA90', 'SoTA', 'Saves', 'Save_percent', 'W', 'D', 'L', 'CS', 'CS_percent',
    'PKatt_Penalty', 'PKA_Penalty', 'PKsv_Penalty', 'PKm_Penalty', 'Save_percent_Penalty'
]

# Identify the columns for wages and injuries
wages_injuries_columns = ['Weekly_Wages_Pounds', 'Annual_Wages_Pounds', 'Career_Injuries_Duration',
'Career_Games_Missed']

# Identify the basic player information columns
basic_info_columns = ['Player', 'Squad', 'Age', 'Pos', 'Nation', 'Season_End_Year', 'Comp', 'Born', 'Mins_Per_90_Playing']

# Combine all the necessary columns for goalkeepers
gk_all_columns = basic_info_columns + gk_columns + wages_injuries_columns

# Create a boolean mask for goalkeepers
gk_mask = dataforanalysis_cleaned['Pos'] == 'GK'

# Create the GK dataframe with all relevant columns
gk_data = dataforanalysis_cleaned[gk_mask][gk_all_columns]

# Create the non-GK dataframe with all other columns
outfield_data = dataforanalysis_cleaned[~gk_mask].drop(columns=gk_columns)

# Save the new dataframes as CSV files
gk_data.to_csv('gk_data.csv', index=False)
outfield_data.to_csv('outfield_data.csv', index=False)

from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer

# Remove rows with missing Nation or Age in outfield_data
outfield_data = outfield_data.dropna(subset=['Nation', 'Age'])

# Define the imputer
iterative_imputer = IterativeImputer()

# Impute missing values in numeric columns for gk_data
gk_numeric_columns = gk_data.select_dtypes(include=['float64', 'int64']).columns
imputed_gk = gk_data.copy()
```

```
imputed_gk[gk_numeric_columns] = iterative_imputer.fit_transform(gk_data[gk_numeric_columns])

# Impute missing values in numeric columns for outfield_data
outfield_numeric_columns = outfield_data.select_dtypes(include=['float64', 'int64']).columns
imputed_outfield = outfield_data.copy()
imputed_outfield[outfield_numeric_columns] = iterative_imputer.fit_transform(outfield_data[outfield_numeric_columns])

# Save the imputed datasets with new names
imputed_gk.to_csv('imputed_gk.csv', index=False)
imputed_outfield.to_csv('imputed_outfield.csv', index=False)
```

R.

Data pre-processing & Random Forest, GBM, BART modelling (Forwards example):

```
## ATTACKERS
# Load necessary libraries
library(dplyr)
library(caret)
library(rJava)

# Set Java memory for bartMachine (adjust as needed)
options(java.parameters = "-Xmx16g")
library(bartMachine)
library(readr)

# Load the dataset
imputed_outfield <- read_csv("imputed_outfield.csv")

# Define the features relevant to attackers
attacker_features <- c(
  'Season_End_Year', 'Squad', 'Comp', 'Player', 'Nation', 'Pos', 'Age', 'Born', 'Mins_Per_90_Playing', 'Gls', 'G.A',
  'G_minus_PK', 'PK', 'PKatt', 'xAG_Expected', 'npxG.xAG_Expected', 'PrgC_Progression', 'PrgP_Progression',
  'PrgR_Progression', 'Gls_Per', 'Ast_Per', 'G.A_Per', 'G_minus_PK_Per', 'G.A_minus_PK_Per', 'xG_Per', 'xAG_Per',
  'xG.xAG_Per', 'npxG_Per', 'npxG.xAG_Per', 'Cmp_Total', 'Att_Total', 'Cmp_percent_Total', 'TotDist_Total',
  'PrgDist_Total', 'Cmp_Short', 'Att_Short', 'Cmp_percent_Short', 'Cmp_Medium', 'Att_Medium', 'Cmp_percent_Medium',
  'Cmp_Long', 'Att_Long', 'Cmp_percent_Long', 'xAG', 'xA_Expected', 'A_minus_xAG_Expected', 'KP', 'Final_Third',
  'PPA', 'CrsPA', 'PrgP', 'Touches_Touches', 'Def.Pen_Touches', 'Def.3rd_Touches', 'Mid.3rd_Touches',
  'Att.3rd_Touches', 'Att.Pen_Touches', 'Live_Touches', 'Att_Take', 'Succ_Take', 'Succ_percent_Take', 'Tkld_Take',
  'Tkld_percent_Take', 'Carries_Carries', 'TotDist_Carries', 'PrgDist_Carries', 'PrgC_Carries', 'Final_Third_Carries',
  'CPA_Carries', 'Mis_Carries', 'Dis_Carries', 'Rec_Receiving', 'PrgR_Receiving', 'MP_Playing.Time',
  'Min_Playing.Time', 'Mn_per_MP_Playing.Time', 'Min_percent_Playing.Time', 'Mins_Per_90_Playing.Time',
  'Starts_Starts', 'Mn_per_Start_Starts', 'Compl_Starts', 'Subs_Subs', 'Mn_per_Sub_Subs', 'unSub_Subs',
  'PPM_Team.Success', 'onG_Team.Success', 'onGA_Team.Success', 'plus_per__minus__Team.Success',
  'plus_per__minus_90_Team.Success', 'On_minus_Off_Team.Success', 'onxG_Team.Success..xG.',
  'onxGA_Team.Success..xG', 'xGplus_per__minus__Team.Success..xG', 'xGplus_per__minus_90_Team.Success..xG',
  'On_minus_Off_Team.Success..xG', 'Gls_Standard', 'Sh_Standard', 'SoT_Standard', 'SoT_percent_Standard',
  'Sh_per_90_Standard', 'SoT_per_90_Standard', 'G_per_Sh_Standard', 'G_per_SoT_Standard', 'Dist_Standard',
  'FK_Standard', 'PK_Standard', 'PKatt_Standard', 'npxG_per_Sh_Expected', 'G_minus_xG_Expected',
  'np.G_minus_xG_Expected', 'SCA_SCA', 'SCA90_SCA', 'PassLive_SCA', 'PassDead_SCA', 'TO_SCA', 'Sh_SCA',
'Fld_SCA',
  'Def_SCA', 'GCA_GCA', 'GCA90_GCA', 'PassLive_GCA', 'PassDead_GCA', 'TO_GCA', 'Sh_GCA', 'Fld_GCA',
'Def_GCA',
  'X2CrdY', 'Fls', 'Fld', 'Off', 'Crs', 'Int', 'TklW', 'PKwon', 'PKcon', 'OG', 'Recov', 'Won_Aerial', 'Lost_Aerial',
  'Won_percent_Aerial', 'Att', 'Live_Pass', 'Dead_Pass', 'FK_Pass', 'TB_Pass', 'Sw_Pass', 'Crs_Pass', 'TI_Pass',
  'CK_Pass', 'In_Corner', 'Out_Corner', 'Str_Corner', 'Cmp_Outcomes', 'Off_Outcomes', 'Blocks_Outcomes',
  'Career_Injuries_Duration', 'Career_Games_Missed', 'Weekly_Wages_Pounds', 'Annual_Wages_Pounds', 'Row_Number'
)
```

```
# Create attackers DataFrame with the relevant features
attackers_df <- imputed_outfield %>%
  filter(grepl('FW', Pos)) %>%
  select(all_of(attacker_features), Annual_Wages_Pounds, Comp, Nation)

# Log-transform the Annual_Wages_Pounds (target variable)
attackers_df$log_Annual_Wages <- log(attackers_df$Annual_Wages_Pounds)

# Drop the original Annual_Wages_Pounds from the predictors
attackers_df <- attackers_df %>% select(-Annual_Wages_Pounds)

# Encode categorical variables
attackers_df$Comp <- as.factor(attackers_df$Comp)
attackers_df$Nation <- as.factor(attackers_df$Nation)

# One-hot encode the categorical variables
dummies_att <- dummyVars(~ Comp + Nation, data = attackers_df)
encoded_df_att <- predict(dummies_att, newdata = attackers_df)

# Convert to dataframe
encoded_df_att <- as.data.frame(encoded_df_att)

# Rename columns to replace any problematic characters
colnames(encoded_df_att) <- make.names(colnames(encoded_df_att), unique = TRUE)

# Combine with the original data (excluding Comp and Nation)
attackers_encoded <- cbind(encoded_df_att, attackers_df %>% select(-Comp, -Nation))

# Drop irrelevant columns that should not be predictors
irrelevant_columns <- c('Pos', 'Squad','Row_Number', 'Weekly_Wages_Pounds',
                        'Born', 'Season_End_Year', 'Player', 'log_Annual_Wages')
attackers_encoded <- attackers_encoded %>% select(-all_of(irrelevant_columns))

# Add back the target variable (log_Annual_Wages)
attackers_encoded$log_Annual_Wages <- attackers_df$log_Annual_Wages

# Split the data into training and testing sets
set.seed(123)
train_indices_att <- sample(seq_len(nrow(attackers_encoded)), size = 0.8 * nrow(attackers_encoded))
train_data_att <- attackers_encoded[train_indices_att, ]
test_data_att <- attackers_encoded[-train_indices_att, ]

# Step 3: Identify top features but exclude the target variable

# Run the Random Forest model to get feature importance
rf_att_model <- randomForest(log_Annual_Wages ~ ., data = train_data_att, importance = TRUE)

# Extract feature importance
rf_importance <- importance(rf_att_model, type = 1)
rf_importance_df <- data.frame(Feature = rownames(rf_importance), Importance = rf_importance[, 1])

# Sort features by importance
rf_importance_df <- rf_importance_df[order(-rf_importance_df$Importance), ]

# Initialize variables to store the best performance
best_r2_rf <- -Inf
best_r2_gbm <- -Inf
```

```
best_r2_bart <- -Inf
best_num_features_rf <- 0
best_num_features_gbm <- 0
best_num_features_bart <- 0
best_rf_model <- NULL
best_gbm_model <- NULL
best_bart_model <- NULL

# Loop through subsets of features for each model
for (num_features in seq(10, nrow(rf_importance_df), by = 5)) {

 # Select top `num_features` based on importance
 top_features <- rf_importance_df$Feature[1:num_features]

 # Subset the training and testing data
 train_data_subset <- train_data_att[, c(top_features, "log_Annual_Wages")]
 test_data_subset <- test_data_att[, c(top_features, "log_Annual_Wages")]

 # Random Forest Model
 rf_model_subset <- randomForest(log_Annual_Wages ~ ., data = train_data_subset, mtry = rf_grid_search$bestTune$mtry)
 rf_predictions_subset <- predict(rf_model_subset, newdata = test_data_subset)
 r2_rf <- cor(test_data_subset$log_Annual_Wages, rf_predictions_subset)^2

 if (r2_rf > best_r2_rf) {
   best_r2_rf <- r2_rf
   best_num_features_rf <- num_features
   best_rf_model <- rf_model_subset
 }

 # GBM Model
 gbm_model_subset <- gbm(
   log_Annual_Wages ~ .,
   data = train_data_subset,
   n.trees = gbm_grid_search$bestTune$n.trees,
   interaction.depth = gbm_grid_search$bestTune$interaction.depth,
   shrinkage = gbm_grid_search$bestTune$shrinkage,
   n.minobsinnode = gbm_grid_search$bestTune$n.minobsinnode,
   distribution = "gaussian",
   verbose = FALSE
 )
 gbm_predictions_subset <- predict(gbm_model_subset, newdata = test_data_subset, n.trees =
gbm_grid_search$bestTune$n.trees)
 r2_gbm <- cor(test_data_subset$log_Annual_Wages, gbm_predictions_subset)^2

 if (r2_gbm > best_r2_gbm) {
   best_r2_gbm <- r2_gbm
   best_num_features_gbm <- num_features
   best_gbm_model <- gbm_model_subset
 }

 # BART Model
 bart_model_subset <- bartMachine(
   X = train_data_subset[, top_features],
   y = train_data_subset$log_Annual_Wages,
   num_trees = best_params_bart$num_trees,
   num_burn_in = best_params_bart$num_burn_in,
   num_iterations_after_burn_in = best_params_bart$num_iterations_after_burn_in,
   use_missing_data = TRUE
```

```
  )
  bart_predictions_subset <- predict(bart_model_subset, new_data = test_data_subset[, top_features])
  r2_bart <- cor(test_data_subset$log_Annual_Wages, bart_predictions_subset)^2

  if (r2_bart > best_r2_bart) {
    best_r2_bart <- r2_bart
    best_num_features_bart <- num_features
    best_bart_model <- bart_model_subset
  }
}

# Print the optimal number of features and the corresponding R² for each model
cat("Best number of features for Random Forest:", best_num_features_rf, "\n")
cat("Best R² for Random Forest:", best_r2_rf, "\n")

cat("Best number of features for GBM:", best_num_features_gbm, "\n")
cat("Best R² for GBM:", best_r2_gbm, "\n")

cat("Best number of features for BART:", best_num_features_bart, "\n")
cat("Best R² for BART:", best_r2_bart, "\n")

# Rebuild the best models using the optimal number of features
top_features_rf <- rf_importance_df$Feature[1:best_num_features_rf]
train_data_best_rf <- train_data_att[, c(top_features_rf, "log_Annual_Wages")]
test_data_best_rf <- test_data_att[, c(top_features_rf, "log_Annual_Wages")]

top_features_gbm <- rf_importance_df$Feature[1:best_num_features_gbm]
train_data_best_gbm <- train_data_att[, c(top_features_gbm, "log_Annual_Wages")]
test_data_best_gbm <- test_data_att[, c(top_features_gbm, "log_Annual_Wages")]

top_features_bart <- rf_importance_df$Feature[1:best_num_features_bart]
train_data_best_bart <- train_data_att[, c(top_features_bart, "log_Annual_Wages")]
test_data_best_bart <- test_data_att[, c(top_features_bart, "log_Annual_Wages")]

# Final models with the optimal number of features
final_rf_model <- randomForest(log_Annual_Wages ~ ., data = train_data_best_rf, mtry = rf_grid_search$bestTune$mtry)
final_gbm_model <- gbm(
  log_Annual_Wages ~ .,
  data = train_data_best_gbm,
  n.trees = gbm_grid_search$bestTune$n.trees,
  interaction.depth = gbm_grid_search$bestTune$interaction.depth,
  shrinkage = gbm_grid_search$bestTune$shrinkage,
  n.minobsinnode = gbm_grid_search$bestTune$n.minobsinnode,
  distribution = "gaussian",
  verbose = FALSE
)
final_bart_model <- bartMachine(
  X = train_data_best_bart[, top_features_bart],
  y = train_data_best_bart$log_Annual_Wages,
  num_trees = best_params_bart$num_trees,
  num_burn_in = best_params_bart$num_burn_in,
  num_iterations_after_burn_in = best_params_bart$num_iterations_after_burn_in,
  use_missing_data = TRUE
)

# Evaluate the final models
rf_att_predictions <- predict(final_rf_model, newdata = test_data_best_rf)
```

```
gbm_att_predictions <- predict(final_gbm_model, newdata = test_data_best_gbm, n.trees =
gbm_grid_search$bestTune$n.trees)
bart_att_predictions <- predict(final_bart_model, new_data = test_data_best_bart[, top_features_bart])

# R² for Random Forest on test data
rf_att_r_squared <- cor(test_data_best_rf$log_Annual_Wages, rf_att_predictions)^2
cat("Final R² for Random Forest:", rf_att_r_squared, "\n")

# Top 10 feature importances for Random Forest
rf_importance_final <- importance(final_rf_model, type = 1)
rf_importance_final_df <- rf_importance_final_df[order(-rf_importance_final_df$Importance), ]
top_10_rf_features <- head(rf_importance_final_df, 10)
cat("Top 10 Feature Importances for Random Forest:\n")
print(top_10_rf_features)

# R² for GBM on test data
gbm_att_r_squared <- cor(test_data_best_gbm$log_Annual_Wages, gbm_att_predictions)^2
cat("Final R² for GBM:", gbm_att_r_squared, "\n")

# Top 10 feature importances for GBM
gbm_importance_final <- summary(final_gbm_model, plot = FALSE)
top_10_gbm_features <- head(gbm_importance_final, 10)
cat("Top 10 Feature Importances for GBM:\n")
print(top_10_gbm_features)

# R² for BART on test data
bart_att_r_squared <- cor(test_data_best_bart$log_Annual_Wages, bart_att_predictions)^2
cat("Final R² for BART:", bart_att_r_squared, "\n")

# Top 10 feature importances for BART
bart_importance_final <- investigate_var_importance(final_bart_model)
top_10_bart_features <- head(bart_importance_final$avg_var_props, 10)
cat("Top 10 Feature Importances for BART:\n")
print(top_10_bart_features)
```